Insights into circovirus host range from the genomic fossil record.

Running title: Circovirus host range

Authors: Tristan P.W. Dennis[1*], Peter J. Flynn[2,3*], William Marciel de Souza[1,4], Joshua B. Singer[1], Corrie S. Moreau[2], Sam J. Wilson[1], Robert J. Gifford[1]

Affiliations:

[1] MRC-University of Glasgow Centre for Virus Research, Glasgow, UK

[2] Field Museum of Natural History, Department of Science and Education, Chicago, IL 60605, USA

[3]University of Chicago, Committee on Evolutionary Biology, Chicago IL 60637, USA

[4] Virology Research Center, School of Medicine of Ribeirão Preto of University of São Paulo, Ribeirão Preto, Brazil

* These authors contributed equally

Corresponding author:

Robert J. Gifford
MRC-University of Glasgow Centre for Virus Research
464 Bearsden Road
Glasgow, Scotland, UK
E-mail: robert.gifford@glasgow.ac.uk

Abstract

A diverse range of DNA sequences derived from circoviruses (family *Circoviridae*) have been identified in samples obtained from humans and domestic animals, often in association with pathological conditions. In the majority of cases, however, little or nothing is known about the natural biology of the viruses from which these sequences are derived. Endogenous circoviral elements (CVe) are DNA sequences derived from circoviruses that occur in animal genomes and provide a useful source of information about circovirus-host relationships. In this study we screened whole genome sequence data of 684 animal species to identify CVe. We identify numerous novel circovirus-related sequences in invertebrate genome assemblies, including the first examples of CVe derived from cycloviruses. We confirmed the presence of these CVe in the germline of the elongate twig ant (*Pseudomyrmex gracilis*) via PCR, thereby establishing the first concrete evidence of a host association for the *Cyclovirus* genus. We examined the evolutionary relationships between CVe and contemporary circoviruses, showing that when sequences derived from circovirus isolates and CVe are examined alone, the host species associations of circovirus clades appear relatively stable, at least at higher taxonomic levels (i.e. phylum, class, order). However, when sequences generated via metagenomic sequencing are included, associations are randomly distributed across the phylogeny, particularly in the clade corresponding to the *Cyclovirus* genus, suggesting that contamination may be an issue. Based on the robust grouping of CVe from ants and mites with cycloviruses in phylogenies, we propose that cycloviruses occur commonly in the environment as infections of arthropods, and may frequently contaminate vertebrate samples as a consequence. Our study shows how endogenous viral sequences can inform metagenomics-based virus discovery. In addition, it raises important questions about the role of cycloviruses as pathogens of humans and other vertebrate species.

Importance

Advances DNA sequencing have dramatically increased the rate at which new viruses are being identified. However, the host species associations of most virus sequences identified in metagenomic samples are difficult to determine. Our analysis indicates that viruses proposed to infect vertebrates (in some cases being linked to human disease) may in fact be restricted to arthropod hosts. The detection of these sequences in vertebrate samples may reflect their widespread presence in the environment as viruses of parasitic arthropods.

Background

Circoviruses (family *Circoviridae*) are small, non-enveloped viruses with circular, single stranded DNA (ssDNA) genomes ~1.8 to ~2.1 kilobases (kb) in length. Circovirus genomes encode two major proteins: replicase (Rep) and capsid (Cap), responsible for genome replication and particle formation respectively. Transcription is bidirectional with the *rep* gene being encoded in the forward sense, and the *cap* gene being encoded in the complementary sense (1, 2).

The family *Circoviridae* contains two recognised genera: *Circovirus* and *Cyclovirus* (1). The genus *Circovirus* includes pathogenic viruses of vertebrates, such as porcine circovirus 2 (PCV-2), which causes post-weaning multisystemic wasting syndrome in swine. The genus *Cyclovirus*, by contrast, is comprised entirely of viruses that have been identified only via sequencing, and for which host species associations are less clear. Nevertheless, cycloviruses have frequently been associated with pathogenic conditions in humans and domestic mammals (3-8). For example, cyclovirus sequences have been detected in the cerebrospinal fluid of humans suffering from acute central nervous system disease in both Vietnam and Malawi (7-9)

Progress in whole genome sequencing has revealed that sequences derived from circoviruses are present in many eukaryotic genomes (10-13). These endogenous circoviral elements (CVe) are thought to be derived from the genomes of ancient circoviruses that were – by one means or another - ancestrally integrated into the nuclear genome of germline cells (14). CVe can provide unique information about the long term coevolutionary relationships between viruses and hosts (15). For example the identification of ancient CVe in vertebrate genomes shows that viruses in the genus *Circovirus* have been co-evolving with vertebrate hosts for millions of years (13).

We recently reported the results of a study in which we systematically screened vertebrate whole genome sequence (WGS) for CVe (16). In the present study, we expanded this screen to include a total 684 animal genomes, including 308 invertebrate species. We recover sequences related to circoviruses, cycloviruses, and more divergent *rep* encoding viruses in the CRESS-DNA group. We examine the phylogenetic relationships between; (i) well-studied circovirus isolates; (ii) sequences recovered from WGS data; (iii) circovirus-related sequences recovered via metagenomic sequencing of environmental samples or animal tissues. Our analysis raises important questions about the origins of cyclovirus sequences in samples derived from humans and other mammals, and their role in causing disease in these hosts.

Materials and Methods

Sequence data

Whole genome sequence (WGS) assemblies of 684 species (**Table S1**) were downloaded from the National Center for Biotechnology Information (NCBI) website. We obtained a representative set of sequences for the genus *Circovirus*, and a non-redundant set of vertebrate CVe sequences, from an openly accessible dataset we compiled in our previous studies (16). This dataset was expanded to include a broader range of sequences in the family *Circoviridae*, including representative species in the *Cyclovirus* genus, and the more distantly related CRESS-DNA viruses (**Table S1**). We used GLUE - an open, data-centric software environment specialized in capturing and processing virus genome sequence datasets - to collate the sequences, alignments and associated data used in this investigation.

Genome screening in silico

The database-integrated genome screening (DIGS) tool (17), was used to systematically screen WGS genome assemblies for sequences homologous to circovirus Rep and Cap polypeptides. The DIGS procedure comprises two steps. In the first, a probe sequence is used to search a genome assembly file using the basic local alignment search tool (BLAST) program (18). In the second, sequences that produce statistically significant matches to the probe are extracted and classified by BLAST-based comparison to a set of virus reference genomes (see **Table S1**). Results are captured in a MySQL database. The input data, DIGS tool configuration, and database for the genome screens implemented here are included as supplementary data.

Newly identified CVe identified in this study were assigned a unique identifier (ID), following a convention we developed previously (16). The first component is of the ID the classifier ‘CVe’. The second is a composite of two distinct subcomponents separated by a period: the name of CVe group (usually derived from the host group in which the element occurs in (e.g. Carnivora), and the second is a numeric ID that uniquely identifies the insertion. Orthologous copies in different species are given the same number, but are differentiated using the third component of the ID that uniquely identifies the species from which the sequence was obtained. Unique numeric IDs were assigned to novel CVe as appropriate based on reference to ID assignments in the previously assembled dataset (16).

Alignments and phylogenetic analysis

Multiple sequence alignments were constructed using MUSCLE (19), RevTrans 2.0 (20), MACSE (21) and PAL2NAL (22). Manual inspection and adjustment of alignments was performed in Se-Al (23) and AliView (24). Phylogenies were reconstructed using maximum likelihood as implemented in RaxML (25) and the VT protein substitution model (26) as selected using ProTest (27).

Amplification and sequencing

Genomic DNA was extracted from ant tissue samples following the Moreau protocol (28) and a DNAeasy Blood & Tissue Kit (Qiagen). PCR amplification of CVe-*Pseudomyrmex* was performed using two sets of primer pairs designed with Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/), each comprising one primer anchored in the CVe sequence, and another anchored in the genomic flanking sequence (**Table S2)**. PCR primers were tested using illustra PuReTaq Ready-To-Go PCR Beads (GE Lifesciences). A temperature gradient PCR was performed to assess the optimum annealing temperature for the specific primer pairs. PCR was then performed using the genomic DNA ant extractions. The PCR conditions for this run were: an initial denature stage of 5 minutes at 95°C, 30 cycles of 30 seconds denaturing at 95°C, 30 seconds annealing at 49.7°C for Primer Pair 1 and 62°C for Primer Pair 2, and an extension at 72°C for 1 minute, then after 30 cycles a final extension at 72°C for 5 minutes. Each run included a negative control. Amplification products (800-1000bp) for each PCR reaction were excised and run on agarose gels.

Results

Identification of CVe in animal genomes

We screened WGS data of 684 animal species (**Table S3**) *in silico* to identify sequences related to circoviruses. We identified 300 circovirus-related sequences in total, 76 of which have not been reported previously (**Table 1, Table S4**). To investigate the novel sequences identified in our screen, each was virtually translated and incorporated into a multiple sequence alignment that included a representative set of previously reported circoviruses and CVe (**Table S1**). Incorporation of CVe sequences into alignment provided a basis for determining the structure of individual CVe loci, and for investigating the phylogenetic relationships between CVe and circoviruses (**Figure 1, Figure S1**). All of the newly identified sequences were derived from *rep* – no novel sequences derived from circovirus *cap* genes were detected*.*

We identified two novel CVe derived from viruses in the genus *Circovirus* in fish genomes (**Table 1**). One of these, identified in the tomato clownfish (*Amphiprion frenatus*), appeared to an ortholog of a CVe locus previously identified in perciform fish. The other, identified in a mormyrid fish, was clearly related to other fish CVe, but as it comprised a relatively short fragment of the *rep* gene, its more precise phylogenetic relationship to these CVe could not be determined with confidence.

We identified 98 circovirus-derived sequences in invertebrate genome assemblies, 78 of which have not been reported previously (**Table 1**, **Table S4**). Of these, 72 exhibited coding potential, and six were also predicted to express messenger RNA (mRNA) (**Table S4**). Some occurred on short contigs, and could potentially have been derived from contaminating virus. However, we found that in many cases, at least one of the circovirus-related sequences was incorporated into a contig large enough to contain an entire circovirus genome. If all the sequences were closely related, we assumed all derived from CVe. On this basis, we estimate that 86 of the 91 sequences we identified in invertebrate genomes are likely to be derived from CVe. However, in almost all cases, the sequences flanking regions of *rep* homology disclosed no unambiguous similarity to previously sequenced genomes, and there consequently remains a degree of uncertainty regarding their provenance. Potentially, they could represent CVe derived from contaminating DNA of another species.

Maximum likelihood (ML) phylogenies were reconstructed using an alignment of Rep proteins, and disclosed two robustly supported, monophyletic clades corresponding to the recently defined *Circovirus* and *Cyclovirus* genera (1). In line with our previous investigations (16), we found that all Rep-related sequences from vertebrate WGS grouped within the *Circovirus* clade, with the exception of a highly divergent sequence recovered from the WGS of the inshore hagfish (*Eptatretus burgeri*). All sequences derived from invertebrate WGS grouped within the *Cyclovirus* clade), or with divergent CRESS-DNA viruses (e.g. Avon-Heathcote Estuary associated circular virus 24) (data not shown). CRESS-DNA virus-like sequences from distinct species tended to emerge on relatively long branches, and bootstrap support for branching patterns in this region of the phylogeny were generally quite low. The low resolution in this part of the phylogeny likely reflects the lack of adequate sampling of viruses from invertebrate species.

Some sequences from invertebrate WGS were observed to cluster within the cyclovirus clade in phylogenies. Most of these sequences occur within relatively short contigs, whereas others occur on contigs that are easily large enough contain genomic flanking sequences, but similarity to previously sequenced arthropod genomes could be detected. They include sequences identified in WGS data of two parasitic mite species: *Varroa destructor* and *Tropilaelaps mercedesae*. Those identified in *V. destructor* have previously been reported as CVe (12). A DNA sequence homologous to the cyclovirus *rep* gene was also identified in the genome of the elongate twig ant (*Pseudomyrmex gracilis*). This sequence – hereafter referred to as CVe-*Pseudomyrmex* - was no more distantly related to contemporary cycloviruses than many of them are to one another, including some that are associated with vertebrates (at least superficially) **(Figure 1**). Because this seemed a little surprising, we sought to confirm the presence of CVe-*Pseudomyrmex* in the twig ant germline.

PCR confirmation of CVe-Pseudomyrmex

We obtained genomic DNA from four species of ant within the *Pseudomyrmex* genus (*P. gracilis*, *P. elongatus*, *P. spinicola*, and *P. oculatus)*, including three distinct populations of *P. gracilis*, We then used polymerase chain reaction (PCR) to amplify a region encompassing part of the CVe, and part of the genomic flanking sequence. We obtained an amplicon of the expected size in all three DNA samples of *P.gracilis*, all other samples were negative (**Figure 2**). The fact that we did not detect the CVe-*Pseudomyrmex* sequence in other members of the genus suggests it was incorporated into the *P. gracilis* germline after this species diverged from *P. elongatus, P. spinicola, and P. oculatus* in the mid-Miocene (29). However, we cannot rule out that the failure to obtain an amplicon in these species is due to sequence divergence in the regions targeted by PCR primers.

Mapping host associations in the Circovirus and Cyclovirus genera

In phylogenies based on Rep, clades corresponding to the *Circovirus* and *Cyclovirus* genera contained a mixture of; (i) CVe from WGS assemblies; (ii) sequences obtained from virus isolates; (iii) sequences obtained from metagenomic samples (**Figure 1, Table S1**). In the clade representing the *Circovirus* genus, associations at the level of class appear relatively stable. For example, beak and feather disease virus (BFDV) groups robustly with a CVe that entered the germline of passeriforme birds, while barbel circovirus groups (BarbCV) groups robustly with CVe from the genome of the golden line barbell, in a well-supported clade containing numerous CVe from ray-finned fish. The only sequence that superficially seems to contradict this pattern is 'chimpanzee circovirus', which groups robustly within a cluster of avian viruses. However, the name of this sequence is misleading. It was recovered from chimpanzee feces, and in fact the possibility that it reflected environmental contamination with an avian virus was noted at the time it was reported (30).

The *Cyclovirus* clade contains three well-supported sublineages, here labelled cyclovirus 1-3 (see **Figure 1**). In the *Cyclovirus* genus as a whole, the only confirmed host associations are with arthropods, via CVe, as reported above. Many of the cyclovirus sequences that have been identified via metagenomic sequencing are associated with arthropod species, such as dragonflies (31). However, others are associated with vertebrates, having names such as 'bat cyclovirus'. Cyclovirus 1 is exclusively comprised of viruses from vertebrate samples. In cyclovirus groups 2 and 3, however, sequences from vertebrate and invertebrate samples are extensively intermingled (**Figure 1**), and clade structure does not reflect these host associations in any obvious way. Sequences from each host group appear to be dispersed randomly, and the branch lengths separating vertebrate from invertebrate viruses (and CVe) are relatively short in many cases.

Discussion

In this study we expand the catalogue of CVe that have been identified in animal genomes. In addition, we identify a diverse range of circovirus-related sequences that may represent either novel CVe or novel circoviruses.

Most of the novel circovirus sequences described here were identified in invertebrate genome assemblies. Many of these are highly divergent, and are likely derived from uncharacterised CRESS-DNA virus lineages that infect invertebrate species. Interestingly, a large proportion lacked frameshifting mutations or in-frame stop codons, indicating that they are evolving under purifying selection, and are thus likely to represent relatively recently integrated CVe, assuming that they do indeed represent CVe (as opposed to contaminating virus), which we could not ascertain in every case. However, we note that CVe-*Pseudomyrmex*, which we show here to be a *bona fide* endogenous element (**Figure 2**), encodes an intact *rep* gene product, and is predicted by genome annotation software to express mRNA (**Table S4**). The occurrence of an apparently fixed, intact, and expressed circovirus *rep* gene in an ant genome provides further evidence that these genes have been co-opted or exapted by host species for as yet unknown functions.

The aim of our study was to examine the host associations of circovirus sequences in the context of their evolutionary relationships. Importantly, we distinguished sequences for which the host associations are well established (i.e. CVe and viruses that have been investigated using methods besides only sequencing), and sequences recovered from metagenomic samples. Prior to this study, the only host associations that had been robustly demonstrated were within the genus *Circovirus*. Circoviruses have been isolated from vertebrates, and in phylogenies based on Rep proteins, these isolates group together with vertebrate CVe in a well-supported *Circovirus* clade. Within this clade, host associations appear quite stable, with ancient CVe from particular host orders or classes seen grouping together with contemporary viruses. The only sequence that seems to contradict this pattern is derived from a stool sample and likely reflects environmental contamination, as observed in when it was first reported (30).

We do, however, see evidence for potential inter-class transmission within one *Circovirus* sub-lineage that contains sequences obtained from birds (waterfowl) and mammals (mink, bats, dogs and pigs) as well as CVe from reptile genomes (see **Figure 1**). Within this clade, sequences from viruses of mink are robustly separated from those obtained from porcine and canine viruses by a CVe that was incorporated into the serpentine germline >72 million years ago, based on its presence as an orthologous insertion in multiple species (16). However, it should be noted that while the clustering patterns observed in this clade do suggest potential transmission of circoviruses between vertebrate classes, they also indicate that such events have occurred relatively infrequently during evolution. Furthermore, they can also be accounted for by alternative evolutionary scenarios that do not entail inter-class transmission.

Interestingly, the limited evidence available regarding the zoonotic potential of circoviruses suggests they lack the capacity to be transmitted between relatively distantly related hosts (i.e. hosts in distinct classes or orders). During the 1990s and early 2000s, porcine circovirus 1 (PCV-1) was inadvertently introduced into batches of live attenuated rotavirus vaccine as an adventitious agent. These vaccines were administered to millions of people (32), yet PCV-1 is not thought to have infected any humans as a result, indicating that powerful barriers to cross-species transmission are probably in effect.

If cross-species transmission of circoviruses between distinct mammalian orders does not occur readily, then transmission between arthropod and vertebrate hosts appears extremely unlikely. But if we take the reported host species associations of sequences in the *Cyclovirus* clade at face value, we might conclude it occurs frequently, particularly within one subclade, here referred to as ‘cyclovirus 3’ (**Figure 1**). Notably, CVe-*Pseudomyrmex* groups robustly within this clade and discloses relatively close relationships with other clade members.

Since all other taxa within the *Cyclovirus* clade have been recovered via metagenomic screening, CVe-*Pseudomyrmex* provides the first unambiguous evidence of a host-association for cycloviruses, establishing with a high degree of confidence that they do infect arthropods. Furthermore, since the only proven associations for cycloviruses so far are with arthropods, contamination of vertebrate samples with viruses derived from arthropods is perhaps the most parsimonious explanation for the host associations observed here. Contamination from arthropod sources can presumably occur fairly easily given their ubiquity – for example, diverse varieties of mite live on animal skin and in house dust. Intriguingly with respect to this, we identified putative cyclovirus CVe in the genomes of two distinct mite species (**Figure 1**, **Table S4**). Since there is always a risk of being misled by contamination when identifying viruses via sequencing-based approaches, we propose that host associations of circoviruses identified via sequencing should be viewed with caution where they are found to strongly contradict established host associations within well-defined clades, particularly at higher taxonomic levels (e.g. phylum, class, order).

Whereas the weight of evidence may favour cyclovirus clades 2 and 3 being exclusively arthropod viruses that frequently contaminate vertebrate samples, the status of cyclovirus clade 1 is more equivocal. This lineage, which is positioned basally within the *Cyclovirus* clade as a whole, is comprised exclusively of sequences obtained from mammalian samples, and includes cycloviruses proposed to cause disease in humans (cyclovirus VN and human cyclovirus VS5700009). Conceivably, these sequences could represent a mammal or vertebrate-specific lineage of cycloviruses that is distinct from arthropod-infecting lineages. Notably, however false-positive detection of human cyclovirus VS5700009 has been reported (33).

Virus sequences recovered from metagenomic samples can be investigated by examining their phylogenetic relationships to other viruses for which host associations have been established. The work performed here demonstrates the value of endogenous virus sequences in this process. This approach can be generalized to inform metagenomics-based virus discovery and diversity mapping efforts for any virus group that has generated endogenous sequences.

FIGURES & LEGENDS

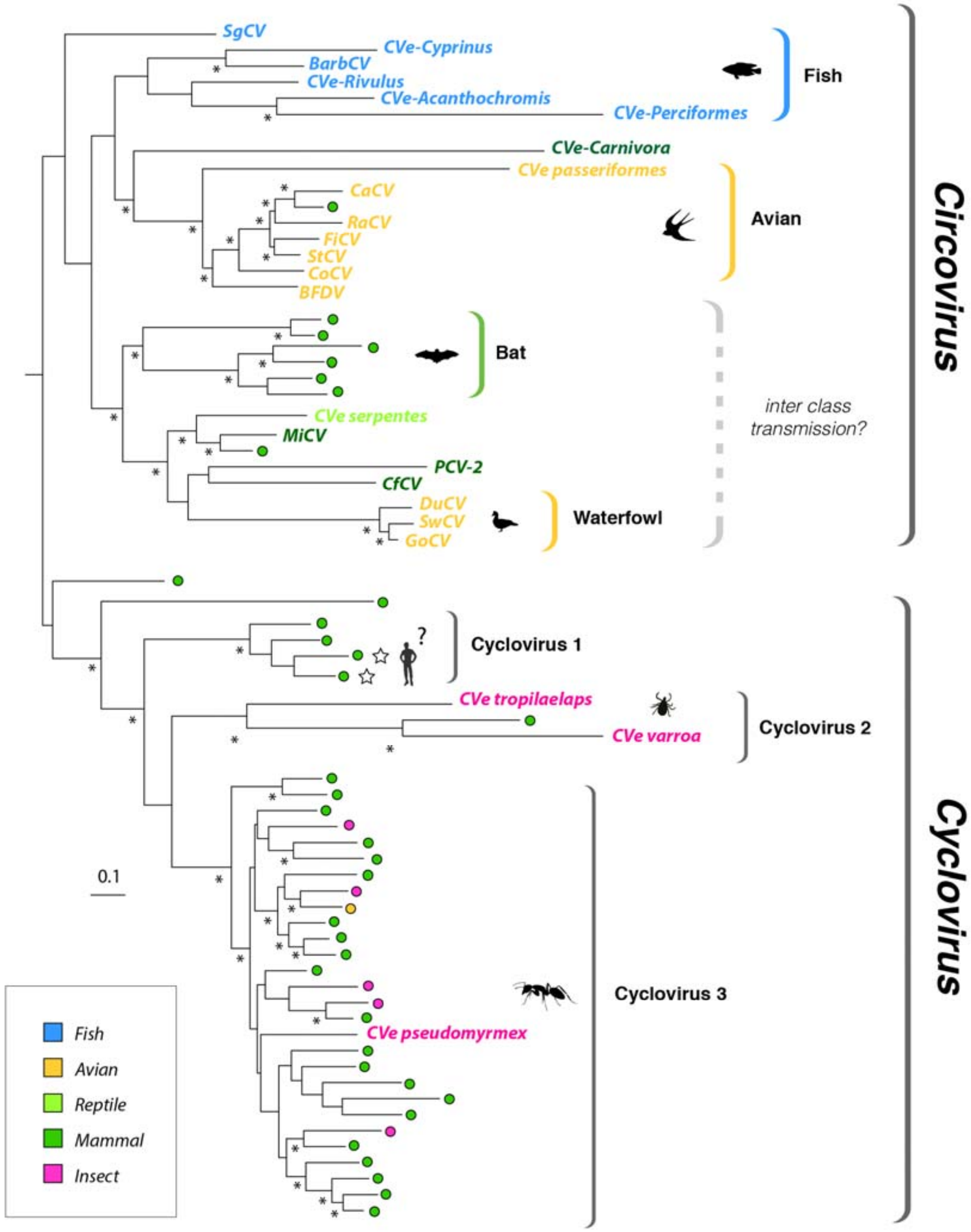

Figure 1. Phylogeny of exogenous and endogenous circovirus Rep sequences.

Maximum-likelihood phylogeny reconstructed from an alignment of Rep-associated protein sequences. Asterisks indicate nodes with >70% bootstrap support. Scale bar indicates evolutionary distance in substitutions per site. Sequences derived from metagenomic samples are indicated by colored circles. Taxa names are shown for sequences derived from viruses and CVe, and are coloured to indicate associations with host species groups, as shown in the key. Sequences derived from metagenomic sampling are indicated by circles coloured according to indicate sample associations with host species group. Stars indicate viral taxa that have been linked to human disease. See **Table S1** and **Figure S1** for accession numbers and details of taxa shown here.

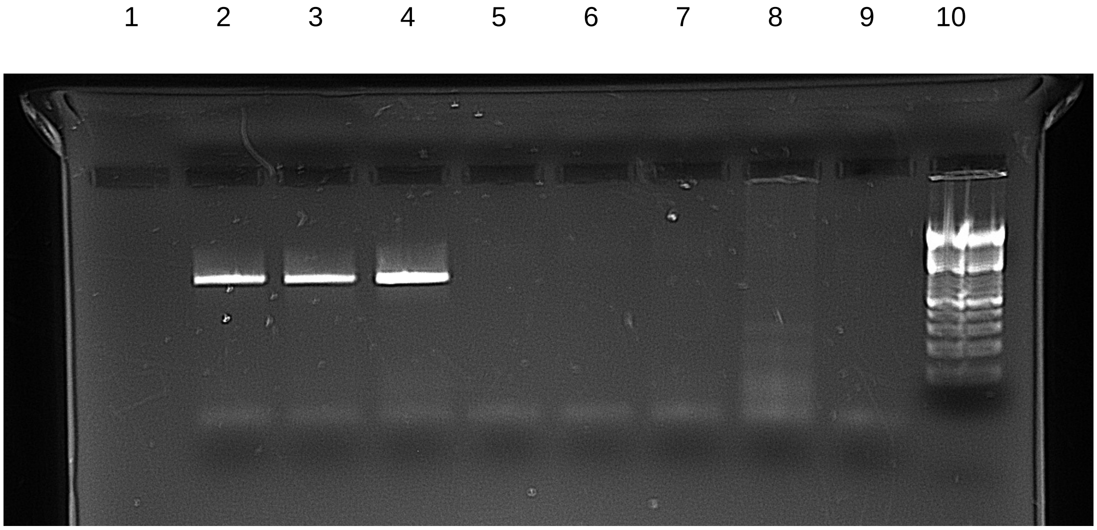

Figure 2. **PCR confirmation of CVe-*Pseudomyrmex* presence in three populations of *Pseudomyrmex gracilis*.** Columns: (1) negative control; (2) *Pseudomyrmex gracilis*, from the Florida Keys; (3) *P. gracilis*, from mainland Florida, USA; (4) *P. gracilis*, from Texas, USA; (5) *P. elongatus*, from the Florida Keys, USA; (6) *P. spinicola*, from Guanacaste Province, Costa Rica; (7) *P. oculatus*, from Cusco, Peru ; (8) *Cephalotes atratus,* from Cusco, Peru; (9) negative control; (10) ladder.

Table 1. Novel CVe identified in this study.

Common name	Scientific name	Class	Order	# Seq.
Circovirus				
Tomato clownfish	*Amphiprion frenatus*	**Vertebrata**	Perciformes	1
Elephantfish	*Paramormyrops kingsleyae*	**Vertebrata**	Osteoglossiformes	1
Cyclovirus				
Asian bee mite	*Tropilaelaps mercedesae*	**Arthropoda**	Arachnida	7
Elongate twig ant	*Pseudomyrmex gracilis*	**Arthropoda**	Insecta	1
CRESS				
Myxosporean parasite	*Thelohanellus kitauei*	**Cnidaria**	Myxosporea	1
Philippine horse mussel	*Modiolus philippinarum*	**Mollusca**	Bivalvia	4
Mediterranean mussel	*Mytilus galloprovincialis*	**Mollusca**	Bivalvia	4
Freshwater snail	*Biomphalaria glabrata*	**Mollusca**	Gastropoda	1
Tribble’s cone	*Conus tribblei*	**Mollusca**	Gastropoda	3
Western predatory mite	*Galendromus occidentalis*	**Arthropoda**	Arachnida	1
Phytoseiid predatory mite	*Metaseiulus occidentalis*	**Arthropoda**	Arachnida	1
Brown recluse spider	*Loxosceles reclusa*	**Arthropoda**	Arachnida	19
Scarab beetle	*Oryctes borbonicus*	**Arthropoda**	Insecta	1
Drifting brine fly	*Ephydra gracilis*	**Arthropoda**	Insecta	10
Alkali fly	*Ephydra hians*	**Arthropoda**	Insecta	8
Amphipod crustacean	*Parhyale hawaiensis*	**Arthropoda**	Malacostraca	3
Sea louse	*Caligus rogercresseyi*	**Arthropoda**	Maxillopoda	4
Tadpole shrimp	*Triops cancriformis*	**Arthropoda**	Branchiopoda	1

ACKNOWLEDGEMENTS

TPWD, JBS, SJW, and RJG were supported by a grant from the UK Medical Research Council (No. MC_UU_12014/10). WMS was supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo, Brazil (Scholarships No. 12/24150-9; 15/05778-5; 17/13981-0). SM and PJF were supported by a grant from the National Science Foundation (NSF DEB 1442316).

References

1. **Rosario K, Breitbart M, Harrach B, Segales J, Delwart E, Biagini P, Varsani A.** 2017. Revisiting the taxonomy of the family Circoviridae: establishment of the genus Cyclovirus and removal of the genus Gyrovirus. Arch Virol **162:**1447-1463.
2. **Breitbart M, Delwart E, Rosario K, Segales J, Varsani A, Ictv Report C.** 2017. ICTV Virus Taxonomy Profile: Circoviridae. J Gen Virol **98:**1997-1998.
3. **Phan TG, Luchsinger V, Avendano LF, Deng X, Delwart E.** 2014. Cyclovirus in nasopharyngeal aspirates of Chilean children with respiratory infections. J Gen Virol **95:**922-927.
4. **Phan TG, Mori D, Deng X, Rajindrajith S, Ranawaka U, Fan Ng TF, Bucardo-Rivera F, Orlandi P, Ahmed K, Delwart E.** 2015. Small circular single stranded DNA viral genomes in unexplained cases of human encephalitis, diarrhea, and in untreated sewage. Virology **482:**98-104.
5. **Phan TG, Giannitti F, Rossow S, Marthaler D, Knutson TP, Li L, Deng X, Resende T, Vannucci F, Delwart E.** 2016. Detection of a novel circovirus PCV3 in pigs with cardiac and multi-systemic inflammation. Virol J **13:**184.
6. **Matthews PC, Sharp C, Simmonds P, Klenerman P.** 2017. Human parvovirus 4 'PARV4' remains elusive despite a decade of study. F1000Res **6:**82.
7. **Tan le V, van Doorn HR, Nghia HD, Chau TT, Tu le TP, de Vries M, Canuti M, Deijs M, Jebbink MF, Baker S, Bryant JE, Tham NT, NT BK, Boni MF, Loi TQ, Phuong le T, Verhoeven JT, Crusat M, Jeeninga RE, Schultsz C, Chau NV, Hien TT, van der Hoek L, Farrar J, de Jong MD.** 2013. Identification of a new cyclovirus in cerebrospinal fluid of patients with acute central nervous system infections. MBio **4:**e00231-00213.
8. **Smits SL, Zijlstra EE, van Hellemond JJ, Schapendonk CM, Bodewes R, Schurch AC, Haagmans BL, Osterhaus AD.** 2013. Novel cyclovirus in human cerebrospinal fluid, Malawi, 2010-2011. Emerg Infect Dis **19**.
9. **Macera L, Focosi D, Vatteroni ML, Manzin A, Antonelli G, Pistello M, Maggi F.** 2016. Cyclovirus Vietnam DNA in immunodeficient patients. J Clin Virol **81:**12-15.
10. **Gibbs MJ, Smeianov VV, Steele JL, Upcroft P, Efimov BA.** 2006. Two families of rep-like genes that probably originated by interspecies recombination are represented in viral, plasmid, bacterial, and parasitic protozoan genomes. Mol Biol Evol **23:**1097-1100.
11. **Kapoor A, Simmonds P, Lipkin WI.** 2010. Discovery and characterization of mammalian endogenous parvoviruses. J Virol **84:**12628-12635.
12. **Liu H, Fu Y, Li B, Yu X, Xie J, Cheng J, Ghabrial SA, Li G, Yi X, Jiang D.** 2011. Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. BMC Evol Biol **11:**276.
13. **Katzourakis A, Gifford RJ.** 2010. Endogenous viral elements in animal genomes. PLoS Genet **6:**e1001191.
14. **Holmes EC.** 2011. The evolution of endogenous viral elements. Cell Host Microbe **10:**368-377.
15. **Feschotte C, Gilbert C.** 2012. Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet **13:**283-296.

16. **Dennis TPW, Souza WM, Marsile-Medun S, Singer JB, Wilson SJ, Gifford RJ.** 2018. The evolution, distribution and diversity of endogenous circoviral elements. bioRxiv doi:10.1101/207399.
17. **Gifford RJ.** 2015. The database-integrated genome screening (DIGS) tool, https://giffordlabcvr.github.io/DIGS-tool/.
18. **Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DL.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nuc Acids Res **25:**3389-3402.
19. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32:**1792-1797.
20. **Wernersson R, Pedersen AG.** 2003. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res **31:**3537-3539.
21. **Ranwez V, Harispe S, Delsuc F, Douzery EJ.** 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PLoS One **6:**e22594.
22. **Suyama M, Torrents D, Bork P.** 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res **34:**W609-612.
23. **Rambaut A.** 2002. SE-AL Sequence Alignment Editor, vv2 0a11. University of Oxford, Oxford, UK.
24. **Larsson A.** 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics **30:**3276-3278.
25. **Stamatakis A.** 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22:**2688-2690.
26. **Muller T, Vingron M.** 2000. Modeling amino acid replacement. J Comput Biol **7:**761-776.
27. **Darriba D, Taboada GL, Doallo R, Posada D.** 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics **27:**1164-1165.
28. **Moreau CS.** 2014. A practical guide to DNA extraction , PCR , and gene-based DNA sequencing in insects. Halteres **5:**32–42.
29. **Gomez-Acevedo S, Rico-Arce L, Delgado-Salinas A, Magallon S, Eguiarte LE.** 2010. Neotropical mutualism between Acacia and Pseudomyrmex: phylogeny and divergence times. Mol Phylogenet Evol **56:**393-408.
30. **Li L, Kapoor A, Slikas B, Bamidele OS, Wang C, Shaukat S, Masroor MA, Wilson ML, Ndjango JB, Peeters M, Gross-Camp ND, Muller MN, Hahn BH, Wolfe ND, Triki H, Bartkus J, Zaidi SZ, Delwart E.** 2010. Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces. J Virol **84:**1674-1682.
31. **Rosario K, Dayaram A, Marinov M, Ware J, Kraberger S, Stainton D, Breitbart M, Varsani A.** 2012. Diverse circular ssDNA viruses discovered in dragonflies (Odonata: Epiprocta). J Gen Virol **93:**2668-2681.
32. **Victoria JG, Wang C, Jones MS, Jaing C, McLoughlin K, Gardner S, Delwart EL.** 2010. Viral nucleic acids in live-attenuated vaccines: detection of minority variants and an adventitious virus. J Virol **84:**6033-6040.
33. **Chan MC, Kwok SW, Chan PK.** 2015. False-positive PCR detection of cyclovirus Malawi strain VS5700009 in human cerebrospinal fluid. J Clin Virol **68:**76-78.