VIRUS EVOLUTION

Virus Evolution, 2016, 2(2): vew025

doi: 10.1093/ve/vew025
Research article

Diversity and comparative genomics of chimeric viruses in *Sphagnum*-dominated peatlands

Achim Quaiser,[1,*] Mart Krupovic,[2] Alexis Dufresne,[1] André-Jean Francez,[1] and Simon Roux[3,†]

[1]Université de Rennes 1, CNRS UMR6553 EcoBio, Rennes, France, [2]Department of Microbiology, Institut Pasteur, Paris, France and [3]Department of Microbiology, Ohio State University, Columbus, OH, USA

*Corresponding author: E-mail: achim.quaiser@univ-rennes1.fr
†http://orcid.org/0000-0002-5831-5895

Abstract

A new group of viruses carrying naturally chimeric single-stranded (ss) DNA genomes that encompass genes derived from eukaryotic ssRNA and ssDNA viruses has been recently identified by metagenomic studies. The host range, genomic diversity, and abundance of these chimeric viruses, referred to as cruciviruses, remain largely unknown. In this article, we assembled and analyzed thirty-seven new crucivirus genomes from twelve peat viromes, representing twenty-four distinct genome organizations, and nearly tripling the number of available genomes for this group. All genomes possess the two characteristic genes encoding for the conserved capsid protein (CP) and a replication protein. Additional ORFs were conserved only in nearly identical genomes with no detectable similarity to known genes. Two cruciviruses possess putative introns in their replication-associated genes. Sequence and phylogenetic analyses of the replication proteins revealed intragene chimerism in at least eight chimeric genomes. This highlights the large extent of horizontal gene transfer and recombination events in the evolution of ssDNA viruses, as previously suggested. Read mapping analysis revealed that members of the 'Cruciviridae' group are particularly prevalent in peat viromes. Sequences matching the CP ranged from 0.6 up to 10.9 percent in the twelve peat viromes. In contrast, from sixty-nine available viromes derived from other environments, only twenty-four contained cruciviruses, which on average accounted for merely 0.2 percent of sequences. Overall, this study provides new genome information and insights into the diversity of chimeric viruses, a necessary first step in progressing toward an accurate quantification and host range identification of these new viruses.

Key words: virus ecology; viral metagenomics; virus diversity; ssDNA viruses.

1. Introduction

Viruses are the most abundant biological entity on Earth (Wommack and Colwell 2000) and present in all kinds of ecosystems. They regulate the structure of microbial communities and influence food–web interactions and global geochemical cycles (Fuhrman 1999; Suttle 2007). Although our knowledge of viral diversity is still largely incomplete, recent advances in sequencing technologies combined with culture independent molecular techniques enable the identification of numerous novel viral genera and the analysis of countless novel viral genomes. Single-stranded (ss) DNA viruses are among the smallest viruses infecting prokaryotes and eukaryotes. Viruses with ssDNA genomes infect hosts from all three domains of life and are classified into eleven families and one unassigned genus (Krupovic and Forterre 2015; Krupovic et al. 2016). Eight of these taxa include viruses infecting various eukaryotes, from fungi to human beings. However, many groups of ssDNA viruses, primarily detected by metagenomics studies, remain unclassified. With the notable exception of insect-infecting viruses of the *Bidnaviridae* family, all eukaryotic ssDNA viruses encode homologous rolling-circle replication-initiation proteins (RC-Rep), with characteristic N-terminal endonuclease domains and C-terminal superfamily three helicase domains (Krupovic 2013). The RC-Reps from different families of eukaryotic ssDNA viruses display moderate sequence conservation and are often used for phylogenetic analyses and taxonomic affiliation. This conservation of the Replication-associated proteins led to the designation of these viruses as “circular Rep-encoding single-strand DNA” viruses, or CRESS-DNA viruses (Rosario, Duffy, and Breitbart 2012). All known eukaryotic ssDNA viruses also form icosahedral capsids (Krupovic 2013). Notably, unlike in the case of RC-Reps, capsid proteins (CPs) encoded by viruses belonging to different families do not display recognizable sequence similarity.

Recently, a new group of ssDNA viruses tentatively called RNA-DNA hybrid viruses (Diemer and Stedman 2012) or chimeric viruses (Roux et al. 2013) was identified in aquatic ecosystems (for an overview see Stedman 2015). Since then, related chimeric viruses genomes were identified in samples collected from sewage treatment oxidation pond (Kraberger et al. 2015) marine water (Mcdaniel et al. 2014), aquatic arthropods (Hewson et al. 2013; Dayaram et al. 2016; Steel et al. 2016), animal fecal matter (Steel et al. 2016) and, unexpectedly, in nucleic acid extraction spin-columns (Krupovic et al. 2015). Overall, all RC-Reps proteins from chimeric viruses show homology to eukaryotic ssDNA viruses from the families *Geminiviridae*, *Circoviridae*, and *Nanoviridae*, whereas the CP is related to those of ssRNA viruses from the family *Tombusviridae* and unclassified oomycete-infecting viruses. Notably, the genomes of chimeric viruses are considerably larger than those of other known eukaryotic ssDNA viruses but comparable in size to the genomes of ssRNA viruses from the *Tombusviridae* family (Roux et al. 2013). All known chimeric viruses also show relatively high similarity in their CP proteins, suggesting a recent divergence from a common ancestor.

In this study, we analyze thirty-seven new complete chimeric viruses genomes reconstructed through *de novo* assembly of sequences from twelve viromes obtained from peat soil water samples collected in a *Sphagnum*-dominated peatland. Detailed gene content comparisons and phylogenetic analyses provide new insights into the diversity, distribution, and abundance of this novel group of ssDNA viruses.

2. Material and Methods

2.1. Sampling, accession to twelve peat viromes, assembly, annotation, and comparative genome analysis

The sampling of peat and virome production was conducted as previously described (Quaiser et al. 2015; Ballaud et al. 2016). In short, twelve samples were recovered from a *Sphagnum*-dominated peatland at “les Pradeaux mire” in the French Massif Central (3°55′E; 45°32′N) at an altitude of 1,250 m. Peat water was sampled in Fen and Bog dynamic states at three different dates (in June, August, and October 2011) without replicates and from Fen and Bog (March 2012) in biological triplicates. Viruses were concentrated using PEG precipitation (Colombet et al. 2007). Triplicate whole genome amplification using Phi29 DNA polymerase, library construction, sequencing assembly, annotation, and comparative analysis was performed as previously described (Quaiser et al. 2015). The identification of circular genomes was analyzed using MUMmer for genome recruitment (Kurtz et al. 2004). Only when at least one sequence covered the beginning and the end of the contig with 100 percent sequence identity was the contig considered as circular (Quaiser et al. 2015). General sequence manipulation was done on the Biogenouest Galaxy web platform (Le Bras et al. 2013).

2.2. Phylogenetic and similarity analysis

Sequences of full-length CPs and RC-Reps were aligned using MUSCLE (Edgar 2004) or PROMALS3D (Pei, Kim, and Grishin 2008), respectively. Alignments were manually edited using ARB (Ludwig et al. 2004). Gaps and ambiguously aligned positions were excluded from phylogenetic analysis. Maximum likelihood trees for CP were reconstructed using TREEFINDER (Jobbvon Haeseler, and Strimmer 2004) applying a JTT model (amino acid) and GTR3 (nucleic acid) of sequence evolution with a four category discrete approximation of a gamma distribution plus invariant sites. The best-fit models were determined using TREEFINDER. Maximum likelihood trees for the RC-Rep were reconstructed using PHYML 3.1 (Guindon et al. 2010) using the automatic model selection (VT+G6+I+F). Maximum likelihood bootstrap proportions were inferred using 1,000 replicates.

Sequence similarity between the CPs of chimeric viruses was analyzed with the Sequence Demarcation Tool (SDT v1.2) (Muhire, Varsani, and Martin 2014). The similarity between the CPs of chimeric viruses was further investigated in the context of their 3D structure. To this end, sequence similarity was mapped onto the previously generated 3D model of the CP of CHIV10 (Roux et al. 2013) using UCSF Chimera (Pettersen et al. 2004).

2.3. Virome read mapping to chimeric viruses

We compiled a database of full-length CP sequences (total of sixty-eight sequences) and attributed group affiliations according to the phylogenetic analysis. BLASTx analysis against the CP sequences was performed for the twelve *Sphagnum*-peat viromes and sixty-nine viromes from public databases that were equally obtained by whole genome amplification. To assure the reliability of matches a strict cut-off value of 10^{-10} was applied. The number of matches was normalized to the total number of reads in the different viromes.

2.4. Accession numbers

Chimeric viruses genomes reported in this publication have been deposited in GenBank under the accession numbers KX388494-KX388530 and in MetaVir Project id: 8195.

3. Results and Discussion

3.1. Structure and genome content of complete chimeric genomes

To analyze the genome content and structure of chimeric viruses, sequences from twelve peat viromes were assembled separately into contigs as described previously (Quaiser et al. 2015). To enrich for small ssDNA viral genomes, we used whole genome amplification, known to amplify preferentially small circular nucleic acid templates (Kim and Bae 2011). Given that all known chimeric viruses carry relatively large genomes compared with those of other known eukaryotic ssDNA viruses (Roux et al. 2013), only contigs larger than 3,000 bp were considered (Supplementary Table S1). In total, twenty-two contigs from Bog samples and fifteen contigs from Fen samples that encoded for a replication protein linked to a CHIV-like CP encoding gene were identified. The circularity of the sequences was verified using MUMmer for

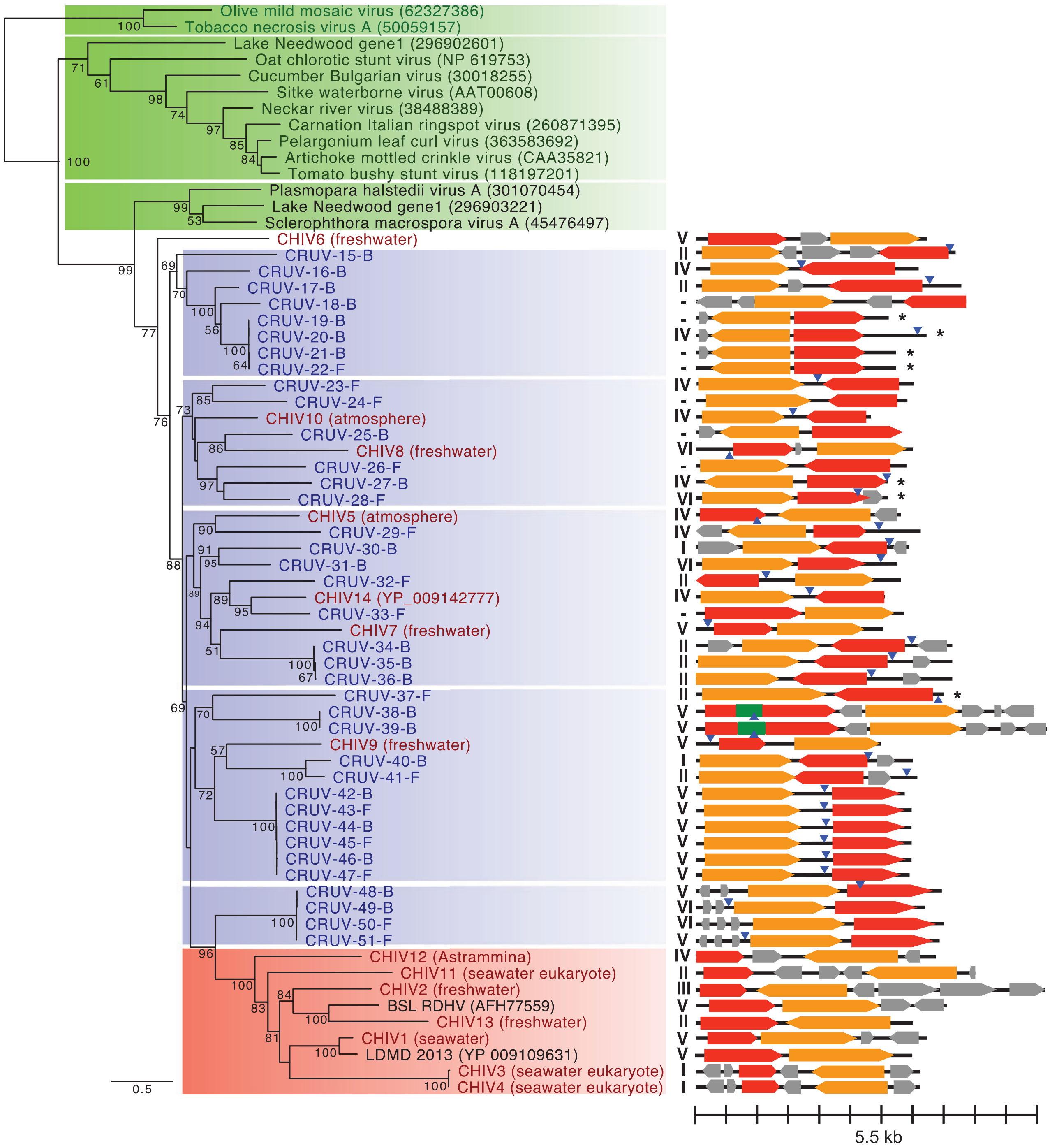

Figure 1. Maximum-likelihood phylogenetic analysis of full-length CP sequences and associated genome structure of crucivirus-like genomes from *Sphagnum*-dominated peat viromes. A total of 333 unambiguously aligned positions from sixty-seven sequences were used in the analysis. Bootstrap values above 50 are indicated at the nodes. The scale bar indicates the number of substitutions per position for a unit branch length. Three groups are highlighted by different background colors. Green: ssRNA viruses (*Tombusviridae*); blue: peat origin; red: RDHV or chimeric viruses origin. Genome organizations are drawn to scale. Red: replication protein (RC-Rep); yellow: CP; green: intron in RC-Rep, grey: other ORFs. Blue triangle: potential replication origin. Star (*): linear genomes. I–VI: CRESS virus types.

sequence recruitment to reconstructed genomes (Quaiser et al. 2015). In total, twenty-nine from thirty-seven contigs could be closed and were considered as circular (Fig. 1).

All these assembled genomes contained genes that encode for two conserved proteins, tombusvirus-like CP and RC-Rep, as typical of chimeric viruses (Roux et al. 2013). No other genes, beside CP and RC-Rep were conserved throughout the genomes. Additional ORFs were shared only by nearly identical contigs. The comparison of these ORFs to protein sequence databases revealed no significant matches. Although official classification of these viruses will require isolation of cultivable representatives, we propose referring to this viral assemblage as 'Cruciviridae' (crucis: cross in Latin) as these viruses originate from a "cross" (i.e., recombination) between two widely different groups of viruses. The new name is advantageous in that it avoids the vagueness and negative connotation of such terms as “hybrid” and “chimeric”, both of which are widely used in the fields of Genetics and Molecular Biology.

The genome names were thus constructed using the abbreviated virus group name (CRUV) followed by a number and by the letters B (for bog) or F (for fen) indicating their origin. Four of the open contigs, deriving from different samples, show nearly identical sequences (CRUV19–22) and no contig in this group could be circularized indicating that these could represent the first cruciviruses with linear genomes. However, no terminal inverted repeats typical of linear ssDNA viral genomes (e.g., parvoviruses) could be identified. Thus, the actual structure of these genomes remains uncertain. The number of ORFs per genome varied from two to seven. The genomic organization (i.e. gene order, gene direction, and gene composition) varied considerably and was only conserved in nearly identical genomes (Fig. 1). Putative introns were identified in the RC-Rep gene of contigs CRUV-38-B and CRUV-39-B. In both cases, the conserved splicing motifs 5′ XGT and AG 3′ are present. In addition, the RCR motifs I–II and III are separated by this putative intron. While introns in replication proteins of *Geminiviridae* (Wright et al. 1997; Bernardo et al. 2013) as well as in certain circular replication associated protein (Rep)-encoding single stranded DNA viruses (CRESS) (Dayaram et al. 2016; Male et al. 2016) were found before, this is the first evidence of potential introns in cruciviruses. Together, these characteristics suggest that we identified new viral genomes from the ‘Cruciviridae’ group.

3.2. Diversity of cruciviruses based on phylogenetic analysis of the capsid protein

The affiliation of newly assembled contigs to the ‘Cruciviridae’ group was first verified by phylogenetic analysis of the ORFs encoding for the CP. The thirty-seven peat derived CPs (CRUV15–51) and sixteen previously identified sequences affiliated to chimeric viruses (Supplementary Table S1) as well as ssRNA virus CP homologs were included in a phylogenetic tree (Fig. 1). The CPs from the peat-derived viruses were most closely related to those of cruciviruses identified in other environments (Roux et al. 2013). In the CP phylogeny, all cruciviruses formed a monophyletic clade, which branched as a sister group to the ssRNA viruses infecting oomycetes, namely *Plasmopara halstedii* virus A and *Sclerophthora macrospora* virus A. Within the ‘Cruciviridae’ branch, several distinct clades can be defined. A large group is formed by eight previously described chimeric viruses sequences (CHIV1–4, CHIV11–13) from aquatic environments, and also includes BSL_RDHV and LDMD-2013 (Fig. 1, red). The CPs of peat-derived cruciviruses (CRUV) form five distinct subclades but with relatively low bootstrap support (~70%) (Fig. 1, blue). Three of these subclades, besides the peat-derived cruciviruses, included six sequences from other environments. CHIV6 was found at the base of the ‘Cruciviridae’ clade and did not group with other sequences. Notably, several subgroups contain nearly identical crucivirus sequences. However, given that none of these genomes is derived from the same replicate, we can exclude potential assembly biases. This consistent detection of nearly identical cruciviruses across different samples suggests that the hosts of these viruses, although unknown, are likely abundant and permanently present in Fen and Bog peat.

3.3. Structure and conservation of Cruciviridae capsid protein

Since the tombusvirus-like CP represents the distinguishing feature of cruciviruses, which allows discriminating from other CRESS-DNA viruses, these CPs were analyzed in more detail. The length of the CPs from different members of the ‘Cruciviridae’ group was rather uniform (466aa ± 42), in line with previous observations (Roux et al. 2013). Similar to tombusviruses and the two oomycete-infecting ssRNA viruses, the CPs from cruciviruses display the characteristic 3-domain organization. The S-domain is involved in the formation of the icosahedral capsid, while the P-domain, less conserved, might be involved in virus–host interaction, and the R-domain likely interacts with the encapsidated viral genome (Roux et al. 2013). The level of divergence of CPs was analyzed by pairwise comparison of all available crucivirus CPs, and mapped on the previously obtained model of CHIV10 (Roux et al. 2013). Pairwise comparison of the full-length protein sequences showed a low level of conservation (Fig. 2, Supplementary Fig. S1) with an average of 34.1 percent identity (ranging from 21.6 to 77.7%) including all non-identical chimeric viruses and 36.9 percent identity (ranging from 28.3 to 77.7%) among peat-derived cruciviruses. However, sequence conservation was not uniformly distributed along the CP length. As previously observed (Diemer and Stedman 2012; Krupovic et al. 2015), the S-domain from peat-derived CPs showed higher sequence conservation compared with the P-domain (Fig. 2, Supplementary Fig. S1), consistent with the functional prediction. Average identity between the S-domains was 44.4 percent (ranging from 21.4 to 90.2%) for all non-identical chimeric viruses and 49.1 percent (ranging from 39.2 to 89.5%) for peat-derived cruciviral genomes. Interestingly, even chimeric viruses from the same samples, excluding near-identical genomes, showed high divergence, with minimum identity score of 39.2 percent (39.2–75.5%; average 49.2%) indicating a high level of diversity even within a single sample.

3.4. Second layer of chimerism in the cruciviral genomes

While the CPs of the cruciviruses form a monophyletic clade, their RC-Reps are polyphyletic and related to three different groups of eukaryotic ssDNA viruses, the *Geminiviridae*, *Nanoviridae*, or *Circoviridae* (Roux et al. 2013; Krupovic et al. 2015). In addition, detailed protein domain analysis of the RC-Rep revealed that even the two major protein domains, the N-terminal endonuclease domain and the C-terminal SF3 helicase domain, can have different origins within the same protein indicating intra-gene chimerism (Krupovic et al. 2015).

The phylogenetic analysis of the full-length RC-Reps from the peat-derived cruciviruses confirmed that they include sequences affiliated to the three different ssDNA virus families (Fig. 3). Domain-specific phylogenetic analysis showed that in eight of the thirty-seven RC-Reps from the peat cruciviruses, the N-terminal endonuclease and C-terminal SF3 helicase domains are affiliated to different virus families (Fig. 3, Supplementary Fig. S2). Since the branching in particular in the N-terminal endonuclease domain phylogenetic tree is not well supported, the affiliation of several peat-derived RC-Rep domains could not be unequivocally determined. Nevertheless, the N- and C-terminal domains of several RC-Reps display high degree of sequence identity to the homologous domains of RC-Reps from distinct viral families. For instance, the N-terminal nuclease and the C-terminal helicase domains of CRUV-26-F, CRUV-27-B, and CRUV-28-F are respectively affiliated

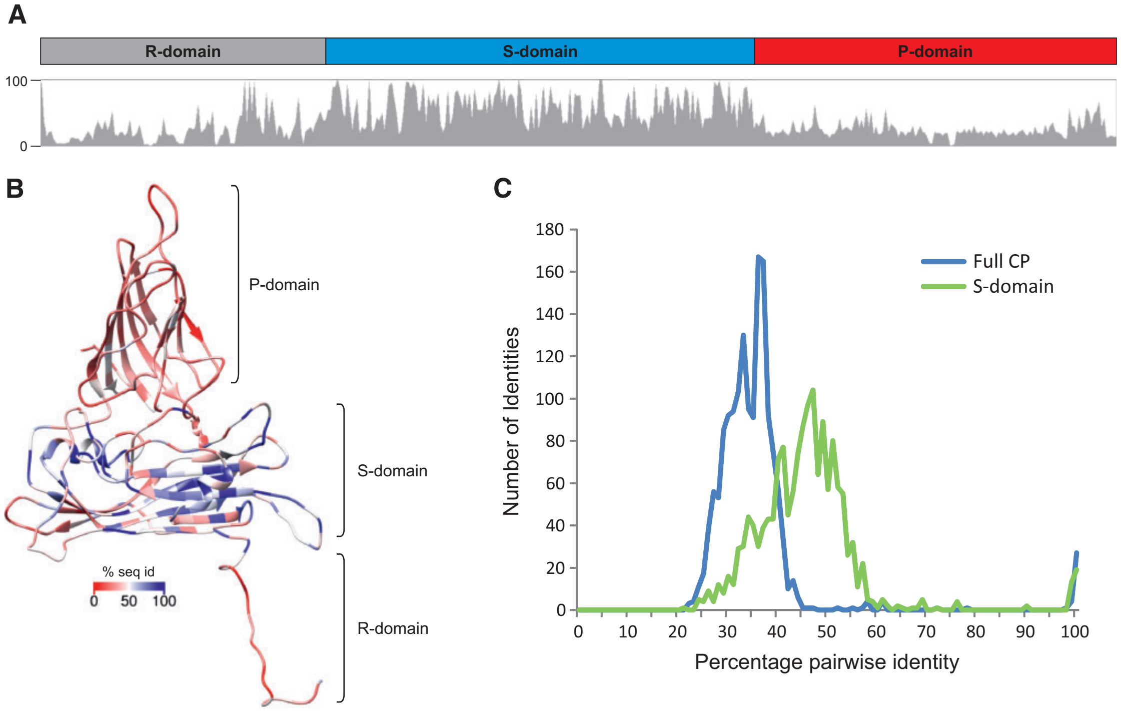

Figure 2. Characterization of CP diversity. (A) Domain organization of the tombusvirus-like CPs of chimeric viruses. The nucleic acid binding (R), shell (S), and projection (P) domains are indicated. The pattern of sequence conservation of fifty-six CP sequences is shown underneath the schematic domain organization cartoon. (B) Structural model of CHIV10 (Roux et al. 2013) showing the unequal distribution of sequence conservation in the context of the protein structure. The color key for the sequence identity is provided underneath the 3D model. (C) Comparison of the sequence conservation within the S-domain with that of the full-length CPs.

to the N- and C-terminal domains of *Circoviridae* and *Nanoviridae*, whereas the opposite is observed in CRUV19-22 and CRUV-32-F where the two domains are affiliated to *Nanoviridae* and *Circoviridae*, respectively. These results suggest that at least some of the peat-derived RC-Reps are chimeric with different origins of the endonuclease and the helicase domain.

3.5. Relative abundance of cruciviruses in peatland and other habitats

The quantification of viruses is a necessary first step for evaluation of their ecological impact on a microbial community in an ecosystem. The enumeration of viral particles and cells is typically achieved by epifluorescence microscopy. However, this approach can only produce bulk measurements while the quantification of specific viral groups remains very difficult. On the other hand, when virome sequences are available, the detection of signature sequences, such as CPs for chimeric viruses, can be used as a proxy for comparisons of relative abundance of viral groups in different ecosystems. Although the use of whole-genome amplification is known to introduce a bias towards small circular ssDNA templates (Kim and Bae 2011), the bias is expected to be uniform for viruses of the same group, such as cruciviruses in this case. Consequently, we consider that, despite the WGA step, comparison of the relative abundance of cruciviruses in different environments can provide biologically relevant information.

To estimate the abundance of cruciviruses in the twelve peat viromes and in sixty-nine available viromes from other environments, all the viromes were searched for crucivirus-like CP homologs with BLASTx. Best matches were counted using strict count conditions (*e*-value 10^{-10}) and normalized by the total number of sequences in each virome (Fig. 4). The affiliation to the putative chimeric groups was determined according to the four groups established in the CP phylogenetic analysis (Fig. 1), namely ssRNA viruses, CHIV6, BSL/CHIV and the peat group. On average, 4.13 percent of the twelve peat virome sequences matched to the CP from the 'Cruciviridae' group, representing 46,012 matches in total. Most of these (99%) matched to the peat group, 0.47 percent (216 matches) to the ssRNA viruses, 0.08% to CHIV6 (thirty-six matches) and 0.45 percent (209 matches) to BSL/CHIV group. Considerable variations were observed between the peat viromes with matches ranging from 0.6 percent in vBog_Oct11 to 10.9 percent in vBog_June11. Normalization to the average size of chimeric viruses genomes suggests that members of the 'Cruciviridae' could on average represent 11.37 percent of reads in the twelve peat viromes (4.13% × 2.75 factor, i.e. ratio of average genome size to average capsid gene size) and up to 30 percent in the vBog_June11 virome. Of the sixty-nine viromes from other environments, forty-five did not show matches to cruciviruses (Supplementary Table S2), whereas in the twenty-four remaining viromes, crucivirus CP matches accounted on average for only 0.2 percent, with the highest number of hits in the Airborne rain virome (2.9%), from which CHIV5 and CHIV10 were previously assembled (Roux et al. 2013). In the lake Bourget virome, the origin of CHIV6 and CHIV7, and the lake Pavin virome (origin of CHIV2, CHIV8, CHIV13), cruciviruses accounted for 0.34 and 0.22 percent or the reads, respectively. In the virome RW_Nursery_DNA, the origin of CHIV9, these viruses accounted for 0.02 percent of the reads.

Members of the 'Cruciviridae' thus appear to be highly abundant in *Sphagnum*-peat samples, representing the most copious single viral group identified. This suggests that, in contrast to other habitats where they represent only a minority, cruciviruses might play an important role in these ecosystems that are driven by temporal succession of the engineer species *Sphagnum* spp.

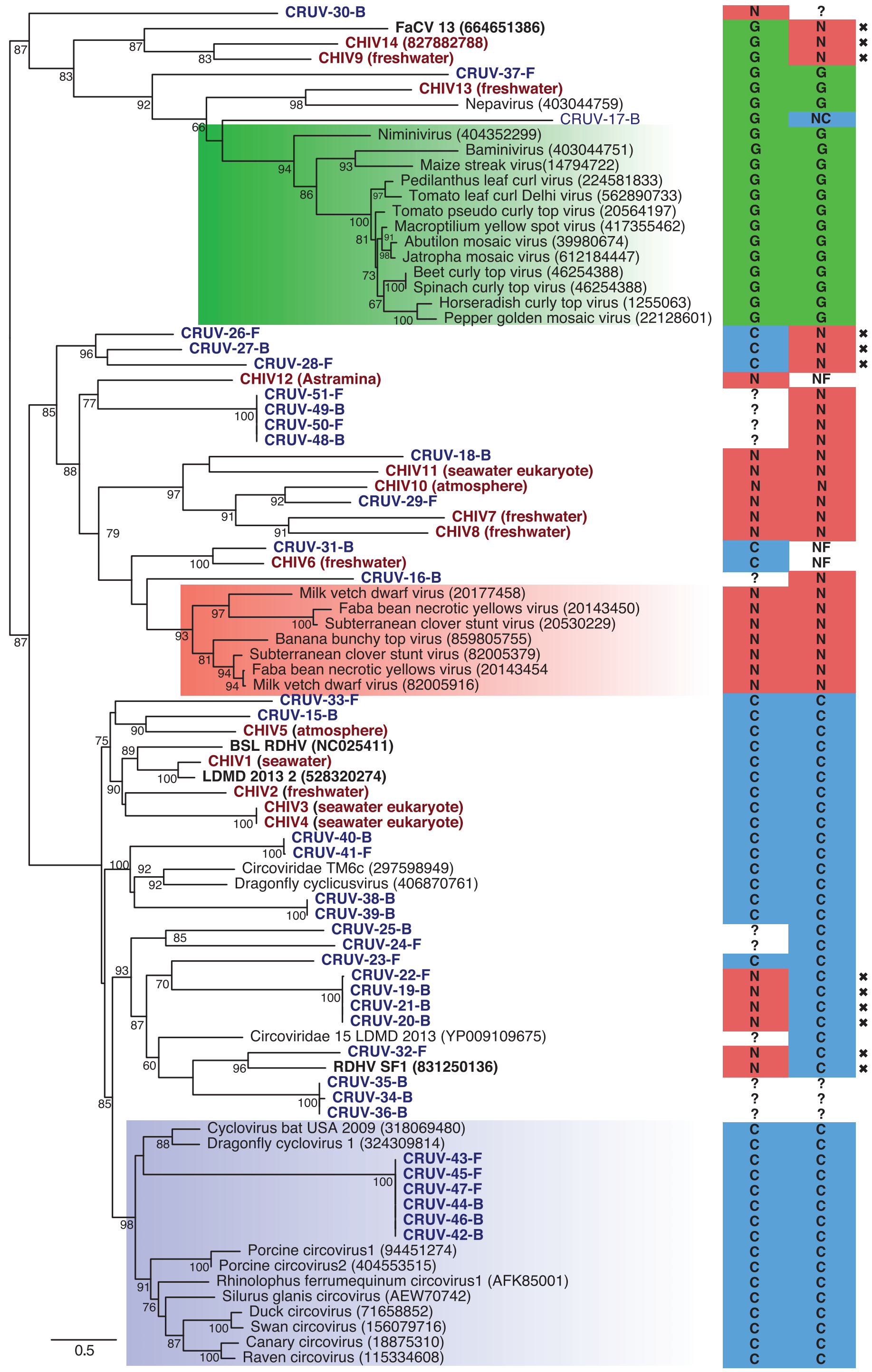
CRUV-30-B N ?
FaCV 13 (664651386) G N ✖
CHIV14 (827882788) G N ✖
CHIV9 (freshwater) G N ✖
CRUV-37-F G G
CHIV13 (freshwater) G G
Nepavirus (403044759) G G
CRUV-17-B G NC
Niminivirus (404352299) G G
Baminivirus (403044751) G G
Maize streak virus(14794722) G G
Pedilanthus leaf curl virus (224581833) G G
Tomato leaf curl Delhi virus (562890733) G G
Tomato pseudo curly top virus (20564197) G G
Macroptilium yellow spot virus (417355462) G G
Abutilon mosaic virus (39980674) G G
Jatropha mosaic virus (612184447) G G
Beet curly top virus (46254388) G G
Spinach curly top virus (46254388) G G
Horseradish curly top virus (1255063) G G
Pepper golden mosaic virus (22128601) G G
CRUV-26-F C N ✖
CRUV-27-B C N ✖
CRUV-28-F C N ✖
CHIV12 (Astramina) N NF
CRUV-51-F ? N
CRUV-49-B ? N
CRUV-50-F ? N
CRUV-48-B ? N
CRUV-18-B N N
CHIV11 (seawater eukaryote) N N
CHIV10 (atmosphere) N N
CRUV-29-F N N
CHIV7 (freshwater) N N
CHIV8 (freshwater) N N
CRUV-31-B C NF
CHIV6 (freshwater) C NF
CRUV-16-B ? N
Milk vetch dwarf virus (20177458) N N
Faba bean necrotic yellows virus (20143450) N N
Subterranean clover stunt virus (20530229) N N
Banana bunchy top virus (859805755) N N
Subterranean clover stunt virus (82005379) N N
Faba bean necrotic yellows virus (20143454 N N
Milk vetch dwarf virus (82005916) N N
CRUV-33-F C C
CRUV-15-B C C
CHIV5 (atmosphere) C C
BSL RDHV (NC025411) C C
CHIV1 (seawater) C C
LDMD 2013 2 (528320274) C C
CHIV2 (freshwater) C C
CHIV3 (seawater eukaryote) C C
CHIV4 (seawater eukaryote) C C
CRUV-40-B C C
CRUV-41-F C C
Circoviridae TM6c (297598949) C C
Dragonfly cyclicusvirus (406870761) C C
CRUV-38-B C C
CRUV-39-B C C
CRUV-25-B ? C
CRUV-24-F ? C
CRUV-23-F C C
CRUV-22-F N C ✖
CRUV-19-B N C ✖
CRUV-21-B N C ✖
CRUV-20-B N C ✖
Circoviridae 15 LDMD 2013 (YP009109675) ? C
CRUV-32-F N C ✖
RDHV SF1 (831250136) N C ✖
CRUV-35-B ? ?
CRUV-34-B ? ?
CRUV-36-B ? ?
Cyclovirus bat USA 2009 (318069480) C C
Dragonfly cyclovirus 1 (324309814) C C
CRUV-43-F C C
CRUV-45-F C C
CRUV-47-F C C
CRUV-44-B C C
CRUV-46-B C C
CRUV-42-B C C
Porcine circovirus1 (94451274) C C
Porcine circovirus2 (404553515) C C
Rhinolophus ferrumequinum circovirus1 (AFK85001) C C
Silurus glanis circovirus (AEW70742) C C
Duck circovirus (71658852) C C
Swan circovirus (156079716) C C
Canary circovirus (18875310) C C
Raven circovirus (115334608) C C
87 87 83 83 92 98 66 94 93 86 97 100 91 81 98 73 100 67 100 96 85 77 100 88 97 92 91 91 79 100 97 100 87 93 81 94 94 90 75 89 100 90 100 100 92 92 100 85 93 70 100 87 96 60 100 85 88 100 98 100 91 76 100 87 100
0.5

Figure 3. Maximum-likelihood phylogenetic analysis of full-length replication protein sequences and associated endonuclease and helicase domain affiliation. A total of 207 unambiguously aligned positions from eighty-nine sequences were used in the full-length phylogenetic analysis. Bootstrap values above 50 are indicated at the nodes. The scale bar indicates the number of substitutions per position for a unit branch length. Three groups are highlighted by different background colors. Green: *Geminiviridae*; blue: *Circoviridae*; red: *Nanoviridae*. Columns indicate the closest relative of the endonuclease and the helicase protein domains as determined by domain specific phylogenetic analysis. C: *Circoviridae*-like domain (blue), G: *Geminivirus*-like domain (green), N: *Nanoviridae*-like domain (red). Bold: all confirmed crucivirus-like viral sequences. Bold blue: chimere-like viral sequences from peat samples. X: potential chimeric replication protein. NF: Non-functional Walker B motif in the SF3 helicase domain.

4. Conclusion

The lack of universal viral marker gene and the high viral genetic diversity limit our current understanding of environmental viral diversity. Here we characterized the genomes of thirty-seven new members of the 'Cruciviridae' from peat soil water, thereby tripling the number of genomes for this new group of viruses. Their affiliation to the 'Cruciviridae' group was shown by several characteristics: (a) they encode tombusvirus-like CPs and RC-

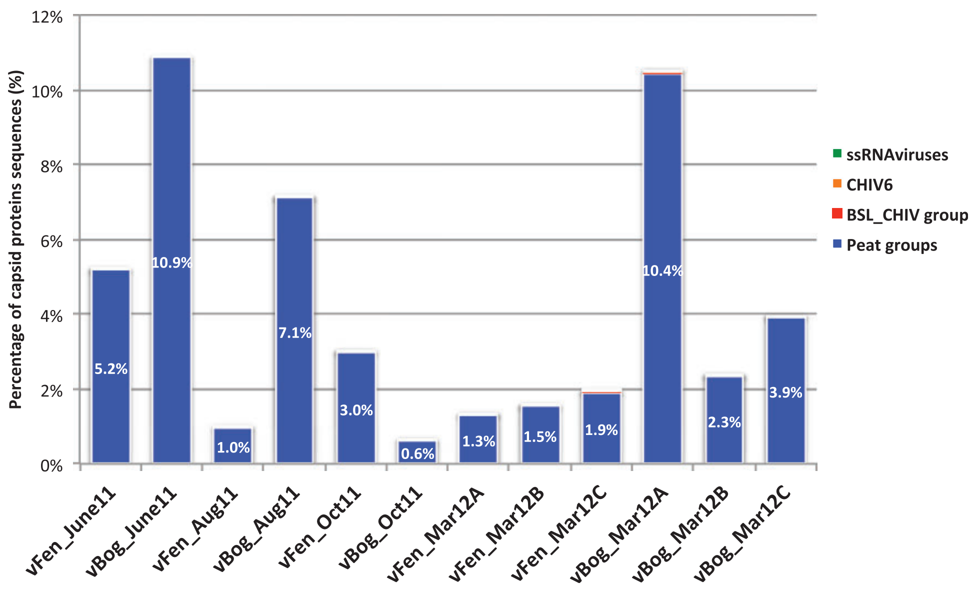

Figure 4. Relative proportions of crucivirus-like CP encoding sequences in six fen and six bog viromes. The viromes were analyzed by BLASTx against sixty-eight CP sequences. Best matches were counted and normalized to the number of sequences in each virome. The grouping is based on the phylogenetic affiliation of the CPs.

Reps homologous to the corresponding proteins of ssDNA viruses from families *Circoviridae*, *Nanoviridae*, and *Geminiviridae*, (b) several RC-Reps are chimeric with respect to the endonuclease and helicase domains, (c) the genome size is significantly larger than that of other CRESS-DNA viruses that possess RC-Rep homologs. The detection of several chimeric RC-Reps is consistent with previous observations (Krupovic et al. 2015) and indicates that RC-Rep domain shuffling is relatively widespread within the 'Cruciviridae'. Our results indicate a high diversity of cruciviruses even within a single sample. When considered alongside the multiple types of RC-Rep genes found associated with the CP genes in the cruciviral genomes (Roux et al. 2013), such sequence- and genome-level diversity might indicate that shortly after the emergence of the ancestral chimeric viruses, the 'Cruciviridae' group has experienced an evolutionary radiation event which led to the emergence of the contemporary diversity of these viruses. An important finding of this study is the unexpectedly high abundance of cruciviruses in all peat samples, both fen and bog. This additional description of genomic and genetic diversity within the 'Cruciviridae' as well as identification of the preferential habitat of these viruses will help to identify their host(s), which to date remain unknown.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Acknowledgments

We thank members of the 'Environmental and Functional Genomics Platform' (Biogenouest Genomics) for sequencing and the molecular ecology facility (UMR 6553 ECOBIO) for support. We thank Sophie Coudouel (OSU Rennes) and Flore Ballaud for technical assistance.

Funding

This study was supported by funding from the INSU-CNRS EC2CO program to A.-J.F. ("Virtou") (Program MICROBIEN, 2011–2012) and grants from the University of Rennes/CNRS and from the "Observatoire des Sciences de l'Univers de Rennes"(OSUR) to A.Q.

Conflict of interest: None declared.

References

Ballaud, F. C., et al. (2016) 'Dynamics of Viral Abundance and Diversity in a Sphagnum-Dominated Peatland: Temporal Fluctuations Prevail Over Habitat', *Frontiers in Microbiology*, 6: 1494.

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

Colombet, J., et al. (2007) 'Virioplankton 'Pegylation': Use of PEG (Polyethylene Glycol) to Concentrate and Purify Viruses in Pelagic Ecosystems', *Journal of Microbiological Methods*, 71: 212–9.

Dayaram, A., et al. (2016) 'Diverse Circular Replication-Associated Protein Encoding Viruses Circulating in Invertebrates Within a Lake Ecosystem', *Infection, Genetics and Evolution*, 39: 304–16.

Diemer, G. S. and Stedman, K. M. (2012) 'A Novel Virus Genome Discovered in an Extreme Environment Suggests Recombination Between Unrelated Groups of RNA and DNA Viruses', *Biology Direct*, 7: 13.

Edgar, R. C. (2004) 'MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput', *Nucleic Acids Research*, 32: 1792–7.

Fuhrman, J. A. (1999) 'Marine Viruses and Their Biogeochemical and Ecological Effects', *Nature*, 399: 541–8.

Guindon, S., et al. (2010) 'New Alogrithms and Methods to Estimate Maximum-Likelihoods Phylogenies: Assessing the Performance of PhyML 3.0', *Systematic Biology*, 59: 307–21.

Hewson, I., et al. (2013) 'Metagenomic Identification, Seasonal Dynamics, and Potential Transmission Mechanisms of a Daphnia-Associated Single-Stranded DNA Virus in Two Temperate Lakes', *Limnology and Oceanography*, 58: 1605–20.

Jobb, G., von Haeseler, A., and Strimmer, K. (2004) 'TREEFINDER: a Powerful Graphical Analysis Environment for Molecular Phylogenetics', *BMC Evolutionary Biology*, 4: 18.

Kim, K.-H. and Bae, J.-W. (2011) 'Amplification Methods Bias Metagenomic Libraries of Uncultured Single-Stranded And Double-Stranded DNA Viruses', *Applied and Environmental Microbiology*, 77: 7663–8.

Kraberger, S., et al. (2015) 'Characterisation of a Diverse Range of Circular Replication-Associated Protein Encoding DNA Viruses Recovered From a Sewage Treatment Oxidation Pond', *Infection, Genetics and Evolution*, 31: 73–86.

Krupovic, M. (2013) 'Networks of Evolutionary Interactions Underlying the Polyphyletic Origin of ssDNA Viruses', *Current Opinion in Virology*, 3: 578–86.

—— and Forterre, P. (2015) 'Single-Stranded DNA Viruses Employ a Variety of Mechanisms for Integration into Host Genomes', *Annals of the New York Academy of Sciences*, 1341: 41–53.

——, et al. (2015) 'Multiple Layers of Chimerism in a Single-Stranded DNA Virus Discovered by Deep Sequencing', *Genome Biology and Evolution*, 7: 993–1001.

——, Ghabrial, S. A., Jiang, D., and Varsani, A. (2016) 'Genomoviridae: a New Family of Widespread Single-Stranded DNA Viruses', *Archives of Virology*, 161: 2633.

Kurtz, S., et al. (2004) 'Versatile and Open Software for Comparing Large Genomes', *Genome Biology*, 5: R12.

Le Bras, Y., et al. (2013). 'Towards a Life Sciences Virtual Research Environment'. *e-biogenouest.org*. https://www.e-biogenouest.org/resources/129/download/jobim_YLeBras_2013.pdf.

Ludwig, W., et al. (2004) 'ARB: A Software Environment for Sequence Data', *Nucleic Acids Research*, 32: 1363–71.

Male, M. F., et al. (2016) 'Cycloviruses, Gemycircularviruses and Other Novel Replication-Associated Protein Encoding Circular Viruses in Pacific Flying Fox (*Pteropus tonganus*) Faeces', *Infection, Genetics and Evolution*, 39: 279–92.

Mcdaniel, L. D., et al. (2014) 'Comparative Metagenomics: Natural Populations of Induced Prophages Demonstrate Highly Unique, Lower Diversity Viral Sequences', *Environmental Microbiology*, 16: 570–85.

Muhire, B. M., Varsani, A., and Martin, D. P. (2014) 'SDT: A Virus Classification Tool Based on Pairwise Sequence Alignment and Identity Calculation', *PLoS One*, 9: e108277.

Pei, J., Kim, B. H., and Grishin, N. V. (2008) 'PROMALS3D: A Tool for Multiple Protein Sequence and Structure Alignments', *Nucleic Acids Research*, 36: 2295–300.

Pettersen, E. F., et al. (2004) 'UCSF Chimera—a Visualization System for Exploratory Research and Analysis', *Journal of Computational Chemistry*, 25: 1605–12.

Quaiser, A., et al. (2015) 'Diversity and Comparative Genomics of Microviridae in Sphagnum-Dominated Peatlands', *Frontiers in Microbiology*, 6: 375.

Rosario, K., Duffy, S., and Breitbart, M. (2012) 'A Field Guide to Eukaryotic Circular Single-Stranded DNA Viruses: Insights Gained From Metagenomics', *Archives of Virology*, 157: 1851–71.

Roux, S., et al. (2013) 'Chimeric Viruses Blur the Borders Between the Major Groups of Eukaryotic Single-Stranded DNA Viruses', *Nature Communications*, 4: 2700.

Stedman, K. M. (2015) 'Deep Recombination: RNA and ssDNA Virus Genes in DNA Virus and Host Genomes', *Annual Review of Virology*, 2: 203–17.

Steel, O., et al. (2016) 'Circular Replication-Associated Protein Encoding DNA Viruses Identified in the Faecal Matter of Various Animals in New Zealand', *Infection, Genetics and Evolution*, 43: 151–64.

Suttle, C. A. (2007) 'Marine Viruses—Major Players in the Global Ecosystem', *Nature Reviews Microbiology*, 5: 801–12.

Wommack, K. E. and Colwell, R. R. (2000) 'Virioplankton: Viruses in Aquatic Ecosystems', *Microbiology and Molecular Biology Reviews*, 64: 69–114.

Wright, E. A., et al. (1997) 'Splicing Features in Maize Streak Virus Virion- and Complementary-Sense Gene Expression', *The Plant Journal*, 12: 1285–97.