10-6-2021

A Bioinformatic and Biochemical Analysis of Cruciviruses

George William Kasun
Portland State University

A Bioinformatic and Biochemical Analysis of Cruciviruses

by

George William Kasun

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
in
Biology

Dissertation Committee:
Kenneth Stedman, Chair
Michael Bartlett
John Perona
Anna-Louise Reysenbach

Portland State University
2021

Abstract

Cruciviruses are novel ssDNA viruses discovered through metagenomics and direct environmental DNA amplification and cloning. The genomes of cruciviruses suggest that gene transfer between RNA and DNA viruses occurred due to the presence of putative protein-encoding genes that are homologous to both ssRNA and ssDNA viruses. In order to gain a better understanding of this group of viruses both bioinformatic analyses and in vitro biochemical experiments were employed. The results of the bioinformatic analyses show that cruciviruses are a highly diverse group of ssDNA viruses. Their placement within established ssDNA phylogenies is difficult due to heterogeneity in their putative replication-associated protein (Rep) that exceeds that of other ssDNA viruses. The results of biochemical experiments show that the putative Rep of the first discovered crucivirus, Boiling Spring Lake RNA-DNA hybrid virus (BSL-RDHV), displays activities consistent with the initiation and completion of rolling circle replication of the ssDNA genome. Specifically, it is demonstrated that recombinant BSL-RDHV Rep is capable of ATP hydrolysis, binding the putative origin of replication, covalently attaching to ssDNA containing a putative nick site, and is released from this covalent attachment in the presence of a pre-formed acceptor ssDNA. Together, these results represent significant progress towards a better understanding of this novel group of viruses. While many questions regarding cruciviruses remain unanswered, this work will enable future research to better characterize the evolution and biochemical capacities of cruciviruses.

Acknowledgements

The work contained in this dissertation would not have been possible without contributions from numerous members of the Extreme Virus Lab. I am thankful for the guidance and mentorship provided by my advisor, Dr. Kenneth Stedman. I am also grateful for his encouraging me to simultaneously pursue a broad diversity of my scientific interests, both inside and outside of a lab setting. Perhaps most importantly I am grateful for his patience and willingness to always make time to discuss ideas and results. Dr. Ignacio de la Higuera provided valuable insights and discussion that pushed this work forward and was able to talk me out of more than one silly idea. There are numerous other members of the lab who provided thought provoking discussions: Dr. Eric Iverson, David Goodman, Ellis Torrance, Max Larson, Alyssa Pratt, and Adam Jones all played instrumental roles in the work presented here.

I am thankful for the guidance provided by the members of my committee, Dr. Michael Bartlett, Dr. John Perona, and Dr. Anna-Louise Reysenbach. Their patience in the completion of this work has my utmost appreciation.

I am eternally indebted to my wife, Robyn Kasun, for her support, encouragement, and steady demeanor. Without her and our two children, Claire and Mark, this work would have been much more difficult. I love all of you, very much.

Table Of Contents

List of Tables

Chapter 3

Chapter 4

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter One

Introduction To Circular Rep Encoding Single-Stranded DNA Viruses

Viruses are small obligate intracellular replicators which infect organisms from all three domains of life (1–6). Viruses are known to be the most abundant biological entities on the planet with an estimated 10^{31} virions present in the earth's oceans alone, most of which likely infect bacterial organisms (7–9). Viruses and viral infection are drivers of worldwide biogeochemical processes, nutrient cycling, microbial turnover, and cellular evolution (9–14). These viral driven processes and the viruses that drive them have historically been understudied compared to the more prominent "disease causing viruses".

The advent of high throughput or "next generation" sequencing technologies over the past two decades has led to an exponential increase in the number of viral sequences deposited in publicly available databases and have revealed large amounts of "viral dark matter", unique viral sequences that lack similarity to any sequences available in public databases (15,16). This increase in available data has revealed not only the large number of viruses present on Earth, but also made it clear that viral diversity is far greater than previously thought (15). It is also now apparent that viruses have been actively shaping life on Earth since the earliest emergence of cellular organisms, and likely emerged before or very soon after the last universal common ancestor (1,3,4,13,17–20). Not only have viruses influenced cellular evolution, but as replicating biological

units, viruses are subject to evolutionary forces. Despite their ubiquitous presence and the increase in available viral sequences a unified picture of initial viral emergence(s) and subsequent evolution remains unclear and widely debated (3,4,19). As such, viruses have been historically classified broadly on the nature of their encapsidated genome: DNA or RNA, single stranded or double stranded, plus or minus sense in the case of RNA based genomes, and monopartite or multipartite genomes (21). The ambiguity of viral evolution can, in part, be attributed to the diversity of nucleic acids making up viral genomes, the lack of a universal hallmark gene in viruses equivalent to cellular rRNA, relatively high rates of mutation, frequent reassortment and recombination, large genetic diversity, and potentially more than one initial emergence of viruses (3,4,13,17,21–23).

Of particular interest to virologists has been the surprisingly large number of circular replication associated protein encoding single stranded DNA viruses (CRESS-DNA viruses) discovered using modern sequencing technologies (24). This group of viruses was historically believed to be relatively uncommon, but deep sequencing has revealed CRESS-DNA viruses to be omnipresent in diverse environments, in numbers larger than previously believed, and in association with various known and likely hosts (24–29). Furthermore, the development of the use of phi29 polymerase to amplify entire CRESS-DNA virus genomes has made the amplification and cloning of such viral genomes easier, thus supporting metagenomic data (30).

Eukaryotic CRESS-DNA viruses constitute a diverse and widespread group of viruses with circular genomes divided into the families *Geminiviridae, Circoviridae*, *Nanoviridae*, *Alphasatellitidae*, *Genomoviridae*, *Bacilladnaviridae*, *Smacoviridae* and *Redondoviridae,* as well as five unclassified clades (CRESSV1-CRESSV5) (31). Recently, the CRESS-DNA viruses have been taxonomically assigned by the International Committee on Taxonomy of Viruses (ICTV) into the phylum *Cressdnaviricota* which contains two distinct clades: Clade 1 contains the families *Geminiviridae, Genomoviridae,* and the unclassified CRESSV6 (**Fig. 1.1**) (31).

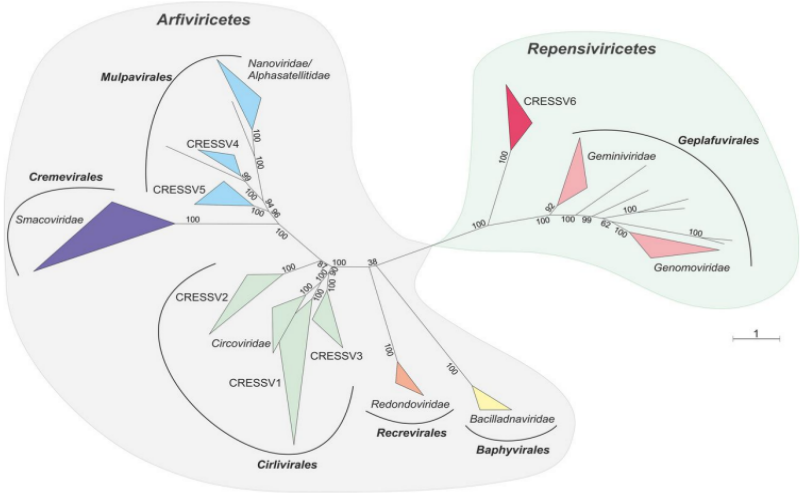

Figure 1.1. Unrooted maximum likelihood phylogenetic tree of the phylum *Cressdnaviricota*. Tree is based on Rep proteins of CRESS-DNA viruses. Adapted from Krupovic et al., 2020 (31).

Clade 2 contains the remaining classified (*Circoviridae, Samcoviridae, Nanoviridae/Alphasatellitidae, Bacilladnaviridae,* and *Redondoviridae)* and unclassified (CRESSV1-5) CRESS-DNA viruses (31). Since work on this

dissertation began it has been hypothesized that eukaryotic CRESS-DNA viruses emerged on at least two occasions when Rep encoding plasmids obtained a single jelly roll capsid gene from RNA viruses (32). These initial emergences have been followed by apparently rampant intergenic and intragenic recombination of Rep genes, particularly in the unclassified CRESS-DNA viruses (32,33).

Some CRESS-DNA viruses are pathogenic and in turn agriculturally important, such as circoviruses of animals and geminiviruses which infect a wide range of plants. Porcine circovirustype 2 (PCV2) (*Circoviridae*) is the causal agent of post weaning disease in pigs, while tomato yellow leaf curl (TYLC) virus (*Geminiviridae*) is responsible for at least tens of millions of dollars in yield losses in the United Sates alone (34–36) . Despite these economically important conditions associated with CRESS-DNA virus infection, CRESS-DNA viruses are also found in high abundance in apparently healthy organisms as is the case with smacoviruses found associated with both chickens and human fecal samples (25). Additionally, it has become apparent that CRESS-DNA viruses are capable of integration into host genomes (37–40). The examination of eukaryotic genomes has uncovered integrated/endogenous ssDNA viral sequences in plants, fungi, protists and animals (37,41). Genomic and transcriptomic work of yams revealed the presence of endogenous Geminivirus sequences which are actively transcribed (42).

All characterized CRESS-DNA viruses package their genomes in small virions, 20-40nm in size with icosahedral symmetry, or in the case of *Geminiviridae* in twinned icosahedral, T=1 capsids, comprised of multiple copies of the capsid protein (CP) encoded in their genome (24,43) . This small virion size makes the CRESS-DNA viruses among the smallest known viral particles. The CP of CRESS-DNA viruses appears to fold into an eight-strand ß-barrel that conforms to the single jelly-roll (SJR) architecture, which is widespread in eukaryotic RNA viruses (22). CP genes of eukaryotic CRESS-DNA viruses are often highly divergent, making the identification of CP genes solely on the basis of sequence identity difficult (24).

Eukaryotic CRESS-DNA virus genomes can be monopartite, bipartite or multipartite (44). Begomoviruses are members of the *Geminiviridae* whose genomes can display a bipartite arrangement in which two different ssDNA genome segments (DNA-A and DNA-B) enclosed in separate virions must both enter a cell to bring about a successful infection (45). Similarly, members of the *Nanoviridae* display a multipartite genome arrangement in which 6-8 individually packaged genome segments are required for infection (46). The multipartite genomes of the *Nanoviridae* can be as large as 10kb, but single segments of approximately 1kb are packaged into individual virions (46). In both the begomoviruses and *Nanoviridae* these genome segments are composed of unique sequences, save for a conserved region of about 200nt involved in initiation and completion of genome replication (45,46). The packaged genomes

of monopartite eukaryotic CRESS-DNA viruses are small, ranging in size from 1.7 to 6kb ssDNA (24). CRESS-DNA virus genomes may contain as few as two ORFs which are usually found in an ambisense orientation: one encoding for the replication associated protein (Rep), involved in the initiation and completion of rolling circle replication (RCR), and the other for the viral CP (24). The presence of ORFs encoding Rep and CP is conserved across all CRESS-DNA virus genomes. However, various CRESS-DNA viruses may contain up to 10 open reading frames (ORFs), and protein coding ability can vary greatly (24). While the CP of various eukaryotic CRESS-DNA viruses can be highly divergent, Rep is well conserved and exhibits a high amino acid pairwise identity across the seven families of eukaryotic CRESS-DNA viruses (24,31). Additionally, CRESS-DNA genomes contain an intragenic stem-loop that serves as the origin of replication (Ori) for virion encapsidated ssDNA (24,47–50). These structures contain a stem region of approximately 15 nucleotides with a usually nonanucleotide loop structure that features the initiation and termination point for ssDNA synthesis (47,50,51).

The hallmark and unifying feature of eukaryotic CRESS-DNA viruses is the presence of a *Rep* gene encoding the conserved replication associated protein (Rep) (24,31). Rep proteins found in eukaryotic CRESS-DNA viruses contain two domains: an N-terminal endonuclease belonging to the HUH endonuclease family and a C-terminal superfamily-3-helicase (SF3 helicase) (52,53). Members of the HUH endonuclease family are found in all three domains

of life as well as viruses and are involved in the processing of a broad range of mobile genetic elements such as viral genomes or prokaryotic plasmids, in a manner similar to the more well-known tyrosine recombinases (53,54). SF3 helicases, similar to the AAA+ family of proteins, are found in a variety of DNA and RNA viruses (32,55,56). This fusion of an HUH endonuclease and SF3 helicase domains is unique among viruses to those ssDNA viruses (linear and circular genomes) which infect eukaryotes but can also be observed in Mob relaxases involved in bacterial conjugation and in some transposases (53).

The HUH endonuclease domain of CRESS-DNA virus Reps contains three conserved motifs, Motif I, II and III. Motif I and immediately adjacent amino acids have been predicted to be involved in origin of replication (ori) recognition and binding (57). Motif II contains the namesake HUH motif (two histidines separated by a hydrophobic amino acid) is involved in metal ion binding (52,58,59). The HUH motif is known to be substituted for various amino acid arrangements in some eukaryotic CRESS-DNA viruses (HxQ in porcine circovirus type 2) (53). Structural studies have shown that this metal ion binding is likely involved in correctly positioning the scissile phosphate for nucleophilic attack by a tyrosine residue in Motif III (58,59).

Motif III resides on an α-helix and contains the catalytic tyrosine residue responsible for introducing a single stranded nick to initiate replication and is also responsible for sealing newly replicated viral genomes (**Fig 1.2**) (24,60). The binding of Rep to dsDNA has been shown to take place near stem loop

structures which are flanked by repeated DNA sequences called iterons (51,57,62). In the case of porcine circovirus type 1 these iterated DNA sequences (hexamer repeats) and the right arm of the stem loop represent the minimal binding site for Rep, while the presence of a stem loop structure (sequence non-specific) seems to be critical for Rep mediated sealing of newly replicated genomes (47,63). The ssDNA nick to initiate viral replication has been shown to take place within nonanucleotide loops typically located in the apex of stem-loops structures for members of the *Circoviridae, Geminiviridae, and Nanoviridae* (46,47,51,60). This nick to initiate replication and subsequent joining to complete replication occurs in sequences similar to NANTATT/AC (where / represents the nick site) (47).

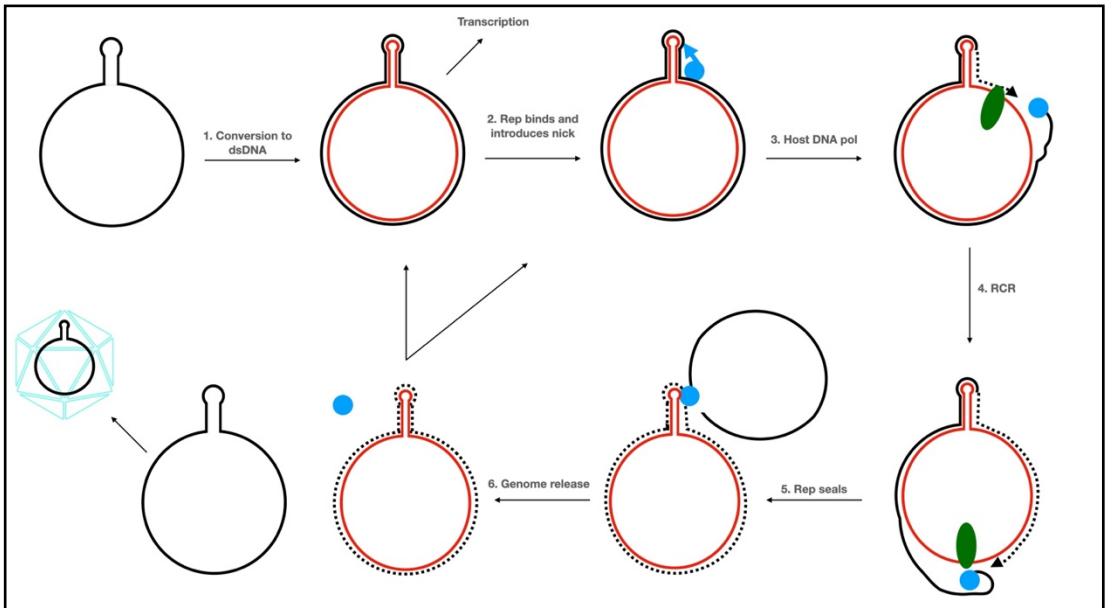

Figure 1.2. An overview of rolling circle replication in CRESS-DNA viruses. Rep is shown in blue and the host polymerase in green. Dotted line is newly synthesized DNA.

Reps found in geminiviruses and genomoviruses also contain a fourth conserved N terminal motif absent in other eukaryotic CRESS-DNA viruses (64). Aptly named as Geminivirus Rep Sequence (GRS), mutation of the GRS motif renders Golden Mosaic virus non-infectious in plants (64). *In vitro* studies utilizing Rep and Rep' of porcine circovirus type 2 have demonstrated that a spliced variant of Rep, Rep' (an identical N-terminal region to Rep with a differing C-terminus), is capable of the same dsDNA binding, ssDNA nicking, and joining activities exhibited by full length Rep (47). In addition to initiating and completing RCR, Rep but not Rep' of PCV1 has been shown to be involved in the regulation of transcription, similar to the roles observed for NS1 (Rep) of parvoviruses (65–67).

The C-terminal portion of Rep in eukaryotic CRESS-DNA viruses contains three motifs conserved in SF3 helicases involved in ATPase activity: Walker A, Walker B, and the C motif (24,52,53,55,68). Walker A contains a p-loop structure involved in NTP binding (55). Walker B coordinates divalent metal ions and hydrolyzes NTP (55). An arginine finger has also been observed to be present in a number of eukaryotic CRESS-DNA Reps which is predicted to be involved in ATPase activity necessary for helicase activity (53). Two additional C-terminal domains, B' and an arginine finger, are generally present in Reps of eukaryotic CRESS-DNA viruses (31,69).

Eukaryotic CRESS-DNA viruses display nucleotide substitution rates that are comparable to those observed by RNA viruses (24,70–72). This observation

in RNA viruses can be explained by the error-prone process of replicating RNA with the viral encoded RNA-dependent RNA polymerase (73). This error prone replication leads to RNA viruses often existing in a quasispecies state which allows RNA viruses to rapidly and continuously explore fitness landscapes (74,75). However, the reasons underlying high nucleotide substitution rates in eukaryotic CRESS-DNA viruses are less clear (24). In part this observed high mutation rate may be explained by the propensity for ssDNA to accumulate mutations due to oxidative damage, potentially while encapsidated (70,76). However, this process alone likely does not account for the difference in expected and observed mutation rates in ssDNA viruses and other yet to be determined mechanisms are likely important (24). Additionally, eukaryotic CRESS-DNA viruses undergo frequent recombination events (77). This may be in part explained by the ambisense nature of some genomes coupled with RCR. When transcription and replication complexes are active on the same DNA template the pausing of cell provided DNA polymerase can lead to template switching events which drive recombination (77). It has previously been noted that these types of events become more common in hosts that are deficient in metal ion transport into the nucleus (78). Finally, the multipartite nature (see above) of some members of the *Geminiviridae* and all *Nanoviridae* members makes reassortment another contributor to evolutionary processes (77).

The historical view of viral gene transfer was that only viruses containing similar genomes underwent recombination or reassortment events, i.e., ssDNA

viruses were confined to these events with other ssDNA viruses. In 2012 the Stedman Lab published the results of a metaviromic survey of Boiling Springs Lake, a hot and acidic lake in Lassen Volcanic National Park located in the volcanic Cascade Range of northern California (79). This analysis revealed the genome of a new type of circular ssDNA virus. The genome of this virus is that of a CRESS-DNA virus based on the apparent circularity of its genome, the presence of a putative Rep gene, and a predicted stem-loop structure with a conserved nonanucleotide sequence that serves as an origin for CRESS-DNA virus RCR (79) (**Fig 1.3**). Interestingly, the CP encoded by this genome is homologous to those encoded by plant infecting ssRNA viruses in the family *Tombusviridae* (79). Named “Boiling Springs Lake RNA-DNA Hybrid Virus” (BSL-RDHV), it has been hypothesized that this virus originated by the acquisition of a capsid gene from an RNA virus through a yet to be demonstrated RNA virus-DNA virus gene exchange (37,79,80). This genome (**Fig 1.3**) represented the first direct evidence that viruses which contain genomes that consist of different nucleic acids may be capable of exchanging genetic material.

While others had predicted that the acquisition of an RNA CP by a CRESS-DNA virus had occurred based on protein fold analysis, BSL-RDHV was the first CRESS-DNA virus whose CP was clearly homologous to those of a ssRNA virus (79,81). Since the initial discovery of a “Hybrid Virus”, approximately 80 additional circular sequences from diverse environments encoding a putative protein homologous to a CRESS-DNA virus Rep and a CP homologous to

tombusvirus CPs have been described in the literature when this work was in its infancy (some of which had been overlooked or ignored) (44,69,82–87). This growing group of viruses have been renamed as cruciviruses, as they imply the crossing between DNA based nucleic acids and ssRNA tombusviruses (69). This naming scheme also removes the potential ambiguity and potential misinterpretations associated with “hybrid virus”. Because BSL-RDHV was so named and published as such prior to this change in naming scheme we have maintained this name in this text and publications.

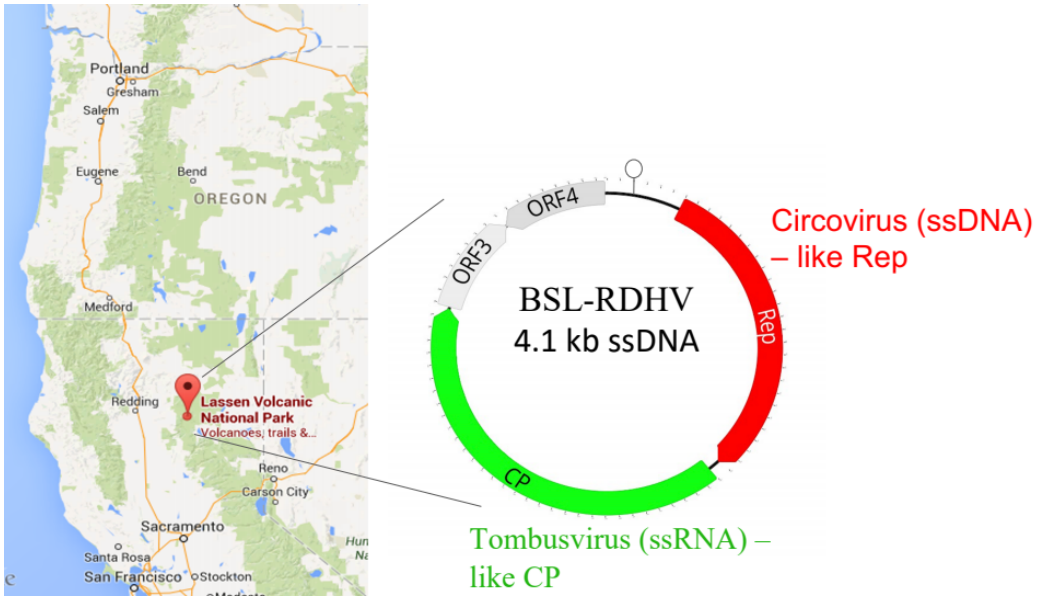

Figure 1.3. Genome of Boiling Springs Lake RNA-DNA Hybrid Virus. The genome is 4.1kb ssDNA. The putative Rep (red) is homologous to those found in CRESS-DNA viruses, while the putative CP (green) is homologous to those found in ssRNA viruses. ORF3 and ORF4 do not contain detectable homology to any publicly available sequences. The stem-loop structure is noted between ORF 4 and Rep.

Cruciviruses have been found associated with forams, alveolates hosted by isopods on the coast of Oregon, arthropods, and in peatland ecosystems (27,69,82,88). But to date no definitive hosts for cruciviruses have been

elucidated. However, the architecture of Rep found in all cruciviruses to date and the work described above seems to suggest that a eukaryotic host is likely (53).

As CRESS-DNA viruses that contain a CP homologous to those found in ssRNA viruses, cruciviruses present a unique opportunity to gain insights to viral evolution. The work presented in chapter two represents the first large scale analysis of crucivirus genomes. Through collaboration with Drs. Francois Enault and Arvind Varsani, leading experts in the field of CRESS-DNA virus metagenomics, we uncovered and analyzed 461 new and crucivirus genomes. In chapter two we show that cruciviruses are highly diverse CRESS-DNA viruses. The putative Rep that is encoded by various cruciviruses span the diversity of CRESS-DNA virus Reps, and as such their placement within established CRESS-DNA virus phylogenies is difficult. We show that this may in part be due to frequent intergenic and intragenic recombination events between cruciviruses and other CRESS-DNA viruses.

Chapter three details efforts to firmly place cruciviruses within CRESS-DNA phylogenies. Even with the discovery of 331 additional crucivirus genomes this task remains difficult. As such I took a more targeted approach in an attempt to group cruciviruses on the basis of shared characteristics of Rep and their origins of replication. I show that methods previously used in other CRESS-DNA viruses seem to not be applicable to cruciviruses.

The work presented in chapter four was originally undertaken to explore the potential biochemical mechanisms that mediated the acquisition of a tombusvirus (RNA virus) CP by a ssDNA virus (80), as it has been hypothesized

that Rep of a ssDNA virus may be involved in this process. While this initial question was not definitively answered, chapter 4 presents the first demonstration of biochemical functions associated with Rep of a crucivirus. I show that purified recombinant Rep of BSL-RDHV is capable of the activities generally associated with initiation and completion of RCR. Specifically, I demonstrate that Rep hydrolyzes ATP indicative of helicase activity, likely binds to the predicted origin of replication, and becomes covalently attached to ssDNA carrying the nonanucleotide of BSL-RDHV indicative of nicking activity. I also show that the use of a pre-formed acceptor oligonucleotide results in the release of Rep from the covalent product mentioned above, assumed to be due to the catalysis of a joining reaction between ssDNAs, typical of the completion of RCR. To our knowledge chapter 4 is also the first demonstrated instance of biochemical activity for a member of the unclassified CRESS-DNA viruses.

Appendix A details the discovery and cloning of three new crucivirus genomes from environmental DNA samples of soil and water from Woodburn, Oregon. I also show that crucivirus genomes were not recovered from a variety of aquatic sediments in Oregon, USA.

References

1. Koonin E V., Senkevich TG, Dolja V V. The ancient virus world and evolution of cells. Biol Direct. 2006;1:1–27.
2. C. E, Keen. A century of phage research: Bacteriophages and the shaping of modern biology. Bioessays. 2015;37(1):139–48.
3. Forterre, P. The origin of viruses and their possible roles in major evolutionary transitions. Virus Res. 2006;117(1):5–16.
4. Forterre P, Prangishvili D. The origin of viruses. Res Microbiol. 2009;160(7):466–72.
5. Krupovic M, Cvirkaite-Krupovic V, Iranzo J, Prangishvili D, Koonin E V. Viruses of archaea: Structural, functional, environmental and evolutionary genomics. Virus Res. 2018;244:181–93.
6. Scholthof KBG, Adkins S, Czosnek H, Palukaitis P, Jacquot E, Hohn T, Hohn B, Saunders K, Candresse T, Ahlquist P, HemmenwayC, Foster GD. Top 10 plant viruses in molecular plant pathology. Mol Plant Pathol. 2011;12(9):938–54.
7. Suttle C. Viruses in the Sea. Nature. 2005;437:356–61.
8. Bergh Ø, Børsheim KY, Bratbak G, Heldal M. High abundance of viruses found in aquatic environments. Nature. 1989;340:467–8.
9. Wommack KE, Colwell RR. Virioplankton: Viruses in aquatic ecosystems. Microbiol Mol Biol Rev [Internet]. 2000;64(1):69–114.
10. Suttle CA. Marine viruses - Major players in the global ecosystem. Nat Rev Microbiol. 2007;5(10):801–12.
11. Proctor LM, Fuhrman JA. Mortality of marine bacteria in response to enrichments of the virus size fraction from seawater. Mar Ecol Prog Ser. 1992;87(3):283–93.
12. Procto, LM, Fuhrman JAL. Viral mortality of marine bacteria and cyanobacteria. Nature. 1990;343:60–62.
13. Nasir A, Forterre P, Kim KM, Caetano-Anolles G. The distribution and impact of viral lineages in domains of life. Front Microbiol. 2014;5:1–5.
14. Suttle CA, Chan AM, Cottrell MT. Infection of phytoplankton by viruses and reduction of primary productivity. Nature. 1990;347:467–9.
15. Brister JR, Ako-Adjei D, Bao Y, Blinkova O. NCBI viral Genomes resource. Nucleic Acids Res. 2015;43(D1):D571–7.
16. Delwart EL. Viral metagenomics. Rev Med Virol. 2007;17(2):115–31.

17. Forterre P. The two ages of the RNA world, and the transition to the DNA world: A story of viruses and cells. Biochimie. 2005;87(9–10):793–803.

18. Rice G, Tang L, Stedman K, Roberto F, Spuhler J, Gillitzer E, et al. The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans all domains of life. Proc Natl Acad Sci U S A. 2004;101(20):7716–20.

19. Koonin E V., Dolja V V., Krupovic M. Origins and evolution of viruses of eukaryotes: The ultimate modularity. Virology. 2015;479–480:2–25.

20. Holmes EC. What does virus evolution tell us about virus origins? J Virol. 2011;85(11):5247–51.

21. Baltimore D. Expression of animal virus genomes. Bacteriol Rev. 1971;35(3):235–41.

22. Krupovic M, Koonin E V. Multiple origins of viral capsid proteins from cellular ancestors. Proc Natl Acad Sci U S A. 2017;114(12):E2401-E2410

23. Durzyńska J, Goździcka-Józefiak A. Viruses and cells intertwined since the dawn of evolution Emerging viruses. Virol J. 2015;12(1):1–10.

24. Zhao L, Rosario K, Breitbart M, Duffy S. Eukaryotic circular rep-encoding single-stranded DNA (CRESS DNA) viruses: ubiquitous viruses with small genomes and a diverse host range. 1st ed. Advances in Virus Research. Elsevier Inc.; 2018:1–63.

25. Varsani A, Krupovic M. Smacoviridae: a new family of animal-associated single-stranded DNA viruses. Arch Virol. 2018;163(7):2005–15.

26. Chabi-Jesus C, Najar A, Fontenele RS, Kumari SG, Ramos-González PL, Freitas-Astúa J, et al. Viruses representing two new genomovirus species identified in citrus from Tunisia. Arch Virol. 2020;165(5):1225–9.

27. Bistolas K, Besemer R, Rudstam L, Hewson I. Distribution and inferred evolutionary characteristics of a chimeric ssDNA virus associated with intertidal marine isopods. Viruses. 2017;9(12):361.

28. Levy H, Fontenele RS, Harding C, Suazo C, Kraberger S, Schmidlin K, et al. Identification and distribution of novel cressdnaviruses and circular molecules in four penguin species in South Georgia and the Antarctic peninsula. Viruses. 2020;12(9):1-21.

29. Dayaram A, Goldstien S, Zawar-Reza P, Gomez C, Harding JS, Varsani A. Novel ssDNA virus recovered from estuarine Mollusc (*Amphibola crenata*) whose replication associated protein (Rep) shares similarities with Rep-like sequences of bacterial origin. J Gen Virol. 2013;94(Part 5):1104–10.

30. Dean FB, Nelson JR, Giesler TL, Lasken RS. Rapid amplification of plasmid and phage DNA using Phi29 DNA polymerase and multiply-primed rolling circle amplification. Genome Res. 2001;11(6):1095–9.

31. Krupovic M, Varsani A, Kazlauskas D, Breitbart M, Delwart E, Yutin N, et al. Cressdnaviricota: a virus phylum unifying 7 families of Rep-encoding viruses with single- stranded, circular DNA genomes. 2020;94(12):e00582-20.

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

33. Kazlauskas D, Varsani A, Krupovic M. Pervasive chimerism in the replication-associated proteins of uncultured single-stranded DNA viruses. Viruses. 2018;10(4):1–11.

34. Esendugue G, Fonsah, Yu C, Diffie S, Srinivansan R, Riley D. Economic productivity and profitability analysis for whiteflies and tomato yellow leaf curl virus (TYLCV) Management Options. J Agric Environ Sci. 2018;7(1):1–9.

35. Chae C. A review of porcine circovirus 2-associated syndromes and diseases. Vet J. 2005;169(3):326–36.

36. Picó B, Díez MJ, Nuez F. Viral diseases causing the greatest economic losses to the tomato crop. II. The tomato yellow leaf curl virus - A review. Sci Hortic (Amsterdam). 1996;67(3–4):151–96.

37. Stedman KM. Deep Recombination: RNA and ssDNA virus genes in DNA virus and host genomes. Annu Rev Virol. 2015;2(1):203–17.

38. Dennis TPW, Flynn PJ, Marciel de Souza W, Singer JB, Moreau CS, Wilson SJ, et al. Insights into circovirus host range from the genomic fossil record. J Virol. 2018;92(16):e00145145-18.

39. Dennis TPW, de Souza WM, Marsile-Medun S, Singer JB, Wilson SJ, Gifford RJ. The evolution, distribution and diversity of endogenous circoviral elements in vertebrate genomes. Virus Res. 2019;262(March)15-23.

40. Krupovic M, Forterre P. Single-stranded DNA viruses employ a variety of mechanisms for integration into host genomes. Ann N Y Acad Sci. 2015;1341(1):41–53.

41. Liu H, Fu Y, Li B, Yu X, Xie J, Cheng J, et al. Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. BMC Evol Biol. 2011;26(11).

42. Filloux D, Murrell S, Koohapitagtam M, Golden M, Julian C, Galzi S, et al. The genomes of many yam species contain transcriptionally active endogenous geminiviral sequences that may be functionally expressed. Virus Evol. 2015;1(1):1–17.

43. Saunders K, Richardson J, Lawson DM, Lomonossoff GP. Requirements for the packaging of geminivirus circular single-stranded DNA: Effect of DNA length and coat protein sequence. Viruses. 2020;12(11).

44. Rosario K, Duffy S, Breitbart M. A field guide to eukaryotic circular single-stranded DNA viruses: Insights gained from metagenomics. Arch Virol. 2012;157(10):1851–71.

45. Saeed M, Zafar Y, Randles JW, Rezaian MA. A monopartite begomovirus-associated DNA β satellite substitutes for the DNA B of a bipartite begomovirus to permit systemic infection. J Gen Virol. 2007;88(10):2881–9.

46. Gronenborn B. Nanoviruses: Genome organisation and protein function. Vet Microbiol. 2004;98(2):103–9.

47. Steinfeldt T, Finsterbusch T, Mankertz A. 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. 2006;80(13):6225–34.

48. Cheung AK. Rolling-circle replication of an animal circovirus genome in a theta-replicating bacterial plasmid in *Escherichia coli*. 2006;80(17):8686–94.

49. Selth LA, Randles JW, Rezaian MA. *Agrobacterium tumefaciens* supports DNA replication of diverse geminivirus types. FEBS Lett. 2002;516(1–3):179–82.

50. Timchenko T, de Kouchkovsky F, Katul L, David C, Vetten HJ, Gronenborn B. A single Rep protein initiates replication of multiple genome components of faba bean necrotic yellows virus, a single-stranded DNA virus of plants. J Virol. 1999;73(12):10173–82.

51. Laufs J, Jupin I, David C, Schumacher S, Heyraud-Nitschke F, Gronenborn B. Geminivirus replication: Genetic and biochemical characterization of Rep protein function, a review. Biochimie. 1995;77(10):765–73.

52. Ilyina T V, Koonin E V. Conserved sequence motifs in the initiator proteins for rolling circle dna-replication encoded by diverse replicons from eubacteria, eukaryotes and archaebacteria. Nucleic Acids Res. 1992;20(13):3279–85.

53. Chandler M, Cruz F De, Dyda F, Hickman AB. Breaking and joining single-stranded DNA : the HUH endonuclease superfamily. Nat Rev Microbiol. 2013;11(8):525–38.

54. Yang W. Topoisomerases and site-specific recombinases: Similarities in structure and mechanism. Crit Rev Biochem Mol Biol. 2010;45(6):520–34.

55. Hickman AB, Dyda F. Binding and unwinding: SF3 viral helicases. Curr Opin Struct Biol. 2005 Feb;15(1):77-85.

56. Snider J, Thibault G, Houry WA. The AAA+ superfamily of functionally diverse proteins. Genome Biol. 2008;9(4):1–8.

57. Londoño A, Riego-Ruiz L, Argüello-Astorga GR. DNA-binding specificity determinants of replication proteins encoded by eukaryotic ssDNA viruses are adjacent to widely separated RCR conserved motifs. Arch Virol. 2010;155(7):1033–46.

58. Vega-Rocha S, Byeon IJL, Gronenborn B, Gronenborn AM, Campos-Olivas R. Solution structure, divalent metal and DNA binding of the endonuclease domain from the replication initiation protein from porcine circovirus 2. J Mol Biol. 2007;367(2):473–87.

59. Luo G, Zhu X, Lv Y, Lv B, Fang J, Cao S, et al. Crystal structure of the dimerized N-terminus of porcine circovirus type 2 replicase protein reveals a novel antiviral interface. J Virol. 2018;92(18):JVI.00724-18.

60. Kai J, Chiaolong C, Jian H, Wu S, Yi S, Chi L, et al. Characterization of the endonuclease activity of the replication - associated protein of beak and feather disease virus. Arch Virol. 2019 Aug;164(8):2091-2106

61. Chatterji A, Padidam M, Beachy RN, Fauquet CM. Identification of replication specificity determinants in two strains of tomato leaf curl virus from New Delhi. J Virol. 1999;73(7):5481–9.

62. Steinfeldt T, Finsterbusch T, Mankertz A. Rep and Rep' protein of porcine circovirus type 1 bind to the origin of replication in vitro. Virology. 2001;291(1):152–60.

63. Cheung AK. 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. 2007;363(1):229–35.

64. Nash TE, Dallas MB, Reyes MI, Buhrman GK, Ascencio-Ibanez JT, Hanley-Bowdoin L. Functional analysis of a novel motif conserved across geminivirus Rep proteins. J Virol. 2011;85(3):1182–92.

65. Cheung AK. Porcine circovirus: Transcription and DNA replication. Virus Res. 2012;164(1–2):46–53.

66. Doerig C, Hirt B, Antonietti JP, Beard P. Nonstructural protein of parvoviruses B19 and minute virus of mice controls transcription. J Virol. 1990;64(1):387–96.

67. Cotmore SF, Gottlieb RL, Tattersall P. Replication initiator protein NS1 of the parvovirus minute virus of mice binds to modular divergent sites distributed throughout duplex viral DNA. J Virol. 2007;81(23):13015–27.

68. Walker JE, Saraste M, Runswick MJ, Gay NJ. Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and

other ATP-requiring enzymes and a common nucleotide binding fold. EMBO J. 1982;1(8):945–51.

69. Quaiser A, Krupovic M, Dufresne A, Roux S. Diversity and comparative genomics of chimeric viruses in Sphagnum- dominated peatlands. 2016;2(2):1–8.

70. Duffy S, Holmes EC. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. J Virol. 2008;82(2):957–65.

71. Duffy S, Holmes EC. Validation of high rates of nucleotide substitution in geminiviruses: Phylogenetic evidence from East African cassava mosaic viruses. J Gen Virol. 2009;90(6):1539–47.

72. Van Der Walt E, Martin DP, Varsani A, Polston JE, Rybicki EP. Experimental observations of rapid maize streak virus evolution reveal a strand-specific nucleotide substitution bias. Virol J. 2008;5:1–11.

73. Duffy S. Why are RNA virus mutation rates so damn high? PLoS Biol. 2018;16(8):1–6.

74. Eigen M. On the nature of virus quasispecies. Trends Microbiol. 1996;4(6):216–8.

75. Domingo E, Sabo D, Taniguchi T, Weissmann C. Nucleotide sequence heterogeneity of an RNA phage population. Cell. 1978;13(4):735–44.

76. Ritchie PA, Anderson IL, Lambert DM. Evidence for specificity of psittacine beak and feather disease viruses among avian hosts. Virology. 2003;306(1):109–15.

77. Martin DP, Biagini P, Lefeuvre P, Golden M, Roumagnac P, Varsani A. Recombination in eukaryotic single stranded DNA viruses. Viruses. 2011;3(9):1699–738.

78. Interactions V, Kovalev N, Pogany J, Nagy PD. Interviral recombination between plant, insect, and fungal RNA viruses: Role of the intracellular Ca2/Mn2 pump. 2020;94(1):1–20.

79. Diemer GS, Stedman KM. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. Biol Direct. 2012;7(13):1–14.

80. Stedman K. Mechanisms for RNA capture by ssDNA viruses: Grand theft RNA. J Mol Evol. 2013;76(6):359–64.

81. Gibbs MJ, Weiller GF. Evidence that a plant virus switched hosts to infect a vertebrate and then recombined with a vertebrate-infecting virus. Proc Natl Acad Sci U S A. 1999;96(14):8022–7.

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

83. Kazlauskas D, Dayaram A, Kraberger S, Goldstien S, Varsani A, Krupovic M. Evolutionary history of ssDNA bacilladnaviruses features horizontal acquisition of the capsid gene from ssRNA nodaviruses. Virology. 2017;504(January):114–21.

84. Rosario K, Nilsson C, Lim YW, Ruan Y, Breitbart M. Metagenomic analysis of viruses in reclaimed water. Environ Microbiol. 2009;11(11):2806–20.

85. Whon TW, Kim M-S, Roh SW, Shin N-R, Lee H-W, Bae J-W. Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere. J Virol. 2012;86(15):8221–31.

86. Hewson I, Ng G, Li WF, LaBarre BA, Aguirre I, Barbosa JG, et al. Metagenomic identification, seasonal dynamics, and potential transmission mechanisms of a Daphnia-associated single-stranded DNA virus in two temperate lakes. Limnol Oceanogr. 2013;58(5):1605–20.

87. Mcdaniel LD, Rosario K, Breitbart M, Paul JH. Comparative metagenomics: Natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. Environ Microbiol. 2014 Feb;16(2):570-85.

88. de la Higuera I, Torrance EL, Pratt AP, Kasun GW, Maluenda A, Stedman KM. Genome sequences of three cruciviruses found in the Willamette Valley. (Oregon). 2019; (June):18–20.

Chapter Two

Unveiling Crucivirus Diversity by Mining Metagenomic Data

This chapter is modified from: **Unveiling Crucivirus Diversity by Mining Metagenomic Data.** Ignacio de la Higuera, George W. Kasun, Ellis L.Torrance, Alyssa A. Pratt, Amberlee Maluenda, Jonathan Colombet, Maxime Bisseux, Viviane Ravet, Anisha Dayaram, Daisy Stainton, Simona Kraberger, Peyman Zawar-Reza, Sharyn Goldstien, James V. Briskie, Robyn White, Helen Taylor, Christopher Gomez, David G Ainley, Jon S. Harding, Rafaela S. Fontenele, Joshua Schreck, Simone G. Ribeiro, Stephen A. Oswald, Jennifer M. Arnold, François Enault, Arvind Varsani, and Kenneth M. Stedman. *mBio Sep 2020, 11 (5) e01410-20; DOI: 10.1128/mBio.01410-20*

Abstract

Cruciviruses are a novel group of circular Rep-encoding single-stranded DNA (ssDNA) (CRESS-DNA) viruses that encode capsid proteins 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 capsid genes between DNA and RNA viruses. Here, we identified and characterized 451 new crucivirus genomes and 10 capsid-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 intergenic recombination. Additionally, we suggest members of the stramenopiles/alveolates/Rhizaria supergroup as possible crucivirus hosts. Altogether, we provide a comprehensive and descriptive characterization of cruciviruses.

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 many 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; Rep being the replication-associated protein) viruses may have been derived from plasmids on multiple occasions by acquiring capsid genes from RNA viruses (9–11). Eukaryotic CRESS-DNA viruses, recently classified into the phylum *Cressdnaviricota*, 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*, in addition to vast numbers of unclassified viruses (12–14) (**Fig. 1.1**). Universal to all CRESS-DNA viruses is the initiator of rolling circle replication protein (Rep), which is involved in the initiation and completion of the viral genome replication through rolling-circle replication (RCR) (12) (**Fig. 1.2**). Rep homologues are also encoded in plasmids (13–15) . Some pathogenic CRESS-DNA viruses are agriculturally important, such as porcine circoviruses, and nanoviruses and geminiviruses that infect a wide range of plant hosts (14). However, many CRESS-DNA viruses have been identified in apparently healthy organisms, and metagenomic studies have revealed their presence in most environments (13).

In 2012, a metagenomic survey of a hot and acidic lake in the volcanic Cascade Range of the western United States uncovered a new type of circular DNA virus (16). The genome of this virus appears to make it 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 RCR (16–18) (**Figs. 1.1 and 1.2**). Interestingly, the amino acid sequence of the capsid protein encoded by this genome resembles those encoded by RNA viruses in the family *Tombusviridae* (16). 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 (16,19). Since the discovery of this putatively "chimeric virus," 80 circular sequences encoding a Rep that shares homology to ssDNA viruses and a capsid protein that shares homology to tombusvirus capsid proteins have been found in different environments around the globe (20–32). This growing group of viruses have been branded "cruciviruses," as they imply crossing between CRESS-DNA viruses and RNA tombusviruses (31). Cruciviruses have been found associated with forams (21), alveolates hosted by isopods (30), arthropods (20,26) and in peatland ecosystems (31), but no host for cruciviruses has been elucidated to date.

The circular genome of previously described cruciviruses is variable in size, ranging from 2.7 to 5.7 kb, and often contains open reading frames (ORFs) in addition to the Rep and capsid genes, which have been found in either a unisense or an ambisense orientation (21,31). The function of additional

crucivirus ORFs is unclear due to their lack of sequence similarity with any characterized protein. The genome replication of CRESS-DNA viruses is initiated by the Rep protein, which binds to direct repeats present just downstream of the stem of the *ori*-containing stem-loop structure and nicks the ssDNA (33,34). The exposed 3′ OH serves as a primer for cellular enzymes to replicate the viral genome via RCR (34,35). 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 (34–37).

Rep has a domain in the N terminus which belongs to the HUH endonuclease superfamily (15). This family of proteins is characterized by a metal ion binding 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 (15,33,38). CRESS-DNA virus Reps also contain a third conserved motif in the N-terminal portion of the protein (motif I), likely responsible for double-stranded DNA (dsDNA) binding specificity (39). In many CRESS-DNA viruses, the HUH of motif II has been replaced with a similar motif that lacks the second histidine residue (e.g., circoviruses have replaced HUH with HLQ) (10,15). The C-terminal portion of eukaryotic CRESS-DNA virus Reps contains a superfamily 3 helicase domain (S3H) that may be responsible for unwinding dsDNA replicative intermediates (40,41). 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 (42). 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 (21,31). In some cruciviruses, these differences in phylogeny have been observed between the individual domains of a single Rep protein (25,31). The apparent polyphyly of crucivirus Reps suggests recombination events involving cruciviruses and other CRESS-DNA viruses, and even intragenic recombination within Reps (21,25).

All characterized CRESS-DNA viruses package their DNA into small capsids with icosahedral symmetry or their geminate variants (twinned particles found in *Geminiviridae)*, built from multiple copies of the capsid protein encoded in their genome (14,43). The capsid protein of these CRESS-DNA viruses appears to fold into an eight-strand β-barrel that conforms to the single jelly-roll (SJR) architecture, which is also commonly found in eukaryotic RNA viruses (44). The capsid protein 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 capsid protein of tombusviruses (16,21,25). Three domains can be distinguished in tombusviral capsid proteins (45). From the N to the C terminus, they are (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 (46,47); (ii) the shell or S-domain containing the single jelly-roll fold and the architectural base of the capsid (45); and (iii) the protruding or P-domain, which decorates the surface of the virion and is involved in host transmission (48). In tombusviruses, the S-domains of 180 capsid protein

subunits interact with each other to assemble around the viral RNA in a T=3 fashion, forming an ~35-nm virion (45,49).

The discovery of cruciviruses by our group 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 (CruV) genomes and cruci-like circular genetic elements (CruCGEs) identified in metagenomic data obtained from different environments and organisms. This study provides a comprehensive analysis of this greatly expanded data set and explores the extent of cruciviral diversity—mostly due to Rep heterogeneity—impacted by rampant recombination.

Methods

Assembly and recovery of crucivirus genomes: A total of 461 crucivirus-related sequences were identified from 1,168 metagenomic surveys (available as supplementary material in de la Higuera et. al. 2020). One thousand one hundred sixty-seven viromes from 57 published data sets and one unpublished virome were obtained from different environments: aquatic systems (freshwater, seawater, hypersaline ponds, thermal springs, and hydrothermal vents), engineered systems (bioreactor and food production), and eukaryote-associated flora (human, insect and other animal feces, human saliva and fluids, cnidarians, and plants). Raw reads from metagenomes were assembled using multiple different programs by our collaborators. New potential cruciviral genomes were

identified from these assembled viromes by screening circular contigs for the presence of capsid proteins from previously known cruciviruses and tombusviruses, using a BLASTx bit-score threshold of 50. The selected genomes are assumed to be complete and circular due to the terminal redundancy identified in the *de novo*-assembled genomes. These assembled potential crucivirus genomes were then passed to our group for annotation and analysis.

The sequences of five potential crucivirus genomes (CruV-240, CruV-300, CruV-331, CruV-338, and CruV-367) were retrieved as assembled contigs from the Joint Genome Institute (JGI)'s IMG/VR repository (50), by searching scaffolds with a function set including the protein family pfam00729, corresponding to the S-domain of tombusvirus capsids. Sequences with an RNA dependent RNA polymerase coding region were excluded (as this suggests an RNA genome), and the circularity of the sequences, as well as the presence of an ORF encoding a tombusvirus-like capsid protein, was confirmed with Geneious 11.0.4

Annotation of crucivirus genomes: The 461 cruciviral sequences were annotated and analyzed in Geneious 11.0.4. Coding sequences (CDSs) were semiautomatically annotated from a custom database of protein sequences of published cruciviruses, and close homologues obtained from GenBank, using Geneious 11.0.4's annotation function with a 25% nucleotide similarity threshold. Annotated CDSs were rechecked with the 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 (51) 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 homologues and no overlap with capsid protein or Rep-like ORFs were annotated as “putative ORFs.”

Putative origin of replication annotation: Stem-loop structures which could serve as an origin of replication (34,53) for circular ssDNA viruses were identified and annotated using StemLoop-Finder developed by Alyssa Pratt (Pratt, Torrance, Kasun, Stedman and de la Higuera, in revision). The 461 cruciviral sequences were scanned for the presence of conserved nonanucleotide motifs described for other CRESS-DNA viruses (NANTANTAN, NAKWRTTAC, TAWWDHWAN, and TRAKATTRC). The integrated ViennaRNA 2.0 library (54) was used to predict secondary structures of DNA around the detected motif, including the surrounding 15 to 20 nucleotides on either side. Predicted structures with a stem longer than 4 bp 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 bp 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 they attained a greater score, as well as those that appeared to have a nonanucleotide within four bases of their stem-loop structure’s first or last nucleotide. These stem-loop annotations were then visually inspected and checked using the Mfold webserver (http://www.unafold.org/).

Conservation analysis and visualization: The pairwise identity between the protein sequence from translated cruciviral genes was calculated with SDTv1.2 (55), with MAFFT alignment option for capsid proteins and S-domains and MUSCLE alignment options for Reps. The raw data were further analyzed with Prism v8.4.3.

Multiple sequence alignments: Capsid protein sequences were aligned using MAFFT (56) 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 (57). Rep alignments were manually inspected and corrected in Geneious 11.0.4 and trimmed using TrimAl v1.3 (58) with a *strict plus* setting. To produce separate alignments of the endonuclease and helicase domains, the full-length trimmed alignments were split at the first residue of the Walker A motif (42).

Phylogenetic trees: Phylogenetic trees containing the entire data set of cruciviral sequences were built in Geneious using the FastTree plugin (59). For the analysis of sequence subsets, trees were inferred with the PhyML 3.0 web server (http://www.atgc-montpellier.fr/phyml/) (60) using an aLRT SH-like support (61) and automatic model selection.

Intergenic and Intragenic Recombination Detection: Tanglegrams were built using Dendroscope v3.5.10 (62) to compare the phylogenies between different genes (CP and Rep) or domains (endonuclease and helicase of Rep) within the same set of crucivirus genomes.

Sequence similarity networks. A total of 540 capsid amino acid sequences and 600 Rep amino acid sequences were uploaded to the EFI–EST web server (https://efi.igb.illinois.edu/efi-est/) (63). A specific alignment score cutoff was established for each data set (E value < 10^{-20} for CP and E value < 100^{-10} for Rep) and xgmml files generated by EFI-EST were visualized and edited in Cytoscape v3.7.2.

Sequence logos: Sequence logos representing the frequency of bases in nonanucleotides at the putative origin of replication and amino acid residues in conserved Rep motifs were made using the WebLogo server (http://weblogo.threeplusone.com/).

Results and Discussion

Expansion of the crucivirus group: To broaden our understanding of the diversity and relationships of cruciviruses, 461 uncharacterized circular DNA sequences containing predicted coding sequences (CDSs) with sequence similarity to the capsid protein 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 (available as supplementary material in de la Higuera et. al. 2020).

The new crucivirus 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 (16, 20–32). The average GC content of the

newly described cruciviral sequences is 42.9% ± 4.9% (**Fig. 2.1B**) with genome lengths spanning from 2,474 to 7,947 bases (**Fig. 2.1A**), some exceeding the size of described bacilladnaviruses (≤6,000 nucleotides [nt] (64), the largest CRESS-DNA viruses known (12).

Of the 461 sequences that contain a capsid protein ORF, 451 have putative coding regions with sequence similarity to Rep of CRESS-DNA viruses (10). The capsid protein 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 10 CruCGEs with no clear Rep gene. Five of these CruCGEs contain a predicted origin of RCR, indicating that they are circular genomes that undergo rolling-circle replication characteristic of other CRESS-DNA virus genomes (17,18).

Prediction of stem-loop structures: Stem-loop structures with conserved nonanucleotide motifs as putative origins of replication were predicted and annotated in 277 cruciviral sequences with StemLoop-Finder (Pratt, Torrance, Kasun, Stedman and de la Higuera, in revision) (**Fig. 2.1C**). 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 predicted stem-loop with a nonanucleotide with a NANTANTAN pattern, with the most common sequence being the canonical circovirus motif TAGTATTAC (**Fig. 2.1C**), found in 64 of the genomes (65). The majority of the 54 sequences that do not correspond to NANTANTAN contain a TAWWDHWAN nonanucleotide motif, typical of genomoviruses (66). The

frequency of bases at each position in the nonanucleotide sequence is given in Figure 1C and reflects similarity to motifs found in other eukaryotic CRESS-DNA viruses (10).

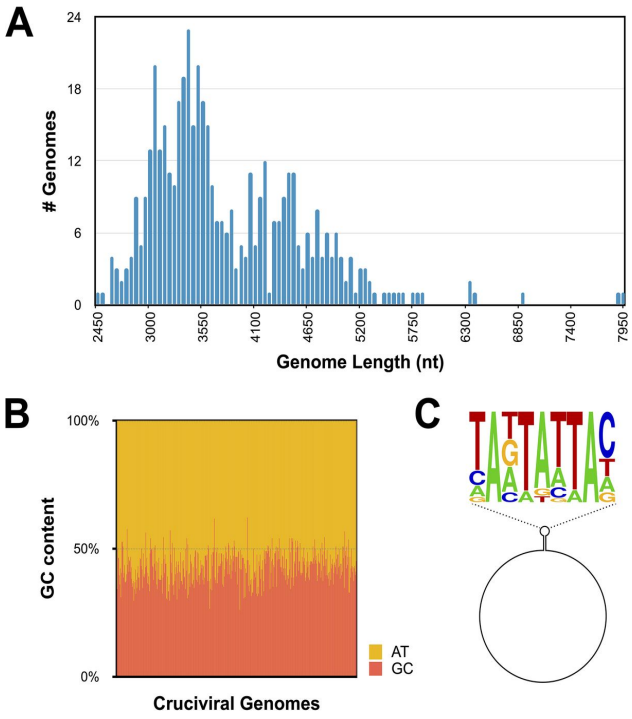

Figure 2.1. Genome properties of 461 new cruciviral circular sequences. A) Histogram of cruciviral genome lengths categorized in 50-nt bins. **B)** Percentage of G+C content versus A+T in each of the sequences described in this study. **C)** Relative abundance of nucleotides in the conserved nonanucleotide sequence of the 211 stem-loops and putative origins of replication represented predicted with StemLoop-Finder (A. A. Pratt et al., in revision) in Sequence Logo format. Constructed by Ignacio de la Higuera.

Crucivirus Rep: The Reps of eukaryotic CRESS-DNA viruses typically contain an N-terminal endonuclease domain characterized by motifs I, II, and III belonging to the HUH (two histidine residues separated by a hydrophobic residue) endonuclease superfamily (15). Members of the HUH endonuclease family catalyze nicking and joining reactions to initiate and complete RCR, respectively (15,33,35). In the case of eukaryotic CRESS-DNA viruses the N-

terminal endonuclease domain is fused to a C-terminal helicase domain with Walker A and B motifs, motif C, and an Arg finger (13–15) (**Fig. 2.2**). Of the 461 sequences that contain a capsid protein ORF, 451 have putative coding regions with sequence similarity to Rep of CRESS-DNA viruses. The remaining ~2% correspond to 10 CruCGEs with no clear Rep CDS. The majority (85.9%) of the crucivirus genomes described in the first data set of 461 genomes contain all of the expected Rep motifs (**Fig. 2.2**). However, five genomes (CruCGE-110, CruCGE-296, CruCGE-436, CruCGE-471, and CruCGE-533) with overall sequence homology to other previously annotated and publicly deposited Reps (35.8, 32.7, 49.7, 60.2, and 57.2% pairwise identity with other putative Reps in the databases, respectively) lack any detectable conserved motifs within their sequence. Thus, these sequences are considered capsid-encoding crucivirus-like circular genetic elements (CruCGEs).

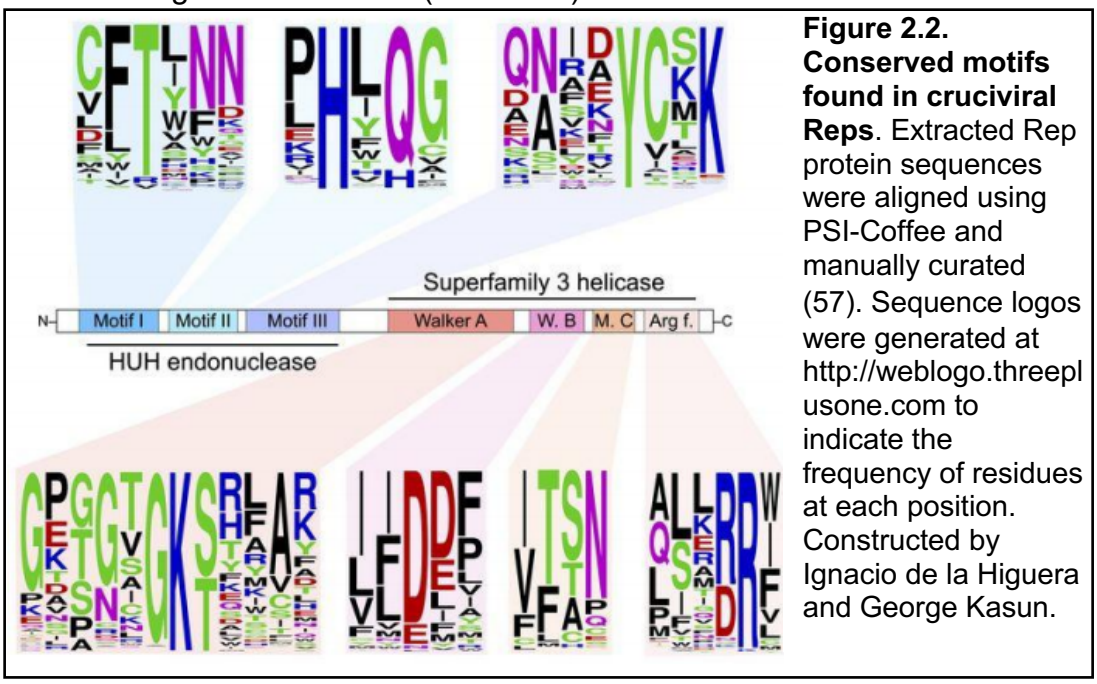

Figure 2.2. Conserved motifs found in cruciviral Reps. Extracted Rep protein sequences were aligned using PSI-Coffee and manually curated (57). Sequence logos were generated at http://weblogo.threeplusone.com to indicate the frequency of residues at each position. Constructed by Ignacio de la Higuera and George Kasun.

Possibly, some cruciviruses are bipartite (5 CrucGE mentioned above with no Rep CDS), requiring more than one virion-encapsidated DNA molecule to bring about a successful infection. The begomoviruses are well characterized members of the *Geminiviridae* in which most members' genomes are bipartite, requiring two distinct, approximately 2.5kbp DNAs (DNA-A and DNA-B) to initiate a successful infection (67). DNA-A contains genes for CP and Rep and DNA-B contains genes involved in intra-host spread (movement protein) and host symptom development (67).These two DNAs are distinct except for an approximately 200bp common region (CR), which contains the origin of replication (67). Five of these CruCGE's contain a predicted origin of RCR, further indicating that they are indeed circular genomes that undergo RCR characteristic of other CRESS-DNA virus genomes.

While Rep and CP of bipartite begomoviruses are encoded on the same DNA, the bipartite *nanoviruses* exhibit a multipartite genome arrangement in which Rep and CP are on distinct circular ssDNA molecules (68). Moreover, some ssRNA tombunodaviruses, including *Plasmopara halstedi*i virus A and *Sclerophthora macrospora* virus A—viruses that contain the capsid sequences most similar to cruciviral capsids—also have multipartite genomes (69). These observations support the potential for multipartite genomes of some cruciviruses. Unfortunately, truly robust or definitive methods do not currently exist to match different sequences belonging to the same multipartite virus in metagenomes, making identification of multipartite or segmented viruses from metagenomic data challenging.

Motif II of the endonuclease domain, which contains the HUH sequence and is located on a beta sheet (70), was identified in 441 of the genomes, 95.2% of which had an alternative to HUH, with the most common arrangement being HUQ (70.0%), also found in circoviruses and nanoviruses (10,15, 28) (**Fig. 2.2**). Crucivirus motif II deviates from the HUH motif by additionally replacing the second hydrophobic residue (U) with a polar amino acid in 26.2% of genomes (**Fig. 2.2**), with 53 Reps with the sequence HYQ (12.0%) also found in smacoviruses (10,29,42).

Motif III lies on an alpha helix and contains the catalytic tyrosine residue responsible for initiating and terminating viral DNA replication, by nicking ssDNA in the conserved nonanucleotide sequence, and subsequent ligation of replicated ssDNA genomes (15,34,35,53,70). In eukaryotic CRESS-DNA viruses this motif typically contains one tyrosine residue and as such their Rep proteins are broadly classified as being “Y1” members of the HUH superfamily (15). Other members of the family include transposases and bacterial relaxases (e.g. MobP, MobQ, MobV) that exhibit this same Y1 architecture (15). Other members of the HUH superfamily Reps contain two tyrosine residues in motif II (“Y2”) such as those found in adeno-associated virus (AAV), phi-x-174, as well as a number of transposases and MobF relaxases (15,35). In some cases of the Y2 architecture only one of the two tyrosine residues is required for catalysis, as in the case of RepB from pMV158, a bacterial plasmid as well as Rep of AAV (71). Other Y2 Reps require both conserved tyrosine residues for complete enzymatic activity as is the case for MobF relaxases (15). One tyrosine residue is involved in initiation

via nicking and a second is involved in sealing of newly replicated genomes (72). We identified 30 crucivirus genomes that conform to this Y2 arrangement, while the rest exhibit a Y1 motif more consistent with other eukaryotic-CRESS DNA viruses. This Y2 arrangement has also been noted in a number of members of the *Smacoviridae* and *Genomoviridae* (14,73,74). Future biochemical and structural studies would be needed to examine the actual mechanisms of RCR initiation and termination in these cruciviruses to confirm if both tyrosine residues are necessary for replication, making them “true” Y2 members of the HUH endonuclease family.

Thirteen putative Reps were identified in these crucivirus genomes that lack all four motifs typically found in S3H helicases (e.g., CruV-166, CruV-202, and CruV-499). 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 (75). 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 through yet to be characterized mechanisms. Alternatively, these activities may be provided by host factors (76), or by a viral replication-enhancer protein—as is the case with the AC3 protein of begomoviruses (77).

We identified 36 crucivirus genomes whose putative *Rep* genes contain in-frame stop codons or in which 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 (51) were identified in these sequences, and the putatively spliced Reps were 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.

No geminivirus Rep sequence (GRS) motifs—which have been biochemically characterized as necessary for geminivirus replication (78) and have also been found in genomoviruses (66)—were detected in Reps in our data set. We were unable to detect any unique conserved Rep motifs present in cruciviruses that are absent in other eukaryotic CRESS-DNA viruses. However, a number of crucivirus Reps contain a large number of amino acids in their N-terminus prior to Motif I. For example, Rep of BSL-RDHV contains 86 amino acids in its N-terminus prior to the first residue of Motif I, while the putative Rep of CruV-484 contains 156 amino acids in this region. Other eukaryotic CRESS-DNA virus Reps generally have less than 40 amino acids preceding the start of Motif I (based on our alignments). This N-terminal region, while seemingly unique to cruciviruses, does not contain conserved amino acid residues. This N-terminal region may be removed via splicing prior to translation. Alternatively, this N-terminal region, seemingly unique to cruciviruses, may be an artifact related to our annotation strategies. Regardless, the general conservation of Rep motifs in these newly described cruciviruses suggests that most are active in rolling-circle replication.

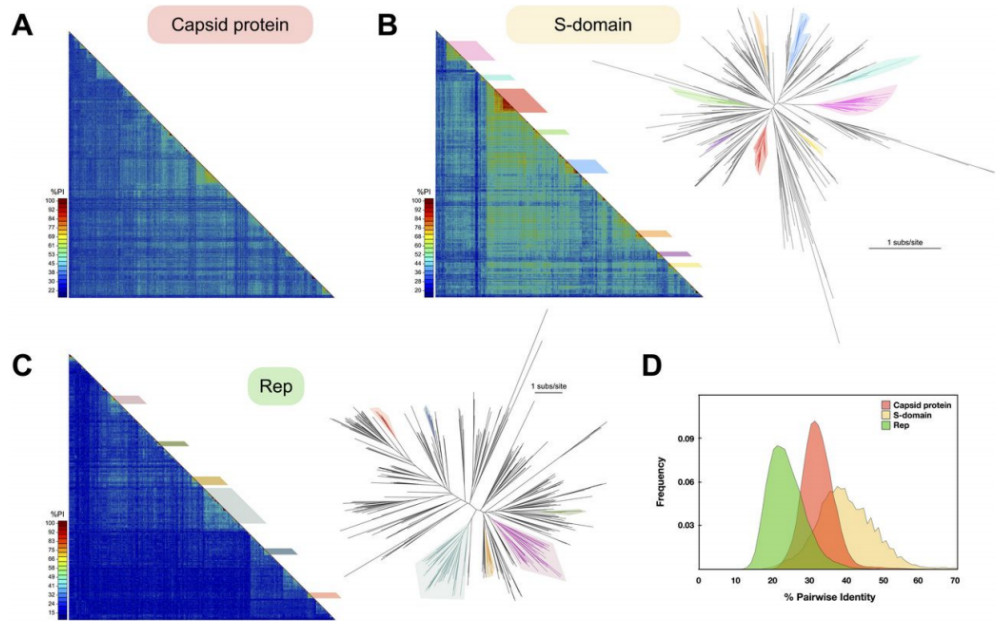

Figure 2.3. **Diversity of cruciviral proteins. A)** Capsid protein diversity. Pairwise amino acid identity (%PI) between the capsid proteins predicted for 461 cruciviral sequences. The alignment and analysis were carried out with SDT, using the integrated MAFFT algorithm. **B)** S-domain diversity. (Left) Pairwise identity matrix between the capsid protein predicted S-domains of the 461 sequences described in this study. The alignment and analysis were carried out with SDT, using the integrated MAFFT algorithm (55). The colored boxes indicate the different clusters of sequences used to create the capsid protein-based cluster 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) (59,60). The colored branches represent the different clusters observed in the matrix. Scale bar indicates substitutions per site. **C)** Rep diversity. (Left) Pairwise identity 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 (55). (Right) Unrooted phylogenetic tree obtained with FastTree from a PSI-Coffee alignment of the translated sequences of Rep trimmed with TrimAl v1.3 (57,58). The colored branches represent the different clusters that contain the Rep-based cluster sequence subset. Scale bar indicates substitutions per site. **D)** Pairwise identity frequency distribution. The frequency of pairwise identity values for each of the putative proteins or domains analyzed is shown. Constructed by Ignacio de la Higuera and George Kasun.

Cruciviruses capsid proteins share higher genetic identity than their Rep proteins: To assess the diversity in the proteins of cruciviruses, the percent pairwise identity between the protein sequences was calculated for capsid

protein and Rep using SDTv1.2 (**Fig. 2.3**). The average pairwise identity for the capsid protein was found to be 33.1% ± 4.9% (mean ± SD) (**Fig. 2.3A and 2.3D**), likely due to the high levels of conservation found in the S-domain (40.5% ± 8.4%) (**Fig. 2.3B and 2.3D**), while the average pairwise identity for Rep is quite low at 24.7% ± 5.6% (**Fig. 2.3C and 2.3D**). The differences in average pairwise identities between Rep, capsid protein, and S-domain are statistically significant (one-way analysis of variance [ANOVA]; $P < 0.0001$). The high variation of the Rep protein sequence relative to the capsid protein in cruciviruses correlates with a previous observation on a smaller data set (21).

To compare cruciviruses to other viral groups with homologous proteins, sequence similarity networks were built for the capsid protein and Rep (**Fig. 2.4**). For the capsid protein, 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 of $<10^{-20}$, sufficient to connect all cruciviruses and tombusviruses, with the exception of CruV-523 (**Fig. 2.4A**). However, using BLASTp, CruV-523 showed similarity to other RNA viruses with an E value of $<10^{-9}$, which were not included in the analysis. The capsid protein sequence similarity network analysis demonstrates the apparent homology of the capsid proteins in our data set with the capsid protein of RNA viruses: specifically, to unclassified RNA viruses that have RNA-dependent RNA polymerases (RdRPs) similar to those of either tombusviruses—also described as tombus-like viruses (79–81)—or nodaviruses. The latter RNA

viruses are proposed to belong to a chimeric group of viruses named tombunodaviruses (82).

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, 2.4B**). Due to the heterogeneity of Rep (**Fig. 2.3C**), the score cutoff for the network was relaxed to an E value of $<10^{-10}$; nonetheless, 10 divergent sequences lacked sufficient similarity to form connections within the network. While the Reps 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. 2.4B**). While still highly divergent from each other, the conserved motifs in Rep still share the most sequence similarity with CRESS-DNA viruses (**Fig. 2.2**).

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 (21). 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 (1,12–14). In contrast, here crucivirus sequences are

selected according to the presence of a tombusvirus-like capsid protein. Moreover, the Rep of cruciviruses could be subject to higher substitution rates than the capsid protein (30). It is possible that sequence divergence in capsid protein is more limited than in the Rep due to structural constraints.

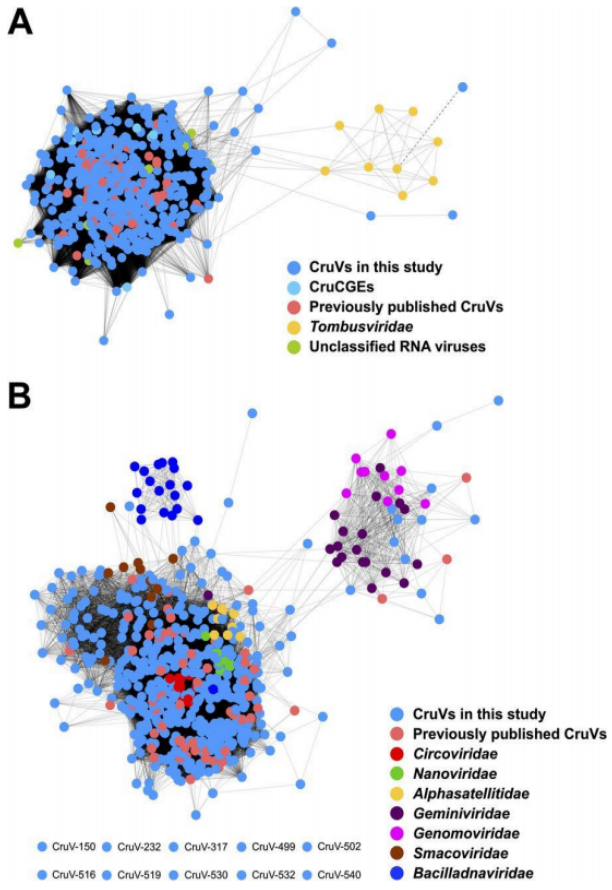

Figure 2.4. Sequence similarity networks of cruciviral proteins with related viruses. **A)** Capsid proteins represented by colored dots are connected with a solid line when the pairwise similarity, as assessed by the EFI-EST web server (63), has an E value of $<10^{-20}$. The dashed line represents an E value of 6×10^{-7} between the nodes corresponding to the capsid protein 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 pairwise similarity has an E value of $<10^{-10}$. 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 (63). Constructed by George Kasun.

Horizontal gene transfer among cruciviruses: To gain insight into the evolutionary history of cruciviruses, we carried out phylogenetic analyses of their capsid proteins and Reps. Due to the high sequence diversity in the data set, two smaller subsets of sequences were analyzed.

(i) Capsid protein-based clusters: Clusters with more than six nonidentical capsid protein sequences whose S-domains share a pairwise identity greater

than 70% were visually identified from **Fig. 2.3B**. This resulted in the identification of seven clusters, and one more divergent, yet clearly distinct, cluster was included (pink in **Fig. 2.3B**). A total of 47 genomes from the eight different clusters were selected for sequence comparison. The protein sequences of capsid and Rep were extracted and aligned, and their phylogenies were inferred and analyzed using tanglegrams (**Fig. 2.5**). The capsid protein phylogeny shows that the sequences from the eight capsid protein-based clusters form separate clades (**Fig. 2.5A**). On the other hand, the phylogeny of Rep shows a different pattern of relatedness between those genomes (**Fig. 2.5A**). This suggests different evolutionary histories for the capsid and Rep proteins, which could be due to recombination events between cruciviruses, as previously proposed with smaller data sets (21,25).

Rep-based clusters: To account for the possible bias introduced by selecting genomes from capsid protein cluster groups and to increase the resolution in the phylogeny of the Rep sequences, clusters of crucivirus genomes with more than six Rep sequences sharing pairwise identity of >45% and <98% were identified. The cutoff values were chosen to allow for the selection of six clusters containing a total of 53 genomes (**Fig. 2.3C**), whose capsid and Rep protein sequences were analyzed. The phylogeny of Reps shows distinct clades between the sequences from different Rep-based clusters (**Fig. 2.5B**). When the phylogeny of Rep was compared to that of their corresponding capsid proteins, we observed cruciviruses that group together in both Rep and capsid protein phylogenies. Discrepancies in topology between Rep and capsid protein trees were observed

as well, particularly in the capsid protein clade marked with an asterisk in **Fig. 2.5B**. This clade corresponds to the highly homogeneous red capsid protein-based cluster shown in **Fig. 2.3B** and suggests that gene transfer is more common in cruciviruses with a more similar capsid protein, 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.

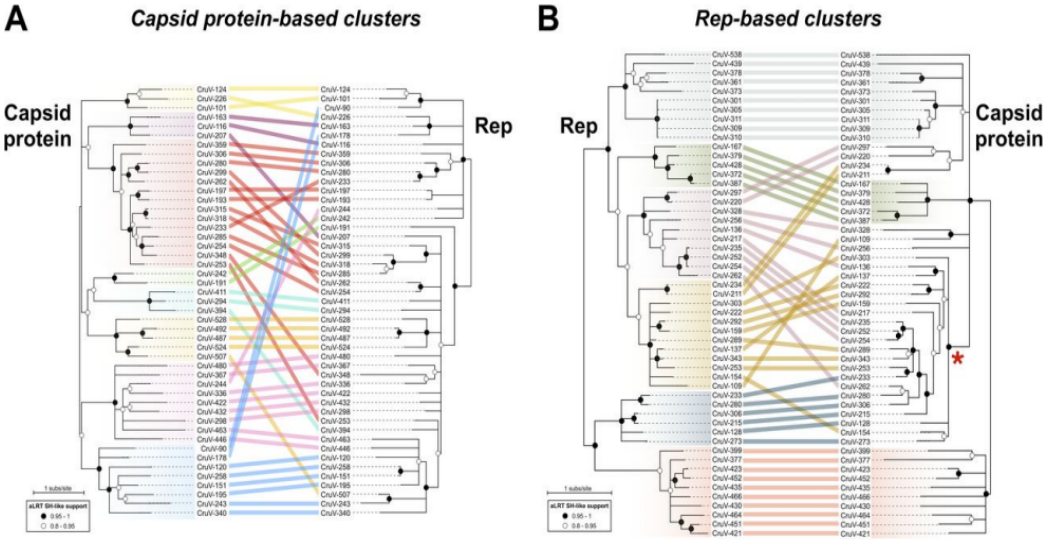

Figure 2.5. Comparison of phylogenies of capsid and Rep proteins of representative cruciviruses. (**A**) Tanglegram calculated with Dendroscope v3.5.10 from phylogenetic trees generated with PhyML from capsid protein (PhyML automatic model selection LG+G+I+F) and Rep (PhyML automatic model selection RtREV+G+I) alignments (60,62). The tips corresponding to the same viral genome are linked by lines that are color coded according to the clusters obtained from Fig. 3A (capsid protein-based clusters). (**B**) Tanglegram calculated with Dendroscope v3.5.10 from phylogenetic trees generated with PhyML from capsid protein (PhyML automatic model selection LG+G+I+F) and Rep (PhyML automatic model selection RtREV+G+I) alignments (62). 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 a red asterisk is formed by members of the red capsid protein-based cluster. Branch support is given according to aLRT SH-like (60). All nodes with an aLRT SH-like branch support inferior to 0.8 were collapsed with Dendroscope prior to constructing the tanglegram. Constructed by Ignacio de la Higuera and George Kasun.

To investigate possible exchanges of individual Rep domains among cruciviruses, the Rep alignments of the analyses of the capsid protein-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 capsid protein-based clusters, we observed incongruence in the phylogenies between endonuclease and helicase domains (**Fig. 2.6A**), suggesting recombination within crucivirus Reps, as has been previously hypothesized with a much smaller data set (25). This incongruency is not observed in the analyzed Rep-based clusters (**Fig. 2.6B**). 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 subjected. The detection of capsid protein/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 stramenopiles/alveolates/Rhizaria (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 an S3H helicase domain is observed in other CRESS-DNA viruses which are known to infect eukaryotes (15). This is distinct from Reps found in prokaryote-infecting CRESS-DNA viruses—which lack a fused S3H helicase domain (83)—and other related HUH endonucleases

involved in plasmid rolling-circle replication and HUH transposases (15). Additionally, the capsid protein of cruciviruses, a suggested determinant of tropism (84,85), 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 (79).

Cruciviruses have been found as contaminants of spin columns made of diatomaceous silica (25), in aquatic metagenomes enriched with unicellular algae (21), in the metagenome of *Astrammina rara*—a foraminiferan protist part of the Rhizaria (21)—and associated with epibionts of isopods, mainly comprised of apicomplexans and ciliates, both belonging to the alveolates (30). These pieces of evidence point toward the stramenopiles/alveolates/Rhizaria (SAR) supergroup as a candidate taxon to contain potential crucivirus hosts (86). No host prediction can be articulated from our sequence data. However, at least five of the crucivirus genomes render complete translated capsid protein and Rep sequences only 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 a glutamine (87). The usage of an alternative genetic code seems evident in CruV-502—found in the metagenome from seawater collected above diseased coral colonies (88) 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.

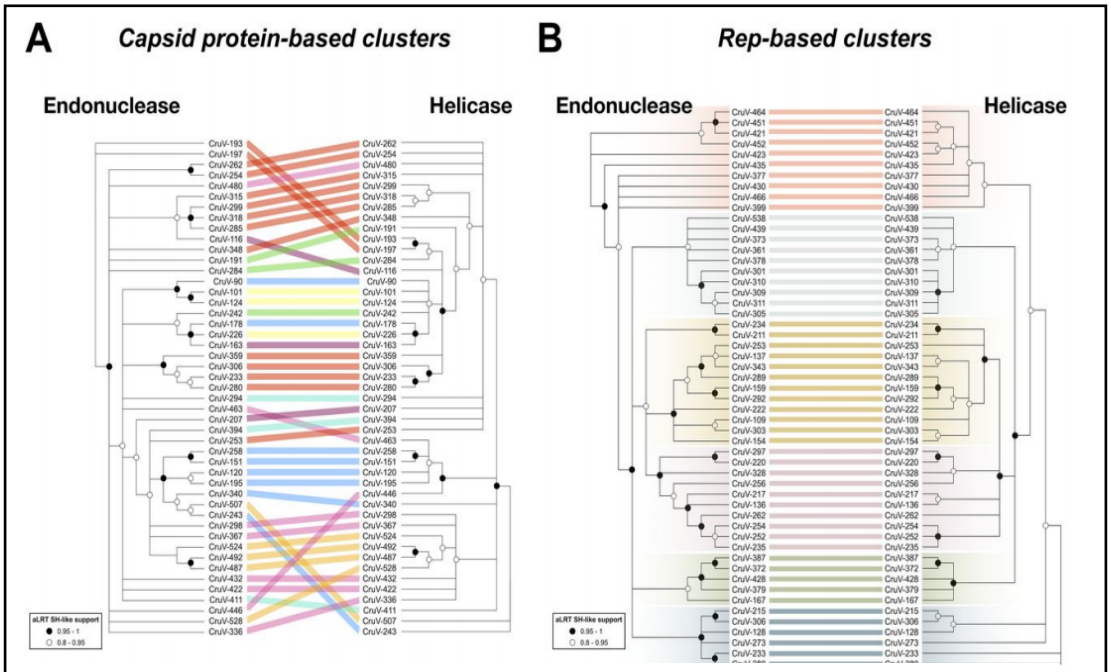

Figure 2.6. Comparison of phylogenies between the endonuclease and helicase domains of Reps from representative cruciviruses. **A)** Tanglegram calculated with Dendroscope v3.5.10 from phylogenetic trees generated with PhyML from separate alignments of Rep endonuclease and helicase domains (60,62). The tips corresponding to the same viral genome are linked by lines that are color coded according to the clusters obtained from Fig. 2.3A (capsid protein-based clusters). **B)** Same as panel A but with sequences from the clusters obtained from Fig. 2.3B (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 (62). Constructed by Ignacio de la Higuera and George Kasun.

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 underpin the placement of cruciviruses within the CRESS-DNA viruses.

The classification of CRESS-DNA viruses is primarily based upon the phylogeny of the Rep proteins, although commonalities in capsid protein and genome organization are also considered (12,13). This taxonomic criterion is challenging in cruciviruses, whose Rep proteins are highly diverse. Whether the use of proteins involved in replication for virus classification should be preferred over structural proteins has been previously questioned (89).

The capsid of cruciviruses, as well as the capsid of other CRESS-DNA virus families like circoviruses, geminiviruses, and bacilladnaviruses, possesses the single jelly-roll architecture (44). However, there is no obvious sequence similarity between the capsid protein of cruciviruses and that of classified CRESS-DNA viruses. The crucivirus capsid protein—homologous to the capsid of tombusviruses—is an orthologous trait within the CRESS-DNA viruses. Hence, the capsid protein constitutes a synapomorphic character that demarcates this group of viruses from the rest of the CRESS-DNA viral families.

CRESS-DNA viruses appear to have multiple origins from plasmids. Their Rep proteins appear to have arisen from these plasmids, and the viruses have diverged into different ssDNA virus groups on acquisition of nonorthologous capsid proteins from RNA viruses (10,90). Cruciviruses, however, are classified as such due to shared capsid protein genes but encode Rep proteins that span many different viral clusters within the phylum *Cressdnaviricota*, as we have shown. Thus, it seems unlikely that cruciviruses will form a formal taxon, as they appear to be a collection of viruses from multiple *Cressdnaviricota* groups. However, like Baltimore classes, the label crucivirus does aid in understanding

virus evolution, particularly the transfer of capsid protein genes, which appears to have been prevalent not only in ssDNA viruses but throughout the virosphere.

Concluding Remarks: Cruciviruses are a growing group of CRESS-DNA viruses that encode a putative capsid protein 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 putative capsid proteins are relatively homogenous, the putative Reps are relatively diverse among the cruciviruses, spanning the diversity of all classified CRESS-DNA viruses. Cruciviruses seem to have recombined with each other to exchange functional modules between themselves, 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.

References

1. Simmonds P, Adams MJ, Benk M, Breitbart M, Brister JR, Carstens EB, et al. Consensus statement: Virus taxonomy in the age of metagenomics. Nat Rev Microbiol. 2017;15(3):161–8.
2. Suttle C. Viruses in the Sea. Nature. 2005;437(7057):356-61
3. Berliner AJ, Mochizuki T, Stedman KM. Astrovirology: viruses at large in the universe. Astrobiology. 2018;18(2):207–23.
4. Koonin E V., Krupovic M. The depths of virus exaptation. Curr Opin Virol. 2018;31:1–8.
5. Koonin E V, Dolja V V. A virocentric perspective on the evolution of life. 2020; 3(5):546-57.
6. Domingo E, Sheldon J, Perales C. Viral quasispecies evolution. Microbiol Mol Biol Rev. 2012;76(2):159–216.
7. Baltimore D. Expression of animal virus genomes. Bacteriol Rev. 1971;35(3):235–41.
8. Koonin E V., Senkevich TG, Dolja V V. The ancient virus world and evolution of cells. Biol Direct. 2006;1:1–29.
9. Krupovic M, Ravantti JJ, Bamford DH. Geminiviruses: A tale of a plasmid becoming a virus. BMC Evol Biol. 2009;9(1):1–11.
10. Kazlauskas D, Varsani A, Koonin E V., Krupovic M. Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. Nat Commun. 2019;10(1):1–12
11. Krupovic M. Recombination between RNA viruses and plasmids might have played a central role in the origin and evolution of small DNA viruses. BioEssays. 2012;34(10):867–70.
12. Krupovic M, Varsani A, Kazlauskas D, Breitbart M, Delwart E, Yutin N, et al. Cressdnaviricota: a virus phylum unifying 7 families of Rep-encoding viruses with single- stranded, circular DNA genomes 3. 2020;1;94(12):e00582-20.
13. Rosario K, Duffy S, Breitbart M. A field guide to eukaryotic circular single-stranded DNA viruses: Insights gained from metagenomics. Arch Virol. 2012;157(10):1851–71.
14. Zhao L, Rosario K, Breitbart M, Duffy S. Eukaryotic Circular Rep-Encoding Single-Stranded DNA (CRESS-DNA) viruses: Ubiquitous viruses with small genomes and a diverse host range. 1st ed. Advances in Virus Research. Elsevier Inc.; 2018. 1–63.

15. Chandler M, Cruz F De, Dyda F, Hickman AB. Breaking and joining single-stranded DNA : the HUH endonuclease superfamily. Nat Rev Microbiol. 2013;11(8):525–38.

16. Diemer GS, Stedman KM. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. Biol Direct. 2012;7(13):1–14.

17. Cheung AK. Porcine circovirus: Transcription and DNA replication. Virus Res. 2012;164(1–2):46–53.

18. Laufs J, Jupin I, David C, Schumacher S, Heyraud-Nitschke F, Gronenborn B. Geminivirus replication: Genetic and biochemical characterization of Rep protein function, a review. Biochimie. 1995;77(10):765–73.

19. Stedman K. Mechanisms for RNA capture by ssDNA viruses: Grand theft RNA. J Mol Evol. 2013;76(6):359–64.

20. Rosario K, Dayaram A, Marinov M, Ware J, Kraberger S, Stainton D, et al. Diverse circular ssDNA viruses discovered in dragonflies (*Odonata: Epiprocta*). J Gen Virol. 2012; ;93(Pt 12):2668-2681.

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

22. de la Higuera I, Torrance EL, Pratt AP, Kasun GW, Maluenda A, Stedman KM. Genome sequences of three cruciviruses found in the Willamette Valley (Oregon). 2019;(June):18–20.

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

24. Tisza MJ, Pastrana D V., Welch NL, Stewart B, Peretti A, Starrett GJ, et al. Discovery of several thousand highly diverse circular DNA viruses. eLife. 2020;9:1–26.

25. Krupovic M, Zhi N, Li J, Hu G, Koonin E V., Wong S, et al. Multiple layers of chimerism in a single-stranded DNA virus discovered by deep sequencing. Genome Biol Evol. 2015;7(4):993–1001.

26. Hewson I, Ng G, Li WF, LaBarre BA, Aguirre I, Barbosa JG, et al. Metagenomic identification, seasonal dynamics, and potential transmission mechanisms of a Daphnia-associated single-stranded DNA virus in two temperate lakes. Limnol Oceanogr. 2013;58(5):1605–20.

27. Steel O, Kraberger S, Sikorski A, Young LM, Catchpole RJ, Stevens AJ, et al. Circular replication-associated protein encoding DNA viruses identified

in the faecal matter of various animals in New Zealand. Infect Genet Evol. 2016;43:151–64.

28. Mcdaniel LD, Rosario K, Breitbart M, Paul JH. Comparative metagenomics: Natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. Environ Microbiol. 2014;16(2):570-85.

29. Dayaram A, Galatowitsch ML, Argüello-Astorga GR, van Bysterveldt K, Kraberger S, Stainton D, et al. Diverse circular replication-associated protein encoding viruses circulating in invertebrates within a lake ecosystem. Infect Genet Evol. 2016;39:304–16

30. Bistolas K, Besemer R, Rudstam L, Hewson I. Distribution and inferred evolutionary characteristics of a chimeric ssDNA virus associated with intertidal marine isopods. Viruses. 2017;26(12):361.

31. Quaiser A, Krupovic M, Dufresne A, Roux S. Diversity and comparative genomics of chimeric viruses in Sphagnum- dominated peatlands. Virus Evol. 2016;2(2):1–8.

32. Salmier A, Tirera S, De Thoisy B, Franc A, Darcissac E, Donato D, et al. Virome analysis of two sympatric bat species (*Desmodus rotundus* and *Molossus molossus*) in French Guiana. PLoS One. 2017;12(11):1–25.

33. Brown DR, Schmidt-Glenewinkel T, Reinberg D, Hurwitz J. DNA sequences which support activities of the bacteriophage phi X174 gene A protein. J Biol Chem. 1983;258(13):8402–12.

34. Steinfeldt T, Finsterbusch T, Mankertz A. 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. 2006;80(13):6225–34.

35. Roth MJ, Brown DR, Hurwitz J. Analysis of bacteriophage phiX174 gene A protein-mediated termination and reinitiation of phiX DNA synthesis. II. Structural characterization of the covalent phiX A protein-DNA complex. J Biol Chem. 1984;259(16):10556–68.

36. Cheung AK. 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. 2007 Jun 20;363(1):229–35.

37. Stenger DC, Revington GN, Stevenson MC, Bisaro DM. 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. 1991 15;88(18):8029-33.

38. Ilyina T V, Koonin E V. Conserved sequence motifs in the initiator proteins for rolling circle DNA-replication encoded by diverse replicons from Eubacteria, Eukaryotes and Archaebacteria. Nucleic Acids Res. 1992;20(13):3279–85.

39. Londoño A, Riego-Ruiz L, Argüello-Astorga GR. DNA-binding specificity determinants of replication proteins encoded by eukaryotic ssDNA viruses are adjacent to widely separated RCR conserved motifs. Arch Virol. 2010;155(7):1033–46.

40. Clérot D, Bernardi F. DNA helicase activity is associated with the replication initiator protein Rep of tomato yellow leaf curl geminivirus. J Virol. 2006;80(22):11322–30.

41. Hickman AB, Dyda F. Binding and unwinding: SF3 viral helicases. Current Opinion in Structural Biology. 2005;15(1):77-85.

42. Kazlauskas D, Varsani A, Krupovic M. Pervasive chimerism in the replication-associated proteins of uncultured single-stranded DNA viruses. Viruses. 2018;10(4):1–11.

43. Zhang W, Olson NH, Baker TS, Faulkner L, Agbandje-McKenna M, Boulton MI, et al. Structure of the maize streak virus geminate particle. Virology. 2001;279(2):471–7.

44. Krupovic M, Koonin E V. Multiple origins of viral capsid proteins from cellular ancestors. Proc Natl Acad Sci U S A. 2017;114(12):E2401-E2410.

45. Structures NCM, Clover R,. Near atomic resolution cryo-electron microscopy structures ofof cucumber leaf spot virus and red clover necrotic mosaic virus: Evolutionary divergence at the icosahedral three-fold axes. Virology. 2020;(August):1–13.

46. Park SH, Sit TL, Kim KH, Lommel SA. 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. 2013;176(1–2):107–18.

47. Alam SB, Reade R, Theilmann J, Rochon DA. 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. 2017;512(August):83–94.

48. Ohki T, Akita F, Mochizuki T, Kanda A, Sasaya T, Tsuda S. 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. 2010;402(1):129–34.

49. Llauró A, Coppari E, Imperatori F, Bizzarri AR, Castón JR, Santi L, et al. Calcium ions modulate the mechanics of tomato bushy stunt virus. Biophys J. 2015;109(2):390–7.

50. Paez-Espino D, Chen IMA, Palaniappan K, Ratner A, Chu K, Szeto E, et al. IMG/VR: A database of cultured and uncultured DNA viruses and retroviruses. Nucleic Acids Res. 2017;45(D1):D457–65.

51. Wright, E A, T Heckel, J Groenendijk, J W Davies MIB. Splicing features in maize streak virus virion- and complementary-sense gene expression.

52. Higuera I de la, Kasun GW, Torrance EL, Pratt AA, Maluenda A, Colombet J, et al. Unveiling crucivirus diversity by mining metagenomic data. mBio. 2020 Sep 1;11(5):e01410-20.

53. Hafner GJ, Stafford MR, Wolter LC, Harding RM, Dale JL. Nicking and joining activity of banana bunchy top virus replication protein in vitro. J Gen Virol. 1997; 78 (Pt 7):1795-9.

54. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6(1):1–14.

55. Muhire BM, Varsani A, Martin DP. SDT: A virus classification tool based on pairwise sequence alignment and identity calculation. PLoS One. 2014;9(9).

56. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.

57. Floden EW, Tommaso PD, Chatzou M, Magis C, Notredame C, Chang JM. 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. 2016;44(W1):W339–43.

58. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–3.

59. Price MN, Dehal PS, Arkin AP. Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009;26(7):1641–50.

60. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst Biol. 2010;59(3):307–21.

61. Anisimova M, Gascuel O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst Biol. 2006; 55(4):539-52

62. Huson DH, Scornavacca C. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. Syst Biol. 2012;61(6):1061–1067.

63. Zallot R, Oberg N, Gerlt JA. From the bench the efi web resource for genomic enzymology web tools : leveraging protein , genome , and metagenome databases to discover novel enzymes and metabolic pathways. Biochemistry. 2019; 15;58(41):4169-4182.

64. Kazlauskas D, Dayaram A, Kraberger S, Goldstien S, Varsani A, Krupovic M. Evolutionary history of ssDNA bacilladnaviruses features horizontal acquisition of the capsid gene from ssRNA nodaviruses. Virology. 2017;504(January):114–21.

65. Rosario K, Mettel KA, Benner BE, Johnson R, Scott C, Yusseff-Vanegas SZ, et al. Virus discovery in all three major lineages of terrestrial arthropods highlights the diversity of single-stranded DNA viruses associated with invertebrates. PeerJ. 2018;2018(10):1–36.

66. Varsani A, Krupovic M. Sequence-based taxonomic framework for the classification of uncultured single-stranded DNA viruses of the family Genomoviridae. Virus Evol. 2017;3(1):1-14.

67. Yadava P, Suyal G, Mukherjee SK. Begomovirus DNA replication and pathogenicity. Curr Sci. 2010;98(3):360–8.

68. Gronenborn B. Nanoviruses: Genome organisation and protein function. Vet Microbiol. 2004;98(2):103–9.

69. Varsani A, Lefeuvre P, Roumagnac P, Martin D. Notes on recombination and reassortment in multipartite/segmented viruses. Curr Opin Virol. 2018;33(December):156–66.

70. Vega-Rocha S, Byeon IJL, Gronenborn B, Gronenborn AM, Campos-Olivas R. Solution structure, divalent metal and DNA binding of the endonuclease domain from the replication initiation protein from porcine circovirus 2. J Mol Biol. 2007;367(2):473–87.

71. Moscoso M, Eritja R, Espinosa M. Initiation of replication of plasmid pMV158: Mechanisms of DNA strand-transfer reactions mediated by the initiator RepB protein. J Mol Biol. 1997;268(5):840–56.

72. Gonzalez-Perez B, Lucas M, Cooke LA, Vyle JS, De La Cruz F, Moncalián G. Analysis of DNA processing reactions in bacterial conjugation by using suicide oligonucleotides. EMBO J. 2007;26(16):3847–57.

73. Varsani A, Krupovic M. *Smacoviridae*: a new family of animal-associated single-stranded DNA viruses. Arch Virol. 2018;163(7):2005–15.

74. Chabi-Jesus C, Najar A, Fontenele RS, Kumari SG, Ramos-González PL, Freitas-Astúa J, et al. Viruses representing two new genomovirus species identified in citrus from Tunisia. Arch Virol. 2020;165(5):1225–9.

75. Huang SW, Liu HP, Chen JK, Shien YW, Wong ML, Wang CY. Dual ATPase and GTPase activity of the replication-associated protein (Rep) of beak and feather disease virus. Virus Res. 2016;213:149–61.

76. Dennis TPW, Flynn PJ, Marciel de Souza W, Singer JB, Moreau CS, Wilson SJ, et al. Insights into circovirus host range from the genomic fossil record. J Virol. 2018;92(June):JVI.00145-18.

77. Pasumarthy KK, Choudhury NR, Mukherjee SK. 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. 2010; 2010;14;7:128.

78. Nash TE, Dallas MB, Reyes MI, Buhrman GK, Ascencio-Ibanez JT, Hanley-Bowdoin L. Functional Analysis of a Novel Motif Conserved across Geminivirus Rep Proteins. J Virol. 2011;85(3):1182–92.

79. Grasse W, Spring O. ssRNA viruses from biotrophic Oomycetes form a new phylogenetic group between Nodaviridae and Tombusviridae. Arch Virol. 2017;162(5):1319–24.

80. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, et al. Redefining the invertebrate RNA virosphere. Nature. 2016;540(7634):539–43.

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

82. Greninger AL, DeRisi JL. Draft genome sequence of tombunodavirus UC1. Genome Announc. 2015;3(4):e00655-15

83. Krupovic M. Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. Curr Opin Virol. 2013;3(5):578–86.

84. Allison AB, Organtini LJ, Zhang S, Hafenstein SL, Holmes EC, Parrish CR. Single mutations in the VP2 300 loop region of the three-fold spike of the carnivore parvovirus capsid can determine host range. J Virol. 2016;90(2):753–67.

85. Carbonell A, Maliogka VI, De Jesús Pérez J, Salvador B, San León D, García JA, et al. 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. 2013;26(10):1211–24.

86. Beakes GW, Glockling SL, Sekimoto S. The evolutionary phylogeny of the oomycete “fungi.” Protoplasma. 2012;249(1):3–19.

87. Hanyu N, Kuchino Y, Nishimura S, Beier H. 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. 1986;5(6):1307–11.

88. Soffer N, Brandt ME, Correa AMS, Smith TB, Thurber RV. Potential role of viruses in white plague coral disease. ISME J. 2014;8(2):271–83

89. Krupovic M, Bamford DH. Order to the viral Universe. *J Virol*. 2010;84(24):12476- 12479.

90. Koonin EV, Dolja V, Krupovic M, Varsani A, Wolf YI, Yutin N, Zerbini FM, Kuhn JH. Global organization and proposed megataxonomy of the virus world. Microbiol Mol Biol Rev. 2020;84(2):1–33.

Chapter Three

Analysis of Crucivirus Evolution and Origin of Replication Associated DNA Sequences

Abstract

Phylogenies of CRESS-DNA viruses are based upon the relationships of their replication associated protein (Rep) which has led to the recent establishment of the phylum *Cressdnaviricota*. As CRESS-DNA viruses, cruciviruses seemingly should be able to be incorporated into these phylogenies. Here it is shown that despite being ssDNA viruses and encoding a putative Rep, the phylogenetic placement of cruciviruses among other CRESS-DNA viruses presents significant challenges. Despite employing multiple phylogenetic techniques including new genomes, cruciviruses do not fit CRESS-DNA phylogenies. This leads to unresolved questions about the origins and subsequent evolution of this new group of viruses. To address this, an attempt was made to more accurately define relationships that cruciviruses display between one another. Specifically, it was attempted to locate specific amino acids in Rep that may play a role in dsDNA binding specificity during the initiation of rolling circle replication. These preliminary results suggest that Reps of cruciviruses do not conform to patterns previously observed for other CRESS-DNA viruses with respect to the location of dsDNA binding specificity determinants.

Introduction

Over the past decade the ubiquity of circular replication-associated protein encoding single-stranded DNA viruses (CRESS-DNA virus) has been revealed through metagenomic studies. (1,2). Previously believed to be restricted to plant and animal hosts (3), these metagenomic studies have unearthed a growing number of environments that harbor CRESS-DNA viruses as well as organisms that serve as potential hosts for CRESS-DNA viruses (1,2). While it was previously known that CRESS-DNA viruses are economically important pathogens of agricultural crops (4,5), metagenomics have shown CRESS-DNA viruses in association with animals as diverse as pigs (6) and dragonflies (7,8), and in environments ranging from antarctic lakes (5) to sewage oxidation ponds (10). CRESS-DNA viruses encapsidate their genomes in some of the smallest known virions and often contain as few as two genes: one encoding for the capsid protein (CP) and the other encoding for the replication associated protein (Rep) (2). Rep is conserved across all CRESS-DNA viruses (2) and is often the only protein that displays a high degree of conservation (1). While CP of CRESS-DNA viruses display structural similarity, they often display sequence divergence among members within the same family (2). As such, Rep is used in the phylogenetic classification of both previously known and newly discovered CRESS-DNA viruses (11).

Cruciviruses are newly discovered CRESS-DNA viruses whose genomes contain at least two open reading frames: one encoding a putative Rep homologous to those of other CRESS-DNA viruses, and the second encoding a CP homologous to those found in ssRNA viruses. Due to the apparent ssDNA nature of their genome and the presence of an ORF encoding a putative Rep it follows that cruciviruses could be placed within the recently established phylum *Cressdnaviricota* (11). In this chapter attempts to place cruciviruses within *Cressdnaviricota* are performed by expanding the number of representative crucivirus genomes available for analysis. Cruciviruses continue to be dispersed across the *Cressdnaviricota* making their evolutionary history and relationships to other CRESS-DNA viruses unclear.

To address the ambiguity left by our previous approaches (this chapter and chapter 2), a more targeted approach to better understanding cruciviruses was developed. Previous work has shown that many CRESS-DNA viruses contain repeated DNA sequences near their origin of replication (ori) stem-loop structures which are distinct between viruses of different species (8,12–15). These repeated DNA sequences, known as iterons, have been shown to play an important role in determining specificity of interactions between Rep and a given ori (13,16).

During the initiation of rolling circle replication (RCR) Rep binds to dsDNA near a stem-loop structure in a manner dependent on iterons of a specific sequence (17–19). Following binding, Rep introduces a single-stranded nick

within a conserved nonanucleotide sequence located in the loop portion of the stem-loop exposing a 3’ OH from which cellular polymerases can replicate ssDNA viral genomes (20–22). Previous work in bipartite begomoviruses has demonstrated that there are constraints on reassortment (pseudorecombination) events which are seemingly dictated by the compatibility of *cis*-acting iterons and *trans*-acting factors of Rep (16). The ability to form viable viral progeny by pseudorecombination is generally limited to genome components of viruses whose Reps contain conserved residues in typically non-conserved regions, known as specificity determinants (SPDs), and whose iterons are similar, suggesting that the ability to complete RCR is dependent on this compatibility (16,23–25). While the presence of similar iterons and SPDs appears to be necessary for viable pseudorecombination, it also appears likely that compatibility of the movement protein (used for intra-host spread) encoded by one virus with the genome of another virus plays a role in the formation of pseudorecombinant begomoviruses (26). It has also been observed that the replication of betasatellites of begomoviruses can be carried out in a promiscuous fashion in which many different Reps are capable of replication of these virally associated DNAs, but that greatest replication occurs when iterons are similar between the betasatellite and a given helper virus (13,27).

Previous in silico work has predicted the presence of SPDs in Reps of various CRESS-DNA viruses (12,14,16). These SPDs consist of conserved amino acids located within otherwise variable regions of Rep adjacent to the

widely conserved motif I and motif II and have been predicted to be involved in iteron/ori discrimination (12,18). Others have previously used these apparent relationships between Rep and iterons to classify closely related CRESS-DNA viruses discovered through metagenomics (8). To explore this potential relationship of iterons and SPDs in cruciviruses, a prediction and annotation script was developed to locate potential iterons associated with stem-loops. This allowed for iterated DNA sequences to be uncovered in cruciviruses and for relationships between iterons and potential SPDs to be explored.

Methods

New Crucivirus Genome Annotation: Following the publication of de la Higuera et al. 2020 (28) (Chapter two) our collaborators provided an additional 425 circular sequences identified in metagenomic studies that appear to be cruciviruses. From 331 genomes were annotated in a fashion similar to what was described in Chapter two and de la Higuera et al., 2020 (28). This second set of crucivirus genomes and associated additional analyses are in preparation for publication.

Rep Alignments: Alignments of crucivirus Reps were generated using MAFFT (29) and an automatic model selection with a BLOSUM 80 scoring matrix in Geneious Prime 2020.0.5 (https://www.geneious.com/) and were subsequently manually curated. Alignment views were generated and edited in Jalview v1.0 (30). Reps of other CRESS-DNA viruses and Reps encoded by plasmids described by Kazlauskas et al., 2020 were retrieved from NCBI by accession

number (31). The same alignment procedure was followed to produce alignments of crucivirus and CRESS-DNA virus Reps.

Phylogenetic Trees: Alignments containing a total of 1,178 Rep sequences (394 CRESS-DNA virus and 784 crucivirus) were trimmed using in TrimAl v1.3 using a strict plus setting (32). Phylogenetic trees were generated in the PhyML 3.0 webserver (http://www.atgc-montpellier.fr/phyml/) (33). Automatic model selection (RtREV +G+F) and aLRT SH-like branch support were used (33). Trees were annotated using the interactive tree of life webserver (https://itol.embl.de/) (34).

Sequence Similarity Network: A total of 1,503 Rep sequences (325 plasmid, 394 CRESS-DNA virus, and 784 crucivirus) were uploaded to the EFI-EST webserver and e-value of $<10^{-10}$ was selected (35). Resulting output files were annotated in Cytoscape 3.8.1 (36).

Stem-Loop Prediction: The presence of stem-loops and associated nonanucleotides in crucivirus genomes was predicted using StemLoop-Finder (Pratt, Torrance, Kasun, Stedman and de la Higuera, 2021) and confirmed with mfold (37).

Crucivirus Iteron Search Tool: In an effort to identify iterons present in ori regions of cruciviruses, Crucivirus iteron search tool (CRUISE, in preparation) was developed. This program searches for and annotates iterons in the region surrounding stem loops detected with StemLoop-Finder. Searches are constrained to the fifty nucleotides on either side of the detected nonanucleotide

sequence. Potential iterons are annotated as such if they consist of: repeated nucleotides at least four bases in length, separated by a number of bases equal to or less than the length of the repeat itself, and contain two or more unique bases. These parameters do not encompass previously predicted iteron diversity (8), but rather represent iteron arrangements for which biochemical experiments have been conducted (13,17). CRUISE has been tested on a set of 37 CRESS-DNA virus genomes in which iterated DNA sequences have been previously found manually and correctly annotates those repeats as iterons, indicating that it is an effective tool for finding iterated DNA sequences within the constraints mentioned above. The program is also capable of annotating iterons from a customizable database of CRESS-DNA virus genomes in which iterons have been previously identified (**Table 1**).

Results and Discussion

Cruciviruses Blur the Lines of Established CRESS-DNA Phylogenies: In attempting to place cruciviruses within the established phylum *Cressdnaviricota*, it becomes apparent that crucivirus Reps both span the diversity of the phylum and disrupt some previously well supported clades (11) (**Fig. 3.1 and Fig. 1.1**). Cruciviruses are spread throughout the tree in multiple often poorly supported branches (bootstrap <0.5). The use of different phylogenetic tree models were unsuccessful in producing better supported branches. As more crucivirus genome sequences become available construction of phylogenetic trees may offer deeper insights to these relationships. Cruciviruses appear to occupy a

unique space in CRESS-DNA classification schemes in which their putative CP often display a higher degree of conservation than does their putative Rep (2,28).

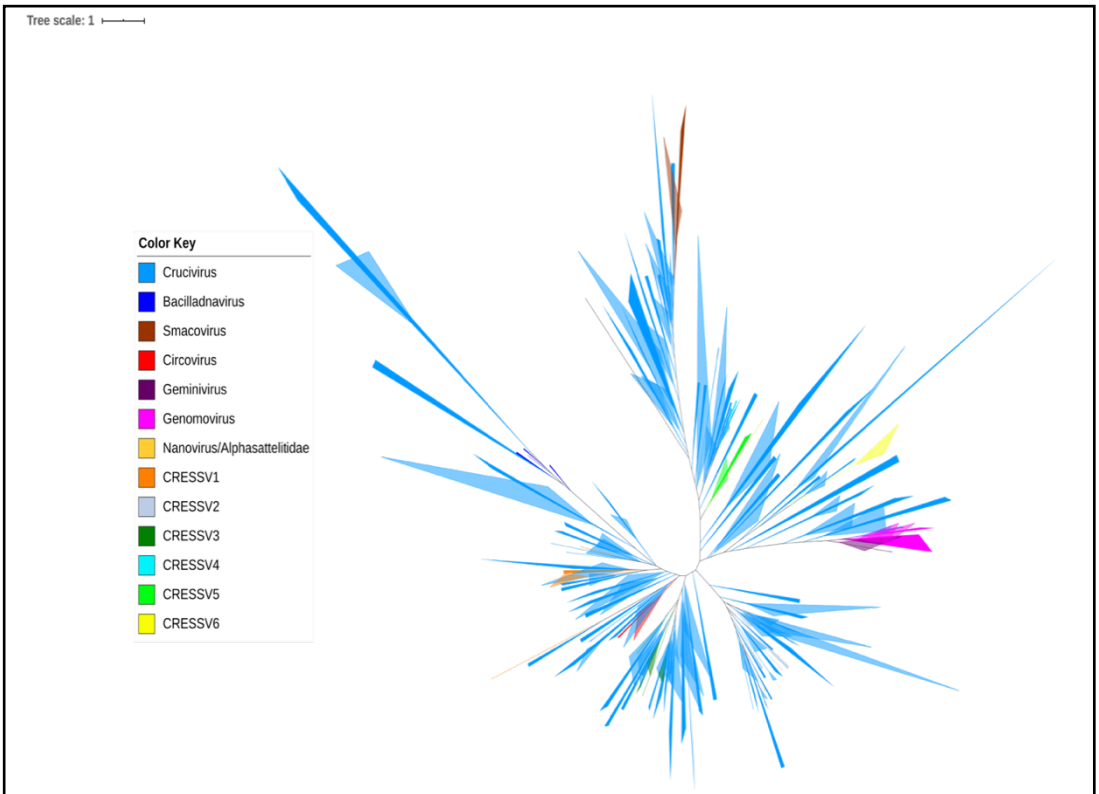

Figure 3.1. Unrooted maximum likelihood tree of crucivirus and other members of the *Cressdnaviricota* based on Rep. Tree was calculated using aLRT SH-like support and automatic model selection. The tree was colored manually by CRESS-DNA virus type (family level or unclassified) using the itol webserver (https://itol.embl.de/). Cruciviruses (light blue) are widespread across the different families of both classified and unclassified members of the *Cressdnaviricota*. Nodes are collapsed to contain only one CRESS-DNA virus group (unclassified or family level classification) or cruciviruses.

One approach that may resolve some of these ambiguities involves closer examination of crucivirus genomes. CRESS-DNA viruses discovered through metagenomics have been diverse genome arrangements. CRESS-DNA virus genomes have been shown to contain both ambisense and unisense arrangements of putative genes encoding the capsid protein (CP) and Rep,

stem-loops containing nonanucleotides of varying sequence have been identified in various orientations with respect to CP and *Rep* genes, and genome size can vary greatly (1,2,7,15,38). Others have previously grouped these genomes on the basis of shared genome arrangements (1). Based on annotations of a large number of genomes it should be possible to form groups that have shared characteristics. For example, genomes which display an ambisense orientation of putative CP and Rep ORFs would be placed in one group (or type) and genomes displaying a unisense orientation into a second type (1). These groups could be further resolved by examining the orientation of the nonanucleotide relative to the aforementioned ORFs. For example, ambisense genomes containing a nonanucleotide sequence on the Rep-encoding strand would be separated from those ambisense genomes which contain the nonanucleotide sequence on the CP encoding strand (1). This general approach of limiting the number of genomes analyzed based on similarities did help to resolve some of the issues encountered in the work of de la Higuera et al., 2020 (28) and is being further examined.

Previous work has indicated the CRESS-DNA viruses likely emerged from plasmids on more than one occasion (31,39,40). Previous analyses have also indicated that the lack of a superfamily-3-helicase (SF3) domain in Reps encoded by plasmids is likely an ancestral state, suggesting that Reps of pCRESS4-8 (Rep encoding plasmids) emerged from plasmids similar to pE198 and pMV158 (**Fig. 3.2**) which lack the SF3 domain following the acquisition and

fusion of a SF3 helicase domain to an HUH-endonuclease domain (31). The acquisition of an RNA virus CP containing a jelly roll fold by a member of pCRESS4-8 then gave rise to geminiviruses and genomoviruses (31). A similar event between plasmids pCRESS1-3 and an RNA virus likely led to the emergence of members of the *Circoviridae*, *Smacoviridae* and *Nanoviridae (31)*. In order to explore the relationships cruciviruses display between Rep encoding plasmids and other CRESS-DNA viruses we constructed a sequence similarity network (**Fig. 3.2**).

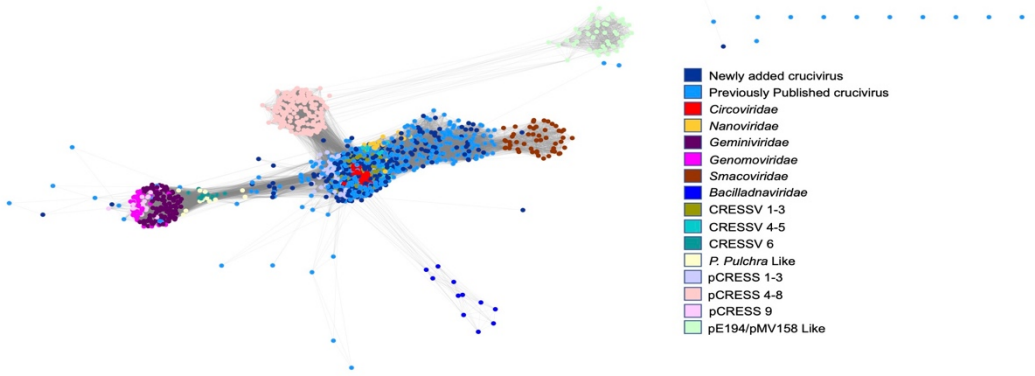

Figure 3.2. Expanded sequence similarity network of Rep. In addition to the original crucivirus (light blue) sequences this network includes 331 additional new crucivirus Reps (dark blue), 290 classified CRESS-DNA virus Reps, 102 unclassified CRESS-DNA virus Reps, and 327 plasmid encoded Reps. Colored nodes representing Rep sequences are connected with a solid line when the pairwise similarity has an E value of $<10^{-10}$. Cruciviruses appear to be persistent in their widespread distribution in the network.

The placement of cruciviruses within this network (**Fig. 3.2)** is highly dispersed in contrast to both Rep encoded by plasmids (pCRESS) and CRESS-DNA viruses (family name or CRESSV), further supporting the observations that Reps of cruciviruses are more diverse than those of the established CRESS-

DNA virus families and other unclassified CRESS-DNA viruses alike. That this pattern is still apparent after introducing additional Reps of cruciviruses further confirms previous observations related to the broad sequence space inhabited by Rep of cruciviruses (Chapter 2). Additionally, there are a number of orphan Rep sequences that do not form connections (e-value $<10^{-10}$) all of which are crucivirus Reps.

The sequence similarity network (**Fig. 3.2**) may offer potential insights to the origins of cruciviruses. While connections are formed between Reps encoded by plasmids and other CRESS-DNA viruses, these connections are more common between cruciviruses and a broad range of plasmid encoded Reps. Given that other CRESS-DNA virus families have been predicted to have descended from specific Rep encoding plasmids (31,39), the diversity of crucivirus-plasmid connections may imply that cruciviruses emerged from plasmids on more than one occasion. Of course, these connections may simply represent the shared evolutionary history of plasmid and viral Reps, or the transfer of Rep from a crucivirus to a plasmid, previously predicted in the evolution of CRESS-DNA viruses (31).

The origin or origins of cruciviruses remains an open and intriguing question. The most parsimonious explanation for their origin(s) appears to be the acquisition of a capsid protein gene by a DNA based genetic element from an RNA virus. It has been previously hypothesized that this event occurred between RNA and DNA viruses (41,42). But the work presented here and in chapter two is

unable to definitively rule out the possibility that this event took place between a plasmid and an RNA virus as seems likely for other CRESS-DNA viruses (31). The heterogeneity that we have described for crucivirus Reps may be explained by multiple initial emergences of cruciviruses, which could be further resolved by exploring more distant relationships between CP of cruciviruses and ssRNA viruses. This heterogeneity could also be explained by a single emergence of cruciviruses followed by multiple recombination events involving Rep with other CRESS-DNA viruses or plasmids.

Crucivirus Iteron Prediction and Analysis: The search for iterons within crucivirus genomes was undertaken in an attempt to develop a more reliable method for classifying relationships between cruciviruses. Given that both phylogenetic trees and sequence similarity networks result in inconclusive results taking a more targeted approach may improve the resolution of similarities and differences amongst cruciviruses. Of the 277 crucivirus genomes (Chapter two) (28) that contain a predicted stem-loop structure CRUISE was able to predict the presence of iterons in 257 genomes. The same stem-loop and iteron analyses on the second set of currently unpublished genomes identified an additional 230 genomes (out of 331 currently annotated) in which both a stem loop and associated iterons could be predicted. In this set of 487 total genomes 138 genomes that contain at least one iteron that is identical to those previously described for other eukaryotic CRESS-DNA viruses in addition to their own unique crucivirus iterons were also identified (**Table 3.1**). The lack of detectable stem-loops in some genomes may be a function of the manner in which

StemLoop-Finder predicts stem-loops. By first searching for nonanucleotides in the apex of stems it may miss stem-loops that do not display this canonical structure, a number of which have been previously reported in CRESS-DNA virus genomes isolated from feces of pigs, and other animals (6,43). However, these types of stem-loops have not been experimentally demonstrated to be ori structures. Similarly, the lack of detected iterons in some genomes with predicted stem-loops may be an artifact related to the relatively stringent search parameters utilized.

Iteron Sequence 5'-3'	**Virus**	**Unique CruV Occurrences**	**Identical Nonanucleotide Occurrences**
GGTGTC	Tomato leaf curl virus - New Delhi A2 (geminivirus)	34	4
GGCGT	Tomato Leaf Curl - New Delhi Cucumber (geminivirus)	40	3
GGAGT	Tomato mottle virus (geminivirus)	45	4
GGTGTC	Tomato mottle virus (geminivirus)	14	0
GAGGACC	Tobacco curly shoot betasatellite (begomovirus)	6	1
GGAGCCAC	Starling circovirus	0	0
GGAACCAC	Finch circovirus	1	1
GGAGCCAC	Raven/Canine circovirus	5	0
GGGGCCAT	Gull circovirus	0	0
GTACTCC	Duck circovirus	0	0
GTACTCC	Goose circovirus	6	2
CGGCAG	Porcine circovirus 1 and 2	9	6
GGGGCACC	Beak and feather disease virus 1 and 2 (circovirus)	3	1

Table 3.1. Iterons present in other CRESS-DNA viruses identified in cruciviruses. The number of unique occurrences of a given iteron sequence is noted as is the number of genomes that contain an identical nonanucleotide sequence to the virus from which an iteron was identified.

Previous work identified iterons adjacent to stem loops as playing a critical role in the replication of geminiviruses, likely determining specificity of dsDNA binding by Rep (13,44). Similarly, Rep of porcine circovirus type 1 (PCV1) has a preferential minimal binding (in vitro) site consisting of two direct hexamer repeats (iterons) and the right arm of the stem loop (17). If that DNA region is not available for binding, Rep can bind further downstream to a set of two additional hexamer repeats (17). The Boiling Springs Lake RNA-DNA Hybrid Virus (BSL-RDHV) genome contains only one set of hexamer repeats suggesting that origin binding by BSL-RDHV Rep may be dependent on the presence of this sole set of hexamer repeats as compared to PCV1 (41) (**Fig. 3.3A**). Additionally, previous work in geminiviruses has shown that the presence of imperfect repeats does not lead to a complete loss of replication, rather imperfect repeats can lead to lower levels of progeny virus production, implying that Rep may be capable of sub-optimal ori binding (45–47).

Based on previous work (8,12,14) an attempt was made to identify SPDs in Rep from BSL-RDHV and other cruciviruses containing iterons of a similar nature. The BSL-RDHV iteron sequence (5'-CGGCAG-3') was identified in 9 crucivirus genomes with one crucivirus genome, CruV-425, containing an imperfect repeat of the BSL-RDHV iteron (**Fig. 3.3A**). Based on the similarity of their iterated DNA sequences it was predicted that Rep of BSL-RDHV and CruV-425 would contain conserved amino acids in SPD regions. However, when Rep sequences of BSL-RDHV and CruV-425 were aligned, these previously identified

regions did not contain well conserved amino acids (**Fig. 3.3B**). Similarly, CruV-53 and NW_Brin_10B_C131 (a currently unpublished genome) contain identical predicted iterons (**Fig. 3.4A**) but no apparent SPDs within their Rep proteins in the previously predicted SPD regions (**Fig. 3.4B**). This inability to detect SPDs in crucivirus genomes was consistent across the pairs of genomes that were examined in detail (CruV-207 and CruV-341, CruV-108 and CruV-420 not presented).

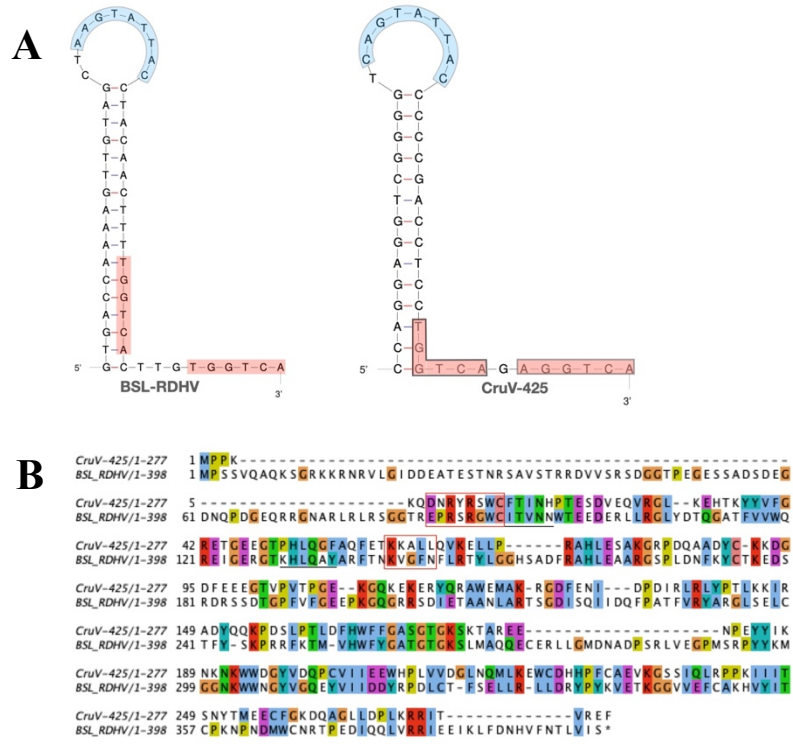

Figure 3.3. A) Predicted stem-loops structures of BSL-RDHV and CruV-425. Nonanucleotides are shaded in blue and iterons are shown in red. **B) Alignment of Rep from BSL-RDHV and CruV-425.** Boxed in red are previously predicted SPDs of CRESS-DNA viruses. The highly conserved motif I and motif II are underlined in black.

While it has been previously predicted that amino acids constituting SPDs would reside in the same region of diverse Reps, due to apparent common ancestry, (12,48), this does not appear to be the case for cruciviruses. It may follow that these SPDs are in different regions of Rep in cruciviruses. Further

analyses of other regions of Rep may be able to uncover these potential SPD regions. The inability to link DNA iterons with SPDs in Rep of cruciviruses could be a result of the large diversity observed in their Rep sequences relative to Reps of other CRESS-DNA viruses (28) (Chapter 2), a number of crucivirus Reps (in both genome sets) that apparently require splicing prior to translation based on the presence of intergenic stop codon(s). It may follow that additional crucivirus Reps, even those lacking an intergenic stop codon, require splicing which could result in different amino acids in SPDs from what these analyses have identified. The possibility also exists that iterons have been mis-annotated . This seems unlikely as the ori regions in which we identified iterons have been analyzed using a DNA repeat finder (https://www.novoprolabs.com/tools/repeats-sequences-finder) which did not identify additional repeated sequences left unannotated by CRUISE. It may prove useful to take the opposite approach to what was described above. Specifically, searching small groups of Reps of cruciviruses for conserved amino acids in regions that are not widely conserved may serve to better identify potential SPDs.

The presence of previously biochemically characterized iterons in these newly described crucivirus genomes coupled with the presence of unique crucivirus iterated DNA sequences is a puzzling observation. Others have not reported other CRESS-DNA viruses whose ori regions contain apparent iterons of multiple viruses. The occurrence of more than one distinct set of iterons in an ori region may suggest that Rep from one virus may be able to bind to the ori of a different virus. During a coinfection event it may be possible that Rep of one

crucivirus (or more broadly a CRESS-DNA virus) may mediate the initiation and termination of RCR for a different virus, perhaps sub-optimally.

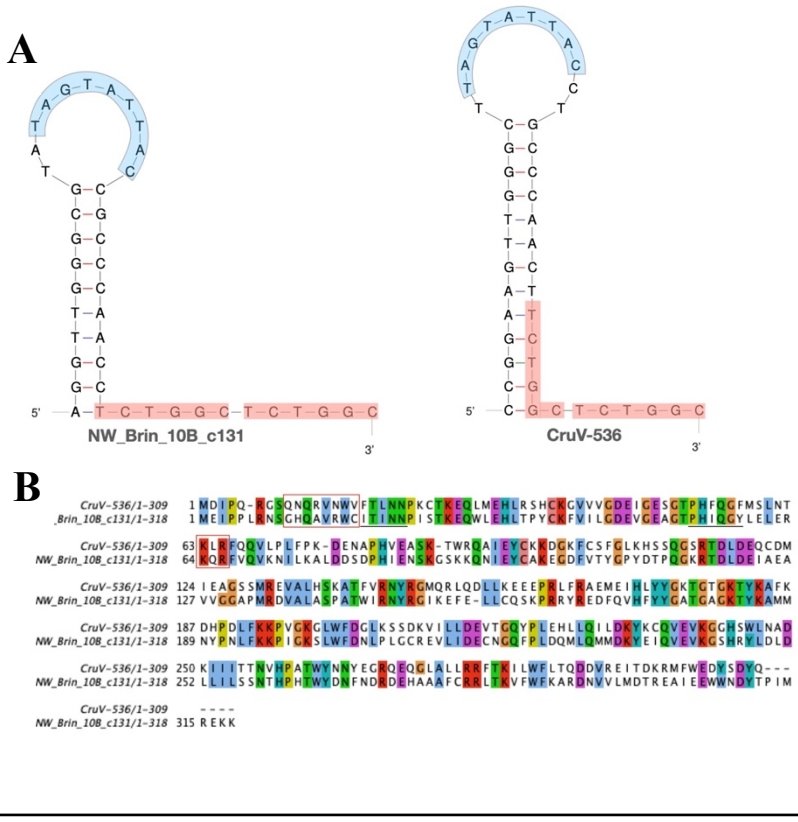

Figure 3.4. A) Predicted stem-loops structures of NW_Brin_10B_c131 and CruV-536. Nonanucleotides are shaded in blue and iterons are shown in red. **B) Alignment of Rep from NW_Brin_10B_c131 and CruV-536.** Boxed in red are previously predicted SPDs of CRESS-DNA viruses. The highly conserved motif I and motif II are underlined in black.

This hypothesis may be partially supported by observations made in the plant infecting begomoviruses. Monopartite begomoviruses have increasingly been found in association with betasatellites, small ssDNA molecules that play a critical role in disease symptom development (27,49,50). These betasatellites rely on the “helper” begomovirus for replication and encapsidation for intra host spread (27). It has been demonstrated that Rep from various helper begomoviruses (with differing iterons) are capable of replicating a given betasatellite, suggesting a “promiscuous” interaction between Rep and iterated DNA sequences (46,51). Perhaps the presence of more than one unique iterated

sequence in cruciviruses points to a similar possibility. Additionally, the ability of Rep to be active on more than one ori sequence may provide clues as to why these regions appear to be recombination hot-spots in CRESS-DNA viruses (52,53). While this is apparently a slightly different situation than those observed in begomoviruses and their betasatellites, the possibility exists that similar promiscuity of Rep exists in cruciviruses. This hypothesis could be explored with in vitro biochemical experiments utilizing purified Rep and varying iterated sequences near a stem-loop. Similarly, the successful development of an *Escherichia coli* or *Agrobacterium tumefaciens* system supporting RCR of crucivirus genomes (Chapter four) could effectively explore this hypothesis (54,55).

References

1. Rosario K, Duffy S, Breitbart M. A field guide to eukaryotic circular single-stranded DNA viruses: Insights gained from metagenomics. Arch Virol. 2012;157(10):1851–71.
2. Zhao L, Rosario K, Breitbart M, Duffy S. Eukaryotic circular Rep-encoding single-stranded DNA (CRESS-DNA) viruses: ubiquitous viruses with small genomes and a diverse host range. Adv Virus Res. Elsevier Inc.; 2018. 1–63.
3. Lefkowitz EJ, Dempsey DM, Hendrickson RC, Orton RJ, Siddell SG, Smith DB. Virus taxonomy: The database of the International Committee on Taxonomy of Viruses (ICTV). Nucleic Acids Res. 2018;46(D1):D708–17.
4. Scholthof KBG, Adkins S, Czosnek H, Palukaitis P, Jacquot E, Hohn T, et al. Top 10 plant viruses in molecular plant pathology. Mol Plant Pathol. 2011;12(9):938–54.
5. Picó B, Díez MJ, Nuez F. Viral diseases causing the greatest economic losses to the tomato crop. II. The tomato yellow leaf curl virus - a review. Sci Hortic (Amsterdam). 1996;67(3):151–96.
6. Cheung AK, Ng TFF, Lager KM, Alt DP, Delwart E, Pogranichniy RM. Identification of several clades of novel single-stranded circular DNA viruses with conserved stem-loop structures in pig feces. Arch Virol. 2015;160(1):353–8.
7. Rosario K, Dayaram A, Marinov M, Ware J, Kraberger S, Stainton D, et al. Diverse circular ssDNA viruses discovered in dragonflies (Odonata: Epiprocta). J Gen Virol. 2012; 2012;93(Pt 12):2668-2681.
8. Dayaram A, Potter KA, Moline AB, Rosenstein DD, Marinov M, Thomas JE, et al. High global diversity of cycloviruses amongst dragonflies. J Gen Virol. 2013;94(Pt 8):1827–40.
9. López-bueno A, Tamames J, Velázquez D, Moya A, Alcamí A, Quesada A. High Diversity of the viral community from an Antarctic Lake. Science. 2009; 6;326(5954):858-61.
10. Kraberger S, Argüello-Astorga GR, Greenfield LG, Galilee C, Law D, Martin DP, et al. Characterisation of a diverse range of circular replication-associated protein encoding DNA viruses recovered from a sewage treatment oxidation pond. Infect Genet Evol. 2015;31:73–86.
11. Krupovic M, Varsani A, Kazlauskas D, Breitbart M, Delwart E, Yutin N, et al. *Cressdnaviricota*: a virus phylum unifying 7 families of Rep-encoding viruses with single-stranded, circular DNA genomes. J Virol. 2020; 1;94(12):e00582-20.

12. Londoño A, Riego-Ruiz L, Argüello-Astorga GR. DNA-binding specificity determinants of replication proteins encoded by eukaryotic ssDNA viruses are adjacent to widely separated RCR conserved motifs. Arch Virol. 2010;155(7):1033–46.

13. Xu X, Qian Y, Wang Y, Li Z, Zhou X. Iterons homologous to helper geminiviruses are essential for efficient replication of betasatellites. J Virol. 2018;93(5):1–22.

14. Ruiz-medrano R, Irapuato U. An iteron-related domain is associated to Motif 1 in the replication proteins of geminiviruses: identification of potential interacting amino acid-base pairs by a comparative approach. Arch Virol. 2001;146(8):1465-85.

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

16. Ramos PL, Guevara-González RG, Peral R, Ascencio-Ibañez JT, Polston JE, Argüello-Astorga GR, et al. Tomato mottle Taino virus pseudorecombines with PYMV but not with ToMoV: Implications for the delimitation of cis- and trans-acting replication specificity determinants. Arch Virol. 2003;148(9):1697–712.

17. Steinfeldt T, Finsterbusch T, Mankertz A. Rep and Rep' protein of porcine circovirus type 1 bind to the origin of replication in vitro. Virology. 2001;291(1):152–60.

18. Orozco BM, Gladfelter HJ, Settlage SB, Eagle PA, Gentry RN, Hanley-Bowdoin L. Multiple cis elements contribute to geminivirus origin function. Virology. 1998;242(2):346–56.

19. Fontes EPB, Eagle PA, Sipe PS, Luckow VA, Hanley-Bowdoin L. Interaction between a geminivirus replication protein and origin DNA is essential for viral replication. J Biol Chem. 1994;269(11):8459–65.

20. Steinfeldt T, Finsterbusch T, Mankertz A. 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. 2006;80(13):6225–34.

21. Hafner GJ, Stafford MR, Wolter LC, Harding RM, Dale JL. Nicking and joining activity of banana bunchy top virus replication protein in vitro. J Gen Virol. 1997; 78 (Pt 7):1795-9

22. Gronenborn B. Nanoviruses: Genome organisation and protein function. Vet Microbiol. 2004;98(2):103–9.

23. Gilbertson RL, Hidayat SH, Paplomatas EJ, Rojas MR, Hou YM, Maxwell DP. Pseudorecombination between infectious cloned DNA components of tomato mottle and bean dwarf mosaic geminiviruses. J Gen Virol. 1993;74(1):23–31.

24. Unseld S, Ringel M, Höfer P, Höhnle M, Jeske H, Bedford ID, et al. Host range and symptom variation of pseudorecombinant virus produced by two distinct bipartite geminiviruses. Arch Virol. 2000;145(7):1449–54.

25. Sung YK, Coutts RHA. Pseudorecombination and complementation between potato yellow mosaic geminivirus and tomato golden mosaic geminivirus. J Gen Virol. 1995;76(11):2809–15.

26. Unseld S, Ringel M, Konrad A, Lauster S, Frischmuth T. Virus-specific adaptations for the production of a pseudorecombinant virus formed by two distinct bipartite geminiviruses from Central America. Virology. 2000;274(1):179–88.

27. Zhou X. Advances in understanding begomovirus satellites. Annu Rev Phytopathol. 2013;51:357–81.

28. de la Higuera I, Kasun GW, Torrance EL, Pratt AA, Maluenda A, Colombet J, et al. Unveiling crucivirus diversity by mining metagenomic data. mBio. 2020;11(5):1–17.

29. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.

30. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25(9):1189–91.

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

32. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–3.

33. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst Biol. 2010;59(3):307–21.

34. Letunic I, Bork P. Interactive Tree of Life (iTOL) v4: Recent updates and new developments. Nucleic Acids Res. 2019;47(W1):256–9.

35. Zallot R, Oberg N, Gerlt JA. From the bench: The EFI web resource for genomic enzymology web tools : leveraging protein , genome , and

metagenome databases to discover novel enzymes and metabolic pathways. Biochem. 2019; 15;58(41):4169-4182.

36. Shannon P, Markiel A, Owen Ozier, Nitin S. Baliga, 1Jonathan T. Wang DR, Amin N, Benno Schwikowski and TI, Er. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498-504

37. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003;31(13):3406–15.

38. Liu Q, Wang H, Ling Y, Yang SX, Wang XC, Zhou R, et al. Viral metagenomics revealed diverse CRESS-DNA virus genomes in faeces of forest musk deer. Virol J. 2020;17(1):1–9.

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

40. Krupovic M. Recombination between RNA viruses and plasmids might have played a central role in the origin and evolution of small DNA viruses. BioEssays. 2012;34(10):867–70.

41. Diemer GS, Stedman KM. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses Biol Direct. 2012;7(13):1–14.

42. Stedman K. Mechanisms for RNA capture by ssDNA viruses: Grand theft RNA. J Mol Evol. 2013;76(6):359–64.

43. Li L, Kapoor A, Slikas B, Bamidele OS, Wang C, Shaukat S, et al. Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces. J Virol. 2010;84(4):1674–82.

44. Timchenko T, de Kouchkovsky F, Katul L, David C, Vetten HJ, Gronenborn B. A single Rep protein initiates replication of multiple genome components of faba bean necrotic yellows virus, a single-stranded DNA virus of plants. J Virol. 1999;73(12):10173–82.

45. Dry IB, Krake LR, Rigden JE, Rezaian MA. A novel subviral agent associated with a geminivirus: The first report of a DNA satellite. Proc Natl Acad Sci U S A. 1997;94(13):7088–93.

46. Mansoor S, Briddon RW, Bull SE, Bedford ID, Bashir A, Hussain M, et al. Cotton leaf curl disease is associated with multiple monopartite begomoviruses supported by single DNA β. Arch Virol. 2003;148(10):1969–86.

47. Saunders K, Briddon RW, Stanley J. Replication promiscuity of DNA-β satellites associated with monopartite begomoviruses; deletion

mutagenesis of the Ageratum yellow vein virus DNA-β satellite localizes sequences involved in replication. J Gen Virol. 2008;89(12):3165–72.

48. Ilyina T V, Koonin E V. Conserved sequence motifs in the initiator proteins for rolling circle dna-replication encoded by diverse replicons from Eubacteria, Eukaryotes and Archaebacteria. Nucleic Acids Res. 1992;20(13):3279–85.

49. Jose J, Usha R. Bhendi yellow vein mosaic disease in India is caused by association of a DNA β satellite with a begomovirus. Virology. 2003;305(2):310–7.

50. Briddon RW, Mansoor S, Bedford ID, Pinner MS, Saunders K, Stanley J, et al. Identification of DNA components required for induction of cotton leaf curl disease. Virology. 2001;285(2):234–43.

51. Kon T, Rojas MR, Abdourhamane IK, Gilbertson RL. Roles and interactions of begomoviruses and satellite DNAs associated with okra leaf curl disease in Mali, West Africa. J Gen Virol. 2009;90(4):1001–13.

52. Lefeuvre P, Lett J-M, Varsani A, Martin DP. Widely conserved recombination patterns among single-stranded DNA viruses. J Virol. 2009;83(6):2697–707.

53. Martin DP, Biagini P, Lefeuvre P, Golden M, Roumagnac P, Varsani A. Recombination in eukaryotic single stranded DNA viruses. Viruses. 2011;3(9):1699–738.

54. Stenger DC, Revington GN, Stevenson MC, Bisaro DM. 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. 1991; 15;88(18):8029-33.

55. Cheung AK. Rolling-Circle Replication of an Animal Circovirus Genome in a Theta-Replicating Bacterial Plasmid in *Escherichia coli*. J Virol. 2006;80(17):8686–94.

Chapter Four

Biochemical Activities of the Replication Associated Protein of Boiling Springs Lake RNA DNA Hybrid Virus

Abstract

Cruciviruses are currently unclassified circular Rep-encoding single-stranded DNA viruses (CRESS-DNA virus) whose genomes suggest gene transfer between RNA and DNA viruses due to a putative capsid protein gene closely related to capsid protein genes found in single stranded RNA viruses. Boiling Springs Lake RNA-DNA hybrid virus (BSL-RDHV) contains a putative intergenic stem-loop structure that may serve as an origin of rolling circle replication (RCR), and a putative replication-associated protein (Rep) similar to other CRESS-DNA viruses. In this study it is shown, for the first time, that a crucivirus Rep is capable, in vitro, of the predicted biochemical activities associated with initiation and completion of RC, including ATP hydrolysis, DNA binding, DNA nicking and joining. The results of this study confirm, biochemically, that BSL-RDHV likely replicates its genome by RCR.

Introduction

Cruciviruses are currently unclassified single stranded DNA (ssDNA) viruses discovered exclusively through metagenomic studies and direct environmental DNA amplification and cloning (1–15). Their genomes contain at least two open reading frames (ORFs): one encoding a putative replication associated protein (Rep) similar to those found in members of the circular Rep-

encoding ssDNA viruses (CRESS-DNA), while the other conserved ORF encodes a putative capsid protein (CP) similar to those found in the ssRNA plant infecting *Tombusviridae* (1,16). This arrangement in which a single viral genome contains genes apparently from both DNA and RNA viruses is the first direct evidence of gene exchange between viruses containing disparate genome types (1,17,18). Hypothetical biochemical mechanisms for such a recombination event dependent on the Rep protein have been proposed, but no biochemical activities of any cruciviral Rep protein have been demonstrated. Gaining a better understanding of crucivirus biology and biochemistry could shed light on a very poorly understood mechanism of viral evolution, specifically recombination between viruses harboring genomes composed of different nucleic acids (17,19,20).

The first described crucivirus genome was discovered during a metagenomic survey of Boiling Springs Lake (BSL); a high temperature and low pH lake in Lassen Volcanic National Park (1). Named Boiling Springs Lake RNA-DNA Hybrid Virus (BSL-RDHV), this 4.1kb ssDNA genome contains four ORFs: a putative Rep gene similar to those found in the *Cressdnaviricota* (ssDNA viruses) (1,21), a putative capsid gene (*cap*) homologous to *cap* genes found in *Tombusviridae*, and two ORFs that do not contain significant sequence homology to publicly available sequences (1). The putative *Rep* and *cap* genes are arranged unidirectionally “head to tail” on the predicted virion sense DNA strand (1) (**Fig. 4.1**). Since this initial discovery approximately 500 additional crucivirus

genomes have been identified in metagenomes from diverse environments (2-15).

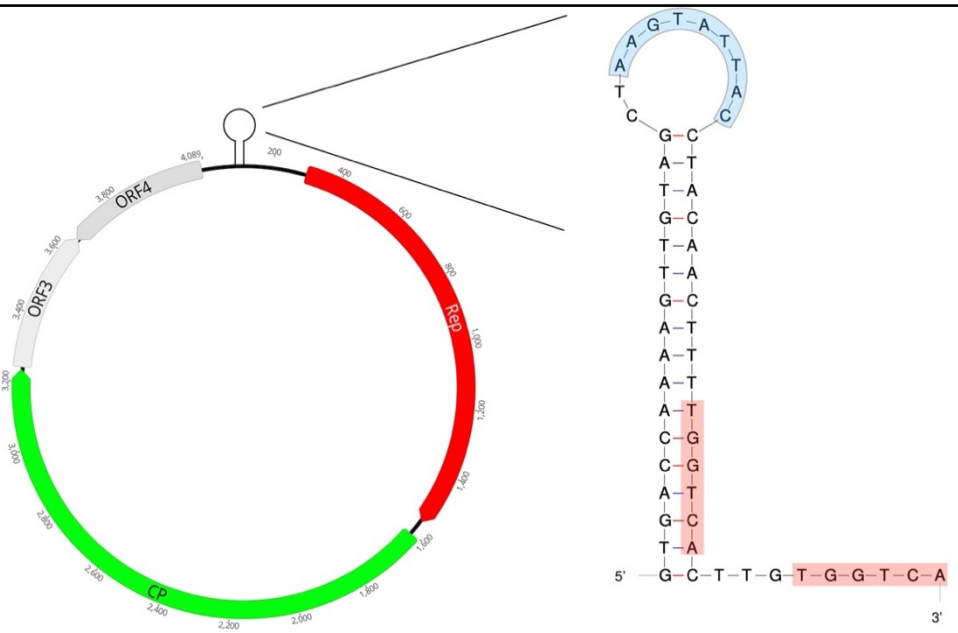

Figure 4.1. Position of the putative stem-loop within the genome of BSL-RDHV (1). The detailed view highlights the stem loop structure and associated nonanucleotide loop in blue and the direct hexamer repeats (iterons) in red.

The hallmark and unifying feature of viruses belonging to the *Cressdnaviricota* is the presence of a *Rep* gene whose translated protein's N-terminal region belongs to the HUH endonuclease superfamily (16,21–23). This family of proteins, similar to the tyrosine recombinases in function, is widespread across all three domains of life and is involved in initiation and completion of replication for a diverse range of mobile genetic elements via the catalysis of cleavage and joining reactions on ssDNA substrates through the action of one or two tyrosine residues located in the N-terminus, specifically motif III in CRESS-DNA viruses (22–26) (**Fig. 2.2 and Fig. 4.2A**). Two additional motifs, motifs I and

II, are found in the N-terminus of CRESS-DNA Reps. Motif I has been predicted to be involved in dsDNA binding specificity (27–29), while motif II contains the namesake HUH arrangement (two histidine residues separated by a large hydrophobic amino acid) which is active in metal ion coordination (22–24,30,31). Many CRESS-DNA viruses exhibit a non-canonical motif II sequence (HUH) such as the circoviruses which display an HUQ architecture (30,32). Members of the *Geminiviridae* and *Genomoviridae* contain a fourth Rep motif, geminivirus Rep sequence (GRS), which has been demonstrated to be required for replication of geminiviruses in plants (33). The C-terminal domain of CRESS-DNA virus Reps contains a superfamily-3-helicase (SF3), found in small DNA and RNA viruses (34). SF3 helicases are characterized by the presence of four conserved motifs: Walker A/P-loop and Walker B involved in ATPase activity, B', and motif C (35). Additionally, CRESS-DNA viruses contain a conserved arginine-finger in their C-terminal domain (36).

Following host cell entry CRESS-DNA virus genomes are converted to supercoiled double-stranded DNA (dsDNA), which serves as a template for both the replication of the ssDNA viral genome and the transcription of viral genes (16,37–39) (**Fig. 1.3)**. The origin of replication (ori) for a number of CRESS-DNA viruses such as members of the *Circoviridae, Geminiviridae*, and *Nanoviridae* has been localized to intergenic stem loops, which consist of an inverted repeat stem section and a loop at the stem apex 9-11 nucleotides in length (38,40–43) (**Fig. 3.3A and Fig. 4.1**). The loop located at the stem apex typically contains 9

nucleotides (nonanucleotide) with sequences similar to TAGTATTAC (*Circoviridae*) or TAATATTAC (*Geminiviridae*) (16). Rep proteins of some CRESS-DNA viruses have been shown to bind dsDNA near these stem loops dependent on repeated DNA sequences called iterons (28,32,44). Once bound, Rep induces conformational changes in the nonanucleotide region apparently leading to a region of ssDNA in the loop portion of the stem loop (41,42). This region of ssDNA serves as a substrate for cleavage of the phosphodiester bond (nick) between the 7th and 8th positions of the nonanucleotide (TAGTATT/AC or TAATATT/AC in which / represents the nick site) (32,38,40,41,45). Rep remains covalently bound to the newly generated 5' end via a phosphotyrosine bond, while the newly exposed 3'-OH serves as the initiation point for unidirectional DNA replication by host cell DNA polymerases via rolling circle replication (RCR) (25). The helicase activity associated with the C-terminal SF3 domain of Rep likely functions in unwinding viral DNA, allowing cellular enzymes to polymerize a new viral genome (46–48). Following one or more rounds of RCR, viral genomes are sealed by a joining reaction also mediated by Rep (32,41,49). In CRESS-DNA viruses these nicking and joining reactions are mediated by a single catalytic tyrosine residue located in conserved motif III of the N-terminus of Rep (22,26,40,41). While Rep alone has been shown to be required and indispensable for replication of diverse geminiviruses and nanoviruses, a second spliced form of Rep containing identical N-terminal motifs and a frameshifted C-

terminus, known as Rep', has been demonstrated to also be required for genome replication of porcine circovirus type 1 (PCV1) (38,50,51).

Previous work identified a putative intergenic stem-loop structure and associated hexamer repeats (iterons) in the genome of BSL-RDHV that may serve as an ori of RCR (1) (**Fig. 4.1**). This observation coupled with the presence of a putative Rep containing all necessary N-terminal motifs for initiation and completion of RCR, and the C-terminal SF3 helicase domain (**Fig. 4.2A**) led to the hypothesis that Rep of BSL-RDHV should be capable of demonstrating RCR initiation and completion activities in vitro. Specifically Rep of BSL-RDHV should: (i) demonstrate ATP hydrolysis indicative of helicase activity, (ii) bind to dsDNA containing the predicted stem loop and associated iterons, (iii) nick ssDNA containing the BSL-RDHV nonanucleotide sequence to allow the initiation of RCR, and (iv) join nicked ssDNA characteristic of the completion of ssDNA replication by RCR. It has been previously hypothesized that Rep of cruciviruses, or more broadly Rep of CRESS-DNA viruses, may be involved in the novel and uncharacterized DNA-RNA recombination event that led to the initial emergence of cruciviruses (17). While this hypothesis was neither supported or rejected through biochemical experiments, this chapter presents initial biochemical characterization of activities necessary RCR initiation and termination reactions catalyzed by Rep of BSL-RDHV.

This chapter demonstrates, for the first time, that purified recombinant Rep of a crucivirus (BSL-RDHV) is capable, in vitro, of the activities necessary for

initiation and completion of RCR, strongly suggesting that cruciviruses are truly ssDNA viruses that replicate their genomes via RCR. This is also the first demonstration of biochemical activities related to both dsDNA and ssDNA associated with a member of the very large number of unclassified *Cressdnaviricota* genomes (21).

Methods

BSL-RDHV-Rep Overexpression Results in Insoluble Protein: Preliminary BSL-RDHV-Rep overexpression work showed that the overexpression of two variants of BSL-RDHV Rep in a number of *Escherichia coli* overexpression strains (BL21DE3, BL21DE3 Rosetta, BL21 DE3 ArcticExpress) under varying growth temperatures, varying growth media, varying isopropyl β-d-1-thiogalactopyranoside (IPTG) concentrations, and utilizing varying cell lysis buffers resulted in the overexpressed protein remaining insoluble. Briefly, the stop codon immediately downstream of BSL-RDHV Rep Motif II was removed using Gibson Assembly to generate an overexpression vector with a 6x N-terminal histidine (6x His) purification tag; pET30b-6HN-RepΔ133-153 and pET30b-6HN-Rep-*138W (* represents a stop codon). Two additional C-terminal 6x His constructs were made using restriction enzyme cloning, pET21b-6HC-RepΔ133-153. pET21b-6HC-Rep-*138W. However, no soluble BSL-RDHV Rep could be recovered using these constructs and varying overexpression conditions. Next an attempt to clone (via Gibson Assembly) BSL-RDHV-Rep into pFastBac (ThermoFisher Scientific) to establish an insect cell expression system was carried out, but no successful constructs were obtained.

Head to tail RDHV construction: To explore the possibility of replicational release of a portion of BSL-RDHV genome from a theta-replicating plasmid, due to the activity of Rep, head to tail constructs containing 2 copies of the portion of the BSL-RDHV genome that contains the *Rep* gene and predicted origin of replication were cloned in pBluescript KS+. Briefly, the WT *Rep* gene of BSL-RDHV, which contains an intragenic stop codon, was replaced with RepΔ133-153 by Gibson cloning. Two copies of the BSL-RDHV genome segment containing ORF4, RepΔ133-153, and the predicted stem loop were inserted in a head to tail fashion in the theta replicating pBluescript KS+ by Gibson cloning to generate pBluescript KS+RDHV (**Fig. 4.4A**). Constructs were confirmed to be free of PCR misincorporations by Sanger sequencing (Eurofins Genomics). pBluescript KS+RDHV was transformed into Top10 *E. coli* and grown overnight in LB with 100µg/ml ampicillin. Plasmids were extracted using the GET Plasmid Miniprep Kit (G-Biosciences), digested with XhoI and resolved on a 0.7% agarose gel (**Fig. 4.4B**).

Codon Optimization: Due to the presence of approximately 70 rare codons (for *E. coli*) in BSL-RDHV, a codon-optimized BSL-RDHV-RepΔ133-153 (RepD1) containing an N-Terminal 6x histidine tag and an in frame terminal stop codon was synthesized by Integrated DNA Technologies (IDT). pUC-IDT-Optimized-RepΔ133-153 (pUC-RepD1) was purchased from IDT using their *E. coli* codon optimization tool to design the appropriate DNA sequence.

Construction of pET21b-BSL-RDHV-RepΔ133-153 expression vector: Codon optimized RepD1 was cloned from pUC-RepD1 into pET21-b using

Gibson Cloning to create an *E. coli* overexpression vector. pET21b was linearized and amplified by polymerase chain reaction (PCR). 1ng of pET21b was used in a 50µl PCR that consisted of 0.5µM pET21b_F and 0.5µM pET21b_R (**Table 4.1)**, 200µmM dNTP's, 0.5U Phusion Polymerase (New England Biolabs), and 1X Phusion Reaction Buffer. Reactions were initially denatured for 10 minutes at 96 °C followed by 35 cycles of 96 °C for 30 seconds, 52.5 °C for 30 seconds, and 72 °C for 2 minutes. A final extension of 72 °C for 10 minutes was employed to complete amplification. PCR products of the expected size were confirmed by 0.7% agarose gel electrophoresis. pET21b amplification/linearization reactions were digested with 10U DpnI (New England Biolabs) at 37 °C for two hours in order to avoid transforming circularized (empty vector) PCR template. DpnI digests were purified using the Monarch PCR Cleanup Kit (New England Biolabs) and quantified using a NanoDrop instrument.

1ng of pUC-RepD1 was used in a 20µl PCR that consisted of 0.5µM Opt_Rep_21b_F and Opt_Rep_21b_R (**Table 4.1),** 200µM dNTP's, 1U Phusion Polymerase (New England Biolabs), and 1X Phusion Reaction Buffer. Reactions were initially denatured for 5 minutes at 95 °C followed by 35 cycles of 95 °C for 30 seconds, 62.5 °C for 30 seconds, 72 °C for 20 seconds. A final extension of 10 minutes at 72 °C was included to complete amplification. PCR products of the expected size were confirmed by 0.7% agarose gel electrophoresis. Reactions were purified and quantified as above.

Gibson cloning reactions were carried out using 150ng of linearized pET21b, a threefold molar excess (104ng) of PCR amplified RepD1, and 1x

Gibson Assembly Master Mix (New England Biolabs) in a 20µl reaction. Reactions were incubated at 50 °C for one hour. Gibson assembly reactions were diluted 4 fold in water and 2µl was transformed by heat shock at 42 °C for 30 seconds into chemically competent TOP10 *E. coli* and grown on LB plates with 100µg/ml ampicillin. Resulting colonies were screened for inserts using colony PCR and patched onto LB plates. Cells collected on a pipette tip from each patch was resuspended in 50µl of water, and then heated to 98 °C for 10 minutes. 1µl of this heated cell suspension was then used as a template for PCR using T7F and T7R primers (New England Biolabs Taq MasterMix). Samples from patches showing the correct PCR product size were grown overnight in 10ml of liquid LB containing 100µg/ml ampicillin. Plasmids were extracted the following morning using the GET Plasmid Miniprep Kit (G-Biosciences). pET21b-6HN-BSL-RDHV-RepΔ133-153 constructs (pET21b-RepD1) (**Fig. 4.3A**) were confirmed to be free of PCR misincorporations by Sanger sequencing (Eurofins Genomics).

Construction of pET21b-BSL-RDHV-RepΔ133-162 : 1ng of pET21b-RepD1 was used in a 20µl "inverse PCR" consisting of 0.5µM Motif_2_F and Motif_2_R (**Table 1**, see page 123)**,** 200µM dNTP's, 0.5U Phusion Polymerase (New England Biolabs), and 1X Phusion Reaction Buffer. Reactions were initially denatured for 10 minutes at 96 °C followed by 35 cycles of 95 °C for 30 seconds, 62.5 °C for 30 seconds, 72 °C for 4 minutes. A final extension of 10 minutes at 72 °C was included to complete amplification. PCR products of the expected size were confirmed by 0.7% agarose gel electrophoresis. Reactions were digested

with DpnI, and the entire reaction was precipitated and quantified via a nanodrop instrument.

Linearized PCR products were phosphorylated using the New England Biolabs Quick Blunting kit. 1µg purified PCR product was used in a reaction following the manufacturer's protocol for PCR products. Phosphorylated PCR products were then ligated overnight at 15 °C using the New England Biolabs Quick Ligation Kit. Ligations were precipitated and transformed into Top10 *E. coli.* pET21b-BSL-RDHV-RepΔ133-162 (pET21b-RepD2) (**Fig. 4.3A**) were confirmed to be free of PCR misincorporations by Sanger sequencing (Eurofins Genomics).

Protein Expression: pET21b-RepD1 and pET21b-RepD2 were separately transformed into chemically competent *E. coli* ArcticExpress (DE3) (Agilent Technologies) cells by heat shock transformation as described above. Cells were grown overnight at 37 °C on LB plates containing 100µg/ml ampicillin. The following evening an individual colony was picked and grown overnight at 37 °C with shaking at 250RPM in liquid LB containing 100µg/ml ampicillin, 20µg/ml gentamicin, 0.2% w/v glucose. The following morning the overnight culture was diluted into 500ml LB containing 0.2% w/v glucose in a 2L baffled flask and grown for 3 hours at 30 °C with shaking at 250RPM. Cultures were cooled for approximately 1 hour in an incubator set to 8 °C placed in a 4 °C cold room with shaking at 250RPM. Protein (RepD1 and RepD2) overexpression was induced by the addition of IPTG to a final concentration of 1mM. Protein overexpression was allowed to proceed at 8 °C for 36 hours.

Protein Purification: Cells were harvested by centrifugation at 10,000xg for 30 minutes at 4 °C. Cell pellets were resuspended in 10ml of column buffer (50mM NaH_2PO_4, pH 7.5, 300mM NaCl, 10mM imidazole) per gram of wet cell weight supplemented just before use with 20U DNase I (New England Biolabs) and 1mM PMSF. Cells were lysed by 12 sonication bursts of approximately 10 seconds using 75% power with a semi-micro tip in an ice water bath. Lysates were cleared by centrifugation at 20,000xg for 30 minutes at 4 °C.

Cleared lysates were applied to 2.5ml of pre-equilibrated Ni-NTA beads in a 25ml column. Binding was achieved by allowing the cleared lysates to flow over the Ni-NTA resin twice, at a rate of approximately 1ml per minute. Bound protein was washed with 25ml column buffer. In order to remove Cpn60/10 co-purifying contaminant, beads with bound protein were washed with 50ml ATP wash buffer (20 mM HEPES, pH 7.0, 10 mM $MgCl_2$, 5 mM ATP, 150 mM KCl) (52). Beads with bound protein were then washed with 25ml each of increasing imidazole concentrations in column buffer (20, 40, 80, and 100mM imidazole). Protein was eluted and collected in 1ml fractions (50mM NaH_2PO_4, pH 7.5, 300mM NaCl, 300mM imidazole). Fractions were analyzed by 12% SDS-PAGE followed by Coomassie brilliant blue staining.

Ni-NTA Fractions of similar purity were pooled and subjected to a salt-out procedure. Briefly, ammonium sulfate was added to pooled fractions in 5% of saturation steps. Ammonium sulfate was added slowly while the pooled fractions were stirred gently in an ice-water bath. Following ammonium sulfate dissolution precipitated proteins were collected by centrifugation for 45 minutes at 20,000xg

and 4C. Greatest purities and concentrations of RepD1 and RepD2 were achieved at 25% ammonium sulfate saturation as determined by 12% SDS-PAGE followed by Coomassie staining. The identity of the SDS-PAGE band of the apparent correct MW (RepD1) (**Fig. 4.2B**) was verified by mass spectrometry by the Oregon Health and Science University Proteomics Shared Resource core facility. Purified RepD1 and RepD2 were concentrated and rebuffered using 30kDa cutoff spin concentrators at 3260xg and 4 °C. Purified RepD1 and RepD2

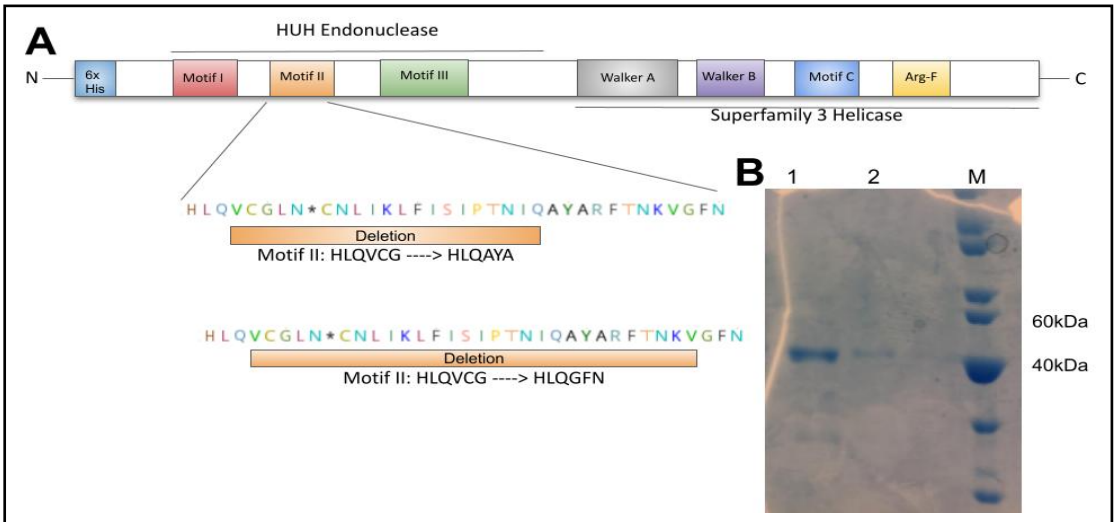

Figure 4.2. RepD1 cloning strategy and purification. A) The cloning strategy to generate RepD1 and RepD2 is shown. RepD1 was generated based on the presence of donor and acceptor nucleotides predicted to be involved in a splicing event to remove the stop codon at position 138. RepD2 was generated based on better conservation of the motif II sequence with other CRESS-DNA viruses. **B)** Ni-NTA purification and ammonium acetate precipitation of RepD1 (lane 1) and RepD2 (lane 2) results in purified protein as verified by mass spectrometry. (M) Molecular weight marker.

were quantified via Bradford assay (or with a nanodrop instrument) and rebuffered into 25mM TRIS pH 7.5, 50% glycerol for storage at -20 °C.

Construction of pET21-b MBP-RDHV Rep Fusion Constructs: Rep of BSL-RDHV was fused to maltose binding protein (MBP) to increase recombinant

protein yields, increase solubility, and provide a secondary purification tag (53). pET21b was linearized and amplified by polymerase chain reaction (PCR). 1ng of pET21b was used in a 50µl PCR that consisted of 0.5µM pET21b_F and 0.5µM pET21b_R (**Table 4.1,** see page 122), 200µM dNTP's, 0.5U Phusion Polymerase (New England Biolabs), and 1X Phusion Reaction Buffer. Reactions were initially denatured for 10 minutes at 96 °C followed by 35 cycles of 96 °C for 30 seconds, 52.5 °C for 30 seconds, and 72 °C for 2 minutes. A final extension of 72 °C for 10 minutes was employed to complete amplification.

The MBP gene along with an N-terminal 6x histidine tag, and a C-terminal tobacco etch virus (TEV) protease site was amplified by PCR from pLIC-HMK-MBP (supplied by Erik Chow, University of California, San Francisco). MBP was amplified with 5' overhang for cloning into pET21b and 3' overhang into codon-optimized BSL-RDHV Rep (above). A 20µl reaction containing 5 ng of pLIC-HMK, 0.5µM each MBP_F and MBP_R (**Table 4.1**), 200µM dNTP's, 0.5U Phusion Polymerase (New England Biolabs), and 1X Phusion Reaction Buffer. Reactions were initially denatured at 95 °C for 10 minutes followed by 35 cycles of 95 °C for 30 seconds, 62.5 °C for 30 seconds, 72 °C for 30 seconds. A final extension at 72 °C for 10 minutes completed the amplification. The resulting PCR product contained a 5' overhang into pET21b.

1ng of pUC-RepD1, or 1ng of pET30b-RepD2 (Chapter 3) was used in a 20µl PCR that consisted of 0.5µM each OR_OLMBP_F and OR_OL21b_R (**Table 4.1**), 200µM dNTP's, 1U Phusion Polymerase (New England Biolabs),

and 1X Phusion Reaction Buffer. Reactions were initially denatured for 5 minutes at 95 °C followed by 35 cycles of 95 °C for 30 seconds, 62.5 °C for 30 seconds, 72 °C for 20 seconds. A final extension of 10 minutes at 72 °C was included to complete amplification. The resulting PCR products contained a 5' overhang for cloning into MBP and a 3' overhang for cloning into pET21b.

PCR products of the expected size for all reactions were confirmed by 0.7% agarose gel electrophoresis. All reactions were digested overnight at 37 °C with 5U of DpnI to remove circular template. All reactions were purified (New England Biolabs Monarch PCR Purification Kit) and quantified using a NanoDrop instrument.

Gibson cloning reactions were carried out in a 20µl final volume containing 150ng of linearized pET21b, a 5-fold molar excess of both MBP and RepD1 or RepD2, and 1X HiFi DNA assembly mastermix (New England Biolabs). Reactions were incubated at 50 °C for 1 hour and then precipitated with sodium acetate/ethanol. The precipitated reaction was resuspended in 4µl water and was transformed by heat shock into chemically competent TOP10 *Escherichia coli.* Inserts were initially confirmed by colony PCR (as described above) and were shown to be free of PCR misincorporations by Sanger sequencing (**Fig. 4.3A**) (Eurofins Genomics).

Overexpression of MBP-RepD1/D2 Chemically competent BL21 DE3 pLysS *E. coli* (Novagen) were transformed with pET21b-MBP-RepD1 or pET21b-MBP-RepD2 or pLIC-HMK-MBP by 30s heat shock at 42 °C. Cells were plated on lysogeny broth (LB) plates containing 100µg/mL ampicillin and 25µg/mL

chloramphenicol and grown overnight at 37 °C. The following evening an individual colony was picked and inoculated in liquid rich media (LB with 0.5X salt) supplemented with 100µg/ml ampicillin, 25µg/ml chloramphenicol and 0.2% w/v glucose. Cultures were grown overnight at 37 °C with shaking at 250RPM . The next morning 10ml overnight cultures were diluted in 500mL rich media containing 100µg/ml ampicillin, 25µg/ml chloramphenicol and 0.2% w/v glucose in 2L baffled flasks. Two cultures totaling 1L were used for each overexpression. Cultures were grown for approximately 3 hours at 37 °C, 250RPM shaking. For the last 30 minutes (total initial growth of approximately 3.5 hours) flasks were incubated at room temperature with 250RPM shaking until an OD600nm of 0.5 was reached. Protein overexpression was induced by the addition of sterile 100mM IPTG to a final concentration of 1mM. Protein overexpression was allowed to proceed for 16 hours at room temperature.

MBP-RepD1/D2 Purification: Overexpression cultures were collected by centrifugation at 10,000xg and 4 °C for 30 minutes. Pelleted cells were resuspended in amylose column buffer (5ml/g cell weight) (50mM NaH2PO4, pH 7.5, 300mM NaCl) and supplemented with 10mg/ml lysozyme and 20U DNase I immediately prior to use. Cells were gently agitated for 30 minutes at 37 °C. Cell lysis was completed by 12 sonication bursts of approximately 10 seconds each, using 75% power and a semi-micro tip in an ice water bath. Lysates were cleared by centrifugation at 20,000xg for 30 minutes at 4 °C.

Cleared lysates were applied to 10ml of amylose resin (New England Biolabs) that had been pre-equilibrated with amylose column buffer. Cleared

lysates were allowed to flow over the resin twice at a rate of approximately 1ml per minute. Amylose resin with bound protein was washed with 100ml amylose column buffer. Approximately 30 1ml fractions were eluted with amylose elution buffer (50mM NaH2PO4, pH 7.5, 300mM NaCl, 10mM maltose). Fractions were analyzed by 10% SDS-PAGE followed by Coomassie staining. MBP-RepD1 and MBP-RepD2 Fractions of a similar purity were pooled and dialyzed against Ni-NTA buffer (50mM NaH2PO4, pH 7.5, 300mM NaCl, 10mM imidazole) overnight. Purified MBP was concentrated to approximately 10mg/ml (per Bradford assay and NanoDrop) using 30kDa cutoff spin concentrators.

Dialyzed MBP-RepD1 and MBP-RepD2 were further purified using AKTA FPLC and a 1ml Ni-NTA column. All FPLC protocols were carried out at 4 °C. Protein was loaded on the column at a rate of 0.25ml per minute and washed with 100ml of Ni-NTA buffer at a rate of 1ml per minute. Protein was eluted using a linear gradient of imidazole ranging from 20mM to 500mM with a 2.5% increase in imidazole concentration per minute. UV absorbance was monitored and fractions with UV-absorbance were again analyzed by SDS-PAGE. Fractions of a similar purity were pooled, concentrated and rebuffered into heparin column buffer (50mM NaH2PO4, pH 7.5, 200mM NaCl) using 100kDa cut-off spin concentrators (Amicon).

MBP-RepD1 and MBP-RepD2 were further purified using an AKTA-FPLC and a 1ml heparin column. Protein was loaded on the column at a rate of 0.25mL per minute and washed with 100ml of heparin column buffer at a rate of 1ml per minute. Protein was eluted using a linear NaCl gradient (200mM-1M) with a 2.5%

increase in NaCl concentration per minute. Fractions of a similar purity were concentrated to approximately 4mg/ml (as determined by Bradford assay and NanoDrop) by 100kDa spin concentrators and rebuffered in 25mM Tris pH 7.5 containing 50% glycerol (**Fig. 4.3B and 4.3C**)

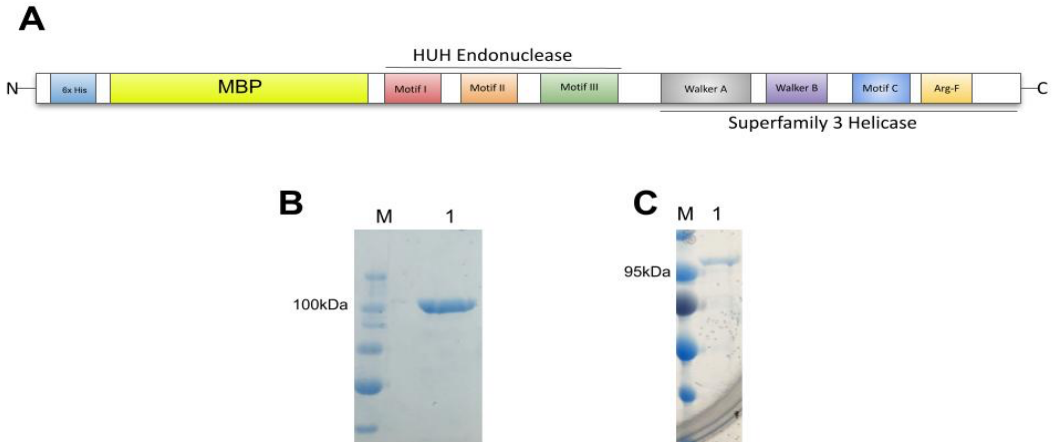

Figure 4.3. A) Construction of MBP-RepD1 and MBP-RepD2 fusion proteins. RepD1 and RepD2 (**Fig. 4.3**) were cloned as fusions to MBP to increase solubility and provide an additional purification tag. MBP and Rep are not to scale. **B)** Purification of MBP-RepD1 (lane 1). Protein was purified via amylose, Ni-NTA, and heparin affinity resins. **C)** Purification of MBP-RepD2 (lane 1) as described for MBP-RepD1. Molecular weight markers (M) are different between **B** and **C**.

Fusion Protein Cleavage: 150pmol of partially purified (Ni-NTA only) MBP-RepD1 and MBP-RepD2 were diluted in 25mM Tris. 30µl reactions containing approximately 150µM MBP-RepD1 or 15µg MBP-RepD2, 1µl of tobacco etch virus (TEV) protease (New England Biolabs) and a final TEV protease reaction buffer concentration of 1X were incubated for 1 hour at 30 °C and then allowed to continue overnight at 4 °C. While the fusion protein was cleaved (not shown), precipitates assumed to be RepD1 and RepD2 were observed in these reactions following the overnight incubation. Varying the amounts of fusion protein

substrate and the buffer had no impact on the presence of this precipitate. Due to this result all DNA based assays in this chapter were performed with the purified MBP-RepD1 or MBP-RepD2 fusion proteins.

Measurement of ATPase and GTPase Activities: To measure the release of free phosphate by RepD1 and RepD2 the QuantiChrom ATPase Assay Kit (BioAssay Systems) was used according to the manufacturer's instructions with a number of modifications. 1pmol purified RepD1 and RepD2 were resuspended in ATPase buffer (50mM Tris pH7.5, 100mM NaCl, 1mM DTT) which was supplemented with one of the following: 10mM $MgCl_2$, $MnCl_2$, $CaCl_2$, or $ZnCl_2$. ATP or GTP was added to a final concentration of 500µm and 40µl reactions were incubated for 30 minutes to three hours at room temperature unless otherwise indicated. After the specified time, 200µl of "Reagent" from the assay kit was added to the reactions and was allowed to incubate at room temperature for 30 minutes. Reactions were placed in microcuvettes and the OD620nm was read using a spectrophotometer. The kit-supplied phosphate standard was used to construct a standard curve (0-50µM free phosphate), and spectrophotometer readings were converted to phosphate concentrations as specified by the curve. For inhibition assays, RepD1 was treated for 30 minutes with 0.1, 1, or 10mM sodium orthovanadate or sodium azide prior to the addition of ATP substrate.

dsDNA Oligonucleotide Construction: ssDNA oligonucleotides containing the plus and minus strands of the putative BSL-RDHV ori (**Fig. 4.1** and **Table 4.2**) were purchased from Integrated DNA Technologies. To make dsDNA ori fragments, oligos representing the plus (5' 6-FAM labelled) and unlabeled minus

strand were mixed in an equimolar ratio and placed in a water bath set to 95 °C for 10 minutes. The power was turned off to the water bath and oligos were allowed to cool to room temperature in the water bath. Table 4.2 contains oligos used for binding, nicking, and joining assays in this study, while Figure 4.1 highlights their location in the BSL-RDHV genome and predicted stem loop. This same procedure was followed for generating dsDNA for nicking assays, with the exception that neither ssDNA oligonucleotide contained a 5' 6-FAM label.

dsDNA Oligonucleotide Binding and Electrophoretic Mobility Shift Assay:

To examine the potential of purified MBP-RepD1 and MBP-RepD2 to bind dsDNA carrying the predicted BSL-RDHV ori (dsRDHV-Ori) a binding protocol and electrophoretic mobility shift assay (EMSA) were developed. Binding reactions consisted of 2.5pmol of dsRDHV-Ori with a 2.5, 5, and 10 fold excess of purified MBP-RepD1 or MBP-RepD2 in a binding buffer consisting of 50mM TRIS pH 7.5, 100mM NaCl, 10mM $MgCl_2$, 5mM DTT, 5% glycerol, 1.5µg poly(dI dC). 30µl reactions were incubated at room temperature for 30 minutes. After 30 minutes reactions were mixed with gel loading buffer (25mM Tris, pH 7.5, 0.1% w/v bromophenol blue, 30% glycerol), and 5µl was loaded onto a 3.5% native polyacrylamide gel (0.5xTBE, 4% polyacrylamide, 10% glycerol) using gel loading tips. Prior to sample loading, wells were thoroughly flushed with 1x TBE and gels were pre-run in an ice water bath at 20V for approximately 2 hours to remove residual APS from the wells. Following the pre-run wells were again thoroughly flushed with 1x TBE. Electrophoresis was carried out using 1x TBE as running buffer and was allowed to proceed at 25V for approximately 8 hours (or

until the dye front was approximately 90% migrated to the bottom of the gel). Gels were then visualized using a Typhoon Imager (GE Life Sciences).

ssDNA Oligonucleotide Nicking: Nicking assays consisted of 75pmol of purified MBP-RepD1 and 16.6μM ssDNA oligonucleotides (**Table 4.2,** see page 123) in 50mM Tris pH 7.5, 100mM NaCl, 10mM divalent metal ions, 5mM DTT, 1mM EDTA, 5% glycerol. 30μl reactions were incubated at 37 °C for 2 hours and were subsequently resolved by 12% SDS-PAGE and Coomassie staining. Denaturing polyacrylamide gels consisted of 12% polyacrylamide, 7M urea, 1x TBE. Prior to electrophoresis reactions were digested with proteinase K overnight at 37 °C. Denaturing electrophoresis was carried out at 50 °C.

ssDNA Oligonucleotide Joining: Joining reactions consisted of 75pmol of purified MBP-RepD1, 16.6μM of ssRDHV-Ori, and either 32μM or 64μM of a single-stranded preformed acceptor oligonucleotide (ssAcceptor) (**Table 4.2**). Reactions were carried out and resolved as described for ssDNA nicking assays.

ssRNA Oligonucleotide Nicking: An RNA oligo containing the BSL-RDHV ori sequence was ordered from Integrated DNA Technologies. To test the ability of MBP-RepD1 to nick this oligo, an identical protocol to that which was employed for ssDNA nicking was carried out. 75pmol of purified MBP-RepD1 was incubated with 10, 20 or 40μM of ssRNA-Ori for 2 hours at 37 °C in 50mM TRIS pH 7.5, 100mM NaCl, 10mM $MgCl_2$, 5mM DTT, 1mM EDTA, 5% glycerol.

Protein Modelling: The 3D structure of BSL-RDHV Rep was predicted using SWISS-MODEL (54). The predicted structure was built using the crystal structure of Rep of PCV1 as a template (30). Models were visualized and edited using

PyMOL 2.1.4 (The PyMOL Molecular Graphics System, Schrödinger, LLC). The DNA binding surface was predicted using the APBS electrostatics plugin tool (55).

Statistical Analyses: Statistical analyses for ATP hydrolysis experiments (one-way ANOVA and Student's t-tests) and figure construction were performed in GraphPad Prism version 8.4.3 for MacOS, GraphPad Software, San Diego, California USA, (www.graphpad.com).

Results and Discussion

No RCR Products Detected In a Head to Tail BSL-RDHV Construct: Previous reports have shown that the insertion of tandem "head to tail" repeats of the CRESS virus genomes of porcine circovirus type 1 and 2 and diverse geminiviruses in bacterial plasmids results in the replicative release of viral genomes from the plasmid (56–59). This replicational, release apparently due to RCR, has been shown to be dependent on the presence of viral replication factors: an unmutated *Rep* gene coupled with two stem loops and associated nonanucleotide and iterons of correct sequence (57,58). In the case of geminiviruses this replicational release in *Agrobacterium tumefaciens* has been proposed to support the hypothesis that geminiviruses evolved from an ancestral prokaryotic plasmid (58) which has been subsequently supported by large scale sequence analyses for many members of the *Cressdnaviricota* (60,61). When a portion of the BSL-RDHV genome was cloned in a similar fashion (**Fig. 4.4A**), no RCR products or genome release were detected by agarose gel electrophoresis

following the extraction and XhoI digestion of plasmids from *E. coli* (**Fig. 4.4B**) Only one 6.9kb band was detected, indicating that only the complete pBluescript KS+-RDHV was replicated in *E. coli* and no RCR products had been released (**Fig. 4.4B**). If RCR products been present these products would have been 4.9kb linear DNA and 1.9kb circular DNA (**Fig 4.4A and Fig. 4.4B**).

A number of possibilities exist that may explain this result. Perhaps the

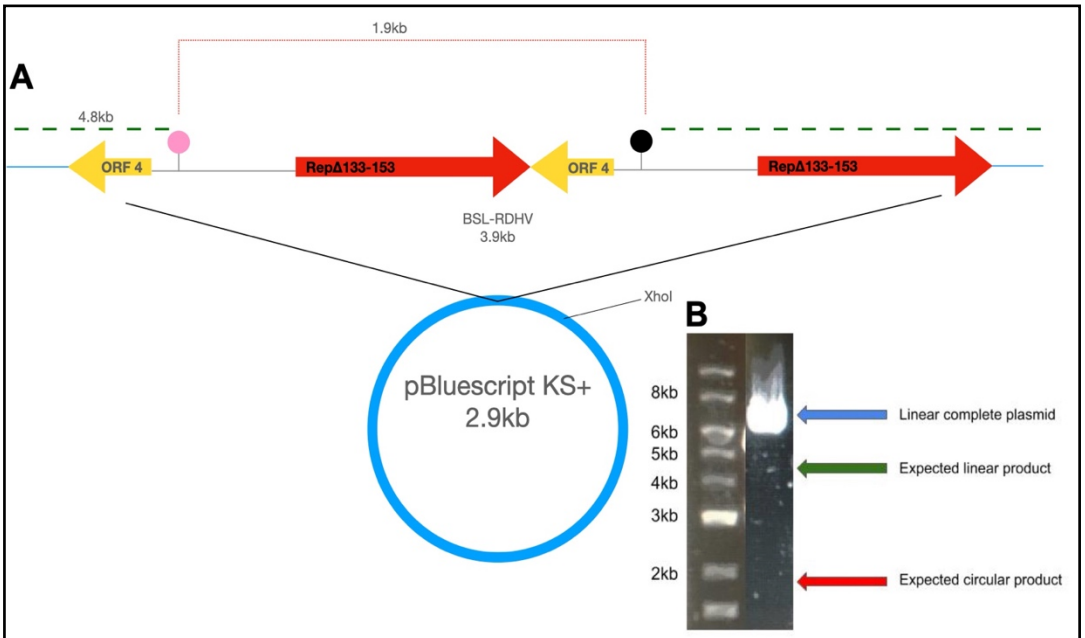

Figure 4.4. Construction of head to tail RDHV genome fragment in pBluescript KS+ (pBluescript KS+-RDHV). **A)** Highlighted in red-dashed line is the potential RCR product from initiation at the pink stem-loop and termination and the black stem-loop. Highlighted in green-dashed line is the potential RCR product resulting from initiation at the black-stem loop and termination at the pink stem-loop. This RCR product would also contain the complete pBluescript backbone. **B)** 0.7% agarose gel of *XhoI* digest of pBluescript KS+-RDHV. Only a band corresponding to the full length plasmid is present at 6.9kb. The expected size of RCR products is highlighted by arrows whose color corresponds to the expected RCR product highlighted in **A**.

genome portion that was cloned is insufficient to support RCR and other portions of BSL-RDHV are necessary for replication. This could be explored in future by inserting two complete BSL-RDHV genomes (again containing RepΔ133-153 to

remove the stop codon) head to tail in pBluescript KS+. Secondly, transcription and translation of RepΔ133-153 may not take place from pBluescript KS+-RDHV in *E. coli.* While no mechanism for the transcription and translation of circovirus or geminivirus Reps has been proposed in *E. coli* or *A. tumefaciens*, the presence of a complete *Rep* gene has been shown to be required to detect RCR products, suggesting that these processes are active (57,58). Finally, perhaps as was observed for overexpression of RepΔ133-153 in *E. coli,* the produced protein is simply insoluble and in turn inactive leading to no RCR products. This could be potentially be overcome by supplying an MBP fusion form of Rep or perhaps the codon optimized ORF. Whatever the case, more studies in this area could potentially demonstrate that replicational release takes place for BSL-RDHV.

Purified RepD1 and RepD2 Hydrolyze ATP and GTP: To examine the hydrolysis of NTPs a colorimetric assay was used. The free phosphate released was calculated as the excess free phosphate as compared to reactions incubated in the absence of necessary divalent metal ions. The release of free phosphate from ATP and GTP in the presence of 10mM $MgCl_2$ after 30 minutes at room temperature and 37 °C for RepD1 was found to be 12.32 +/- .74μM and 43.67 +/- 2.36μM, respectively (**Fig 4.5a**). This was significantly more ($p < .05$) than the phosphate released from the hydrolysis of GTP, 9.59 +/- 1.4μM and 39.35 +/- .82μM at room temperature and 37 °C respectively (**Fig 4.5a**). Similarly,

over the course of 3 hours more free phosphate was released from ATP hydrolysis than GTP hydrolysis (**Fig 4.5c**).

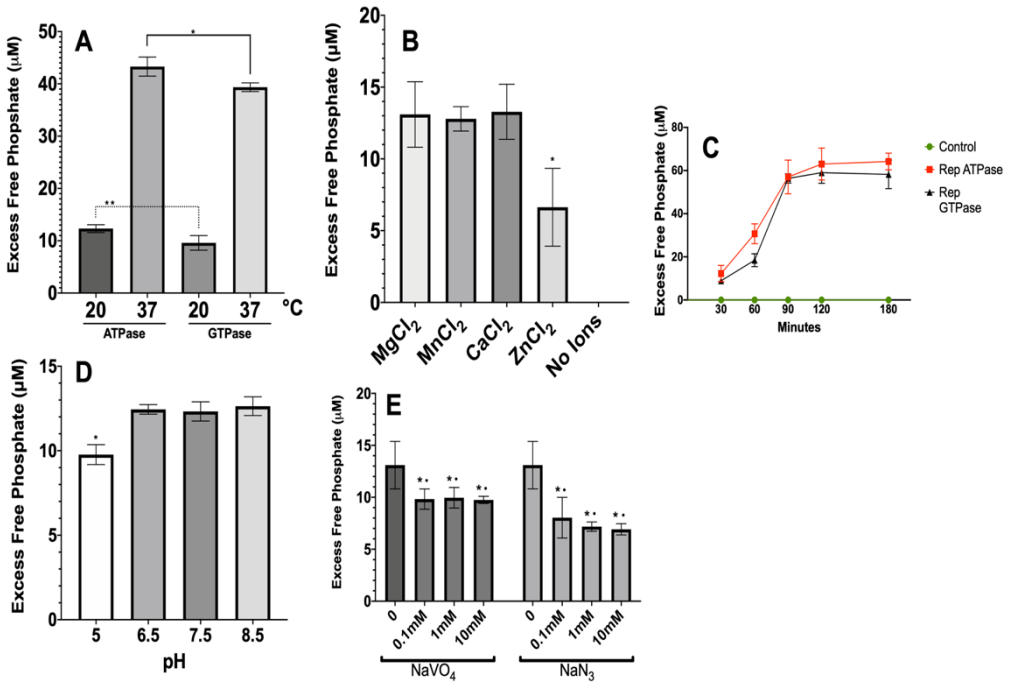

Figure 4.5. Phosphatase activity of RepD1. A) Excess free phosphate (µM) released following 30 minute incubation of 500µM ATP or 500µM GTP with 25nM RepD1 at 20 °C and 37 °C. **B)** Excess free phosphate (µM) released following 30 minute incubation of 500µM ATP with 25nM RepD1 at 20 °C in the presence of various divalent metal ions (10mM). **C)** Time course of excess free phosphate (µM) released at 20 °C from the incubation of 500µM ATP with 25nM RepD1. **D)** Excess free phosphate (µM) released at pH 5, 6.5, 7.5 and 8 following 30 minute incubation of 500µM ATP with 25nM RepD1 at 20 °C. **E)** Excess free phosphate (µM) released following 30 minute incubation of 500µM ATP with 25nM RepD1 at 20 °C following pre-treatment of RDHV-Rep with 0.1, 1, or 10mM sodium orthovandate or sodium azide. Single asterisks in panels **A, B,** and **D** represent significant differences ($p<.05$). Double asterisks in panel **A** represent very significant differences ($p<.01$). Asterisk and circles in panel **E** indicate significant differences ($p<.05$) between inhibitor and the control, as well as significant difference between the two inhibitors.

Previous reports have demonstrated than Rep of beak and feather disease virus (BFDV) displays a significant preference for $MgCl_2$ during in vitro

ATP hydrolysis experiments, but we were unable to detect statistically significant differences ($p<.05$) for different divalent metal ions, with the exception that the presence of $ZnCl_2$ led to a significant decrease ATPase activity exhibited by Rep (**Fig 4.5b**) (62). Similar $ZnCl_2$ inhibition has been demonstrated by both Rep of BFDV and Avian reovirus core protein μA (62,63). The only pH that resulted a significant decrease in release of excess free phosphate was pH 5 (**Fig 4.5d**).

RepD1 and RepD2 displayed no significant differences in ATP hydrolysis activities (GTP not examined) at room temperature (not shown). This was expected as motif II is not expected to play a direct role in NTP hydrolysis. There are some reports that deletion of motif II or specific amino acid substitutions within motif II can lead to increased ATPase activity possibly due to relieving conformational tensions, thus making the ATP binding site more accessible (30,32).

ATPases can be broadly classified based on their specific sensitivity to a variety of inhibitors (64). When RepD1 was preincubated with increasing concentrations of sodium azide or sodium orthovanadate before addition of ATP substrate there was a statistically significant ($p<.05$) decrease in release of free phosphate (**Fig 4.5e**). Of the two inhibitors, sodium azide led to significantly less free phosphate being released (**Fig 4.5e**). This would seem to suggest that Rep of BSL-RDHV is most similar to a V-Type ATPase (64). This is similar to what has been previously observed for Rep of BFDV (62).

The detection of ATPase activity supports the hypothesis that the C-terminal domain of Rep of BSL-RDHV functions as a helicase. Perhaps surprisingly, others have shown that the addition of both ss and dsDNA does not exert a stimulating effect on ATPase activity exhibited by Rep of other CRESS-DNA viruses, suggesting that helicase activity may be partially or completely supplied by host factor (62,65). Furthermore, direct helicase activity (DNA unwinding) has only been demonstrated for a small number of CRESS-DNA virus Reps (47,48). Future studies of the C-terminal domain of Rep of BSL-RDHV could look in more detail for direct helicase activity to confirm if the ATPase activity of the protein is truly indicative of DNA unwinding activity.

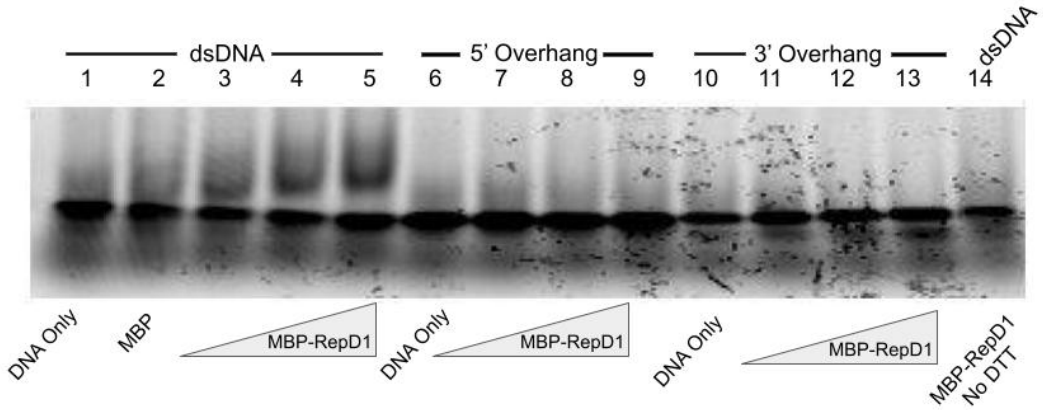

Figure 4.6. Binding of MBP-RepD1 to the putative ori of BSL-RDHV. 2.5pmol dsRDHV-Ori migrates freely (lane 1) as does dsRDHV-Ori incubated with 4µg MBP (lane 2). As the concentration of MBP-RepD1 is increased a shifted species becomes apparent (lanes 3-5). When MBP was incubated with 2.5pmol partially double stranded oligonucleotides carrying the BSL-RDHV ori sequence (pds3Over and pds5Over) no shifted species is apparent (lanes 7-9 and 11-13). When 5mM DTT was omitted from the binding buffer no shifted species is present following incubation of MBP-RepD1 with 2.5pmol dsRDHV-Ori (lane 14).

MBP-RepD1 and MBP-RepD2 Bind dsDNA: When MBP-RepD1 (**Fig. 4.3**) was

incubated with a double-stranded oligo containing the BSL-RDHV predicted stem loop and associated iterons (dsRDHV-Ori) (**Fig. 4.1** and **Table 4.2**) a shifted band was resolved by EMSA, indicating that MBP-RepD1 binds to dsRDHV-Ori (**Fig 4.6 lanes 3-5**). As the concentration of MBP-RepD1 was increased the shifted species becomes more apparent. MBP did not induce this same shift indicating that RepD1 is solely responsible for observed shift (**Fig 4.6 lane 2**). This binding of Rep near a stem-loop structure has been demonstrated to be the likely first step in RCR of various CRESS-DNA viruses (25,32,38,43,44,66–68). This binding is also implicated in transcriptional regulation of CRESS-DNA viruses as the nonanucleotide at the apex of the stem loop contains the TATA box involved in transcriptional regulation of virion sense DNA (36,39,69–71). MBP-RepD2 also induced a similar (more poorly resolved) band shift, indicating that the motif II mutation to 130 HLQGF 134 does not abolish dsDNA binding (**Fig. 4.7**). This result was expected as the RepD2 motif II sequence more similar to other motif II sequences in other CRESS-DNA viruses and does not contain

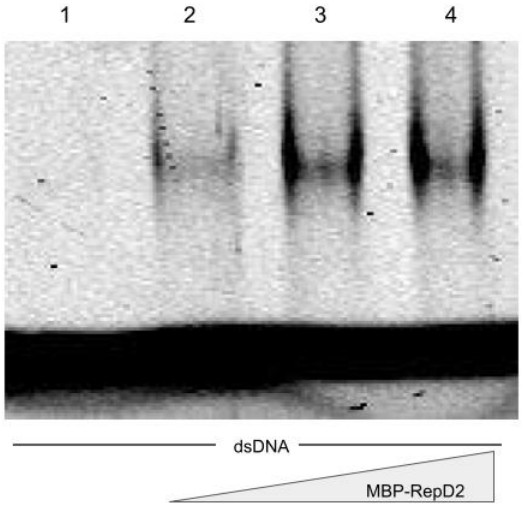

Figure 4.7. Binding of MBP-RepD2 to the putative BSL-RDHV ori. 2.5pmol dsRDHV-Ori migrates freely (lane 1). As the concentration of MBP-RepD2 increases (lanes 2-4) the intensity of the shifted species increases.

residues predicted by structural studies to be directly involved in dsDNA binding (2,30,72). The deletion of motif II of BFDV Rep has been previously shown to lead to an increase in dsDNA ori binding, while the introduction of an alanine residue (similar to RepD1) had no effect on dsDNA ori binding (32).

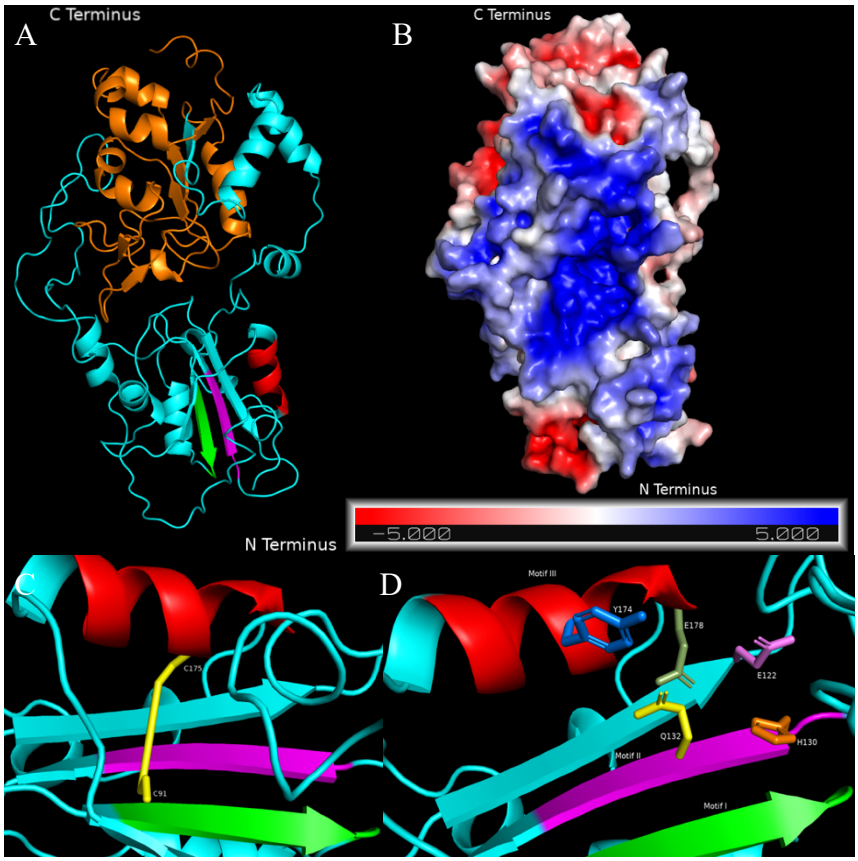

Figure 4.8. Models of Rep of BSL-RDHVA) Structure of Rep of BSL-RDHV. The N-terminal endonuclease domain is shown in cyan, and the C-terminal helicase domain is shown in orange. Motifs I, II, and III of the endonuclease domain are shown in green, magenta, and red, respectively. **B)** Predicted DNA binding surface (blue) results from basic amino acids present in both domains. **C**) Disulfide bond between C91 (motif I) and (C175 motif III) is highlighted in yellow. **D)** Residues of motif II and III that appear correctly positioned to mediate divalent metal ion coordination are highlighted.

When MBP-RepD1 was incubated with a partially dsDNA oligo containing a 3' or 5' overhang that resulted in single-stranded hexamer repeats (pds3Over and pds5Over) no band shift was observed, implying that MBP-RepD1 does not bind these oligos (**Fig 4.6 lanes 6-9 and 11-13**). This observation is consistent with previous observations that two double stranded repeats are required for Rep ori binding of other CRESS-DNA viruses (28,44,68). The binding of MBP-RepD1 appears to be dependent on the inclusion of 5mM dithiothreitol (DTT) in the binding buffer. When DTT was excluded from the binding buffer no band shift of dsRHVOri was observed (**Fig 4.6 lane 14**). Given the presence of two cysteine residues predicted to form a disulfide bridge in motif I (C91) and motif III (C175), this DTT inclusion requirement suggests that the C91-C175 disulfide bond may alter the structure of the dsDNA binding surface (**Fig. 4.8b and 4.8c**).

MBP-RepD1 Becomes Covalently Attached to ssDNA Containing the Putative BSL-RDHV Stem Loop: Following incubation of MBP-RepD1 with a 54nt single stranded oligo containing the predicted BSL ori (ssRDHV-Ori) (**Fig. 4-1** and **Table 4-2**) two bands were observed on an SDS-polyacrylamide gel, one at the predicted size of 100kDa, and a second shifted band at approximately 110kDa (**Fig. 4-9**). The presence of a shifted band implies that MBp-RepD1 has been covalently attached to the newly generated 5' end following a ssDNA nicking event, characteristic of the initiation of RCR in other CRESS-DNA viruses (30,32,49). When a single-stranded oligo of the same sequence but containing a 5' 6-FAM label (ssRDHV-Ori*) was incubated with Rep-D1 little to no shift was apparent, implying that the 5' fluorophore may partially interfere with this nicking

activity (**Fig 4.9**). This may explain why resolving nicked ssDNA using denaturing polyacrylamide gels yielded inconclusive results (not shown).

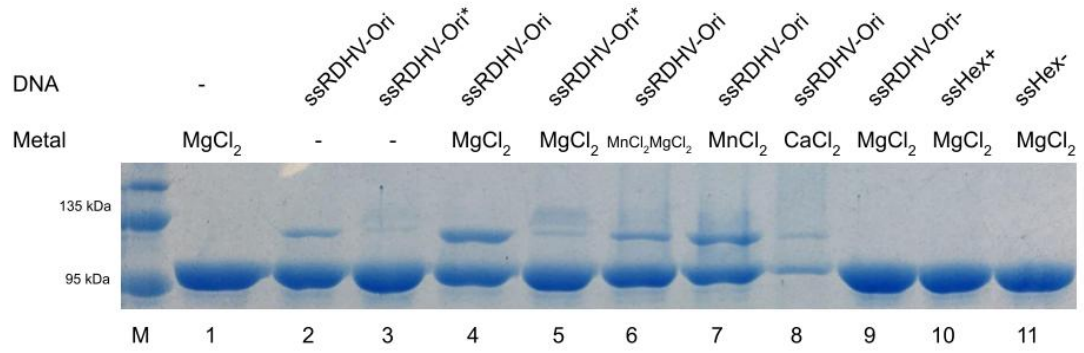

Figure 4.9. SDS-PAGE indicating MBP-RepD1 covalently attaches to a single stranded oligonucleotide carrying the predicted BSL-RDHV ori sequence (ssRDHV-Ori). MBP-RepD1 (lane 1, 100kDa) was incubated with 16.6μM ssRDHV-Ori in the absence of divalent metal ions (lane 2) a shifted band (approximately 110kDa) emerges. When MBP-RepD1 was incubated with 16.6μM ssRDHV-Ori* the intensity of the shifted band is reduced (lane 3). Lanes 4-8 demonstrate the effect of different divalent metal ions (10mM) on the intensity of the shifted band. No shifted band is apparent following MBP-RepD1 incubation with ssRDHV-Ori-, ssHex+, or ssHex- (lanes 9-11).

When the nicking buffer was supplemented with $MgCl_2$ the intensity of the shifted band increased (**Fig 4.9**). This apparent stimulation of nicking activity by $MgCl_2$ can likely be explained by the presence of positively charged metal ions overcoming the effect of unfavorable interactions between MBP-RepD1 and ssDNA due to negatively charged side chains present in Rep (73,74). TrwC (bacterial relaxase), NS1 of minute virus of mice, and U94 of human herpesvirus are all members of the HUH endonuclease family that have demonstrated increased nicking activity in the presence of divalent metal ions, but not an apparent strict requirement for their inclusion in in vitro nicking buffers, as is expected for DNA nicking enzymes employing a catalytic tyrosine residue (74–

77). Despite containing an active site tyrosine in motif III, there have been published reports of Reps of CRESS-DNA viruses strictly requiring the inclusion of metal ions in nicking buffers to observe activity (30,49,74). Rep of BFDV, whose motif II contains the alternative HLQ sequence, seems to not be dependent on the inclusion of metal ions for nicking activity (32). However, the deletion of BFDV Rep motif II was shown to completely abolish ssDNA nicking, but not dsDNA binding by Rep (32). Furthermore, mutation of motif II of BFDV Rep from 51 HLQGY 55 to 51 HLQGA 55 does result in an apparent decrease in nicking activity (32). Given that the motif II sequence of MBP-RepD1 is 130 HLQAY 134, the observed covalent attachment/nicking activity in this present study may be less than optimal due to decreased metal affinity compared to the unknown actual BSL-RDHV Rep motif II sequence. Until a host for BSL-RDHV is developed or discovered this actual WT sequence of motif II seems difficult to ascertain. However, it does seem that Q132 in motif II of BSL-RDHV RepD1/D2 would not participate directly in metal ion coordination normally mediated by His side chains, and it would be predicted to maintain ssDNA nicking activity (30).

To evaluate the potential effect different divalent metal ions have on the nicking activity of MBP-RepD1 nicking buffer was supplemented with $MnCl_2$, $CaCl_2$, and a combination of $MgCl_2$/$MnCl_2$ in individual nicking assays. Based on the intensity of the shifted band (**Fig 4.9**) it appears that MBP-RepD1 exhibits the most nicking activity in the presence of $MgCl_2$ > $MnCl_2$ > $MgCl_2$/$MnCl_2$. $CaCl_2$ inclusion consistently resulted in fainter shifted and unshifted bands for reasons

that are not clear. A previous assay using Rep of porcine circovirus type 2 (PCV2) in the presence of $CaCl_2$ also resulted in a similarly distorted band (30). Rep of PCV1 and BFDV exhibit an in vitro preference for $MnCl_2$, but both still exhibit nicking activity in the presence of the more biologically available Mg^{2+} (30,32).

Partial Localization of Nick Site To more accurately localize the nick and covalent attachment site for BSL-RDHV Rep a number of single-stranded oligonucleotides were incubated with purified MBP-RepD1. A single-stranded oligonucleotide containing the predicted minus-sense strand ori sequence ssRDHV-Ori(-) did not generate a shift, indicating that the predicted plus-sense strand has been correctly identified and that there is no nick site present in the minus-sense strand (**Fig 4.9**). Additionally, oligos containing only the plus-sense or minus-sense hexamer repeats (ssRDHV-Hex+ and ssRDHV-Hex- respectively) did not generate a shift implying that they also do not contain a nick site (**Fig 4.9**). Along with previous work on CRESS-DNA virus nick sites, these data tentatively confirm that the nick site for the BSL-RDHV putative ori is located in the loop portion of the stem loop (32,40,41,66). These data alone do not exclude the nick site being in the 5' portion of the stem structure, but sequence similarity favors the nick site being located in the 5' AAGTATT/AC 3' nonanucleotide loop (where / represents the predicted nick site).

ssDNA Sequence and Structure Requirements For Nicking: In order to investigate the sequence requirements for nicking, the same shift assay was performed with a single-stranded oligo (ssPCV1) containing the nonanucleotide sequence 5'- TAGTATTAC-3' (the BSL-RDHV nonanucleotide sequence is 5'-AAGTATTAC-3'). This oligo contains a BSL-RDHV stem sequence and associated iterons, but the first position of the nonanucleotide has been substituted (underlined above) resulting in the nonanucleotide found in PCV1(78). After incubation with ssPCV1 a shifted MBP-RepD1 band becomes apparent via SDS-PAGE, suggesting that MBP-RepD1 is capable of nicking and covalently attaching to an oligo carrying a mutation in position 1 of the nonanucleotide (**Fig 4.10**). This data is in agreement with previous in vitro work in other CRESS-DNA viruses using purified recombinant Rep that has demonstrated that mutagenesis of the first nucleotide of the nonanucleotide does not abolish nicking activity (41).

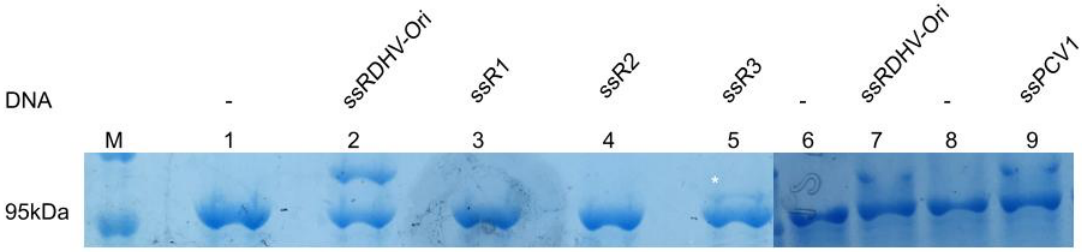

Figure 4.10. SDS-PAGE indicating MBP-RepD1 covalently attaches to a variety of ssDNA substrates. When MBP-RepD1 (lanes 1, 6, and 8 100kDa) was incubated with 16.6μM ssRDHV-Ori (lane 2) a shifted band (approximately 110kDa) emerges. No shifted band is apparent following incubation of MBP-RepD1 with ssR1 and ssR2 (lanes 3 and 4). A shifted band emerges (white asterisk) following the incubation of MBP-RepD1 with ssR3. A shifted band also emerges following incubation of MBP-RepD1 with ssPCV1.

Previous work in cell culture revealed that PCV1 is tolerant of mutations in the nonanucleotide loop as well as those nucleotides directly upstream of the conserved nonanucleotide and that viruses carrying these mutations still produce viral progeny (79). However, reversion to WT nonanucleotide sequences is generally observed after a small number of passages in cell culture (79). These observations led to testing of how mutable the nucleotide positions in and surrounding the 5' AAGTATT/AC 3' predicted nonanucleotide and nick site in BSL-RDHV are. Three oligos, ssR1, ssR2, and ssR3 (**Table 4.2**), available in the Stedman Lab oligo library contained the potential minimal nicking sequence of 5'-TATTAC-3', previously identified in other CRESS-DNA viruses (32,41). The potential minimal nick site in these oligos are flanked by nucleotides that are irrelevant to this study.

After incubation with MBP-RepD1, ssR1 and ssR2 did not generate a band shift, indicative of a lack of nicking and covalent attachment of MBP-RepD1 (**Fig 4.10, lanes 3 and 4**). However, ssR3 does generate a faint band shift, indicating that this oligo is capable of being nicked (**Fig 4.10, lane 5**). While the shifted band is faint, it does appear to be correctly positioned at approximately half the shift observed for ssRDHV-Ori, due to the predicted covalent attachment of 14nt as compared to a 28nt attachment in the case of ssRDHV-Ori. In the ssR3 oligo the first and third nucleotides of the BSL-RDHV nonanucleotide (5' **A**A**G** 3') have been substituted for 5' **G**A**T** 3'. These data suggest that positions 1 and 3 within the BSL-RDHV nonanucleotide are tolerant of at least some

nucleotide substitutions and do not abolish nicking activity. However, the relative intensity of this shifted band being less than that of the positive control may suggest that this oligo is not nicked as effectively as one carrying a wt ori. Taken together the ssR3 and ssPCV1 data (from above) further suggest that nucleotides 1 and 3 within the nonanucleotide are tolerant of at least some mutations while position 2 (mutated in ssR1 and ssR2 but not ssR3 or ssPCV1) appears to be intolerant of an A to C and A to T substitution. More work is needed to fully confirm this tentative conclusion. However, this does echo previous work on PCV1 that demonstrated in cell culture the mutable positions of the nonanucleotide that were capable of generating progeny virus were positions 1 and 3 (5'-xAxTATTAC-3' in which x represents positions tolerant of mutation) (79). Additionally, the predicted nick site of ssR3 is located in a predicted stem structure, while the predicted nick site of BSL-RDHV and other CRESS-DNA viruses is at the loop apex (**Fig. 4.1**). The interpretation that a loop is not required for nicking by Rep of BSL-RDHV is supported by previous work which demonstrated no stem-loop structure is required by Rep of PCV1 or Rep of BFDV for detection of in vitro nicking activity (32,80). ssR3 results also suggest that the presence of at least some irrelevant nucleotides outside the conserved minimal nick site (5'-TATT/AC-3') do not completely abolish nicking activity. The presence of "non-interfering" nucleotides coupled with nucleotide substitution tolerance in and surrounding the nonanucleotide may help to explain why CRESS-DNA virus ori regions appear to be recombination hotspots, as well as

why some CRESS-DNA viruses contain replication factors that appear to be interchangeable (81–83).

The covalent attachment of MBP-RepD1 to ssR3 suggests that the iterons downstream of the stem loop structure are not required for nicking activity (**Fig 4.10**). However, our dsDNA binding results from above indicate that double-stranded iterons are required for MBP-RepD1 binding to dsDNA. Thus, it appears that the minimal binding site for dsDNA, consisting of the double-stranded stem-loop and associated double-stranded hexamer repeats, and the minimal nicking site for ssDNA, consisting of the final six nucleotides of the nonanucleotide, are different. Together these data suggest that the initial binding of BSL-RDHV Rep to the ori is likely the more stringently controlled event in the initiation of RCR.

MBP-RepD1 Exhibits Joining Activity: MBP-RepD1 was used in a band shift assay that included the ssRDHV-Ori and an increasing amount of a preformed acceptor oligonucleotide (ssAcceptor). The acceptor oligo contains the 5’ portion of the BSL-RDHV stem loop as it would appear after a nicking reaction and should serve as a suitable substrate for a joining reaction. As the concentration of ssAcceptor was increased the intensity of the unshifted MBP-RepD1 band increased while the shifted band intensity decreased (**Fig. 4.11**). This result suggests that the addition of the preformed acceptor results in the release of MBP-RepD1 form the covalent protein-ssRDHV-Ori adducts formed during a nicking reaction. This release can be explained by MBP-RepD1 mediating a joining reaction between the ssAcceptor oligonucleotide and the ssRDHV-Ori.

This nucleotidyl transferase reaction is characteristic of the completion of RCR in other CRESS-DNA viruses (32,38,41,84).

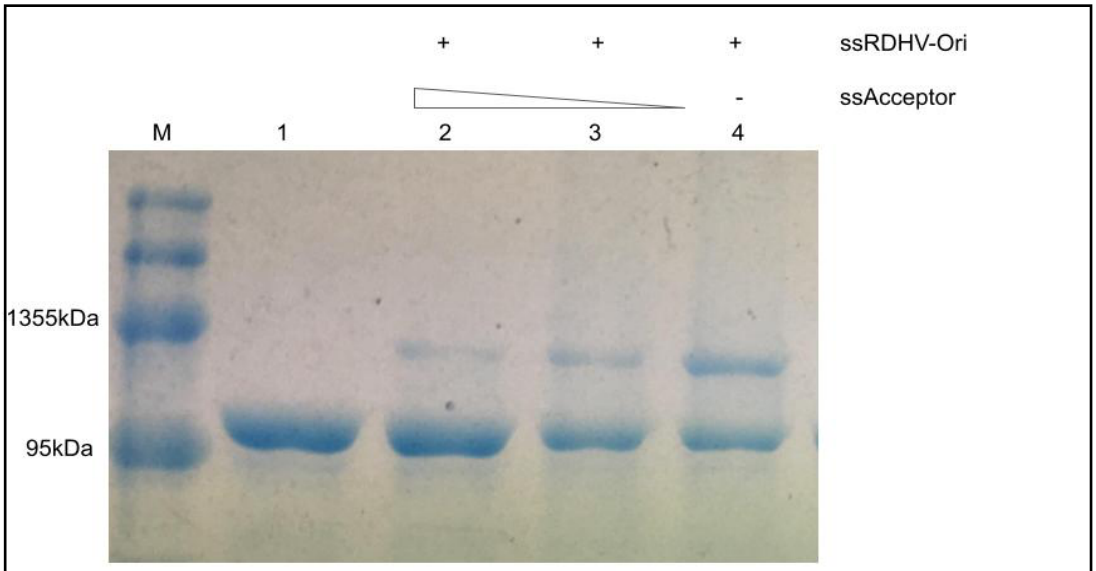

Figure 4.11. MBP-RepD1 exhibits joining activity in vitro. When MBP-RepD1 (lanes 1, 100kDa) was incubated with 16.6μM ssRDHV-Ori (lane 4) and an increasing concentration (32μM and 64μM) of a preformed acceptor oligonucleotide (ssAcceptor) the intensity of the unshifted band increased, while the intensity of the shifted band decreased (lanes 2 and 3).

MBP-RepD1 Does Not Appear to Nick ssRNA: A number of hypotheses regarding the recombination event that led to the emergence of cruciviruses involve Rep mediating an RNA-DNA recombination event (17). To examine the potential biochemical activity Rep exhibits on RNA, MBP-RepD1 was incubated with an increasing concentration of an RNA oligo containing the BSL-RDHV ori sequence (ssRNA-Ori). After incubation, there was only one band present on an SDS-PAGE gel at the predicted size of unbound MBp-RepD1 indicating that no nicking reaction had taken place (**Figure 4.12**).

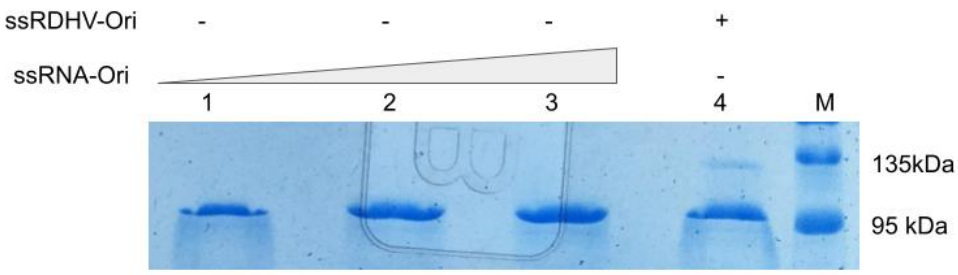

Figure 4.12. MBP-RepD1 does not covalently attach to a ssRNA oligonucleotide. When MBP-RepD1 was incubated with 10, 20 or 40μM of ssRNA-Ori (lane1-4) no shifted band is apparent. Lane 4 shows the characteristic shifted band following incubation of 7.5μg of MBP-RepD1 with 16.6μM ssRDHV-Ori.

The lack of detectable RNA nicking activity does not preclude Rep from being involved in RNA-DNA recombination. Previous work with PCV1 has demonstrated that there is a requirement for secondary structure in ssDNA in order for Rep to mediate a joining reaction, but no strict requirement for what DNA sequence gives rise to the structure (80). Perhaps MBP-RepD1 is not capable of nicking RNA but would be capable of joining ssDNA-RNA if the RNA contains an appropriate acceptor site and flanking secondary structure. This could be tested in future experiments using a pre-formed acceptor ssRNA oligo similar to what was described above for ssDNA joining reactions.

A number of ligases are known to be capable of RNA-DNA ligation reactions, T4 DNA ligase can carry out DNA-RNA ligations under prolonged incubations at elevated temperatures, while supplementing the reaction with less ATP than usual (85,86). *Paramecium bursaria* chlorella virus DNA ligase has been shown to exhibit much more efficient DNA-DNA ligation activities (as

compared to T4 DNA ligase) utilizing RNA-splinted DNA substrates when reactions are incubated under low salt conditions in the presence of Mn^{2+} (87). Similar “non-optimal” conditions, differing from those developed for the ssDNA assays described in this chapter, may be helpful in detecting RNA nicking, or RNA-DNA joining by Rep in vitro.

Name	Sequence 5'-3'	Use
pET21b_F	TCCGAATTCGAGCTCCG TC	Linearizes pET21b for Gibson Cloning.
pET21b_R	ATGTATATCTCCTTCTTA AAGTTAAAC	Linearizes pET21b for Gibson Cloning.
Opt_Rep_21b_F	TTTGTTTAACTTTAAGAA GGAGATATACATATGCA CCATCATCACCAC	Amplifies codon optimized BSL-RDHV Rep, contains pET21b overlap
Opt_Rep_21b_R	GCAAGCTTGTCGACGGA GCTCGAATTCGGATTAG CTAATCACAAGGGTGTT	Amplifies codon optimized BSL-RDHV Rep, contains pET21b overlap
MBP_F	TTTGTTTAACTTTAAGAA GGAGATATACATATGGG T TCTTCTCACCATC	Amplifies MBP from pLIC-HMK for Rep fusion construction, contains pET21b overlap.
MBP_R	GTGGTGATGATGGTGTC CACTTCCAA ATTGGA	Amplifies MBP from pLIC-HMK for Rep fusion construction, contains optimized Rep overlap.
OR_OLMBP_F	AATATTGGAAGTGGACA CCATCATCACCACCATT C	Amplifies codon optimized BSL-RDHV Rep, contains MBP overlap.
OR_OL21b_R	GCAAGCTTGTCGACGGA GCTCGAATTCGGATTAG CTAATCACAAGGGTG	Amplifies codon optimized BSL-RDHV Rep, contains pET21b overlap.

Table 4.1. PCR primers used in this study.

Name	Sequence 5’-3’	Use	Note
ssRDHV-Ori*	*GTGACCAAAAGTTGTAGCTAAGTATTACCTACAACTTTTGGTCACTTGTGGTCA	Annealed to ssRDHV-Ori- and ssRDHVNoHex- for EMSA. Annealed to ssRDHV-Ori* for EMSA. Used as single stranded for nicking assays.	Predicted plus strand of BSL-RDHV ori with associated hexamer repeats. Contains a 5’ FAM label (*)
ssRDHV-Ori-	TGACCACAAGTGACCAAAAGTTGTAGGTAATACTTAGCTACAACTTTTGGTCAC	Annealed to ssRDHV-Ori* and ssRDHV-OriNoHex for EMSA. Used as single stranded for ssDNA nicking assays.	Predicted minus strand of BSL-RDHV ori with associated hexamer repeats.
ssRDHV-OriNoHex*	*GTGACCAAAAGTTGTAGCTAAGTATTACCTACAACTTT	Annealed to ssRDHV-Ori- for EMSA.	Predicted plus strand of BSL-RDHV ori lacking associated hexamer repeats. Contains a 5’ FAM label.
ssRDHV-Ori-NoHex	AAAGTTGTAGGTAATACTTAGCTACAACTTTTGGTCAC	Annealed to ssRDHV-Ori* for EMSA.	Predicted minus strand of BSL-RDHV ori lacking associated hexamer repeats.
ssRDHV-Ori	GTGACCAAAAGTTGTAGCTAAGTATTACCTACAACTTTTGGTCACTTGTGGTCA	Used in ssDNA nicking assays.	Predicted plus strand of BSL-RDHV ori with associated hexamer repeats.
ssHex+	TGGTCACTTGTGGTCA	Used in ssDNA nicking assays.	Contains predicted plus strand hexamer repeats.
ssHex-	TGACCACAAGTGACCA	Used in ssDNA nicking assays.	Contains predicted minus strand hexamer repeats.

Table 4.2. Oligonucleotides used for dsDNA binding and ssDNA attachment assays in this study. See Figure 4.1 for positions in BSL-RDHV genome.

Name	Sequence 5’-3’	Use	Note
ssR1	TGGTG**TATTAC**AGTCAATAACTG	Used in ssDNA nicking assays.	Predicted minimal nick site bolded.
ssr2	CTCACCC**TATTAC**TGATACGCTAC	Used in ssDNA nicking assays.	Predicted minimal nick site bolded.
ssR3	CAAGCTTGTCGGACGGAGCTCGAATTCGGAT**TATTAC**GGGCATGTAATG	Used in ssDNA nicking assays.	Predicted minimal nick site bolded.
ssPCV1	GTGACCAAAAGTTGTAGCTTAGTATTACCTACAACTTTTGGTCACTTGTGGTCA	Used in ssDNA nicking assays.	First position of nonanucleotide A to T (underlined) substitution.
ssRNA-Ori	GUGACCAAAAGUUGUAGCUAAGTAUUACCUACAACUUUUGGTCACUUGUGGUCA	Used in ssDNA nicking assays.	

Table 4.2 Continued. Oligonucleotides used for dsDNA binding and ssDNA attachment assays in this study. Predicted minimal nick site is shown in bold.

References

1. Diemer GS, Stedman KM. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. Biol Direct. 2012;7(13):1–14.
2. de la Higuera I, Kasun GW, Torrance EL, Pratt AA, Maluenda A, Colombet J, et al. Unveiling crucivirus diversity by mining metagenomic data. mBio. 2020;11(5):1–17.
3. Bistolas K, Besemer R, Rudstam L, Hewson I. Distribution and inferred evolutionary characteristics of a chimeric ssDNA virus associated with intertidal marine isopods. Viruses. 2017 Nov 26;9(12):361
4. Quaiser A, Krupovic M, Dufresne A, Roux S. Diversity and comparative genomics of chimeric viruses in Sphagnum dominated peatlands. Virus Evol. 2016;2(2):1–8.
5. Salmier A, Tirera S, De Thoisy B, Franc A, Darcissac E, Donato D, et al. Virome analysis of two sympatric bat species (*Desmodus rotundus* and *Molossus molossus*) in French Guiana. PLoS One. 2017;12(11):1–25.
6. Kraberger S, Argüello-Astorga GR, Greenfield LG, Galilee C, Law D, Martin DP, et al. Characterisation of a diverse range of circular replication-associated protein encoding DNA viruses recovered from a sewage treatment oxidation pond. Infect Genet Evol. 2015;31:73–86.
7. Tisza MJ, Pastrana D V., Welch NL, Stewart B, Peretti A, Starrett GJ, et al. Discovery of several thousand highly diverse circular DNA viruses. eLife. 2020;9:1–26.
8. de la Higuera I, Torrance EL, Pratt AP, Kasun GW, Maluenda A, Stedman KM. Genome sequences of three cruciviruses Found in the Willamette Valley (Oregon). Microbiol Resour Announc. 2019;(June):18–20.
9. Rosario K, Dayaram A, Marinov M, Ware J, Kraberger S, Stainton D, et al. Diverse circular ssDNA viruses discovered in dragonflies (*Odonata: Epiprocta*). J Gen Virol. 2012;93(Pt 12):2668-2681.
10. Roux S, Enault F, Bronner G, Vaulot D, Forterre P, Krupovic M. Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses. Nat Commun. 2013;4:2700.
11. Krupovic M, Zhi N, Li J, Hu G, Koonin E V., Wong S, et al. Multiple layers of chimerism in a single-stranded DNA virus discovered by deep sequencing. Genome Biol Evol. 2015;7(4):993–1001.
12. Hewson I, Ng G, Li WF, LaBarre BA, Aguirre I, Barbosa JG, et al. Metagenomic identification, seasonal dynamics, and potential transmission

mechanisms of a Daphnia-associated single-stranded DNA virus in two temperate lakes. Limnol Oceanogr. 2013;58(5):1605–20.

13. Steel O, Kraberger S, Sikorski A, Young LM, Catchpole RJ, Stevens AJ, et al. Circular replication-associated protein encoding DNA viruses identified in the faecal matter of various animals in New Zealand. Infect Genet Evol. 2016;43:151–64.

14. Mcdaniel LD, Rosario K, Breitbart M, Paul JH. Comparative metagenomics: Natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. Environ Microbiol. 2014 Feb;16(2):570-85.

15. Dayaram A, Galatowitsch ML, Argüello-Astorga GR, van Bysterveldt K, Kraberger S, Stainton D, et al. Diverse circular replication-associated protein encoding viruses circulating in invertebrates within a lake ecosystem. Infect Genet Evol. 2016;39:304–16.

16. Rosario K, Duffy S, Breitbart M. A field guide to eukaryotic circular single-stranded DNA viruses: Insights gained from metagenomics. Arch Virol. 2012;157(10):1851–71.

17. Stedman K. Mechanisms for RNA capture by ssDNA viruses: Grand theft RNA. J Mol Evol. 2013;76(6):359–64.

18. Gibbs MJ, Weiller GF. Evidence that a plant virus switched hosts to infect a vertebrate and then recombined with a vertebrate-infecting virus. Proc Natl Acad Sci U S A. 1999;96(14):8022–7.

19. Krupovic M, Zhi N, Li J, Hu G, Koonin E V, Wong S, et al. Multiple layers of chimerism in a single-stranded DNA virus discovered by deep sequencing. 2015;7(4):993–1001.

20. Varsani A, Lefeuvre P, Roumagnac P, Martin D. Notes on recombination and reassortment in multipartite/segmented viruses. Curr Opin Virol. 2018;33(September):156–66.

21. Krupovic M, Varsani A, Kazlauskas D, Breitbart M, Delwart E, Yutin N, et al. Cressdnaviricota: a virus phylum unifying 7 families of rep-encoding viruses with single- stranded, circular DNA genomes. J Virol. 2020 Jun 1;94(12):e00582-20.

22. Chandler M, Cruz F De, Dyda F, Hickman AB. Breaking and joining single-stranded DNA : the HUH endonuclease superfamily. Nat Rev Microbiol. 2013;11(8):525–38.

23. Ilyina T V, Koonin E V. Conserved sequence motifs in the initiator proteins for rolling circle DNA-replication encoded by diverse replicons from Eubacteria, Eukaryotes and Archaebacteria. Nucleic Acids Res. 1992;20(13):3279–85.

24. Koonin E V, Ilyina T V. Computer-assisted dissection of rolling circle DNA replication. Vol. 30, BioSystems. 1993;30(1-3):241-68.

25. Timchenko T, de Kouchkovsky F, Katul L, David C, Vetten HJ, Gronenborn B. A single Rep protein initiates replication of multiple genome components of faba bean necrotic yellows virus, a single-stranded DNA virus of plants. J Virol. 1999;73(12):10173–82.

26. Laufs J, Schumacher S, Geisler N, Jupin I, Gronenborn B. Identification of the nicking tyrosine of geminivirus Rep protein. FEBS Lett. 1995;377(2):258–62.

27. Londoño A, Riego-Ruiz L, Argüello-Astorga GR. DNA-binding specificity determinants of replication proteins encoded by eukaryotic ssDNA viruses are adjacent to widely separated RCR conserved motifs. Arch Virol. 2010;155(7):1033–46.

28. Xu X, Qian Y, Wang Y, Li Z, Zhou X. Iterons homologous to helper geminiviruses are essential for efficient replication of betasatellites. J Virol. 2018;93(5):1–22.

29. Ruiz-medrano R, Irapuato U. An iteron-related domain is associated to Motif 1 in the replication proteins of geminiviruses: identification of potential interacting amino acid-base pairs by a comparative approach. Arch Virol. 2001;1465–85.

30. Vega-Rocha S, Byeon IJL, Gronenborn B, Gronenborn AM, Campos-Olivas R. Solution structure, divalent metal and DNA binding of the endonuclease domain from the replication initiation protein from porcine circovirus 2. J Mol Biol. 2007;367(2):473–87.

31. Luo G, Zhu X, Lv Y, Lv B, Fang J, Cao S, et al. Crystal structure of the dimerized N-terminus of porcine circovirus type 2 replicase protein reveals a novel antiviral interface. J Virol. 2018;92(18):JVI.00724-18.

32. Kai J, Chiaolong C, Jian H, Wu S, Yi S, Chi L, et al. Characterization of the endonuclease activity of the replication - associated protein of beak and feather disease virus. Arch Virol. 2019;Aug;164(8):2091-2106.

33. Nash TE, Dallas MB, Reyes MI, Buhrman GK, Ascencio-Ibanez JT, Hanley-Bowdoin L. Functional analysis of a novel motif conserved across geminivirus Rep proteins. J Virol. 2011;85(3):1182–92.

34. Gorbalenya AE, Koonin E V., Wolf YI. A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses. FEBS Lett. 1990;262(1):145–8.

35. Koonin E V. A common set of conserved motifs in a vast variety of putative nucleic acid-dependent ATPases including MCM proteins involved in the

initiation of eukaryotic DNA replication. Nucleic Acids Res. 1993;21(11):2541–7.

36. Zhao L, Rosario K, Breitbart M, Duffy S. Eukaryotic circular Rep-encoding single-stranded DNA (CRESS-DNA) viruses: ubiquitous viruses with small genomes and a diverse host range. 1st ed. Advances in Virus Research. Elsevier Inc.; 2018. 1–63 p.

37. Faurez F, Dory D, Grasland B, Jestin A. Replication of porcine circoviruses. Virol J. 2009;6:1–8.

38. Gronenborn B. Nanoviruses: Genome organisation and protein function. Vet Microbiol. 2004;98(2):103–9.

39. Mankertz A, Çaliskan R, Hattermann K, Hillenbrand B, Kurzendoerfer P, Mueller B, et al. Molecular biology of Porcine circovirus: Analyses of gene expression and viral replication. Vet Microbiol. 2004;98(2):81–8.

40. Hafner GJ, Stafford MR, Wolter LC, Harding RM, Dale JL. Nicking and joining activity of banana bunchy top virus replication protein in vitro. J Gen Virol. 1997 Jul;78 (Pt 7):1795-9.

41. Steinfeldt T, Finsterbusch T, Mankertz A. 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. 2006;80(13):6225–34.

42. Cheung AK. Detection of template strand switching during initiation and termination of dna replication of porcine circovirus. J Virol. 2004;78(8):4268–77.

43. Orozco BM, Hanley-Bowdoin L. A DNA structure is required for geminivirus replication origin function. J Virol. 1996;70(1):148–58.

44. Steinfeldt T, Finsterbusch T, Mankertz A. Rep and Rep‘ protein of porcine circovirus type 1 bind to the origin of replication in vitro. Virology. 2001;291(1):152–60.

45. Stanley J. Analysis of African cassava mosaic virus recombinants suggests strand nicking occurs within the conserved nonanucleotide motif during the initiation of rolling circle DNA replication. Virology. 1995;206(1):707–12.

46. Hickman AB, Dyda F. Binding and unwinding: SF3 viral helicases. Current Opinion in Structural Biology. 2005 Feb;15(1):77-85.

47. Clérot D, Bernardi F. DNA helicase activity is associated with the replication initiator protein Rep of tomato yellow leaf curl geminivirus. J Virol. 2006;80(22):11322–30.

48. Choudhury NR, Malik PS, Singh DK, Islam MN, Kaliappan K, Mukherjee SK. The oligomeric Rep protein of Mungbean yellow mosaic India virus

(MYMIV) is a likely replicative helicase. Nucleic Acids Res. 2006;34(21):6362–77.

49. Laufs J, Trautt W, Heyraudt F, Matzeitt V, Rogersf SG, Schellt J, et al. In vitro cleavage and joining at the viral origin of replication by the replication initiator protein of tomato yellow leaf curl virus. 1995;92(April):3879–83.

50. Mankertz A, Hillenbrand B. Replication of porcine circovirus type 1 requires two proteins encoded by the viral Rep gene. Virology. 2001;279(2):429–38.

51. Rizvi I, Choudhury NR, Tuteja N. Insights into the functional characteristics of geminivirus rolling-circle replication initiator protein and its interaction with host factors affecting viral DNA replication. Arch Virol. 2015;160(2):375–87.

52. Belval L, Marquette A, Mestre P, Piron MC, Demangeat G, Merdinoglu D, et al. A fast and simple method to eliminate Cpn60 from functional recombinant proteins produced by E. coli Arctic Express. Protein Expr Purif. 2015;109:29–34.

53. Kapust RB, Waugh DS. Escherichia coli maltose-binding protein is effective in promoting solubility. Protein Sci. 1999;(8):1668–74.

54. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: Homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46(W1):W296–303.

55. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. Electrostatics of nanosystems: Application to microtubules and the ribosome. Proc Natl Acad Sci U S A. 2001;98(18):10037–41.

56. Stenger DC, Revington GN, Stevenson MC, Bisaro DM. 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. 1991 Sep 15;88(18):8029-33.

57. Cheung AK. Rolling-Circle Replication of an animal circovirus genome in a theta-replicating bacterial plasmid in *Escherichia coli*. J Virol. 2006;80(17):8686–94.

58. Rigden JE, Dry IB, Krake LR, Rezaian MA. Plant virus DNA replication processes in Agrobacterium: Insight into the origins of geminiviruses? Proc Natl Acad Sci U S A. 1996;93(19):10280–4.

59. Selth LA, Randles JW, Rezaian MA. *Agrobacterium tumefaciens* supports DNA replication of diverse geminivirus types. FEBS Lett. 2002;516(1–3):179–82.

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

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

62. Huang SW, Liu HP, Chen JK, Shien YW, Wong ML, Wang CY. Dual ATPase and GTPase activity of the replication-associated protein (Rep) of beak and feather disease virus. Virus Res. 2016;213:149–61.

63. Su YP, Shien JH, Liu HJ, Yin HS, Lee LH. Avian reovirus core protein μA expressed Escherichia coli possesses both NTPase and RTPase activities. J Gen Virol. 2007;88(6):1797–805.

64. Sarin J, Aggarwal S, Chaba R, Varshney GC, Chakraborti PK. B-subunit of phosphate-specific transporter from *Mycobacterium tuberculosis* is a thermostable ATPase. J Biol Chem. 2001;276(48):44590–7.

65. Steinfeldt T, Finsterbusch T, Mankertz A. Functional analysis of cis- and trans-acting replication factors of porcine circovirus type 1. J Virol. 2007;81(11):5696–704.

66. Merits A, Fedorkin ON, Guo D, Kalinina NO, Morozov SY. Activities associated with the putative replication initiation protein of Coconut foliar decay virus, a tentative member of the genus Nanovirus. J Gen Virol. 2000;81(12):3099–106.

67. Orozco BM, Gladfelter HJ, Settlage SB, Eagle PA, Gentry RN, Hanley-Bowdoin L. Multiple cis elements contribute to geminivirus origin function. Virology. 1998;242(2):346–56.

68. Fontes EPB, Eagle PA, Sipe PS, Luckow VA, Hanley-Bowdoin L. Interaction between a geminivirus replication protein and origin DNA is essential for viral replication. J Biol Chem. 1994;269(11):8459–65.

69. Muñoz-Martín A, Collin S, Herreros E, Mullineaux PM, Fernández-Lobato M, Fenoll C. Regulation of MSV and WDV virion-sense promoters by WDV nonstructural proteins: A role for their retinoblastoma protein-binding motifs. Virology. 2003;306(2):313–23.

70. Hefferon KL, Moon YS, Fan Y. Multi-tasking of nonstructural gene products is required for bean yellow dwarf geminivirus transcriptional regulation. FEBS J. 2006;273(19):4482–94.

71. Cheung AK. Transcriptional analysis of porcine circovirus type 2. Virology. 2003;305(1):168–80.

72. Kazlauskas D, Varsani A, Krupovic M. Pervasive chimerism in the replication-associated proteins of uncultured single-stranded DNA viruses. Viruses. 2018;10(4):1–11.

73. Dupureur CM. NMR studies of restriction enzyme-DNA interactions: Role of conformation in sequence specificity. Biochemistry. 2005;44(13):5065–74.

74. Cowan JA. Metal activation of enzymes in nucleic acid biochemistry. Chem Rev. 1998;98(3):1067–87.

75. Tewary SK, Liang L, Lin Z, Lynn A, Cotmore SF, Zhao H, et al. Structures of minute virus of mice replication initiator protein Nterminal domain: insights into DNA nicking and origin binding. Virology. 2016;61–71.

76. Trempe F, Gravel A, Dubuc I, Wallaschek N, Collin V, Gilbert-Girard S, et al. Characterization of human herpesvirus 6A/B U94 as ATPase, helicase, exonuclease and DNA-binding proteins. Nucleic Acids Res. 2015;43(12):6084–98.

77. Boer R, Russi S, Guasch A, Lucas M, Blanco AG, Pérez-Luque R, et al. Unveiling the molecular mechanism of a conjugative relaxase: the structure of TrwC complexed with a 27-mer DNA comprising the recognition hairpin and the cleavage site. J Mol Biol. 2006;358(3):857–69.

78. Mankertz A, Persson F, Mankertz J, Blaess G, Buhk HJ. Mapping and characterization of the origin of DNA replication of porcine circovirus. J Virol. 1997;71(3):2562–6.

79. Cheung AK. Detection of rampant nucleotide reversion at the origin of DNA replication of porcine circovirus type 1. Virology. 2005;333(1):22–30.

80. Cheung AK. 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. 2007 Jun 20;363(1):229–35.

81. Lefeuvre P, Lett J-M, Varsani A, Martin DP. Widely conserved recombination patterns among single-stranded DNA viruses. J Virol. 2009;83(6):2697–707.

82. Mankertz A, Mueller B, Steinfeldt T, Schmitt C, Finsterbusch T. New reporter gene-based replication assay reveals exchangeability of replication factors of porcine circovirus types 1 and 2. J Virol. 2003;77(18):9885–93.

83. Ramos PL, Guevara-González RG, Peral R, Ascencio-Ibañez JT, Polston JE, Argüello-Astorga GR, et al. Tomato mottle Taino virus pseudorecombines with PYMV but not with ToMoV: Implications for the delimitation of cis- and trans-acting replication specificity determinants. Arch Virol. 2003;148(9):1697–712.

84. Laufs J, Jupin I, David C, Schumacher S, Heyraud-Nitschke F, Gronenborn B. Geminivirus replication: Genetic and biochemical characterization of Rep protein function, a review. Biochimie. 1995;77(10):765–73.

85. Cherepanov A V., De Vries S. Kinetics and thermodynamics of nick sealing by T4 DNA ligase. Eur J Biochem. 2003;270(21):4315–25.

86. Bullard DR, Bowater RP. Direct comparison of nick-joining activity of the nucleic acid ligases from bacteriophage T4. Biochem J. 2006;398(1):135–44.

87. Lohman GJS, Zhang Y, Zhelkovsky AM, Cantor EJ, Evans TC. Efficient DNA ligation in DNA-RNA hybrid helices by Chlorella virus DNA ligase. Nucleic Acids Res. 2014;42(3):1831–44

Chapter Five

Concluding Remarks and Future Directions

The work presented in this dissertation was undertaken to better understand cruciviruses, a novel group of circular replication associated protein encoding single-stranded DNA viruses (CRESS-DNA virus) (1). To do this we combined basic bioinformatic analyses with biochemical studies. These studies enabled us to answer some basic questions about cruciviruses. We specifically addressed the properties that unite cruciviruses with other CRESS-DNA viruses, I showed that a putative protein encoded on their genomes is active in vitro, and identified crucivirus genomes in various environments.

Chapter two of this dissertation presented the first large scale analysis of crucivirus genomes from various metagenomes: acquatic systems, engineered systems, eukaryote-associated flora. Our analyses showed that cruciviruses are a diverse group of CRESS-DNA viruses whose placement within established CRESS-DNA phylogenies is difficult and blurs the lines of established CRESS-DNA phylogenies. The difficulty of placement can be in part attributed to the unusual amount of diversity in the putative replication associated protein (Rep) encoded by their genomes. Traditionally this protein has been used to classify CRESS-DNA viruses (2). However, it is now apparent that, as a group, cruciviruses contain putative *Rep* genes that span previously classified as well as unclassified CRESS-DNA viruses (2,3). This can likely be explained by intergenic as well as intragenic recombination of cruciviruses with a variety of CRESS-DNA

viruses (3), in addition to more than one RNA virus capsid acquisition event by plasmids encoding diverse replication associated proteins (Rep) resulting in multiple initial emergences of CRESS-DNA viruses, and potentially cruciviruses (4).

In these studies, we did not examine any possible influences that geographic location or metagenome type may have on types of crucivirus genomes uncovered. Given that cruciviruses have been found in a wide range of geographic locations, geographic analysis of crucivirus genomes could provide insights as to what types of cruciviruses are present in distinct parts of the world. Since cruciviruses encode putative Rep proteins which are similar to a wide variety of members of the *Cressdnaviricota,* it may also prove useful to evaluate these metagenomes for the presence of other CRESS-DNA viruses which may influence recombination patterns. Given that CRESS-DNA viruses display high recombination rates (5), a more detailed analysis of these patterns in cruciviruses could provide more insight into their seemingly tangled evolutionary patterns.

Attempting to elucidate a crucivirus host is an attractive follow up to the environmental and metagenomic work that first identified these cruciviruses. The questions of “who is doing what?” or “who is infecting whom?” are often left unanswered by viral metagenomic studies. The definitive association of a crucivirus with a host would begin to answer this question. To date, no host for any crucivirus has been determined despite studies forming loose and inferred associations with various potential hosts (6–8). A eukaryotic host appears likely based on the architecture of Rep; an HUH endonuclease domain fused to a

superfamily-3-helicase (1,3). This fusion has only been observed in CRESS-DNA viruses infecting eukaryotes and is absent in those CRESS-DNA viruses which infect prokaryotes (9,10). The advent and subsequent refinement of single cell sequencing technologies and digital droplet PCR may provide useful tools for this endeavor (11). Single cell genomics technologies have been employed to elucidate virus-host relationships in marine environments suggesting that they may be particularly applicable in the hunt for a crucivirus host (12,13).

Chapter three of this work sought to better understand the relationship that exists between putative Reps and their putative origins of replication (ori) in cruciviruses. Previously characterized relationships between Rep of CRESS-DNA viruses and their iterated DNA sequences near stem loops (iterons) identified a small number of amino acids, deemed specificity determinants (SPDs) that putatively play a role in ori binding discrimination (14,15). It has also been predicted that due to likely common ancestry of Reps spanning cellular and viral lineages alike that these regions adjacent to conserved motifs should harbor SPDs (14,16,17). The aforementioned heterogeneity observed in both Rep of cruciviruses, and their iterated DNA sequences made this task difficult. It appears that the regions previously identified in Rep of CRESS-DNA viruses remain largely variable even among cruciviruses harboring similar iterons. It is possible that SPDs for cruciviruses are in different regions which could be explored with future biochemical analyses. Mutation of the residues adjacent to motif I and

motif II in Rep coupled with electrophoretic mobility shift assays could provide hints as to whether these regions are SPDs in crucivirus Reps.

Chapter four presented the first demonstration of biochemical activities associated with a putative protein found in a crucivirus genome, Rep of BSL-RDHV. These results indicate that Rep of BSL-RDHV is capable of the activities associated with initiation and completion of rolling circle replication characterized in other CRESS-DNA viruses (18–22). Results showed that Rep hydrolyzes ATP, but I did not attempt to detect helicase activity assumed to be associated with that activity. Demonstrating that Rep is capable of helicase activity could be accomplished through the use of partially dsDNA templates containing overhangs of various lengths. This line of inquiry would aid in better understanding the mechanism by which BSL-RDHV replicates.

My results indicate that Rep of BSL-RDHV appears to be dependent on double stranded iterated DNA sequences for ori binding. Future studies could be employed to further define the minimal binding site for Rep. The presence of imperfect repeats downstream of the predicted stem-loop may be capable of mediating Rep binding as has been shown in other CRESS-DNA viruses (19). Additionally, the oligomerization state of Rep could be explored. Previous studies have shown that the disruption of formation of Rep dimers leads to a decrease of replication of porcine circovirus type 2 (23), suggesting that similar higher order Rep structures may be important in the case of BSL-RDHV.

Chapter four also demonstrated that Rep covalently attaches to ssDNA indicative of the initiation of RCR. Additionally, chapter four demonstrated that when Rep is incubated with an oligonucleotide containing the predicted BSL-RDHV ori and an increasing concentration of a preformed acceptor oligonucleotide that the shifted band indicative of covalent attachment appears to be diminished, suggestive of a joining reaction (termination of RCR) taking place. The nonanucleotide substitution work in Chapter four suggests that covalent attachments of Rep to ssDNA following a nicking reaction are capable of forming when the first and third positions of the nonanucleotide are mutated. These data also indicates that Rep of BSL-RDHV becomes covalently attached to a ssDNA oligonucleotide carrying the nonanucleotide of porcine circovirus type 1, suggestive of the ability to initiate RCR in a promiscuous manner. Further work to fully characterize the DNA sequence (and potential structure) requirements for these activities could be undertaken with the methods presented in Chapter four. A fuller examination of the promiscuity of Rep for nicking and joining may also help to support the idea presented in Chapter two that some cruciviruses are bipartite.

This work was unable to offer definitive insight into the putative role that Rep may play in RNA-DNA recombination, one possible scenario for the initial emergence of cruciviruses. Rep does not appear to nick RNA, the ability of Rep to mediate joining reactions between RNA and DNA has not been examined. This topic should be explored more fully. This dissertation and the associated methods presented in Chapter four may provide a good starting point for those

studies. The use of acceptor RNA oligonucleotides with an associated “protein shift” assay, similar to what was described for ssDNA joining activity in Chapter four, is a logical starting point for these studies.

Appendix A presented both successful and unsuccessful attempts to directly amplify and clone crucivirus genomes from various environments. Our results indicate that cruciviruses are present in soil and water samples taken from Woodburn, Oregon while they are undetectable in a variety of aquatic sediments. Metagenomic studies could be employed in the environments from which crucivirus genomes were cloned, which may lead to the discovery of more genomes. Similarly, the environments that did not produce crucivirus genomes by direct amplification and cloning could be subjected to deep sequencing which may clarify whether cruciviruses are indeed absent from those locations.

This research originally sought to uncover the biochemical mechanism that may have led to a capsid protein gene of an RNA virus being acquired by a DNA virus. While this question remains unanswered, the work described in this dissertation has emphasized that the cruciviruses are a unique group of viruses and are worthy of further study. Future studies proposed in this concluding chapter would provide insight to the poorly understood topic of viral evolution as well as the process of rolling circle replication in ssDNA viruses as a whole.

References

1. Diemer GS, Stedman KM. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses Biol Direct. 2012;7(13):1–14.
2. Krupovic M, Varsani A, Kazlauskas D, Breitbart M, Delwart E, Yutin N, et al. Cressdnaviricota: a virus phylum unifying 7 families of Rep-encoding viruses with single- stranded, circular DNA genomes 3. 2020;94(12):e00582-20.
3. de la Higuera I, Kasun GW, Torrance EL, Pratt AA, Maluenda A, Colombet J, et al. Unveiling crucivirus diversity by mining metagenomic data. mBio. 2020;11(5):1–17.
4. Kazlauskas D, Varsani A, Koonin E V., Krupovic M. Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. Nat Commun. 2019;10(1):1–12.
5. Lefeuvre P, Lett J-M, Varsani A, Martin DP. Widely conserved recombination patterns among single-stranded DNA viruses. J Virol. 2009;83(6):2697–707.
6. Roux S, Enault F, Bronner G, Vaulot D, Forterre P, Krupovic M. Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses. Nat Commun. 2013;4:2700.
7. Krupovic M, Zhi N, Li J, Hu G, Koonin E V., Wong S, et al. Multiple layers of chimerism in a single-stranded DNA virus discovered by deep sequencing. Genome Biol Evol. 2015;7(4):993–1001.
8. Bistolas K, Besemer R, Rudstam L, Hewson I. Distribution and Inferred Evolutionary Characteristics of a Chimeric ssDNA Virus Associated with Intertidal Marine Isopods. Viruses. 2017;26(12):361.
9. Krupovic M. Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. Curr Opin Virol. 2013;3(5):578–86.
10. Chandler M, Cruz F De, Dyda F, Hickman AB. Breaking and joining single-stranded DNA : the HUH endonuclease superfamily. Nat Rev Microbiol. 2013;11(8):525–38.
11. Quan PL, Sauzade M, Brouzes E. DPCR: A technology review. Sensors (Switzerland). 2018;18(4):1271.
12. Martinez-Hernandez F, Garcia-Heredia I, Gomez ML, Maestre-Carballa L, Martínez JM, Martinez-Garcia M. Droplet digital PCR for estimating absolute abundances of widespread pelagibacter viruses. Front Microbiol. 2019;10(JUN):1–13.

13. Martinez-Hernandez F, Fornas Ò, Lluesma Gomez M, Garcia-Heredia I, Maestre-Carballa L, López-Pérez M, et al. Single-cell genomics uncover Pelagibacter as the putative host of the extremely abundant uncultured 37-F6 viral population in the ocean. ISME J. 2019;13(1):232–6.

14. Londoño A, Riego-Ruiz L, Argüello-Astorga GR. DNA-binding specificity determinants of replication proteins encoded by eukaryotic ssDNA viruses are adjacent to widely separated RCR conserved motifs. Arch Virol. 2010;155(7):1033–46.

15. Ruiz-medrano R, Irapuato U. An iteron-related domain is associated to Motif 1 in the replication proteins of geminiviruses: identification of potential interacting amino acid-base pairs by a comparative approach. Arch Virol. 2001;1465–85.

16. Ilyina T V, Koonin E V. Conserved sequence motifs in the initiator proteins for rolling circle dna-replication encoded by diverse replicons from Eubacteria, Eukaryotes and Archaebacteria. Nucleic Acids Res. 1992;20(13):3279–85.

17. Koonin E V. A common set of conserved motifs in a vast variety of putative nucleic acid-dependent ATPases including MCM proteins involved in the initiation of eukaryotic DNA replication. Nucleic Acids Res. 1993;21(11):2541–7.

18. Steinfeldt T, Finsterbusch T, Mankertz A. 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. 2006;80(13):6225–34.

19. Steinfeldt T, Finsterbusch T, Mankertz A. Rep and Rep' protein of Porcine circovirus type 1 bind to the origin of replication in vitro. Virology. 2001;291(1):152–60.

20. Hafner GJ, Stafford MR, Wolter LC, Harding RM, Dale JL. Nicking and joining activity of banana bunchy top virus replication protein in vitro. J Gen Virol. 1997 Jul;78 (Pt 7):1795-9.

21. Timchenko T, de Kouchkovsky F, Katul L, David C, Vetten HJ, Gronenborn B. A single Rep protein initiates replication of multiple genome components of faba bean necrotic yellows virus, a single-stranded DNA virus of plants. J Virol. 1999;73(12):10173–82.

22. Kai J, Chiaolong C, Jian H, Wu S, Yi S, Chi L, et al. Characterization of the endonuclease activity of the replication - associated protein of beak and feather disease virus. Arch Virol. 2019 Aug;164(8):2091-2106

23. Luo G, Zhu X, Lv Y, Lv B, Fang J, Cao S, et al. Crystal structure of the dimerized N-terminus of porcine circovirus type 2 replicase protein reveals a novel antiviral interface. J Virol. 2018;92(18):JVI.00724-18

Appendix A

Isolation of Crucivirus Genomes From Environmental DNA Samples

This chapter has been modified from: **Genome Sequences of Three Cruciviruses Found in the Willamette Valley (Oregon).** Ignacio de la Higuera, Ellis L. Torrance, Alyssa A. Pratt, George W. Kasun, Amberlee Maluenda, Kenneth M. Stedman. Microbiology Resource Announcements Jun 2019, 8 (23) e00447-19; DOI: 10.1128/MRA.00447-19

Data availability: The information and genomic sequences of CruV-MC1, CruV-MC2, and CruV-MC3 were deposited at DDBJ/ENA/GenBank under the accession numbers MK679543, MK679544, and MK679545, respectively.

Abstract

Cruciviruses are single-stranded DNA (ssDNA) viruses whose genomes suggest the possibility of gene transfer between DNA and RNA viruses. Many crucivirus genome sequences have been found in metagenomic data sets, although no crucivirus has been isolated. Here, we present the complete genome sequences of three cruciviruses recovered and cloned from environmental samples from Oregon, as well as the results of other environmental sampling that did not uncover crucivirus genomes. We also present basic genome analyses showing that newly uncovered cruciviruses are similar to previously described cruciviruses and other single-stranded DNA viruses.

Introduction

The advent of next-generation or deep sequencing technologies in the last 15 years has led to an exponential increase in the number of viral genomes in public databases such as the Viral Genome Resource Center (National Center for Biotechnology Information) and ViralZone (Swiss Institute of Bioinformatics) (1,2). Viral metagenomics has revealed the presence of “viral dark matter”, a large number of apparent viral genomes with no similarity to sequences currently deposited in public databases (2). At the same time, the number of viral reference genomes has grown significantly. This availability of both reference and novel genomes has allowed researchers easier paths to annotating newly discovered viral genomes, which has increased our understanding of the ambiguous topic of viral evolution (3,4).

The development of the use of phi29 polymerase with random hexamer primers in rolling circle amplification (RCA) has led to a similar increase in the number of circular single-stranded DNA (ssDNA) virus genomes discovered through metagenomics (5–12). RCA allows for the selective amplification of small amounts of ssDNA molecules from within an environmental or clinical sample, which allows for their subsequent detection using deep-sequencing (13,14). These genomes can then be confirmed and cloned through inverse polymerase chain reaction (PCR) utilizing back to back primers (15,16) Because RCA is efficient in amplifying complete genomes this technique has also been adopted in the direct amplification and cloning of complete ssDNA genomes (16,17). This

practice of confirming genomes assembled through metagenomics with conventional PCR seems prudent, when possible, due to the potential for the introduction of chimeric reads due to the RCA process (18). Initially used for the diagnosis of geminivirus disease in plants (14), this technique has revealed the omnipresent nature of circular replication associated protein encoding ssDNA viruses (CRESS-DNA virus) globally (6,7). Additionally, RCA has been previously used to directly amplify and subsequently clone novel CRESS-DNA virus genomes (15).

One such novel CRESS-DNA virus genome identified through phi29 amplification paired with deep sequencing, and subsequently cloned, is that of Boiling Springs Lake RNA-DNA hybrid virus (15, 22, 28). The first described member of the growing crucivirus group, Boiling Springs Lake RNA-DNA Hybrid Virus (BSL-RDHV) was discovered through a metagenomic study of a high temperature and low pH lake. This genome represented the first direct evidence of a protein homologue in both RNA and DNA viruses, namely a putative capsid protein most similar to those found in ssRNA viruses (15). Since this initial discovery approximately 800 additional crucivirus genomes have been uncovered through deep sequencing (8–10,12,19–27, Chapter Two of this work). Despite this growing number of cruciviruses discovered in metagenomes, only a small number have been verified through inverse PCR and subsequent cloning (15,22,28). Here, we report that crucivirus genomes were not recovered from a variety of additional aquatic sediments using RCA. We also present the results of

sampling different environments which led to the discovery and subsequent cloning of three novel crucivirus genomes. Basic analysis of these new crucivirus genomes show that they are similar to previously described crucivirus and CRESS-DNA virus genomes.

Methods

Environmental DNA Extraction: Approximately 20l of water and sediment was collected from the north end of Boiling Springs Lake (Lassen Volcanic National Park, California) in July of 2014. The sediment portion accounted for approximately one third of the total volume collected. DNA was extracted the following week as previously described (15,29). Additional environmental DNA was isolated from Boiling Springs Lake the following August (2015) using the PowerLyzer PowerSoil DNA isolation kit. DNA was extracted from Deschutes River (Oregon, 2015) sediment, Mirror Lake (Oregon, 2016) sediment , Clackamas River (Oregon, 2016) sediment, and Woodburn soil (Oregon, 2018) using the PowerLyzer PowerSoil DNA Isolation Kit (MoBio) as per the manufacturer's instructions. All extracted DNA was stored at -20 °C until further use.

phi29 Rolling Circle Amplification: Extracted environmental DNA was used as a template for isothermal rolling circle amplification. Reactions consisting of 30U phi29 polymerase (New England Biolabs), phi29 reaction buffer supplemented with BSA to final concentration of 200μg/ml, 2.5μM either random hexamer primers, CP-phi-F or CP-phi-R and water were assembled in a clean PCR hood. In the case of reactions containing CP-phi-F and CP-phi-R reactions were

incubated at 30 °C for 30 minutes to clean up potential contaminating DNA in reaction components. After 30 minutes, dNTPs were added to a final concentration of 1mM, and 5-20ng template DNA was added. Reactions were incubated at 35 °C for 5 min, 34 °C for 10 min, 33 °C for 15 min, 32 °C for 20 min, 31 °C for 30 min, and 30 °C for 16 hours (30). Amplification was confirmed by 0.7% agarose gel electrophoresis, and reactions were extracted and DNA was concentrated by ethanol/sodium acetate precipitation. This phi29 amplified DNA was then used in a PCR containing degenerate primers targeting the conserved S-domain of putative crucivirus capsid protein genes. Varying degenerate primer pairs (**Table A.1**) were used in a PCR that contained varying amounts of phi29 amplified environmental DNA. Because samples from Boiling Springs and Mirror Lakes sediment, and Deschutes and Clackamas Rivers sediment never resulted in a band of correct size no PCR products were cloned.

While none of the sediment DNA samples that I tested generated appropriately sized amplicons using conserved crucivirus sequences, other members of the group were successful in the subsequent amplification of complete crucivirus genomes from soil and water collected in Woodburn, Oregon. CruV-MC1, CruV-MC2, and CruV-MC3 were cloned in a PCR cloning vector, pMINIT2.0 (New England Biolabs) and subsequently sequenced.

Results and Discussion

No Crucivirus Genomes From Various Environments: phi29 amplification with random hexamer primers of environmental DNA extracted from Deschutes River (Oregon) water and sediment, Mirror Lake (Oregon) sediment, Boiling

Springs Lake (California) sediment, and Clackamas River (Oregon) sediment resulted in DNA amplification. However, no PCR products of expected size were recovered when the phi29 amplified DNA was used in a PCR with degenerate primers targeting the conserved Crucivirus CP S-domain. The lack of detectable crucivirus PCR products may be because the genomic sequences are too divergent to amplify. The DNA templates used in the initial phi29 RCA procedures were at least one year old by the time we developed a robust method that reliably amplified crucivirus sequences. Perhaps in that time the extracted DNA (stored at -20 °C) became unsuitable for detection of rare crucivirus genomes. Of course, it is also possible that these environments simply do not harbor cruciviruses.

This second explanation is seemingly at odds with the discovery of BSL-RDHV in the metagenome of Boiling Springs Lake. However, since the initial discovery of BSL-RDHV cruciviruses have been isolated from a wide variety of environments (8–10,12,19–27), but none with a pH as low or a temperature as high as that of Boiling Springs Lake. Perhaps the initial discovery of BSL-RDHV was fortuitous in its timing and the host within Boiling Springs Lake was simply not present in subsequent years at the time of sample collection in numbers sufficient to extract crucivirus genomes from extracellular virions. The original environmental DNA from which BSL-RDHV was cloned was extracted from 20l of Boiling Springs Lake sediment (15, 29). Perhaps the amount of DNA isolated from approximately 6l of sediment (above) simply did not contain sufficient BSL-RDHV DNA to result in successful amplification. Finally, it may be possible that

BSL-RDHV was initially detected in metagenomic surveys due to a “contaminant” genome that does not truly reside in Boiling Springs Lake. However, we were able to detect and sequence PCR products from Boiling Springs Lake environmental DNA (both non phi29 amplified as well as amplified) that are similar to those of VP2, the structural protein gene of Sulfolobus spindle shaped virus 1 (31).

Name	Sequence 5’-3’	Use
CP-phi-F	RTNGARTG*Y*G	Phi-29 Amplification
CP-phi-R	KCRCAYTC*N*A	Phi-29 Amplification
Random Hexamers	NNNNNN	Phi-29 Amplification
ChiV-F	GGTWCWRTHATWATGKCTACTSAWTAYAA	Degenerate primer targeting conserved crucivirus capsid domain
ChiV-R	TTRTAWTSAGTAGMCATWATDAYWGWACC	Degenerate primer targeting conserved crucivirus capsid domain
ChiV-CP-F	ATGKCTACTSAWTAYRAYKCT	Degenerate primer targeting conserved crucivirus capsid domain
ChiV-CP-R	KKRTCRCATTCAACWSCRTG	Degenerate primer targeting conserved crucivirus capsid domain
B2B_iDegF	GGCWACKNAWTATAATGCWWC	Inverse degenerate primers to amplify complete crucivirus genome
B2B_iDegR	ATAAYWACWGKWCCHARWGC	Inverse degenerate primers to amplify complete crucivirus genome

Table A.1. PCR primers used in this study. Asterisks indicate 3’- phosphorothioation.

Cruciviruses Discovered in Woodburn, Oregon: Using similar techniques other members of the group were successful in detecting and subsequently cloning complete crucivirus genomes from samples from Woodburn, Oregon. The genomes of CruV-MC1 and CruV-MC2 were recovered from soil samples, while CruV-MC3 was found in a water sample from a creek adjacent to the soil samples collected above. All three crucivirus genomes contain two major ORFs (>300 amino acids) arranged in an ambisense orientation, similar to the arrangement of circoviruses and cycloviruses (2). One ORF encodes a putative replication associated protein (Rep) that contain all motifs necessary for endonuclease and ATPase activity associated with initiation and completion of rolling circle replication (RCR) in CRESS-DNA viruses (9–11). The second ORF encodes a putative capsid protein (CP) similar to those of previously described cruciviruses.

StemLoop-Finder (Pratt, Torrance, Kasun, Stedman and de la Higuera, 2021) identified putative stem loop structures and associated nonanucleotides in the intergenic region of CruV-MC2 and CruV-MC3 which putatively serve as origins of replication. CruV-MC2 and CruV-MC3 contain the nonanucleotides 5'-TAGTATTAC-3' and 5'-GAGTATTAC-3' respectively, which are both found in a number of circoviruses (**Fig. A.1**) (33). A manual search of the CruV-MC1 genome revealed the presence of a 5'-TAGTATTAC-3 nonanucelotide in the 3' region of the putative *Rep* gene. Other putative stem-loops have been localized to intragenic portions of putative CP genes, but to our knowledge none have

been identified in putative *Rep* genes (13). The nonaucleotide of CruV-MC1 was predicted to be partially located in a stem-loop by mFold (**Fig. A.1**) (14).

When CruV-MC2 and CruV-MC3 were analyzed with CRUISE for the presence of potential iterons (Chapter Three), no likely repeated sequences that could function as iterons were found. When CruV-MC1 was analyzed by CRUISE the sequence 5'-GCGCCT-3' was found to be repeated twice in the 5' region of the predicted stem-loop with a spacer of 3 bases (**Fig. A.1**). The predicted hexamer repeats overlap with the first position of the nonanucleotide (**Fig. A.1**).

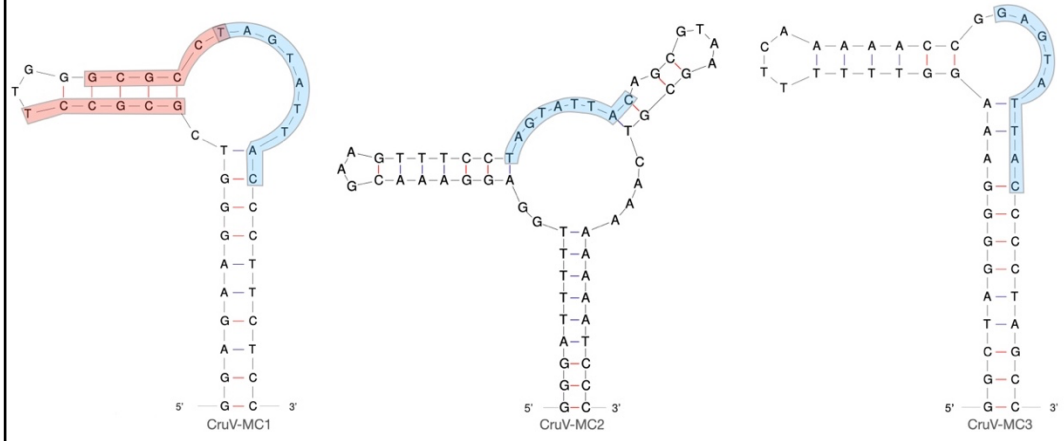

Figure A.1. Predicted stem loop structures of CruV-MC1, CruV-MC2, and CruV-MC3. Nonanucleotides are highlighted in blue, and predicted iterons are highlighted in red. Stem loop structures were predicted with StemLoop-Finder (CruV-MC2 and CruV-MC3) and the mFold webserver (CruV-MC1) (14). Iterons were predicted with CRUISE (in preparation).

To our knowledge this type of arrangement has not been observed in other CRESS-DNA viruses. However, no other repeated sequences are present in the region of the predicted stem-loop leaving this repeat as the only potential iteron. Given the arrangement of the nonanucleotides within these potential stem loops

relative to other CRESS-DNA viruses these may not serve as origins of replication.

References

1. Brister JR, Ako-Adjei D, Bao Y, Blinkova O. NCBI viral Genomes resource. Nucleic Acids Res. 2015;43(D1):D571–7.
2. Krishnamurthy SR, Wang D. Origins and challenges of viral dark matter. Virus Res. 2017;239:136–42.
3. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: Mining viral signal from microbial genomic data. PeerJ. 2015;2015(5):1–20.
4. Wang S, Sundaram JP, Spiro D. VIGOR, an annotation program for small viral genomes. BMC Bioinformatics. 2010;11:451.
5. Binga EK, Lasken RS, Neufeld JD. Something from (almost) nothing: The impact of multiple displacement amplification on microbial ecology. ISME J. 2008;2(3):233–41.
6. Rosario K, Duffy S, Breitbart M. A field guide to eukaryotic circular single-stranded DNA viruses: Insights gained from metagenomics. Arch Virol. 2012;157(10):1851–71.
7. Zhao L, Rosario K, Breitbart M, Duffy S. Eukaryotic circular Rep-encoding single-stranded DNA (CRESS DNA) viruses: ubiquitous viruses with small genomes and a diverse host range. 1st ed. Advances in Virus Research. Elsevier Inc.; 2018. 1–63 p.
8. de la Higuera I, Kasun GW, Torrance EL, Pratt AA, Maluenda A, Colombet J, et al. Unveiling crucivirus diversity by mining metagenomic data. mBio. 2020;11(5):1–17.
9. Quaiser A, Krupovic M, Dufresne A, Roux S. Diversity and comparative genomics of chimeric viruses in Sphagnum-dominated peatlands. Virus Evol. 2016;2(2):1–8.
10. Roux S, Enault F, Bronner G, Vaulot D, Forterre P, Krupovic M. Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses. Nat Commun. 2013;4:2700.
11. Liu Q, Wang H, Ling Y, Yang SX, Wang XC, Zhou R, et al. Viral metagenomics revealed diverse CRESS-DNA virus genomes in faeces of forest musk deer. Virol J. 2020;17(1):1–9.
12. Tisza MJ, Pastrana D V., Welch NL, Stewart B, Peretti A, Starrett GJ, et al. Discovery of several thousand highly diverse circular DNA viruses. eLife. 2020;9:1–26.
13. Kim KH, Bae JW. Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. Appl Environ Microbiol. 2011;77(21):7663–8.

14. Haible D, Kober S, Jeske H. Rolling circle amplification revolutionizes diagnosis and genomics of geminiviruses. J Virol Methods. 2006;135(1):9–16.

15. Diemer GS, Stedman KM. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses Biol Direct. 2012;7(13):1–14.

16. Rosario K, Duffy S, Breitbart M. Diverse circovirus-like genome architectures revealed by environmental metagenomics. J Gen Virol. 2009;90(10):2418–24.

17. Inoue-Nagata AK, Albuquerque LC, Rocha WB, Nagata T. A simple method for cloning the complete begomovirus genome using the bacteriophage φ29 DNA polymerase. J Virol Methods. 2004;116(2):209–11.

18. Tu J, Guo J, Li J, Gao S, Yao B, Lu Z. Systematic characteristic exploration of the chimeras generated in multiple displacement amplification through next generation sequencing data reanalysis. PLoS One. 2015;10(10):1–10.

19. Bistolas K, Besemer R, Rudstam L, Hewson I. Distribution and Inferred Evolutionary characteristics of a chimeric ssDNA virus associated with intertidal marine isopods. Viruses. 2017 Nov 26;9(12):361.

20. Salmier A, Tirera S, De Thoisy B, Franc A, Darcissac E, Donato D, et al. Virome analysis of two sympatric bat species (*Desmodus rotundus* and *Molossus molossus*) in French Guiana. PLoS One. 2017;12(11):1–25.

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

22. Krupovic M, Zhi N, Li J, Hu G, Koonin E V., Wong S, et al. Multiple layers of chimerism in a single-stranded DNA virus discovered by deep sequencing. Genome Biol Evol. 2015;7(4):993–1001.

23. Rosario K, Dayaram A, Marinov M, Ware J, Kraberger S, Stainton D, et al. Diverse circular ssDNA viruses discovered in dragonflies (Odonata: Epiprocta). J Gen Virol. 2012;93(Pt 12):2668-2681.

24. Hewson I, Ng G, Li WF, LaBarre BA, Aguirre I, Barbosa JG, et al. Metagenomic identification, seasonal dynamics, and potential transmission mechanisms of a Daphnia-associated single-stranded DNA virus in two temperate lakes. Limnol Oceanogr. 2013;58(5):1605–20.

25. Steel O, Kraberger S, Sikorski A, Young LM, Catchpole RJ, Stevens AJ, et al. Circular replication-associated protein encoding DNA viruses identified

in the faecal matter of various animals in New Zealand. Infect Genet Evol. 2016;43:151–64.

26. Mcdaniel LD, Rosario K, Breitbart M, Paul JH. Comparative metagenomics: Natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. Environ Microbiol. 2014 Feb;16(2):570-85.

27. Dayaram A, Galatowitsch ML, Argüello-Astorga GR, van Bysterveldt K, Kraberger S, Stainton D, et al. Diverse circular replication-associated protein encoding viruses circulating in invertebrates within a lake ecosystem. Infect Genet Evol. 2016;39:304–16.

28. Rosario K, Nilsson C, Lim YW, Ruan Y, Breitbart M. Metagenomic analysis of viruses in reclaimed water. Environ Microbiol. 2009;11(11):2806–20.

29. Diemer GS. The Boiling Springs Lake Metavirome: Charting the viral sequence-space of an extreme environment microbial ecosystem. Dissertation. 2014.

30. Dean FB, Nelson JR, Giesler TL, Lasken RS. Rapid amplification of plasmid and phage DNA using Phi29 DNA polymerase and multiply-primed rolling circle amplification. Genome Res. 2001;11(6):1095–9.

31. Reiter, WD., Palm, P., Henschen, A. *et al.* Identification and characterization of the genes encoding three structural proteins of the *Sulfolobus* virus-like particle SSV1. Mol Gen Genet. (1987) 206**,** 144–153.