



HAL
open science

SARS-CoV-2-like viruses from captive Guangdong pangolins generate circular RNAs

Alexandre Hassanin, Huw Jones, Anne Ropiquet

► **To cite this version:**

Alexandre Hassanin, Huw Jones, Anne Ropiquet. SARS-CoV-2-like viruses from captive Guangdong pangolins generate circular RNAs. 2020. hal-02616966

HAL Id: hal-02616966

<https://hal.science/hal-02616966>

Preprint submitted on 25 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **SARS-CoV-2-like viruses from captive Guangdong pangolins generate circular RNAs**

2 Alexandre Hassanin^{1*}, Huw Jones² & Anne Ropiquet²

3

4 1. Institut de Systématique, Evolution, Biodiversité, UMR 7205 CNRS, MNHN, Sorbonne
5 Université, EPHE, Université des Antilles, Muséum National d'Histoire Naturelle, CP 51,
6 57 rue Cuvier, 75231 PARIS Cedex 05 France.

7 *Corresponding author: alexandre.hassanin@mnhn.fr

8

9 2. Department of Natural Sciences, Faculty of Science and Technology, Middlesex
10 University, The Burroughs, London NW4 4BT United Kingdom.

11

12 **Viruses closely related to SARS-CoV-2, the virus responsible for the Covid-19**
13 **pandemic, have previously been identified in several Sunda pangolins (*Manis javanica*)**
14 **seized in Guangdong and Guangxi provinces of China¹⁻⁴. The virus extracted from**
15 **Guangdong pangolins is particularly intriguing as the amino-acid sequence of the**
16 **receptor binding domain of the spike protein is nearly identical to that of the human**
17 **SARS-CoV-2 virus^{2,3}. This characteristic suggests it is capable of binding to the human**
18 **ACE2 receptor and may therefore be able to mediate infection of human cells.**

19 **Here, we assembled SARS-CoV-2-like genomes from the Sequence Read Archive**
20 **(SRA) data available in May 2020 for ten Guangdong pangolins. As previously**
21 **described⁵ the genome assemblies were of poor quality, having high levels of missing**
22 **data and with possible DNA contamination from human, mouse and tiger. We found 10**
23 **circular RNAs (circRNAs) of 278-776 nucleotides (nt) containing SARS-CoV-2-like**

24 **sequences in three pangolin lung samples. This is the first time this class of molecule is**
25 **reported to be generated from an RNA-genome virus.**

26

27 Next-generation sequencing (NGS) runs previously published for 16 pangolin samples
28 were analysed (**Table 1**)¹⁻³. A large number of SARS-CoV-2-like reads ($\geq 184,870$) were
29 found in SRA datasets generated from lungs and intestines of Guangxi pangolins, whereas a
30 limited number of reads ($\leq 1,078$) were found in SRA datasets generated from lungs of
31 Guangdong pangolins. These differences suggest higher RNA degradation in Guangdong
32 samples and explain why it was not possible to assemble a full virus genome from these data.
33 Assuming the same virus infected Guangdong pangolins^{3,4}, we pooled together the 4688
34 SARS-CoV-2-like reads found in the ten Guangdong datasets to assemble a consensus
35 sequence. The resulting genome remained partial, with 15 fragments and 4% missing data
36 (mean coverage 16.6X). Two recent studies^{3,4} had filled genome gaps by sequencing multiple
37 PCR products. However, these two published genomes differ at 15 nucleotide positions
38 (Supplementary Information) indicating the quality of Guangdong pangolin virus sequence
39 still needs to be improved⁵.

40 In three SRA datasets generated from lung samples of three Guangdong pangolins
41 (**Table 1**), we discovered several circRNAs by mapping NGS reads to the reference genome
42 (GISAID accession number EPI_ISL_410721³) using Geneious Prime® 2020.0.3 and the
43 “low sensitivity” option (maximum mismatch: 10%). For the *GD/P5L* dataset³ (**Table 1**),
44 most SARS-CoV-2-like reads (566 of all 600 reads) mapped to a small portion of the ORF1ab
45 gene (**Figure 1A**). However, the 5'-and 3'-flanking regions of the contig were found to
46 contain chimeric reads, in which the extremities could not be aligned to the reference genome.
47 A preliminary consensus sequence of 748 nt was extracted and then used as a reference to
48 extend the contig by mapping *GD/P5L* reads (three iterations in Geneious Prime® 2020.0.3).

49 The final contig contained 635 reads, and the consensus sequence of 862 nt was then
50 extracted for further analysis. As shown in **Figure 1B**, its flanking regions (highlighted in
51 yellow) are parts of an original circRNA of 733-nt long, herein referred to as *circRNA-*
52 *GDP5L-1*: the 5'-flanking region is identical to the 3'-end of *circRNA-GDP5L-1* (positions
53 685-733), whereas the 3'-flanking region is identical to the 5'-end of *circRNA-GDP5L-1*
54 (positions 1-80). Such repetitions of identical sequences are consistent with a circular nature
55 of the original RNA molecule. In support of this finding, the high number of reads found for
56 *circRNA-GDP5L-1* (635 reads) suggests the molecule was still circular before reverse
57 transcription (see protocol used for library construction³) as multiple cDNA copies of a single
58 circRNA can be generated through rolling circle amplification⁶. The structure of *circRNA-*
59 *GDP5L-1* is characterized by the presence of a 33-nt motif (positions 1-33, shown in red in
60 **Figure 1B**) which cannot be aligned to the reference genome, but which is the full reverse
61 complement of the sequence located 20-nt downstream (underlined sequence in **Figure 1B**).
62 Based on computer modeling tools⁷, we can predict *circRNA-GDP5L-1* to form a highly
63 stable secondary structure (Supplementary Information). We found nine other highly
64 structured circRNAs showing similar features: two further in the *GD/P5L* dataset, two in the
65 *GD/P7L* dataset, and five in the *GD/P8L* dataset. Their full sequences, characteristics and
66 secondary structures are detailed in Supplementary Information. The length of these
67 circRNAs varies between 278 and 776 nt. Only a single circRNA contained a viral sequence
68 without additional motif. The nine other circRNAs are composed of a viral genomic sequence
69 followed by the full reverse complement of a viral motif (length between 5 and 96 nt) found a
70 few nucleotides downstream (between 14 and 62 nt). Notably, a substantial proportion of
71 these genomic sequences are protein-coding-fragments (see details in Supplementary
72 Information).

73 CircRNAs form a large class of RNA molecules that are widespread in eukaryotes⁸.
74 They play important functions in gene regulation as they can have interactions with RNA-
75 binding proteins, serve as translation templates, and act as microRNA sponges or
76 transcriptional regulatory factors⁸⁻¹⁰. Recent studies have reported that DNA-virus genomes
77 are also capable of generating circRNAs^{11,12}. To our knowledge, these molecules have not
78 previously been described as by-products of RNA-virus genomes. It is therefore important to
79 know whether these circRNAs are natural or not, since some circRNAs have been described
80 as possible artifacts of the reverse transcription process¹³. In addition, an intriguing issue is
81 that the three SRA datasets in which circRNAs were found are the same three datasets
82 containing human and mouse DNA (**Table 1**)⁵. The comparison between NGS runs
83 performed on RNAs extracts from pangolin lungs (**Table 1**) clearly show much more SARS-
84 CoV-2-like reads for Guangxi datasets than for Guangdong datasets ($\geq 184,870$ versus
85 $\leq 1,078$). Two hypotheses can be formulated to explain these differences: Guangdong samples
86 contain lower concentrations of SARS-CoV-2-like virus; or RNA molecules extracted from
87 Guangdong samples were much more degraded due to poor sample preservation. In support of
88 the second hypothesis, it has been found that transcriptome degradation can result in circRNA
89 enrichment as linear RNAs decay more rapidly than circRNAs¹⁴, suggesting that the 10
90 circRNAs discovered in the three Guangdong pangolin samples are true by-products of
91 SARS-CoV-2 like RNA-genomes. Another argument for the natural origin of these circRNAs
92 is provided by the alignment of their 5'-end with homologous sequences extracted from
93 SARS-CoV-2 like genomes (Supplementary Information). In the case of *circRNA-GDP5L-1*,
94 the two additional base-pairing motifs involved in the secondary structure (highlighted in
95 green in **Figure 3C**) are found to be highly conserved among SARS-CoV-2-like viruses,
96 whereas the genomic fragment is a part of a protein-coding gene. In this context, the fact that
97 canonical base-pairing motifs have been restored through compensatory mutations in the virus

98 genome from Guangxi pangolins (position 2: G=>A and position 19: C=>T in **Figure 3C**) is
99 particularly relevant, as it suggests strong natural selection pressure for maintaining a RNA-
100 stem-loop (hairpin) structure. This hypothesis is also corroborated by the high conservation of
101 canonical base-pairing motifs in all other circRNAs and by the detection of similar
102 compensatory mutations in three other circRNAs (*circRNA-GDP7L-1*, *circRNA-GDP7L-2*,
103 and *circRNA-GDP8L-3*; Supplementary Information), suggesting that the hairpin is an
104 important structure that may be directly involved in the circularization process.

105 But what can the role of viral circRNAs be during host infection? It has been
106 demonstrated that viral infection triggers the degradation of host-produced circRNAs which
107 in turn activates the innate immune response¹⁵. This mechanism suggests viral circRNAs
108 could have an important role, either to delay the host's immune response, or to initiate viral
109 proliferation in host cells. Our discovery of circRNAs generated from pangolin SARS-CoV-2-
110 like virus opens up new perspectives for understanding virus-host interactions and has
111 possible implications in a wider range of RNA-virus infections including Covid-19.

112

113 **References**

- 114 1. Liu, P. *et al.* Viral Metagenomics Revealed Sendai Virus and Coronavirus Infection of
115 Malayan Pangolins (*Manis javanica*). *Viruses* **11**, (2019).
- 116 2. Lam, T.T. *et al.* Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins.
117 *Nature* doi: 10.1038/s41586-020-2169-0. (2020).
- 118 3. Xiao, K. *et al.* Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins.
119 *Nature* (2020).
- 120 4. Liu, P. *et al.* Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-
121 CoV-2)? *PLoS Pathog.* **16**, (2020).

- 122 5. Hassanin, A. The SARS-CoV-2-like virus found in captive pangolins from Guangdong
123 should be better sequenced. (submitted; doi: <https://doi.org/10.1101/2020.05.07.077016>).
- 124 6. Szabo, L., Salzman, J. Detecting circular RNAs: bioinformatic and experimental
125 challenges. *Nat Rev Genet.* **17**, (2016).
- 126 7. Gruber, A.R. *et al.* The ViennaRNA web services. *Methods Mol Biol.* **1269**, (2015).
- 127 8. Memczak, S. *et al.* Nature. Circular RNAs are a large class of animal RNAs with
128 regulatory potency. **495**, (2013).
- 129 9. Chen, L.L. *et al.* The expanding regulatory mechanisms and cellular functions of circular
130 RNAs. *Nat Rev Mol Cell Biol.* Online ahead of print (2020).
- 131 10. Xie, R. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency.
132 *Nature* **495**, (2020).
- 133 11. Ungerleider, N. *et al.* The Epstein Barr virus circRNAome. *PLoS Pathog.* **14**, (2018).
- 134 12. Toptan, T. *et al.* Circular DNA tumor viruses make circular RNAs. *Proc Natl Acad Sci U*
135 *S A* **115**, (2018).
- 136 13. Chen I, *et al.* Biogenesis, identification, and function of exonic circular RNAs. *Wiley*
137 *Interdiscip Rev RNA.* **6**, (2015).
- 138 14. Alhasan, A.A. *et al.* Circular RNA enrichment in platelets is a signature of transcriptome
139 degradation. *Blood* **127**, (2016).
- 140 15. Liu, C.X. *et al.* Structure and Degradation of Circular RNAs Regulate PKR Activation in
141 Innate Immunity. *Cell* **177**, (2019).

142

143

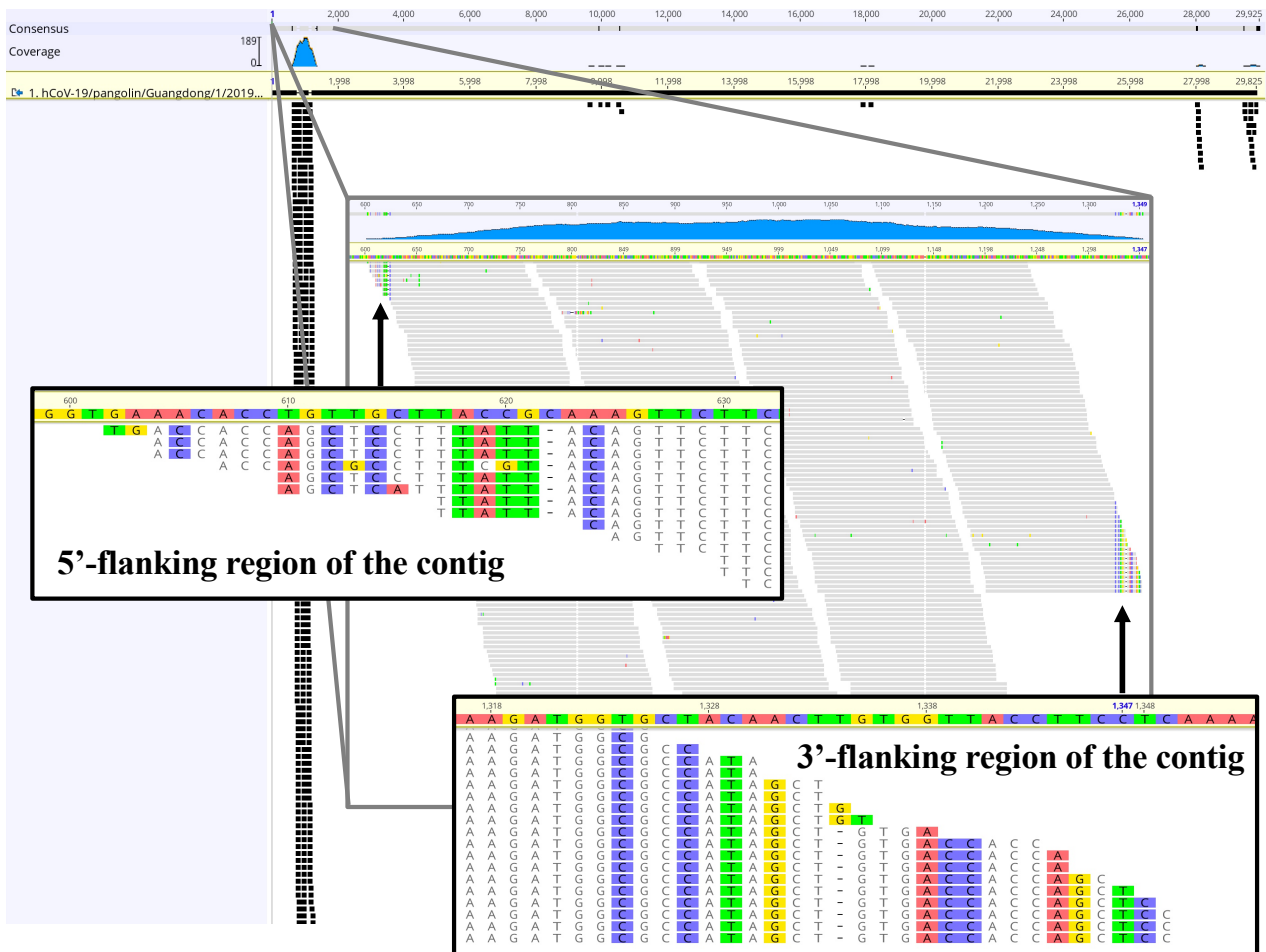
144 **Figure 1. Identification and characterization of *circRNA-GDP5L-1***

145 (A) The reads of the *GD/P5L* SRA dataset were mapped to the SARS-CoV-2-like genome
146 sequenced from Guangdong pangolins (GISAID accession number EPI_ISL_4107214) using
147 Geneious Prime® 2020.0.3. The preliminary consensus sequence was then used as a reference
148 for three mapping iterations in order to extend the contig.

149 (B) The final consensus sequence of 862 nucleotides was constructed from a contig of 635
150 reads. The *circRNA-GDP5L-1* is 733-nt long (red and black nucleotides not highlighted in
151 yellow). The 5'- and 3'-flanking regions, highlighted in yellow, support the circular nature of
152 the RNA molecule, as they are perfectly identical to the 3'- and 5'-ends of the *circRNA-*
153 *GDP5L-1* sequence, respectively. The 33-nt motif shown in red cannot be aligned to the
154 reference genome but is the perfect reverse complement of the underlined motif, located 20-
155 nucleotides downstream.

156 (C) The alignment of the genomic sequences located between the two complementary motifs
157 shows the existence of additional base-pairing motifs (highlighted in green) which are highly
158 conserved among SARS-CoV-2-like genomes, i.e., GX/P (same sequence found in five virus
159 genomes from Guangxi pangolins; GenBank accession numbers: MTO40333-6 and
160 MT072864), SARS-CoV-2 (NC_045512, human reference genome), RaGT13 (MN996532,
161 virus genome from the bat species *Rhinolophus affinis*), VZC45 (MG772933) and VZXC21
162 (MG772934), two virus genomes from the bat species *Rhinolophus sinicus*.

A. GD/P5L reads mapped to the reference genome (hCov-19/pangolin/Guangdong)



B. CircRNA-GDP5L-1 sequence

TTAGAGCCACTTGTGAGTTCGTGGTACTGAAAATTTGACTAAAGATGGCGCCATAGCTGTGACCACCAGCTCCT
 TTATTACAGTTCTTCTTCGCAAGAACGGTAATAAAGGAGCTGGTGGTCACAGCTATGGCGCCGATCTAAAGTCCT
 ATGACTTAGGTGACGAGCTGGGCACCTGATCCTTATGAAGATTATCAAGAAAACCTGGAACACTAAACATGGCAGTG
 GTGTAACCTCGTGAGCTCATGCGTGAGCTTAATGGGGCGCATACTCGCTATGTCGATAACAACACTACTGTGGCC
 CTGATGGCTACCCTCTTGAGTGCATTAAGACTTGTGGCGGTGCTGGTAAAGCTTCTTGCACTTTGTCCGAAC
 AACTGGACTTTCTTGACACTAAGAGAGGTGTGACTGCTGCCGTGAGCATGACCATGAAATTGCATGGTACACGG
 AAfCGCTCTGATAAGAGTTATGAATTGCAGACACCTTTTGAATTAACCTGGCAAAGAAATTTGACACTTTTACT
 GGGGAGTGCCCAAATTTGTATTCCTCTTAATTCACCTATCAAGACTATTCACCTAGAGTTGAAAGGAAAAAG
 CTTGATGGCTTTATGGGTAGGATTCGATCTGTCTACCTGTTGCTTACCTAATGAATGCAACCAAATGTGCCTG
 TCAACTCTCATGAAGTGAACCAATTGTGGTGAACCTTCATGGCAGACAGGCGATTTTGTAGAGCCACTTGTGAG
 TTCTGTGGTACTGAAAATTTGACTAAAGATGGCGCCATAGCTGTGACCACCAGCTCCTTTATTACAGTTCTTCTT
 CGCAAGAACGGTAATAAAGGAGCTGGTGGTCACAGCT

C. Alignment of the genomic sequences located between the two complementary motifs

| | | |
|------------|----------------------|----------------------|
| | | 11111111112 |
| | | 12345678901234567890 |
| GD/P5L | AGTTCTTCTTCGCAAGAACG | |
| GX/P | -A-----C-T-----T- | |
| SARS-CoV-2 | G-----T-----T----- | |
| RaGT13 | G-----T-----T----- | |
| VZC45 | -----T-----T----- | |
| VZXC21 | -----T-----T----- | |

Table 1. Analyses of SRA data available for SARS-CoV-2-like viruses detected in captive pangolins from Guangdong (GD) and Guangxi (GX).

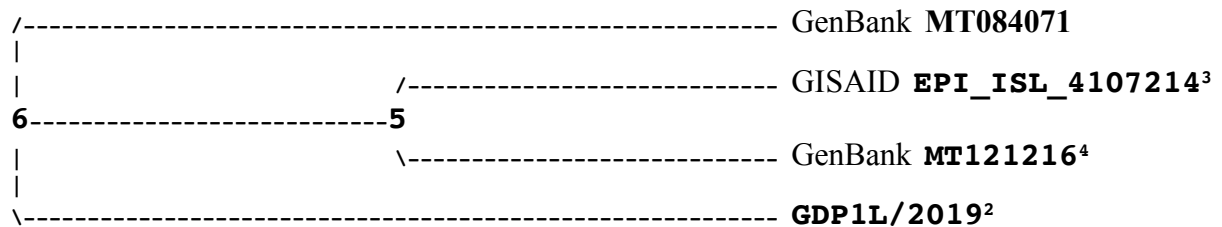
| Illumina sequencing run | | | Viral genome assembly | | | Reads mapped to mitochondrial genomes | | | |
|----------------------------------|-------------|-------------|-----------------------|---------|-----|---------------------------------------|--------------------|--------------------|--------------------|
| Code - Tissue | NCBI SRA | Reads | Reads | MC | MD | pangolin NC_026781 | human NC_012920 | tiger NC_010642 | mouse NC_005089 |
| GD/P7L ¹ - Lung 07 | SRR10168378 | 38,091,846 | 285* | 1.4X | 57% | 98,226 | 28 | 0 | 1,634 |
| GD/P8L ¹ - Lung 08 | SRR10168377 | 32,829,850 | 1,078* | 5.3X | 19% | 7,727 | 1,333 | 0 | 183 |
| GD/P9L ¹ - Lung 09 | SRR10168376 | 36,135,230 | 36 | 0.2X | 88% | 13,770 | 47 | 3,447 | 0 |
| GD/P11L ¹ - Lung 11 | SRR10168375 | 44,440,374 | 10 | 0.2X | 99% | 807,747 | 24 | 1,394 | 3 |
| GD/P2S ² – Scale P2 | SRR11093265 | 2,633 | 2,604 | 6.5X | 29% | 0 | 0 | 0 | 0 |
| GD/P1L ³ - Lung Mj1 | SRR11119766 | 46,486,240 | 10 | 0.1X | 97% | 1,952,841 | 0 | 0 | 0 |
| GD/P2L ³ - Lung Mj2 | SRR11119765 | 424,322,500 | 0 | NA | NA | 492,015 | 0 | 0 | 0 |
| GD/P3L ³ - Lung Mj3 | SRR11119763 | 393,522,404 | 8 | 0.1X | 97% | 319,594 | 0 | 112 | 0 |
| GD/P4L ³ - Lung Mj4 | SRR11119762 | 547,302,862 | 58 | 0.3X | 83% | 281,251 | 0 | 0 | 0 |
| GD/P5L ³ - Lung Mj5 | SRR11119759 | 214,534,718 | 600* | 3.0X | 92% | 333,399 | 9,637 | 0 | 20 |
| GX/P1E ² - Intestine | SRR11093266 | 470,344 | 450,433 | 510.5X | 1% | 25 | 0 | 0 | 0 |
| GX/P2V ² - Vero cells | SRR11093271 | 4,550,437° | 57,356 | 289.8X | 0% | 4 | 0 | 0 | 0 |
| GX/P3B ² - Blood | SRR11093270 | 19,607,536 | 1,794 | 7.3X | 16% | 52,875 | 1,174 | 17 | 9 |
| GX/P4L ² - Lung | SRR11093269 | 520,254 | 184,870 | 2002.1X | 3% | 7 | 0 | 0 | 0 |
| GX/P5E ² - Intestine | SRR11093268 | 416,659 | 401,378 | 3958.2X | 1% | 8 | 0 | 0 | 0 |
| GX/P5L ² - Lung | SRR11093267 | 340,661 | 274,982 | 2749.3X | 2% | 59 | 0 | 0 | 0 |

*: datasets in which circRNA reads were detected (see main text for details). Abbreviations: MC: mean coverage; MD: missing data.

Supplementary information 1

Nucleotide differences between the four SARS-CoV-2-like genomes sequenced from Guangdong pangolins.

Apomorphy lists associated to the following parsimonious tree reconstructed in PAUP* version 4.0b10 (Swofford, 2003).



| Branch | Character | CI | Change |
|---------------------|-----------|---------|---------|
| node_6 --> MT084071 | 5836 | 0.500 | A --> G |
| | 5838 | 0.500 | A --> T |
| | 5851 | 0.500 | A --> C |
| | 8870 | 1.000 | G ==> A |
| | 8875 | 1.000 | T ==> A |
| | 8878 | 1.000 | T ==> C |
| | 8883 | 1.000 | A ==> C |
| | 8887 | 1.000 | C ==> G |
| | 8889 | 1.000 | T ==> C |
| | 8891 | 1.000 | C ==> G |
| | 8892 | 1.000 | A ==> G |
| | 8893 | 1.000 | T ==> C |
| | 8894 | 1.000 | T ==> C |
| | 8895 | 1.000 | T ==> G |
| | 8898 | 1.000 | T ==> G |
| | 8900 | 1.000 | C ==> G |
| | 10216 | 1.000 | A --> G |
| | 11865 | 1.000 | A ==> T |
| | 25027 | 1.000 | T ==> A |
| | 25028 | 1.000 | T ==> A |
| 25535 | 1.000 | A --> G | |
| 29420 | 1.000 | C ==> T | |
| 29421 | 1.000 | T ==> C | |
| 29423 | 1.000 | C --> T | |
| 29427 | 1.000 | A --> T | |
| node_6 --> node_5 | 4057 | 1.000 | C --> T |
| | 4168 | 1.000 | G ==> T |
| | 4171 | 1.000 | C ==> T |
| | 4173 | 1.000 | A ==> C |
| | 4174 | 1.000 | A ==> C |
| | 4175 | 1.000 | C ==> G |

| | | | |
|----------------------------|-------|-------|---------|
| | 4176 | 1.000 | T ==> A |
| | 4177 | 1.000 | G ==> A |
| | 4178 | 1.000 | T ==> A |
| | 4181 | 1.000 | T ==> C |
| | 4182 | 1.000 | C --> T |
| | 4183 | 1.000 | C --> T |
| | 6143 | 1.000 | A ==> G |
| | 6920 | 1.000 | G --> T |
| | 7240 | 1.000 | A --> T |
| | 8948 | 1.000 | A --> C |
| | 15390 | 1.000 | A --> G |
| | 22542 | 1.000 | A --> C |
| | 25051 | 1.000 | C --> T |
| | 25052 | 1.000 | A --> C |
| | 25536 | 1.000 | C --> T |
| | 25538 | 1.000 | G ==> A |
| | 25539 | 1.000 | C ==> T |
| | 29422 | 1.000 | C --> G |
| | 29424 | 1.000 | C --> T |
| | 29425 | 1.000 | C --> G |
| | 29428 | 1.000 | A --> G |
| | 29800 | 1.000 | C --> T |
| | 29801 | 1.000 | C --> G |
| | 29802 | 1.000 | G --> T |
| node_5 --> EPI_ISL_4107214 | 5837 | 0.500 | C --> G |
| | 9042 | 1.000 | A ==> G |
| | 9089 | 1.000 | A ==> C |
| | 23090 | 1.000 | A --> T |
| | 23508 | 1.000 | A ==> T |
| | 24233 | 1.000 | C ==> T |
| node_5 --> MT121216 | 2193 | 1.000 | A --> G |
| | 2730 | 1.000 | A ==> G |
| | 4993 | 0.500 | C --> T |
| | 5836 | 0.500 | A --> G |
| | 5838 | 0.500 | A --> T |
| | 5851 | 0.500 | A --> C |
| | 18094 | 1.000 | C --> T |
| | 23433 | 1.000 | A --> C |
| | 26536 | 1.000 | G --> T |
| node_6 --> GDP1L/2019 | 4967 | 1.000 | G ==> T |
| | 4993 | 0.500 | C --> T |
| | 5837 | 0.500 | C --> G |
| | 5840 | 1.000 | G ==> T |
| | 8226 | 1.000 | C --> T |
| | 8253 | 1.000 | G --> T |
| | 19388 | 1.000 | T ==> C |
| | 23294 | 1.000 | C --> T |
| | 23395 | 1.000 | C --> T |
| | 29799 | 1.000 | A --> C |

Supplementary Information 2

Sequences and characteristics of the 10 circRNAs detected in SRA datasets generated from Guangdong pangolin samples in which SARS-Cov-2-like virus were identified.

As shown in **Figure 1**, the reads were mapped to the SARS-CoV-2-like genome sequenced from Guangdong pangolins (GISAID accession number EPI_ISL_410721³). All contigs with chimeric reads at the 5'- and 3'-flanking regions were considered as possible circRNAs, and further studied.

The characteristics of the 10 circRNAs here detected are described by the following data:

- (1) the viral sequence identical to the reference genome is written in black;
- (2) 5'- and 3'-flanking sequences are highlighted in yellow;
- (3) the additional motif is shown in red;
- (4) its reverse complement in the viral genomic sequence is underlined; and
- (5) the viral sequence between these two complementary motifs is generally characterized by additional base-pairing motifs, which are highlighted in green.

The evolution of these additional base-pairing motifs was analyzed by aligning the following SARS-CoV-2 genomes: GX/P (representing the five full genomes sequenced from Guangxi pangolins MTO40333-6 + MT072864); SARS-CoV-2 (NC_045512, human reference genome); RaGT13 (genome sequenced from the bat species *Rhinolophus affinis*); and VZC45 (MG772933) VZXC21(MG772934), two genomes isolated from the bat species *Rhinolophus sinicus*.

For each sequence, the centroid secondary structure was predicted using ViennaRNA web services⁷. In this structure, the location of the two full complementary motifs is indicated by a red arrow, whereas the location of additional base-pairing motifs is indicated by a green arrow.

Characteristics of the 10 circRNAs identified in this study in three lung samples from Guangdong pangolins (length in nucleotides)

| circRNA name | VGF length | SARS gene(s) | Genomic position* | Additional motif length | Distance (nt) from its RC |
|-----------------|------------|--------------------|-------------------|-------------------------|---------------------------|
| circRNA-GDP5L-1 | 700 | ORF1ab | 625-1323 | 33 | 20 |
| circRNA-GDP5L-2 | 268 | ORF8 | 28003-28270 | 10 | 14 |
| circRNA-GDP5L-3 | 412 | N + ORF10 + 3' end | 29424-29890 | 96 | 19 |
| circRNA-GDP7L-1 | 302 | ORF1ab | 13629-13930 | 35 | 32 |
| circRNA-GDP7L-2 | 329 | ORF1ab | 18694-19022 | 5 | 62 |
| circRNA-GDP8L-1 | 610 | ORF1ab | 12120-12729 | NA | NA |
| circRNA-GDP8L-2 | 512 | ORF1ab | 16230-16741 | 14 | 18 |
| circRNA-GDP8L-3 | 681 | ORF1ab | 18694-19374 | 10 | 62 |
| circRNA-GDP8L-4 | 758 | ORF1ab | 20677-21434 | 18 | 17 |
| circRNA-GDP8L-5 | 555 | N | 28581-29135 | 5 | 22 |

Abbreviations: VGF=viral genomic fragment; RC= reverse complement in the VGF.

*reference genome = hCoV-19/pangolin/Guangdong/EPI_ISL_410721³.

°partial RC in the VGF.

>circRNA-GD/P5L-1 [733 nt; contig of 635 reads]

```

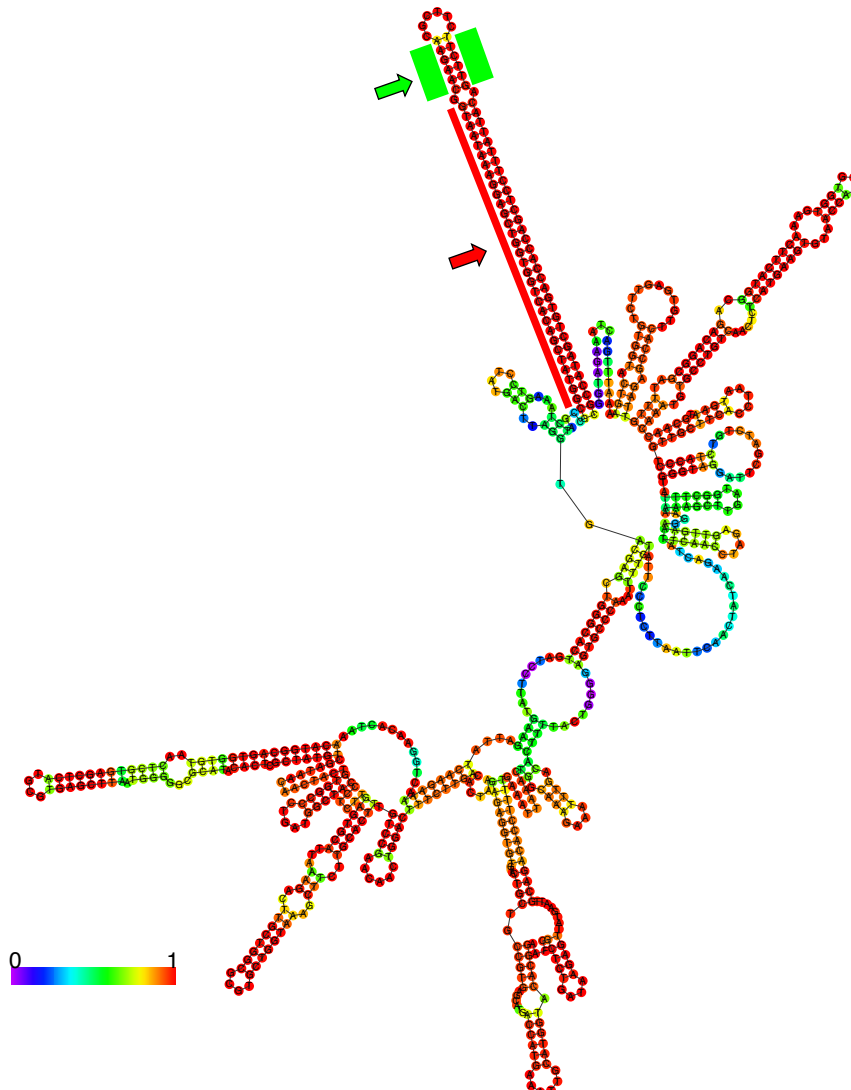
...CTGTGGTACTGAAAATTTGACTAAAAGATGGCGCCATAGCTGTGACCACCAGCTCCTTTATTACAAGTTCTTCTTC
GCAAGAACGGTAATAAAGGAGCTGGTGGTCACAGCTATGGCGCCGATCTAAAGTCCATGACTTAGGTGACGAGC
TGGGCACCTGATCCTTATGAAGATTATCAAGAAAACCTGGAACACTAAACATGGCAGTGGTGTAACTCGTGAGCTCA
TGCCTGAGCTTAATGGGGGCGCATACTCGCTATGTCGATAACAACACTACTGTGGCCCTGATGGCTACCCCTTTG
AGTGCATTAAAGACTTGCTGGCGCGTGTGGTAAAGCTTCTTGCACCTTTGTCCGAACAACCTGGACTTTCCTTGACA
CTAAGAGAGGTGTGTACTGCTGCCGTGAGCATGACCATGAAAATTGCATGGTACACGGAAFCGCTCTGATAAGAGT
TATGAATTGCAGACACCTTTTGAATTAACCTGGCAAAGAAAATTTGACACTTTTACTGGGGAGTGCCCAAATTTT
GTATTCCCTCTTAATTCAACTATCAAGACTATTCAACCTAGAGTTGAAAGGAAAAAGCTTGATGGCTTTATGGGT
AGGATTCGATCTGTCTACCCTGTTGCTTACCTAATGAATGCAACCAAATGTGCCTGTCAACTCTCATGAAGTGT
AACCATTGTGGTGAACCTTCATGGCAGACAGGCGATTTTGTTAGAGCCACTTGTGAGTTCTGTGGTACTGAAAAT
TTGACTAAAGATGGCGCCATAGCTGTGACCACCAGCTCCTTTATTACAGTTCTTCTTCGCAAGAACGGTAATAA...
  
```

Alignment of genomic sequences located between the two complementary motifs

| | |
|------------|----------------------|
| GD/P | AGTTCTTCTTCGCAAGAACG |
| GX/P | -A-----C--T-----T- |
| SARS-CoV-2 | G-----T-----T----- |
| RaGT13 | G-----T-----T----- |
| VZC45 | -----T-----T----- |
| VZXC21 | -----T-----T----- |

The base-pairing motifs (the reverse complement being underlined) are fully conserved in other SARS-CoV-2-like genomes. In GX/P genomes, the canonical base-pairing motifs have been restored through compensatory mutations (G=>A and C=>T).

Centroid secondary structure (minimum free energy=256.20 kcal/mol)



>circRNA-GDP5L-2 [278 nt; contig of 17 reads]

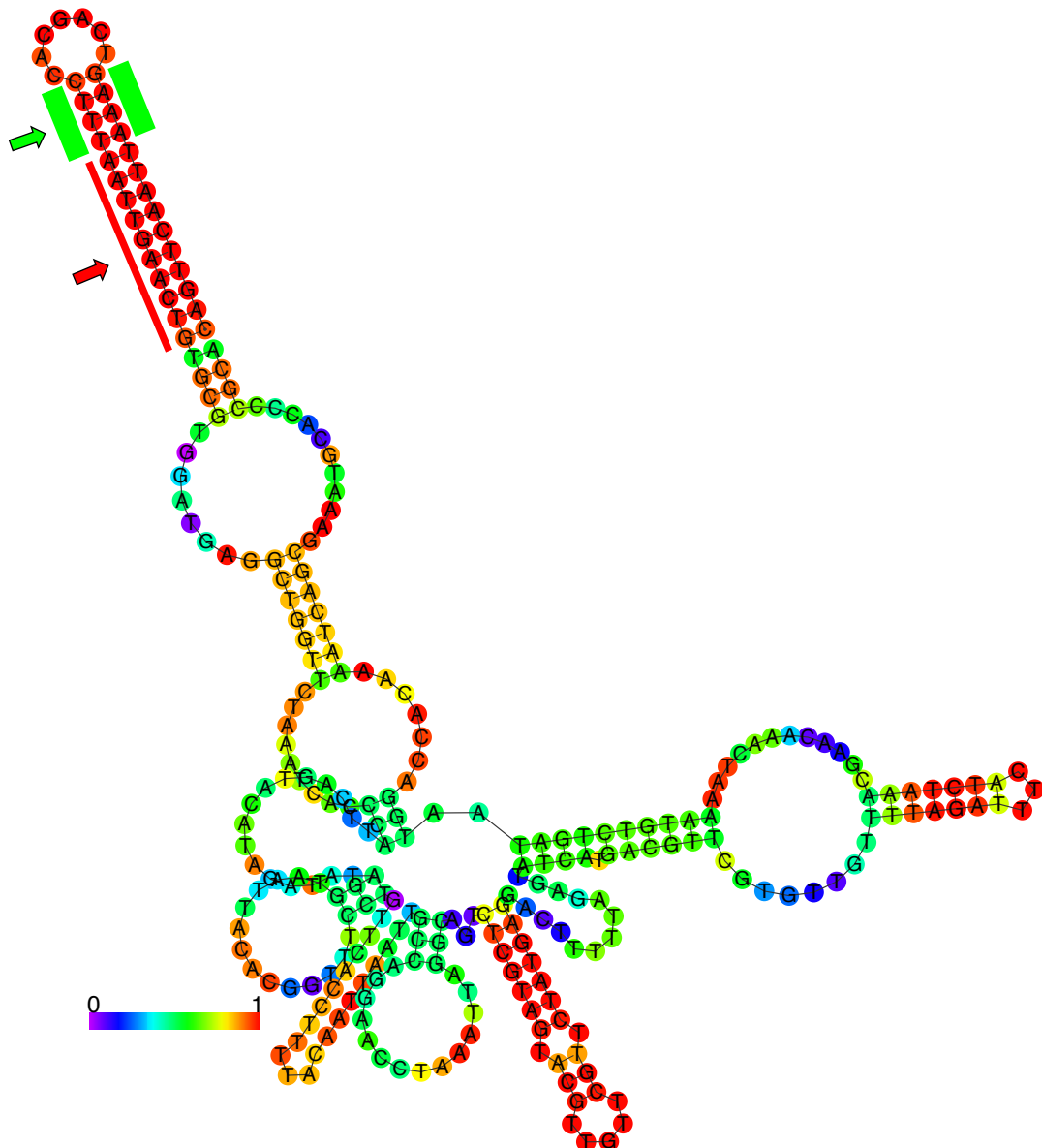
TCTGATAATGGACCACAAAATCAGCGAAAATGCACCCCGCACAGTTCAATTAAAGTCAGCACCTTTAATTGAACTG
 TGCGTGGATGAGGCTGGTTCTAAATCACCCATTCAGTACATAGATATAGGTAATTACACGGTTTCCTGTTACCT
 TTTACAATTAATTGCCAGGAACCTAAATTAGGCAGTCTCGTAGTACGTTGTTGTTCTATGAGGACTTTTATAGAG
 TATCATGACGTTTCGTGTTGTTTATAGATTTTCATCTAAACGAACAAACTAAAATGTCTGATAATGGACCACAAAATC
 AGCGAAAATGCACCCCGCACAGTTCAATTAAAGTCAGCACCTTTAATTGAACTGTGCGTGGATGAGGCTGGTTCTA
 AATCACCCATTCAGTACA

Alignment of genomic sequences located between the two complementary motifs

| | |
|------------|----------------------------|
| GD/P | <u>AAAGTCAGCACCTTT</u> |
| GX/P | ---A---T---A--- |
| SARS-CoV-2 | <u>A</u> ----- <u>CTTT</u> |
| RaGT13 | <u>A</u> ----- <u>CTTT</u> |
| VZC45 | <u>A</u> ----- <u>CTTT</u> |
| VZXC21 | <u>A</u> ----- <u>CTTT</u> |

The base-pairing motifs (the reverse complement being underlined) are partially conserved (3/4 nucleotides) in other SARS-CoV-2-like genomes. In GX/P genomes, different base-pairing motifs were found in the same region.

Centroid secondary structure (minimum free energy=49.20 kcal/mol)



>circRNA-GDP5L-3 [508 nt; contig of 17 reads]

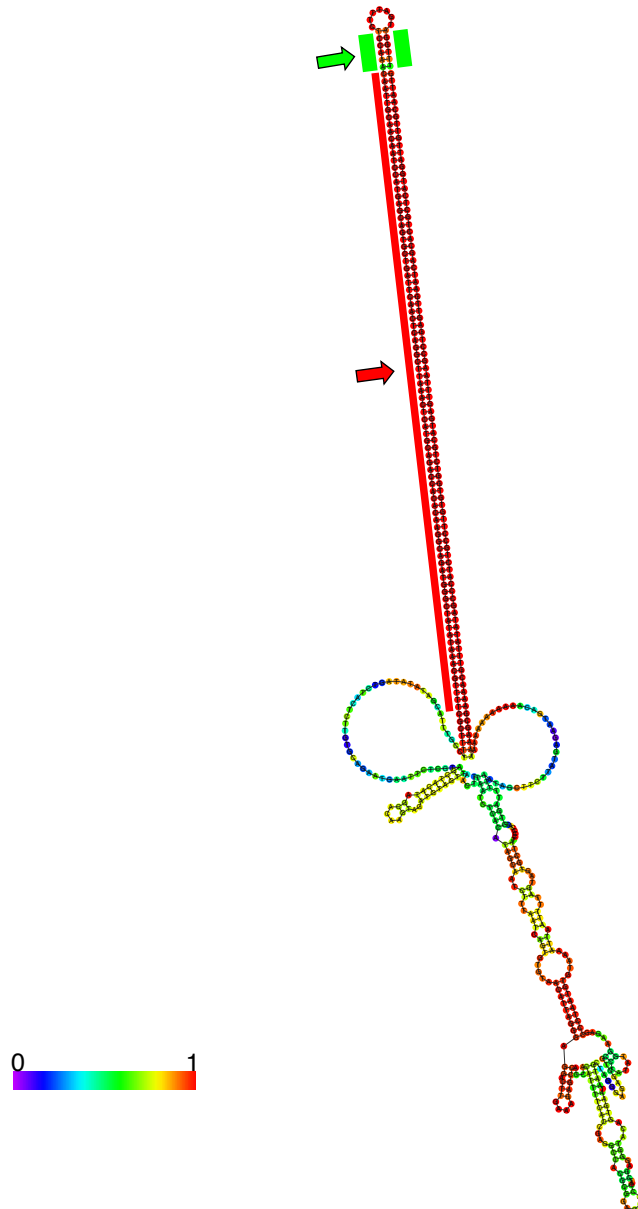
TTGTTTGGATGATTTCTCCAAACAATTGCAACAATCCATGAGCAGTGCTGATTCAACTCAGGCTTAAACTCATGC
 AGACCACACAAGGCAGATGGGCTATATAAACGTTTTTCGCTTTTCCGTTTACGATATATAGTCTACTCTTGTGCAG
 AATGAATTCTCGTAGCTACATAGCACAAAGTAGATGTAGTTAACTTTAATCTCACATAGCAATCTTTAATCAGTGT
 GTAACATTAGGGAGGACTTGAAAGAGCCACCACATTTTCACCGAGGCCACGCGGAGTACGATCGAGGGTACAGTG
 AATAATGCTAGGGAGAGCTGCCTATATGGAAGAGCCCTAATGTGTAAAAATTAATTTTAGTAGTGCTATCCCATG
 TGATTTTAATAGCTTCTTATGAGAATGACAAAAAAAAAAAAAAAAAAAGCGAAAACGTTTATATAGCCCATCTGCCTTG
TGTGGTCTGCATGAGTTTAAGCCTGAGTTGAATCAGCACTGCTCATGGATTGTTGCAATTGTTGGATGATTCT
CCAAAC

Alignment of genomic sequences located between the two complementary motifs

| | | |
|------------|---------------|---------------|
| GD/P | TTTGGATGATTTC | <u>TCCAAA</u> |
| GX/P | ----- | ----- |
| SARS-CoV-2 | ----- | ----- |
| RaGT13 | ----- | <u>C</u> |
| VZC45 | ----- | <u>A</u> |
| VZXC21 | ----- | <u>A</u> |

The base-pairing motifs (the reverse complement being underlined) are fully conserved in other SARS-CoV-2-like genomes.

Centroid secondary structure (minimum free energy=230.30 kcal/mol)



>circRNA-GDP7L-1 [337 nt; contig of 8 reads]

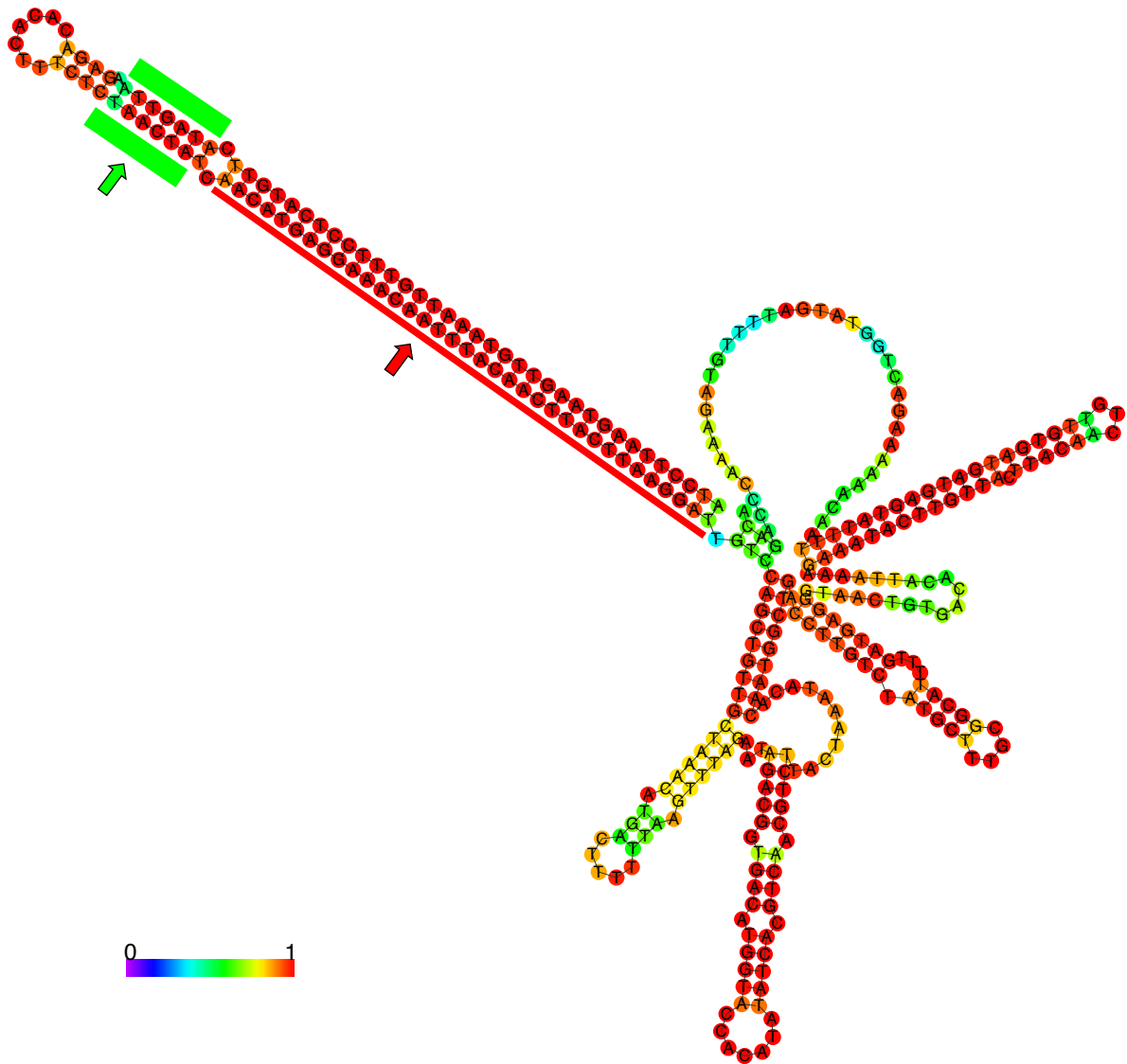
AACCCAGACAATCCTTAAGTAAGTTGTAAATTGTTTCCTCATGTTCCATAGTTAAGAGACACACTTTTCTCTAACTA
 TCAACATGAGGAAACAATTTACAACCTTACTTAAGGATTGTCCAGCTGTTGCTAAACATGACTTTTTTAAGTTTAG
 AATAGACGGTGACATGGTACCACATATATCACGTCAACGCTTACTAAATACACAATGGCTGACCTTGTCTATGC
 TTTGCGGCATTTTGTATGAGGGTAACTGTGACACATTAAAAGAAATACTTGTACTTACAACCTGTTGTGATGATGA
 GTATTTTAACAAAAAGACTGGTATGATTTTGTAGAAAACCCAGACAATCCTTAAGTAAGTTGTAAATTGTTTC
 TCATGTTTCATAG

Alignment of genomic sequences located between the two complementary motifs

| | |
|------------|-----------------------------------|
| GD/P | CATAGTTAAGAGACACACTTTTCTCTAACTATC |
| GX/P | -----T----- |
| SARS-CoV-2 | <u>TC</u> ----- <u>GC</u> |
| RaGT13 | <u>TC</u> ----- <u>GC</u> |
| VZC45 | <u>C</u> -----GAA-G-C |
| VZXC21 | <u>C</u> -----GAA-G-C |

The base-pairing motifs (the reverse complement being underlined) are fully or partially conserved (6/7 nucleotides) in other SARS-CoV-2-like genomes. In pangolin viral genomes, the canonical base-pairing motifs have been restored through compensatory mutations (G=>A and C=>T).

Centroid secondary structure (minimum free energy=119.00 kcal/mol)



>circRNA-GDP7L-2 [332 nt; contig of 6 reads]

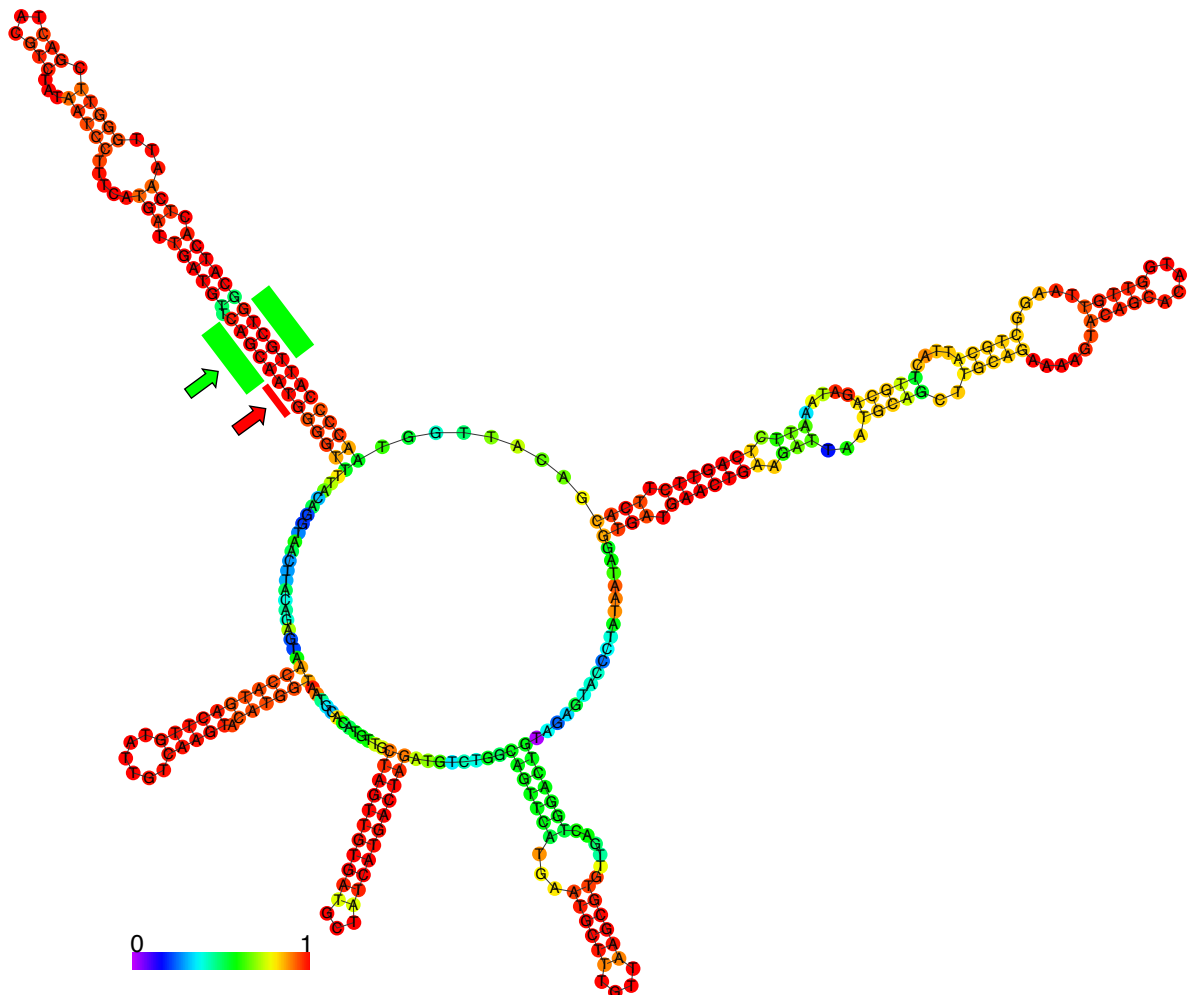
ACATGGTTGTTAAGGCTGCATTACTTGCAGATAAAATTCTCAGTTCTTCACGACATTGGTAACCCCATTGCTGGCA
 TCACTCAATTGGGTTTCGACTACGTCTATAATCCTTTTCATGATTGATGTTCAGCAATGGGGTTTTACAGGTAACCTT
 ACAGAGTAACCATGACATGCTATTGTCAAGTACATGGTAATGCACATGTTGCTAGTTGTGATGCTATCATGACTAG
 ATGTCTGGCAGTTCAATGAATGCTTTGTTAAGCGTGTGACTGGACTGTAGAGTACCCATAATAGGTGATGAAC
 GAAGATTAATGCAGCTTGCAGAAAAGTACAGCACATGGTTGTTAAGGCTGCATTACTTGCAGATAAAATTCAGT
 TCTTCACGACATTGGTAACCCCATTGCTGGCATCACTCAATTGGGTTCGACTACGTCTATAATCCTTTTCATGATT
GATGT

Alignment of genomic sequences located between the two complementary motifs

| | |
|------------|--|
| GD/P | <u>TGCTGGCATCACTCAATTGGGTT</u> C.....TTTCATGATTGATGTT <u>CAGCA</u> |
| GX/P | <u>T</u> -----G-----A-T----- <u>A</u> |
| SARS-CoV-2 | <u>T</u> -----T-T-----A-T.....G-T----- <u>A</u> |
| RaGT13 | <u>T</u> -----T-T-----A-T.....G-T----- <u>A</u> |
| VZC45 | <u>T</u> -----A-----TG-G-C-T.....G-T-----C |
| VZXC21 | <u>T</u> -----A-----TG-G-C-T.....A-----C |

The base-pairing motifs (the reverse complement being underlined) are fully or partially conserved (4/5 nucleotides) in other SARS-CoV-2-like genomes. In the Guangdong pangolin viral genome, the canonical base-pairing motifs have been restored through compensatory mutations (T=>C and A=>G).

Centroid secondary structure (minimum free energy=97.20 kcal/mol)



>circRNA-GDP8L-1 [610 nt; contig of 89 reads]

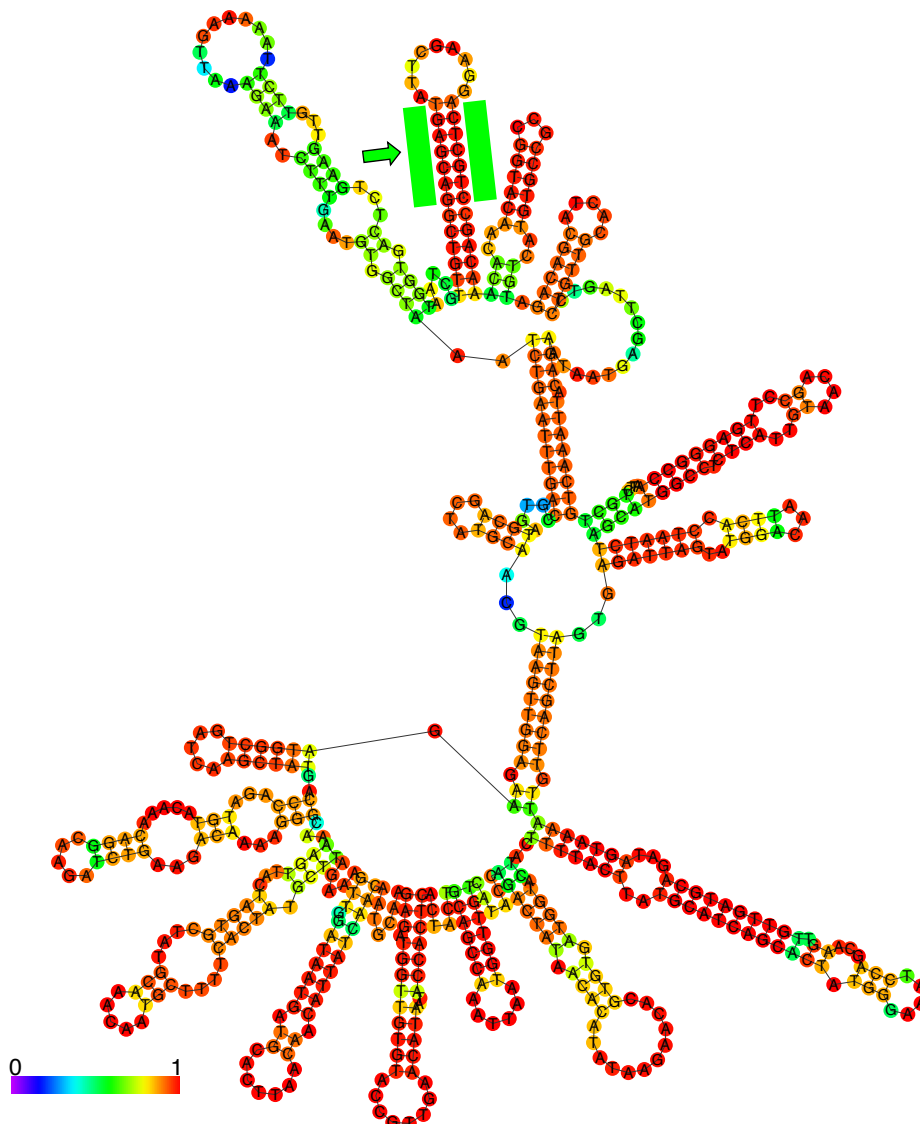
GTTGCACTACGACAGATGTCATGTGCCGCCGGTACAACACAAACAGCCTGCTCAGGAAGCTTAGAGCAGGCTGT
 TGCTAATGGTGACTCTGAAGTTGTTCTTAAAAAGTTAAAGAAATCTTTGAATGTGGCTAAATCTGAATTTGACCC
 TGATGCAGCTATGCAACGTAAGTTGGAGAAGATGGCTGATCAAGCTATGACCCAGATGTACAAACAGGCAAGATC
 TGAAGACAAAAGGGCAAAGTTACTAGTGTCTAGTGCATCAACAATGCTTTTCACTATGCTTAGAAAGTTGGATAATGA
 TGCACCTAACAACATTATCAACAATGCAAGAGATGGTTGTGTACCGTTGAACATAATACCACCTCACTACTGCAGC
 CAAATTAATGGTTGTCATACCAGACTATAACACATATAAGAACACGTGTGATGGTACTACTTTTACTTATGCATC
 AGCACTATGGGAAATCCAGCAAGTTGTTGATGCAGATAGTAAAATTTGTTTCAGCTTAGTGAGATTAGTATGGACAA
 TTCACCTAATCTAGCATGGCCTCTCATTGTAACAGCCTTGAGGGCCAATTTCTGCTGTCAAATTACAGAATAATGA
 GCTTAGTCCTGTTGCACTACGACAGATGTCATGTGCCGCCGGTACAACACAAACAGCCTGCTCAGGAAGCTTATG
 AGCAGGCTGTTGCTAATGGTGACTCTGAAGTTGTTCTTAAAAAGTTAAAGAAATCTTTGAATG

Alignment of genomic sequences in the 5'-part of the circRNA

| | |
|------------|-------------------------------|
| GD/P | ACTGCTCAGGAAGCTTATGAGCAGGCTGT |
| GX/P | ---A---A-----G---A---G--- |
| SARS-CoV-2 | -----A-----G----- |
| RaGT13 | ---C---A-----G--- |
| VZC45 | -----A---G----- |
| VZXC21 | ---A---C---A---G----- |

The base-pairing motifs (the reverse complement being underlined) are fully or partially conserved (at least 5/7 nucleotides) in other SARS-CoV-2-like genomes. In the VZXC21 genome, different base-pairing motifs were found in the same region.

Centroid secondary structure (minimum free energy=165.30 kcal/mol)



>circRNA-GDP8L-2 [526 nt; contig of 55 reads]

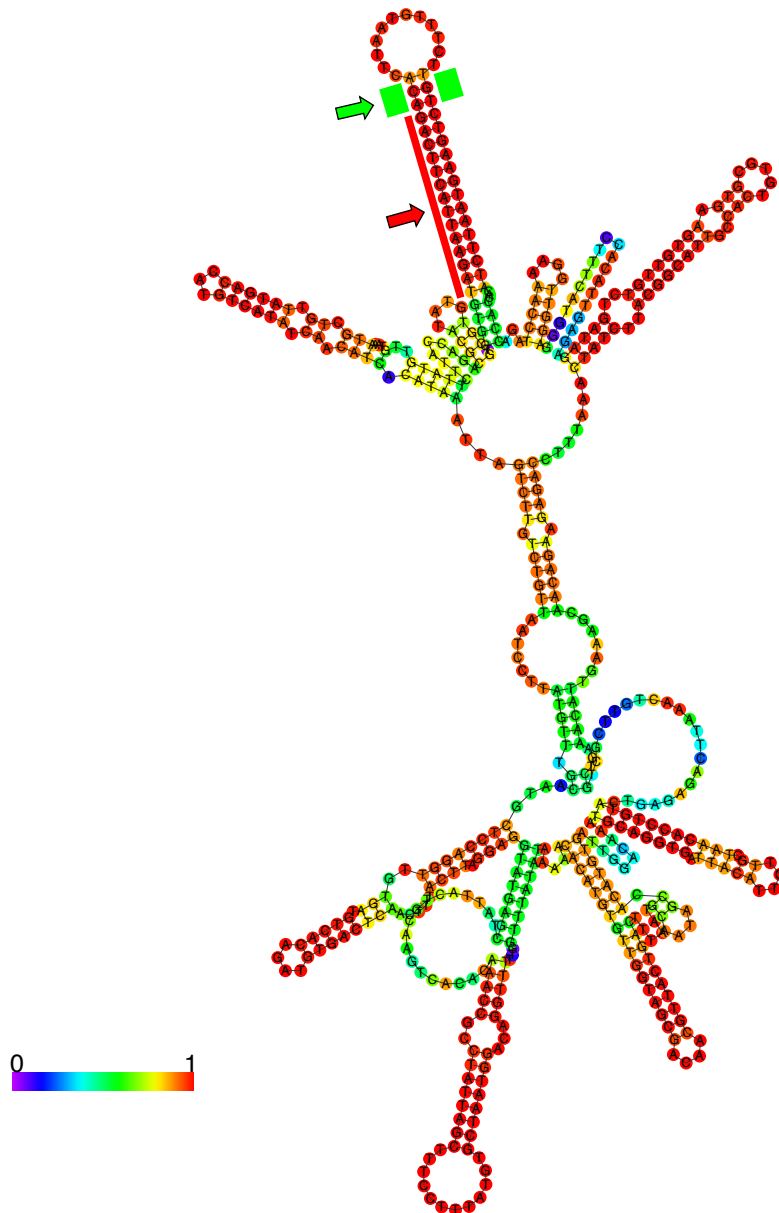
CCANCCACATCTTAATGAAGTCTGTTCTTTGTAATTCACAGACTTCATTAAGATGTGGTGCGTGTATACGCAGAC
 CATTCTTATGTTGTAATGCTGTTATGACCATGTCATATCAACATCACATAAATTAGTCTTGTCTGTTAATCCTT
 ATGTTTGCAATGCTCCAGGTTGTGATGTCACAGATGTGACTCAACTTACTTAGGAGGTATGAGCTATTACTGCA
 AGTCACACAAACCGCTATTAGCTTTCCTTTATGTGCTAATGGACAGGTTTTTGGTTTATATAAAAAACACATGTG
 TTGGTAGCGACAACGTTACTGACTTTAATGCAATAGCCACATGTGATTGGACAAATGCAGGTGATTACATTCCTG
 CTAACACCTGTACTGAGAGACTTAAACTGTTTCGCTGCTGAAAACATTGAAAACAACAGAAGAGACCTTTAAACTAT
 CTTACGGCATTGCCACTGTGCGTGAAGTGTGCTGATAGAGAGTTACACCTTTCATGGGAGGTTGGAAAACCTA
 GACCACCACATCTTAATGAAGTCTGTTCTTTGTAATTCACAGACTTCATTAAGATGTGGTGCGTGTATACGCAG...

Alignment of genomic sequences located between the two complementary motifs

| | |
|------------|--------------------|
| GD/P | TGTTCTTTGTAATTCACA |
| GX/P | -----C-C----- |
| SARS-CoV-2 | -----C----- |
| RaGT13 | -----C----- |
| VZC45 | GT-A-C----- |
| VZXC21 | AT-A----- |

The base-pairing motifs (the reverse complement being underlined) are fully conserved in other SARS-CoV-2-like genomes.

Centroid secondary structure (minimum free energy=137.80 kcal/mol)



>circRNA-GDP8L-3 [691 nt; contig of 85 reads]

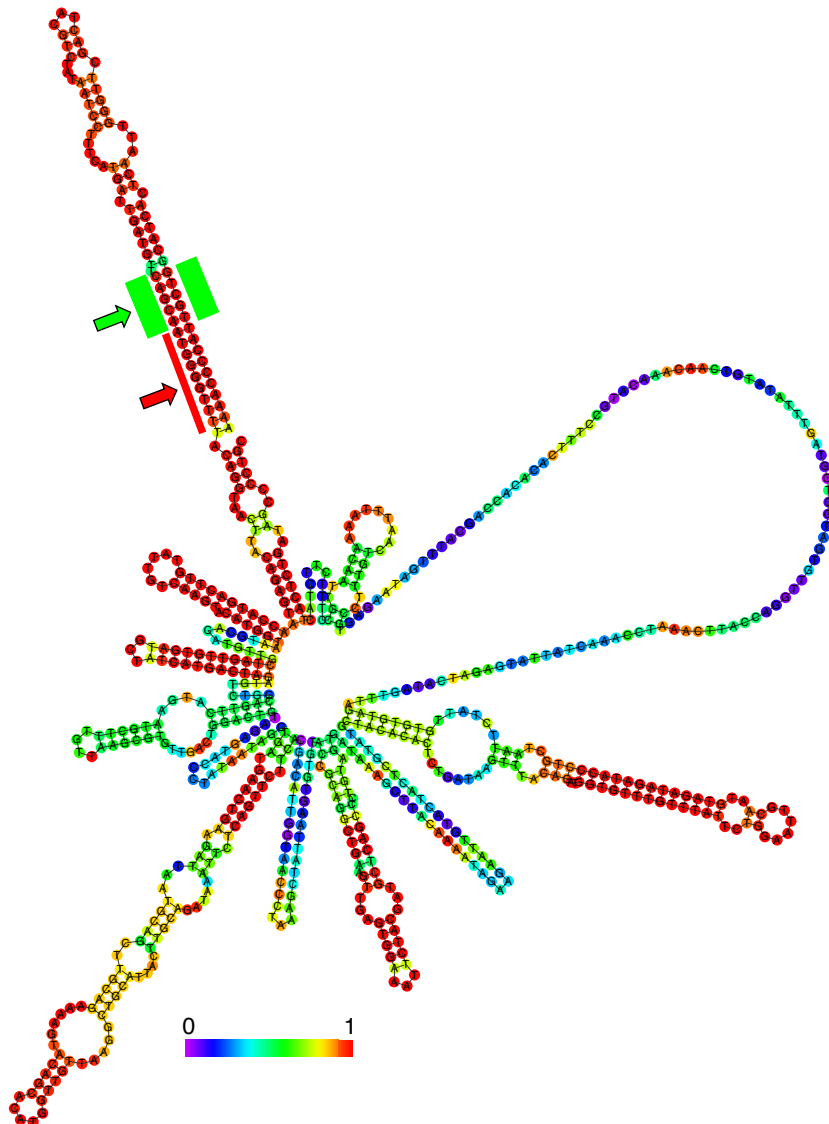
...CTACTACTCTGATAGCCCCTGCAAAACCCCATTGCTG GCATCACTCAATTGGGTTGACTACGTCTATAATCCT
 TTCATGATTGATGTTCAGCAATGGGGTTTTACAGGTAACCTTACAGAGTAACCATGACTTGTATTGTCAAGTACAT
 GGTAAATGCACATGTTGCTAGTTGTGATGCTATGATGATGCTAGATGCTGCGCAGTTCATGAATGCTTTGTTAAGCGT
 GTTGACTGGACTGTAGAGTACCCTATAATAGGTGATGAAGATTAAATGCAGCTTGCAGAAAAGTACAGCAC
 ATGGTTGTTAAGGCTGCATTACTTGCAGATAAAATTCTCAGTTCTTACAGACATTGGTAACCCATAAGCTATTAAG
 TGTGTACCGCAGGCTGAAGTTGAGTGGAAATTTCTACGATGCTCAGCCCTGTAGTGATAAAGCTTACAAAATAGAA
 GAATTGTACTACTCGTATGCTACACACTCTGATAAGTTTACAGATGGTGTGTTGTTTATTTCTGGAATTGCAATGTA
 GATAGATACCCTGCTAATTCTATTGTGTGTAGATTTGATACTAGAGTATTATCAAACCTAAACTTACCAGGTTGT
 GATGGTGGTAGTTTATATGTCAACAAACATGCCTTTTACACACCAGCATTGATAAGAGTGCCTTTGTCAATTTA
 AAACAATTGCCTTTCTTCTACTACTCTGATAGCCCCTGCAAAACCCCATTGCTGAACATCAATCATGAAAGG...

Alignment of genomic sequences located between the two complementary motifs

| | | |
|------------|-----------------------------------|--------------|
| GD/P | TGCTGGCATCACTCA.....CATGATTGATGTT | <u>CAGCA</u> |
| GX/P | ----- | <u>A</u> |
| SARS-CoV-2 | -----T-----T-----T-----T----- | <u>A</u> |
| RaGT13 | -----T-----T-----T-----T----- | <u>A</u> |
| VZC45 | -----T-----A-----T-----T----- | <u>C</u> |
| VZXC21 | -----T-----A-----T----- | <u>C</u> |

The base-pairing motifs (the reverse complement being underlined) are fully or partially conserved (4/5 nucleotides) in other SARS-CoV-2-like genomes. In the Guangdong pangolin viral genome, the canonical base-pairing motifs have been restored through compensatory mutations (T=>C and A=>G).

Centroid secondary structure (minimum free energy=166.30 kcal/mol)



>circRNA-GDP8L-4 [776 nt; contig of 40 reads]

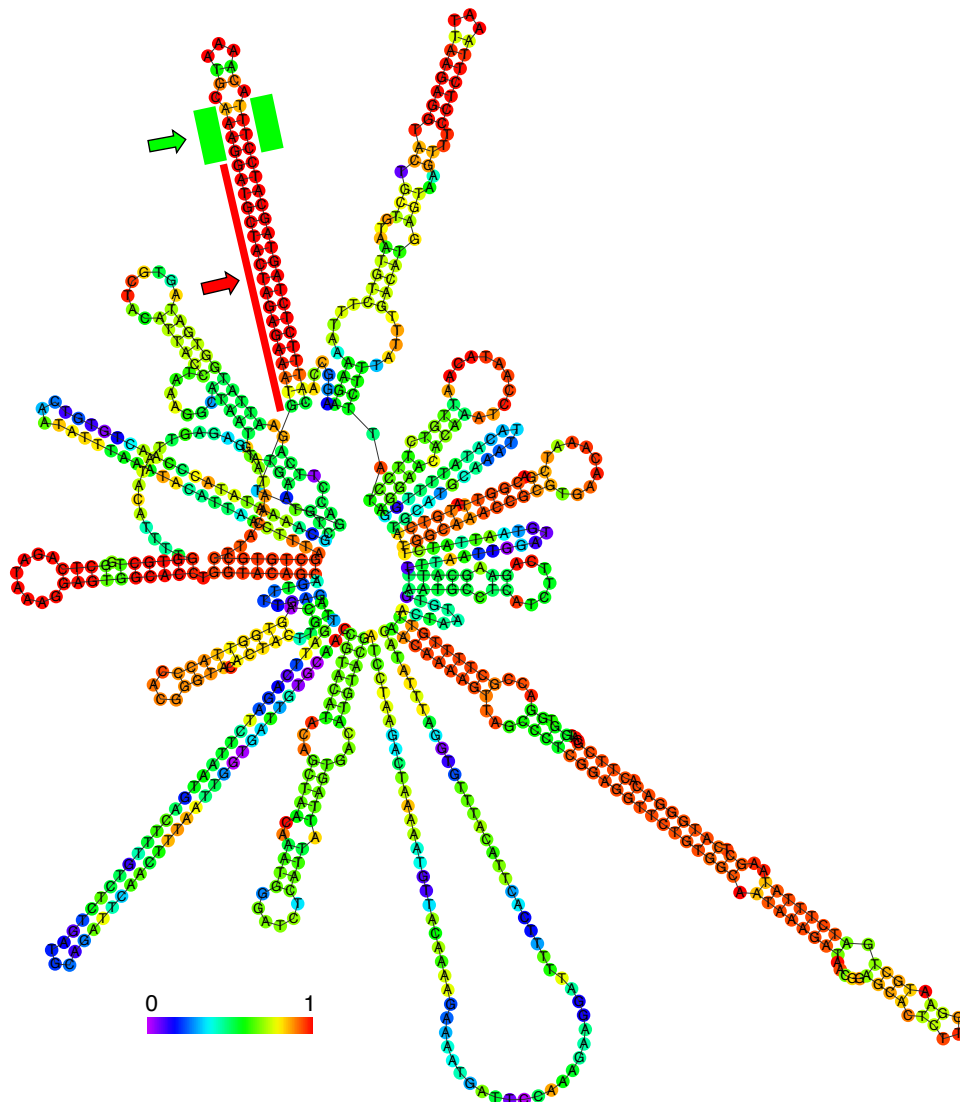
TAAAAGAAGGCCA**CATTTCTCTAGTAGCATC**CTTTACAAAATGC**AAAAG**GATGCTACTAGAGAAATGTGACCTTCA
 GAATTATGGTGATAGTGCTACATTACCTAAAGGCATAATGATGAATGTCGCAAAATATACCCAACCTGTGTCAATA
 TTTAAATACATTAACCTTTAGCTGTGCCTTACAATATGAGAGTTATACATTTTGGTGCTGGCTCAGATAAAGGAGT
 GGCACCTGGTACAGCAGTTTTGAGACAGTGGTTACCCACGGGTACACTACTTGTGATTGATTGATCCTTAATGACTT
 TGTCTCTGATGCAGATTCAACTTTAATTGGTGATTGTGCAACCGTACATACAGCTAACAAATGGGATCTCATTAT
 TAGTGACATGTACGATCCTAAGACTAAAAATGTTACAAAAGAAAATGATTCCAAAAGAAGGATTTTTCACTTACAT
 TTGTGGATTTATAACAACAAAAGTTAGCCCTCGGAGGTTCTGTGGCAATAAAGATAACGGAGCACTCTTGGAAATGC
 TGATCTTTATAAGCTCATGGGACACTTCGCATGGTGGACCGCTTTTGTACTAATGTGAATGCCTCATCTTCAGA
 AGCATTTTTAATTGGATGTAATTATCTTGGCAAACCGCGTGAACAAATCGACGGTTATGTCATGCATGCAAATTA
 CATATTTTGGAGGAACACAAATCCAATACAATTGTCTTCTTCTTATTTGACATGAGTAAGTTTCTCTTAA
 ATTAAGAGGTACTGCTGTAATGTCTTTAAAAGAAGGCCA**AATTTCTCTAGTAGCATC**CTTTACAAAATGCAAAGG
ATGCTACT

Alignment of genomic sequences located between the two complementary motifs

| | |
|------------|-------------------------------------|
| GD/P | CTTTACAAAATGCAAAAG |
| GX/P | ----- G ----- |
| SARS-CoV-2 | ----- G ----- |
| RaGT13 | C ----- G ----- |
| VZC45 | C ----- G ----- |
| VZXC21 | ----- T ----- G ----- |

The base-pairing motifs (the reverse complement being underlined) are fully or partially conserved (at least 3/4 nucleotides) in other SARS-CoV-2-like genomes.

Centroid secondary structure (minimum free energy=164.30 kcal/mol)



>circRNA-GDP8L-5 [560 nt; contig of 92 reads]

AAGAATTAATCAGACAAGGAAGTATTACAAACAATGGCCGCA**TTGTTAGC**TGGACTTCCCTATGGT**GCT**AACAA
 AGAAGGCATCATATGGGTTGCAAATGAGGGAGCCTTGAATACACCTAAAGATCATTGGCACCCGAAATCCTGC
 TAACAATGCTGCAATCGTGTCTACAATTCCTCAAGGAACAACATTGCCAAAAGGCTTCTACGCAGAAGGGAGCAG
 AGCGCGCAGTCAAGCTTCTCTCGTTCCTCATCAGTAGTCGCAACAGTTCGAAGAAACACAACCTCCAGGCAGCAG
 CAGGGGAACCTTCTCCTGCTAGGATGGCTGGCAATGGTGGTGATGCTGCTCTTGCTTTGCTGCTGCTTGACAGGTT
 GAACCAACTTGAGAGCAAAAATGTCTGGTAAAGGCCAACAACAACAAGGCCAAAAGTGCCTAAGAAAATCCGCTGC
 AGAGGCTTCTAAGAAACCTCGCCAAAACGTAAGTCCACCAACAACAACAATGTAACACAAGCTTTTGGCAGACG
 TGGTCCAGAACAACCCAAGGAACTTTGGGGATCAAGAATTAATCAGACAAGGAACTGATTACAAACAATGGCC
 GCAT**TTGTTAGCTGGACTTCCCTATGGTGCTAACAAGAAGGCATCATATGGGTTGCAAATGAGGGAGCCTTGAA...**

Alignment of genomic sequences located between the two complementary motifs

| | |
|------------|--------------------------------|
| GD/P | AGCTGGACTTCCCTATGGT GCT |
| GX/P | ----- |
| SARS-CoV-2 | ----- |
| RaGT13 | ----- |
| VZC45 | ----- |
| VZXC21 | ----- |

The sequence was found to be fully conserved in all SARS-CoV-2-like genomes.

Centroid secondary structure (minimum free energy=135.80 kcal/mol)

