Truly ubiquitous CRESS DNA viruses scattered across the eukaryotic tree of life

Lele Zhao, Erik Lavington, Siobain Duffy*

Department of Ecology, Evolution and Natural Resources, School of Environmental and Biological Sciences, Rutgers, the State University of New Jersey.

Running title: endogenized CRESS DNA virus Reps

*Corresponding Author:
duffy@sebs.rutgers.edu
phone +1 (848) 932-6299
fax +1 (732) 932-8578

Acknowledgements: We thank two anonymous reviewers for their thoughtful comments and the editors of Journal of Evolutionary Biology for their patience during the COVID-19 pandemic. This work was supported by the US National Science Foundation's Assembling the Tree of Life program DEB1240049 and DEB1545553. We thank the members of the Duffy lab for helpful feedback on this manuscript.

Truly ubiquitous CRESS DNA viruses scattered across the eukaryotic tree of life

Abstract

Until recently, most viruses detected and characterized were of economic significance, associated with agricultural and medical diseases. This was certainly true for the eukaryote-infecting circular Rep (replication-associated protein)-encoding single-stranded DNA (CRESS DNA) viruses, which were thought to be a relatively small group of viruses. With the explosion of metagenomic sequencing over the past decade and increasing use of rolling-circle replication for sequence amplification, scientists have identified and annotated copious numbers of novel CRESS DNA viruses – many without known hosts but which have been found in association with eukaryotes. Similar advances in cellular genomics have revealed that many eukaryotes have endogenous sequences homologous to viral Reps, which not only provide "fossil records" to reconstruct the evolutionary history of CRESS DNA viruses but also reveal potential host species for viruses known by their sequences alone. The Rep protein is a conserved protein that all CRESS DNA viruses use to assist rolling circle replication that is known to be endogenized in a few eukaryotic species (notably tobacco and water yam). A systematic search for endogenous Rep-like sequences in GenBank's non-redundant eukaryotic database was performed using tBLASTn. We utilized relaxed search criteria for the capture of integrated Rep sequence within eukaryotic genomes, identifying 93 unique species with an endogenized fragment of Rep in their nuclear, plasmid (1 species), mitochondrial (6 species) or chloroplast (8 species) genomes. These species come from 19 different phyla, scattered across the eukaryotic tree of life. Exogenous and endogenous CRESS DNA viral Rep tree topology suggested potential hosts for one family of uncharacterized viruses and supports a primarily fungal host range for genomoviruses.

Keywords: paleovirology, single-stranded DNA virus, integration, CRESS DNA virus

Introduction

Recent metagenomics advances have widened our knowledge of all viruses, including a previously understudied group of viruses, circular Rep-encoding ssDNA (CRESS DNA) viruses (Zhao et al., 2019b, Krupovic et al., 2020). The homologous Rep protein these viruses share is a replication-associated protein that facilitates rolling-circle replication (Rosario et al., 2012b). Several families of viruses with circular ssDNA genomes (CRESS DNA viruses) have been recently united into the order *Cressdaviricota*, including the plant-infecting *Geminivridae* and *Nanoviridae*, the animal-infecting *Circoviridae,* the fungal-infecting *Genomoviridae,* the diatom-infecting *Bacilladnaviridae,* and two families without a confirmed host range: *Smacoviridae* and *Redondoviridae* (Krupovic et al., 2020). Among these the nanoviruses are multipartite, with one ORF for each of their 6-9 genomic segments, and the geminiviruses can be monopartite or have two segments to encode 4-8 proteins. The remaining families are all monopartite and encode fewer ORFs; for instance, smacoviruses only encode a Rep and a capsid protein, while bacilladnaviruses have four ORFs (Krupovic et al. 2020). This new order will likely expand shortly, as sequences have been identified that are not closely related to any of these genera (Kazlauskas et al., 2018; Kinsella et al., 2020). The ubiquitous presence of these CRESS DNA viruses in different environments has been confirmed by numerous sequencing efforts, but seldom have cellular hosts been identified in these metagenomic studies. Many of these viruses may not be very virulent in their hosts (Roossinck and Bazán, 2017), and isolating these hundreds of unclassified viruses to screen against thousands to millions of potential hosts is a daunting task with low probability of success. Another way to narrow the potential host range for viruses known by sequence alone could be through finding “fossil records” inside host genomes, which would indicate that a related virus infected that host some time ago (Dennis et al., 2018b; Patel et al., 2011, Kinsella et al, 2020).

Many viruses integrate themselves inside host genomes during an infection, including retro-transcribing viruses replicating through an integrated DNA intermediate (Nisole and Saïb, 2004). Eight percent of the human genome consists of retroviral elements because they were inserted inside germline cells (Hayward and Katzourakis, 2015); the same process is currently ongoing in the koala genome (Stoye, 2006). Phages, with both dsDNA and ssDNA genomes, can be equipped with integrases and transposases to facilitate endogenization (Krupovic and Forterre, 2015). These interactions can be helpful to the host for a short time, for instance, by providing protection against related lytic viruses, or over millions of years, as endogenized viruses can provide genetic novelty for improved fitness such as in the mammalian placenta (Mi et al., 2000). In eukaryotes, viruses that replicate in the host's nucleus will have a better chance of endogenization than others (Gilbert and Feschotte, 2010). While the mechanisms for retrovirus and dsDNA viral integration are well-studied, how other viral sequences become endogenized is an active area of research (Tu et al., 2017). The mechanism by which the CRESS DNA viruses integrate into their eukaryotic host genomes is still not clear (Krupovic and Forterre, 2015), but the conventional wisdom is that viral use of host replication machinery inside the nucleus facilitates illegitimate recombination between viral and host genome (Belyi et al., 2010; Gilbert and Feschotte, 2010). Integration must sometimes occur in germline cells to allow CRESS DNA endogenous viral elements (EVE) to be transmitted and persist in the host's lineage.

Previous studies have found strong evidence of endogenous CRESS DNA virus sequences in some eukaryotic genomes. One of the earliest studies identified 35 species with Rep-like sequences after a BLAST search using circovirus, geminivirus and nanovirus Rep proteins as queries (Liu et al., 2011). Circovirus-like Rep sequences were found in vertebrates such as cat, dog, panda, frog, opossum, and sloth, and assuming these came from a single genomic integration event, it was dated to ~55 million years ago (Belyi et al., 2010). More recently, another group searched for circovirus-like endogenous elements in vertebrate genome assemblies, confirming

the previously observed shared integration event, and suggesting that circovirus-like sequences had been introduced nineteen times into vertebrate germlines (Dennis et al., 2018a). Geminivirus-like Rep sequences have been found in a number of plant species. Multiple studies have reported *Begomovirus* (a genus within the plant-infecting *Geminiviridae*) -derived sequences inside several tobacco *Nicotiana* species: *N. tabacum, N. tomentosiformis, N. tomentosa* and *N. kawakamii*, (Ashby et al., 1997; Bejarano et al., 1996; Kenton et al., 1995; Murad et al., 2004). Evidence showed geminivirus might have integrated more than once into the ancestors of *Nicotiana* species (Murad et al., 2004). Two endogenized Rep fragments similar to geminiviruses have also been found in the genome of the water yam (*Dioscorea alata*) and 22 other *Dioscorea* species. These Rep fragments appear to be actively expressed: they are under purifying selection and small RNAs and the expressed proteins have been detected in *Dioscorea* (Filloux et al., 2015). Recently, large magnitude surveys were conducted searching for traces of all non-retrotranscribing viral sequences in over four thousand eukaryotic genomes (Kryukov et al., 2018). While CRESS DNA viruses were not the only focus of that study, it showed the distribution of endogenous sequences among diverse eukaryotic taxa. However, there was neither discussion nor detailed presentation of their CRESS DNA virus-like endogenized sequence results.

Paleovirology is the emergent field studying ancient extinct viruses through endogenized sequences or viral “fossil records.” These sequences are not only useful in studying the origin and evolution of viruses, but also have many implications for how viruses have shaped the evolution of their hosts (Feschotte and Gilbert, 2012). Endogenized viral sequences are used to answer questions concerning evolutionary time-scale of the exogenous viruses, such as ancient host-shifting events and determining long-term substitution rates (Gilbert and Feschotte, 2010). Endogenized lentiviruses were used to estimate endogenization events about 4.2 million years ago, and suggested that endogenous sequences are very useful in studying ancestral host-

120 pathogen dynamics and reconstructing ancient viruses (Gilbert et al., 2009). Paleovirology has
121 already yielded important insights for CRESS DNA viruses. While studies of extant crop virus
122 nucleotide sequences often coalesce around the time of the dawn of agriculture (~10,000 years
123 ago), with the evidence of an endogenized Rep sequence from *Nicotiana* spp., we know that
124 geminiviruses originated more than ten million years ago (Gibbs et al., 2006; Lefeuvre et al.,
125 2011). As endogenous sequences can serve as evidence of past host use, further detection of
126 endogenous Rep-like sequences can deduce likely hosts for uncharacterized CRESS DNA viruses
127 (Aiewsakun and Katzourakis, 2015)

128 In this study, we performed a relaxed sequence identity search (tBLASTn) of the non-redundant
129 eukaryotic nucleotide database for endogenized CRESS DNA viral Reps, including a family of
130 Rep-encoding alphasatellites that are known to be related to CRESS DNA viral Reps. We
131 detected endogenous CRESS DNA Rep sequences in 434 unique accession entries from 93
132 unique eukaryotic species. The endogenous Rep fragments came from species of 19 different
133 phyla, scattered across the eukaryotic tree of life. All viral families displayed intriguing findings;
134 for example, genomovirus Reps were closely related to endogenous sequences from fungal
135 genomes, showing that fungal species might indeed serve as hosts for uncharacterized
136 genomoviruses. The circovirus tree showed strong intermingling of exogenous and endogenous
137 sequences, suggesting that extant circovirus diversity might still not be well-sampled.
138 Geminivirus Reps were, as expected, surrounded by endogenous sequences from *Nicotiana* and
139 *Dioscorea* spp. Endogenous sequences found by searching with nanoviruses and related
140 alphasatellite (associated with CRESS DNA viral infection of plants) Reps were from a wide
141 range of hosts, not just their current plant host range. The few endogenous sequences found by
142 searching with bacilladnavirus Reps were distantly related to exogenous viral sequences, which
143 did not help explicate the evolutionary history of these undersampled viruses. Finally, Reps from
144 *Smacoviridae,* a CRESS DNA virus family without a single cultured member, were unexpectedly

found to be similar to a sequence from a diatom, which is not an animal species, with which most smacovirus sequences are found in association.

Material and Methods

BLAST searches

Local tBLASTn runs were carried out with the replication-associated protein (Rep) sequences of CRESS DNA viruses from the RefSeq database (downloaded December 2017) as queries, against the non-redundant (nr) eukaryote nucleotide database (taxid: 2759; downloaded March 2018) from NCBI. The queries dataset includes 66 Reps from *Alphasatellitidae*, 8 Reps from *Bacilladnaviridae*, 154 Reps from *Circoviridae*, 416 Reps from *Geminiviridae*, 67 Reps from *Genomoviridae*, 8 Reps from *Nanoviridae*, 26 Reps from *Smacoviridae*, and 164 Reps from Unclassified ssDNA viruses (Table S1). Family *Redondoviridae* had not been proposed until 2019, and thus it was not included in the scope of this project (Abbas et al., 2019). The tBLASTn search ran with the following less stringent criteria: BLOSUM50 matrix, word size 6, e-value threshold 0.001, gap penalty 15 and extension penalty 1 (Altschul et al., 1990).

BLAST results processing

Repetitive and overlapping hits of the same accession entry resulted from queries from the same family were manually merged into one consensus sequence using Seaview (Gouy et al., 2010). Consensus sequences were omitted if they were less than 50 amino acids in length. We then conducted a reciprocal BLAST analysis, where putative endogenous Rep-like sequences were used as queries for a tblastn search of the full nt/nr GenBank database (June 2020). Only sequences that had at least one high-ranking hit to a CRESS DNA virus were retained for further analysis. To provide more confidence that the sequences identified are truly endogenous, five hundred nucleotides up and down stream of the consensus sequences were extracted from Genbank and scanned for repetitive elements using WSCensor (http://www.girinst.org/censor/,

169 Kohany et al., 2006). The presence of repetitive elements was taken as strong evidence that the
170 Rep-like sequence was integrated into a eukaryotic genome. Sequences that did not have evidence
171 of transposable elements near their sequence were then examined for the size of genomic
172 fragment they had been assembled into. As CRESS DNA virus genomic segments are generally
173 very compact (<6000 bases, Krupovic et al., 2020), we took the presence of more than 6000 bases
174 in either direction on the sequence as evidence that the Rep-like sequence was not in a CRESS
175 DNA virus and instead was integrated into a host's genome. Finally, some of the sequences
176 identified here have been previously noted by other groups that conducted experiments that
177 demonstrate the location of the sequence in a eukaryotic genome (Ashby et al., 1997; Lefeuvre et
178 al., 2011; Theze et al., 2014; Filloux et al., 2015; Metegnier et al., 2015). The sequences that
179 fulfilled at least one of these criteria and the information about their reciprocal BLAST hits are
180 given in Supplementary File 1.

181 Endogenous and viral sequence alignment and tree generation

182 All endogenous consensus sequences and the viral Reps used to search for them were aligned
183 using MUSCLE (default maximum 16 iterations, Edgar, 2004) and trimmed using TrimAl (-
184 gappyout) (Capella-Gutierrez et al., 2009). 256 begomoviruses were taken out of the dataset,
185 leaving 100 representative members of *Begomovirus* within the geminivirus and endogenous
186 sequences dataset. This was to save computation time and avoid overrepresentation of
187 *Begomovirus*, the most speciose genus of all classified viruses, within the alignment. All trimmed
188 endogenous and viral sequence alignments were inputs to PhyML 3.0 (Guindon and Gascuel,
189 2003) to estimate maximum likelihood trees using CRESS+G+F model (Zhao et al., 2019a). The
190 Shimodaira-Hasegawa approximate likelihood ratio test (SH-aLRT, or SH-like) statistic was
191 selected as the branch support option from PhyML, which is a more efficient choice for larger
192 data sets with good accuracy (Guindon et al., 2010). Trees were visualized and colored using
193 Figtree (http://tree.bio.ed.ac.uk/software/figtree/).

Results and Discussion

tBLASTn results

With our relaxed search criteria and using 908 CRESS DNA viral Rep amino acid sequences as queries, we were able to obtain 111,344 raw hits after the search, which collapsed to 434 unique accession entries, 93 unique species and 19 eukaryotic phyla (Table 1 and Figure 1). Bacilladnavirus Reps found the smallest number of similar sequences in eukaryotic genomes and geminivirus Reps found the greatest number of hits per viral Rep. However, circovirus Reps were the most widespread sequences, as they were found in 69 unique species genomes. These endogenous Rep sequences spread across Plantae, Chromalveolates, Unikonts, and Excavates across the eukaryotic tree of life. The phyla with the most representing species in our study are Magnoliophyta, Ascomycota and Chordata. Some of the species were identified with multiple Rep queries, and thus are present in multiple phylogenetic analyses, aligned with different extant CRESS DNA viruses in different trees. In some cases this reflects the homology of Rep among CRESS DNA virus families. This is best illustrated with the geminiviruses and genomoviruses, as the latter was named the Gemini-like No Movement protein viruses based on similarities to geminiviruses including a closely related Rep protein (Varsani and Krupovic, 2017). In others, it perhaps reflects that the integrated sequences may be most related to an as yet unclassified CRESS DNA virus family, as could be the case for the bulk of hits in the *Entamoeba* spp. (Kinsella et al., 2020).

Maximum likelihood trees

Geminivirus Reps and endogenous sequences

Seven maximum likelihood trees were built with PhyML3 using the CRESS+G+F model (Zhao et al., 2019a), one for each family of Reps used to query the database. The geminivirus and similar endogenous sequences ML tree is shown in Figure 2 (a version with accession numbers for all

sequences is shown in Supplementary Tree S2). While the majority of eukaryotic species in which CRESS DNA endogenous elements were found were plants, there are some surprising results showing geminivirus-like sequences integrated into protists, oomycetes and fungi. The unclassified species Niminivirus and Baminivirus are in a well-supported clade containing species from the fungal groups Ascomycota and Basidiomycota (SH-like support 0.964). One large clade (SH-like support 0.956) containing all Reps from *Begomovirus*, *Curtovirus*, *Topocurtovirus*, and *Turncurtovirus* includes only one endogenous Rep sequence, from the common sunflower (*Helianthus annuus*). The sister group to this clade is composed entirely of *Nicotiana* species (tobacco), in which endogenous geminivirus Rep homologues have been long described (Ashby et al., 1997; Gibbs et al., 2006; Lefeuvre et al., 2011). The sister group to these two clades exclusively contains more endogenized sequences, mostly *Dioscorea* species (water yam, previously described by Filloux et al., 2015), three additional sequences from *Nicotiana* and two from the mitochondrion of *Amborella trichopoda* (understory shrubs, KF754803). These two clades of endogenized sequences and the clade dominated by extant begomovirus sequences formed a distinct group (SH-like 0.787). Deeper in the tree, the intron-containing (spliced) and non-intron-containing (unspliced) geminivirus Rep sequences are quite separated (as in Filloux et al 2015, Zhao et al., 2019a), and there are no endogenous sequences from plant genomes that group with the spliced Reps. More distantly related geminivirus-like Rep sequences were found in another plant (narrow-leafed ash *Fraxinus angustifolia,* as had been noted by Filloux et al., 2015), the chloroplast genomes of two species (*Euglena garcilis* and *Paradoxia multiseta*), several *Entamoeba* species, an algal plasmid, and the mitochondria of two omycetes (*Phytophthora infestans, Peronospora tabacina*).

Endogenous sequences similar to geminivirus Reps have previously been found not only in many *Nicotiana* and *Dioscorea* spp., but also in assorted other plants like black cottonwood (*Populus trichocarpa*, Liu et al., 2011, Filloux et al., 2015), lettuce (*Lactuca sativa*, Filloux et al., 2015)

and Coffee (Sharma et al., 2020). We can add common sunflower (*Helianthus annuus*) to this list, but we did not find these previously established homologs in our search. This is due to the non-redundant eukaryotic database excluding the whole genome sequencing projects that produced the genomic sequences for these plants. Our complementary analyses are conservative, and our relaxed search strategy identified homologs in a much wider range of taxa than previous studies (e.g., putative endogenized geminivirus sequences in fungus, compared to the few in Liu et al., 2011).

Circovirus Reps and endogenous sequences

The circovirus Reps and similar endogenous sequences are shown in a tree in Figure 3 (a version with accession numbers for all sequences is shown in Supplementary Tree S3). The close relationships and intermingling of exogenous viral sequences and endogenous elements justifies the significant research done on EVEs of this viral family (Liu et al., 2011; Theze et al., 2014; Metegnier et al., 2015); circoviruses leave marks in the genomes of their hosts. Endogenous sequences similar to circovirus Rep proteins were found in 69 eukaryotic species from 17 phyla. Unlike the geminiviruses (Figure 2), the extant circoviruses are not together in large groups, and close relationships between extant and endogenous sequences are apparent. An exception to this would be the well-supported clade (SH-like support 0.963) containing most of the sequences assigned to the genus *Cyclovirus* (Rosario et al., 2017). However, this clade still includes endogenous Rep-like elements from four species: an ant (*Pseudomyrmex gracilis*) and three tapeworms (*Hymenolepis diminuta*, *Spirometra erinaceieuropaei*, *Taenia asiatica*). This provides independent support for the invertebrate host range inferred for many of cyclovirus species (Rosario et al., 2012a; Rosario et al., 2018).

The official members of genus *Circovirus* are dispersed throughout the tree, with classified members forming clades with both unclassified circoviruses and endogenized eukaryotic

sequences. Some clades reinforced similar host ranges for extant circoviruses and their related integrated sequences. For instance, the clade containing Barbel circovirus includes sequences from four fish species (and one sequence from the parasitic mite *Varroa jacobsoni*, SH-like support 0.952). However, this was not the only part of the tree that included fish – salmon and carp sequences both also appear elsewhere in the tree, and a sequence from spiny chromis damselfish forms a weakly supported clade with a circovirus isolated from a bird (Garrulus glandarius associated virus 1). Other well-supported clades suggest hosts for uncultured viruses. For instance, sequences from two related parasitic flukes (*Opisthorchis viverrini* and *Dicrocoelium dendriticum*) group well with two aquatic animal-associated extant viral sequences (SH-like support 0.946). These circovirus sequences obtained from marine animals in their natural habitat may reflect viral infections of their parasites such as flukes. The unclassified circovirus species had sister groups from a much wider range than either of the two genera, suggesting a very wide host range for this family among animals, commensurate with the high sequence diversity among unclassified circoviruses (Dayaram et al., 2014; Li et al., 2010; Steel et al., 2016).

The wide net cast by our search strategy is evident by the large representation on the tree of sequences related to geminiviruses (especially those in *Nicotiana* and *Dioscorea*), but these were not closely grouped with extant circoviruses and the results are understandable due to the homology between geminivirus and circovirus Rep proteins. More surprising was that circovirus Rep queries yielded more chloroplast hits than plant-infecting geminivirus Rep queries (Figure 2). This could be another artifact of our relaxed search approach, as reciprocal BLAST searches for the chloroplast sequences in this tree produced strong hits to geminiviruses (4 sequences), alphasatellites (2 sequences) with only the chloroplast of the diatom *Pseudo nitzschia multiseries* having a hermit crab-associated circovirus as its top viral hit (Supplementary File 1). The relatively close groupings of some chloroplast sequences with circulating circoviruses (e.g., *Cylindrotheca closterium*'s chloroplast clustering with the fiddler crab associated circular virus,

SH-like support 0.861) could reflect a wider host range of sequenced exogenous circoviruses than currently understood, or bizarre parallel evolution towards circovirus-like Rep sequences after an ancestral integration of a CRESS DNA virus Rep that was associated with plants. There also exists the possibility that not all of the data in GenBank is accurate – either in sequence, or in the declared organism the nucleic acid came from. For instance, we found that a handful of sequences that we used in our analysis have since been retracted from GenBank at the submitters' request (e.g., *Taenia asiatica*, which grouped with the exogenous cycloviruses).

Nanovirus Reps and endogenous sequences

In Figure 4 (a version with accession numbers for all sequences is shown in Supplementary Tree S4), the small number of sequenced nanovirus Reps formed their own clade (SH-like support 0.933). The eukaryotic sequences identified by the nanovirus Rep queries were largely those that were found in the circovirus analysis (Figure 3) and do not reflect the plant host range of extant nanoviruses. The circulating nanoviruses are most closely related to the Rep-like sequence of the liver fluke *O. viverrini*, and are more distantly related to sequences found in Varroa mites and marine chordates, then to pillbugs and a crustacean. The other clade in the tree includes more diverse taxa: additional marine invertebrates, a fungus, a green alga (*Micromonas pusilla*) and three chloroplast sequences. This tree is not very informative about the historical host range of the nanoviruses because the current sequences form a clade and it is likely that the other sequences are more closely related to Reps from other families. For instance, the sister taxon *O. vierrrini*'s best viral reciprocal BLAST hit was to a circovirus not a nanovirus (Supplementary File 1).

Genomovirus Reps and endogenous sequences

Most genomovirus Reps are in a strongly supported clade (SH-like support 0.975) in the viral and eukaryotic sequence tree in Figure 5 (a version with accession numbers for all sequences is

shown in Supplementary Tree S5). Together with sequences from two Basidiomycotes (mushroom bicoloured deceiver [*Laccaria bicolor*] and dry rot fungus [*Serpula lacrymans* var. lacrymans S7.9]) and one Ascomycete (*Exophilala spinifera*), all the exogenous genomoviruses form a well-supported clade (SH-like support 0.981). The more distantly related sequences to the genomoviruses, understandably, resemble a subset of the sequences related to geminiviruses: plant endogenous sequences (from *Nicotiana* and *Dioscorea*, *Helianthus annuus* and *Fraxinus angustifolia*), *Entamoeba* spp, the mitochondria of oomycetes and *Amborella trichopoda*, the red algal plasmid and one of the chloroplast sequences. There are a few more fungal taxa, including taxa previously identified by Liu et al (2011): *Aspergillus nidulans, Nectria haematococca*, *Magnaporthe oryzae*, and some novel Rep-like sequences (*Cordyceps militaris*, *Verticillium dahlia, Colletotrichum higginsianum, Metarhizium majus*) but no other endogenous sequences are closely related to the genomoviruses. All of these fungal sequences were also identified in the geminivirus tree (Figure 2) but were closely related only to the unclassified Niminivirus and Baminivirus sequences – distantly related to all the recognized genera of *Geminiviridae*. These results suggest that these viral sequences might be misclassified as geminiviruses, and might fit better within *Genomoviridae*.

The only cultivated genomovirus was isolated from a fungus (SsHADV-1, which infects the ascomycete *Sclerotinia sclerotiorum,* Yu et al., 2010), and this tree suggests that the current and extinct genomovirus host range is primarily restricted to fungi. Recently, a study showed SsHADV-1 is able to infect fungi and the mycophagous insects feeding on them (Liu et al., 2016). Virus-like particles with genome sequences most similar to genomoviruses have also been found in fungus-farming termites (Kerr et al., 2018). These findings suggest members of *Genomoviridae* are primarily fungal infecting but may also infect fungal-feeding predators. Host range is hard to infer from the presence of viral genomes, since, for example, fungal viruses are often part of a meal of fungus that an insect would eat. This is one of the major reasons that many

sequences of CRESS DNA viruses are named with the word “associated” since it is increasingly rare that the identified virus has a confirmed ability to infect any host (Zhao et al., 2019b). It has been previously proposed that SsHADV-1 could be a biological fungal pathogen control agent (Yu et al., 2010). Our findings would undermine that application, since they indicate genomoviruses likely have infected fungi for a long period of time, and the integrated genomovirus-like sequences inside fungal hosts suggest some fungi may be able to resist viral infection through the production of small antiviral RNAs (Campo et al., 2016).

Bacilladnavirus Reps and endogenous sequences

There were very few eukaryotic sequences identified using the Reps of bacilladnaviruses as queries (Figure 6, a version with accession numbers for all sequences is shown in Supplementary Tree S6). This could be due to the small number of representative Reps from this newly codified family (Kazlauskas et al., 2017), compared to the diversity within the geminivirus, circovirus and genomovirus query sets. Alternatively, this could reflect the true relationship between the bacilladnavirus Reps and endogenous sequences: that no close relatives to bacilladnaviruses have integrated into sequenced eukaryotes, or that such events happened so long ago as to erase any close sequence relationship. The extant viruses formed a well-supported clade (SH-like support 0.918), and only two other species joined them on this tree: the parasitic Loa loa worm and a parasitic alveolate, *Gregarina niphandrodes*. These sequences were also identified with circovirus queries and are very distantly related to the heterokont diatoms that bacilladnaviruses are known to infect, so this analysis does not expand our current, limited understanding of this group’s host range.

Smacovirus Reps and endogenous sequences

Smacovirus Reps also did not produce many hits in eukaryotic genomes and many of the few identified endogenous sequences were also found when querying with Reps from other families. A sequence from the chloroplast of a diatom (*Pseudo-nitzschia multiseries*), which had been identified in the circovirus search, was nested within the extant smacovirus Reps, implying a closer relationship to smacoviruses than circoviruses (Figure 7, a version with accession numbers for all sequences is shown in Supplementary Tree S7). Therefore, the smacovirus-like endogenous sequences were detected in a basal metazoan, an arthropod, a fungus, an alga, a sea snail, and the California two-spot octopus. However, the best viral reciprocal BLAST hits for these species were neither in nanovirus nor smacovirus (they were variously circovirus, geminivirus, alphasatellite, and unclassified CRESS DNA viruses, Supplemental File 1), indicating that only the diatom chloroplast sequence likely has a more recent relative with extant smacoviruses. As this viral family has no confirmed hosts, it was disheartening that other potential hosts were not identified in this study. This could be because hosts that have endogenized sequences related to smacoviruses have not yet been sequenced, or there might be no benefit to hosts to maintain smacovirus Rep-like sequences. Perhaps smacoviruses are less likely to experience sequence endogenization compared to other CRESS DNA viruses. Since the definitive hosts of smacoviruses have not yet been identified, any explanation for the single diatom species' chloroplast genome with sequence similarity is speculative. These results do not shed light on why viruses that have been so widely associated with vertebrate and invertebrate samples (Varsani and Krupovic, 2018), might have an endogenous fossil in diatoms.

Alphasatellite Reps and endogenous sequences

The alphasatellite Reps, which group with nanovirus Reps in phylogenies of CRESS DNA viruses (Simmonds et al., 2017) found sequences similar to BLAST hits to both the nanoviruses and geminiviruses with which they associate (Figure 8, a version with accession numbers for all

sequences is shown in Supplementary Tree S8). Most of the extant geminivirus- and nanovirus-associated alphasatellites formed a clade with strong support, with no endogenous sequences (SH-like support 0.991). With a relative diverse set of exogenous sequences, the alphasatellite Rep queries were able to identify divergent homologs in eukaryotic genomes because they still found the *Nicotiana* and *Dioscorea* endogenous elements that are unambiguously due to geminivirus integration events. Alphasatellite Reps are much more closely related to nanovirus Reps than geminivirus Reps (Zhao et al., 2019b), and nanovirus queries found relatively few Rep-like sequences in eukaryotic genomes, so it was surprising that alphasatellites were able to find geminivirus-like Rep-like sequences. This is likely due to the small number of queries used in the nanovirus endogenous search compared to the alphasatellite queried search; very few nanovirus species have been identified and sequenced. The alphasatellite tree shared many of the same eukaryotic sequences as both the nanoviruses and geminiviruses, evincing both their close ancestry with nanoviruses and the obligately shared host range of the alphasatellites with the nanoviruses and geminiviruses. These included terrestrial and marine animals found on the circovirus (Figure 3) and bacilladnavirus trees (Figure 6), which was unexpected from exclusively plant-associated viruses and satellites (Briddon et al., 2018). Perhaps the host range of alphasatellites extends to animals, and these elements might be found in association with CRESS DNA viruses of animals in the future. Four chloroplast sequences were also found with the alphasatellite dataset, but none of the sequences were closely grouped with alphasatellite sequences and it is much more likely that the chloroplast sequences descended from viral Reps instead of these satellite Reps.

Endogenous elements in organelles

There are three mitochondria and seven chloroplasts from eukaryotes that appear in these seven trees, often in more than one tree (three additional mitochondrial sequences and an eighth

418 chloroplast sequence were identified by using unclassified CRESS DNA virus Reps as queries,
419 Supplementary File 1). The mechanism of how eukaryotic CRESS DNA viruses would integrate
420 into these erstwhile prokaryotes is an open area of inquiry, though it is thought that some
421 replication genes in mitochondria trace back to double-stranded DNA T-odd-like phages (Shutt
422 and Gray, 2006). Other researchers who have found CRESS DNA virus-like sequences in
423 mitochondria did not attempt a mechanistic explanation (Liu et al., 2011), but there are some
424 theoretical ways Rep proteins might be useful in an organelle. Mitochondrial genomes and their
425 plasmids have been detected in single-stranded states, undergoing rolling-circle replication in
426 higher plants (Backert et al., 1996). This suggests the requirement of a protein with a similar
427 function as Rep, which would not have been ancestral to mitochondrial genomes. Similarly,
428 strand displacement during mtDNA synthesis has also been suggested as a mode of human
429 mitochondrial DNA replication (Miralles Fusté et al., 2014). Independent of these obvious uses of
430 a rolling-circle replication enhancing enzyme in mitochondria, it is likely that these Rep-like
431 sequences have a function in the organelles, since these organelles are so genome-reduced that it
432 is hard to imagine useless integrated sequence lasting over long evolutionary times (Smith and
433 Keeling, 2015). It is unlikely that these Rep-like sequences are descended from a prokaryotic
434 virus with a Rep protein, because all phage RepA proteins are quite divergent from eukaryotic
435 CRESS DNA virus Reps – they are more different from the Reps studied here than the CRESS
436 DNA viral Reps are from one another (Koonin and Ilynia 1993). Although no direct evidence of a
437 Rep from any CRESS DNA virus has been shown to be harnessed by mitochondria, we identified
438 several potential candidates for such investigations. The mitochondria of *Amborella trichopoda*,
439 *Peronospora tabacina*, *and Phytophthora infestans* and the chloroplasts of *Pediastrum duplex*,
440 *Dunaliella salina*, *Euglena gracilis*, *Pediastrum duplex*, *Paradoxia multiseta*, *Pseudo-nitzschia*
441 *multiseries*, and *Cylindrotheca closterium* might hold intriguing evidence showing tangents of the
442 evolution of the CRESS DNA viral Rep protein.

The Reps of eukaryotic CRESS DNA viruses form a clade with the Reps of ssDNA plasmids, including those of red algae, and it has been proposed that the viruses evolved from the plasmid (Krupovic et al., 2009; Saccardo et al., 2011). When a Rep-like sequence was first observed in the oomycete *P. infestans*, it grouped with a Rep from ssDNA algal plasmids (Liu et al., 2011), and a plasmid from *Pyropia pulchra* is in the same clade with the *Phytophthora* mitochondrial sequences. These eukaryotic sequences set apart from the more geminivirus-like clades (Figure 2) could well represent Rep elements from plasmids, not viruses. While plasmids are more expected than eukaryotic viruses inside organelles, this cannot explain the Rep-like sequences that are very closely related to exogenous viruses, such as one of the integration events into the *Amborella trichopoda* mitochondrion or the *Cylindrotheca closterium* chloroplast.

Conclusions

Despite the proliferation of papers discussing endogenized CRESS DNA viral sequences, we were able to find more endogenous circovirus Rep sequences and fragments than previous studies (Dennis et al., 2018a; Liu et al., 2011; Metegnier et al., 2015; Theze et al., 2014). This is because we used relaxed search criteria, allowing for distantly related sequences to be included, which also accounts for the rapid evolution of proteins in ssDNA viruses (Duffy and Holmes, 2008; Firth et al., 2009) and diversification over the potential millions of years since integration (Gibbs et al., 2006; Lefeuvre et al., 2011). Many of the sequences identified here have been previously observed to be related to CRESS DNA viruses of eukaryotes and were found through multiple different viral family searches in our study (e.g., *Phytophthora infestans*, Liu et al., 2011; *Entamoeba histolytica* HM-1:IMSS, Gibbs et al., 2006). All of the Rep-like sequences identified here, especially those in supported clades with exogenous CRESS DNA viral Reps, either are the product of very recent integration events or are likely under some purifying selection, since all homologs were identified by conserved sequence similarity. While we cannot quantify the lineages that may have been harmed by integrating CRESS DNA Rep sequences, the existence of

so many Rep-like sequences in contemporary times, some of them reflecting integration events that occurred millions of years ago, means that these sequences derived from viruses may have been beneficial for at least some of the hosts.

The preponderance of CRESS DNA viruses known by sequence alone has stymied our understanding of the ecological impact of this group. Although all genomic sequences inferred from metagenomes that we used in this study were verified with PCR and Sanger sequencing to obtain the complete genome sequences, we lack the biological isolates to conduct any host range screening for nearly all of these species (Male et al., 2016; Rosario et al., 2015; Steel et al., 2016). We hope the host taxa showing evidence of a historical host range for these viruses will help researchers target the hosts of a given CRESS DNA virus family, and increase the odds of isolating viral particles for in depth characterization. We find evidence to support the host range assertions of several CRESS DNA virus groups, including arthropod-infecting circoviruses (Dayaram et al., 2014; Dayaram et al., 2013; Rosario et al., 2012a; Rosario et al., 2011; Rosario et al., 2018), the fungal host range of genomoviruses (Yu et al., 2010), and the closely related species that likely expand the host range of vertebrate-infecting circoviruses (Dennis et al., 2018a; Dennis et al., 2018b). We look forward to further screening of the genomes and transcriptomes that will be sequenced in the coming years as we anticipate our relaxed search approach will help reveal many more endogenized Rep-like sequences in the future.

Data Availability

All sequence data was retrieved from NCBI GenBank (accession numbers in supplemental figures and files). Alignments and phylogenetic trees can be accessed from Dryad. https://doi.org/10.5061/dryad.280gb5mqs

References

Abbas, A.A., Taylor, L.J., Dothard, M.I., Leiby, J.S., Fitzgerald, A.S., Khatib, L.A., Collman, R.G. & Bushman, F.D. (2019) Redondoviridae, a family of small, circular DNA viruses of the

493 human oro-respiratory tract associated with periodontitis and critical illness. Cell Host and
494 Microbe, 25, 719-729.
495 Aiewsakun, P., & Katzourakis, A. (2015). Endogenous viruses: Connecting recent and
496 ancient viral evolution. Virology, 479-480, 26-37.
497 Altschul, S.F., Gish, W., Miller, W., Myers, E.W., & Lipman, D.J. (1990). Basic local
498 alignment search tool. Journal of Molecular Biology, 215, 403-410.
499 Ashby, M.K., Warry, A., Bejarano, E.R., Khashoggi, A., Burrell, M., & Lichtenstein,
500 C.P. (1997). Analysis of multiple copies of geminiviral DNA in the genome of four
501 closely related Nicotiana species suggest a unique integration event. Plant Molecular
502 Biology, 35, 313-321.
503 Backert, S., Dörfel, P., Lurz, R., & Börner, T. (1996). Rolling-circle replication of
504 mitochondrial DNA in the higher plant Chenopodium album (L.). Molecular and Cellular
505 Biology, 16, 6285-6294.
506 Bejarano, E.R., Khashoggi, A., Witty, M., & Lichtenstein, C. (1996). Integration of
507 multiple repeats of geminiviral DNA into the nuclear genome of tobacco during
508 evolution. Proceedings of the National Academy of Sciences, 93, 759-764.
509 Belyi, V.A., Levine, A.J., & Skalka, A.M. (2010). Sequences from ancestral single-
510 stranded DNA viruses in vertebrate genomes: the parvoviridae and circoviridae are more
511 than 40 to 50 million years old. Journal of Virology, 84, 12458-12462.
512 Briddon, R.W., Martin, D.P., Roumagnac, P., Navas-Castillo, J., Fiallo-Olive, E.,
513 Moriones, E., Lett, J.M., Zerbini, F.M., Varsani, A., 2018. Alphasatellitidae: a new
514 family with two subfamilies for the classification of geminivirus- and nanovirus-
515 associated alphasatellites. Arch Virol. 163, 2587-2600.
516 Campo, S., Gilbert, K.B., & Carrington, J.C. (2016). Small RNA-Based Antiviral
517 Defense in the Phytopathogenic Fungus Colletotrichum higginsianum. PLoS Pathogens,
518 12, e1005640-e1005640.
519 Capella-Gutierrez, S., Silla-Martinez, J.M., & Gabaldon, T. (2009). trimAl: a tool for
520 automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics
521 (Oxford, England), 25, 1972-1973.

Dayaram, A., Galatowitsch, M., Harding, J.S., Arguello-Astorga, G.R., & Varsani, A. (2014). Novel circular DNA viruses identified in Procordulia grayi and Xanthocnemis zealandica larvae using metagenomic approaches. Infect. Genet. Evol., 22, 134-141.

Dayaram, A., Potter, K.A., Moline, A.B., Rosenstein, D.D., Marinov, M., Thomas, J.E., Breitbart, M., Rosario, K., Arguello-Astorga, G.R., & Varsani, A. (2013). High global diversity of cycloviruses amongst dragonflies. J. Gen. Virol., 94, 1827-1840.

Dennis, T.P.W., de Souza, W.M., Marsile-Medun, S., Singer, J.B., Wilson, S.J., & Gifford, R.J. (2018a). The evolution, distribution and diversity of endogenous circoviral elements in vertebrate genomes. Virus Research, 262, 15-30.

Dennis, T.P.W., Flynn, P.J., Marciel de Souza, W., Singer, J.B., Moreau, C.S., Wilson, S.J., &Gifford, R.J. (2018b). Insights into circovirus host range from the genomic fossil record. Journal of Virology, 92, e00145-18.

Duffy, S., & Holmes, E.C. (2008). Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. J Virol 82, 957-965.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res, 32, 1792-1797.

Feschotte, C., & Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. Nature Reviews Genetics, 13, 283-296.

Filloux, D., Murrell, S., Koohapitagtam, M., Golden, M., Julian, C., Galzi, S., Uzest, M., Rodier-Goud, M., D'Hont, A., Vernerey, M.S., Wilkin, P., Peterschmitt, M., Winter, S., Murrell, B., Martin, D.P. & Roumagnac, P. (2015). The genomes of many yam species contain transcriptionally active endogenous geminiviral sequences that may be functionally expressed. Virus Evolution, 1, vev002.

Firth, C., Charleston, M.A., Duffy, S., Shapiro, B., & Holmes, E.C. (2009). Insights into the Evolutionary History of an Emerging Livestock Pathogen: Porcine Circovirus 2. Journal of Virology, 83, 12813.

Gibbs, M.J., Smeianov, V.V., Steele, J.L., Upcroft, P., & Efimov, B.A. (2006). Two families of rep-like genes that probably originated by interspecies recombination are represented in viral, plasmid, bacterial, and parasitic protozoan genomes. Molecular Biology and Evolution, 23, 1097-1100.

553 Gilbert, C., & Feschotte, C. (2010). Genomic Fossils Calibrate the Long-Term Evolution
554 of Hepadnaviruses. PLOS Biology, 8, e1000495.
555 Gilbert, C., Maxfield, D.G., Goodman, S.M., & Feschotte, C. (2009). Parallel germline
556 infiltration of a lentivirus in two Malagasy lemurs. PLoS Genetics, 5, e1000425.
557 Gouy, M., Guindon, S., & Gascuel, O. (2010). SeaView Version 4: A Multiplatform
558 Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building.
559 Molecular Biology and Evolution, 27, 221-224.
560 Guindon, S., & Gascuel, O. (2003). A Simple, Fast and Accurate Algorithm to Estimate
561 Large Phylogenies by Maximum Likelihood. Syst Biol, 52, 696-704.
562 Hayward, A., & Katzourakis, A. (2015). Endogenous retroviruses. Curr Biol, 25, R644-
563 646.
564 Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M. Hordijk, W. & Gascuel, O.
565 (2010). New slgorithms and methods to estimate maximum-likelihood phylogenies:
566 assessing the performance of PhyML 3.0. Syst Biol, 59, 307-321.
567 Kazlauskas, D., Dayaram, A., Kraberger, S., Goldstien, S., Varsani, A., & Krupovic, M.
568 (2017). Evolutionary history of ssDNA bacilladnaviruses features horizontal acquisition
569 of the capsid gene from ssRNA nodaviruses. Virology, 504, 114-121.
570 Kazlauskas, D., Varsani, A. & Krupovic, M. (2018). Pervasive chimerism in the
571 replication-associated proteins of uncultured single-stranded DNA viruses. Viruses, 10,
572 187.
573 Kenton, A., Khashoggi, A., Parokonny, A., Bennett, M.D., & Lichtenstein, C. (1995).
574 Chromosomal location of endogenous geminivirus-related DNA sequences in Nicotiana
575 tabacum L. Chromosome research, 3, 346-350.
576 Kerr, M., Rosario, K., Baker, C.C.M., & Breitbart, M. (2018). Discovery of Four Novel
577 Circular Single-Stranded DNA Viruses in Fungus-Farming Termites. Genome
578 Announcements, 6, e00318-00318.
579 Kinsella, C.M., Bart, A., Deijs, M., Broekhuizen, P., Kaczorowska, P., Jebbink, M.F.,
580 van Gool, T., Cotton, M. & van der Hoek, L. (2020). Entamoeba and Giardia parasites
581 implicated as hosts of CRESS viruses. Nature Communications, 11, 4620.

582 Kohany, O., Gentles, A.J., Hankus, L. & Jurka, J. (2006). Annotation, submission and
583 screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC
584 Bioinformatics, 7, 747.
585 Krupovic, M., & Forterre, P. (2015). Single-stranded DNA viruses employ a variety of
586 mechanisms for integration into host genomes. Annals of the New York Academy of
587 Sciences, 1341, 41-53.
588 Krupovic, M., Ravantti, J.J., & Bamford, D.H. (2009). Geminiviruses: a tale of a plasmid
589 becoming a virus. BMC Evol. Biol., 9, 112.
590 Krupovic, M., Varsani, A., Kazlauskas, D., Breitbart, M., Delwart, E., Rosario, K., Yutin,
591 N., Wolf, Y.I., Harrach, B., Zerbini, F.M., Dolja, V.V., Kuhn, J.H., & Koonin, E.V.
592 (2020) Cressdnaviricota: a virus phylum unifying several families of Rep-encoding
593 viruses with single-stranded, circular DNA genomes. J Virol, 94, e00582-20.
594 Kryukov, K., Ueda, M.T., Imanishi, T., & Nakagawa, S. (2018). Systematic survey of
595 non-retroviral virus-like elements in eukaryotic genomes. Virus Research, 262, 30-36.
596 Lefeuvre, P., Harkins, G.W., Lett, J.-M., Briddon, R.W., Chase, M.W., Moury, B., &
597 Martin, D.P. (2011). Evolutionary Time-Scale of the Begomoviruses: Evidence from
598 Integrated Sequences in the Nicotiana Genome. PLOS ONE, 6, e19193.
599 Li, L., Kapoor, A., Slikas, B., Bamidele, O.S., Wang, C., Shaukat, S., Masroor, M.A.,
600 Wilson, M.L., Ndjango, J.B., Peeters, M., Gross-Camp, N.D., Muller, M.N., Hahn, B.H.,
601 Wolfe, N.D., Triki, H., Bartkus, J., Zaidi, S.Z., & Delwart, E. (2010). Multiple diverse
602 circoviruses infect farm animals and are commonly found in human and chimpanzee
603 feces. J Virol, 84, 1674-1682.
604 Liu, H.Q., Fu, Y.P., Li, B., Yu, X., Xie, J.T., Cheng, J.S., Ghabrial, S.A., Li, G.Q., Yi,
605 X.H., & Jiang, D.H. (2011). Widespread Horizontal Gene Transfer from Circular Single-
606 stranded DNA Viruses to Eukaryotic Genomes. BMC Evol. Biol., 11, 15.
607 Liu, S., Xie, J., Cheng, J., Li, B., Chen, T., Fu, Y., Li, G., Wang, M., Jin, H., & Wan, H.
608 (2016). Fungal DNA virus infects a mycophagous insect and utilizes it as a transmission
609 vector. Proceedings of the National Academy of Sciences, 113, 12803-12808.
610 Male, M.F., Kraberger, S., Stainton, D., Kami, V., & Varsani, A. (2016). Cycloviruses,
611 gemycircularviruses and other novel replication-associated protein encoding circular

612 viruses in Pacific flying fox (Pteropus tonganus) faeces. Infect. Genet. Evol., 39, 279-
613 292.
614 Metegnier, G., Becking, T., Chebbi, M.A., Giraud, I., Moumen, B., Schaack, S., Cordaux,
615 R., & Gilbert, C. (2015). Comparative paleovirological analysis of crustaceans identifies
616 multiple widespread viral groups. Mobile DNA, 6, 16.
617 Miralles Fusté, J., Shi, Y., Wanrooij, S., Zhu, X., Jemt, E., Persson, Ö., Sabouri, N.,
618 Gustafsson, C.M., & Falkenberg, M. (2014). In Vivo Occupancy of Mitochondrial
619 Single-Stranded DNA Binding Protein Supports the Strand Displacement Mode of DNA
620 Replication. PLoS Genetics, 10, e1004832.
621 Murad, L., Bielawski, J.P., Matyasek, R., Kovarik, A., Nichols, R.A., Leitch, A.R., &
622 Lichtenstein, C.P. (2004). The origin and evolution of geminivirus-related DNA
623 sequences in Nicotiana. Heredity, 92, 352-358.
624 Nisole, S., & Saïb, A. (2004). Early steps of retrovirus replicative cycle. Retrovirology, 1,
625 9.
626 Patel, M.R., Emerman, M., & Malik, H.S. (2011). Paleovirology—ghosts and gifts of
627 viruses past. Current Opinion in Virology, 1, 304-309.
628 Roossinck, M.J., & Bazán, E.R. (2017). Symbiosis: Viruses as Intimate Partners. Annual
629 Review of Virology, 4, 123-139.
630 Rosario, K., Breitbart, M., Harrach, B., Segales, J., Delwart, E., Biagini, P., & Varsani,
631 A. (2017). Revisiting the taxonomy of the family Circoviridae: establishment of the
632 genus Cyclovirus and removal of the genus Gyrovirus. Archives of Virology, 162, 1447-
633 1463.
634 Rosario, K., Dayaram, A., Marinov, M., Ware, J., Kraberger, S., Stainton, D., Breitbart,
635 M., & Varsani, A. (2012a). Diverse circular ssDNA viruses discovered in dragonflies
636 (Odonata: Epiprocta). Journal of General Virology, 93, 2668-2681.
637 Rosario, K., Duffy, S., & Breitbart, M. (2012b). A field guide to eukaryotic circular
638 single-stranded DNA viruses: insights gained from metagenomics. Archives of Virology,
639 157, 1851-1871.
640 Rosario, K., Marinov, M., Stainton, D., Kraberger, S., Wiltshire, E.J., Collings, D.A.,
641 Walters, M., Martin, D.P., Breitbart, M., & Varsani, A. (2011). Dragonfly cyclovirus, a

642 novel single-stranded DNA virus discovered in dragonflies (Odonata: Anisoptera).
643 Journal of General Virology, 92, 1302-1308.
644 Rosario, K., Mettel, K.A., Benner, B.E., Johnson, R., Scott, C., Yusseff-Vanegas, S.Z.,
645 Baker, C.C.M., Cassill, D.L., Storer, C., Varsani, A., & Breitbart, M. (2018). Virus
646 discovery in all three major lineages of terrestrial arthropods highlights the diversity of
647 single-stranded DNA viruses associated with invertebrates. PeerJ, 6, e5761.
648 Rosario, K., Schenck, R.O., Harbeitner, R.C., Lawler, S.N., & Breitbart, M. (2015).
649 Novel circular single-stranded DNA viruses identified in marine invertebrates reveal high
650 sequence diversity and consistent predicted intrinsic disorder patterns within putative
651 structural proteins. Frontiers in Microbiology, 6, 13.
652 Saccardo, F., Cettul, E., Palmano, S., Noris, E., & Firrao, G. (2011). On the alleged origin
653 of geminiviruses from extrachromosomal DNAs of phytoplasmas. BMC Evol. Biol., 11,
654 185.
655 Shutt, T.E., & Gray, M.W. (2006). Bacteriophage origins of mitochondrial replication
656 and transcription proteins. Trends in Genetics, 22, 90-95.
657 Simmonds, P., Adams, M.J., Benko, M., Breitbart, M., Brister, J.R., Carstens, E.B.,
658 Davison, A.J., Delwart, E., Gorbalenya, A.E., Harrach, B., Hull, R., King, A.M., Koonin,
659 E.V., Krupovic, M., Kuhn, J.H., Lefkowitz, E.J., Nibert, M.L., Orton, R., Roossinck,
660 M.J., Sabanadzovic, S., Sullivan, M.B., Suttle, C.A., Tesh, R.B., van der Vlugt, R.A.,
661 Varsani, A., & Zerbini, F.M. (2017). Consensus statement: Virus taxonomy in the age of
662 metagenomics. Nat Rev Microbiol, 15, 161-168.
663 Smith, D.R., & Keeling, P.J. (2015). Mitochondrial and plastid genome architecture:
664 Reoccurring themes, but significant differences at the extremes. Proceedings of the
665 National Academy of Sciences, 112, 10177-10184.
666 Steel, O., Kraberger, S., Sikorski, A., Young, L.M., Catchpole, R.J., Stevens, A.J.,
667 Ladley, J.J., Coray, D.S., Stainton, D., Dayarama, A., Julian, L., van Bysterveldt, K., &
668 Varsani, A. (2016). Circular replication-associated protein encoding DNA viruses
669 identified in the faecal matter of various animals in New Zealand. Infect. Genet. Evol,.
670 43, 151-164.
671 Stoye, J.P. (2006). Koala retrovirus: a genome invasion in real time. Genome Biology, 7,
672 241-241.

673 Theze, J., Leclercq, S., Moumen, B., Cordaux, R., & Gilbert, C. (2014). Remarkable
674 diversity of endogenous viruses in a crustacean genome. Genome Biology and Evolution,
675 6, 2129-2140.
676 Tu, T., Budzinska, M.A., Shackel, N.A., & Urban, S. (2017). HBV DNA Integration:
677 Molecular Mechanisms and Clinical Implications. Viruses, 9, 75.
678 Varsani, A., & Krupovic, M. (2017) Sequence-based taxonomic framework for the
679 classification of uncultured single-stranded DNA viruses of the family *Genomoviridae*.
680 Virus Evolution, 3, vew037.
681 Varsani, A., & Krupovic, M. (2018). Smacoviridae: a new family of animal-associated
682 single-stranded DNA viruses. Arch Virol., 163, 2005-2015.
683 Yu, X., Li, B., Fu, Y.P., Jiang, D.H., Ghabrial, S.A., Li, G.Q., Peng, Y.L., Xie, J.T.,
684 Cheng, J.S., Huang, J.B., & Yi, X.H. (2010). A geminivirus-related DNA mycovirus that
685 confers hypovirulence to a plant pathogenic fungus. Proceedings of the National
686 Academy of Sciences, 107, 8387-8392.
687 Zhao, L., Lavington, E., & Duffy, S. (2019a). A comprehensive genealogy of replication
688 associated protein of CRESS DNA viruses reveals a single origin of intron-containing
689 Rep. bioRxiv. BIORXIV/2019/687855
690 Zhao, L., Rosario, K., Breitbart, M., & Duffy, S. (2019b). Eukaryotic Circular Rep-
691 Encoding Single-Stranded DNA (CRESS DNA) Viruses: Ubiquitous Viruses With Small
692 Genomes and a Diverse Host Range. Advances in virus research, 103, 71-133.
693
694

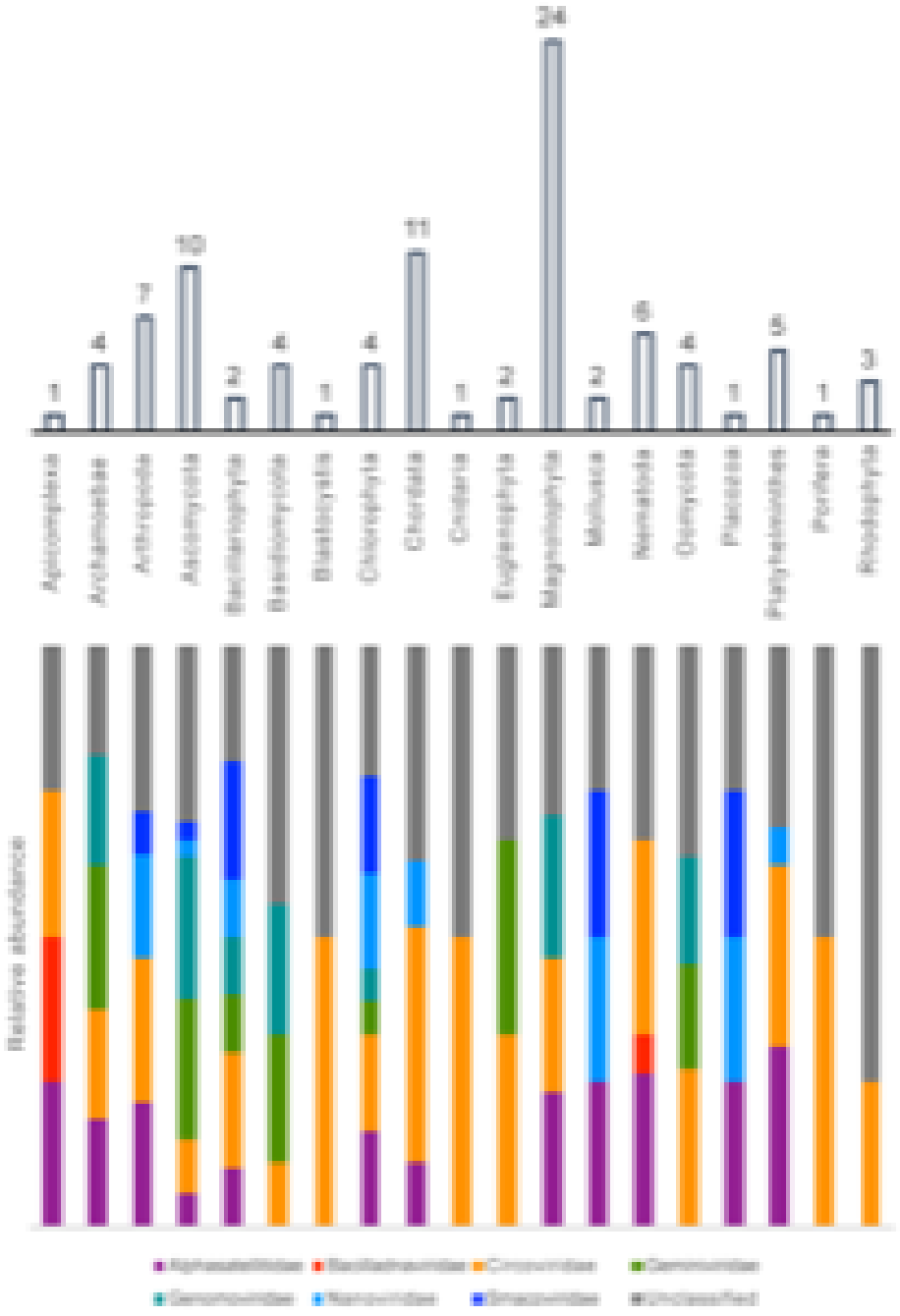
695 Table 1 Summary of results from tBLASTn using viral Rep queries.

Virus family	Queries	Initial raw hits	Unique Species	Number of Species hit more than once	Consensus sequences (cutoff 50 amino acids)
Alphasatellitidae	66	2396	50	24	189
Bacilladnaviridae	8	55	2	1	3
Circoviridae	153	15680	69	39	290
Geminiviridae	416	71708	41	23	191
Genomoviridae	67	7290	37	19	148
Nanoviridae	8	359	17	7	55
Smacoviridae	26	233	11	2	15
Unclassified	164	13623	91	46	N/A

696

697

698 Figure 1. Distribution of endogenous Reps across eukaryotic phyla. The numbers (top) represent
699 number of species containing endogenous viral sequences. The percentages (bottom) represent
700 the relative amounts of identified endogenous sequence by each viral family.



701

Figure 2. Geminivirus Reps and endogenous sequences midpoint-rooted maximum likelihood tree. Eukaryote sequences are colored in grey, geminivirus Reps are colored green, mitochondrial sequences are colored dark red, chloroplast sequences are colored bright green, plasmids from red algae are colored dark yellow. Open circles indicate SH-like support between 0.75 and 0.90, while filled circles indicate SH-like support ≥0.90. A version of this tree with protein accession numbers is in Supplementary Tree S2.

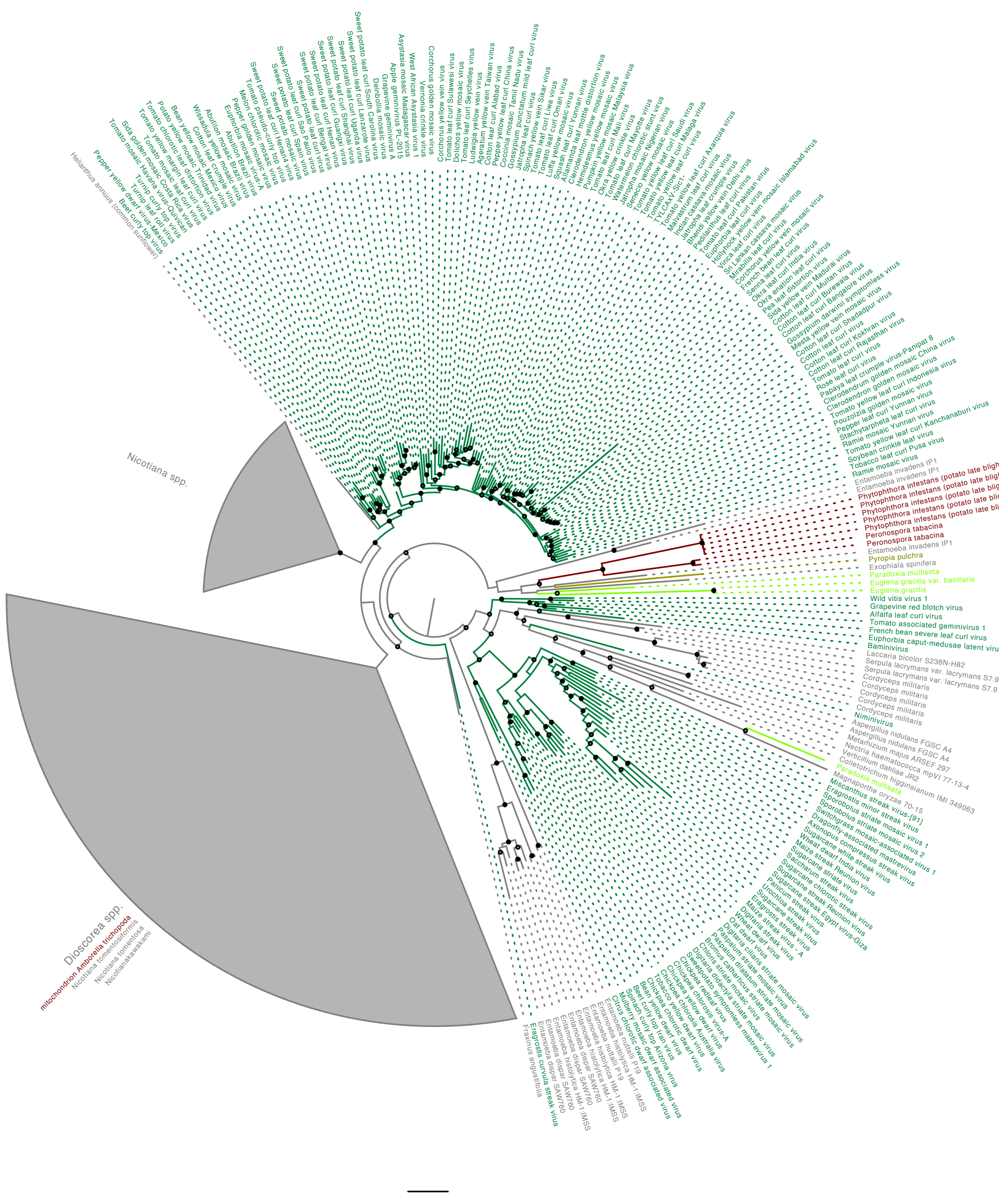


Figure 3. Circovirus Reps and endogenous sequences midpoint-rooted maximum likelihood tree. Eukaryote sequences are colored in grey, circovirus Reps are colored orange, mitochondrial sequences are colored dark red, chloroplast sequences are colored bright green, plasmids from red algae are colored dark yellow. Open circles indicate SH-like support between 0.75 and 0.90, while filled circles indicate SH-like support ≥0.90. A version of this tree with protein accession numbers is in Supplementary Tree S3.

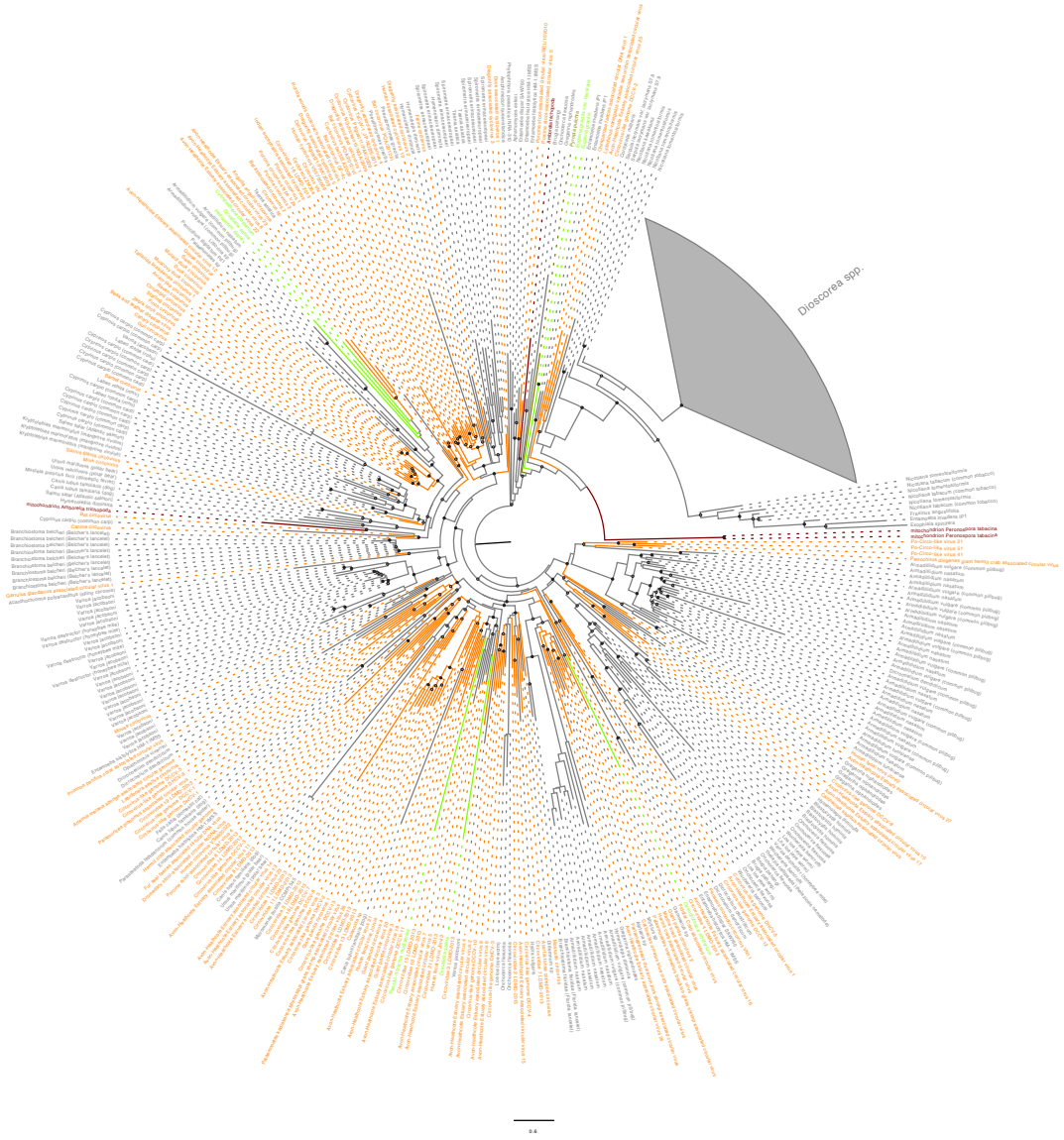
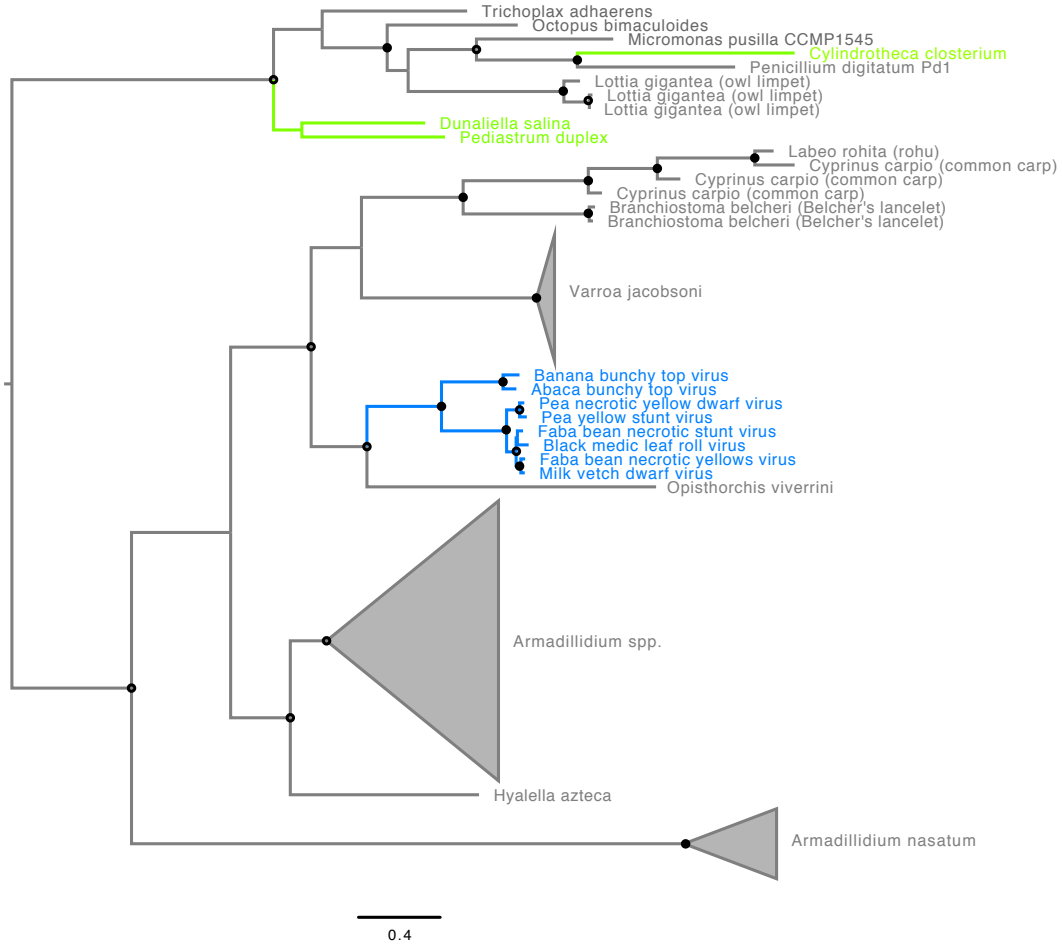
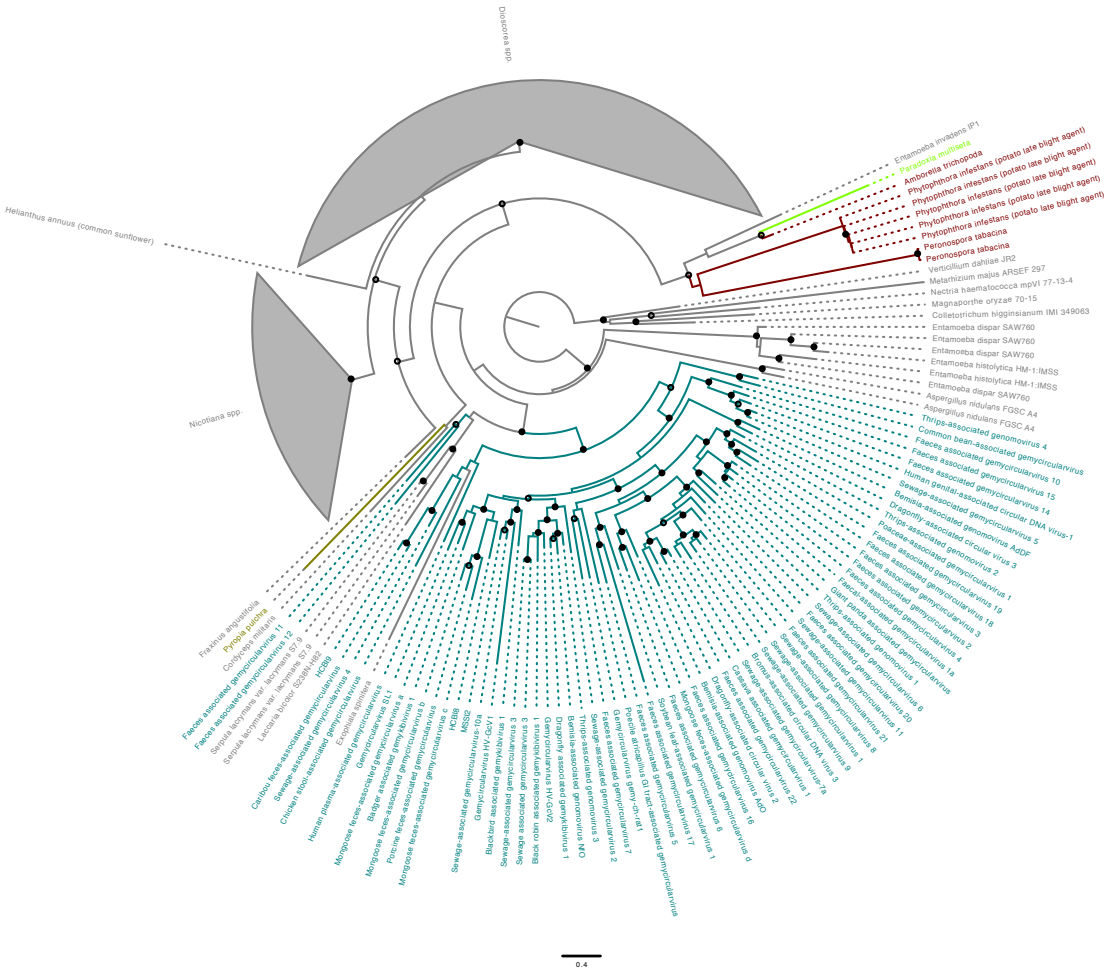


Figure 4. Nanovirus Reps and endo sequences midpoint-rooted maximum likelihood tree. Eukaryote sequences are colored in grey, nanovirus Reps are light blue, chloroplast sequences are colored bright green. Open circles indicate SH-like support between 0.75 and 0.90, while filled circles indicate SH-like support ≥0.90. A version of this tree with protein accession numbers is in Supplementary Tree S4.



722 Figure 5. Genomovirus Reps and endo sequences midpoint-rooted maximum likelihood
723 tree.Eukaryote sequences are colored in grey, genomovirus Reps are teal, mitochondrial
724 sequences are colored dark red, chloroplast sequences are colored bright green, plasmids from red
725 algae are colored dark yellow. Open circles indicate SH-like support between 698 0.75 and 0.90,
726 while filled circles indicate SH-like support ≥0.90. A version of this tree with protein accession
727 numbers is in Supplementary Tree S5.



728

Figure 6. Bacilladnavirus Reps and endo sequences midpoint-rooted maximum likelihood tree. Eukaryote sequences are colored in grey, bacilladnavirus Reps are red. Open circles indicate SH-like support between 0.75 and 0.90, while filled circles indicate SH-like support ≥0.90. A version of this tree with protein accession numbers is in Supplementary Tree S6.

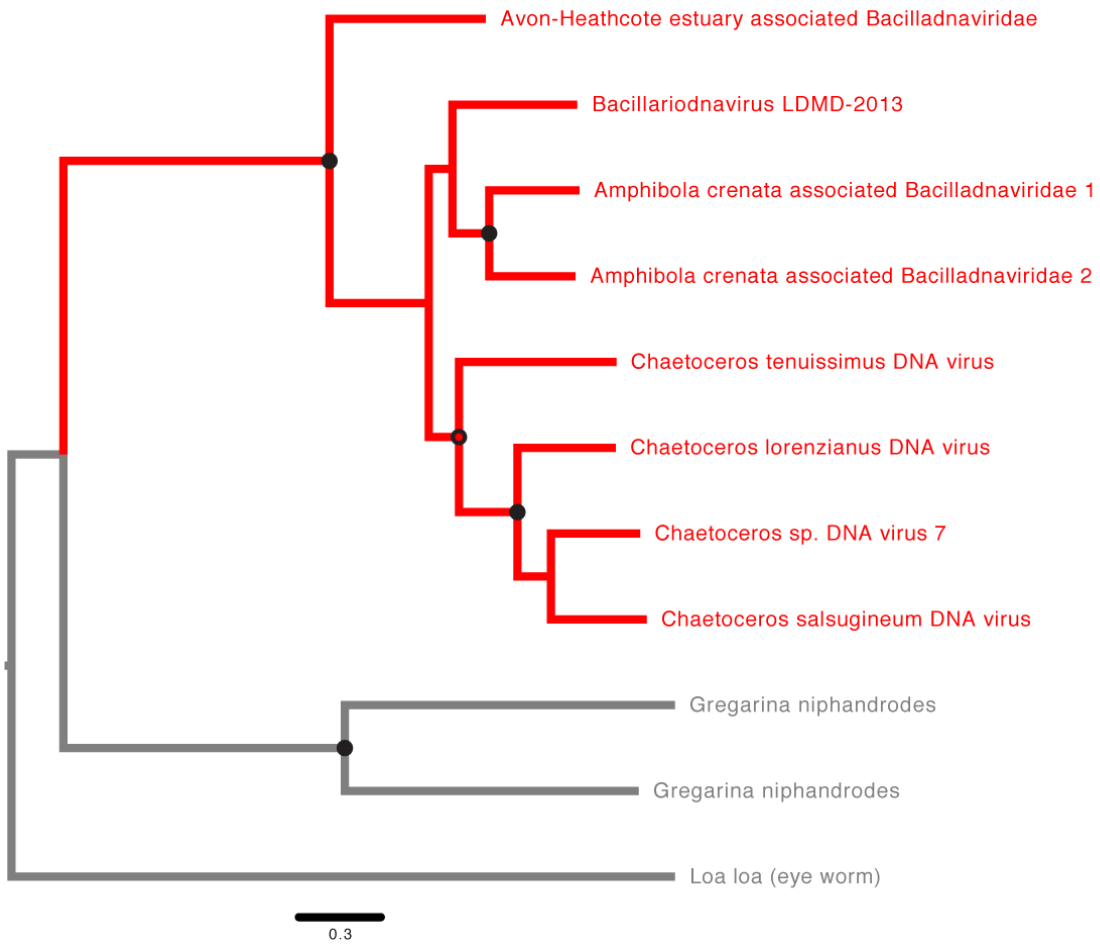


Figure 7. Smacovirus Reps and endo sequences unrooted maximum likelihood tree. Eukaryote sequences are colored in grey, smacovirus Reps are blue, chloroplast sequences are colored bright green. Open circles indicate SH-like support between 0.75 and 0.90, while filled circles indicate SH-like support ≥0.90. A version of this tree with protein accession numbers is in Supplementary Tree S7.

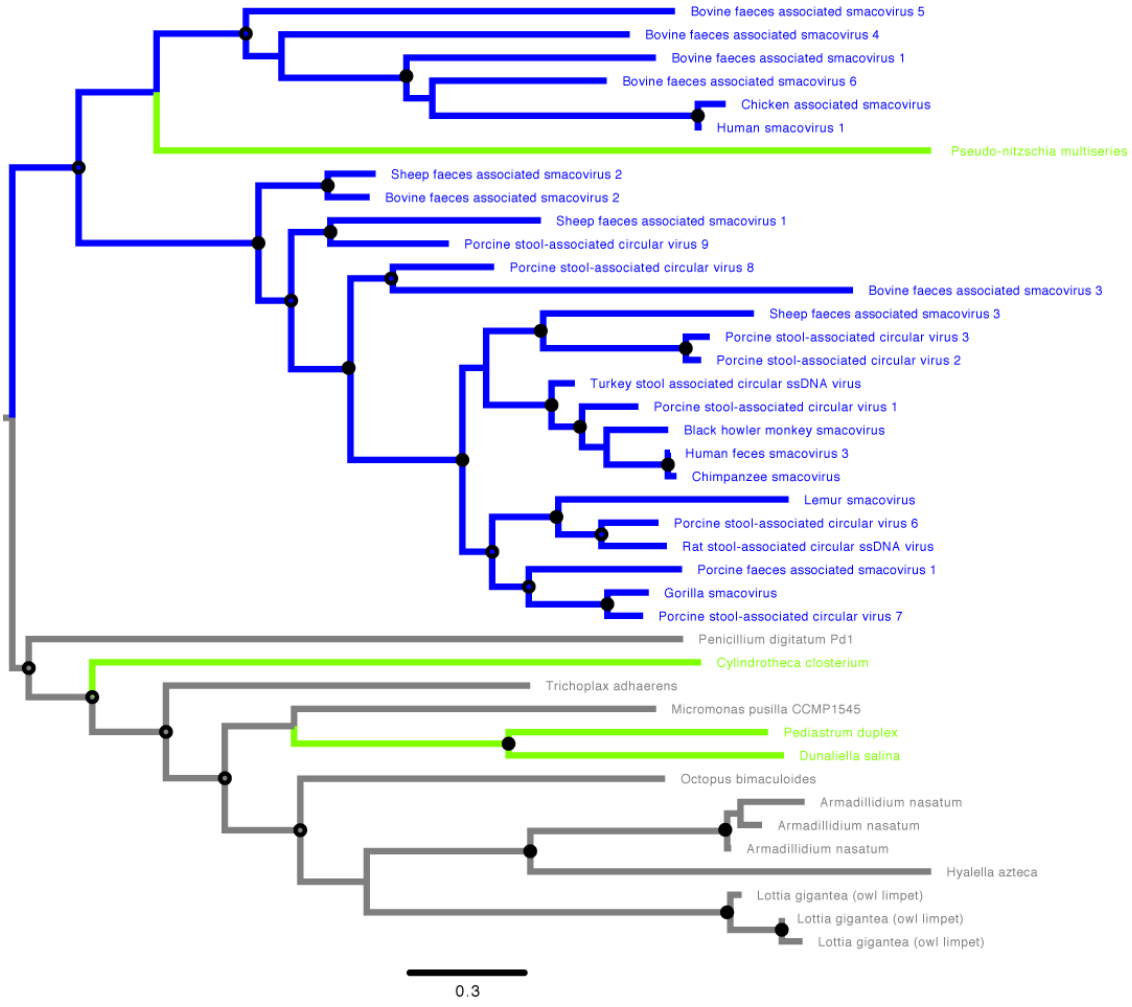


Figure 8. Alphasatellites Reps and endo sequences midpoint-rooted maximum likelihood tree. Eukaryote sequences are colored in grey, alphasatellite Reps are purple, chloroplast sequences are colored bright green. Open circles indicate SH-like support between 0.75 and 0.90, while filled circles indicate SH-like support ≥0.90. A version of this tree with protein accession numbers is in Supplementary Tree S8.