

1 **The SARS-CoV-2-like virus found in captive pangolins from Guangdong should be**  
2 **better sequenced.**

3

4 Alexandre Hassanin

5

6 Institut de Systématique, Evolution, Biodiversité, UMR 7205 CNRS, MNHN, Sorbonne

7 Université, EPHE, Université des Antilles, Muséum National d'Histoire Naturelle, CP 51, 57

8 rue Cuvier, 75231 PARIS Cedex 05 France.

9 Email: [alexandre.hassanin@mnhn.fr](mailto:alexandre.hassanin@mnhn.fr)

10

11

12 **Viruses closely related to SARS-CoV-2, which is the virus responsible of the**  
13 **Covid-19 pandemic, were sequenced in several Sunda pangolins (*Manis javanica*) seized**  
14 **in the Guangdong and Guangxi provinces of China between 2017 and 2019<sup>1-3</sup>. These**  
15 **viruses belong to two lineages: one from Guangdong (*GD/P*) and the other from**  
16 **Guangxi (*GX/P*). The *GD/P* viruses are particularly intriguing as the amino-acid**  
17 **sequence of the receptor binding domain of the spike protein is very similar to that of**  
18 **the human SARS-CoV-2 virus (97.4%)<sup>2</sup>. This characteristic suggests that *GD/P* viruses**  
19 **are capable of binding human ACE2 receptor and may therefore be able to mediate**  
20 **infection of human cells. Whereas all six *GX/P* genomes were deposited as annotated**  
21 **sequences in GenBank, none of the two *GD/P* genomes assembled in previous studies<sup>2,3</sup>**  
22 **are currently available. To overcome this absence, I assembled these genomes from the**  
23 **Sequence Read Archive (SRA) data available for SARS-CoV-2-like viruses detected in**  
24 **five captive pangolins from Guangdong. I found the genome assemblies of *GD/P* virus of**  
25 **poor quality, having high levels of missing data. Additionally, unexpected reads in the**

26 **Illumina sequencing data were identified. The *GD/P2S* dataset<sup>2</sup> contains reads that are**  
27 **identical to SARS-CoV-2, suggesting either the coexistence of two SARS-CoV-2-like**  
28 **viruses in the same pangolin or contamination by the human virus. In the four other**  
29 ***GD/P* datasets<sup>1</sup> many mitochondrial reads from pangolin were identified, as well as from**  
30 **three other species, namely, human, mouse and tiger. Importantly, I only identified**  
31 **three polymorphic nucleotide sites between the five *GD/P* sequences. Such low levels of**  
32 **polymorphism may reasonably be accounted for by sequencing errors alone, thus**  
33 **raising the possibility that the five pangolins seized in Guangdong in March 2019 were**  
34 **infected by the same virus strain, most probably during their captivity.**

35

36 For each of the five *GD/P* samples sequenced on Illumina platforms (**Table 1**), I  
37 mapped the reads to the reference genome of the human SARS-CoV-2 virus (GenBank  
38 accession number: NC\_045512)<sup>4</sup> using Geneious Prime® 2020.0.3 and the “High sensitivity”  
39 option (maximum mismatch: 40%). Then, mapped reads were used for *de novo* assembly. All  
40 contigs were aligned to the SARS-CoV-2 genome and assembled into a consensus sequence  
41 used as reference to discover more reads in each *GD/P* dataset. All five *GD/P* genome  
42 assemblies (*GD/P2S*, *GD/P7L*, *GD/P8L*, *GD/P9L*, and *GD/P11L*) were of poor-quality  
43 having been previously sequenced at low depth (mean coverage between 0.2 and 6.5X) and  
44 therefore containing high levels (between 19% and 99%) of missing data (**Table 1**).

45 All 2633 reads sequenced for *GD/P2S*<sup>2</sup> were mapped on the SARS-CoV-2 genome,  
46 indicating that all non-viral reads were removed. Curiously, within this pangolin dataset, I  
47 found 11 reads identical to the human SARS-CoV-2 genome (numbered 62, 412, 514, 786,  
48 787, 1417, 1440, 1498, 2222, 2231, and 2403) and four reads very similar to SARS-CoV-2  
49 (only a single mutation in reads 102, 502, 1390, and 1882) whereas several homologous reads  
50 (between 3 and 35) were found more divergent (between 2 and 9 mutations) but identical to

51 other *GD/P* sequences. Two hypotheses can be proposed to explain this result: (1) the  
52 *GD/P2S* sample contained two different SARS-CoV-2-like viruses; or (2) the sample had  
53 been contaminated by the human SARS-CoV-2 virus. To choose between these two  
54 hypotheses the full raw dataset for *GD/P2S* is required. All Illumina reads generated for  
55 *GD/P2S* should consequently be deposited by the authors of the study<sup>2</sup> in NCBI without any  
56 filtration process.

57 Less filtered SRA data were provided for other pangolin samples, i.e., *GD/P7L*,  
58 *GD/P8L*, *GD/P9L*, and *GD/P11L*<sup>1</sup>. It was therefore possible to extract mitochondrial  
59 sequences from the host (the Sunda pangolin) in order to determine the geographic origin of  
60 the seized animals. As shown in **Table 1**, many mitochondrial reads of pangolin ( $\geq 7727$ )  
61 were found for these four *GD/P* samples. Pairwise distances between assembled  
62 mitochondrial genomes were between 0.12% - 0.41%, confirming the four samples were  
63 collected on different pangolins, and suggesting they came from different localities in  
64 Southeast Asia. It should be noted mitochondrial reads could not be analysed for *GD/P2S*  
65 because the authors<sup>2</sup> have removed all non-viral reads. It is therefore impossible to prove that  
66 the pangolin (*GD/P2S*) analysed by Lam *et al.*<sup>2</sup> differs from those studied by Liu *et al.*<sup>1</sup>.  
67 Surprisingly, numerous mitochondrial reads from other species (nucleotide identity = 100%)  
68 were also detected in the sequencing data. The *GD/P7L* dataset contains 1634 reads of mouse  
69 (*Mus musculus*), representing 2% of pangolin reads. The *GD/P8L* dataset contains 183 mouse  
70 reads (2%), and 1333 human reads (17%) (M7b haplogroup, which is found in humans from  
71 China and Southeast Asia). The *GD/P9L* dataset includes 3447 reads of tiger, subspecies  
72 *Panthera tigris altaica* (25%) and the *GD/P11L* dataset includes 1394 tiger reads (<1%). This  
73 unexpected range of mammalian species that I have identified clearly warrants an  
74 explanation. The most likely hypotheses are that laboratory experiments were contaminated

75 by RNA molecules from multiple organisms, or that different RNA extractions were pooled  
76 into the same library.

77         When the five partial *GD/P* genomes were compared to each other, only three  
78 nucleotide sites were found to be polymorphic: (1) position 1807: A in 21 *GD/P2S* reads  
79 *versus* C in four *GD/P8L* reads; (2) position 5228: A in two *GD/P9L* reads, one *GD/P7L* read,  
80 and one *GD/P8L* read *versus* C in three *GD/P2S* reads; (3) position 24979: G in 15 *GD/P2S*  
81 reads and one *GD/P8L* read *versus* A in three *GD/P7L* reads. Considering the low-coverage  
82 of *GD/P* genomes, these differences can be interpreted as sequencing errors. I suggest  
83 therefore that all five pangolins were infected by the same *GD/P* virus strain, approximately  
84 at the same time, and most probably during their captivity<sup>5</sup>. I decided therefore to assemble a  
85 consensus *GD/P* genome by pooling together all reads sequenced from the five *GD/P*  
86 samples. The quality of the assembled genome is still very low with a mean coverage of 13X  
87 being composed of 26 fragments and containing 6.4 % missing data by comparison with the  
88 human SARS-CoV-2 genome. In particular, the sequence of the gene coding for the spike  
89 protein is composed of three fragments and includes 6.4 % missing data.

90         Reliable whole genome sequences of the virus detected in pangolins from Guangdong  
91 are crucial to better understand the origin of Covid-19. These sequences could be used in  
92 many studies, in particular to estimate mutation and recombination rates during the  
93 evolutionary history of viruses related to SARS-CoV-2. For this reason, I strongly encourage  
94 Chinese researchers to re-sequence the *GD/P* genome more deeply in order to reach a mean  
95 coverage of 30X, as often recommended for genomic studies<sup>6</sup>. In this spirit, I would  
96 encourage editorial boards of relevant journals to maintain their data publication standards by  
97 requiring authors to fully make available their unfiltered data for the benefit of scientific  
98 collaboration in tackling the current pandemic.

99

100 **Acknowledgements**

101 I would like to thank Anne Ropiquet and Huw Jones for helpful comments on the first version  
102 of the manuscript.

103

104 **References**

- 105 1. Liu, P. *et al.* Viral Metagenomics Revealed Sendai Virus and Coronavirus Infection of  
106 Malayan Pangolins (*Manis javanica*). *Viruses* **11**, (2019).
- 107 2. Lam, T.T. *et al.* Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins.  
108 *Nature* doi: 10.1038/s41586-020-2169-0. (2020).
- 109 3. Zhang, T. *et al.* Probable pangolin origin of SARS-CoV-2 associated with the COVID-19  
110 outbreak. *Curr Biol*, **30**, (2020).
- 111 4. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China.  
112 *Nature* **579**, (2020).
- 113 5. Hassanin, A. *et al.* Covid-19: natural or anthropic origin? *Mammalia* in press (2020).
- 114 6. Sims, D. *et al.* Sequencing depth and coverage: key considerations in genomic analyses.  
115 *Nat Rev Genet* **15**, (2014).

**Table 1. Analyses of SRA data available for SARS-CoV-2-like viruses detected in captive pangolins from Guangdong**

Illumina sequencing run			GD/P viral genome			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/P2S <sup>2</sup> - Scale	SRR11093265	2,633	2,604	6.5X	29%	0	0	0	0
GD/P7L <sup>1*</sup> - Lung 07	SRR10168378	38,091,846	285	1.4X	57%	98,226	28	0	1,634
GD/P8L <sup>1*</sup> - Lung 08	SRR10168377	32,829,850	1,078	5.3X	19%	7,727	1,333	0	183
GD/P9L <sup>1</sup> - Lung 09	SRR10168376	36,135,230	36	0.2X	88%	13,770	47	3,447	0
GD/P11L <sup>1</sup> - Lung 11	SRR10168375	44,440,374	10	0.2X	99%	807,747	24	1,394	3

\*: SRA data used by Zhang *et al.*<sup>3</sup>. Abbreviations: MC: mean coverage; MD: missing data.