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

Alexandre Hassanin, Huw Jones, Anne Ropiquet

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

Alexandre Hassanin[1*], Huw Jones[2] & Anne Ropiquet[2]

1. Institut de Systématique, Evolution, Biodiversité, UMR 7205 CNRS, MNHN, Sorbonne Université, EPHE, Université des Antilles, Muséum National d'Histoire Naturelle, CP 51, 57 rue Cuvier, 75231 PARIS Cedex 05 France.
*Corresponding author: alexandre.hassanin@mnhn.fr

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

Viruses closely related to SARS-CoV-2, the virus responsible for the Covid-19 pandemic, have previously been identified in several Sunda pangolins (*Manis javanica*) seized in Guangdong and Guangxi provinces of China[1-4]. The virus extracted from Guangdong pangolins is particularly intriguing as the amino-acid sequence of the receptor binding domain of the spike protein is nearly identical to that of the human SARS-CoV-2 virus[2,3]. This characteristic suggests it is capable of binding to the human ACE2 receptor and may therefore be able to mediate infection of human cells.

Here, we assembled SARS-CoV-2-like genomes from the Sequence Read Archive (SRA) data available in May 2020 for ten Guangdong pangolins. As previously described[5] the genome assemblies were of poor quality, having high levels of missing data and with possible DNA contamination from human, mouse and tiger. We found 10 circular RNAs (circRNAs) of 278-776 nucleotides (nt) containing SARS-CoV-2-like

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

Next-generation sequencing (NGS) runs previously published for 16 pangolin samples were analysed (**Table 1**)[1-3]. A large number of SARS-CoV-2-like reads (≥184,870) were found in SRA datasets generated from lungs and intestines of Guangxi pangolins, whereas a limited number of reads (≤1,078) were found in SRA datasets generated from lungs of Guangdong pangolins. These differences suggest higher RNA degradation in Guangdong samples and explain why it was not possible to assemble a full virus genome from these data. Assuming the same virus infected Guangdong pangolins[3,4], we pooled together the 4688 SARS-CoV-2-like reads found in the ten Guangdong datasets to assemble a consensus sequence. The resulting genome remained partial, with 15 fragments and 4% missing data (mean coverage 16.6X). Two recent studies[3,4] had filled genome gaps by sequencing multiple PCR products. However, these two published genomes differ at 15 nucleotide positions (Supplementary Information) indicating the quality of Guangdong pangolin virus sequence still needs to be improved[5].

In three SRA datasets generated from lung samples of three Guangdong pangolins (**Table 1**), we discovered several circRNAs by mapping NGS reads to the reference genome (GISAID accession number EPI_ISL_410721[3]) using Geneious Prime® 2020.0.3 and the "low sensitivity" option (maximum mismatch: 10%). For the *GD/P5L* dataset[3] (**Table 1**), most SARS-CoV-2-like reads (566 of all 600 reads) mapped to a small portion of the ORF1ab gene (**Figure 1A**). However, the 5'-and 3'-flanking regions of the contig were found to contain chimeric reads, in which the extremities could not be aligned to the reference genome. A preliminary consensus sequence of 748 nt was extracted and then used as a reference to extend the contig by mapping *GD/P5L* reads (three iterations in Geneious Prime® 2020.0.3).

The final contig contained 635 reads, and the consensus sequence of 862 nt was then extracted for further analysis. As shown in **Figure 1B**, its flanking regions (highlighted in yellow) are parts of an original circRNA of 733-nt long, herein referred to as *circRNA-GDP5L-1*: the 5'-flanking region is identical to the 3'-end of *circRNA-GDP5L*-1 (positions 685-733), whereas the 3'-flanking region is identical to the 5'-end of *circRNA-GDP5L*-1 (positions 1-80). Such repetitions of identical sequences are consistent with a circular nature of the original RNA molecule. In support of this finding, the high number of reads found for *circRNA-GDP5L-1* (635 reads) suggests the molecule was still circular before reverse transcription (see protocol used for library construction[3]) as multiple cDNA copies of a single circRNA can be generated through rolling circle amplification[6]. The structure of *circRNA-GDP5L-1* is characterized by the presence of a 33-nt motif (positions 1-33, shown in red in **Figure 1B**) which cannot be aligned to the reference genome, but which is the full reverse complement of the sequence located 20-nt downstream (underlined sequence in **Figure 1B**). Based on computer modeling tools[7], we can predict *circRNA-GDP5L-1* to form a highly stable secondary structure (Supplementary Information). We found nine other highly structured circRNAs showing similar features: two further in the *GD/P5L* dataset, two in the *GD/P7L* dataset, and five in the *GD/P8L* dataset. Their full sequences, characteristics and secondary structures are detailed in Supplementary Information. The length of these circRNAs varies between 278 and 776 nt. Only a single circRNA contained a viral sequence without additional motif. The nine other circRNAs are composed of a viral genomic sequence followed by the full reverse complement of a viral motif (length between 5 and 96 nt) found a few nucleotides downstream (between 14 and 62 nt). Notably, a substantial proportion of these genomic sequences are protein-coding-fragments (see details in Supplementary Information).

CircRNAs form a large class of RNA molecules that are widespread in eukaryotes[8]. They play important functions in gene regulation as they can have interactions with RNA-binding proteins, serve as translation templates, and act as microRNA sponges or transcriptional regulatory factors[8-10]. Recent studies have reported that DNA-virus genomes are also capable of generating circRNAs[11,12]. To our knowledge, these molecules have not previously been described as by-products of RNA-virus genomes. It is therefore important to know whether these circRNAs are natural or not, since some circRNAs have been described as possible artifacts of the reverse transcription process[13]. In addition, an intriguing issue is that the three SRA datasets in which circRNAs were found are the same three datasets containing human and mouse DNA (**Table 1**)[5]. The comparison between NGS runs performed on RNAs extracts from pangolin lungs (**Table 1**) clearly show much more SARS-CoV-2-like reads for Guangxi datasets than for Guangdong datasets (≥184,870 versus ≤1,078). Two hypotheses can be formulated to explain these differences: Guangdong samples contain lower concentrations of SARS-CoV-2-like virus; or RNA molecules extracted from Guangdong samples were much more degraded due to poor sample preservation. In support of the second hypothesis, it has been found that transcriptome degradation can result in circRNA enrichment as linear RNAs decay more rapidly than circRNAs[14], suggesting that the 10 circRNAs discovered in the three Guangdong pangolin samples are true by-products of SARS-CoV-2 like RNA-genomes. Another argument for the natural origin of these circRNAs is provided by the alignment of their 5'-end with homologous sequences extracted from SARS-CoV-2 like genomes (Supplementary Information). In the case of *circRNA-GDP5L-1*, the two additional base-pairing motifs involved in the secondary structure (highlighted in green in **Figure 3C**) are found to be highly conserved among SARS-CoV-2-like viruses, whereas the genomic fragment is a part of a protein-coding gene. In this context, the fact that canonical base-pairing motifs have been restored through compensatory mutations in the virus

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

But what can the role of viral circRNAs be during host infection? It has been demonstrated that viral infection triggers the degradation of host-produced circRNAs which in turn activates the innate immune response[15]. This mechanism suggests viral circRNAs could have an important role, either to delay the host's immune response, or to initiate viral proliferation in host cells. Our discovery of circRNAs generated from pangolin SARS-CoV-2-like virus opens up new perspectives for understanding virus-host interactions and has possible implications in a wider range of RNA-virus infections including Covid-19.

References

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

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

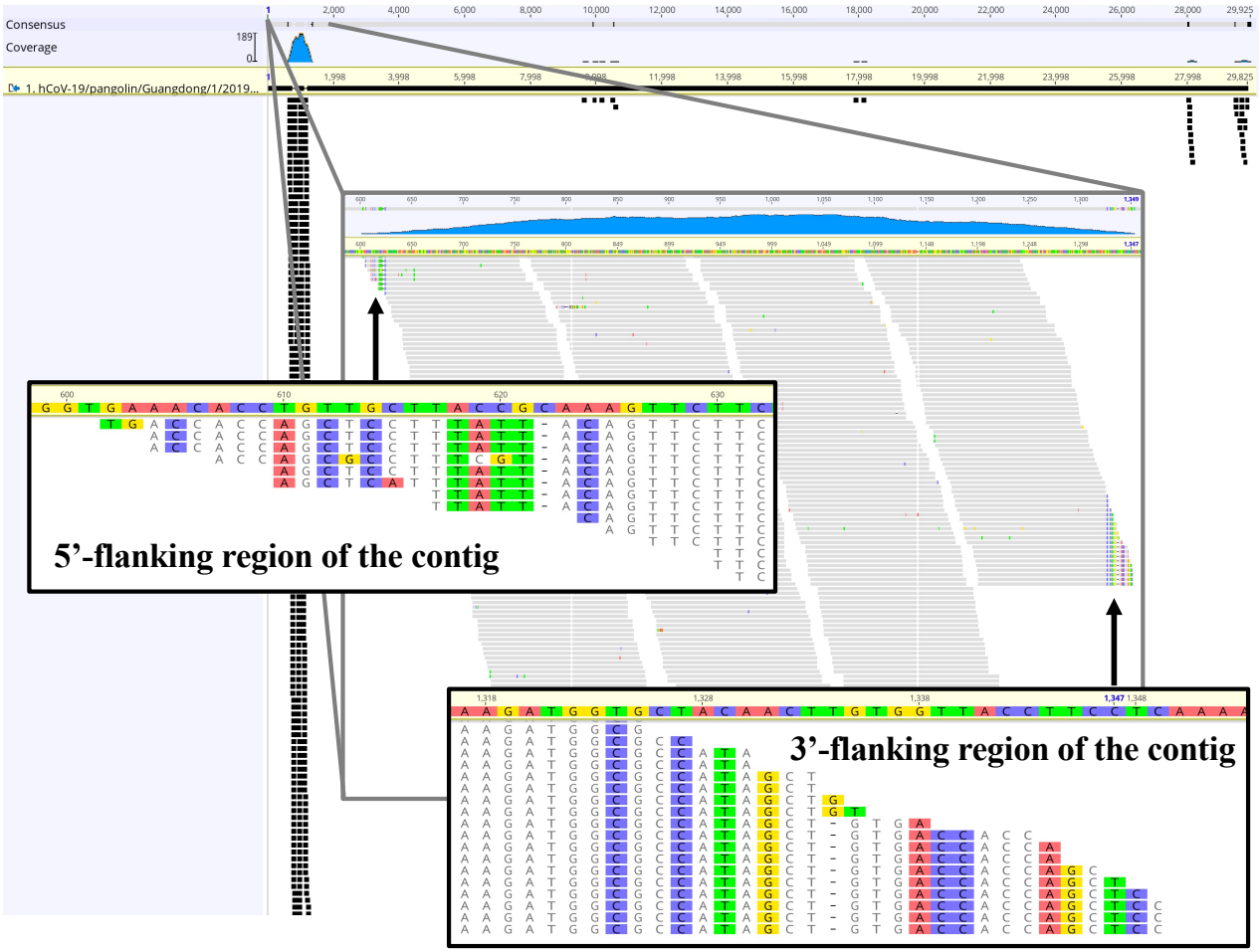
Figure 1. Identification and characterization of *circRNA-GDP5L-1*

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

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

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

TTAGAGCCACTTGTGAGTTCTGTGGTACTGAAAATTTGACTAAAGATGGCGCCATAGCTGTGACCACCAGCTCCT
TTATTACAGTTCTTCTTCGCAAGAACGGTAATAAAGGAGCTGGTGGTCACAGCTATGGCGCCGATCTAAAGTCCT
ATGACTTAGGTGACGAGCTGGGCACTGATCCTTATGAAGATTATCAAGAAAACTGGAACACTAAACATGGCAGTG
GTGTAACTCGTGAGCTCATGCGTGAGCTTAATGGGGGCGCATACACTCGCTATGTCGATAACAACTACTGTGGCC
CTGATGGCTACCCTCTTGAGTGCATTAAAGACTTGCTGGCGCGTGCTGGTAAAGCTTCTTGCACTTTGTCCGAAC
AACTGGACTTTCTTGACACTAAGAGAGGTGTGTACTGCTGCCGTGAGCATGACCATGAAATTGCATGGTACACGG
AAfCGCTCTGATAAGAGTTATGAATTGCAGACACCTTTTGAAATTAAACTGGCAAAGAAATTTGACACTTTTACT
GGGGAGTGCCCAAATTTTGTATTCCCTCTTAATTCAACTATCAAGACTATTCAACCTAGAGTTGAAAGGAAAAAG
CTTGATGGCTTTATGGGTAGGATTCGATCTGTCTACCCTGTTGCTTCACCTAATGAATGCAACCAAATGTGCCTG
TCAACTCTCATGAAGTGTAACCATTGTGGTGAAACTTCATGGCAGACAGGCGATTTTGTTAGAGCCACTTGTGAG
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-------
RaGT13    G----------T-------
VZC45     -----------T-------
VZXC21    -----------T-------
```

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

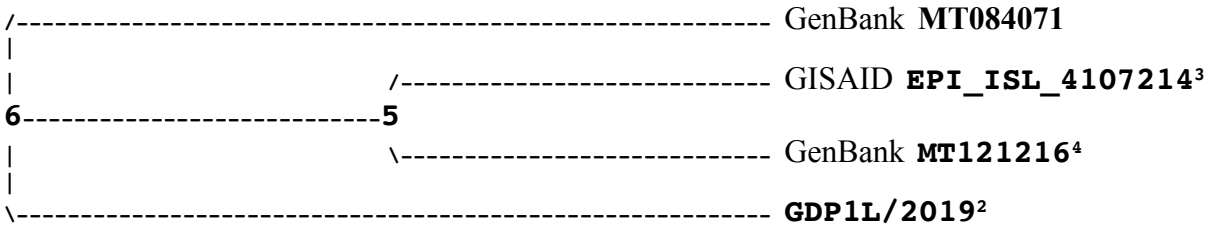
Illumina sequencing run			Viral genome assembly			Reads mapped to mitochondrial genomes			
Code - Tissue	**NCBI SRA**	**Reads**	**Reads**	**MC**	**MD**	**pangolin** NC_026781	**human** NC_012920	**tiger** NC_010642	**mouse** NC_005089
GD/P7L[1] - Lung 07	SRR10168378	38,091,846	285*	1.4X	57%	98,226	28	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	46,4866,240	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_410721[3]). 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 services[7]. 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_410721[3].
°partial RC in the VGF.

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

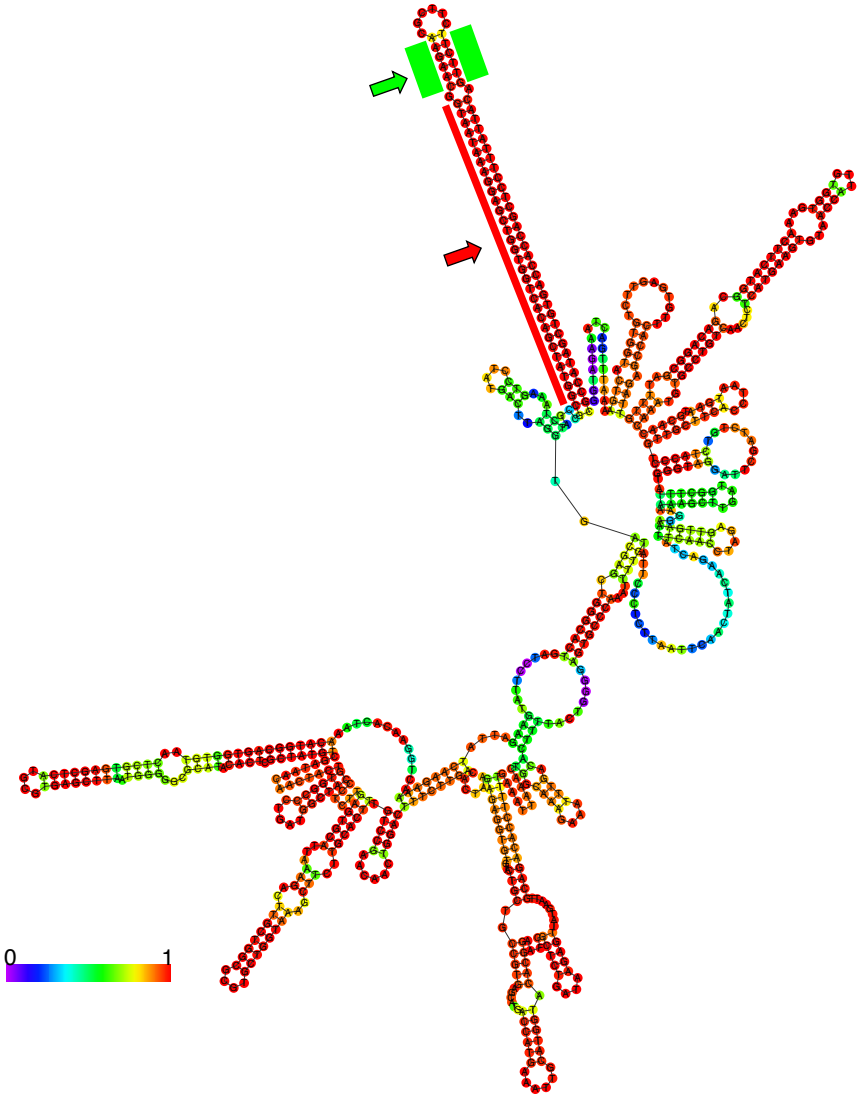
…CTGTGGTACTGAAAATTTGACTAAAGATGGCGCCATAGCTGTGACCACCAGCTCCTTTATTACAGTTCTTCTTC
GCAAGAACGGTAATAAAGGAGCTGGTGGTCACAGCTATGGCGCCGATCTAAAGTCCTATGACTTAGGTGACGAGC
TGGGCACTGATCCTTATGAAGATTATCAAGAAAACTGGAACACTAAACATGGCAGTGGTGTAACTCGTGAGCTCA
TGCGTGAGCTTAATGGGGGCGCATACACTCGCTATGTCGATAACAACTACTGTGGCCCTGATGGCTACCCTCTTG
AGTGCATTAAAGACTTGCTGGCGCGTGCTGGTAAAGCTTCTTGCACTTTGTCCGAACAACTGGACTTTCTTGACA
CTAAGAGAGGTGTGTACTGCTGCCGTGAGCATGACCATGAAATTGCATGGTACACGGAAfCGCTCTGATAAGAGT
TATGAATTGCAGACACCTTTTGAAATTAAACTGGCAAAGAAATTTGACACTTTTACTGGGGAGTGCCCAAATTTT
GTATTCCCTCTTAATTCAACTATCAAGACTATTCAACCTAGAGTTGAAAGGAAAAAGCTTGATGGCTTTATGGGT
AGGATTCGATCTGTCTACCCTGTTGCTTCACCTAATGAATGCAACCAAATGTGCCTGTCAACTCTCATGAAGTGT
AACCATTGTGGTGAAACTTCATGGCAGACAGGCGATTTTGTTAGAGCCACTTGTGAGTTCTGTGGTACTGAAAAT
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-------
RaGT13        G----------T-------
VZC45         -----------T-------
VZXC21        -----------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]

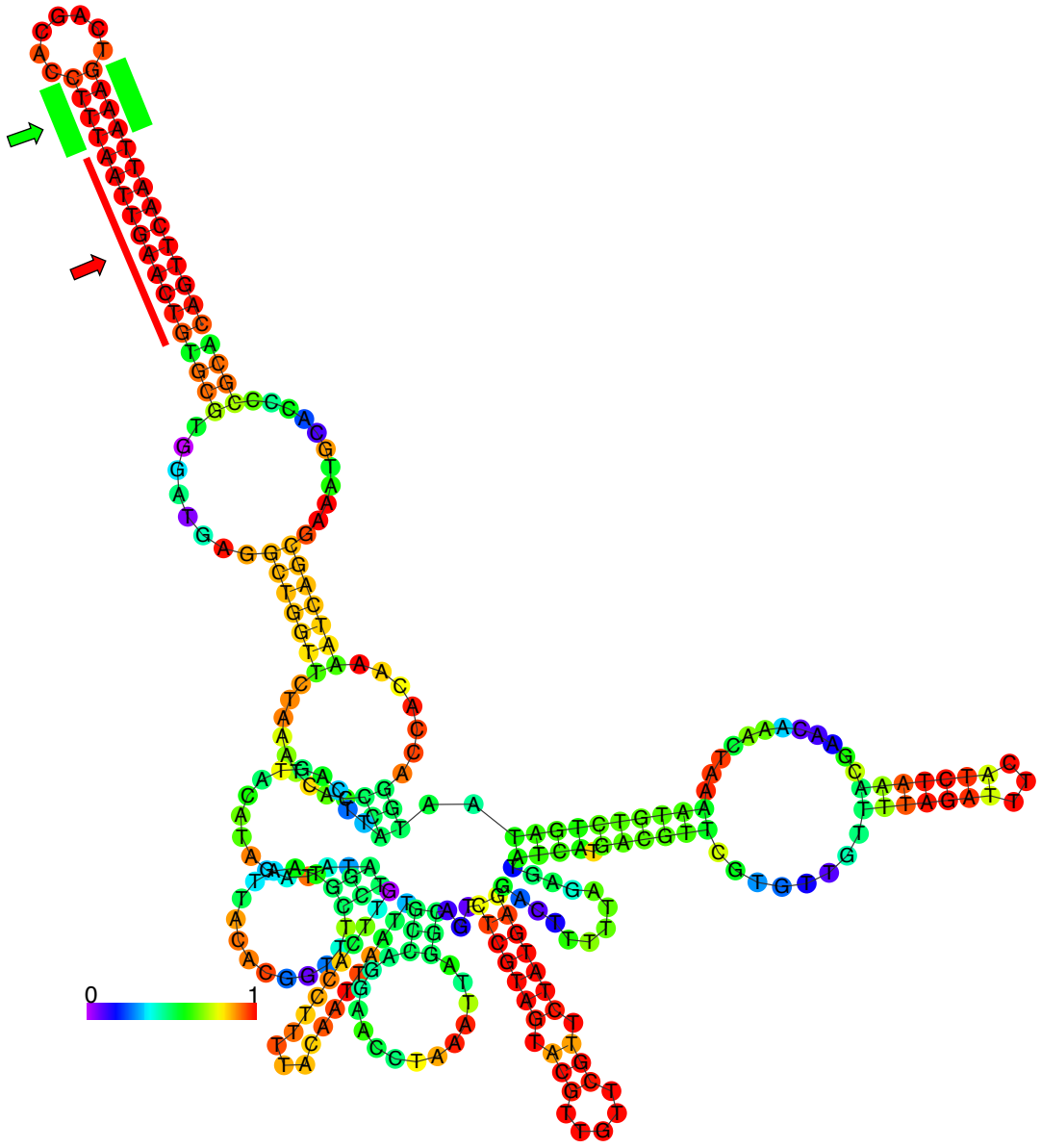
TCTGATAATGGACCACAAAATCAGCGAAATGCACCCCGCACAGTTCAATTAAAGTCAGCACCTTTAATTGAACTG
TGCGTGGATGAGGCTGGTTCTAAATCACCCATTCAGTACATAGATATAGGTAATTACACGGTTTCCTGTTCACCT
TTTACAATTAATTGCCAGGAACCTAAATTAGGCAGTCTCGTAGTACGTTGTTCGTTCTATGAGGACTTTTTAGAG
TATCATGACGTTCGTGTTGTTTTAGATTTCATCTAAACGAACAAACTAAAATGTCTGATAATGGACCACAAAATC
AGCGAAATGCACCCCGCACAGTTCAATTAAAGTCAGCACCTTTAATTGAACTGTGCGTGGATGAGGCTGGTTCTA
AATCACCCATTCAGTACA

Alignment of genomic sequences located between the two complementary motifs

```
GD/P          AAAGTCAGCACCTTT
GX/P          ---A--T-----A--
SARS-CoV-2    ---A------------
RaGT13        ---A------------
VZC45         ---A------------
VZXC21        ---A------------
```

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]

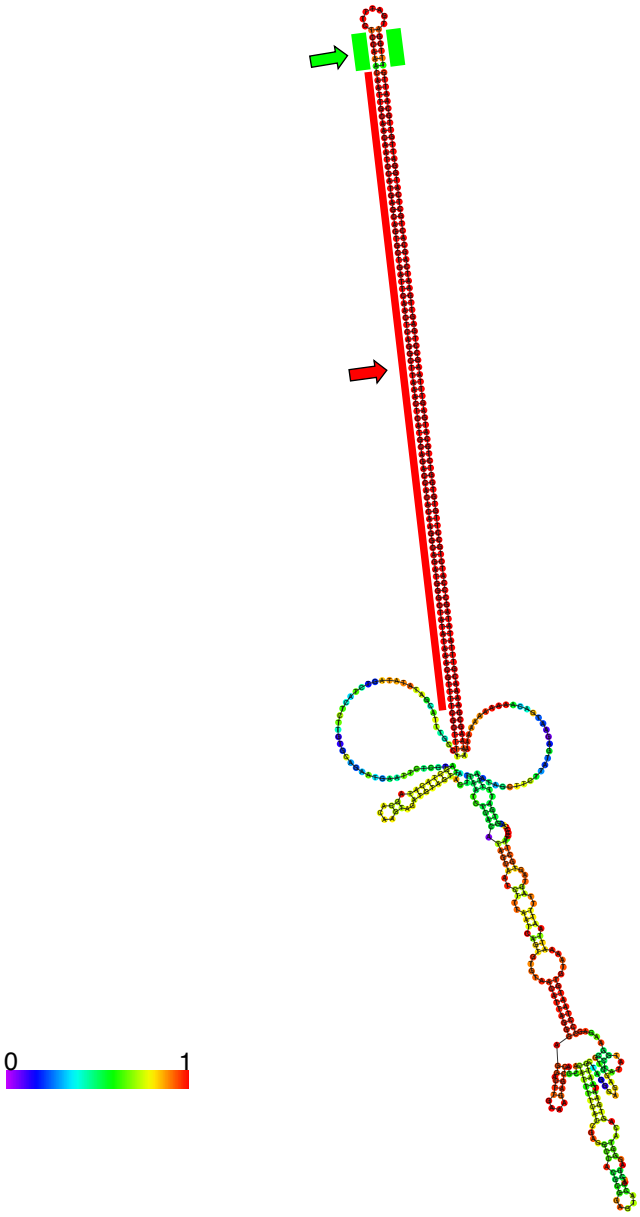
TTGTTTGGATGATTTCTCCAAACAATTGCAACAATCCATGAGCAGTGCTGATTCAACTCAGGCTTAAACTCATGC
AGACCACACAAGGCAGATGGGCTATATAAACGTTTTCGCTTTTCCGTTTACGATATATAGTCTACTCTTGTGCAG
AATGAATTCTCGTAGCTACATAGCACAAGTAGATGTAGTTAACTTTAATCTCACATAGCAATCTTTAATCAGTGT
GTAACATTAGGGAGGACTTGAAAGAGCCACCACATTTTCACCGAGGCCACGCGGAGTACGATCGAGGGTACAGTG
AATAATGCTAGGGAGAGCTGCCTATATGGAAGAGCCCTAATGTGTAAAATTAATTTTAGTAGTGCTATCCCCATG
TGATTTTAATAGCTTCTTATGAGAATGACAAAAAAAAAAAAAAAGCGAAAACGTTTATATAGCCCATCTGCCTTG
TGTGGTCTGCATGAGTTTAAGCCTGAGTTGAATCAGCACTGCTCATGGATTGTTGCAATTGTTTGGATGATTTCT
CCAAAC

Alignment of genomic sequences located between the two complementary motifs

```
GD/P        TTTGGATGATTTCTCCAAA
GX/P        -------------------
SARS-CoV-2  -------------------
RaGT13      ------C-------------
VZC45       ------A-------------
VZXC21      ------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=230.30 kcal/mol)



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

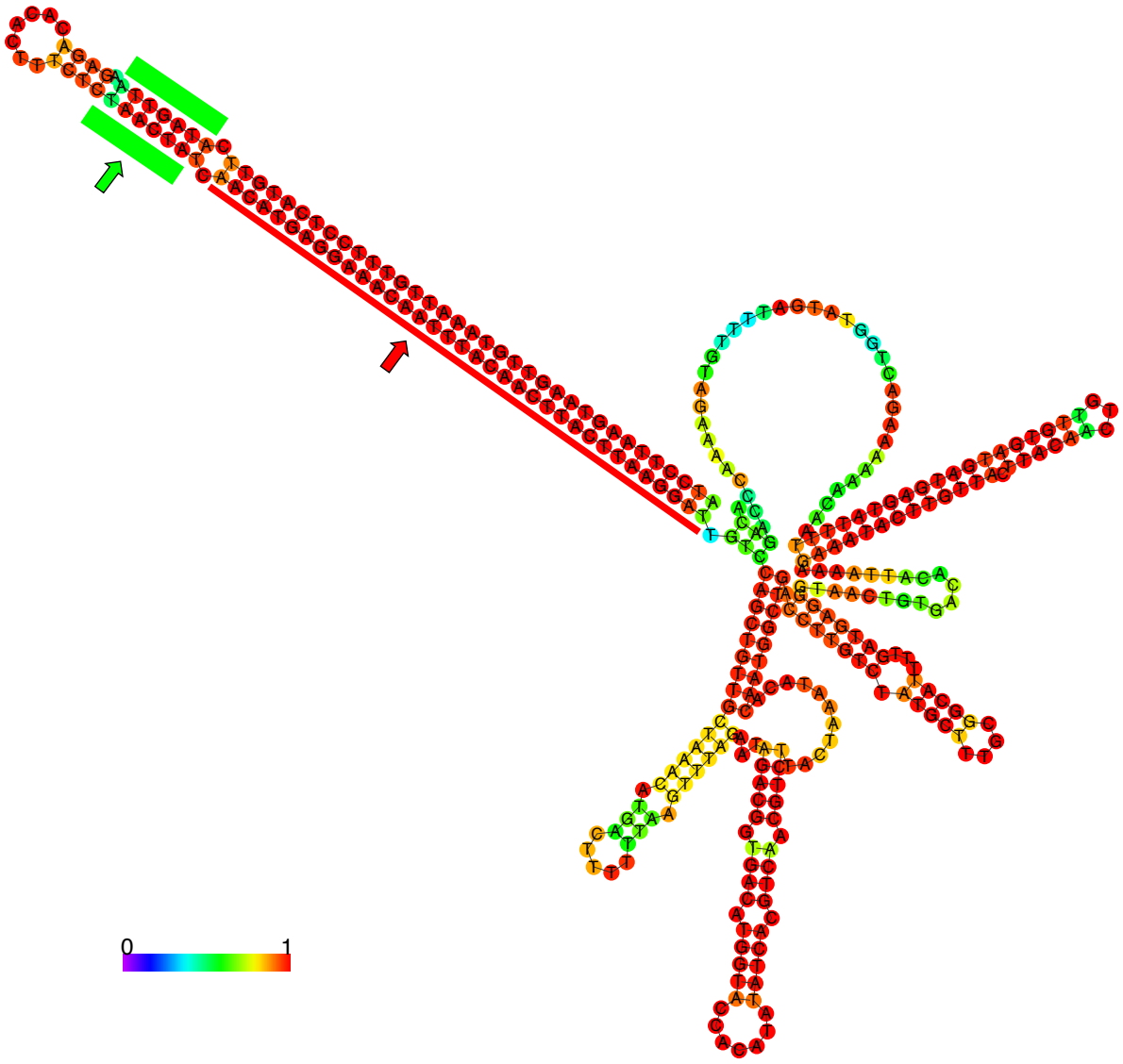
AACCCAGACAATCCTTAAGTAAGTTGTAAATTGTTTCCTCATGTTCATAGTTAAGAGACACACTTTCTCTAACTA
TCAACATGAGGAAACAATTTACAACTTACTTAAGGATTGTCCAGCTGTTGCTAAACATGACTTTTTTAAGTTTAG
AATAGACGGTGACATGGTACCACATATATCACGTCAACGTCTTACTAAATACACAATGGCTGACCTTGTCTATGC
TTTGCGGCATTTTGATGAGGGTAACTGTGACACATTAAAAGAAATACTTGTTACTTACAACTGTTGTGATGATGA
GTATTTTAACAAAAAAGACTGGTATGATTTTGTAGAAAACCCAGACAATCCTTAAGTAAGTTGTAAATTGTTTCC
TCATGTTCATAG

Alignment of genomic sequences located between the two complementary motifs

```
GD/P          CATAGTTAAGAGACACACTTTCTCTAACTATC
GX/P          -------------------T--------------
SARS-CoV-2    TG-----------------------------C-
RaGT13        TG-----------------------------C-
VZC45         -G-----------G-----AA-G--C-----C-
VZXC21        -G-----------G-----AA-G--C-----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]

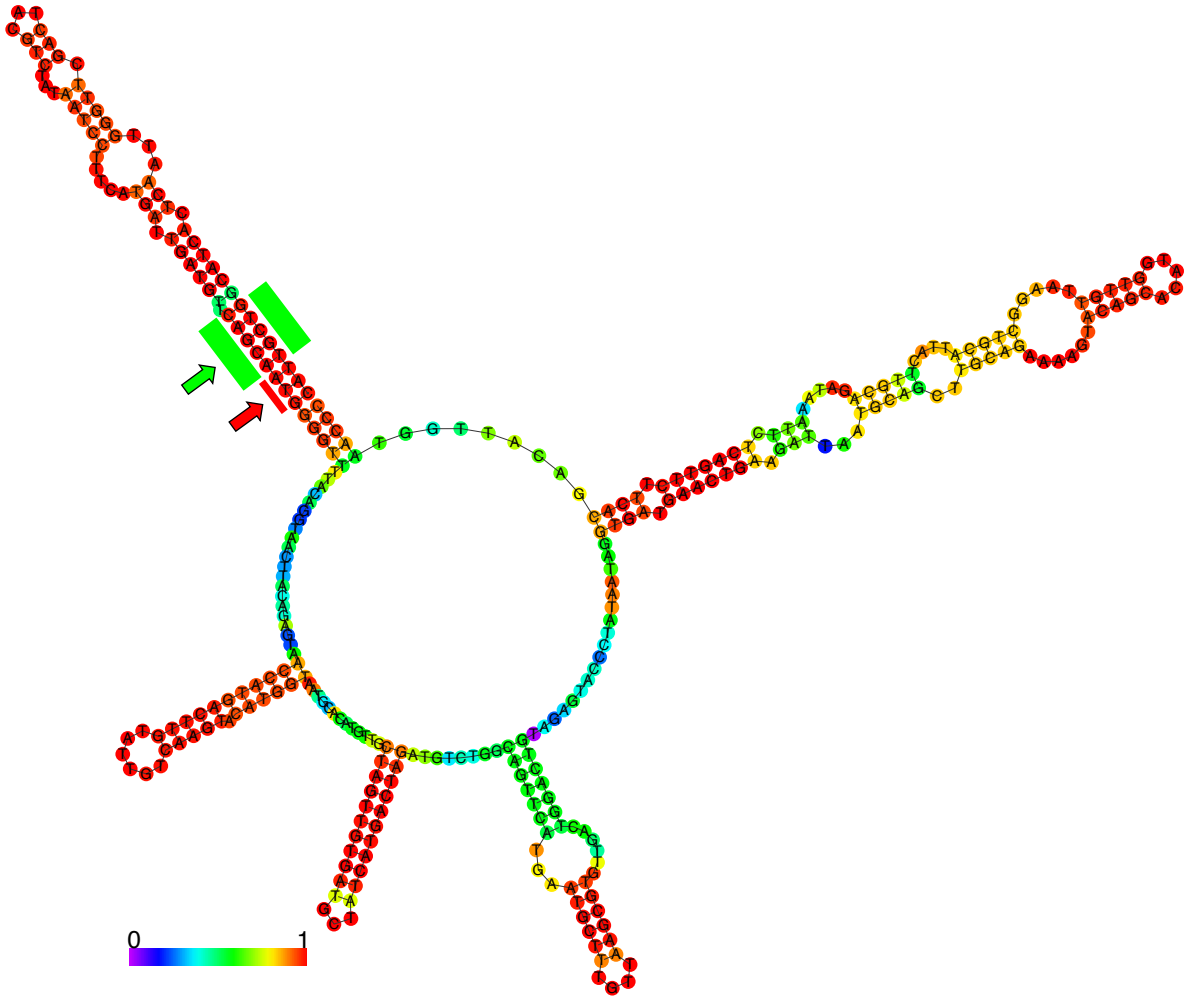
ACATGGTTGTTAAGGCTGCATTACTTGCAGATAAATTCTCAGTTCTTCACGACATTGGTAACCCCATTGCTGGCA
TCACTCAATTGGGTTCGACTACGTCTATAATCCTTTCATGATTGATGTTCAGCAATGGGGTTTTACAGGTAACTT
ACAGAGTAACCATGACTTGTATTGTCAAGTACATGGTAATGCACATGTTGCTAGTTGTGATGCTATCATGACTAG
ATGTCTGGCAGTTCATGAATGCTTTGTTAAGCGTGTTGACTGGACTGTAGAGTACCCTATAATAGGTGATGAACT
GAAGATTAATGCAGCTTGCAGAAAAGTACAGCACATGGTTGTTAAGGCTGCATTACTTGCAGATAAATTCTCAGT
TCTTCACGACATTGGTAACCCCATTGCTGGCATCACTCAATTGGGTTCGACTACGTCTATAATCCTTTCATGATT
GATGT

Alignment of genomic sequences located between the two complementary motifs

```
GD/P        TGCTGGCATCACTCAATTGGGTTC.........TTTCATGATTGATGTTCAGCA
GX/P        --T----------G----A--T.........----------------A--
SARS-CoV-2  --T----------T--T-----A--T.........G--T----------A--
RaGT13      --T----------T--T-----A--T.........G--T----------A--
VZC45       --T---A-------TG-G--C--T.........G--T----------C----
VZXC21      --T---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]

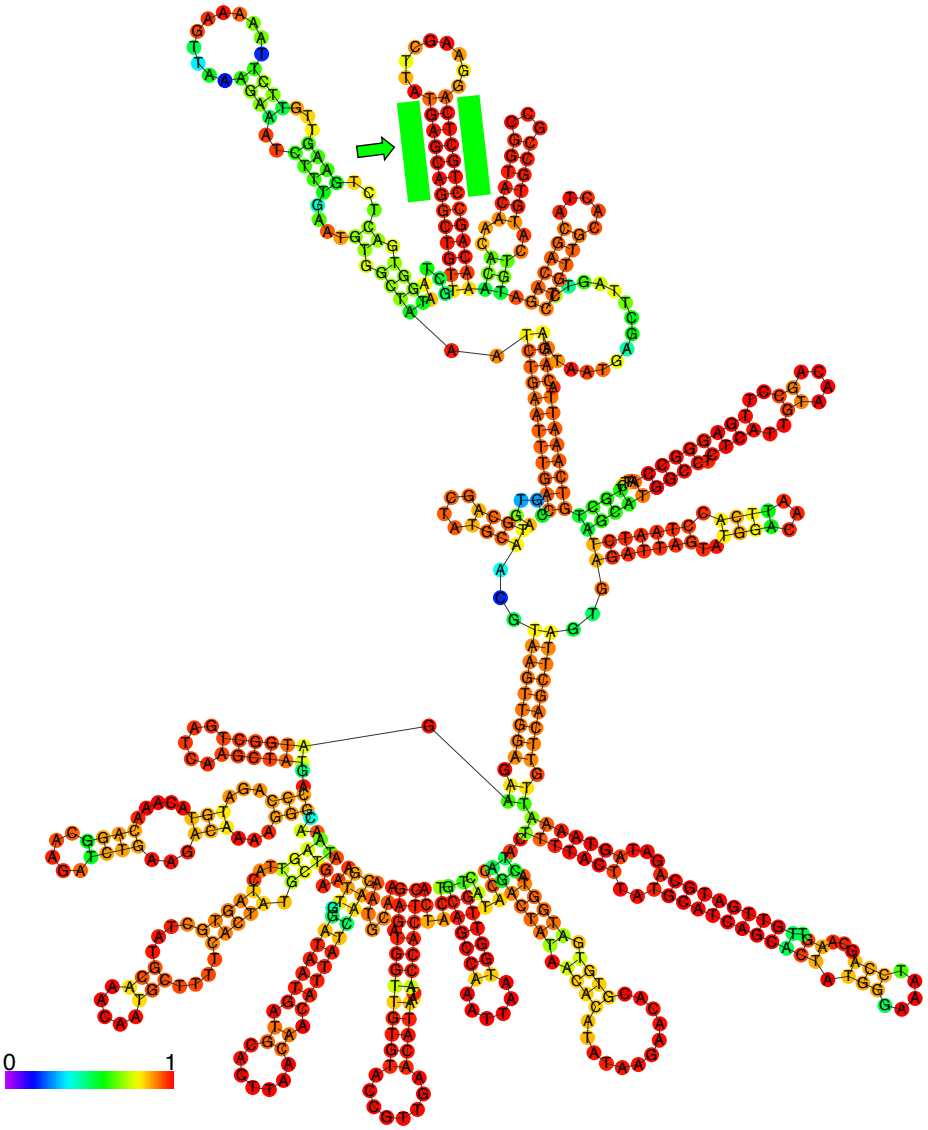
GTTGCACTACGACAGATGTCATGTGCCGCCGGTACAACACAAACAGCCTGCTCAGGAAGCTTATGAGCAGGCTGT
TGCTAATGGTGACTCTGAAGTTGTTCTTAAAAAGTTAAAGAAATCTTTGAATGTGGCTAAATCTGAATTTGACCG
TGATGCAGCTATGCAACGTAAGTTGGAGAAGATGGCTGATCAAGCTATGACCCAGATGTACAAACAGGCAAGATC
TGAAGACAAAAGGGCAAAAGTTACTAGTGCTATGCAAACAATGCTTTTCACTATGCTTAGAAAGTTGGATAATGA
TGCACTTAACAACATTATCAACAATGCAAGAGATGGTTGTGTACCGTTGAACATAATACCACTCACTACTGCAGC
CAAATTAATGGTTGTCATACCAGACTATAACACATATAAGAACACGTGTGATGGTACTACTTTTACTTATGCATC
AGCACTATGGGAAATCCAGCAAGTTGTTGATGCAGATAGTAAAATTGTTCAGCTTAGTGAGATTAGTATGGACAA
TTCACCTAATCTAGCATGGCCTCTCATTGTAACAGCCTTGAGGGCCAATTCTGCTGTCAAATTACAGAATAATGA
GCTTAGTCCTGTTGCACTACGACAGATGTCATGTGCCGCCGGTACAACACAAACAGCCTGCTCAGGAAGCTTATG
AGCAGGCTGTTGCTAATGGTGACTCTGAAGTTGTTCTTAAAAAGTTAAAGAAATCTTTGAATG

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

```
GD/P          ACTGCTCAGGAAGCTTATGAGCAGGCTGT
GX/P          --A------A----------------A--G--
SARS-CoV-2    ---------A---------------------
RaGT13        --C------A---------------------
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]

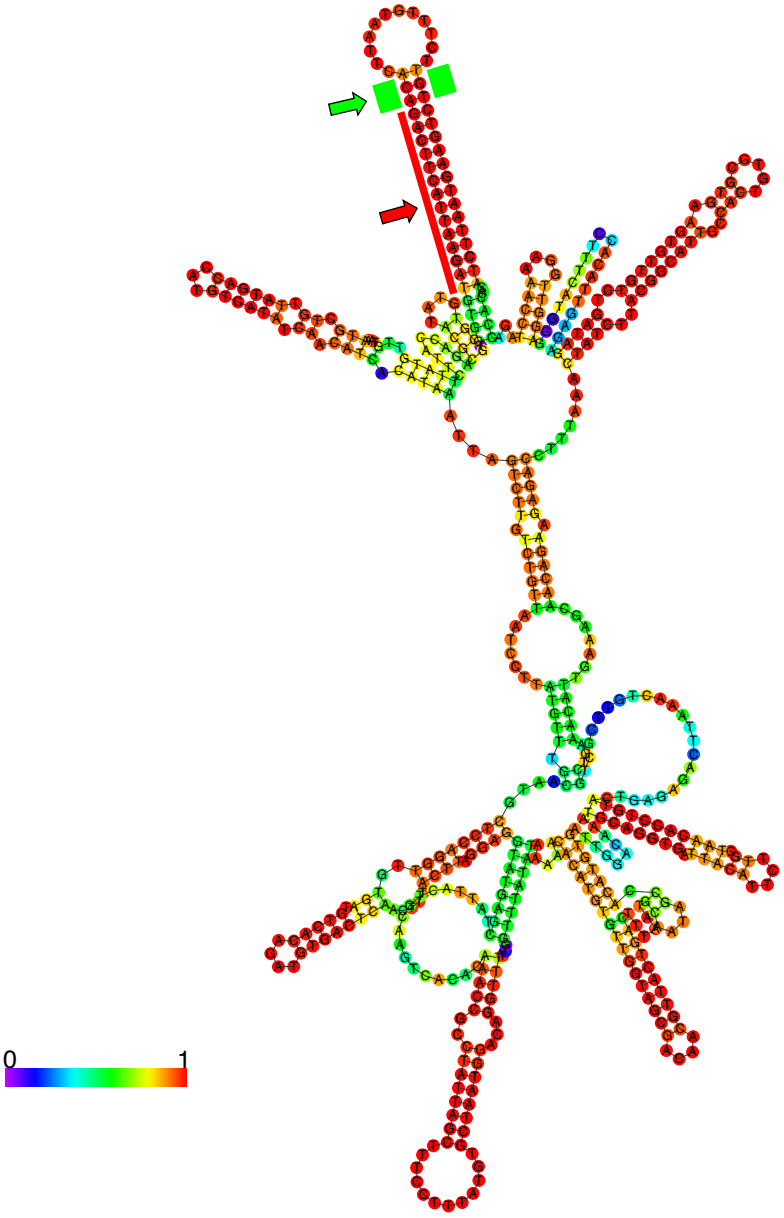
CCANCCACATCTTAATGAAGTCTGTTCTTTGTAATTCACAGACTTCATTAAGATGTGGTGCGTGTATACGCAGAC
CATTCTTATGTTGTAAATGCTGTTATGACCATGTCATATCAACATCACATAAATTAGTCTTGTCTGTTAATCCTT
ATGTTTGCAATGCTCCAGGTTGTGATGTCACAGATGTGACTCAACTTTACTTAGGAGGTATGAGCTATTACTGCA
AGTCACACAAACCGCCTATTAGCTTTCCTTTATGTGCTAATGGACAGGTTTTTGGTTTATATAAAAACACATGTG
TTGGTAGCGACAACGTTACTGACTTTAATGCAATAGCCACATGTGATTGGACAAATGCAGGTGATTACATTCTTG
CTAACACCTGTACTGAGAGACTTAAACTGTTCGCTGCTGAAACATTGAAAGCAACAGAAGAGACCTTTAAACTAT
CTTACGGCATTGCCACTGTGCGTGAAGTGTTGTCTGATAGAGAGTTACACCTTTCATGGGAGGTTGGAAAACCTA
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]

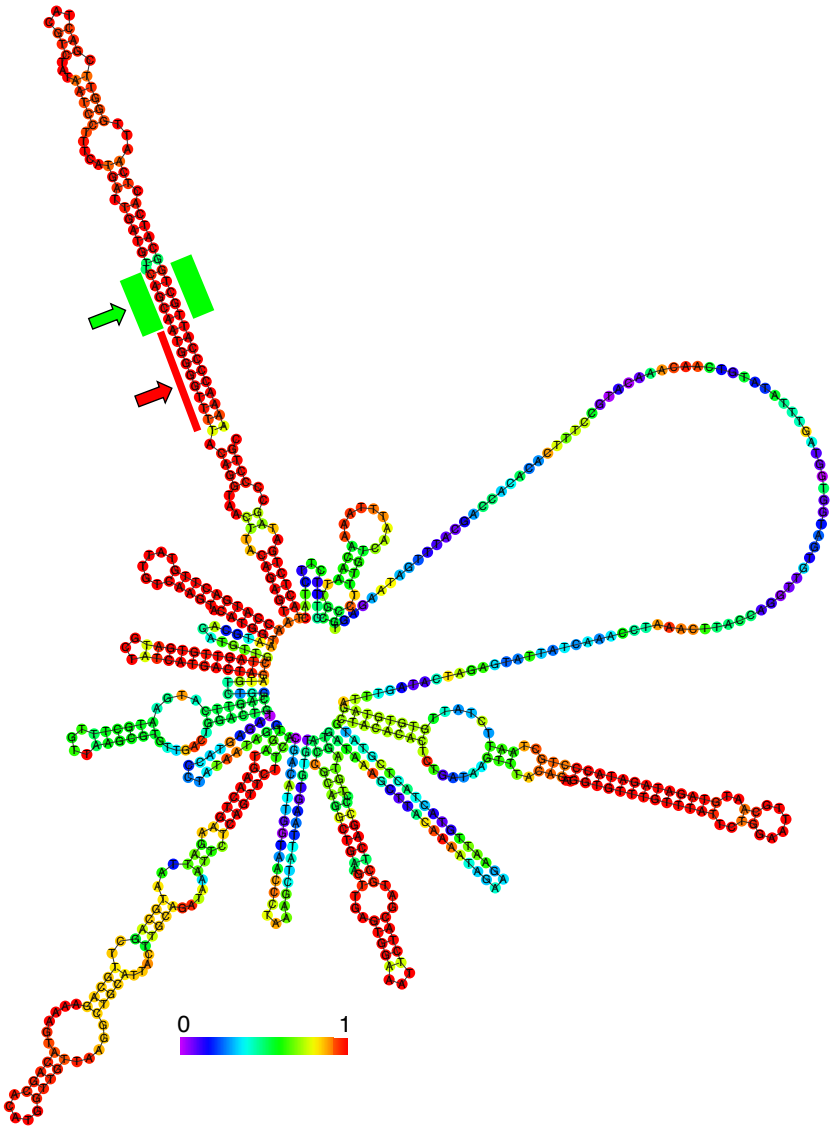
...CTACTACTCTGATAGCCCCTGCAAAACCCCATTGCTGGCATCACTCAATTGGGTTCGACTACGTCTATAATCCT
TTCATGATTGATGTTCAGCAATGGGGTTTTACAGGTAACTTACAGAGTAACCATGACTTGTATTGTCAAGTACAT
GGTAATGCACATGTTGCTAGTTGTGATGCTATCATGACTAGATGTCTGGCAGTTCATGAATGCTTTGTTAAGCGT
GTTGACTGGACTGTAGAGTACCCTATAATAGGTGATGAACTGAAGATTAATGCAGCTTGCAGAAAAGTACAGCAC
ATGGTTGTTAAGGCTGCATTACTTGCAGATAAATTCTCAGTTCTTCACGACATTGGTAACCCTAAAGCTATTAAG
TGTGTACCGCAGGCTGAAGTTGAGTGGAAATTCTACGATGCTCAGCCCTGTAGTGATAAAGCTTACAAAATAGAA
GAATTGTACTACTCGTATGCTACACACTCTGATAAGTTTACAGATGGTGTTTGTTTATTCTGGAATTGCAATGTA
GATAGATACCCTGCTAATTCTATTGTGTGTAGATTTGATACTAGAGTATTATCAAACCTAAACTTACCAGGTTGT
GATGGTGGTAGTTTATATGTCAACAAACATGCCTTTCACACACCAGCATTTGATAAGAGTGCCTTTGTCAATTTA
AAACAATTGCCTTTCTTCTACTACTCTGATAGCCCCTGCAAAACCCCATTGCTGAACATCAATCATGAAAGG...

Alignment of genomic sequences located between the two complementary motifs

```
GD/P        TGCTGGCATCACTCA.....CATGATTGATGTTCAGCA
GX/P        --T----------.....----------A--
SARS-CoV-2  --T--------T--T.....T---------A--
RaGT13      --T--------T--T.....T---------A--
VZC45       --T---A-------T.....T----------C---
VZXC21      --T---A-------T.....----------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=166.30 kcal/mol)



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

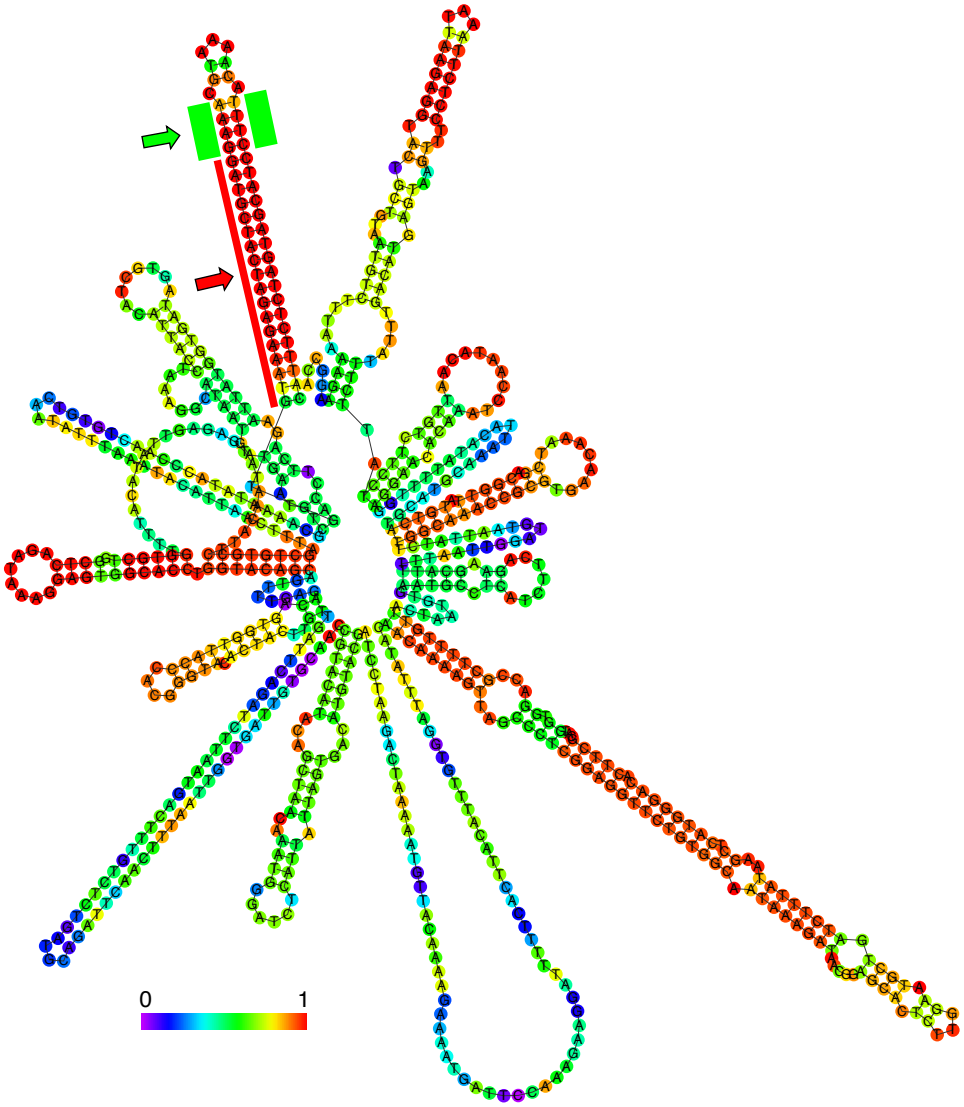
TAAAAGAAGGCCACATTTCTCTAGTAGCATCCTTTACAAAATGCAAAGGATGCTACTAGAGAAATGTGACCTTCA
GAATTATGGTGATAGTGCTACATTACCTAAAGGCATAATGATGAATGTCGCAAAATATACCCAACTGTGTCAATA
TTTAAATACATTAACTTTAGCTGTGCCTTACAATATGAGAGTTATACATTTTGGTGCTGGCTCAGATAAAGGAGT
GGCACCTGGTACAGCAGTTTTGAGACAGTGGTTACCCACGGGTACACTACTTGTTGATTCAGATCTTAATGACTT
TGTCTCTGATGCAGATTCAACTTTAATTGGTGATTGTGCAACCGTACATACAGCTAACAAATGGGATCTCATTAT
TAGTGACATGTACGATCCTAAGACTAAAAATGTTACAAAAGAAAATGATTCCAAAGAAGGATTTTTCACTTACAT
TTGTGGATTTATACAACAAAAGTTAGCCCTCGGAGGTTCTGTGGCAATAAAGATAACGGAGCACTCTTGGAATGC
TGATCTTTATAAGCTCATGGGACACTTCGCATGGTGGACCGCTTTTGTTACTAATGTGAATGCCTCATCTTCAGA
AGCATTTTTAATTGGATGTAATTATCTTGGCAAACCGCGTGAACAAATCGACGGTTATGTCATGCATGCAAATTA
CATATTTTGGAGGAACACAAATCCAATACAATTGTCTTCCTATTCTTTATTTGACATGAGTAAGTTTCCTCTTAA
ATTAAGAGGTACTGCTGTAATGTCTTTAAAAGAAGGCCAAATTTCTCTAGTAGCATCCTTTACAAAATGCAAAGG
ATGCTACT

Alignment of genomic sequences located between the two complementary motifs

```
GD/P          CTTTACAAAATGCAAAG
GX/P          ---------------G--
SARS-CoV-2    -----------------
RaGT13        --C---------------
VZC45         --C--T-------G----
VZXC21        ---------------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]

AAGAATTAATCAGACAAGGAACTGATTACAAACAATGGCCGCATTGTTAGCTGGACTTCCCTATGGTGCTAACAA
AGAAGGCATCATATGGGTTGCAAATGAGGGAGCCTTGAATACACCTAAAGATCACATTGGCACCCGAAATCCTGC
TAACAATGCTGCAATCGTGCTACAACTTCCTCAAGGAACAACATTGCCAAAAGGCTTCTACGCAGAAGGGAGCAG
AGGCGGCAGTCAAGCTTCTTCTCGTTCCTCATCACGTAGTCGCAACAGTTCAAGAAACACAACTCCAGGCAGCAG
CAGGGGAACTTCTCCTGCTAGGATGGCTGGCAATGGTGGTGATGCTGCTCTTGCTTTGCTGCTGCTTGACAGGTT
GAACCAACTTGAGAGCAAAATGTCTGGTAAAGGCCAACAACAACAAGGCCAAACTGTCACTAAGAAATCCGCTGC
AGAGGCTTCTAAGAAACCTCGCCAAAAACGTACTGCCACCAAACAATACAATGTAACACAAGCTTTTGGCAGACG
TGGTCCAGAACAAACCCAAGGAAACTTTGGGGATCAAGAATTAATCAGACAAGGAACTGATTACAAACAATGGCC
GCATTGTTAGCTGGACTTCCCTATGGTGCTAACAAAGAAGGCATCATATGGGTTGCAAATGAGGGAGCCTTGAA...

Alignment of genomic sequences located between the two complementary motifs

```
GD/P        AGCTGGACTTCCCTATGGTGCT
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)