



HAL
open science

Whole genome sequencing and phylogenetic characterization of a novel bat-associated picornavirus-like virus with an unusual genome organization

Sarah Temmam, Vibol Hul, Thomas Bigot, Mael Bessaud, Delphine Chrétien, Thavry Hoem, Christopher Gorman, Veasna Duong, Philippe Dussart, Julien Cappelle, et al.

► To cite this version:

Sarah Temmam, Vibol Hul, Thomas Bigot, Mael Bessaud, Delphine Chrétien, et al.. Whole genome sequencing and phylogenetic characterization of a novel bat-associated picornavirus-like virus with an unusual genome organization. *Infection, Genetics and Evolution*, 2020, 78, 5 p. 10.1016/j.meegid.2019.104130 . hal-02625131

HAL Id: hal-02625131

<https://hal.inrae.fr/hal-02625131>

Submitted on 21 Jul 2022

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.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

1 **Whole genome sequencing and phylogenetic characterization of a novel bat-**
2 **associated picornavirus-like virus with an unusual genome organization.**

3 Sarah Temmam¹, Vibol Hul², Thomas Bigot^{1,3}, Maël Bessaud⁴, Delphine Chrétien¹, Thavry
4 Hoem⁵, Christopher Gorman², Veasna Duong², Philippe Dussart², Julien Cappelle^{5,6,7}, Marc
5 Eloit^{1,8,*}.

6 ¹ Institut Pasteur, Biology of Infection Unit, Pathogen Discovery Laboratory, Inserm U1117, Paris,
7 France.

8 ² Virology Unit, Institut Pasteur du Cambodge, Institut Pasteur International Network, Phnom Penh,
9 Cambodia.

10 ³ Institut Pasteur – Bioinformatics and Biostatistics Hub – Computational Biology department, Institut
11 Pasteur, USR 3756 CNRS – Paris, France.

12 ⁴ Institut Pasteur, Viral Populations and Pathogenesis Unit – WHO Collaborating Center for
13 Enteroviruses, Paris, France.

14 ⁵ Epidemiology and Public Health Unit, Institut Pasteur du Cambodge, Institut Pasteur International
15 Network, Phnom Penh, Cambodia.

16 ⁶ UMR ASTRE, CIRAD, INRA, Université de Montpellier, Montpellier, France.

17 ⁷ UMR EpiA, VetAgro Sup, INRA, Marcy l'Etoile, France.

18 ⁸ National Veterinary School of Alfort, Paris-Est University, Maisons-Alfort, 94704 Cedex, France.

19 * Corresponding author: Pr. Marc Eloit; marc.eloit@pasteur.fr.

20 **Abstract**

21 The order *Picornavirales* is one of the most important viral orders in terms of virus diversity
22 and genome organizations, ranging from a mono- or bi-cistronic expression strategies to the recently
23 described poly-cistronic *Polycipiviridae* viruses. We report here the description and characterization of
24 a novel picorna-like virus identified in rectal swabs of frugivorous bats in Cambodia that presents an
25 unusual genome organization. Kandabadicivirus presents a unique genome architecture and distant
26 phylogenetic relationship to the proposed *Badiciviridae* family. These findings highlight a high
27 mosaicism of genome organizations among the *Picornavirales*.

28 *Keywords:* bats, rectal swabs, *Picornavirales*, phylogeny, nucleotide composition analysis.

29 **Text**

30 The order *Picornavirales* is one of the most important viral orders in terms of virus diversity
31 and genome organization. Viruses belonging to this order consist of non-enveloped small viruses of

32 approximately 30 nm diameter with a pseudo T=3 symmetry and characterized by (i) a positive-sense
33 RNA genome with a covalently linked 5'-terminal protein (called VPg) and a 3' polyA tail, (ii) a
34 polyprotein gene expression strategy, (iii) a structural protein module containing three capsid
35 domains, and (iv) a non-structural replicase module containing the viral helicase, a chymotrypsin-like
36 protease and the RNA-dependent RNA polymerase (RdRP) domains (1). The genomic organization of
37 these modules is highly variable among the order, according to viral families and host spectrum (Figure
38 1). *Dicistroviridae* and *Marnaviridae* families, along with *Bacillarnavirus* and *Labyrnavirus* genera, with
39 a host spectrum ranging from arthropods to algae, present a bi-cistronic genome architecture with the
40 non-structural module (NS-module) in the 5' part, and the structural module (S-module) in the 3' part
41 of the genome, separated by an intergenic region (IGR). The same genome organization is observed in
42 plant-infecting *Secoviridae* viruses except that the two modules could be located in distinct genome
43 segments, depending on the genera. A third type of genome organization, observed in hosts ranging
44 from vertebrates to arthropods and plants, is on the contrary, a localization of the S-module in the 5'
45 part and the NS-module in the 3' part of the viral genome, following a mono-cistronic (*Picornaviridae*,
46 *Iflaviridae* and *Secoviridae*) or poly-cistronic (*Polycipiviridae*) translation strategy (Figure 1) (1). For
47 decades, *Picornaviridae* were considered as mono-cistronic viruses, but recently a second short ORF
48 was discovered within the genome of some of them (2).

49 The knowledge of picornaviruses host range, geographical distribution and genome
50 organization has recently exploded due to the use of high-throughput sequencing and the
51 identification of novel picorna-like viruses in stool samples from various species (3). For example, Yinda
52 *et al.* recently reported the identification of 11 novel picorna-like genomes in bat stool samples,
53 including highly divergent viruses with novel genome architecture: the mono-cistronic bat posalivirus
54 and fisalivirus; and the bi-cistronic bat felisavirus, dicibavirus, and badiciviruses 1 & 2 (4). In this study,
55 we report the identification and phylogenetic characterization of a new picornavirus-like virus in
56 frugivorous bats rectal swabs with distant homology to the previously reported bat badicivirus 1 that
57 presents an unusual genome architecture.

58 A total of 481 *Pteropus lylei* rectal swabs were collected during monthly captures between May
59 2015 and July 2016 in Kandal province, Cambodia. Bats were captured using mist nets. Handling and
60 sampling were conducted following the FAO guideline (5) under the supervision of agents of the
61 Forestry Administration of Cambodia, Ministry of Agriculture, Forestry and Fisheries. Individual swabs
62 were pooled, clarified and further ultracentrifuged at 100,000g for one hour. Total nucleic acids were
63 extracted from the resuspended pellet by the QIAamp cadon Pathogen mini kit (Qiagen) with the
64 substitution of carrier RNA by linear acrylamide (Life Tech). After extraction, DNA was digested with
65 20U Turbo DNase (Ambion) and RNA was purified with the RNeasy cleanup protocol (RNeasy mini kit,

66 Qiagen), analyzed on a Agilent BioAnalyzer and used as template for library preparation using the
67 SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian kit (Clontech). Library was sequenced in
68 pairs in a 2 x 75 bp format onto a NextSeq sequencer at DNAsision Company (Charleroi, Belgium). An
69 in-house bioinformatics pipeline comprising quality check and trimming (based on AlienTrimmer
70 package (6)), *de novo* assembly (using Megahit tool (7)), ORF prediction
71 (https://figshare.com/articles/translateReads_py/7588592) and sequence blasting against the
72 protein Reference Viral database (RVDB, [8]) followed by invalidation of the hits by blast against the
73 whole protein NCBI/nr database, was processed.

74 A large contig of 8 559 nt presented distant homology with the previously reported bat
75 badicivirus 1 (4). Phylogenetic analyses performed on the complete RdRP domain of proposed
76 *Badiciviridae* along with several representative members of *Picornavirales* clustered this contig within
77 the *Badiciviridae* (Figure 1), with the maximum protein identity observed with bat badicivirus 1
78 (54.64%). This genome, tentatively named Kandabadicivirus (accession no MK468720), present an
79 unusual genome architecture, with two predicted 5'-terminal ORFs coding for putative structural
80 proteins of 245 and 606 aa, respectively; and a large 3'-terminal ORF coding for the putative replicase
81 proteins of 1 645 aa. Among this ORF, the RdRP domain corresponds to 478 aa, which is in the range
82 of 450-490 aa observed for all known picornaviruses polymerase domains (8) (Figure 2). The positions
83 of the putative start and stop codons were confirmed either by RACE-PCR or by classical PCR followed
84 by Sanger sequencing, using specific primers flanking these regions and designed on Kandabadicivirus
85 sequence. We defined as possible initiation codons of ORF1 and ORF2 those generating the longest
86 ORFs, and verified this hypothesis by identifying amino-acid homologies of the N-terminal regions of
87 ORF1 and ORF2 with bat badicivirus 1 capsid protein. For example, ORF1 could either start at position
88 224 or at a downstream position (such as the position 437). Annotation of the region comprised
89 between nt 224 and 437 resulted in the identification of a domain (between nt 278 and 436) presenting
90 an amino-acid identity of 62% with bat badicivirus 1. Consequently, ORF1 initiation codon could either
91 be located at positions 224 or at position 248. Since no homologies were identified by Psi-Blast for the
92 region nt 224-248, we hypothesized that the initiation codon of ORF1 was the one generating the
93 longest ORF, and consequently applied a similar approach to identify the possible initiation codon for
94 ORF2, resulting in the identification of two putative ORFs (ORF1 and ORF2) that are overlapping over
95 14 nucleotides (Figure 2). Whether Kandabadicivirus genome follows a tri-cistronic or a bi-cistronic
96 expression strategy (with an obligatory frameshift between ORF1 and ORF2) is still unclear and need
97 further experiments.

98 Although Kandabadicivirus presents a unique genome organization, it presents also several
99 characteristics shared by the *Picornavirales* members: (i) the polyprotein expression strategy; (ii) the

100 three capsid protein domains within the S-module; and (iii) the RNA helicase and RNA-dependent RNA
101 polymerase domains within the NS-module (Figure 2). Two putative capsid proteins were predicted for
102 Kandabadicivirus genome: the first structural ORF contains one rhv-like capsid domain (located in the
103 C-terminal part of ORF1), while the second structural ORF contains two putative capsid domains (the
104 N-terminal part of ORF2 code for a rhv-like capsid domain and the C-terminal part of ORF2 code for a
105 capsid-like domain [pfam08762]). As for its closest relative (bat badicivirus 1), the 3C-like
106 chymotrypsin-like protease domain of Kandabadicivirus was not identified within the NS-module. As
107 described by Yinda *et al.* for bat badicivirus 1 (4), Kandabadicivirus presents several functional motifs
108 signatures of the replicase domain of picornaviruses: the GxxGxGKS helicase motif, and the KDE / KSG
109 / YGDD and FLKR polymerase motifs were retrieved while the D(YSDWD)D polymerase motif
110 characteristic of bat badicivirus 1 was not identified for Kandabadicivirus in which the serine was
111 replaced by a threonine (Figure 2). The serine residue in the active site observed in bat badicivirus 1 in
112 place of the 3C-like proteinase was not found in Kandabadicivirus.

113 Another *Picornavirales* genome characteristic is the presence of highly structured secondary
114 RNA structures at their 5' and 3' termini which constitute: i) the Internal Ribosomal Entry Site (IRES) in
115 the 5' part of the genome, which is necessary for ribosomal recognition; and ii) the 3' UnTranslated
116 Region (UTR) in the 3' part of the genome, which is used to initiate the RNA negative-strand synthesis.
117 Some *Picornavirales* viruses, such as Cricket Paralysis virus, could also present internal IRES (10). The
118 IRES are structurally and functionally classified into 5 types (from I to V) according to viral genera (11-
119 12). We sequenced by RACE-PCRs the 5' and 3' termini of Kandabadicivirus and *in silico* modeled their
120 RNA structure using the RNA Secondary Structure Prediction tool implemented through CLC Genomics
121 Workbench program. We further evaluated the presence of a 5' IRES using the IRESPred program (13).
122 While bat badicivirus 1 presents 5' and 3' termini of 332 and 234 nt respectively, Kandabadicivirus 5'
123 and 3' termini are shorter (223 nt and 197 nt long for the 5' and 3' UTR, respectively). The 5' UTR of
124 Kandabadicivirus was evaluated as a possible viral IRES. Interestingly, the intergenic region (IGR) of
125 Kandabadicivirus presents a RNA secondary structure highly structured that could also possibly
126 constitute (as for Cricket Paralysis virus) a second IRES, as suggested by IRESPred program (Figure 2).
127 Kandabadicivirus isolation is planned and, in case of success, will allow performing experiments
128 needed to confirm these IRES modeling and their functionality.

129 Surprisingly, Kandabadicivirus presents a large domain (*i.e.* "insertion" in Figure 2) of 249 aa
130 within the NS-module that is not present in bat badicivirus 1 (Figure 2). To confirm that this domain
131 was not an artifact during genome assembly, we performed PCR and Sanger sequencing using specific
132 primers flanking this region and designed on Kandabadicivirus sequence. The presence of this insertion
133 was confirmed on the initial pool of RNA. The origin of this domain is however unknown: neither

134 BlastN, Psi-BlastP nor CD-search analyses of this fragment of genome gave significant result. The
135 functional annotation and 3D reconstruction of this putative domain (using Swiss-Model program) (14)
136 did not reveal any putative function. The function of this domain and even its presence after the
137 maturation process of the polyprotein is therefore currently undetermined.

138 By analyzing the dinucleotide composition of *Picornavirales* viruses according to their host
139 spectrum, Yinda *et al.* inferred a plant origin of bat badicivirus 1 (4), possibly reflecting the diet of
140 Eidolon and Epomophorus bats, although *Badiciviridae*-related viruses were only previously reported
141 in Aphididae insects. To infer the host origin of Kandabadicivirus, we analyzed the dinucleotide
142 composition of its genome compared to other *Picornavirales* genomes clustered according to their
143 host spectrum. Briefly, all *Picornavirales* complete genomes whose host information was known were
144 retrieved from the Virus-Host Database (15) on the 26th of October 2018. Segmented *Secoviridae* were
145 concatenated and treated as single genome. The resulted database (*i.e.* 566 full genomes, Additional
146 Table 1) was used to constitute five groups of genomes: arthropods (N=70), birds (N=46), mammals
147 (N=365), mollusks (N=8), and plants (N=61). Sixteen genomes were not included in the analysis because
148 of a lack of a significant number of sequences to constitute a group (*i.e.* algae [N=2], amphibians [N=2],
149 diatoms [N=4], fish [N=5], reptiles [N=2], and fungi [N=1]). The rate defining the composition of
150 dinucleotides for a given genome was determined by counting the frequency of each dinucleotide
151 divided by the total count of dinucleotides of this genome. Each group was therefore characterized by
152 a matrix associating N genomes with their corresponding 16 possible dinucleotide rates. A discriminant
153 analysis was performed to predict Kandabadicivirus host group using R software (available at
154 [<https://doi.org/10.5281/zenodo.3547558>]). A posterior probability greater than 99% was obtained
155 for Kandabadicivirus belonging to the Arthropod group, confirming the host spectrum of previously
156 reported *Badiciviridae* (Figure 3). In addition, Kandabadicivirus was identified in rectal swabs NGS
157 dataset, and further confirmed by SYBR Green RT-qPCR specifically targeting Kandabadicivirus, and not
158 in the corresponding oral swabs or urines of *Pteropus lylei* (neither in the NGS datasets nor after the
159 qPCR), highlighting again a possible diet origin of this virus, for example via the consumption of fruits
160 containing insects, larvae or eggs, as suggested by Webala *et al.* (16).

161 The description of the viral (and genetic) diversity of picorna-like viruses found in bats gut
162 contents, and especially bats in close contact with humans such as *Pteropus lylei* in Cambodia, is
163 important and need further characterizations because new viruses, as Kandabadicivirus, participate to
164 the pool of viruses that may recombine and generate novel picorna-like variants with possible impact
165 on host range.

166 **Acknowledgments**

167 The authors want to thank the agents of the Forestry Administration of Cambodia, Ministry
168 of Agriculture, Forestry and Fisheries, for their supervision and help during captures and sampling of
169 bats; Pr. Francis Delpeyroux for his helpful knowledge on picornaviruses; and Vincent Guillemot for
170 his precious expertise on discriminant analysis.

171 **Funding information**

172 This work was supported by Laboratoire d'Excellence 'Integrative Biology of Emerging
173 Infectious Diseases' (grant no.ANR-10-LABX-62-IBEID), by the Direction Internationale de l'Institut
174 Pasteur, and undertaken in the framework of the ComAcross project with the financial support of the
175 European Union (EuropeAid, INNOVATE contract 315-047).

176 **References**

- 177 1. Zell R, Delwart E, Gorbalenya AE, Hovi T, King AMQ, Knowles NJ, Lindberg AM, Pallansch MA,
178 Palmenberg AC, Reuter G, Simmonds P, Skern T, Stanway G, Yamashita T, ICTV Report
179 Consortium. ICTV Virus Taxonomy Profile: *Picornaviridae*. J Gen Virol. 2017; 98:2421-2422.
- 180 2. Lulla V, Dinan AM, Hosmillo M, Chaudhry Y, Sherry L, Irigoyen N, Nayak KM, Stonehouse NJ,
181 Zilbauer M, Goodfellow I, Firth AE. An upstream protein-coding region in enteroviruses
182 modulates virus infection in gut epithelial cells. Nat Microbiol. 2019; 4:280-292.
- 183 3. Koonin EV, Wolf YI, Nagasaki K, Dolja VV. The Big Bang of picorna-like virus evolution antedates
184 the radiation of eukaryotic supergroups. Nat Rev Microbiol. 2008; 6:925-39.
- 185 4. Yinda CK, Zell R, Deboutte W, Zeller M, Conceição-Neto N, Heylen E, Maes P, Knowles NJ,
186 Ghogomu SM, Van Ranst M, Matthijssens J. Highly diverse population of *Picornaviridae* and
187 other members of the *Picornavirales*, in Cameroonian fruit bats. BMC Genomics. 2017; 18:249.
- 188 5. Food and Agriculture Organisation of the United Nations. Investigating the role of bats in
189 emerging zoonoses: Balancing ecology, conservation and public health interests. FAO Animal
190 Production and Health. Manual No. 12. Rome. 2011. Edited by S.H. Newman, H.E. Field, C.E.
191 de Jong and J.H. Epstein.
- 192 6. Criscuolo A, Brisse S. AlienTrimmer removes adapter oligonucleotides with high sensitivity in
193 short-insert paired-end reads. Commentary on Turner (2014) Assessment of insert sizes and
194 adapter content in FASTQ data from NexteraXT libraries. Front Genet. 2014; 5:130.
- 195 7. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large
196 and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;
197 31:1674-6.
- 198 8. Bigot T, Temmam S, Pérot P and Eloit M. RVDB-prot, a reference viral protein database and 1
199 its HMM profiles [version 1; peer review: awaiting peer review]. F1000Research 2019, 8:530.

- 200 9. Boros Á, Pankovics P, Simmonds P, Pollák E, Mátics R, Phan TG, Delwart E, Reuter G. Genome
201 analysis of a novel, highly divergent picornavirus from common kestrel (*Falco tinnunculus*): the
202 first non-enteroviral picornavirus with type-I-like IRES. *Infect Genet Evol.* 2015; 32:425-31.
- 203 10. Hodgman CE, Jewett MC. Characterizing IGR IRES-mediated translation initiation for use in
204 yeast cell-free protein synthesis. *N Biotechnol.* 2014; 31:499-505.
- 205 11. Palmenberg A, Neubauer D, Skern T. 2010. Chapter 1: Genome organization and encoded
206 proteins. In: Ehrenfeld E, Domingo E, Roos RP (Eds.). *The Picornaviruses*. ASM Press,
207 Washington, DC, pp. 3–17.
- 208 12. Sweeney TR, Dhote V, Yu Y, Hellen CU. A distinct class of internal ribosomal entry site in
209 members of the *Kobuvirus* and proposed *Salivirus* and *Paraturdivirus* genera of the
210 *Picornaviridae*. *J Virol.* 2012; 86:1468-86.
- 211 13. Kolekar P, Pataskar A, Kulkarni-Kale U, Pal J, Kulkarni A. IRESPred: Web Server for Prediction of
212 Cellular and Viral Internal Ribosome Entry Site (IRES). *Sci Rep.* 2016; 6:27436. doi:
213 10.1038/srep27436.
- 214 14. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Gallo Cassarino T,
215 Bertoni M, Bordoli L, Schwede T. SWISS-MODEL: modelling protein tertiary and quaternary
216 structure using evolutionary information. *Nucleic Acids Res.* 2014; 42:W252-8.
- 217 15. Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H, Hingamp P, Goto S,
218 Ogata H. Linking Virus Genomes with Host Taxonomy. *Viruses.* 2016; 8:66.
- 219 16. Webala PW, Musila S, Makau R. Roost Occupancy, Roost Site Selection and Diet of Straw-
220 Coloured Fruit Bats (Pteropodidae: *Eidolon helvum*) in Western Kenya: The Need for Continued
221 Public Education. *Acta Chiropterologica.* 2014; 16: 85-94.
- 222 17. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment,
223 interactive sequence choice and visualization. *Brief Bioinform.* 2017; bbx108.
- 224 18. Lefort V, Longueville JE, Gascuel O. SMS: Smart Model Selection in PhyML. *Mol Biol Evol.* 2017;
225 34:2422-2424.
- 226 19. Miller MA, Pfeiffer W, Schwartz T. "Creating the CIPRES Science Gateway for inference of large
227 phylogenetic trees". Proceedings of the Gateway Computing Environments Workshop (GCE),
228 14 Nov. 2010, New Orleans, LA pp 1 - 8.
- 229 20. Marchler-Bauer A, Bryant SH. CD-Search: protein domain annotations on the fly. *Nucleic Acids*
230 *Res.* 2004; 32: 327-31.

231

232 **Figure legends**

233 **Figure 1.** Phylogenetic reconstruction of Kandabadicivirus and other *Badiviridae* along with
234 representative members of the *Picornavirales* order. The schematic genome organization of
235 *Picornavirales* is presented according to their host spectrum (left panel: (1) vertebrates, (2) mammals,
236 (3) arthropods, (4) plants, and (5) algae) and their phylogenetic position among the order (right panel).
237 Complete amino-acid sequences of RdRP were aligned with MAFFT with the L-INS-I parameter (17).
238 The best amino-acids substitution models that fitted the data were determined with ATGC Start Model
239 Selection (18) implemented in <http://www.atgc-montpellier.fr/phymml-sms/> using the corrected Akaike
240 information criterion. Phylogenetic trees were constructed using Maximum Likelihood (ML) method
241 implemented through RAxML program under the CIPRES Science Gateway portal (19) according to the
242 selected substitution model. Nodal support was evaluated using 1000 bootstrap replicates. Only
243 supported nodes (i.e. with bootstrap values above 50) were represented.

244 **Figure 2.** Genome organization of Kandabadicivirus. Capsid, helicase and RdRP domains were predicted
245 by CD-search (20). The secondary RNA structure of the 5' and 3' UTR regions predicted by the RNA
246 Secondary Structure Prediction tool implemented through CLC Genomics Workbench program are
247 presented, with the predicted initiation codon of the first S-ORF and the stop codon of the NS-ORF
248 highlighted in bold. The genome coverage of Kandabadicivirus along the genome is also presented.

249 **Figure 3.** Discriminant analysis of dinucleotide composition rates clustered by host type. X and Y axes
250 represent the two first factors, with 95% confidence ellipses centered on the centroid of each group.





