Unveiling Crucivirus Diversity by Mining Metagenomic Data

Ignacio de la Higuera[1], George W. Kasun[1], Ellis L. Torrance[1], Alyssa A. Pratt[1], Amberlee Maluenda[1], Jonathan Colombet[2], Maxime Bisseux[2], Viviane Ravet[2], Anisha Dayaram[3], Daisy Stainton[4], Simona Kraberger[5], Peyman Zawar-Reza[6], Sharyn Goldstien[7], James V. Briskie[7], Robyn White[7], Helen Taylor[8], Christopher Gomez[9], David G. Ainley[10], Jon S. Harding[7], Rafaela S. Fontenele[5], Joshua Schreck[5], Simone G. Ribeiro[11], Stephen A. Oswald[12], Jennifer M. Arnold[12], François Enault[2], Arvind Varsani[5, 7, 13] and Kenneth M. Stedman[1].

[1]Department of Biology, Center for Life in Extreme Environments, Portland State University, P.O. Box 751, Portland, OR 97207-0751, USA

[2]Université Clermont Auvergne, CNRS, Laboratoire Microorganismes: Génome et Environnement, UMR 6023, Clermont–Ferrand, France

[3]Institut für Neurophysiology, Charité-Universitätsmedizin, Charitéplatz 1, Berlin 10117, Germany

[4]Department of Entomology and Plant Pathology, Division of Agriculture, University of Arkansas System, Fayetteville, AR 72701, USA

[5]The Biodesign Center for Fundamental and Applied Microbiomics, Center for Evolution and Medicine, School of Life Sciences, Arizona State University, Tempe, AZ 85287, USA

[6]School of Earth and Environment, University of Canterbury, Christchurch, New Zealand

[7]School of Biological Sciences, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand

[8]Department of Anatomy, University of Otago, Lindo Ferguson Building, Great King Street, Dunedin, 9016, New Zealand

[9]Graduate School of Maritime Sciences, Laboratory of Sediment Hazards and Disaster Risk, Kobe University, Kobe City, Japan

[10]HT Harvey and Associates, Los Gatos, CA 95032, USA

[11]Embrapa Recursos Genéticos e Biotecnologia, Brasília, DF 70770-017, Brazil

[12]Division of Science, Pennsylvania State University, Berks Campus, Reading, PA 19619, USA

[13]Structural Biology Research Unit, Department of Clinical Laboratory Sciences, University of Cape Town, Rondebosch, Cape Town, South Africa

Correspondence should be addressed to Kenneth M. Stedman: kstedman@pdx.edu

KEYWORDS: Crucivirus, CRESS-DNA viruses, gene transfer, recombination, virus evolution

ABBREVIATIONS

CruV: Crucivirus; CRESS: Circular Rep-Encoding Single Stranded; CP: Capsid; RdRP: RNA-dependent RNA Polymerase; ORF: Open Reading Frame; SJR: Single Jelly Roll; PI: Pairwise Identity; MDA: Multiple Displacement Amplification; S3H: Superfamily 3 helicase; RCR: Rolling Circle Replication; CDS: Coding Sequence; CruCGE: Cruci-like Circular Genetic Element

ABSTRACT

The discovery of cruciviruses revealed the most explicit example of a common protein homologue between DNA and RNA viruses to date. Cruciviruses are a novel group of circular Rep-encoding ssDNA (CRESS-DNA) viruses that encode capsid proteins (CPs) that are most closely related to those encoded by RNA viruses in the family *Tombusviridae*. The apparent chimeric nature of the two core proteins encoded by crucivirus genomes suggests horizontal gene transfer of CP genes between DNA and RNA viruses. Here, we identified and characterized 451 new crucivirus genomes and ten CP-encoding circular genetic elements through *de novo* assembly and mining of metagenomic data. These genomes are highly diverse, as demonstrated by sequence comparisons and phylogenetic analysis of subsets of the protein sequences they encode. Most of the variation is reflected in the replication associated protein (Rep) sequences, and much of the sequence diversity appears to be due to recombination. Our results suggest that recombination tends to occur more frequently among groups of cruciviruses with relatively similar capsid proteins, and that the exchange of Rep protein domains between cruciviruses is rarer than gene exchange. Altogether, we provide a comprehensive and descriptive characterization of cruciviruses.

IMPORTANCE

Viruses are the most abundant biological entities on Earth. In addition to their impact on animal and plant health, viruses have important roles in ecosystem dynamics as well as in the evolution of the biosphere. Circular Rep-encoding single-stranded (CRESS) DNA viruses are ubiquitous in nature, many are agriculturally important, and are viruses that appear to have multiple origins from prokaryotic plasmids. CRESS-DNA viruses such as the cruciviruses, have homologues of capsid proteins (CPs) encoded by RNA viruses. The genetic structure of cruciviruses attests to the transfer of capsid genes between disparate groups of viruses. However, the evolutionary history of cruciviruses is still unclear. By collecting and analyzing cruciviral sequence data, we provide a deeper insight into the evolutionary intricacies of cruciviruses. Our results reveal an unexpected diversity of this virus group, with frequent recombination as an important determinant of variability.

INTRODUCTION

In the last decade, metagenomics has allowed for the study of viruses from a new angle; viruses are not merely agents of disease, but abundant and diverse members of ecosystems (1, 2). Viruses have been shaping the biosphere probably since the origin of life, as they are important drivers of the evolution of the organisms they infect (3–5). However, the origin of viruses is not entirely clear. Viruses, as replicons and mobile elements, are also subject to evolution. Virus variability is driven by various mutation rates, recombination and reassortment of genetic components (6). These attributes, coupled with types of genomes (RNA or DNA, single or double stranded and circular or linear), lead to a large genetic diversity in the 'viral world'.

Viruses are generally classified based on the nature of their transmitted genetic material (7). Viral genetic information is coded in either RNA or DNA. Moreover, these genomes can be single (positive or negative sense) or double stranded, linear or circular, and can be comprised of a single or multiple molecules of nucleic acid (monopartite or multipartite, respectively). These different groups of viruses have different replication strategies, and they harbor distinct taxa based on their genome arrangement and composition (1). The striking differences between viral groups with disparate genome types suggest polyphyletic virus origins (8).

For example, the highly abundant circular Rep-encoding single-stranded DNA (CRESS-DNA) viruses may have been derived from plasmids on multiple occasions by acquiring capsid genes from RNA viruses (9–11). Eukaryotic CRESS-DNA viruses constitute a diverse and widespread group of viruses with circular genomes –some of them multipartite– that contains the families *Geminiviridae, Circoviridae*, *Nanoviridae*, *Alphasatellitidae*, *Genomoviridae*, *Bacilladnaviridae*, *Smacoviridae* and *Redondoviridae* (ICTV classification for some groups is pending at this time), in addition to vast numbers of unclassified viruses (12, 13). Universal to all CRESS-DNA viruses is the Rep, which is involved in the initiation of the virus' rolling-circle replication. Rep homologues are also encoded in plasmids (13, 14). Some pathogenic CRESS-DNA viruses are agriculturally important, such as porcine circoviruses, and nanoviruses and geminiviruses that infect a wide range of plant hosts (12). However, many CRESS-DNA viruses have been identified in apparently healthy organisms, and metagenomics have revealed their presence in most environments (12).

In 2012, a metagenomic survey of a hot and acidic lake in the volcanic Cascade Range of the western USA uncovered a new type of circular DNA virus (15). The genome of this virus is a CRESS-DNA virus based on the circularity of its sequence, the presence of a *rep* gene, and a predicted stem-loop structure with a conserved nucleotide sequence (*ori*) that serves as an origin for CRESS-DNA virus rolling-circle replication (RCR; reviewed in (16, 17)). Interestingly, the sequence of CP encoded by this genome resembles those encoded by the RNA viruses in the family *Tombusviridae* (15). It was hypothesized that this virus originated by the acquisition of a capsid gene from an RNA virus through a yet to be demonstrated RNA-DNA recombination event (15, 18). Since the discovery of this putatively "chimeric virus", 80 circular sequences encoding a Rep and a CP that share homology to tombusvirus CPs have been found in different environments around the globe (19, 20, 29–31, 21–28). This growing group of viruses have been branded "cruciviruses", as they imply the *crossing* between CRESS-DNA viruses and RNA tombusviruses (27). Cruciviruses have been found associated with forams (20), alveolates hosted by isopods (26), arthropods (19, 22), and in peatland ecosystems (27), but no host for cruciviruses have been elucidated to date.

The circular genome of known cruciviruses is variable in size, ranging from 2.7 to 5.7 kb and often contains ORFs in addition to the Rep and CP, which have been found in either a unisense or an ambisense orientation (20, 27). The function of additional crucivirus ORFs is unclear due to the lack of sequence similarity with any characterized protein. The genome replication of CRESS-DNA viruses is initiated by the Rep protein, that binds to direct repeats present just downstream of the stem of the *ori*-containing stem-loop structure and nicks the ssDNA (32, 33). The exposed 3'OH serves as a primer for cellular enzymes to replicate the viral genome via RCR (33–35). The exact terminating events of CRESS-DNA virus replication are poorly understood for most CRESS-DNA viruses, but Rep is known to be involved in the sealing of newly replicated genomes (33, 35–37).

Rep has a domain in the N-terminus that belongs to the HUH endonuclease superfamily (38). This family of proteins is characterized by a HUH motif (Motif II), in which two histidine residues are separated by a bulky hydrophobic amino acid, and a

Tyr-containing motif (Motif III) that catalyzes the nicking of the ssDNA (38–41). CRESS-DNA virus Reps also contain a third conserved motif in the N-terminal portion of the protein (Motif I), likely responsible for dsDNA binding specificity (42). In many CRESS-DNA viruses, the HUH motif has been substituted for a similar motif that lacks the second histidine residue (e.g. circoviruses have replaced HUH with HLQ) (10, 38). The C-terminal portion of eukaryotic CRESS-DNA virus Reps contain a Superfamily 3 helicase domain (S3H) that may be responsible for unwinding dsDNA replicative intermediates (43, 44). This helicase domain is characterized by Walker A and B motifs, Motif C and an Arg finger. Previous studies have identified evidence of recombination in the endonuclease and helicase domains of Rep, which contributes to the potential ambiguity of Rep phylogenies (45). Interestingly, the Rep proteins of different cruciviruses have been shown to be similar to CRESS-DNA viruses in different families, including circoviruses, nanoviruses, and geminiviruses (20, 27). In some cruciviruses, these differences in phylogeny have been observed between the individual domains of a single Rep protein (21, 27). The apparent polyphyly of crucivirus Reps suggests recombination events involving cruciviruses and other CRESS-DNA viruses, even within Reps (20, 21).

All characterized CRESS-DNA viruses package their DNA into small capsids with icosahedral symmetry or their geminate variants, built from multiple copies of the CP encoded in their genome (12). The CP of these CRESS-DNA viruses appears to fold into an eight-strand ß-barrel that conforms to the single jelly-roll (SJR) architecture, which is also commonly found in eukaryotic RNA viruses (46). The CP of cruciviruses has no detectable sequence similarity with the capsid of other CRESS-DNA viruses, and is predicted to adopt the SJR conformation found in the CP of tombusviruses (15, 20, 21). Three domains can be distinguished in tombusviral CPs (47, 48). From the N- to the C-terminus: i) The RNA-interacting or R-domain, a disordered region that faces the interior of the viral particle to interact with the nucleic acid through abundant basic residues (49, 50), ii) the shell or S-domain containing the single jelly-roll fold and the architectural base of the capsid (48) and iii) the protruding or P-domain, that decorates the surface of the virion and is involved in host transmission (51). In tombusviruses, the S-domain of 180 CP subunits interact with each other to assemble around the viral RNA in a T=3 fashion, forming a Ø~35 nm virion (48, 52).

The study of cruciviruses suggests evidence for the transfer of capsid genes between disparate viral groups, which can shed light on virus origins and the phenotypic plasticity of virus capsids. Here, we document the discovery of 461 new crucivirus genomes (CruV) and cruci-like circular genetic elements (CruCGE) identified in metagenomic data obtained from different environments and organisms. This study provides a comprehensive analysis of this greatly expanded dataset and explores the extent of cruciviral diversity –mostly due to Rep heterogeneity– impacted by rampant recombination.

MATERIALS AND METHODS

Recovery of viral genomes from assembled viromes.

A total of 461 crucivirus-related sequences were identified from 1168 metagenomic surveys (Supp. Tables 1 and 2). 1167 viromes from 57 published datasets and one unpublished virome were obtained from different types of environments: i) aquatic systems (freshwater, seawater, hypersaline ponds, thermal springs and hydrothermal vents), ii) engineered systems (bioreactor, food production), and iii) eukaryote-associated flora (human, insect and other animal feces, human saliva and fluids, cnidarians and plants). New cruciviral sequences were identified in these viromes by screening circular contigs for the presence of CPs from previously known cruciviruses (20) and tombusviruses, using a BLASTx bit-score threshold of 50.

Additionally, sequences CruV-240, CruV-300, CruV-331, CruV-338 and CruV-367 were retrieved from Joint Genome Institute (JGI)'s IMG/VR repository (53), by searching scaffolds with a function set including the protein family pfam00729, corresponding to the S-domain of tombusvirus capsids. The sequences with an RdRP coding region were excluded, and the circularity of the sequences, as well as the presence of an ORF encoding a tombusvirus-like capsid, were confirmed with Geneious 11.0.4 (Biomatters, Ltd).

Annotation of crucivirus putative genes.

The 461 cruciviral sequences were annotated and analyzed in Geneious 11.0.4. Coding sequences (CDSs) were semi-automatically annotated from a custom database (Supp. Table 3) of protein sequences of published cruciviruses and close homologs obtained from GenBank, using Geneious 11.0.4's annotation function with a 25% nucleotide similarity threshold. Annotated CDS were re-checked with GenBank database using BLASTx to identify sequences similar to previously described cruciviruses and putative relatives. Sequences containing in-frame stop codons were checked for putative splicing sites (54), or translated using a ciliate genetic code only when usage rendered a complete ORF with similarity to other putative crucivirus CDSs. Predicted ORFs longer than 300 bases with no obvious homologs and no overlap with CP or Rep-like ORFs were annotated as "putative ORFs".

Putative Stem-loop annotation.

Stem-loop structures that could serve as an origin of replication for circular ssDNA viruses were identified and annotated using StemLoop-Finder (Pratt *et* al., unpublished, (55, 56)). The 461 cruciviral sequences were scanned for the presence of conserved nonanucleotide motifs described for other CRESS-DNA viruses (NANTANTAN, NAKWRTTAC, TAWWDHWAN, & TRAKATTRC) (12). The integrated ViennaRNA 2.0 library was used to predict secondary structures of DNA around the detected motif, including the surrounding 15-20 nucleotides on either side (57, 58). Predicted structures with a stem longer than four base pairs and a loop including seven or more bases were subjected to the default scoring system, which increases the score by one point for each deviation from ideal stem lengths of 11 base pairs and loop lengths of 11 nucleotides. A set of annotations for stem-loops and nonanucleotides was created with StemLoop-Finder for those with a score of 15 or below. Putative stem-loops were excluded from annotation when a separate stem-loop was found with the same first base, but attained a greater score, as well as those that appeared to have a nonanucleotide within four bases of its stem-loop structure's first or last nucleotide.

Conservation analysis and visualization

Pairwise identity matrices. The pairwise identity (PI) between the protein sequence from translated cruciviral genes was calculated with SDTv1.2 (59), with MAFFT alignment option for CPs and S-domains, and MUSCLE alignment options for Reps.

Sequence conservation annotation. CP sequence conservation represented in Fig. 2A was generated with Jalview v2.11.0 (60), and reflects the conservation of the physicochemical properties for each column of the alignment (61).

Sequence logos. Sequence logos showing frequency of bases in nonanucleotides at the origin of replication or residue in conserved Rep motifs were made using the weblogo server (http://weblogo.threeplusone.com/; (62)).

Structural representation of capsid conservation. The 3D structure of CP was modeled with Phyre^2 (63). The generated graphic was colored by sequence conservation with Chimera v.1.13 (64), from the alignment of the 47 capsid sequences from each of the CP-based clusters (Fig. 3B).

Phylogenetic analyses.

Multiple sequence alignments. CP sequences were aligned using MAFFT (65) in Geneious 11.0.4, with a G-INS-i algorithm and BLOSUM 45 as exchange matrix, with an open gap penalty of 1.53 and an offset value of 0.123, and manually curated. Rep protein sequences were aligned using PSI-Coffee [http://tcoffee.crg.cat/; (66)]. Rep alignments were manually inspected and corrected in Geneious 11.0.4, and trimmed using TrimAI v1.3 with a *strict plus* setting (67). To produce individual alignments of the endonuclease and helicase domains the full length trimmed alignments were split at the Walker A motif (45).

Phylogenetic trees. Phylogenetic trees containing the entire dataset of cruciviral sequences were built on Geneious using the FastTree plugin (68). For the analysis of sequence subsets, trees were inferred with PhyML 3.0 web server [http://www.atgc-montpellier.fr/phyml/; (69)], using an aLRT SH-like support (70). The substitution model for each analysis was automatically selected by the program.

Intergene and interdomain comparison. Tanglegrams were made using Dendroscope v3.5.10 (71) to compare the phylogenies between different genes or domains within the same set of crucivirus genomes.

Sequence similarity networks. A total of 540 CP amino acid sequences, and 600 Rep amino acid sequences were uploaded to EFI–EST web server for the calculation of PIs [https://efi.igb.illinois.edu/efi-est/; (72)]. A specific alignment score cutoff was established for each dataset analyzed, and xgmml files generated by EFI-EST were visualized and edited in Cytoscape v3.7.2 (73).

Accession numbers

Provided in Supp. Table 1.

RESULTS & DISCUSSION

Expansion of the crucivirus group.

To broaden our understanding of the diversity and relationships of cruciviruses, 461 uncharacterized circular DNA sequences containing predicted CDSs with sequence similarity to the CP of tombusviruses were compiled from metagenomic sequencing data. The data came from published and unpublished metagenomic studies, carried out in a wide variety of environments, from permafrost to temperate lakes, and on various organisms from red algae to invertebrates (Metagenomes and their metadata are provided in Supp. Table 2). The selected genomes are assumed to be complete and circular based on the terminal redundancy identified in *de novo* assembled genomes.

The cruciviral sequences were named sequentially, beginning with the smallest genome, which was named CruV-81 to account for the 80 crucivirus genomes reported in prior literature (15, 19, 28–31, 20–27). The average GC content of the newly described cruciviral sequences is 42.9 ± 4.9 % (Fig. 1B) with genome lengths spanning from 2,474 to 7,947 bases (Fig. 1A), some exceeding the size of described bacilladnaviruses (≤6,000 nt (74)), the largest CRESS-DNA viruses known (12).

Of the 461 sequences that contain a CP ORF, 451 have putative coding regions with sequence similarity to Rep of CRESS-DNA viruses (10). The CP and Rep ORFs are encoded in a unisense orientation in 40% of the genomes and an ambisense orientation in 58% of the genomes. The remaining ~2% correspond to ten CruCGEs with no clear Rep CDS. Five of these CruCGEs contain a predicted origin of rolling-circle replication (RCR) (Supp. Table 1), indicating that they are circular genomes that undergo RCR characteristic of other CRESS-DNA virus genomes (16, 17).

One possible reason for the lack of a Rep ORF in certain sequences is that some of these may be sub-genomic molecules or possible components of multipartite viruses (75). Some CRESS-DNA viruses, such as geminiviruses and nanoviruses, have multipartite genomes (76). Moreover, some ssRNA tombunodaviruses; including *Plasmopara halstedii* virus A and *Schlerophthora macrospora* virus A –viruses that contain the most similar capsid sequences to cruciviral capsids (15, 27)– also have multipartite genomes (77). Unfortunately, no reliable method yet exists to match different sequences belonging to the same multisegmented virus in metagenomes, making identification of multipartite or segmented viruses from metagenomic data challenging (76).

Stem-loop structures with conserved nonanucleotide motifs as putative origins of replication were predicted and annotated in 277 cruciviral sequences with StemLoop-Finder (Pratt *et* al., unpublished). In some cases, more than one nonanucleotide motif with similar scores were found for a single genome, resulting in more than one stem-loop annotation. Of the annotated genomes, 223 contain a stem-loop with a nonanucleotide with a NANTANTAN pattern, with the most common sequence being the canonical circovirus motif TAGTATTAC, found in 64 of the genomes (Supp. Table 1; (78)). The majority of the 54 sequences that do not correspond to NANTANTAN contain a TAWWDHWAN nonanucleotide motif, typical of genomoviruses (79). The frequency of bases at each position in the nonanucleotide sequence is given in Fig. 1C, and reflects similarity to motifs found in other CRESS-DNA viruses (10).

Crucivirus capsid protein (CP)

The CP of cruciviruses is predicted to have a single jelly-roll (SJR) architecture, based on its homology to tombusvirus CPs for which 3D structures have been determined [Fig. 2A; (80–82)]. The SJR conformation is found in CPs of both RNA and DNA viruses (46). The SJR CP of tombusviruses and cruciviruses contains three distinct domains: the RNA-binding or R-domain, the shell or S-domain, and the protruding or P-domain (Fig. 2A). All 461 crucivirus CPs analyzed in this study contain a complete S-domain. This domain contains a distinct jelly-roll fold and interacts with the S-domain of other capsid subunits in the virion of related tombusviruses (48). The S-domain has greater sequence conservation than the remaining regions of the CP (Fig. 2A), likely due to its functional importance in capsid structure. In tombusviruses, the S-domain contains a calcium binding motif (DxDxxD), which was not identified in previously described cruciviruses (83). However, we detected this Ca-binding motif in 68 CPs of the newly identified cruciviral sequences. These crucivirus sequences form a distinct cluster, shown in red in Fig. 3B. The S-domain is flanked on the N-terminus by the R-domain, which in cruciviruses appears variable in size (up to 320 amino acids long), and appears to be truncated in some of the CP sequences (e.g. CruV-386 and CruV-493). The R-domain is characterized by an abundance of basic residues at the N-terminus, followed by a Gly-rich tract (Fig. 2A). The P-domain, on the C-terminal end of the CP sequence, is generally the largest domain, with the exception of CruV-385, where it appears to be truncated. The conservation of CP suggests a similar structure for all cruciviruses. However, those cruciviruses with larger genomes may assemble their capsids in a different arrangement to accommodate their genome. While the capsids of tombusviruses have been shown to adopt a T=1 icosahedral conformation, rather than the usual T=3, when the R-domain is partially or totally removed (82), we have not seen a correlation between the length of CP domains and genome size in our dataset that could be indicative of alternative capsid arrangements. Furthermore, no packaging dynamics relating genome size and virion T-number arrangement have been determined in CRESS-DNA viruses, although sub-genomic elements of geminiviruses can be packaged in non-geminate capsids (84, 85).

Interestingly, CruV-420 contains not one tombusvirus-related CP, but two. A recent compilation of CRESS-DNA viruses from animal metagenomes also contains four genomes with two different CPs in their capsid (31). Whether these viruses use two different CPs in their capsid (as some RNA viruses do), or whether these are intermediates in the exchange of CP genes, as predicted from the gene capture mechanism proposed by Stedman (2013) (18), is unclear. If the latter is true, CP gene acquisition by CRESS-DNA viruses may be much more common than previously thought.

Crucivirus Rep

The Reps of CRESS-DNA viruses typically contain an endonuclease domain characterized by conserved motifs I, II and III, and a helicase domain with Walker A and B motifs, motif C, and an Arg-finger [Fig. 2B; (12)]. The majority (85.9%) of the crucivirus genomes described in this dataset contain all of the expected Rep motifs (Supp. Table 4). However, five genomes (CruCGE-110, CruCGE-296, CruCGE-436, CruCGE-471 and CruCGE-533) with overall sequence homology to other Reps (35.8, 32.7, 49.7, 60.2 and 57.2 % PI with other putative Reps in the databases, respectively), lack any detectable conserved motifs within their sequence. Thus, these sequences are considered CP-encoding cruci-like circular genetic elements (CruCGEs).

The endonuclease catalytic domain of Rep (motif II), including HUH, was identified in 441 of the genomes of which 95.2% had an alternative HUH, with the most common arrangement being HUQ (70.0%), also found in circoviruses and nanoviruses [(10, 25, 39); Fig. 2B]. 26.2% of the crucivirus motif II deviate from the HUH motif by additionally replacing the second hydrophobic residue (U) with a polar amino acid (Fig. 2B; supp. Table 4), with 53 of Reps with the sequence HYQ (12.0%), also found in smacoviruses (10, 23, 45).

We identified thirteen putative Reps in these crucivirus genomes that lack all four motifs typically found in S3H helicases (e.g. CruV-166, CruV-202, CruV-499; Supp. Table 4). Recent work has shown that the deletion of individual conserved motifs in the helicase domain of the Rep protein of beak and feather disease virus does not abolish ATPase and GTPase activity (86). The absence of all four motifs may prevent these putative Reps from performing helicase and ATPase activity using previously characterized mechanisms. However, it is possible that crucivirus Reps that lack these motifs are still capable of ATP hydrolysis and associated helicase activity. Alternatively, these activities may be provided by host factors (87), or by a viral replication-enhancer protein – as is the case with the AC3 protein of begomoviruses (88).

We identified 36 crucivirus genomes whose putative *rep* genes contain in-frame stop codons or the HUH and SF3 helicase are in different frames, suggesting that their transcripts may require intron splicing prior to translation. Acceptor and donor splicing sites identical to those found in maize streak virus (54) were found in all these sequences, and the putatively spliced Reps annotated accordingly. In five of the 36 spliced Reps, we were unable to detect any of the four conserved motifs associated with helicase/ATPase activity, which are encoded in the predicted second exon in most cases. CruV-513 and CruV-518 also contain predicted splicing sites in their CP gene.

No GRS motifs –which have been identified as necessary for geminivirus replication (89), and have also been found in genomoviruses (90)– were detected in Reps in our dataset. We were unable to detect any conserved Rep motifs present in cruciviruses that are absent in other CRESS-DNA viruses. Given the conservation of Rep motifs in these newly-described cruciviruses, we expect most to be active in RCR.

Crucivirus CPs share higher genetic identity than their Rep proteins

To assess the diversity in the proteins of cruciviruses, the percentage pairwise identity (% PI) between the protein sequences was calculated for CP and Rep using SDTv1.2 (Fig. 3). The average % PI for CP was found to be 33.1± 4.9 % PI (Figs. 3A and 3D), likely due to the high levels of conservation found in the S-domain (40.5 ± 8.4 % PI; Figs. 3B and 3D), while the average % PI for Rep is quite low at 24.7% (± 5.6 SD; Figs. 3C and 3D). The high variation of the Rep protein sequence relative to CP in cruciviruses correlates with a previous observation on a smaller dataset (20).

To compare cruciviruses to other viral groups with homologous proteins, sequence similarity networks were built for CP and Rep (Fig. 4). For the CP, related protein sequences from tombusviruses and unclassified RNA viruses were included. The virus sequences were connected when the similarity between their protein sequence had an e-value < 1e–20, sufficient to connect all cruciviruses and tombusviruses, with the exception of CruV-523 (Fig. 4A). However, using BLASTp, CruV-523 showed similarity to other RNA viruses with an e-value < 1e–9, which were not included in the analysis. The CP sequence similarity network analysis demonstrates the apparent homology of the CPs in our dataset with the CP of RNA viruses; specifically to unclassified RNA viruses that have RdRPs similar to either tombusviruses –also described as tombus-like viruses (77, 91, 92)– or to nodaviruses. The latter RNA viruses are proposed to belong to a chimeric group of viruses named tombunodaviruses (93).

For sequence similarity network analysis of Rep, sequences from CRESS-DNA viruses belonging to the families *Circoviridae, Nanoviridae, Alphasatellitidae, Geminiviridae, Genomoviridae, Smacoviridae* and *Bacilladnaviridae* were used (Fig. 4B). Due to the heterogeneity of Rep (Fig. 3C), the score cutoff for the network was relaxed to an e-value < 1e–10; nonetheless, ten divergent sequences lacked sufficient similarity to form connections within the network. While the Rep of the different viral families clustered in specific regions of the network, the similarity of cruciviral Reps spans the diversity of all CRESS-DNA viruses, and blurs the borders between them. Though there are cruciviruses that appear to be closely related to geminiviruses and genomoviruses, these connections are less common than with other classified CRESS-DNA families (Fig. 4B). While still highly divergent from each other, the conserved motifs in the Rep still share the most sequence similarity with CRESS-DNA viruses (Fig. 2B).

The broad sequence space distribution of cruciviral Rep sequences has been proposed to reflect multiple Rep acquisition events through recombination with viruses from different CRESS-DNA viral families (20). However, the apparent larger diversity of cruciviral Reps relative to classified CRESS-DNA viruses can be due to the method of study, as most classified CRESS-DNA viruses have been discovered from infected organisms and are grouped mainly based on Rep similarity (1). By contrast, here crucivirus sequences are selected according to the presence of a tombusvirus-like CP. Moreover, the Rep of cruciviruses could be subject to higher substitution rates than CP (26). It is possible that sequence divergence in CP is more limited than in the Rep due to structural constraints.

Horizontal gene transfer among cruciviruses.

To gain insight into the evolutionary history of cruciviruses, we carried out phylogenetic analyses of their CPs and Reps. Due to the high sequence diversity in the dataset, two smaller subsets of sequences were analyzed:

i) *CP-based clusters*. Clusters with more than six non-identical CP sequences whose S-domains share a % PI greater than 70% were identified from Fig. 3B. This resulted in the identification of seven clusters, and a more divergent, yet clearly distinct, cluster was included (pink in Fig. 3B). A total of 47 genomes from the eight different clusters were selected for sequence comparison. The protein sequences of CP and Rep were extracted, aligned, and their phylogenies inferred and analyzed using tanglegrams (Fig. 5A). The CP phylogeny shows that the eight CP-based clusters form separate clades (Fig. 5A). On the other hand, the phylogeny of Rep shows a different pattern of relatedness between those genomes (Fig. 5A). This suggests different evolutionary histories for the CP and Rep proteins, which could be due to recombination events between cruciviruses, as previously proposed with smaller datasets (20, 21).

ii) *Rep-based clusters*. To account for the possible bias introduced by selecting genomes from CP cluster groups and to increase the resolution in the phylogeny of the Rep sequences, clusters with more than six Rep sequences sharing PI > 45 and < 98% were identified. A total of 53 genomes from six clusters (Fig. 3C) were selected, and their CP and Rep protein sequences analyzed. The phylogeny of Reps shows distinct clades between the sequences from different clusters (Fig. 5B). When the phylogeny of Rep was compared to that of their corresponding CPs, we observed the presence of groups of cruciviruses that clade with each other in both the Rep and the CP. Discrepancies in topology between Rep and CP were observed as well, particularly in the CP clade marked with an asterisk in Fig. 5B. This clade corresponds to the highly-homogeneous red CP-based cluster shown in Fig. 3B, and suggests that gene transfer is more common in cruciviruses with a more similar CP, likely infecting the same type of organism. On the other hand, the presence of cruciviral groups with no trace of genetic exchange may indicate that lineages within the cruciviral group may have undergone speciation in the course of evolution.

To investigate possible exchanges of individual Rep domains among cruciviruses, the Rep alignments of the analyses of the CP-based and Rep-based clusters were split at the beginning of the Walker A motif to separate endonuclease and helicase domains. From the analysis of the CP-based clusters, we observed incongruence in the phylogenies between endonuclease and helicase domains (Fig. 6A), suggesting recombination within crucivirus Reps, as has been previously hypothesized with a much smaller dataset (21). This incongruency is not observed in the analyzed Rep-based clusters (Fig. 6B). This is likely due to the higher similarity between Reps in this subset of sequences, biased by the clustering based on Rep. We do observe different topologies between the trees, which may be a consequence of different evolutionary constraints to which the endonuclease and helicase domains are subject. The detection of CP/Rep exchange and not of individual Rep domains in Rep-based clusters suggests that the rate of intergenic recombination is higher than intragenic recombination in cruciviruses.

Members of the SAR supergroup are potential crucivirus hosts

While no crucivirus host has been identified to date, the architecture of the Rep protein found in most cruciviruses, as well as the presence of introns in some of the genomes, suggests a eukaryotic host. The fusion of an endonuclease domain to a S3H helicase domain is observed in other CRESS-DNA viruses which are known to infect eukaryotes (38). This is distinct from Reps found in prokaryote-infecting CRESS-DNA viruses –which lack a fused S3H helicase domains (94)– and other related HUH endonucleases involved in plasmid RCR and HUH transposases (38). Additionally, the CP of cruciviruses, a suggested determinant of tropism (95, 96), is homologous to the capsid of RNA viruses known to infect eukaryotes. The RNA viruses with a known host with capsids most similar to cruciviral capsids (tombunodaviruses) infect oomycetes, a group of filamentous eukaryotic stramenopiles (77).

Cruciviruses have been found as contaminants of spin columns made of diatomaceous silica (21), in aquatic metagenomes enriched with unicellular algae (20), in the metagenome of *Astrammina rara* –a foraminiferan protist part of the rhizaria– (20), and associated with epibionts of isopods, mainly comprised of apicomplexans and ciliates, both belonging to the alveolates (26). These pieces of evidence point toward the stramenopiles/alveolates/rhizaria (SAR) supergroup as a candidate taxon to contain potential crucivirus hosts (97). No host prediction can be articulated from our sequence data. However, at least five of the crucivirus genomes only render complete translated CP and Rep sequences when using a relaxed genetic code. Such alternative genetic codes have been detected in ciliates, in which the hypothetical termination codons UAA and UAG encode for a glutamine (98). The usage of an alternative genetic code seems evident in CruV-502 –found in the metagenome from seawater collected above diseased coral colonies (99)– that uses a UAA codon for a glutamine of the S-domain conserved in 33.5% of the sequences. While the data accumulated suggest unicellular eukaryotes and SAR members as crucivirus-associated organisms, the host of cruciviruses remains elusive, and further investigations are necessary.

Classification of cruciviruses

Cruciviruses have circular genomes that encode a Rep protein probably involved in RCR. The single-stranded nature of packaged crucivirus genomes has not been demonstrated experimentally; however, the overall genomic structure and sequence similarity underpins the placement of cruciviruses within the CRESS-DNA viruses.

The classification of the CRESS-DNA viruses is primarily based upon the phylogeny of the Rep proteins, although commonalities in CP and genome organization are also considered (13). This taxonomic criteria is challenging in cruciviruses, whose

Rep proteins are highly diverse and apparently paralogous. Whether the use of proteins involved in replication for virus classification should be preferred over structural proteins has been previously questioned (100).

The capsid of cruciviruses, as well as the capsid of other CRESS-DNA virus families like circoviruses, geminiviruses and bacilladnaviruses, possess the single-jelly roll architecture (46). However, there is no obvious sequence similarity between the CP of cruciviruses and that of classified CRESS-DNA viruses. The crucivirus CP – homologous to the capsid of tombusviruses– is an orthologous trait within the CRESS-DNA viruses. Hence, CP constitutes a synapomorphic character that demarcate this group of viruses from the rest of the CRESS-DNA viral families. Thus, the classification of cruciviruses is challenging.

CONCLUDING REMARKS

Cruciviruses are a growing group of CRESS-DNA viruses that encode CPs that are homologous to those encoded by tombusviruses. Over 500 crucivirus genomes have been recovered from various environments across the globe. These genomes vary in size, sequence and genome organization. While crucivirus CPs are relatively homogeneous, the Reps are relatively diverse amongst the cruciviruses, spanning the diversity of all classified CRESS-DNA viruses. It has been hypothesized that cruciviruses emerged from the recombination between a CRESS-DNA virus and a tombus-like RNA virus (15, 18). Furthermore, cruciviruses seem to have recombined with each other to exchange functional modules between them, and probably with other viral groups, which blurs their evolutionary history. Cruciviruses show evidence of genetic transfer, not just between viruses with similar genomic properties, but also between disparate groups of viruses such as CRESS-DNA and RNA viruses.

ACKNOWLEDGEMENTS

This work was supported by the NASA Exobiology Program, grant 80NSSC17K0301 (I. dl H., G. K., E. T., A. P. and K. S.), the NIH BUILD EXITO Program (A. M.), BUILD EXITO was supported by grants from the National Institutes of Health: UL1GM118964; RL5GM118963; TL4GM118965, and the Portland State

University Ronald E. McNair Scholars Program (E. T.), supported by grants from the U.S. Department of Education and Portland State University. The Antarctic field work was supported by the US National Science Foundation (NSF) under grant ANT-0944411, with logistics supplied by the US Antarctic Program. The fresh water work in New Zealand was supported by a grant (UC-E6007) from the American New Zealand Association (USA) awarded to P. Z-R., C. G., J. S. H. and A. V. The green lipped mussel work was supported by a grant from the Brian Mason Scientific & Technical Trust of New Zealand awarded to S.G. and A.V. EU-s Horizon 2020 Framework Program for Research and Innovation ('Virus-X', project no. 685778) supported F. E.

BIBLIOGRAPHY

1. Simmonds P, Adams MJ, Benk M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B, Hull R, King AMQ, Koonin E V., Krupovic M, Kuhn JH, Lefkowitz EJ, Nibert ML, Orton R, Roossinck MJ, Sabanadzovic S, Sullivan MB, Suttle CA, Tesh RB, Van Der Vlugt RA, Varsani A, Zerbini FM. 2017. Consensus statement: Virus taxonomy in the age of metagenomics. Nat Rev Microbiol 15:161–168.

2. Chow C-ET, Suttle CA. 2015. Biogeography of Viruses in the Sea. Annu Rev Virol 2:41–66.

3. Koonin E V., Dolja V V. 2013. A virocentric perspective on the evolution of life. Curr Opin Virol 3:546–557.

4. Koonin E V., Krupovic M. 2018. The depths of virus exaptation. Curr Opin Virol 31:1–8.

5. Berliner AJ, Mochizuki T, Stedman KM. 2018. Astrovirology: Viruses at Large in the Universe. Astrobiology 18:207–223.

6. Domingo E, Sheldon J, Perales C. 2012. Viral Quasispecies Evolution. Microbiol Mol Biol Rev 76:159–216.

7. Baltimore D. 1971. Expression of animal virus genomes. Bacteriol Rev 35:235–241.

8. Koonin E V., Senkevich TG, Dolja V V. 2006. The ancient virus world and evolution of cells. Biol Direct 1.

9. Krupovic M, Ravantti JJ, Bamford DH. 2009. Geminiviruses: A tale of a plasmid becoming a virus. BMC Evol Biol 9:1–11.

10. Kazlauskas D, Varsani A, Koonin E V., Krupovic M. 2019. Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. Nat Commun 10:1–12.

11. Krupovic M. 2012. Recombination between RNA viruses and plasmids might have played a central role in the origin and evolution of small DNA viruses. BioEssays 34:867–870.

12. Zhao L, Rosario K, Breitbart M, Duffy S. 2019. Eukaryotic Circular Rep-Encoding Single-Stranded DNA (CRESS DNA) Viruses: Ubiquitous Viruses With Small Genomes and a Diverse Host Range. Adv Virus Res, 1st ed. 103:71–133.

13. Rosario K, Duffy S, Breitbart M. 2012. A field guide to eukaryotic circular single-stranded DNA viruses: Insights gained from metagenomics. Arch Virol 157:1851–1871.

14. Cheung AK. 2015. Specific functions of the Rep and Rep' proteins of porcine circovirus during copy-release and rolling-circle DNA replication. Virology 481:43–50.

15. Diemer GS, Stedman KM. 2012. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. Biol Direct 7:13.

16. Cheung AK. 2012. Porcine circovirus: Transcription and DNA replication. Virus Res 164:46–53.

17. LAUFS J. 1995. Geminivirus replication: Genetic and biochemical characterization of rep protein function, a review. Biochimie 77:765–773.

18. Stedman K. 2013. Mechanisms for RNA capture by ssDNA viruses: Grand theft RNA. J Mol Evol 76:359–364.

19. 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.

20. Roux S, Enault F, Bronner G, Vaulot D, Forterre P, Krupovic M. 2013. Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses. Nat Commun 4:2700.

21. Krupovic M, Zhi N, Li J, Hu G, Koonin E V., Wong S, Shevchenko S, Zhao K, Young NS. 2015. Multiple layers of chimerism in a single-stranded DNA virus discovered by deep sequencing. Genome Biol Evol 7:993–1001.

22. Hewson I, Ng G, Li WF, LaBarre BA, Aguirre I, Barbosa JG, Breitbart M, Greco AW, Kearns CM, Looi A, Schaffner LR, Thompson PD, Hairston NG. 2013. Metagenomic identification, seasonal dynamics, and potential transmission mechanisms of a Daphnia-associated single-stranded DNA virus in two temperate lakes. Limnol Oceanogr 58:1605–1620.

23. Steel O, Kraberger S, Sikorski A, Young LM, Catchpole RJ, Stevens AJ, Ladley JJ, Coray DS, Stainton D, Dayaram A, Julian L, van Bysterveldt K, Varsani A. 2016. Circular replication-associated protein encoding DNA viruses identified in the faecal matter of various animals in New Zealand. Infect Genet Evol 43:151–164.

24. McDaniel LD, Rosario K, Breitbart M, Paul JH. 2014. Comparative metagenomics: natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. Environ Microbiol 16:570–585.

25. Dayaram A, Galatowitsch ML, Argüello-Astorga GR, van Bysterveldt K, Kraberger S, Stainton D, Harding JS, Roumagnac P, Martin DP, Lefeuvre P, Varsani A. 2016. Diverse circular replication-associated protein encoding viruses circulating in invertebrates within a lake ecosystem. Infect Genet Evol 39:304–316.

26. Bistolas K, Besemer R, Rudstam L, Hewson I. 2017. Distribution and Inferred Evolutionary Characteristics of a Chimeric ssDNA Virus Associated with Intertidal Marine Isopods. Viruses 9:361.

27. Quaiser A, Krupovic M, Dufresne A, Francez A-J, Roux S. 2016. Diversity and comparative genomics of chimeric viruses in Sphagnum- dominated peatlands. Virus Evol 2:vew025.

28. Salmier A, Tirera S, de Thoisy B, Franc A, Darcissac E, Donato D, Bouchier C, Lacoste V, Lavergne A. 2017. Virome analysis of two sympatric bat species (Desmodus rotundus and Molossus molossus) in French Guiana. PLoS One 12:e0186943.

29. de la Higuera I, Torrance EL, Pratt AA, Kasun GW, Maluenda A, Stedman KM. 2019. Genome Sequences of Three Cruciviruses Found in the Willamette Valley (Oregon). Microbiol Resour Announc 8.

30. Kraberger S, Argüello-Astorga GR, Greenfield LG, Galilee C, Law D, Martin DP, Varsani A. 2015. Characterisation of a diverse range of circular replication-associated protein encoding DNA viruses recovered from a sewage treatment oxidation pond. Infect Genet Evol 31:73–86.

31. Tisza MJ, Pastrana D V, Welch NL, Stewart B, Peretti A, Starrett GJ, Pang Y-YS, Krishnamurthy SR, Pesavento PA, McDermott DH, Murphy PM, Whited JL, Miller B, Brenchley J, Rosshart SP, Rehermann B, Doorbar J, Ta'ala BA, Pletnikova O, Troncoso JC, Resnick SM, Bolduc B, Sullivan MB, Varsani A, Segall AM, Buck CB. 2020. Discovery of several thousand highly diverse circular DNA viruses. Elife 9:555375.

32. Brown DR, Schmidt-Glenewinkelg T, Reinberg D, Hurwitz J. 1983. DNA sequences which support activities of the bacteriophage phiX174 gene A protein.

33. Steinfeldt T, Finsterbusch T, Mankertz A. 2006. Demonstration of Nicking/Joining Activity at the Origin of DNA Replication Associated with the Rep and Rep' Proteins of Porcine Circovirus Type 1. J Virol 80:6225–6234.

34. Gassmann M, Focher F, Buhk HJ, Ferrari E, Spadari S, Hübscher U. 1988. Replication of single-stranded porcine circovirus DNA by DNA polymerases α and δ. BBA - Gene Struct Expr.

35. Roth MJ, Brown DR, Hurwitz J. 1984. Analysis of Bacteriophage phi174 Gene A Protein-mediated Termination and Reinitiation of 6X DNA Synthesis 11. Structural characterization of the covalent 4X A protein-DNA complex.J. Biol. Chern.

36. Cheung AK. 2007. A stem-loop structure, sequence non-specific, at the origin of DNA replication of porcine circovirus is essential for termination but not for initiation of rolling-circle DNA replication. Virology 363:229–235.

37. Stenger DC, Revington GN, Stevenson MC, Bisaro DM. 1991. Replicational release of geminivirus genomes from tandemly repeated copies: Evidence for rolling-circle replication of a plant viral DNA. Proc Natl Acad Sci U S A.

38. Chandler M, de la Cruz F, Dyda F, Hickman AB, Moncalian G, Ton-Hoang B. 2013. Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. Nat Rev Microbiol 11:525–538.

39. Brown DR, Schmidt-Glenewinkelg T, Reinberg D, Hurwitz J. 1983. DNA sequences which support activities of the bacteriophage phiX174 gene A protein 258:8402–8412.

40. Ilyina T V., Koonin E V. 1992. Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. Nucleic Acids Res 20:3279–3285.

41. Koonin E V, Ilyina T V. 1993. Computer-assisted dissection of rolling circle DNA replication. BioSystems 30:241–268.

42. Londoño A, Riego-Ruiz L, Argüello-Astorga GR. 2010. DNA-binding specificity determinants of replication proteins encoded by eukaryotic ssDNA viruses are adjacent to widely separated RCR conserved motifs. Arch Virol 155:1033–1046.

43. Gorbalenya AE, Koonin E V. 1993. Helicases: amino acid sequence comparisons and structure-function relationships. Curr Opin Struct Biol 3:419–429.

44. Clerot D, Bernardi F. 2006. DNA Helicase Activity Is Associated with the Replication Initiator Protein Rep of Tomato Yellow Leaf Curl Geminivirus. J Virol 80:11322–11330.

45. Kazlauskas D, Varsani A, Krupovic M. 2018. Pervasive chimerism in the replication-associated proteins of uncultured single-stranded DNA viruses. Viruses 10:1–11.

46. Krupovic M, Koonin E V. 2017. Multiple origins of viral capsid proteins from cellular ancestors. Proc Natl Acad Sci U S A 114:E2401–E2410.

47. Hopper P, Harrison SC, Sauer RT. 1984. Structure of tomato bushy stunt virus. V. Coat protein sequence determination and its structural implications. J Mol Biol 177:701–713.

48. Sherman MB, Guenther R, Reade R, Rochon D, Sit T, Smith TJ. 2019. Near-Atomic-Resolution Cryo-Electron Microscopy Structures of Cucumber Leaf Spot Virus and Red Clover Necrotic Mosaic Virus: Evolutionary Divergence at the Icosahedral Three-Fold Axes. J Virol 94.

49. Alam SB, Reade R, Theilmann J, Rochon DA. 2017. Evidence for the role of basic amino acids in the coat protein arm region of Cucumber necrosis virus in particle assembly and selective encapsidation of viral RNA. Virology 512:83–94.

50. Park SH, Sit TL, Kim KH, Lommel SA. 2013. The red clover necrotic mosaic virus capsid protein N-terminal amino acids possess specific RNA binding activity and are required for stable virion assembly. Virus Res 176:107–118.

51. Ohki T, Akita F, Mochizuki T, Kanda A, Sasaya T, Tsuda S. 2010. The protruding domain of the coat protein of Melon necrotic spot virus is involved in compatibility with and transmission by the fungal vector Olpidium bornovanus. Virology 402:129–134.

52. Llauró A, Coppari E, Imperatori F, Bizzarri AR, Castón JR, Santi L, Cannistraro S, De Pablo PJ. 2015. Calcium Ions Modulate the Mechanics of Tomato Bushy Stunt Virus. Biophys J 109:390–397.

53. Paez-Espino D, Chen I-MA, Palaniappan K, Ratner A, Chu K, Szeto E, Pillay M, Huang J, Markowitz VM, Nielsen T, Huntemann M, K. Reddy TB, Pavlopoulos GA, Sullivan MB, Campbell BJ, Chen F, McMahon K, Hallam SJ, Denef V, Cavicchioli R, Caffrey SM, Streit WR, Webster J, Handley KM, Salekdeh GH, Tsesmetzis N, Setubal JC, Pope PB, Liu W-T, Rivers AR, Ivanova NN, Kyrpides NC. 2017. IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. Nucleic Acids Res 45:D457–D465.

54. Wright EA, Heckel T, Groenendijk J, Davies JW, Boulton MI. 1997. Splicing features in maize streak virus virion- and complementary-sense gene expression. Plant J 12:1285–1297.

55. Steinfeldt T, Finsterbusch T, Mankertz A. 2006. Demonstration of Nicking/Joining Activity at the Origin of DNA Replication Associated with the Rep and Rep' Proteins of Porcine Circovirus Type 1. J Virol 80:6225–6234.

56. Hafner GJ, Dale JL, Harding RM, Wolter LC, Stafford MR. 1997. Nicking and joining activity of banana bunchy top virus replication protein in vitro. J Gen Virol 78:1795–1799.

57. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. Algorithms Mol Biol 6:26.

58. Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 288:911–940.

59. Muhire BM, Varsani A, Martin DP. 2014. SDT: A Virus Classification Tool Based on Pairwise Sequence Alignment and Identity Calculation. PLoS One 9:e108277.

60. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189–1191.

61. Livingstone CD, Barton GJ. 1993. Protein sequence alignments: A strategy for the hierarchical analysis of residue conservation. Bioinformatics.

62. Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: A sequence logo generator. Genome Res.

63. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10:845–858.

64. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera - A visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612.

65. Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 30:772–780.

66. Floden EW, Tommaso PD, Chatzou M, Magis C, Notredame C, Chang J-M. 2016. PSI/TM-Coffee: a web server for fast and accurate multiple sequence alignments of regular and transmembrane proteins using homology extension on reduced databases. Nucleic Acids Res 44:W339–W343.

67. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics.

68. Price MN, Dehal PS, Arkin AP. 2009. Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 26:1641–1650.

69. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst Biol 59:307–321.

70. Anisimova M, Gascuel O. 2006. Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. Syst Biol 55:539–552.

71. Huson DH, Scornavacca C. 2012. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. Syst Biol.

72. Zallot R, Oberg N, Gerlt JA. 2019. The EFI Web Resource for Genomic Enzymology Web Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. Biochemistry acs.biochem.9b00735.

73. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504.

74. Kazlauskas D, Dayaram A, Kraberger S, Goldstien S, Varsani A, Krupovic M. 2017. Evolutionary history of ssDNA bacilladnaviruses features horizontal acquisition of the capsid gene from ssRNA nodaviruses. Virology 504:114–121.

75. Sicard A, Michalakis Y, Gutiérrez S, Blanc S. 2016. The Strange Lifestyle of Multipartite Viruses. PLoS Pathog 12:1–19.

76. Varsani A, Lefeuvre P, Roumagnac P, Martin D. 2018. Notes on recombination and reassortment in multipartite/segmented viruses. Curr Opin Virol 33:156–166.

77. Grasse W, Spring O. 2017. ssRNA viruses from biotrophic Oomycetes form a new phylogenetic group between Nodaviridae and Tombusviridae. Arch Virol 162:1319–1324.

78. Rosario K, Mettel KA, Benner BE, Johnson R, Scott C, Yusseff-Vanegas SZ, Baker CCM, Cassill DL, Storer C, Varsani A, Breitbart M. 2018. Virus discovery in all three major lineages of terrestrial arthropods highlights the diversity of single-stranded DNA viruses associated with invertebrates. PeerJ 6:e5761.

79. Varsani A, Krupovic M. 2017. Sequence-based taxonomic framework for the classification of uncultured single-stranded DNA viruses of the family Genomoviridae. Virus Evol 3.

80. Harrison SC, Olson AJ, Schutt CE, Winkler FK, Bricogne G. 1978. Tomato bushy stunt virus at 2.9 Å resolution. Nature 276:368–373.

81. Chelvanayagam G, Heringa J, Argos P. 1992. Anatomy and Evolution of Proteins Displaying the Viral Capsid Jellyroll Topology. J Mol Biol 228:220–242.

82. Katpally U, Kakani K, Reade R, Dryden K, Rochon D, Smith TJ. 2007. Structures of T=1 and T=3 Particles of Cucumber Necrosis Virus: Evidence of Internal Scaffolding. J Mol Biol 365:502–512.

83. Campbell JW, Clifton IJ, Greenhough TJ, Hajdu J, Harrison SC, Liddington RC, Shrive AK. 1990. Calcium binding sites in tomato bushy stunt virus visualized by laue crystallography. J Mol Biol 214:627–632.

84. Casado CG, Javier Ortiz G, Padron E, Bean SJ, McKenna R, Agbandje-McKenna M, Boulton MI. 2004. Isolation and characterization of subgenomic DNAs encapsidated in "single" T = 1 isometric particles of Maize streak virus. Virology 323:164–171.

85. Bennett A, Rodriguez D, Lister S, Boulton M, McKenna R, Agbandje-McKenna M. 2018. Assembly and disassembly intermediates of maize streak geminivirus. Virology 525:224–236.

86. Chen JK, Hsiao C, Wu JS, Lin SY, Wang CY. 2019. Characterization of the endonuclease activity of the replication-associated protein of beak and feather disease virus. Arch Virol 164:2091–2106.

87. Rizvi I, Choudhury NR, Tuteja N. 2015. Insights into the functional characteristics of geminivirus rolling-circle replication initiator protein and its interaction with host factors affecting viral DNA replication. Arch Virol 160:375–387.

88. Pasumarthy KK, Choudhury NR, Mukherjee SK. 2010. Tomato leaf curl Kerala virus (ToLCKeV) AC3 protein forms a higher order oligomer and enhances ATPase activity of replication initiator protein (Rep/AC1). Virol J 7:128.

89. Nash TE, Dallas MB, Reyes MI, Buhrman GK, Ascencio-Ibanez JT, Hanley-Bowdoin L. 2011. Functional Analysis of a Novel Motif Conserved across Geminivirus Rep Proteins. J Virol 85:1182–1192.

90. Varsani A, Krupovic M. 2017. Sequence-based taxonomic framework for the classification of uncultured single-stranded DNA viruses of the family Genomoviridae. Virus Evol 3.

91. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, Qin XC, Li J, Cao JP, Eden JS, Buchmann J, Wang W, Xu J, Holmes EC, Zhang YZ. 2016. Redefining the invertebrate RNA virosphere. Nature 540:539–543.

92. Dolja V V., Koonin E V. 2018. Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. Virus Res 244:36–52.

93. Greninger AL, DeRisi JL. 2015. Draft Genome Sequence of Tombunodavirus UC1. Genome Announc 3.

94. Krupovic M. 2013. Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. Curr Opin Virol 3:578–586.

95. Allison AB, Organtini LJ, Zhang S, Hafenstein SL, Holmes EC, Parrish CR. 2016. Single Mutations in the VP2 300 Loop Region of the Three-Fold Spike of the Carnivore Parvovirus Capsid Can Determine Host Range. J Virol.

96. Carbonell A, Maliogka VI, Pérez J de J, Salvador B, León DS, García JA, Simón-Mateo C. 2013. Diverse Amino Acid Changes at Specific Positions in the N-Terminal Region of the Coat Protein Allow Plum pox virus to Adapt to New Hosts. Mol Plant-Microbe Interact 26:1211–1224.

97. Beakes GW, Glockling SL, Sekimoto S. 2012. The evolutionary phylogeny of the oomycete "fungi." Protoplasma 249:3–19.

98. Hanyu N, Kuchino Y, Nishimura S, Beier H. 1986. Dramatic events in ciliate evolution: alteration of UAA and UAG termination codons to glutamine codons due to anticodon mutations in two Tetrahymena tRNAs Gln . EMBO J 5:1307–1311.

99. Soffer N, Brandt ME, Correa AMS, Smith TB, Thurber RV. 2014. Potential role of viruses in white plague coral disease. ISME J 8:271–283.

100. Krupovic M, Bamford DH. 2010. Order to the Viral Universe. J Virol 84:12476–12479.

Figure 1

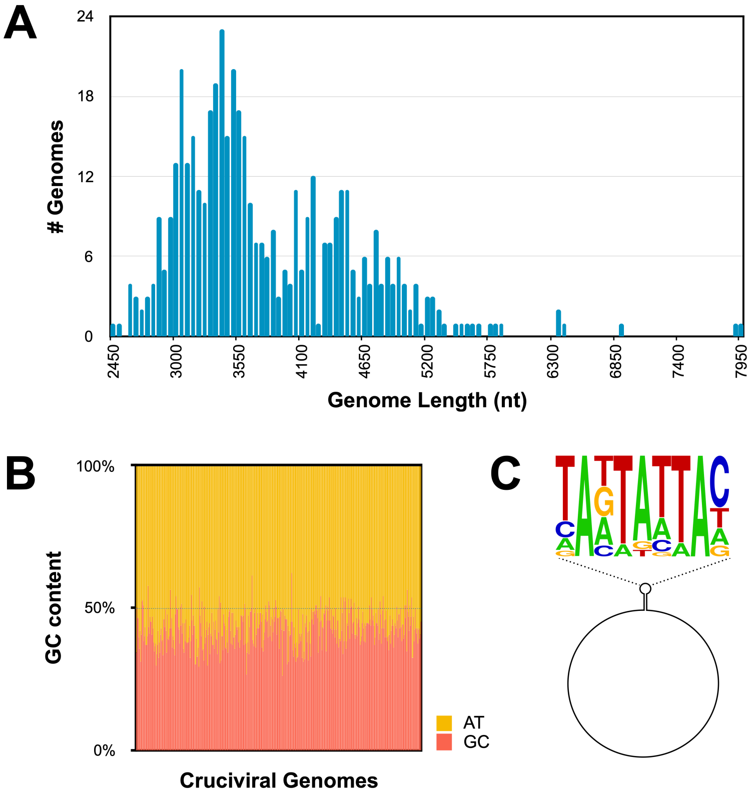

Figure 1. Genome properties of 461 new cruciviral circular sequences. (A) Histogram of cruciviral genome lengths categorized in 50 nt bins. **(B)** Percentage of G + C content versus A + T in each of the sequences described in this study **(C)** Relative abundance of nucleotides in the conserved nonanucleotide sequence of the 211 stem-loops and putative origins of replication represented predicted with StemLoop-Finder (Pratt *et* al., unpublished) in Sequence Logo format.

Figure 2

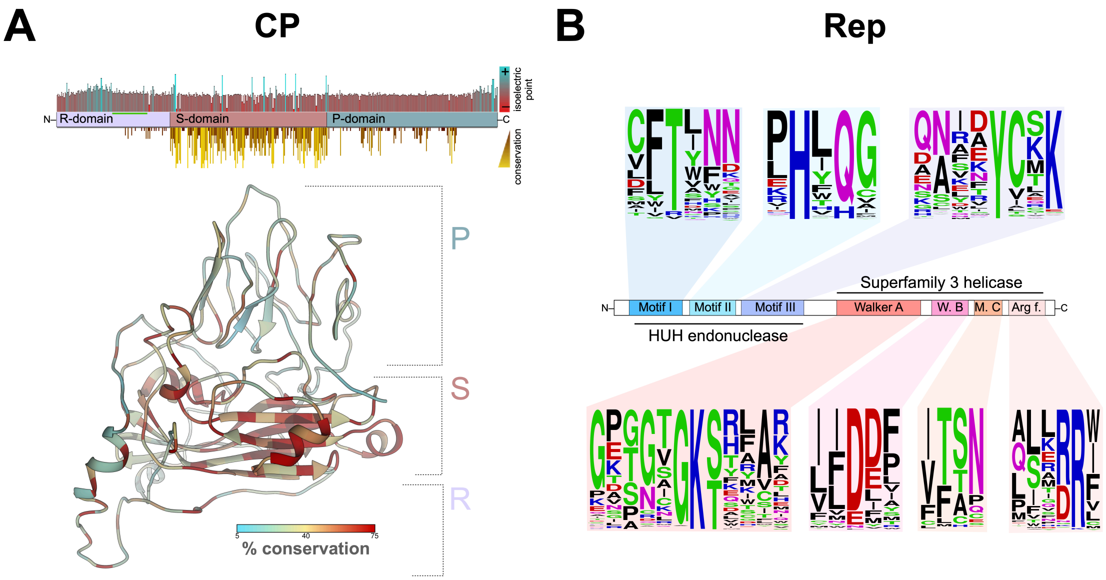

Figure 2. Protein conservation in cruciviruses. (A) Top: distribution of domains, isoelectric point and conservation in a consensus capsid protein (CP). 461 CP protein sequences were aligned in Geneious 11.0.4 with MAFFT (G-INS-i, BLOSUM 45, open gap penalty 1.53, offset 0.123) and trimmed manually. The conservation of the physico-chemical properties at each position was obtained with Jalview v2.11.0, and the isoelectric point was estimated in Geneious 11.0.4. The region of CP rich in glycine is highlighted with a green bar. **Bottom:** Structure of a cruciviral CP (CruV-359) as predicted by Phyre2 showing sequence conservation based on an alignment of the 47 CP protein sequences from the CP-based clusters. **(B)** Conserved motifs found in cruciviral Reps after aligning all the extracted Rep protein sequences using PSI-Coffee. Sequence logos were generated at http:// weblogo.threeplusone.com to indicate the frequency of residues at each position.

Figure 3

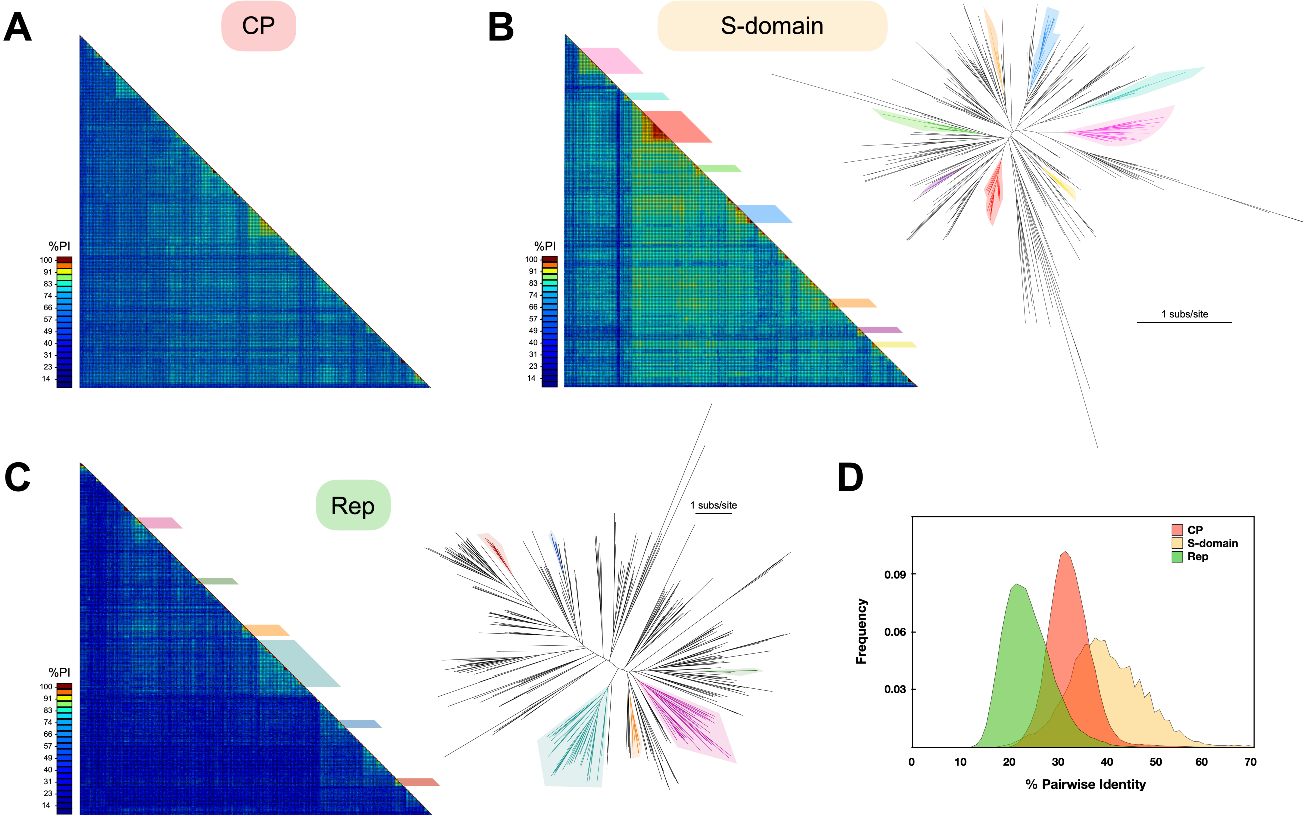

Figure 3. Diversity of cruciviral proteins. (A) CP diversity. Pairwise amino acid identity (PI) between the CPs predicted for 461 cruciviral sequences. The alignment and analysis were carried out with SDT, using the integrated MAFFT algorithm. **(B) S-domain diversity. Left:** PI matrix between the capsid protein (CP) predicted S-domain of the 461 sequences described in this study. The alignment and analysis were carried out with SDT, using the integrated MAFFT algorithm. The colored boxes indicate the different clusters of sequences used to create the CP-based clusters sequence subset. **Right:** Unrooted phylogenetic tree obtained with FastTree from a manually curated MAFFT alignment of the translated sequences of the S-domain (G-INS-i, BLOSUM 45, open gap penalty 1.53, offset 0.123). The colored branches represent the different clusters observed in the matrix. Scale bar indicates substitutions per site. **(C) Rep diversity. Left:** Pairwise identity (PI) matrix between all Reps found in cruciviral genomes in this study. The alignment and analysis were carried out with SDT, using the integrated MUSCLE algorithm. **Right:** Unrooted phylogenetic tree obtained with FastTree from an PSI-Coffee alignment of the translated sequences of Rep trimmed with TrimAl v1.3. The colored branches represent the different clusters that contain *Rep-based clusters* sequence subset. Scale bar indicates substitutions per site. **(D) PI frequency distribution.** The frequency of PI values for each of the putative proteins or domains analyzed is shown.

Figure 4

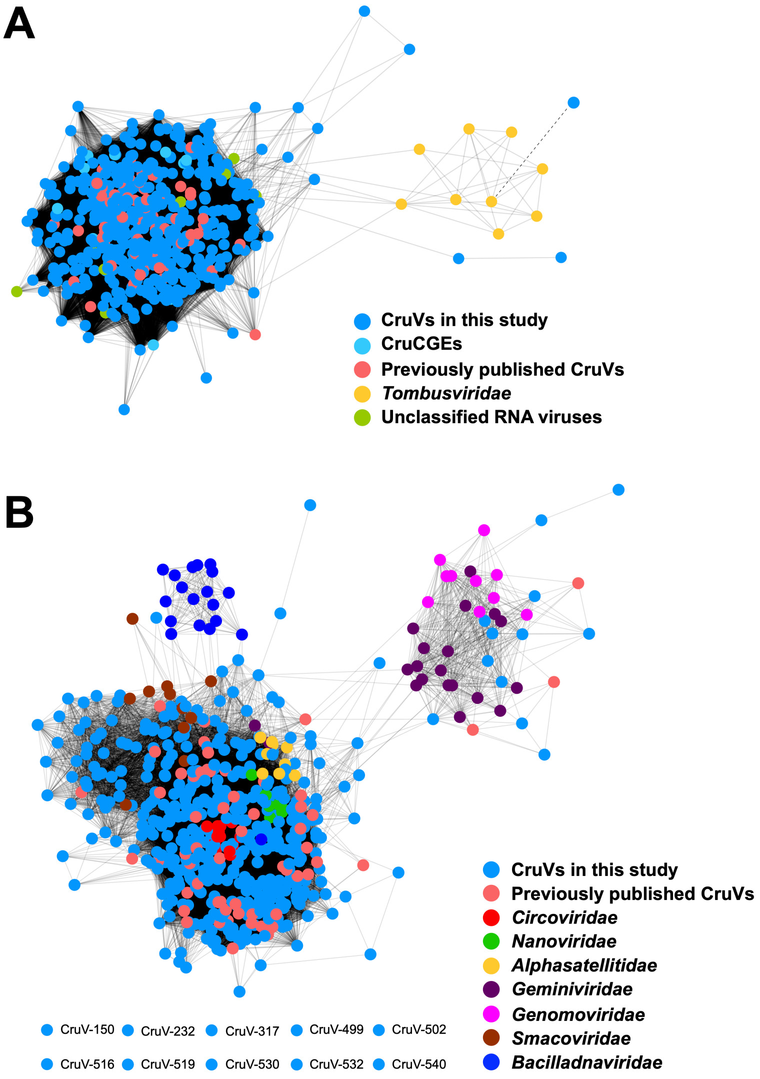

Figure 4. Similarity networks of cruciviral proteins with related viruses. (A) Capsid proteins (CP) represented by colored dots are connected with a solid line when the similarity between them is greater than e-value=$1e^{-20}$. The dashed line represents an e-value = $6e^{-7}$ between the nodes corresponding to the CP of CruV-523 and turnip crinkle virus, as given by BLASTp. **(B)** Replication-associated protein (Rep) translations, represented by colored dots, are connected with a solid line when the similarity between them greater than e-value=$1e^{-10}$. The eight nodes at the bottom left did not connect to any other node. All networks were carried out with pairwise identities calculated in the EFI–EST web server and visualized in Cytoscape v3.7.2.

Figure 5

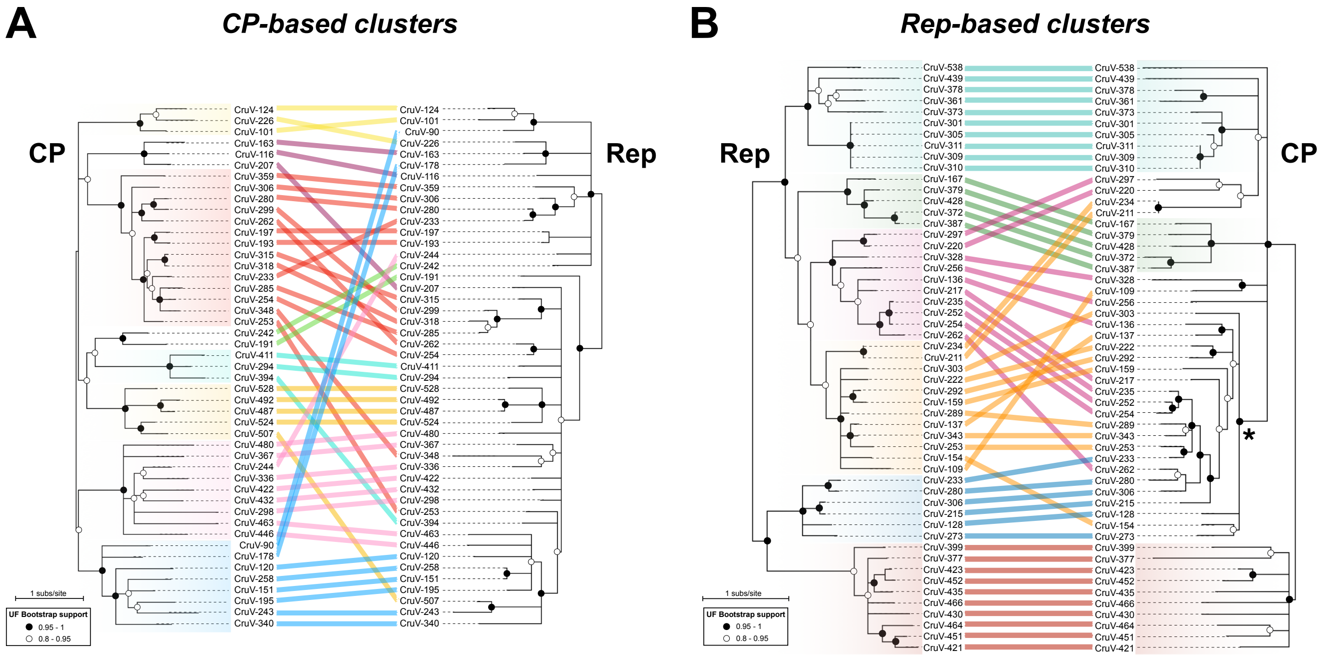

Figure 5. Comparison of phylogenies of CP and Rep proteins of representative cruciviruses. (A) Tanglegram calculated with Dendroscope v3.5.10 from phylogenetic trees generated with PhyML from Cp (PhyML automatic model selection LG+G+I+F) and Rep (PhyML automatic model selection RtREV+G+I) alignments. The tips corresponding to the same viral genome are linked by lines that are color-coded according to the clusters obtained from Fig. 3A (CP-based clusters). **(B)** Tanglegram calculated with Dendroscope v3.5.10 from phylogenetic trees generated with PhyML from Cp (PhyML automatic model selection LG+G+I+F) and Rep (PhyML automatic model selection RtREV+G+I) alignments. The tips corresponding to the same viral sequence are linked by lines that are color-coded according to the clusters obtained from Fig. 3B (Rep-based clusters). The clade marked with an asterisk is formed by members of the red cluster of subset A. Branch support is given according to aLRT SH-like (Anisimova & Gascuel, 2006). All nodes with an aLRT SH-like branch support inferior to 0.8 were collapsed with Dendroscope prior to constructing the tanglegram.

Figure 6

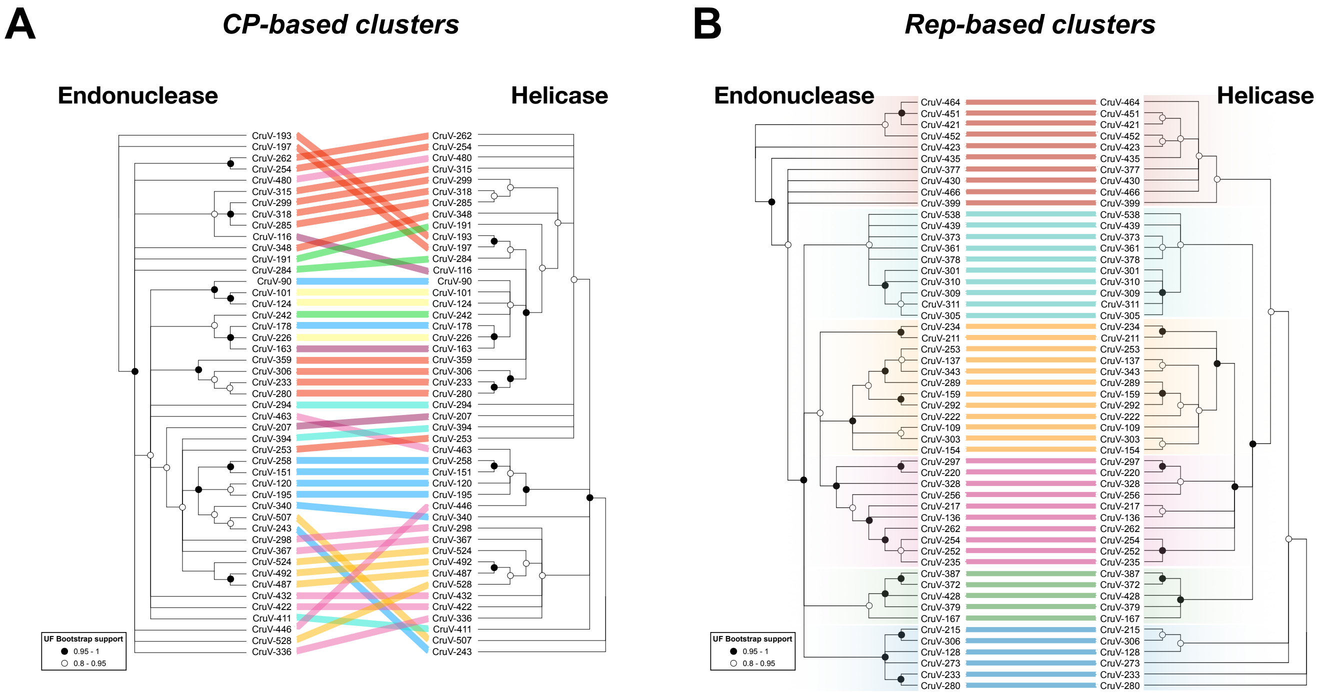

Figure 6. Comparison of phylogenies between the endonuclease and helicase domains of Reps from representative cruciviruses. (A) Tanglegram calculated with Dendroscope v3.5.10 from phylogenetic trees generated with PhyML from separate alignments of Rep endonuclease and helicase domains. The tips corresponding to the same viral genome are linked by lines that are color-coded according to the clusters obtained from Fig. 3A (CP-based clusters). **(B)** Same as A but with sequences from the clusters obtained from Fig. 3B (Rep-based clusters). All nodes with an aLRT SH-like branch support inferior to 0.8 were collapsed with Dendroscope v3.5.10 prior to constructing the tanglegram.