Sequence-based taxonomic framework for the classification of uncultured single-stranded DNA viruses of the family Genomoviridae

Arvind Varsani, Mart Krupovic

To cite this version:

Arvind Varsani, Mart Krupovic. Sequence-based taxonomic framework for the classification of uncultured single-stranded DNA viruses of the family Genomoviridae. Virus Evolution, 2017, 3 (1), pp.vew037. 10.1093/ve/vew037. pasteur-01977359

HAL Id: pasteur-01977359
https://pasteur.hal.science/pasteur-01977359
Submitted on 10 Jan 2019

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
Sequence-based taxonomic framework for the classification of uncultured single-stranded DNA viruses of the family Genomoviridae

Arvind Varsani,1,2,† and Mart Krupovic3,*‡

1The Biodesign Center for Fundamental and Applied Microbiomics, School of Life sciences, Center for Evolution and Medicine, Arizona State University, Tempe, AZ 85287, USA, 2Structural Biology Research Unit, Department of Clinical Laboratory Sciences, University of Cape Town, Observatory 7700, South Africa and 3Unité Biologie moléculaire du Gène chez les Extrêmophiles, Department of Microbiology, Institut Pasteur, 25 rue du Docteur Roux, Paris 75015, France

*Corresponding authors: E-mails: krupovic@pasteur.fr; arvind.varsani@asu.edu
†http://orcid.org/0000-0003-4111-2415
‡http://orcid.org/0000-0001-5486-0098

Abstract

With the advent of metagenomics approaches, a large diversity of known and unknown viruses has been identified in various types of environmental, plant, and animal samples. One such widespread virus group is the recently established family Genomoviridae which includes viruses with small (~2–2.4 kb), circular ssDNA genomes encoding rolling-circle replication initiation proteins (Rep) and unique capsid proteins. Here, we propose a sequence-based taxonomic framework for classification of 121 new virus genomes within this family. Genomoviruses display ~47% sequence diversity, which is very similar to that within the well-established and extensively studied family Geminiviridae (46% diversity). Based on our analysis, we establish a 78% genome-wide pairwise identity as a species demarcation threshold. Furthermore, using a Rep sequence phylogeny-based analysis coupled with the current knowledge on the classification of geminiviruses, we establish nine genera within the Genomoviridae family. These are Gemycircularvirus (n = 73), Gemyduguivirus (n = 1), Gemygorivirus (n = 9), Gemykibiivirus (n = 29), Gemykolovirus (n = 3), Gemykrogvirus (n = 3), Gemykroznavirus (n = 1), Gemytondvirus (n = 1), Gemyvongvirus (n = 1). The presented taxonomic framework offers rational classification of genomoviruses based on the sequence information alone and sets an example for future classification of other groups of uncultured viruses discovered using metagenomics approaches.

Key words: Genomoviridae; CRESS DNA viruses; replication-associated protein; ssDNA viruses.

1. Introduction

Viral metagenomics, fostered by powerful high-throughput sequencing methods, has recently revolutionized our perception of virus diversity in the environment. Many novel groups of uncultivated viruses have been discovered during the past decade, including viruses with small, moderately-sized, and even large genomes (Yau et al. 2011; Roux et al. 2012; Labonte and Suttle, 2013; Dutilh et al. 2014; Yutin et al. 2015; Zhou et al. 2015et al.; Dayaram et al. 2016; Steel et al. 2016). Many of these virus groups remain unclassified. To embrace the constantly growing output from viral metagenomics studies, virus taxonomy is increasingly switching from the traditional classification guided by biological features, such as serology, virion morphology or host range, to predominantly sequence-guided practices (Simmonds et al. 2017). Sequence-guided virus classification is relatively straightforward...
when the new viruses fall into existing taxa, with well-defined demarcation criteria. However, in the absence of isolated representatives and established taxonomic framework, rational definition of appropriate taxonomic ranks, such as families, genera, and species, for novel groups of uncultured viruses might be considerably more complex. Solutions to this problem are perhaps most urgently needed in the case of single-stranded (ss) DNA viruses, which are extremely widespread in nature. Due to their small genomes sizes, high mutation and recombination rates (Duffy and Holmes 2008; Duffy and Holmes 2009; Firth et al. 2009; Harkins et al. 2009, 2014; Grigoras et al. 2010; Martin et al. 2011; Streck et al. 2011; Nguyen et al. 2012; Cadar et al. 2013; Roux et al. 2013), and relative ease of genome amplification, an incredible diversity of these viruses has been discovered through metagenomics studies in all conceivable habitats. ssDNA viruses infect cells from all three domains of life and are currently classified by the International Committee on Taxonomy of Viruses (ICTV) into eleven families and one unassigned genus. Members of the families Microviridae and Inoviridae infect bacteria, viruses of the families Spiraviridae and Pleolipoviridae prey on archaea, whereas eukaryotes host viruses classified into the families Anelloviridae, Birnaviridae, Circoviridae, Geminiviridae, Genomoviridae, Nanoviridae, and Parvoviridae, and the unassigned genus Bacilladnavirus. In addition, several widespread groups of uncultured viruses discovered by viral metagenomics remain unclassified, predominantly those that are circular replication-associated protein encoding single-stranded (CRESS) DNA viruses (Simmonds et al. 2017).

The Genomoviridae family is one of the most recently established families of ssDNA viruses (Adams et al. 2016; Krupovic et al. 2016). The family currently includes a single genus Gymycircularus, which contains a single species, Sclerotinia gymycircularus 1, encompassing a single isolate, Sclerotinia sclerotiorum hypovirulence-associated DNA virus 1 (SsHADV-1). SsHADV-1 was isolated from a plant-pathogenic fungus Sclerotinia sclerotiorum and is the only ssDNA virus known to infect fungi (Yu et al. 2010, 2013). Recently, Liu et al (2016) have shown that SsHADV-1 is able to infect a mycophagous insect (Lycoriella ingenua) which acts as a transmission vector. SsHADV-1 virions are non-enveloped, isometric, 20–22 nm in diameter, and assembled from a single capsid protein (CP) (Yu et al. 2010). The genome is a circular ssDNA molecule of 2,166 nucleotides and contains two genes—for CP and rolling-circle replication initiation protein (Rep). Like in many other ssDNA viruses with circular genomes, the large intergenic region of SsHADV-1 contains a potential stem-loop structure with a nonanucleotide (TAATATTAT) motif at its apex, which is likely to be important for rolling-circle replication initiation. The CP of SsHADV-1 is not recognizable similar to the corresponding proteins from viruses in other taxa. Although SsHADV-1 remains the only isolated and classified member of the Genomoviridae, 121 viral genomes with varying degree of similarity to that of SsHADV-1 have been recovered and sequenced from various environmental, plant- and animal-associated samples, indicating that these viruses are widespread and abundant in the environment (Table 1). However, a proper taxonomic framework and demarcation criteria necessary to accommodate these viruses within the family Genomoviridae are lacking. Here, we explore the diversity and evolution of uncultured SsHADV-1-like viruses and attempt to establish a framework for their classification based on sequence data alone.

2. Genomoviridae diversity and species classification

At the time of the analysis (August, 2016), there were 121 SsHADV-1-like genome sequences in the GenBank database. Each of these genomes encodes two putative proteins homologous to the CP and Rep of SsHADV-1, highlighting strong coherence of this virus assemblage. Nevertheless, there is a considerable sequence divergence within the group (Supplementary Fig. S1). To investigate the extent of genomoviral sequence diversity, we analyzed the distribution of genome-wide pairwise identities (one minus Hamming distances of pairwise aligned sequences with pairwise deletion of gaps) across all 121 available genomes (Fig. 1A) using SDT v1.2 (Muhire, Varsani, and Martin 2014). Most of the virus genomes in our dataset share 56–66 genome-wide pairwise identities and only a handful contained nearly identical relatives (>98% identity), indicating that sequence diversity among SsHADV-1-like viruses remains largely unexplored.

Pairwise comparison of the Rep and CP protein sequences revealed a broader distribution of identity values (Fig. 1B and C). Notably, the CPs were considerably more divergent that the Reps, with the highest proportion of pairwise identities being ~35% (versus ~48% for the Rep). This observation is in line with functional differences of the two proteins and the fact that viral CPs often encompass host recognition determinants which are under constant pressure to co-evolve with the cellular receptors (Kolawole et al. 2014; Shangjin, Cortey, and Segales, 2009). Based on the analysis of distribution of the pairwise identities across genomes, CPs and Reps, we consider a threshold of 78% to be a conservative value for species demarcation. Thus, all viral genomes showing identities higher than this value should be considered as variants of the existing species. Nonetheless, there may be situations where it is difficult to assign species because a particular new sequence is

1. >78% similar to sequences from a particular species but <78% similar to other variants of that same species;
2. >78% similar to sequences from two or more different species.

To resolve the above conflicts, we suggest adopting a similar approach proposed for geminiviruses (Muhire et al. 2013; Varsani et al. 2014a, b; Brown et al. 2015). To resolve conflict 1, we suggest that the new sequence be classified within any species in which it shares >78% identity to any one variant formally classified as belonging to that species, even if it is <78% identical to other viruses within that species. To resolve conflict 2, we suggest that the new sequence be considered as belonging to the species with sequences with which it shares the highest degree of similarity.

3. Rep-based approach for creation of genera

Maximum likelihood phylogenetic analyses based on the Rep of 121 genomoviruses revealed several well-supported clades that could be considered as genera within the family (Fig. 2). We note that the clades obtained in the Rep-based phylogeny are not fully consistent with those obtained in the phylogenetic analysis of the full genome or the more diverse CP sequences (Figs 3 and 4). This is most explicit in the case of the newly proposed genus Gymyklovirus (see below). In the Rep-based tree corresponding sequences form a sister clade to the single representative of the genus Gymysgulavirus (Fig. 2). In contrast, in the whole-genome-based phylogeny, gymykloviruses form a sister group to members of the genus Gymycircularus (Fig. 3). The reason for this incongruence is likely to be intra-familial recombination between different genomovirus genomes resulting in chimeric entries encoding Rep and CP with different evolutionary histories (Kraberg et al. 2015a). Indeed, in the CP-based
### Table 1. Details of all members of the Genomoviridae.

| Genus                     | Species                                      | Accession # | Sequence ID | Isolation source | Common name                  | Sample type | Country    | Reference |
|---------------------------|----------------------------------------------|-------------|-------------|------------------|------------------------------|-------------|------------|-----------|
| Gemycircularvirus         | Blackbird associated gemycircularvirus 1     | KT862243    | 27 Fec79971 1 chicken | Gallus gallus domesticus  | Chicken          | Faeces      | New Zealand | Sikorski et al. (2013) |
| Gemycircularvirus         | Blackbird associated gemycircularvirus 1     | KT862244    | 29 Fec79971 1 Bama | Gallus gallus domesticus  | Blackbird        | Faeces      | New Zealand | Sikorski et al. (2013) |
| Gemycircularvirus         | Blackbird associated gemycircularvirus 1     | KT862246    | 30 Fec9971 1 horse  | Equus ferus caballus     | Horse            | Faeces      | New Zealand | Steel et al. (2016)    |
| Gemycircularvirus         | Blackbird associated gemycircularvirus 1     | KT862247    | 27 Fec1497 1 chicken | Gallus gallus domesticus  | Chicken          | Faeces      | New Zealand | Steel et al. (2016)    |
| Gemycircularvirus         | Dragonfly associated gemycircularvirus       | JX185429    | FL2-5X-2010 | Erythrops simplicollis  | Dragonfly        | Abdomen   | USA         | Rosario et al. (2012)  |
| Gemycircularvirus         | Equine associated gemycircularvirus          | KT862248    | 30 Fle6061 1 horse  | Equus ferus caballus     | Horse            | Faeces      | New Zealand | Steel et al. (2016)    |
| Gemycircularvirus         | Fur seal associated gemycircularvirus        | KT862249    | 27 Fle1 chicken | Gallus gallus domesticus  | Chicken          | Faeces      | New Zealand | Steel et al. (2016)    |
| Gemycircularvirus         | Gerygone associated gemycircularvirus 1      | KT861363    | P24a         | Gerygone a bofrontata    | Chatham Island warbler | Faeces      | New Zealand | Sikorski et al. (2013) |
| Gemycircularvirus         | Gerygone associated gemycircularvirus 2      | KT861364    | P24b         | Gerygone a bofrontata    | Chatham Island warbler | Faeces      | New Zealand | Sikorski et al. (2013) |
| Gemycircularvirus         | Hypericum associated gemycircularvirus       | KF413620    | VN6H-W6 | Hypericum japonicum    | Hypericum        | Leaf       | Vietnam     | Du et al. (2014)       |
| Gemycircularvirus         | Llama associated gemycircularvirus           | KT862245    | 29 Fle6018 1 Bama | Llama glama            | Llama             | Faeces      | New Zealand | Steel et al. (2016)    |
| Gemycircularvirus         | Llama associated gemycircularvirus           | KT862247    | 30 Fle6018 1 horse  | Equus ferus caballus     | Horse            | Faeces      | New Zealand | Steel et al. (2016)    |
| Gemycircularvirus         | Mallard associated gemycircularvirus         | KT861365    | as24         | Anas platyrynchos      | Mallard duck      | Faeces      | New Zealand | Sikorski et al. (2013) |
| Gemycircularvirus         | Miniopterus associated gemycircularvirus 1   | KJ641719    | BatMgCV-2/3/GCDO212 | Miniopterus fuliginosus  | Bat               | Pharyngeal & rectal swabs | China   | Wu et al. (2016)       |
| Mongoose associated gemycircularvirus | Mongoose associated gemycircularvirus 1 | KP935340    | 47B1         | Herpestes ichneumon     | Egyptian mongoose | Faeces    | Portugal    | Conceicao-Neto et al. (2015) |
| Mosquito associated gemycircularvirus | Mosquito associated gemycircularvirus 1 | HQ233506    | SDBV-G       | Culex erythrothorax     | Mosquito samples  | USA        | Ng et al. (2011)       |
| Odonata associated gemycircularvirus | Odonata associated gemycircularvirus 1 | KM598385    | OdiaGiV-1-150-360BC-12 | Ichinura psita      | Damsel fly        | Abdomen   | Dayaram et al. (2015) |
| Odonata associated gemycircularvirus | Odonata associated gemycircularvirus 2 | KM598386    | OdiaGiV-1-150-360R1-12 | Pantala hymenaea     | Dragonfly        | Abdomen   | Dayaram et al. (2015) |
| Odonata associated gemycircularvirus | Odonata associated gemycircularvirus 2 | KM598387    | OdiaGiV-2-1642KW-12 | Aeschna multicolor     | Dragonfly        | Abdomen   | Dayaram et al. (2015) |
| Odonata associated gemycircularvirus | Odonata associated gemycircularvirus 2 | KM598388    | OdiaGiV-2-1642LM2-12 | Libellula sativata     | Dragonfly        | Abdomen   | Dayaram et al. (2015) |
| Poaceae associated gemycircularvirus | Poaceae associated gemycircularvirus 1 | KT125357    | PaGmV-1 / TO STO14-25204 14 | Brachiaaria defixa   | Leaff             | Tonga    | Male et al. (2015)    |
| Poaceae associated gemycircularvirus | Poaceae associated gemycircularvirus 1 | KT125358    | PaGmV-1 / TO STO15-25204 14 | Brachiaaria defixa   | Leaff             | Tonga    | Male et al. (2015)    |
| Poaceae associated gemycircularvirus | Poaceae associated gemycircularvirus 1 | KT125359    | PaGmV-1 / TO STO18-25204 14 | Brachiaaria defixa   | Leaff             | Tonga    | Male et al. (2015)    |
| Poaceae associated gemycircularvirus | Poaceae associated gemycircularvirus 1 | KT125360    | PaGmV-1 / TO STO19-25204 14 | Brachiaaria defixa   | Leaff             | Tonga    | Male et al. (2015)    |
| Porcine associated gemycircularvirus | Porcine associated gemycircularvirus 1 | KT862250    | 49 Fle6006 1 pig | Sus scrofa domestica    | Pig               | Faeces      | New Zealand | Steel et al. (2016)    |
| Porcine associated gemycircularvirus | Porcine associated gemycircularvirus 2 | KF371460    | as5          | Sus scrofa             | Domestic pig    | Faeces      | New Zealand | Sikorski et al. (2013) |
| Peropus associated gemycircularvirus | Peropus associated gemycircularvirus 1 | KT732204    | Tbat45285 | Peropus tonguana     | Peropus tonguana | Bat        | Tonga    | Male et al. (2016)    |
| Peropus associated gemycircularvirus | Peropus associated gemycircularvirus 1 | KT732205    | Tbat4764        | Peropus tonguana     | Peropus tonguana | Bat        | Tonga    | Male et al. (2016)    |
| Peropus associated gemycircularvirus | Peropus associated gemycircularvirus 1 | KT732207    | Tbat113791      | Peropus tonguana     | Peropus tonguana | Bat        | Tonga    | Male et al. (2016)    |
| Peropus associated gemycircularvirus | Peropus associated gemycircularvirus 1 | KT732208    | Tbat130791      | Peropus tonguana     | Peropus tonguana | Bat        | Tonga    | Male et al. (2016)    |
| Peropus associated gemycircularvirus | Peropus associated gemycircularvirus 1 | KT732209    | Tbat130852      | Peropus tonguana     | Peropus tonguana | Bat        | Tonga    | Male et al. (2016)    |
| Peropus associated gemycircularvirus | Peropus associated gemycircularvirus 1 | KT732210    | Tbat130860      | Peropus tonguana     | Peropus tonguana | Bat        | Tonga    | Male et al. (2016)    |
| Peropus associated gemycircularvirus | Peropus associated gemycircularvirus 1 | KT732211    | Tbat12377       | Peropus tonguana     | Peropus tonguana | Bat        | Tonga    | Male et al. (2016)    |
| Peropus associated gemycircularvirus | Peropus associated gemycircularvirus 1 | KT732201    | Tbat12377       | Peropus tonguana     | Peropus tonguana | Bat        | Tonga    | Male et al. (2016)    |
| Peropus associated gemycircularvirus | Peropus associated gemycircularvirus 1 | KT732207    | Tbat130791      | Peropus tonguana     | Peropus tonguana | Bat        | Tonga    | Male et al. (2016)    |
| Peropus associated gemycircularvirus | Peropus associated gemycircularvirus 1 | KT732208    | Tbat130851      | Peropus tonguana     | Peropus tonguana | Bat        | Tonga    | Male et al. (2016)    |
Table 1. Continued

| Accession # | Sequence ID | Common name | Sample type | Country | Reference |
|-------------|-------------|-------------|-------------|---------|-----------|
| KT732790    | Tbat H 103746 | Pteropus associated gemycircularvirus 7 | Bat Faeces Tonga | Male et al. (2016) |
| KT732791    | Tbat H 103909 | Pteropus associated gemycircularvirus 7 | Bat Faeces Tonga | Male et al. (2016) |
| KT732792    | Tbat L 103746 | Pteropus associated gemycircularvirus 7 | Bat Faeces Tonga | Male et al. (2016) |
| KT732793    | Tbat L 103909 | Pteropus associated gemycircularvirus 7 | Bat Faeces Tonga | Male et al. (2016) |
| KT732794    | Tbat 31579    | Pteropus associated gemycircularvirus 10 | Bat Faeces Tonga | Male et al. (2016) |
| KT732795    | Tbat 21383    | Pteropus associated gemycircularvirus 10 | Bat Faeces Tonga | Male et al. (2016) |
| KT732796    | Tbat H 103958 | Pteropus associated gemycircularvirus 10 | Bat Faeces Tonga | Male et al. (2016) |
| KR912221    | Ch-zjrat-01  | Rat associated gemycircularvirus 1 | Rat Blood China | Li et al. (2015) |
| GQ365709    | SsHADV-1 CN  | Sclerotinia gemycircularvirus 1 | Sclerotinia Mycelial samples China | Yu et al. (2010) |
| KF268025    | SsHADV-1 NZ H6 2012 | Sclerotinia gemycircularvirus 1 | River Sediments New Zealand | Kraberger et al. (2013) |
| KF268026    | SsHADV-1 NZ SR1 2012 | Sclerotinia gemycircularvirus 1 | River Sediments New Zealand | Kraberger et al. (2013) |
| KF268027    | SsHADV-1 NZ SR3 2012 | Sclerotinia gemycircularvirus 1 | River Sediments New Zealand | Kraberger et al. (2013) |
| KM598383    | SsHADV-1-US-549DFS-12 | Sclerotinia gemycircularvirus 1 | Erythemis simplicicollis Dragonfly Abdomen USA | Dayaram et al. (2015) |
| KM598384    | SsHADV-1-US-549SR-12 | Sclerotinia gemycircularvirus 1 | Pantala hymenaea Dragonfly Abdomen USA | Dayaram et al. (2015) |
| KM821747    | SaGmV-1 NZ-BS3970-2012 | Sewage derived gemycircularvirus 1 | Sewage New Zealand | Kraberger et al. (2015a) |
| KJ547636    | BS4014 | Sewage derived gemycircularvirus 3 | Sewage New Zealand | Kraberger et al. (2015a) |
| KJ547640    | BS3972 | Sewage derived gemycircularvirus 4 | Sewage oxidation pond New Zealand | Kraberger et al. (2015a) |
| KJ547639    | BS3970 | Sewage derived gemycircularvirus 5 | Sewage oxidation pond New Zealand | Kraberger et al. (2015a) |
| KT862238    | 4 Fec7 duck | Sheep associated gemycircularvirus 1 | Sheep Faeces New Zealand | Steel et al. (2016) |
| KT862239    | 24 Fec7 duck | Sheep associated gemycircularvirus 1 | Sheep Faeces New Zealand | Steel et al. (2016) |
| KT732790    | Tbat A 103952 | Pteropus associated gemycircularvirus 1 | Bat Faeces Tonga | Male et al. (2016) |
| KJ413144    | 349 | Sewage derived gemycircularvirus 1 | Sewage oxidation pond New Zealand | Kraberger et al. (2015a) |
| KF371632    | P14 | Blackbird associated gemycircularvirus 1 | European starling Faeces New Zealand | Sikorski et al. (2013) |
| KP263543    | 588t | Badger associated gemycircularvirus 1 | European badger Faeces Portugal | Conceicao-Neto et al. (2015) |
| KJ547644    | BS3980 | Human associated gemycircularvirus 1 | Sewage oxidation pond New Zealand | Kraberger et al. (2015a) |
| KJ547645    | BS3849 | Human associated gemycircularvirus 1 | Sewage oxidation pond New Zealand | Kraberger et al. (2015a) |
| KP974694    | DB2 | Human associated gemycircularvirus 1 | Sewage oxidation pond New Zealand | Kraberger et al. (2015a) |
| KJ547646    | BS3963 | Human associated gemycircularvirus 2 | Sewage oxidation pond New Zealand | Kraberger et al. (2015a) |
| KJ547647    | BS3952 | Human associated gemycircularvirus 2 | Sewage oxidation pond New Zealand | Kraberger et al. (2015a) |
| KJ547648    | BS3951 | Human associated gemycircularvirus 2 | Sewage oxidation pond New Zealand | Kraberger et al. (2015a) |
| KJ547649    | BS3950 | Human associated gemycircularvirus 2 | Sewage oxidation pond New Zealand | Kraberger et al. (2015a) |
| KJ547650    | BS3949 | Human associated gemycircularvirus 2 | Sewage oxidation pond New Zealand | Kraberger et al. (2015a) |
| KJ547651    | BS3948 | Human associated gemycircularvirus 2 | Sewage oxidation pond New Zealand | Kraberger et al. (2015a) |
| Genus                        | Species                        | Accession # | Sequence ID | Isolation source | Common name       | Sample type     | Country  | Reference                        |
|-----------------------------|--------------------------------|-------------|-------------|------------------|-------------------|----------------|----------|----------------------------------|
| Human associated gemykibivirus 2 | Homo sapiens                  | KP13078     | BZ1         | Homo sapiens     | Human             | Faeces         | Brazil   | Phan et al. (2015)               |
| Human associated gemykibivirus 2 | Homo sapiens                  | KP13079     | BZ2         | Homo sapiens     | Human             | Faeces         | Brazil   | Phan et al. (2015)               |
| Human associated gemykibivirus 2 | Homo sapiens                  | KP13080     | NP          | Untreated sewage | –                 | Faeces         | Nepal    | Phan et al. (2015)               |
| Human associated gemykibivirus 3 | Homo sapiens                  | KP987887    | Gemyc1c     | Homo sapiens     | Human             | Faeces         | Portugal | Conceicao-Neto et al. (2015)     |
| Human associated gemykibivirus 4 | Homo sapiens                  | KT363839    | GeTz1       | Homo sapiens     | Human             | Faeces         | China    | Zhou et al. (2015)               |
| Human associated gemykibivirus 5 | Homo sapiens                  | KU134313    | HV-Gov2     | Homo sapiens     | Human             | Faeces         | France   | Harary et al. (2016)             |
| Mongoose associated gemykibivirus 1 | Herpestes ichneumon          | KP063545    | 160d        | Herpestes ichneumon | Egyptian mongoose | Faeces         | Portugal | Conceicao-Neto et al. (2015)     |
| Pteropus associated gemykibivirus 1 | Tbat A 64418                 | KT32213     | Tbat A 64418 | Pteropus tonganus | Bat               | Faeces         | Portugal | Wu et al. (2016)                 |
| Rhinolophus associated gemykibivirus 1 | Rhinolophus hypposideros     | KJ614726    | BtRh-CV-6/Tbet2013 | Herpestes ichneumon | Egyptian mongoose | Faeces         | Portugal | Conceicao-Neto et al. (2015)     |
| Rhinolophus associated gemykibivirus 2 | Rhinolophus ferrumequinum    | KJ614726    | BtRh-CV-8/NM2013 | Herpestes ichneumon | Egyptian mongoose | Faeces         | China    | Wu et al. (2016)                 |
| Sewage derived gemykibivirus 1 | Sewage oxidation pond        | KJ547643    | BS4149      | –                | Sewage oxidation pond | –                | New Zealand | Knaberger et al. (2015(a)   |
| Sewage derived gemykibivirus 1 | Sewage oxidation pond        | KT862240    | 27 BS14149 chicken | Gallus gallus domesticus | Chicken | Faeces         | New Zealand | Steel et al. (2016)             |
| Sewage derived gemykibivirus 1 | Sewage oxidation pond        | KT862252    | 52 BS14149 cow | Bos taurus       | Cow               | Faeces         | New Zealand | Steel et al. (2016)             |
| Sewage derived gemykibivirus 1 | Sewage oxidation pond        | KT862255    | 56 BS14149 hare | Lepus europaeus  | Hare              | Faeces         | New Zealand | Steel et al. (2016)             |
| Sewage derived gemykibivirus 2 | Sewage oxidation pond        | KJ547642    | BS3911      | –                | Sewage oxidation pond | –                | New Zealand | Knaberger et al. (2015(a)   |
| Pteropus associated gemykolovirus 1 | Tbat A 103779                | KT32798     | Tbat A 103779 | Pteropus tonganus | Bat               | Faeces         | Tonga    | Male et al. (2016)              |
| Pteropus associated gemykolovirus 1 | Tbat A 103779                | KT32799     | Tbat A 103779 | Pteropus tonganus | Bat               | Faeces         | Tonga    | Male et al. (2016)              |
| Pteropus associated gemykolovirus 2 | Tbat A 103891                | KT32800     | Tbat A 103891 | Pteropus tonganus | Bat               | Faeces         | Tonga    | Male et al. (2016)              |
| Bovine associated gemykrogvirus 1 | Gallus gallus domesticus     | LK931484    | HCB982122   | Bos taurus       | Cow               | Serum          | Germany  | Lambert et al. (2014)           |
| Caribou associated gemykrogvirus 1 | FaGmc-CV-13                  | KJ941717    | Rangey tarandus | Caribou          | Faeces         | Canada        | Ng et al. (2014)                |
| Sewage derived gemykrogvirus 1 | Seawage oxidation pond       | KJ547634    | BS3913      | –                | Sewage oxidation pond | –                | New Zealand | Knaberger et al. (2015(a)   |
| Rabbit associated gemykroznavirus 1 | Oryctolagus cuniculus        | KF371631    | as35        | Oryctolagus cuniculus | Rabbit | Faeces         | New Zealand | Sikorski et al. (2015)          |
| Ostrich associated gemykroznavirus 1 | Struthio camelus            | KF371630    | as3         | Struthio camelus | Ostrich         | Faeces         | New Zealand | Sikorski et al. (2015)          |
| Human associated gemyvongvirus 1 | Homo sapiens                  | KP974693    | DB1         | Homo sapiens     | Human             | Plasma         | Germany  | Zhang et al. (2016)             |
Tree gemykloviruses are firmly nested within the large clade including the majority of gemycircularviruses (Fig. 4). Given that CP sequences of genomoviruses are considerably more divergent than the Rep sequences (Fig. 1), it appears reasonable to establish a higher (i.e., above the species level) taxonomic framework using the Rep (Fig. 2). The latter protein is also conserved in other eukaryotic ssDNA viruses (which is not the case for the CP) and can thus be used to assess the place of genomoviruses within the larger community of ssDNA viruses.

To evaluate the taxonomic structure of the Genomoviridae, we took advantage of the fact that in Rep-based phylogenetic analyses, genomoviruses consistently form a sister group to
members of the Geminiviridae (Krupovic et al. 2016), a comprehensively characterized family of plant viruses with circular ssDNA genomes (Varsani et al. 2014b). Thus, using the established taxonomic framework of the Geminiviridae overlaid on the Rep-based phylogeny as a guide, we could define five clades and four additional singletons within the Genomoviridae branch (Fig. 2). The defined groups displayed equivalent intra-family divergence as the established genera within the family Geminiviridae (Varsani et al. 2014b). The nine groups were supported in both nucleotide and protein sequence inferred phylogenies (Supplementary Fig. S2). Consequently, in addition to the existing genus Gemycircularvirus, we propose establishing eight new genera within the family Genomoviridae. The details of the nine genera are summarized in Fig. 5 and briefly outlined below.

### 3.1 Gemycircularvirus

This genus has the largest number of new species (n = 43; seventy-three genomes; Table 1) and includes SsHADV-1, the founding member of the family. Members of the genus display 44% diversity. Viruses within the forty-three species cluster with 99 and 96% branch support values in phylogenetic trees constructed from either Rep or full genome sequences, respectively (Figs 2 and 3).

### 3.2 Gemykibivirus

This is the second most populated genus (n = 16; twenty-nine genomes; Table 1) in the family with 43% diversity among its members. The name of the genus is an acronym of words geminivirus-like and myco-like kibi virus (kibi means circular in Amharic). Sequences within the fifteen species cluster with 93% branch support within phylogenetic trees constructed from Rep (Fig. 2) and two well-supported clades (100 and 96%) within trees constructed from full genome sequences (Fig 3), suggesting that recombination has played an important role in the evolution of this group.

### 3.3 Gemygorvirus

Members of this genus (n = 5; nine genomes; Table 1) display 49% diversity. The name of the genus is an acronym of words geminivirus-like and myco-like gor virus (gor means round in Hindi). Sequences within the five species cluster with 100 and 99% branch support within phylogenetic trees constructed from either Rep or full genome sequences, respectively (Figs 2 and 3).

### 3.4 Gemykolovirus

Members of this genus (n = 2; three genomes; Table 1) display 37% diversity. The name of the genus is an acronym of words geminivirus-like and myco-like kolo virus (kolo means round in Czech). Sequences within the two species cluster with 100 and 89% branch support within phylogenetic trees constructed from either Rep or full genome sequences, respectively (Figs 2 and 3).

### 3.5 Gemykrogvirus

Members of this genus (n = 3; three genomes; Table 1) display 33% diversity. The name of the genus is an acronym of words geminivirus-like and myco-like krog virus (krog means round in Slovenian). Sequences within the three species cluster with 99 and 100% branch support within phylogenetic trees constructed from either Rep or full genome sequences respectively (Figs 2 and 3).
Despite there being evidence of recombination within the genomes. Branches with 75% SH-like branch support have been collapsed.

Figure 3. Maximum likelihood phylogenetic tree of the genomes of viruses in the Genomoviridae family. The tree was inferred using FastTree (Price, Dehal, and Arkin 2010) (GTR = CAT). The numbers at the branches indicate SH-like support values. The topology of tree supports the proposed genera demarcation at the genome level, despite there being evidence of recombination within the genomes. Branches with <.75% SH-like branch support have been collapsed.
3.5 Gemyvongvirus

The name of the genus is an acronym of words geminivirus-like and myco-like vong virus (vong means circular in Lao). The single species Human associated gemyvongvirus 1 (Table 1) within the genus shares between 56 and 62% genome-wide sequence similarity with viruses in other genera and is a divergent taxon in the phylogenetic trees constructed from either Rep or full genome sequences (Figs 2 and 3).

3.6 Gemytondavirus

The name of the genus is an acronym of words geminivirus-like and myco-like tond virus (tond means round in Maltese). The single species Ostrich associated gemytondavirus 1 (Table 1) within the genus shares between 53 and 61% genome-wide sequence similarity with viruses in other genera and is a divergent taxon in the phylogenetic trees constructed from either Rep or full genome sequences (Figs 2 and 3).

3.7 Gemykroznavirus

The name of the genus is an acronym of words geminivirus-like and myco-like krozna virus (krozna means circular in Slovenian). The single species Rabbit associated gemykroznavirus 1 (Table 1) within the genus shares between 55 and 60% genome-wide sequence similarity with viruses in other genera and is a divergent taxon in the phylogenetic trees constructed from either Rep or full genome sequences (Figs 2 and 3).
within the genus shares between 56 and 61% genome-wide sequence similarity with other sequences in other genera and is a divergent taxon in the phylogenetic trees constructed from either Rep or full genome sequences (Figs 2 and 3).

3.8 Gemyduguivirus

The name of the genus is an acronym of words geminivirus-like and myco-like dugui virus (dugui means circular in Mongolian). The single species Dragonfly associated gemyduguivirus 1 (Table 1) within the genus shares between 57 and 62% genome-wide sequence similarity with viruses in other genera and is a divergent taxon in the phylogenetic trees constructed from either Rep or full genome sequences (Figs 2 and 3).

4. Conserved sequence motifs in the Genomoviridae

CRESS DNA viruses replicate through the rolling circle replication (RCR) mechanism which is similar to that used by bacterial plasmids (Khan 1997; Chandler et al. 2013; Ruiz-Maso et al. 2015). RCR is initiated by the Rep, encoded by CRESS DNA viruses, cleaving the dsDNA between positions 7 and 8 of a nonanucleotide located at a putative stem-loop structure at the origin of replication (Heyraud-Nitschke et al. 1995; Laufs et al. 1995b; Timchenko et al. 1999; Rosario, Duffy, and Breitbart, 2012). In the case of genomoviruses, this nonanucleotide is variable (‘TAWWDWRN’) with ‘TAATWYAT’ being the consensus nonanucleotide for gemycircularviruses, whereas gemykibviruses display the greatest variation in this motif—‘WATAWWHAN’ (Fig. 6; Supplementary Data S1). In contrast, we note that within the Geminiviridae family, including all recently described geminiviruses (Varsani et al. 2009; Briddon et al. 2010; Krenz et al. 2012; Loconsole et al. 2012; Bernardo et al. 2013; Heydarnejad et al. 2013; Ma et al. 2015; Bernardo et al. 2016), the consensus nonanucleotide motif is ‘TRAKATTRC’.

The N terminus of the Rep contains motifs that are important for initiating RCR and it is not surprising that some of these motifs are well conserved across many ssDNA viruses, phages, and plasmids that replicate using the RCR mechanism (Ilyina and Koonin, 1992; Vega-Rocha et al. 2007a; Rosario, Duffy, and Breitbart, 2012; Krupovic, 2013). The presence of a single catalytic tyrosine residue in the RCR motif III classifies genomovirus, geminivirus, bacilladnavirus, circovirus and nanovirus Reps as members of superfamily II (Ilyina and Koonin, 1992; Krupovic, 2013).

In genomoviruses, the conserved sequence of the RCR motif I, which is thought to be involved in the recognition of iterative sequences associated with the origin of replication, is predominantly ‘uuTYxQ’ (u denotes hydrophobic residues and x any residue) (Fig. 6; Supplementary Data S1), with the exception of the Reps of currently known gemykolorviruses and gemykrogviruses. The genomovirus RCR motif II, ‘xHxHx’
motifs conserved within the genomoviral genomes further supports the validity of the proposed genera. We also define a species demarcation criterion of 78% genome-wide identity, that is sequences that share >78% pairwise identity with other genomovirus sequences belong to the same species and those that share <78% can be considered as new species. It is worth noting that despite the fact that genomoviruses have been studied for over two decades, the sequence diversity of all known genomoviruses is similar to that of the recently discovered genomoviruses (46 vs 47%, respectively). This observation strongly suggests that the extent of sequence diversity within this expansive virus group remains largely unexplored.

Although the guidelines presented here are tailored for the classification of viral genomes in the family Genomoviridae, a similar sequence-based framework can be easily adapted for other virus clusters identified though metagenomics studies and lacking a pre-existing taxonomic framework, in particular for novel CRiSS DNA viruses. We do acknowledge that this approach deviates from a previous norm that used a set of criteria including biological properties such as host range, pathology, vectors, etc. coupled with sequence data. However, given that the rate at which genome sequences of uncultivated viruses are being identified from various sources, we need to establish more robust classification approaches that can easily be implemented on the bases of sequence data alone. Indeed, this necessity is acknowledged by the ICTV which encourages submissions of taxonomic proposals for classification of viruses that are known exclusively from their genome sequences (Simmonds et al. 2017). This new tide in virus taxonomy is expected to catalyze the comprehension of the diversity, ecology and evolution of the global virome.

Supplementary data

Supplementary data are available at Virus Evolution online.

Disclaimer

This article is based on the taxonomic proposal 2016.001a-agF.U.v5.Genomoviridae which has been considered and approved by the Executive Committee (EC) of the ICTV. AV and MK are elected members of the ICTV EC.

Conflict of interest: None declared.

References

Adams, M. J., et al. (2016) ‘Ratification Vote on Taxonomic Proposals to the International Committee on Taxonomy of Viruses’, Archives of Virology, 161: 2921–49.

Bernardo, P., et al. (2013) ‘Identification and Characterisation of a Highly Divergent Geminivirus: Evolutionary and Taxonomic Implications’, Virus Research, 177: 35–45.

Bridgon, R. W., et al. (2010) ‘Turnip Curly Top Virus, a Highly Divergent Geminivirus Infecting Turnip in Iran’, Virus Research, 152: 169–75.

Brown, J. K., et al. (2015) ‘Revision of Begomovirus Taxonomy Based on Pairwise Sequence Comparisons’, Archives of Virology, 160: 1593–619.

5. Concluding remarks

The Reps of genomoviruses are most closely related to those of geminiviruses and hence here we used a geminivirus taxonomy-informed approach to classify 121 genomoviruses into Rep sequence-based genera. Within the Genomoviridae family we establish eight new genera in addition to the one created previously (Krupovic et al. 2016). Detailed analysis of sequence
Wu, Z., et al. (2016) ‘Deciphering the Bat Virome Catalog to Better Understand the Ecological Diversity of Bat Viruses and the Bat Origin of Emerging Infectious Diseases’, ISME Journal, 10: 609–20.

Yau, S., et al. (2011) ‘Virophage Control of Antarctic Algal Host-Virus Dynamics’, Proceedings of the National Academy of Sciences of the United States of America, 108: 6163–8.

Yu, X., et al. (2010) ‘A Geminivirus-Related DNA Mycovirus that Confers Hypovirulence to a Plant Pathogenic Fungus’, Proceedings of the National Academy of Sciences of the United States of America, 107: 8387–92.

Yutin, N., et al. (2015) ‘A Novel Group of Diverse Polinton-Like Viruses Discovered by Metagenome Analysis’, BMC Biology, 13: 95.

Zhang, W., et al. (2016) ‘Viral Nucleic Acids in Human Plasma Pools’, Transfusion, 56: 2248–55.

Zhou, C., et al. (2015) ‘A Novel Gemycircularvirus in an Unexplained Case of Child Encephalitis’, Virology Journal, 12: 197.