Virus Research 262 (2019) 15–23

Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

The evolution, distribution and diversity of endogenous circoviral elements in vertebrate genomes

Tristan P.W. Dennis[a,1], William Marciel de Souza[a,b,1], Soledad Marsile-Medun[a,c], Joshua B. Singer[a], Sam J. Wilson[a], Robert J. Gifford[a,*]

[a] MRC-University of Glasgow Centre for Virus Research, 464 Bearsden Road, Glasgow, UK
[b] Virology Research Center, School of Medicine of Ribeirão Preto of University of São Paulo, Ribeirão Preto, Brazil
[c] Agrocampus Ouest, 65 Rue de Saint-Brieuc, 35000, Rennes, France

ARTICLE INFO

Keywords:
Circovirus
Evolution
Endogenous
Paleovirology
Taxonomy
Phylogeny

ABSTRACT

Circoviruses (family *Circoviridae*) are small, non-enveloped viruses that have short, single-stranded DNA genomes. Circovirus sequences are frequently recovered in metagenomic investigations, indicating that these viruses are widespread, yet they remain relatively poorly understood. Endogenous circoviral elements (CVe) are DNA sequences derived from circoviruses that occur in vertebrate genomes. CVe are a useful source of information about the biology and evolution of circoviruses. In this study, we screened 362 vertebrate genome assemblies *in silico* to generate a catalog of CVe loci. We identified a total of 179 CVe sequences, most of which have not been reported previously. We show that these CVe loci reflect at least 19 distinct germline integration events. We determine the structure of CVe loci, identifying some that show evidence of potential functionalization. We also identify orthologous copies of CVe in snakes, fish, birds, and mammals, allowing us to add new calibrations to the timeline of circovirus evolution. Finally, we observed that some ancient CVe group robustly with contemporary circoviruses in phylogenies, with all sequences within these groups being derived from the same host class or order, implying a hitherto underappreciated stability in circovirus-host relationships. The openly available dataset constructed in this investigation provides new insights into circovirus evolution, and can be used to facilitate further studies of circoviruses and CVe.

1. Introduction

Circoviruses (family *Circoviridae*) are small, non-enveloped viruses with single-stranded DNA (ssDNA) genomes. Circovirus genomes are typically ~2 kilobases (kb) in length and contain only two open reading frames (ORFs): one encoding a non-structural, replication-associated protein (Rep), and a second encoding the viral capsid (Cap). The family contains two genera: *Circovirus* and *Cyclovirus,* and many of the ssDNA viruses that have been identified in recent years fall into one of these two genera (Rosario et al., 2017). However, very little is known about the novel viruses that have been identified using these approaches. Only a handful of circoviruses have been investigated at a level beyond sequencing: porcine circoviruses 1 and 2 (PCV-1 and PCV-2), which infect swine, and 'beak and feather disease virus' (BFDV), which infects various avian species (Amery-Gale et al., 2017).

Endogenous circoviruses (CVe) provide an unconventional but useful source of information about circovirus distribution, diversity and evolution. These sequences are derived from the genomes of circoviruses that circulated millions of years ago, and became integrated into the host germline (Holmes, 2011; Feschotte and Gilbert, 2012). Relatively robust minimum age estimates can be obtained for CVe via the identification of orthologous copies in distinct host lineages. On this basis, we now know that the association between circoviruses and vertebrates extends back millions of years before the present day (Katzourakis and Gifford, 2010; Belyi et al., 2010).

In this study, we screened vertebrate genomes *in silico* to generate a comprehensive catalog of CVe. We used these data to: (i) extract information about the long-term evolution of circoviruses; (ii) generate an openly accessible data resource that can facilitate the further investigation of CVe and circoviruses.

Abbreviations: CVe, endogenous circoviral element; ORF, open reading frame; Cap, capsid; Rep, replication associated protein; ssDNA, single-stranded DNA

[*] Corresponding author.
E-mail address: robert.gifford@glasgow.ac.uk (R.J. Gifford).

[1] These authors contributed equally.

https://doi.org/10.1016/j.virusres.2018.03.014
Received 15 January 2018; Received in revised form 16 March 2018; Accepted 25 March 2018
Available online 27 March 2018

2. Material & methods

2.1. Identification and analysis of CVe sequences

We used similarity searches to systematically screen genome assemblies of 362 chordate species (Table S1) for sequences homologous to circovirus proteins. Vertebrate genome assemblies and circovirus reference genomes were obtained from the NCBI genomes resource. Screening *in silico* was performed using the database-integrated genome-screening tool (DIGS). The DIGS procedure used to identify CVe comprises two steps. In the first, a virtually translated circovirus probe sequence (e.g. a Cap or Rep protein sequence) is used to search a particular genome assembly file using the basic local alignment search tool (tBLASTn) program (Altschul et al., 1997). In the second, sequences that produce statistically significant matches to the probe are extracted and classified by tBLASTn-based comparison to a set of virus reference genomes (see Table S2). Results are captured in a MySQL database.

We inferred the ancestral ORFs of CVe (and the number of stop codons and frameshifts interrupting these ORFs) via a combination of automated alignment and manual adjustment. Multiple sequence alignments were constructed using MUSCLE (Edgar, 2004) and PAL2NAL (Suyama et al., 2006). Manual inspection and adjustment of alignments was performed in Se-Al (Rambaut, 2002). Phylogenies were constructed using maximum likelihood as implemented in RaxML (Jarvis et al., 2014), and the VT protein substitution model (Muller and Vingron, 2000) as selected using ProTest (Darriba et al., 2011).

2.2. Construction of CVe sequence data resource

We used GLUE - an open, data-centric software environment specialized in capturing and processing virus genome sequence datasets (Singer et al., 2018) – to collate the sequences, alignments and associated data used in this investigation. The aim was to create a standardized data CVe resource that would be openly accessible, and would facilitate the further use and development of the dataset assembled here. The project includes all the CVe identified by our *in silico* screen, as well as a set of representative reference sequences for the *Circovirus* genus (Table S2). All of these sequences are linked to the appropriate auxiliary data; for the virus sequences, this includes information about the sample from which the sequence was obtained; for CVe, it includes the name of genome assembly and contig in which the CVe sequence was identified, and its coordinates and orientation within that contig.

The project also includes the key alignments constructed in this study, linked together using the GLUE ‘alignment tree’ data structure. These include: (i) ‘tip’ alignments in which all taxa are CVe that are known or putative orthologs of one another; (ii) a ‘root’ alignment constructed to represent proposed homologies between the genomes of representative viruses in the genus *Circovirus* and the CVe recovered by our screen. Because each of these alignments is constrained to a standard reference sequence, alignments are linked to one another.

We applied a systematic approach to naming CVe. Each element was assigned a unique identifier (ID) constructed from a defined set of components. The first component is 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. In cases were a CVe element had expanded via duplication, an additional number was appended to the standard ID to discriminate individual duplicates.

3. Results

3.1. Identification and phylogenetic analysis of vertebrate CVe

We systematically screened 362 vertebrate genome assemblies for CVe, and identified a total of 179 CVe sequences (Table S3**)**, in 52 distinct species (Table 2). For each CVe sequence, we determined the regions of the circovirus genome represented, and attempted to identify genomic flanks. Where genomic flanks were present, we compared these with one another to identify potentially orthologous CVe loci. In several cases, it was not possible to determine whether multiple CVe loci within the same species (or group of closely-related species) represented the outcome of distinct incorporation events, or the fragmented remains of a single, ancestrally acquired element. The main causes of uncertainty were; (i) lack of flanking sequences due to short contig length, or undetermined DNA sequences flanking CVe, and; (ii) the presence of multiple CVe that spanned non-overlapping regions of the circovirus genome. Since CVe are comparatively rare in vertebrate genomes (Katzourakis and Gifford, 2010; Cui et al., 2014), we conservatively assumed a single incorporation event had taken place except in cases where it could be demonstrated otherwise. On this basis, we estimate that the 179 CVe identified here represent at least 19–26 distinct germline incorporation events (Tables 1 and 2, Fig. 1), depending on whether CVe in ray-finned fish are taken to represent a single incorporation event, or seven distinct incorporation events, each in a different order (see Section 3.3). The large discrepancy between the number of elements versus the number of events reflects the fact that 101 of the 179 CVe identified in our study (57%) belong to a group of highly duplicated CVe loci in carnivore genomes, all of which derive

Table 1
CVe detected in published vertebrate genome assemblies.

EVE name	Reference	Genes	# Seqs	# Species
Agnatha				
CVe-*Eptatretus*[a]	This study	Rep	7	1
Bony Fish				
CVe-*Anguilla*[a]	This study	Rep		1
CVe-*Characiformes*[a]	This study	Rep		1
CVe-*Clupeiformes*[a]	This study	Rep		1
CVe-*Cypriniformes*	(Feher et al., 2013)	Rep-Cap		2
CVe-*Cyprinodontiformes*[a]	This study	Rep		1
CVe-*Perciformes*[a]	This study	Rep		3
CVe-*Salmoniformes*[a]	This study	Rep		1
Amphibians				
CVe-*Anura*	(Cui et al., 2014)	Rep	2	2
Reptiles				
CVe-*Viperidae*	(Gilbert et al., 2014)	Rep-Cap	16	6
Birds				
CVe-*Tinamou*	(Cui et al., 2014)	Rep-Cap	2	1
CVe-*Psittaciformes*	(Cui et al., 2014)	Rep-Cap	4	3
CVe-*Passeriformes*	(Cui et al., 2014)	Rep	7	5
CVe-*Egretta*	(Cui et al., 2014)	Rep	1	1
CVe-*Gallirallus*[a]	This study	Rep	1	1
CVe-*Picoides*[a]	This study	Rep	1	1
Mammals				
CVe-*Chrysochloris*[a]	This study	Rep, Cap	3	1
CVe-*Carnivora*	(Katzourakis and Gifford, 2010)	Rep	101	13
CVe-*Mus.caroli*[a]	This study	Rep	1	1
CVe-*Heterocephalus*[a]	This study	Rep	1	1
CVe-*Phascolarctos*[a]	This study	Rep	1	1
CVe-*Sarcophilus*[a]	This study	Rep	1	1
CVe-*Monodelphis*		Rep	1	1
CVe-*Galeopterus*[a]	This study	Rep	2	1
CVe-*Manis*[a]	This study	Rep	1	1
CVe-*Choloepus*[a]	This study	Cap	2	1
		Totals	179	53

[a] Indicate newly identified CVe loci.

Table 2
Vertebrate species with CVe.

Latin binomial	Common name	EVE name	1 st.[a]	Copies [b]
Agnatha				
Eptatretus burgeri	Inshore hagfish	CVe-_Eptatretus_*	x	7
Bony Fish				
Anguilla anguilla	European eel	CVe-_Anguilla_	x	1
Pygocentrus nattereri	Red-bellied piranha	CVe-_Characiformes_		1
Clupea harengus	Atlantic herring	CVe-_Clupeiformes_		1
Cyprinus carpio	Common carp	CVe-_Cypriniformes_		4
Sinocyclocheilus grahami	Golden-line barbel	CVe-_Cypriniformes_		2
Kryptolebias marmoratus	Killifish	CVe-_Cyprinodontiformes_	X	4
Micropterus floridanus	American black bass	CVe-_Perciformes_	X	1
Neolamprologus brichardi	Princess of Burundi	CVe-_Perciformes_	X	2
Acanthochromis polyacanthus	Spiny chromis damselfish	CVe-_Perciformes_	X	5
Salmo salar	Atlantic salmon	CVe-_Salmoniformes_	X	3
Amphibians				
Xenopus tropicalis	Western clawed frog	CVe-_Xenopus_		1
Rana catesbeiana	American bullforg	CVe-_Rana_	x	1
Reptiles				
_Pantherophis guttatus_*	Corn snake	CVe-_Viperidae_		1
_Python molurus_*	Indian python	CVe-_Viperidae_		1
Crotalus horridus	Timber rattlesnake	CVe-_Viperidae_		1
Crotalus mitchellii pyrrhus	Mitchell's rattlesnake	CVe-_Viperidae_		1
_Protobothrops mucrosquamatus_*	Brown spotted pit viper	CVe-_Viperidae_		1
Ophiophagus hannah**	King cobra	CVe-_Viperidae_		1
Birds				
_Serinus canaria_*	Atlantic canary	CVe-_Passeriformes_		1
_Setophaga coronata_*	Yellow-rumped warbler	CVe-_Passeriformes_		1
_Sporophila hypoxantha_*	Lined seedeater	CVe-_Passeriformes_		1
_Zonotrichia albicollis_*	White-throated sparrow	CVe-_Passeriformes_		1
Geospiza fortis	Medium ground finch	CVe-_Passeriformes_		1
_Agapornis roseicollis_*	Rosy-faced lovebird	CVe-_Psittaciformes_		1
_Amazona aestiva_*	Turquoise-fronted amazon	CVe-_Psittaciformes_		2
Nestor notabilis	Kea	CVe-_Psittaciformes_		1
Tinamus guttatus	White-throated tinamou	CVe-_Tinamou_		1
Egretta garzetta	Little egret	CVe-_Egretta_		1
_Gallirallus okinawae_*	Okinawa rail	CVe-_Gallirallus_	x	1
_Picoides pubescens_*	Downy woodpecker	CVe-_Picoides_	x	1
Mammals				
_Ailurus fulgens_*	Red panda	CVe-_Carnivora_		4
Canis familiaris	Domestic dog	CVe-_Carnivora_		4
_Lycaon pictus_*	Cape hunting dog	CVe-_Carnivora_		3
_Acinonyx jubatus_*	Cheetah	CVe-_Carnivora_		1
Felis catus	Domestic cat	CVe-_Carnivora_		3
_Panthera tigris altaica_*	Siberian tiger	CVe-_Carnivora_		1
_Enhydra lutris_*	Sea otter	CVe-_Carnivora_		15
Mustela putorius furo	Ferret	CVe-_Carnivora_		32
_Odobenus rosmarus_*	Walrus	CVe-_Carnivora_		15
_Leptonychotes weddellii_*	Weddell seal	CVe-_Carnivora_		4
_Neomonachus schauinslandi_*	Monk seal	CVe-_Carnivora_		11
Ailuropoda melanoleuca	Panda	CVe-_Carnivora_		7
_Ursus maritimus_*	Polar bear	CVe-_Carnivora_		8
_Heterocephalus glaber_*	Naked mole rat	CVe-_Heterocephalus_	x	1
_Mus caroli_*	Ryuku mouse	CVe-_Mus_	x	1
_Manis pentadactyla_*	Chinese pangolin	CVe-_Manis_	x	1
Monodelphis domestica	Opossum	CVe-_Monodelphis_		1
_Sarcophilus harrisii_*	Tasmanian devil	CVe-_Sarcophilus_	x	1
_Phascolarctos cinereus_*	Koala	CVe-_Phascolarctos_	x	1
Choloepus hoffmanni	Hoffmann's two-toed sloth	CVe-_Choloepus_		2
_Galeopterus variegatus_*	Sunda flying lemur	CVe-_Galeopterus_	x	2
_Chrysochloris asiatica_*	Cape golden mole	CVe-_Chrysochloris_	x	3

Asterisks indicate newly identified circovirus EVEs.

[a] CVe that provide the first evidence of circovirus infection in the host order in which they occur are marked with an 'x'.

[b] Tandem repeated elements were considered to represent a single germline incorporation event.

from a single germline incorporation event.

We only identified four cases where CVe encoding both *rep* and *cap* were present in the same species or species group. In most, only *rep*-derived sequences appear to have been incorporated/retained, and in one case only *cap* (Table 1). We constructed a multiple sequence alignment (MSA) that spanned the entire circovirus genome and contained both reference sequences for CVe (these could be based on individual loci, or a consensus), and representative circovirus reference taxa (Table S2). We used this 'root' MSA (see Section 2.2) to infer which regions of the circovirus genome had been incorporated as CVe. Where CVe spanned coding sequence, we inferred the putative ancestral reading frame by comparing CVe and circovirus sequences, and

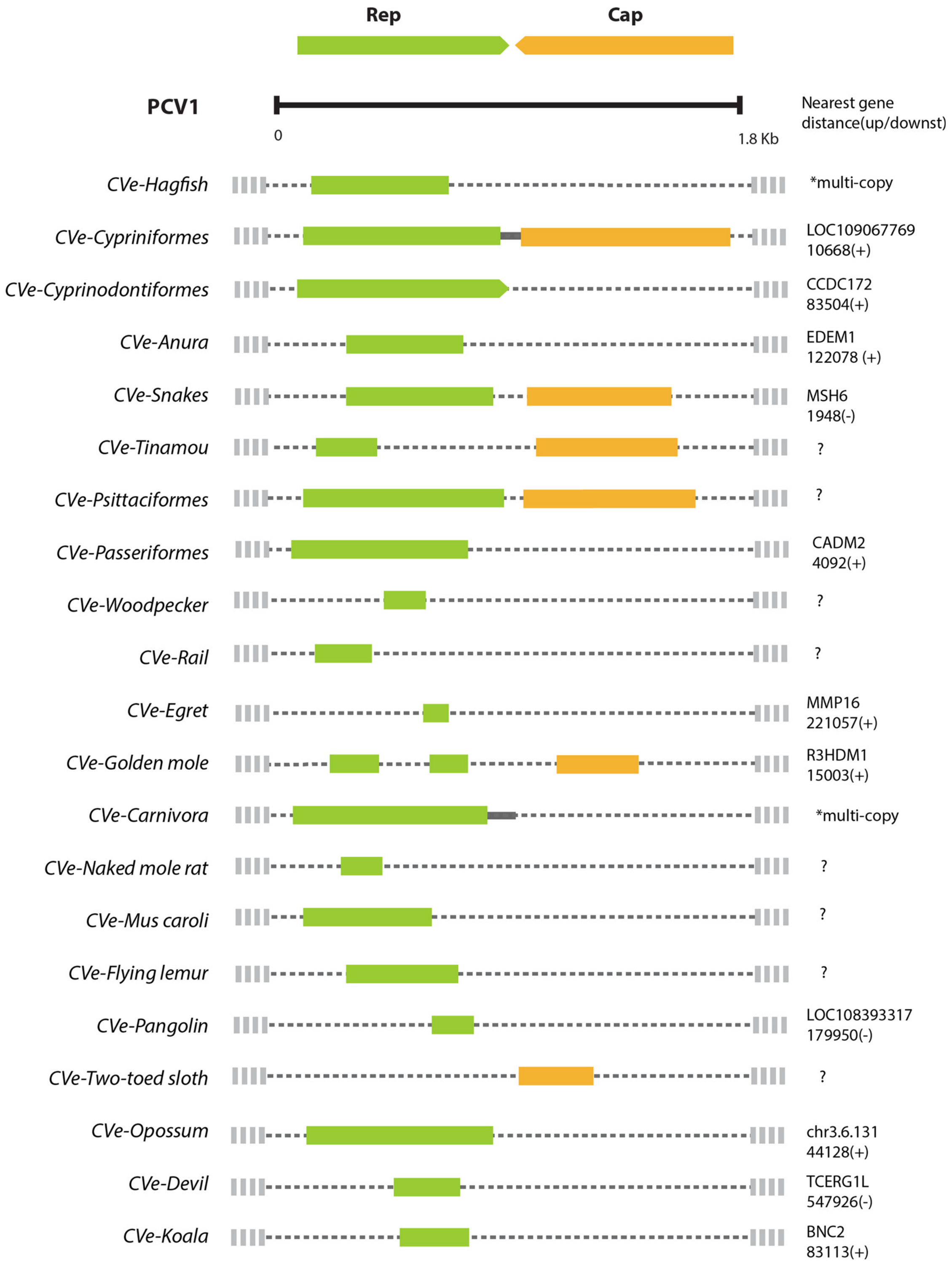

Fig. 1. Genome structures of 21 endogenous circovirus (CVe) elements identified in vertebrate genomes. Genome structures are shown relative to a porcine circovirus 1 (PCV1) reference genome (accession # NC_001792.2). CVe Rep and Cap coding sequences are represented schematically as green and yellow bars. A thickened grey line between the two ORFs indicates internal non-coding region of the circoviral genome. Dotted lines indicate regions of the viral genome that are not represented in CVe. The ID of the nearest annotated gene, where one could be identified, is shown to the right of each element. The distance in base pairs (bp) and the direction relative to the CVe - i.e. upstream (-), downstream (+) are shown. Question marks are shown for CVe that could not be located relative to a host gene (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

attempting to identify likely frameshifting mutations. Most CVe represent only fragments of the genome (Fig. 1), and many are relatively degraded, containing multiple frameshifting indels and stop codons.

Where we identified several CVe from the same species, we compared genomic regions to search for evidence of homology and thereby identify orthologs. Where we were able to identify orthologous CVe insertions, we used these data to create a timeline of circovirus evolution (Fig. 2). In addition, we identified sets of 'potentially orthologous' CVe, where sequence similarity and phylogenetic relationships were consistent with orthology, but this could not be confirmed or ruled out

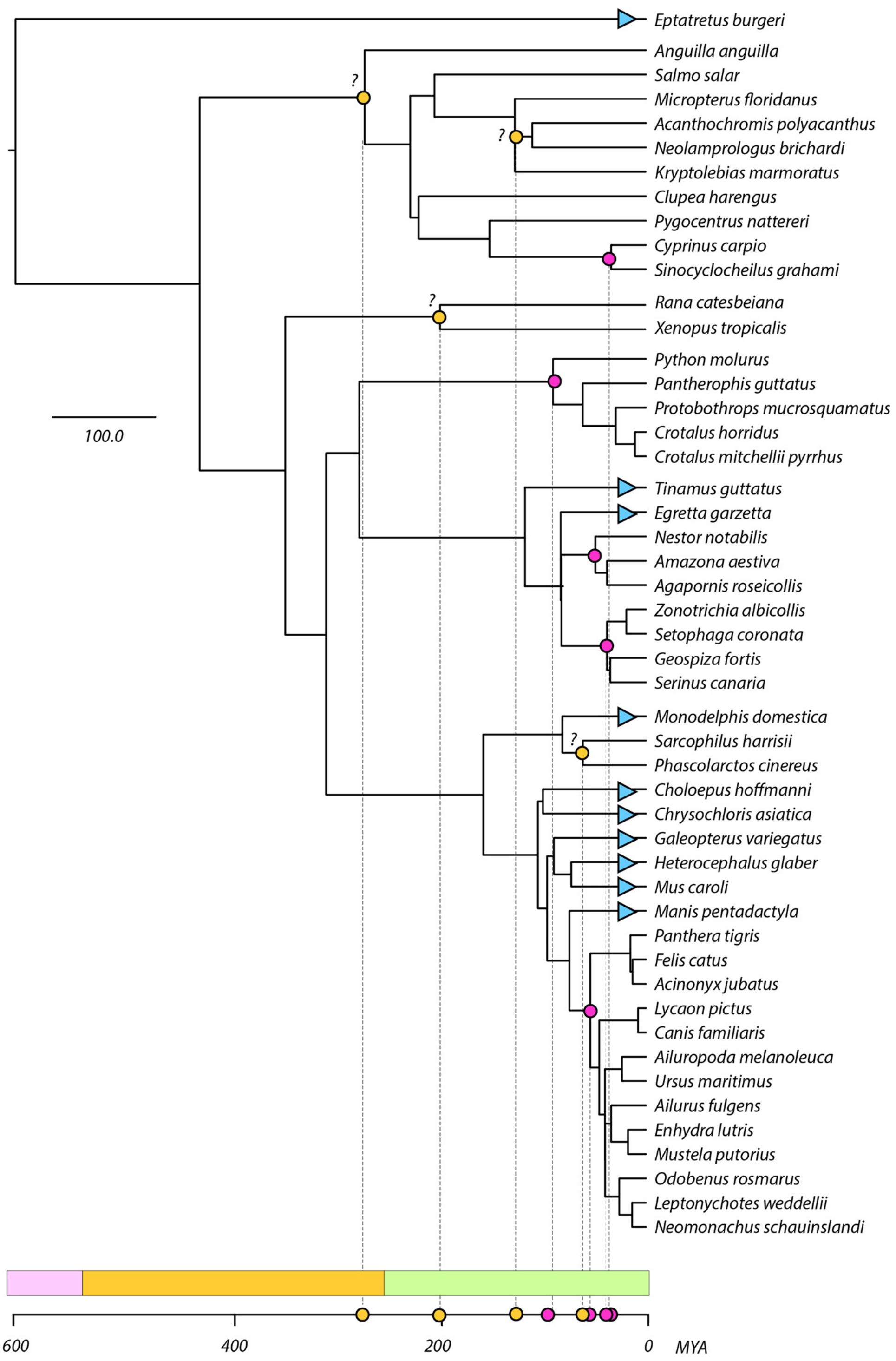

Fig. 2. Evolutionary relationships of vertebrate species in which CVe have been identified, and timeline of CVe evolution. Pink circles indicate confirmed orthologs. Yellow circles indicate the presence of potential orthologs that have not been confirmed. Blue triangles indicate where CVe loci are present, but no information about their ages could be obtained. Phylogeny obtained from the TimeTree database (Kumar et al., 2017) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

based on flanking sequences.

A range of distinct partitions were derived from the virtually translated root MSA (with frameshifts removed), and used to construct bootstrapped ML phylogenies (Fig. 3). In general, support for the deeper branching relationships between CVe and circoviruses was weak, irrespective of which genomic region was used to construct trees. This reflects the fact that most CVe are short and/or highly degraded, and these sequences tend to group distantly from other taxa. However, in phylogenies based on Rep (Fig. 3), several robustly supported subgroupings were observed, three of which – referred to here as mammal 1, cyprinid 1, and avian 1 - included a mixture of CVe and contemporary circoviruses. Notably, in all three of these clades, CVe and circovirus sequences were obtained from the same hosts of the same taxonomic class. The sections that follow describe the distribution and diversity of CVe within individual classes in the subphylum Vertebrata (chordates with backbones).

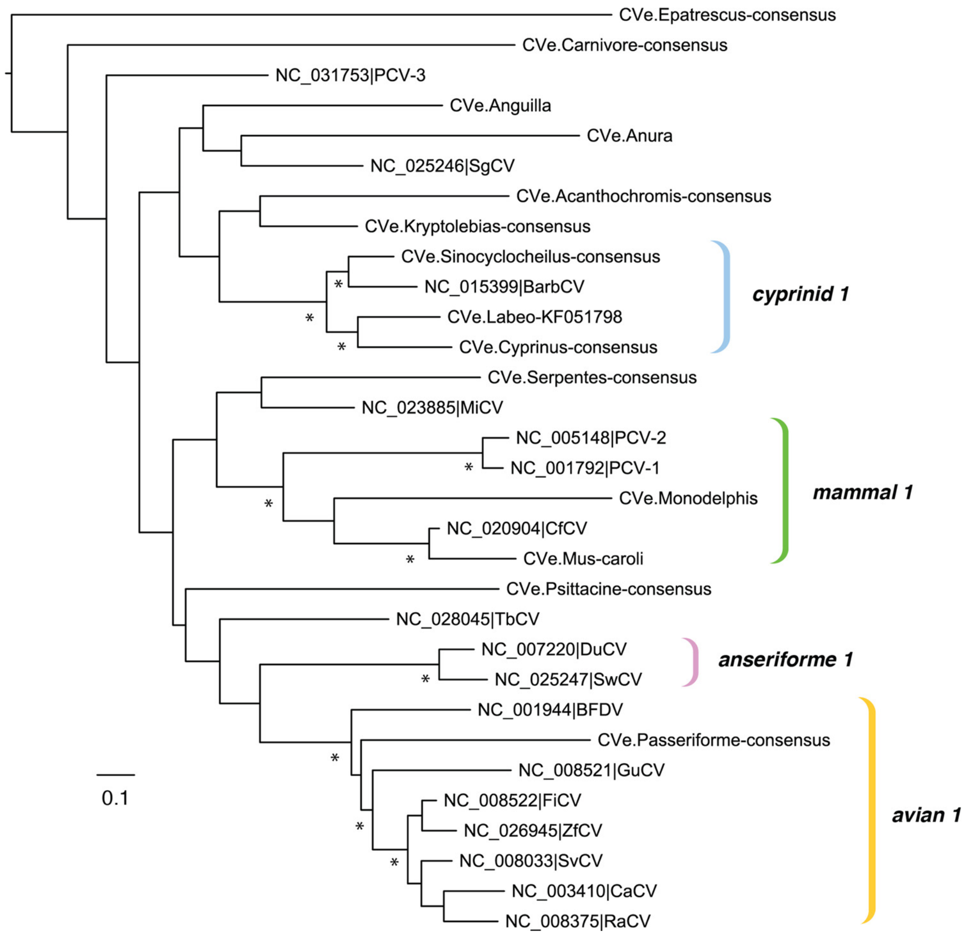

Fig. 3. A maximum likelihood phylogeny showing estimated evolutionary relationships between eCVe and circoviruses. The phylogeny constructed from an alignment spanning ~200 amino acids in Rep. The scale bar shows evolutionary distance in substitutions per site. Bootstrap support was assessed using 1000 bootstrap replicates. Asterisks indicate nodes with bootstrap > 75%.

3.2. CVe in jawless vertebrates

Extant vertebrates are divided into the jawed vertebrates (Gnathostomata) and jawless vertebrates (Agnatha). The Agnatha represent the most basal group of vertebrates and includes the hagfishes (myxinoids) and lampreys (petromyzontids). We identified seven sequences exhibiting homology to *rep* in the genome assembly of the inshore hagfish (*Eptatretus burgeri*). These sequences are relatively distinct from other circoviruses, and also showed relatively high genetic diversity relative to one another, forming three distinct groups in phylogenetic trees (Fig. S1**)**. Notably, the putative Rep polypeptides encoded by these sequences contained several in-frame indels relative to one another. Because such a pattern of variation is unlikely to arise through neutral accumulation of mutations in the germline, this suggests the occurrence of at least three distinct genome incorporation events, each involving distinct, but relatively closely related viruses. However, since we were unable to identify unambiguous genomic flanking sequences for any of these loci, their classification as CVe should for now be considered tentative.

3.3. CVe in ray-finned fish (class Actinopterygii)

Circoviruses are thought to infect barbel fish (*Barbus barbus*) and European catfish (*Silurus glanis*), based on (i) the observation of viral particles in tissues, and the recovery of circovirus sequences from these tissues via nested PCR (Lorincz et al., 2011, 2012). In addition, CVe have been reported in one fish species - the Indian rohu (*Labeo rohita*) (Feher et al., 2013). We identified numerous additional CVe sequence in the genome assemblies of ray-finned fishes (Class *Actinopterygii*) (Tables 2 and S3). We established that at least two of these CVe - occurring in the common carp (*Cyprinus carpio*) and golden-line barbell (*Sinocyclocheilus grahami*) genomes - were orthologs of one another, indicating they were incorporated into the germline of cyprinid fish more than 39 million years ago (Wang et al., 2012; Ren and Mayden, 2016). These CVe were comprised of multiple complete circovirus genomes arranged in tandem, and intriguingly, were observed CVe group as sister taxa to barbel circovirus (BarbCV) in phylogenetic trees, sharing ~70% nucleotide identity (across 1654 nucleotides) with the BarbCV genome.

We also identified matches to *rep* in eight other species of ray-finned fish (Table 2). We could not determine with certainty how many integration events these CVe represented. Interestingly, however, all of these sequences group together in phylogenies (Fig. 3), and the phylogeny constructed for these elements - when rooted on the CVe from the most basal host - the European eel (*Anguilla anguilla*), approximately follows that of the host species, consistent with a single ancestral integration event > 200 million years ago (Fig. 2). Alternatively, the CVe observed in distinct orders might represent distinct incorporation events. This is supported by the placement of CVe.anura in phylogenies, in which it splits the fish CVe from one another, albeit with weak support (Fig. S1). In addition, the observation that CVe elements in order cypriniforme fish (golden-line barbell and carp) occur as full-length tandem genomes, whereas those in Perciformes are derived from more divergent fragments of *rep*, is suggestive of at least two separate incorporation events. Notably one CVe in the mangrove rivulus (*Kryptolebias marmoratus*) encoded a complete intact *rep* gene (Fig. 1) that is predicted to be expressed, suggesting it may have been functionalized in some manner.

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. However, 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.

3.4. CVe in amphibians

Sequences homologous to circovirus *rep* genes have previously been identified in the Western clawed frog (*Xenopus tropicalis*) (Liu et al., 2011). We identified CVe in the genome of the American bullfrog (*Rana catesbeiana*) that partially overlaps that identified in *Xenopus*. Potentially, these sequences could be orthologs of one another, which would imply a minimum age of ~204 MYA (Cannatella, 2015; Roelants et al., 2011) (Fig. 2). However, we were unable to confirm this based on analysis of flanking genomic sequences.

3.5. CVe in reptiles

A pair of orthologous CVe, each covering about 75% of the circovirus genome, have previously been recovered from rattlesnake genomes (*Crotalus spp*) (Gilbert et al., 2014). We identified CVe in four additional snake species (Table 2). Examination of aligned snake CVe sequences indicated that all are likely to be orthologs of those previously reported in rattlesnakes (see Fig. S2), implying that this CVe integrated into the serpentine germline ~72–90 million years ago (Mya) (Fig. 2).

3.6. CVe in birds (class Aves)

CVe have previously been reported in the genomes of several avian species: the little egret (*Egretta garzetta*), white-throated tinamou (*Tinamus guttatus*), medium ground-finch (*Geospiza fortis*), and kea (*Nestor notabilis*) (Cui et al., 2014; Liu et al., 2011). We identified CVe in eight additional species. Some of these appeared likely to be orthologs of CVe reported previously. For example, we identified CVe in two species of psittacine bird that appeared represented orthologs of one another, and possibly of those previously identified in the kea (*Nestor notabilis*) (Cui et al., 2014) (Table 2), which would imply integration into the psittacine germline prior to the divergence of the major extant lineages within the order Psittaciformes (estimated to have occurred 30–60 Mya (Jarvis et al., 2014; Jetz et al., 2012)) (Fig. 2).

We also identified orthologs of the *rep*-derived insertion previously described in the medium ground finch in several additional species in the avian order Passeriformes (songbirds) (Table 2). Identification of these orthologs demonstrates that this particular CVe predates the radiation of avian sub-order Passeroida ~38 Mya (Jarvis et al., 2014; Prum et al., 2015) (Fig. 2).

In addition to identifying the previously reported CVe in the genomes of the white-throated tinamou (*Tinamus guttatus*) and little egret (*Egretta garzetta*) (Cui et al., 2014), we identified previously unreported CVe in the Japanese rail (*Gallirallus okinawae*: order Gruiformes) and downy woodpecker (*Picoides pubescens*: order Piciformes) (Table 2). Both these sequences were relatively short and divergent, and consequently we could not determine their relationships to other CVe and circoviruses with confidence.

3.7. CVe in mammals (class Mammalia)

The majority of CVe identified in our screen were recovered from carnivore genome assemblies. As far as we are able to discern from phylogenetic and comparative analysis, all of these CVe derive from 1 to 4 germline incorporation events involving an ancient carnivore *rep* gene. However, the copy number of these elements has expanded subsequent to their incorporation into the germline, in some cases quite dramatically. The grouping of carnivore CVe in phylogenies (Fig. 4) indicates that at least four CVe insertions were present in the carnivore germline prior to the divergence of extant families within this order. The copy number of one particular element (referred to here as CVe-Carnivora-4) has expanded in some carnivore lineages. As shown in Fig. 4, the phylogenetic relationships between duplicates in the group CVe-Carnivora-4 indicate that these expansions have occurred independently in ursids (bears), pinnipeds (seals and walruses), and mustelids. One plausible explanation for the elevated copy number in certain carnivore lineages is that CVe have become embedded into retroelements and copied along with these sequences when they undergo transposition. To investigate this loci were examined in DFAM (Hubley et al., 2016) to detect repetitive elements flanking CVe (Table S4). This investigation showed that carnivore CVe were overwhelmingly associated with LINE-1 elements. Three of the four expansion groups shown in Fig. 4 were found to be associated with LINE-1 DNA, including the most basal, indicating that the fusion of CVe and LINE-1 occurred relatively early in carnivore evolution.

A novel, relatively well-preserved *rep*-derived CVe was identified in the genome of the Ryukyu mouse (*Mus caroli*) that grouped closely with circoviruses genome recovered from dogs (Li et al., 2013; Decaro et al., 2014). This element presumably arose after this species diverged from the house mouse (*Mus musculus*) ~6–7 Mya, since it is absent from this species.

In the cape golden mole (*Chrysochloris asiaticus*) matches to both *cap* and *rep* were identified. However, these occurred on distinct contigs and did not overlap. Furthermore, both CVe were relatively short and degraded, and were highly divergent relative to other CVe. CVe derived from *cap* were also identified in the genome of Hoffmann's two-toed sloth (*Choloepus hoffmanni*) (Fig. 1).

CVe have previously been identified in the genome of the short-tailed opossum (*Monodelphis domestica*), an American marsupial (Katzourakis and Gifford, 2010). In phylogenies based on rep, this sequence groups together with the porcine circoviruses, canine circovirus, and the CVe we identified in *Mus caroli*. We identified the first examples of CVe from the genomes of Australian marsupial species: the Tasmanian devil (*Sarcophilus harrisii*) and the koala (*Phascolarctos cinereus*). Both these sequences derived from circovirus *rep* genes, and grouped together in phylogenetic trees (Fig. S1). However, their placement relative to other taxa was not supported with confidence, reflecting their short and degraded nature. Several other short and degraded matches to Rep probes were identified in other mammalian species (Tables 1 and 2, Fig. 1). These sequences were relatively distantly related to one another and to contemporary circoviruses.

4. Discussion

4.1. CVe provide retrospective information about circovirus evolution

In this study, we recovered CVe from published vertebrate genomes, determined their genomic structures, and examined their phylogenetic relationships to contemporary circoviruses. Our analysis is the first to examine such a large set of CVe sequences, and to screen so widely within vertebrates. We show that CVe are relatively widespread in vertebrate genomes, though it appears they are absent from some lineages (e.g. primates, in which genome coverage is relatively high).

Several of the CVe loci identified here have been reported previously (Katzourakis and Gifford, 2010; Cui et al., 2014; Feher et al., 2013; Liu et al., 2011), and the majority of novel CVe sequences recovered by our screen were orthologs or duplicates of these loci. Nevertheless, we identified 17 CVe loci that have not been reported before (Tables 1, and S3). These sequences provide the first evidence of (ancestral) circovirus infection for several species (Table 2). In addition, the identification and characterisation of novel orthologs allowed us to establish the first minimum age estimates for some CVe loci, and to markedly extended those of others. Thus, we were able to derive a more accurately calibrated timeline of evolution for the *Circovirus* genus, spanning multiple geological eras (Fig. 2). Unfortunately, many of the

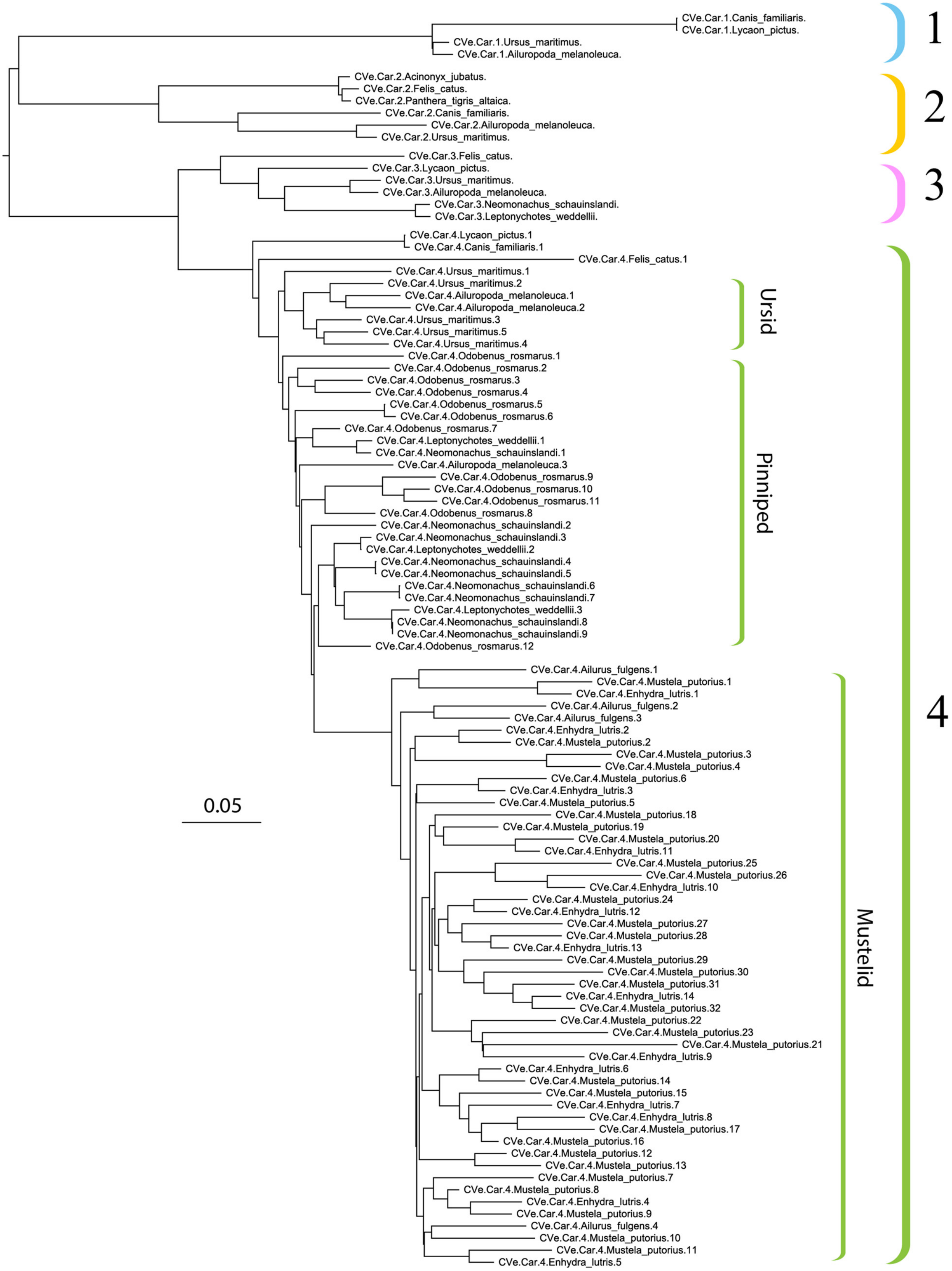
CVe.Car.1.Canis_familiaris.
CVe.Car.1.Lycaon_pictus.
CVe.Car.1.Ursus_maritimus.
CVe.Car.1.Ailuropoda_melanoleuca.
1
CVe.Car.2.Acinonyx_jubatus.
CVe.Car.2.Felis_catus.
CVe.Car.2.Panthera_tigris_altaica.
CVe.Car.2.Canis_familiaris.
CVe.Car.2.Ailuropoda_melanoleuca.
CVe.Car.2.Ursus_maritimus.
2
CVe.Car.3.Felis_catus.
CVe.Car.3.Lycaon_pictus.
CVe.Car.3.Ursus_maritimus.
CVe.Car.3.Ailuropoda_melanoleuca.
CVe.Car.3.Neomonachus_schauinslandi.
CVe.Car.3.Leptonychotes_weddellii.
3
CVe.Car.4.Lycaon_pictus.1
CVe.Car.4.Canis_familiaris.1
CVe.Car.4.Felis_catus.1
CVe.Car.4.Ursus_maritimus.1
CVe.Car.4.Ursus_maritimus.2
CVe.Car.4.Ailuropoda_melanoleuca.1
CVe.Car.4.Ailuropoda_melanoleuca.2
CVe.Car.4.Ursus_maritimus.3
CVe.Car.4.Ursus_maritimus.5
CVe.Car.4.Ursus_maritimus.4
Ursid
CVe.Car.4.Odobenus_rosmarus.1
CVe.Car.4.Odobenus_rosmarus.2
CVe.Car.4.Odobenus_rosmarus.3
CVe.Car.4.Odobenus_rosmarus.4
CVe.Car.4.Odobenus_rosmarus.5
CVe.Car.4.Odobenus_rosmarus.6
CVe.Car.4.Odobenus_rosmarus.7
CVe.Car.4.Leptonychotes_weddellii.1
CVe.Car.4.Neomonachus_schauinslandi.1
CVe.Car.4.Ailuropoda_melanoleuca.3
CVe.Car.4.Odobenus_rosmarus.9
CVe.Car.4.Odobenus_rosmarus.10
CVe.Car.4.Odobenus_rosmarus.11
CVe.Car.4.Odobenus_rosmarus.8
CVe.Car.4.Neomonachus_schauinslandi.2
CVe.Car.4.Neomonachus_schauinslandi.3
CVe.Car.4.Leptonychotes_weddellii.2
CVe.Car.4.Neomonachus_schauinslandi.4
CVe.Car.4.Neomonachus_schauinslandi.5
CVe.Car.4.Neomonachus_schauinslandi.6
CVe.Car.4.Neomonachus_schauinslandi.7
CVe.Car.4.Leptonychotes_weddellii.3
CVe.Car.4.Neomonachus_schauinslandi.8
CVe.Car.4.Neomonachus_schauinslandi.9
CVe.Car.4.Odobenus_rosmarus.12
Pinniped
CVe.Car.4.Ailurus_fulgens.1
CVe.Car.4.Mustela_putorius.1
CVe.Car.4.Enhydra_lutris.1
CVe.Car.4.Ailurus_fulgens.2
CVe.Car.4.Ailurus_fulgens.3
CVe.Car.4.Enhydra_lutris.2
CVe.Car.4.Mustela_putorius.2
CVe.Car.4.Mustela_putorius.3
CVe.Car.4.Mustela_putorius.4
CVe.Car.4.Mustela_putorius.6
CVe.Car.4.Enhydra_lutris.3
CVe.Car.4.Mustela_putorius.5
CVe.Car.4.Mustela_putorius.18
CVe.Car.4.Mustela_putorius.19
CVe.Car.4.Mustela_putorius.20
CVe.Car.4.Enhydra_lutris.11
CVe.Car.4.Mustela_putorius.25
CVe.Car.4.Mustela_putorius.26
CVe.Car.4.Enhydra_lutris.10
CVe.Car.4.Mustela_putorius.24
CVe.Car.4.Enhydra_lutris.12
CVe.Car.4.Mustela_putorius.27
CVe.Car.4.Mustela_putorius.28
CVe.Car.4.Enhydra_lutris.13
CVe.Car.4.Mustela_putorius.29
CVe.Car.4.Mustela_putorius.30
CVe.Car.4.Mustela_putorius.31
CVe.Car.4.Enhydra_lutris.14
CVe.Car.4.Mustela_putorius.32
CVe.Car.4.Mustela_putorius.22
CVe.Car.4.Mustela_putorius.23
CVe.Car.4.Mustela_putorius.21
CVe.Car.4.Enhydra_lutris.9
CVe.Car.4.Enhydra_lutris.6
CVe.Car.4.Mustela_putorius.14
CVe.Car.4.Mustela_putorius.15
CVe.Car.4.Enhydra_lutris.7
CVe.Car.4.Enhydra_lutris.8
CVe.Car.4.Mustela_putorius.17
CVe.Car.4.Mustela_putorius.16
CVe.Car.4.Mustela_putorius.12
CVe.Car.4.Mustela_putorius.13
CVe.Car.4.Mustela_putorius.7
CVe.Car.4.Mustela_putorius.8
CVe.Car.4.Enhydra_lutris.4
CVe.Car.4.Mustela_putorius.9
CVe.Car.4.Ailurus_fulgens.4
CVe.Car.4.Mustela_putorius.10
CVe.Car.4.Mustela_putorius.11
CVe.Car.4.Enhydra_lutris.5
Mustelid
4
0.05

Fig. 4. Phylogeny of CVe Rep amino acid sequences recovered from carnivore genome assemblies. At least four distinct CVe loci are present in the carnivore germline (clades I—IV) as indicated by the coloured brackets. Within group IV, distinct copy number expansions appear to have occurred in ursids (bears), pinnipeds (seals and walruses), and mustelids. The scale bar shows evolutionary distance in substitutions per site. The tree is midpoint rooted for display purposes.

internal nodes in our phylogeny lacked robust support, and this limited our ability to extrapolate calibrations determined for orthologous loci across the phylogeny. However, further sampling of circoviruses and CVe should allow the evolutionary relationships within the *Circovirus* genus to be determined with greater confidence, which will in turn allow more extensive calibrations to be made.

Importantly, we observed that CVe in fish, birds and mammals cluster phylogenetically with exogenous circoviruses identified from the same host class. This implies that there is a degree of stability to the relationship between circoviruses and their hosts, at least at higher taxonomic levels.

4.2. Impact of CVe on host genome evolution

The majority of CVe are derived from *rep* genes. To the extent that CVe have been exapted or co-opted, the predominance of CVe derived from *rep* might reflect that these sequences are more readily functionalised than those derived from *cap*. Furthermore, we identified one CVe in the mangrove rivulus (*Kryptolebias marmoratus*) that encoded an intact *rep* gene that is predicted to express mRNA, suggesting it may have been functionalized in some manner (Fig. 1).

Notably, several distinct examples have now been described of endogenous viral elements (EVEs) that are derived from viral polymerase or replication-associated genes, are expressed, and encode intact ORFs (Katzourakis and Gifford, 2010; Arriagada and Gifford, 2014; Horie et al., 2016). These elements are derived from a range of different viruses, and have clearly arisen in distinct events, suggesting there might be a common mechanism causing EVEs derived from the replicase of distinct viruses to be selected and maintained in different species. Alternatively, it is possible that the discrepancy in numbers simply reflects that *cap*-derived sequences are less conserved and therefore harder to detect.

Curiously, it is rare for more than one CVe to occur in the germline of any jawed vertebrate lineage. Carnivores are an obvious exception, since CVe have been amplified to relatively high copy number (10–20 copies) in several carnivore lineages (Fig. 4), apparently via LINE-mediated retrotransposition (Table S4).

5. Conclusions

We identify the complete repertoire of CVe sequences in published vertebrate genome assemblies. Through comparative analysis of these sequences, we provide the most complete picture yet of how viruses in the genus *Circovirus* have evolved and interacted with their hosts over the course of their evolution. The sequence-based resource implemented here can facilitate further characterisation of circovirus distribution, diversity and evolution as new CVe and circovirus sequence data become available.

Acknowledgements

RJG was funded by the Medical Research Council of the United Kingdom (MC_UU_12014/12). WMS is supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo, Brazil (Scholarships N0. 17/13981-0).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.virusres.2018.03.014.

References

Altschul, S.F., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.
Amery-Gale, J., et al., 2017. A high prevalence of beak and feather disease virus in non-psittacine Australian birds. J. Med. Microbiol. 66 (7), 1005–1013.
Arriagada, G., Gifford, R.J., 2014. Parvovirus-derived endogenous viral elements in two South American rodent genomes. J. Virol. 88 (20), 12158–12162.
Belyi, V.A., Levine, A.J., Skalka, A.M., 2010. Sequences from ancestral single-stranded DNA viruses in vertebrate genomes: the parvoviridae and circoviridae are more than 40 to 50 million years old. J. Virol. 84 (23), 12458–12462.
Cannatella, D., 2015. Xenopus in space and time: fossils, node calibrations, tip-dating, and paleobiogeography. Cytogenet. Genome Res. 145 (3-4), 283–301.
Cui, J., et al., 2014. Low frequency of paleoviral infiltration across the avian phylogeny. Genome Biol. 15 (12), 539.
Darriba, D., et al., 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27 (8), 1164–1165.
Decaro, N., et al., 2014. Genomic characterization of a circovirus associated with fatal hemorrhagic enteritis in dog, Italy. PLoS One 9 (8), e105909.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32 (5), 1792–1797.
Feher, E., et al., 2013. Integrated circoviral rep-like sequences in the genome of cyprinid fish. Virus Genes 47 (2), 374–377.
Feschotte, C., Gilbert, C., 2012. Endogenous viruses: insights into viral evolution and impact on host biology. Nat. Rev. Genet. 13 (4), 283–296.
Gilbert, C., et al., 2014. Endogenous hepadnaviruses, bornaviruses and circoviruses in snakes. Proc. Biol. Sci. 281 (1791), 20141122.
Holmes, E.C., 2011. The evolution of endogenous viral elements. Cell Host Microbe 10 (4), 368–377.
Horie, M., et al., 2016. An RNA polymerase gene in bat genomes derived from an ancient negative-strand RNA virus. Sci. Rep. 6, 25873.
Hubley, R., et al., 2016. The Dfam database of repetitive DNA families. Nucleic Acids Res. 44 (D1), D81–D89.
Jarvis, E.D., et al., 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346 (6215), 1320–1331.
Jetz, W., et al., 2012. The global diversity of birds in space and time. Nature 491 (7424), 444–448.
Katzourakis, A., Gifford, R.J., 2010. Endogenous viral elements in animal genomes. PLoS Genet. 6 (11), e1001191.
Kumar, S., et al., 2017. TimeTree: a resource for timelines, timetrees, and divergence times. Mol. Biol. Evol. 34 (7), 1812–1819.
Li, L., et al., 2013. Circovirus in tissues of dogs with vasculitis and hemorrhage. Emerg. Infect. Dis. 19 (4), 534–541.
Liu, H., et al., 2011. Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. BMC Evol. Biol. 11, 276.
Lorincz, M., et al., 2011. First detection and analysis of a fish circovirus. J. Gen. Virol. 92 (Pt 8), 1817–1821.
Lorincz, M., et al., 2012. Novel circovirus in European catfish (Silurus glanis). Arch. Virol. 157 (6), 1173–1176.
Muller, T., Vingron, M., 2000. Modeling amino acid replacement. J. Comput. Biol. 7 (6), 761–776.
Prum, R.O., et al., 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526 (7574), 569–573.
Rambaut, A., 2002. SE-AL Sequence Alignment Editor. University of Oxford, Oxford, UK.
Ren, Q., Mayden, R.L., 2016. Molecular phylogeny and biogeography of African diploid barbs, 'Barbus', and allies in Africa and Asia (Teleostei: Cypriniformes). Zoologica Scripta 45, 642–649.
Roelants, K., Haas, A., Bossuyt, F., 2011. Anuran radiations and the evolution of tadpole morphospace. Proc. Natl. Acad. Sci. U. S. A. 108 (21), 8731–8736.
Rosario, K., et al., 2017. Revisiting the taxonomy of the family Circoviridae: establishment of the genus Cyclovirus and removal of the genus Gyrovirus. Arch. Virol. 162 (5), 1447–1463.
Singer, J.B., et al., 2018. GLUE: A Flexible Software System for Virus Sequence Data. bioRxiv.
Suyama, M., Torrents, D., Bork, P., 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34, W609–W612 Web Server issue.
Wang, X., et al., 2012. Cyprinid phylogeny based on Bayesian and maximum likelihood analyses of partitioned data: implications for Cyprinidae systematics. Sci. China Life Sci. 55 (9), 761–773.