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.

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

Sarah Temmam[1], Vibol Hul[2], Thomas Bigot[1,3], Maël Bessaud[4], Delphine Chrétien[1], Thavry Hoem[5], Christopher Gorman[2], Veasna Duong[2], Philippe Dussart[2], Julien Cappelle[5,6,7], Marc Eloit[1,8,*].

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

[2] Virology Unit, Institut Pasteur du Cambodge, Institut Pasteur International Network, Phnom Penh, Cambodia.

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

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

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

[6] UMR ASTRE, CIRAD, INRA, Université de Montpellier, Montpellier, France.

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

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

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

Abstract

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

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

Text

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

approximately 30 nm diameter with a pseudo T=3 symmetry and characterized by (i) a positive-sense RNA genome with a covalently linked 5′-terminal protein (called VPg) and a 3′ polyA tail, (ii) a polyprotein gene expression strategy, (iii) a structural protein module containing three capsid domains, and (iv) a non-structural replicase module containing the viral helicase, a chymotrypsin-like protease and the RNA-dependent RNA polymerase (RdRP) domains (1). The genomic organization of these modules is highly variable among the order, according to viral families and host spectrum (Figure 1). *Dicistroviridae* and *Marnaviridae* families, along with *Bacillarnavirus* and *Labyrnavirus* genera, with a host spectrum ranging from arthropods to algae, present a bi-cistronic genome architecture with the non-structural module (NS-module) in the 5′ part, and the structural module (S-module) in the 3′ part of the genome, separated by an intergenic region (IGR). The same genome organization is observed in plant-infecting *Secoviridae* viruses except that the two modules could be located in distinct genome segments, depending on the genera. A third type of genome organization, observed in hosts ranging from vertebrates to arthropods and plants, is on the contrary, a localization of the S-module in the 5′ part and the NS-module in the 3′ part of the viral genome, following a mono-cistronic (*Picornaviridae*, *Iflaviridae* and *Secoviridae*) or poly-cistronic (*Polycipiviridae*) translation strategy (Figure 1) (1). For decades, *Picornaviridae* were considered as mono-cistronic viruses, but recently a second short ORF was discovered within the genome of some of them (2).

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

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

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

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

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

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

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

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

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

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

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

Acknowledgments

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

Funding information

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

References

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

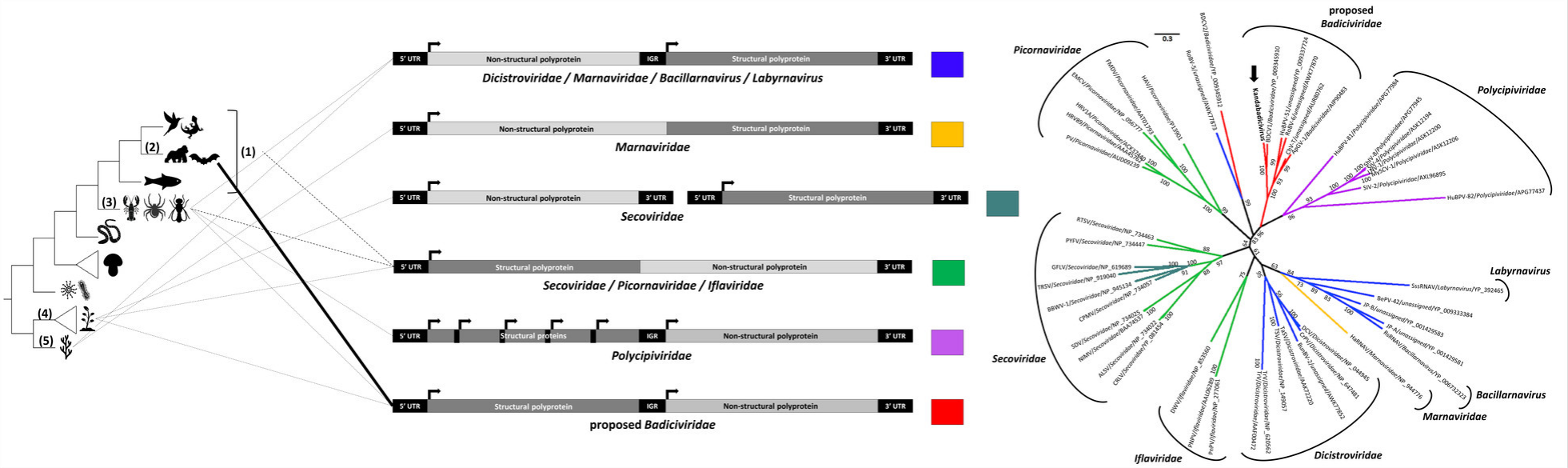
9. Boros Á, Pankovics P, Simmonds P, Pollák E, Mátics R, Phan TG, Delwart E, Reuter G. Genome analysis of a novel, highly divergent picornavirus from common kestrel (Falco tinnunculus): the first non-enteroviral picornavirus with type-I-like IRES. Infect Genet Evol. 2015; 32:425-31.
10. Hodgman CE, Jewett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

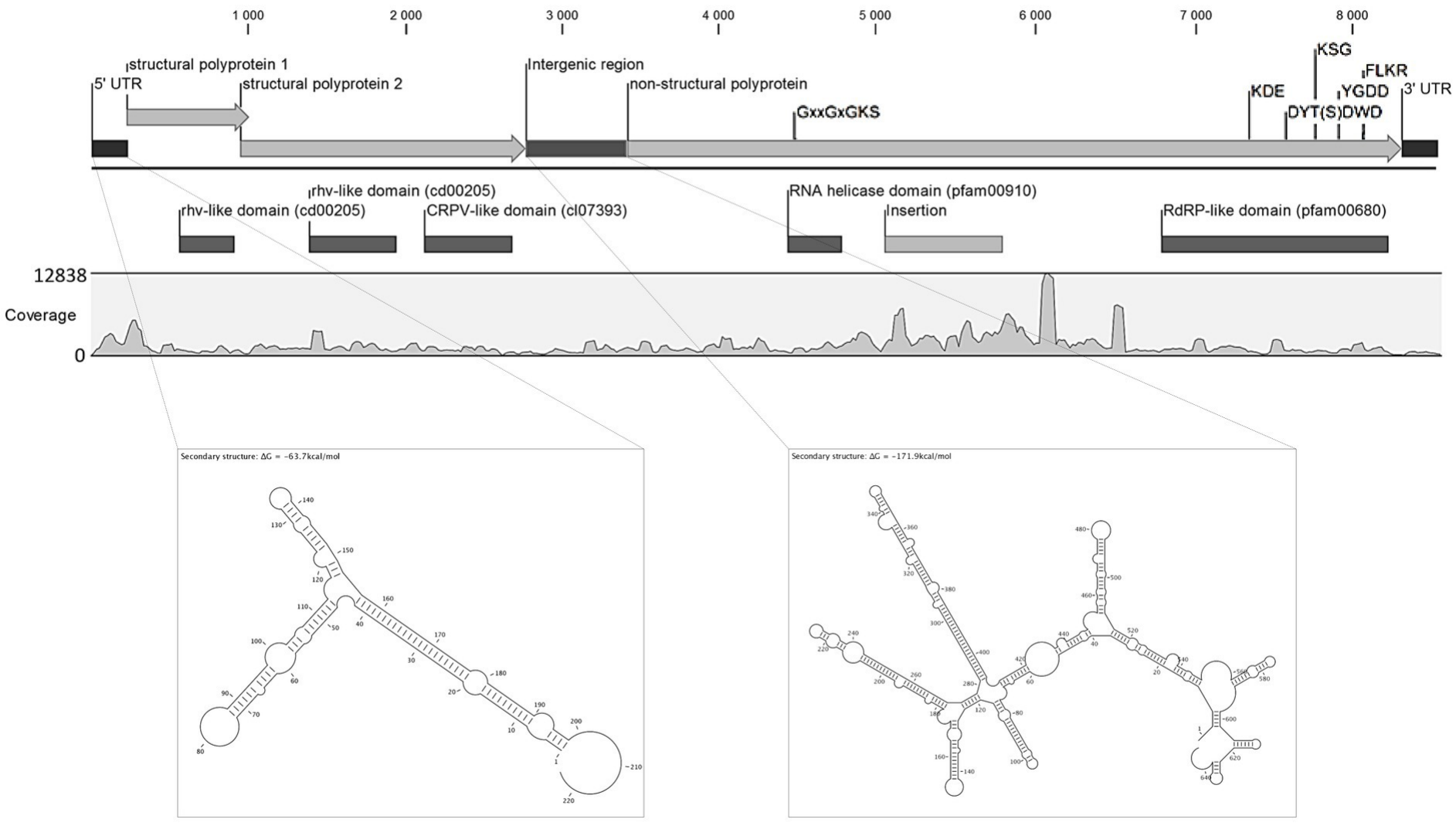
Figure legends

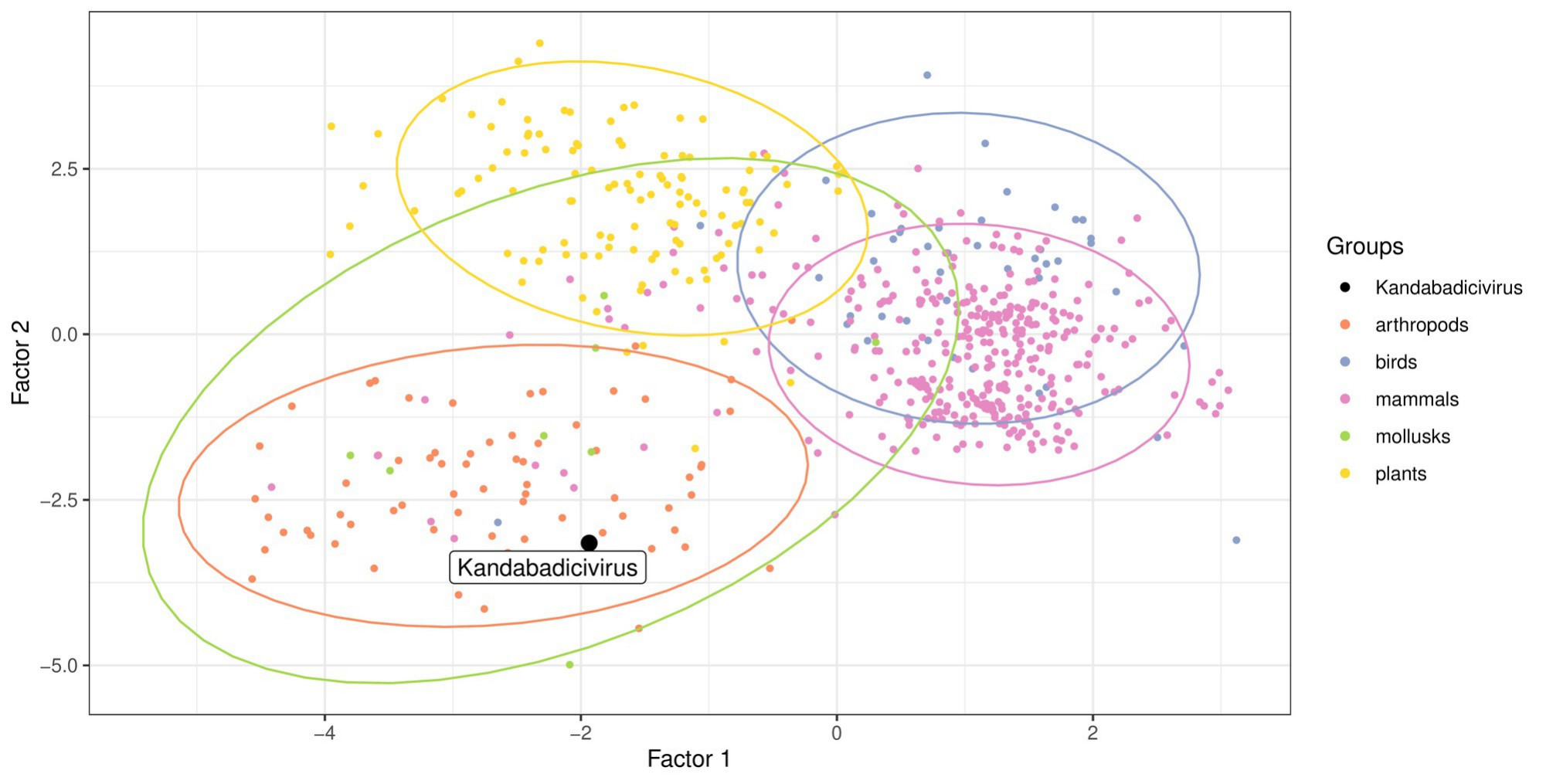
Figure 1. Phylogenetic reconstruction of Kandabadicivirus and other *Badiviridae* along with representative members of the *Picornavirales* order. The schematic genome organization of *Picornavirales* is presented according to their host spectrum (left panel: (1) vertebrates, (2) mammals, (3) arthropods,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

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

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


(1)
(2)
(3)
(4)
(5)
5' UTR
Non-structural polyprotein
IGR
Structural polyprotein
3' UTR
Dicistroviridae / Marnaviridae / Bacillarnavirus / Labyrnavirus
Marnaviridae
Secoviridae
Secoviridae / Picornaviridae / Iflaviridae
Structural proteins
Polycipiviridae
proposed Badiciviridae
0.3
proposed
Badiciviridae
Picornaviridae
Polycipiviridae
Labyrnavirus
Bacillarnavirus
Marnaviridae
Secoviridae
Iflaviridae
Dicistroviridae
Kanbabadiravirus


1 000
2 000
3 000
4 000
5 000
6 000
7 000
8 000
5' UTR
structural polyprotein 1
structural polyprotein 2
Intergenic region
non-structural polyprotein
GxxGxGKS
KDE
KSG
DYT(S)DWD
YGDD
FLKR
3' UTR
rhv-like domain (cd00205)
rhv-like domain (cd00205)
CRPV-like domain (cl07393)
RNA helicase domain (pfam00910)
Insertion
RdRP-like domain (pfam00680)
12838
Coverage
0
Secondary structure: ΔG = -63.7kcal/mol
Secondary structure: ΔG = -171.9kcal/mol


Factor 2
2.5
0.0
−2.5
−5.0
−4
−2
0
2
Factor 1
Kandabadicivirus
Groups
Kandabadicivirus
arthropods
birds
mammals
mollusks
plants