Discovery of novel circular replication-associated protein encoding single-stranded DNA viruses in ecosystems using viral metagenomic approaches

A thesis

submitted in partial fulfilment

of the requirements for the degree

of

Doctor of Philosophy

at the

University of Canterbury

New Zealand

Anisha Dayaram

2014

I would like to dedicate my thesis

to my late Grandpa

Contents

Acknowledgments

I would like to thank the following people for their help, support and encouragement throughout the past four years of my research. Without their involvement my PhD would not have been possible.

Arvind Varsani , words cannot express how grateful I am for taking me on as your PhD student. Your inspiration and enthusiasm encouraged me to stay in science and pursue further postgraduate studies. You have been an amazing mentor throughout my PhD and have given me the best possible supervision that anyone could ever wish for. As well as your ongoing support and guidance, you have changed the way I approach, analyse and question in general. Finishing my PhD I feel that I have gained a broad wealth of knowledge and work ethic that will continue to benefit me in years to come.

Co-supervisor **Renwick Dobson** for your help, support and advice.

Milen Marinov, **Mark Galatowitsch, Sharyn Goldstein** and **Jon Harding** for help with my sample collection, identification and statistical analysis.

Matt Walters for his endless help and patience when creating images for my thesis.

Gerardo Argüello-Astorga with help identifying motifs and iterons in the viral genomes.

My wonderful parents, **Joanne** and **Pradu,** who have been extremely loving and supportive throughout my time at university. As well, the rest of my family including my brothers **Kiran** and **Nikhil** and my cousin **Iona**.

My late **Grandpa**, who always encouraged my interest in science and to pursue my PhD.

A very big thank you to my virology lab partners **Simona Kraberger** and **Daisy Stainton** who have been a part of my PhD from the beginning and took the time to teach me many things, not always science related. Also thanks to my fellow lab mates **Dorien Coray**, **Ryan Catchpole**, **Laurel Julian** and **Katherine van Bysterveldt** for providing encouragement, insight, high quality lunchtime discussions, endless supplies of baked goods and many laughs. I truly feel privileged to have been able to interact with such amazing people throughout my thesis.

Last of all I would like to thank my close friends: **Kirsty Black**, **Elle Soper**, **Mareesa Maynard, Lisa Lay, Sarah Gunn** and **Jess Yardley** who have all been extremely patient and supportive. Thank you for all the late night meals, coffee dates, retail therapy and encouraging me to take much needed breaks throughout my thesis.

Abstract

The introduction of next-generation sequencing (NGS) technologies has dramatically changed the field of virology, with many significant discoveries of novel circular replication-associated protein (Rep) encoding single-stranded (CRESS) DNA viruses. Traditionally, most research into CRESS DNA viruses has often focused on investigating plant and animal pathogens that are of significant economic importance. This research has led to the discovery and establishment of three different CRESS DNA families including *Geminiviridae, Nanoviridae* and *Circoviridae,* which infect eukaryotes. CRESS DNA viruses can have single or multicomponent genomes, with the latter requiring all components for infection. CRESS DNA viruses have circular single-stranded DNA (ssDNA) genomes with at least one protein encoding a Rep which is responsible for viral replication. It has been shown that CRESS DNA viruses are able to evolve rapidly with nucleotide substitution rates that are similar to those observed in RNA viruses. The Rep gene has conserved regions known as motifs which are often used to determine relatedness between CRESS DNA virus.

NGS has expanded our knowledge on the diversity of novel CRESS DNA viruses. Viral genomes are now routinely recovered from different sample types without any prior knowledge of the viral sequence. This has led to the development of the field of viral ecology. This field places an emphasis on viruses being one of the most abundant organisms on earth, and are therefore likely to play a major role in ecosystems. Environmental metagenomic studies have isolated CRESS DNA viruses from sea water, freshwater, faecal matter from various animals, soil, the atmosphere, sediments and sewage; dramatically increasing the known CRESS DNA viral genomes in the public domain. These studies are shedding light on the distribution of CRESS DNA viruses, as well as providing baseline data for future studies to examine virus-host interactions, community structure and ultimately viral evolution.

Vector enable metagenomics (VEM) is another novel approach utilising NGS techniques for discovering CRESS DNA viruses. As many plant-infecting CRESS DNA viruses such as geminiviruses and nanoviruses are vectored by insects, this approach exploits this mechanism by using insect vectors as a surveillance tool to monitor and survey these viruses circulating in ecosystems. Recent studies have used these methods to identify known viral plant pathogens as well as novel viruses circulating in insect vectors such as whiteflies and other higher order insects such a mosquitoes and dragonflies. These approaches successfully demonstrated that VEM can be used as a unique method, with the first mastrevirus discovered in the new world being recovered from dragonfly species *Erythrodiplax fusca* using this approach.

The research in this thesis uses metagenomics to survey CRESS DNA viral diversity in different organisms and environments. Two hundred and sixty eight novel CRESS DNA viruses were recovered and verified in this study from a range of sample types (adult Odonata, Odonata larvae, Mollusca, benthic sediment, water, Oligochaeta and Chironomidae) collected in the United States of America, Australia and New Zealand. All viral genomes isolated had two major proteins encoding for a putative Rep and coat protein (CP), with major Rep motifs identified in most Reps. Phylogenetic analysis of the Reps encoded by the viral genomes highlighted that most were extremely diverse falling outside of the previously described ssDNA viral families. A top-down approach was implemented to recover CRESS DNA viruses and possible viral pathogens from Odonata and their larvae. Thirty six viral genomes were recovered from terrestrial adult dragonflies as well as the twenty four from aquatic larvae. Dragonfly cycloviruses were isolated from the some adult Odonata species which were closely related to the isolates previously described by Rosario *et al*. (2012). The viruses isolated in the aquatic and terrestrial ecosystems differed substantially indicating that different CRESS DNA viromes exist in both land and water.

The diversity of CRESS DNA viruses in seven different mollusc species (*Amphibola crenata, Austrolvenus stutchburyi, Paphies subtriangulata, Musculium novazelandiae, Potamopyrgus antipodarum, Physella acuta* and *Echyridella menziesi*) from Lake Sarah and the Avon-Heathcote estuary both in New Zealand, were also investigated. One hundred and forty nine novel viral genomes were recovered. Two CRESS DNA genomes were recovered from molluscs which have Rep-like sequences most closely related to those found in some bacterial genomes.

Sclerotinia sclerotiorum hypovirulence-associated DNA virus 1 (SsHADV-1) was originally isolated from fungal species *Sclerotinia sclerotiorum* in china and was later found in benthic sediments in New Zealand. As part of this study, SsHADV-1 was recovered from dragonflies (*Erythemis simplicicollis, Ischnura ramburii* and *Pantala hymenaea*) collected in Arizona and Oklahoma, USA suggesting a larger distribution of these viruses and not surprising given the near global distribution of *S. sclerotiorum.*

Dragonfly larvae-associated circular DNA viruses (DflaCVs) that were originally isolated in Odonata larvae samples from three New Zealand lakes were later recovered from water, benthic sediment, worms and molluscs from one of the lakes initially sampled, suggesting that these viruses are ubiquitous in freshwater environments. This study has attempted to generate baseline data of CRESS DNA viruses in certain environments using NGS-informed approaches. This data was used to try and establish whether viral distribution in different samples types can potentially be explained by the food web interactions between different samples types. Although the analysis did not show any significant relationships between sample type interactions and viral distribution a few common associations between Odonata larvae and benthic sediment were evident. This was expected as the larvae live within the sediment so it could be assumed that they potentially have similar CRESS DNA viral distribution. Although the distribution of viruses varied across sample types, molluscs proved the best sampling tool for isolating largest numbers of CRESS DNA viruses in an ecosystem with extensive diversity.

Overall, this research demonstrates the applications of NGS for investigating the diversity of CRESS DNA viruses. It demonstrates that some sample types such as Odonata in terrestrial systems and molluscs in aquatic environments, can be used as effective sampling tool to determine the diversity of CRESS DNA viruses in different environments as well as detecting previously isolated viruses. The CRESS DNA viruses isolated in this body of work provides baseline data that can potentially be used in future research to investigate hosts of these viruses and their interactions with hosts and potential flow in their environments.

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Deputy Vice-Chancellor's Office
Postgraduate Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter 2

Dayaram, A., Potter, K., Moline, A.B., Rosenstein, D.D., Marinov, M., Thomas, J.E., Breitbart, M., Rosario, K., Argüello-Astorga, G.R., Varsani, A. (2013) High global diversity of cycloviruses amongst dragonflies Journal of General Virology 94: 1827-1840

Please detail the nature and extent (%) of contribution by the candidate:

The samples were collected by were collected as a team effort by Anisha Dayaram, Arvind Varsani, Kristen Potter, Angie Moline and Dana Rosenstein. Anisha Dayaram worked with Gerardo Argüello-Astorga to identify the iterons. Milen Marinov identified the Odonata larvae species.

The all molecular work was undertaken by Anisha Dayaram. Arvind Varsani doubled checked all the bioinformatics analysis and the identification of stem loop strictures and conserved motifs in the Rep.

Anisha Dayaram wrote the manuscript and the rest of the authors provided comments / feedback to improve it.

Contribution by Anisha Dayaram: 85%

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all

The undersigned certifys that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Arvind Varsani* Signature: Date: *12th Dec 2014*

UNIVERSITY OF
CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Deputy Vice-Chancellor’s Office
Postgraduate Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter 3

Dayaram, A., Galatowitsch, M., Harding, J.S., Argüello-Astorga, G.R., Varsani, A. (2014) Novel circular DNA viruses identified in Procordulia grayi and Xanthocnemis zealandica larvae. Infection, Genetics and Evolution 22: 134-141

Please detail the nature and extent (%) of contribution by the candidate:

The samples were collected by were collected as a team effort by Anisha Dayaram and Mark Galatowitsch. Anisha Dayaram worked with Gerardo Argüello-Astorga to identify the iterons. Mark Galatowitsch and Jon Harding identified the Odonata larvae species.

The all molecular work was undertaken by Anisha Dayaram. Arvind Varsani doubled checked all the bioinformatics analysis and the identification of stem loop strictures and conserved motifs in the Rep.

Anisha Dayaram wrote the manuscript and the rest of the authors provided comments / feedback to improve it.

Contribution by Anisha Dayaram: 90%

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all

The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Arvind Varsani* Signature: Date: *12th Dec 2014*

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Deputy Vice-Chancellor's Office
Postgraduate Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter 4

Dayaram, A., Potter, A.A., Pailes, R., Marinov, M., Rosenstein, D.D., Varsani, A (In review) Identification of diverse circular Rep-encoding DNA viruses in adult dragonflies and damselflies (Insecta: Odonata) of Arizona and Oklahoma, USA. Infection, Genetics and Evolution

Please detail the nature and extent (%) of contribution by the candidate:

The samples were collected by were collected as a team effort by Arvind Varsani, Kristen Potter, Roberta Pailes, Angie Moline and Dana Rosenstein. Milen Marinov identified the Odonata larvae species.

The all molecular work was undertaken by Anisha Dayaram. Arvind Varsani doubled checked all the bioinformatics analysis and the identification of stem loop strictures and conserved motifs in the Rep.

Anisha Dayaram wrote the manuscript and the rest of the authors provided comments / feedback to improve it.

Contribution by Anisha Dayaram: 85%

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all

The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Arvind Varsani* Signature: Date: *12th Dec 2014*

Deputy Vice-Chancellor's Office
Postgraduate Office

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter 5

Dayaram, A., Goldstien, S., Zawar-Reza, P., Gomez, C., Harding, J.S., Varsani, A. (2013) Novel single stranded DNA virus recovered from estuarine Mollusc (Amphibola crenata) whose replication associated protein (Rep) shares similarities with Rep-like sequences of bacterial origin. Journal of General Virology. 94: 1083-1089

Please detail the nature and extent (%) of contribution by the candidate:

The samples were collected by were collected bySharyn Goldstien.

The all molecular work was undertaken by Anisha Dayaram. Arvind Varsani doubled checked all the bioinformatics analysis and the identification of stem loop strictures and conserved motifs in the Rep.

Anisha Dayaram wrote the manuscript and the rest of the authors provided comments / feedback to improve it.

Contribution by Anisha Dayaram: 90%

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all

The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Arvind Varsani* Signature: Date: *12th Dec 2014*

UC
UNIVERSITY OF
CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Deputy Vice-Chancellor’s Office
Postgraduate Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter 6

Dayaram, A., Goldstien, S., Argüello-Astorga, G. R., Zawar-Reza, P., Gomez, C., Harding, J.S., Varsani, A. (In review) Diverse circular Rep encoding ssDNA viruses circulating amongst molluscs at the Avon-Heathcote Estuary in Christchurch, New Zealand. Infection, Genetics and Evolution

Please detail the nature and extent (%) of contribution by the candidate:

The samples were collected by were collected by Sharyn Goldstien. Anisha Dayaram worked with Gerardo Argüello-Astorga to identify the iterons.

The all molecular work was undertaken by Anisha Dayaram. Arvind Varsani doubled checked all the *bioinformatics analysis and the identification of stem loop strictures and conserved motifs in the Rep.*

Anisha Dayaram wrote the manuscript and the rest of the authors provided comments / feedback to improve it.

Contribution by Anisha Dayaram: 90%

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all

The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the PhD candidate’s contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Arvind Varsani* Signature: Date: *12th Dec 2014*

Chapter 1

Literature review

Contents

1.1 Overview of viruses

Viruses are small infectious entities that require a living host cell for replication. They are abundant in almost all ecosystems and are known to infect all three major domains of life i.e. bacteria, eukaryotes and archaea (Koonin *et al.*, 2006). There are millions of different types of viruses that have huge morphological diversity (Breitbart & Rohwer, 2005). They can differ in size by up to three orders of magnitude with the Pandora virus being the largest currently characterised and circoviruses being some of the smallest known viruses (Koonin *et al.*, 2006; Philippe *et al.*, 2013).

The genomes of viruses also differ greatly. Viral particles known as virons contain genetic material, either DNA or RNA, that make up the viral genome with some viruses having both a RNA and DNA stage in their lifecycles. Viral genomes can be single stranded (ss) or double stranded (ds) with some ss viruses replicating via a ds intermediate. ssRNA and ssDNA genomes can either be positive or negative sense. Some viruses utilise their host's cell membrane to form a protective layer or lipid envelope while others have a protective coat of protein known as a capsid (Alberts *et al.*, 1998). These viral capsids differ in shape, with the most common forms being icosahedral or helical and in the case of bacteriophages can be an elongated icosahedron (Lidmar *et al.*, 2003).

Viruses are some of the most abundant entities on Earth, and thus have a huge influence on ecosystems (Breitbart & Rohwer, 2005; Duhaime & Sullivan, 2012; Suttle, 2007). Viral ecology is a rapidly expanding field which aims to address the distribution of viruses and how they influence their hosts and surrounding ecosystems. Viral ecology can further provide insights into virus host interactions, community structure, evolution and biogeography of environmental viruses. For years viral studies have relied on investigating viruses that are easily cultured in a laboratory setting often resulting in a bias based on propagation using traditional cell culture techniques. Thus, these techniques cannot reveal the true distribution of viruses in nature and their roles in ecosystems. The introduction of new sequencing methods has influenced this field dramatically, allowing for the first time sequencing of nucleic acids from environmental samples without any prior knowledge of the target sequence (Delwart, 2007; Edwards & Rohwer, 2005). The expanding field of viral metagenomics estimates that less the 1% of the viral diversity present on Earth has been identified (Mokili *et al.*, 2012). Advances in metagenomic sequencing have helped develop further interest in the origins and evolution of viruses and the roles these viruses play in the environment. As viruses are constantly being transported through different environments they are thought to play a major role in evolution through their ability to facilitate horizontal gene transfer (HGT) (Sano *et al.*, 2004).

The new sequencing technologies have helped to explore the diversity of giant viruses, with viruses with larger genomes being described in the last decade. The recent isolation of the largest ever described group of viruses, the Pandora viruses, highlights the changing field of virology. Pandora viruses have particles up to 1.5 µm in length and are visible under light microscopes. Their genomes at 2.8Mb long, larger than some intracellular eukaryotes, encode up to 2,500 proteins. One type of Pandora virus recently isolated from the Siberian permafrost showed similarities to icosahedral DNA viruses, highlighting the diversity of large viruses is yet to be explored (Legendre *et al.*, 2014).

1.1.1 Taxonomy

Traditionally viruses are classified on the basis of similarity using the Baltimore classification system; this is based on the mechanisms viruses use to produce mRNA. However, problems arose using this classification system with viruses as their small genome size and high mutation rate make it very difficult to infer their ancestry beyond the order level (Fenner & Maurin, 1976; Kingsbury, 1985; Van Regenmortel & Mahy, 2004). This led to the formation of the International Committee on the Taxonomy of Viruses (ICTV). This committee provides guidelines for viral classification that puts certain weight on different virus properties when classifying them. This viral classification takes into account viron properties such as morphology, mechanisms of replication and genome organisation (Fenner & Maurin, 1976; King *et al.*, 2011). According to the 9th ICTV report, the ICTV currently recognises seven different orders of viruses, as well as many viral families that have not yet been assigned to a Class (King *et al.*, 2011). Amendments to this taxonomic classification are constantly made to accommodate known and novel viral genomes being discovered annually.

1.2 ssDNA viruses

Next generation sequencing technology has allowed rapid growth in the discovery of novel single-stranded DNA (ssDNA) viruses. Although many ssDNA viruses have been extensively studied and are well characterised, these studies have often been limited to viruses that have a direct impact on humans, agriculture and animals as a consequence of disease. ssDNA viruses currently approved by the ICTV include seven families: *Anelloviridae, Geminiviridae, Inoviridae, Microviridae, Nanoviridae* and *Parvoviridae* (King *et al.*, 2011). However, the true extent of ssDNA viruses diversity is only beginning to be recognised (Rosario & Breitbart, 2011).

There are many well documented cases of pathogenic ssDNA viruses that have been isolated from both eukaryotes and prokaryotes. Many of the novel viruses being discovered have not been assigned to a viral family due to the lack of knowledge surrounding their host range and other various factors. NGS methods have enabled the discovery of many novel ssDNA viruses, however, the hosts of most of these are yet to be identified. The recent proposal of two new viral genera, Cyclovirus and Gemycircularvirus, highlights the changing field of ssDNA virus research, with the addition of many novel ssDNA viral isolates to these genera each year (Rosario *et al.*, 2012b). The new sequencing technologies have allowed, for the first time, the discovery of viruses that were previously unknown. This has led to a rapid expansion in the knowledge surrounding viromes in ecosystems.

ssDNA viral genomes can be circular or linear and can have different component arrangements. Multicomponent genomes such as nanoviruses require multiple components for infection, whilst bipartite genomes such as begomoviruses or single component such as circoviruses require two or one component respectively for infection. Entire ssDNA genomes are usually smaller than 9 kb in size (Martin *et al.*, 2011) and are known to encode between two to eight proteins. Replication of these ssDNA viruses utilises the hosts DNA polymerase to synthesise a new genome via a double-stranded DNA intermediate (Sinsheimer, 1959). ssDNA viruses have been shown to evolve at rates similar to those observed for RNA viruses which is thought to be due to high recombination rates (Lefeuvre *et al.*, 2009; Martin *et al.*, 2011).

All ssDNA viruses have circular genomes with the exception of parvoviruses. Circular ssDNA viruses have been found to infect eukaryotic and prokaryotic organisms. The host range and genome organisation of these viruses have been to divide them into plant-infecting viral families *Nanoviridae* and *Geminiviridae* (King *et al.*, 2011), whereas *Circoviridae, Inoviridae, Microviridae, Parvoviridae* , *Anelloviridae* are known to infect vertebrates, invertebrates and bacteria. Many of the viruses that have been discovered recently currently fall under unclassified circular replication-associated protein (Rep) encoding ssDNA (CRESS) viruses. These represent viruses that encode a Rep which was similar at an amino acid level to other ssDNA viruses. This increasingly large group of CRESS DNA viruses have been recovered from various samples throughout the world and have yet to be classified as most are highly diverse and do not fall within any of the current families (Table 1.1).

ssDNA viruses can be monopartite, bipartite or multipartite. The genome organisation of circular ssDNA viruses differs between genera with some having ambisense genome organisation such as circoviruses, whilst others have negative sense genome organisation such as anelloviruses and gyroviruses. Many of the circular ssDNA viruses have been shown to be evolutionarily related when observing the Rep gene, which is responsible for initiating replication of the virus via the rolling circle replication mechanism (Krupovic *et al.*, 2009a).

Table 1.1: Details of unclassified viruses

Acronym	GenBank accession #	Host/environment	Location	Reference
10-LDMD	KF133817	Ocean water	USA	(McDaniel et al., 2014)
11-LDMD	KF133818	Ocean water	USA	(McDaniel et al., 2014)
12-LDMD	KF133819	Ocean water	USA	(McDaniel et al., 2014)
13-LDMD	KF133820	Ocean water	USA	(McDaniel et al., 2014)
14-LDMD	KF133821	Ocean water	USA	(McDaniel et al., 2014)
15-LDMD	KF133822	Ocean water	USA	(McDaniel et al., 2014)
16-LDMD	KF133823	Ocean water	USA	(McDaniel et al., 2014)
17-LDMD	KF133824	Ocean water	USA	(McDaniel et al., 2014)
18-LDMD	KF133825	Ocean water	USA	(McDaniel et al., 2014)
19-LDMD	KF133826	Ocean water	USA	(McDaniel et al., 2014)
1-LDMD	KF133807	Ocean water	USA	(McDaniel et al., 2014)
20-LDMD	KF133827	Ocean water	USA	(McDaniel et al., 2014)
21-LDMD	KF133828	Ocean water	USA	(McDaniel et al., 2014)
2-LDMD	KF133808	Ocean water	USA	(McDaniel et al., 2014)
3-LDMD	KF133810	Ocean water	USA	(McDaniel et al., 2014)
4-LDMD	KF133811	Ocean water	USA	(McDaniel et al., 2014)
5-LDMD	KF133812	Ocean water	USA	(McDaniel et al., 2014)
6-LDMD	KF133813	Ocean water	USA	(McDaniel et al., 2014)
7-LDMD	KF133814	Ocean water	USA	(McDaniel et al., 2014)
8-LDMD	KF133815	Ocean water	USA	(McDaniel et al., 2014)
9-LDMD	KF133816	Ocean water	USA	(McDaniel et al., 2014)
AtCopCV	JQ837277	Acartiatonsa	USA	(Dunlap et al., 2013)
AnCFV	KJ938716	Caribou stool	Greenland	(Ng et al., 2014)
AnCFV	KJ938716	Caribou stool	Greenland	(Ng et al., 2014)
BamiV	JQ898331	Sewage	USA	(Ng et al., 2012)
BatCV-SC703	JN857329	Bat guano	China	(Ge et al., 2012)
BatCV-TM6C	HM228875	Bat Guano		(Li et al., 2010)
BBC-A	FJ959086	British Columbia coastal waters	Canada	(Rosario et al., 2009b)
BoSCV	JN634851	Bovine stool	South Korea	(Kim et al., 2011)
CasCV	JQ412057	Cassava	Ghana	(Dayaram et al., 2012)
CB-A	FJ959082	Chesapeake Bay, Ocean water	USA	(Rosario et al., 2009b)
CB-B	FJ959083	Chesapeake Bay, Ocean water	USA	(Rosario et al., 2009b)
ChiSCV-DP152	GQ351272	Chimpanzee stool	Cameroon	(Blinkova et al., 2010)
ChiSCV-GM476	GQ351274	Chimpanzee stool	Tanzania	(Blinkova et al., 2010)
ChiSCV-GM488	GQ351276	Chimpanzee stool	Tanzania	(Blinkova et al., 2010)
ChiSCV-GM510	GQ351275	Chimpanzee stool	Tanzania	(Blinkova et al., 2010)
CoCV	AF252610	Pigeon	Germany	(Mankertz et al., 2004)
CynNCKV	JX908740	Cyanoramphusauriceps nest	New Zealand	(Sikorski et al., 2013c)
CynNCXV	JX908739	Cyanoramphusauriceps nest	New Zealand	(Sikorski et al., 2013c)
CyVN-cs1	KF031471	Cerebrospinal fluid	Vietnam	(van Doorn et al., 2013)
CyVN-hcf	KF031466	Cerebrospinal fluid	Vietnam	(van Doorn et al., 2013)
CyVN-hcf1	KF031465	Cerebrospinal fluid	Vietnam	(van Doorn et al., 2013)
CyVN-hcf3	KF031467	Cerebrospinal fluid	Vietnam	(van Doorn et al., 2013)
CyVN-hcf4	KF031468	Cerebrospinal fluid	Vietnam	(van Doorn et al., 2013)
CyVN-hcf5	KF031469	Cerebrospinal fluid	Vietnam	(van Doorn et al., 2013)
CyVN-ps1	KF031470	Cerebrospinal fluid	Vietnam	(van Doorn et al., 2013)
DfaCV-1	JX185430	*Miathyriamarcella*	USA	(Rosario et al., 2012)
DfaCV-2	JX185429	*Erythemissimplicicollis*	USA	(Rosario et al., 2012)
DfaCV-3	JX185428	*Pantalaflavescens*	Tonga	(Rosario et al., 2012)
DfCirV	JX185415	*Pantalaflavescens*	Tonga	(Rosario et al., 2012)
DfCyClV	JX185418	*Pantalaflavescens*	USA	(Rosario et al., 2012)
DflaCV-1	KF738873	*Procorduliagrayi*	New Zealand	(Dayaram et al., 2014)
DflaCV-10	KF738884	*Xanthocnemiszealandic*	New Zealand	(Dayaram et al., 2014)
DflaCV-10	KF738885	*Xanthocnemiszealandic*	New Zealand	(Dayaram et al., 2014)
DflaCV-2	KF738874	*Procorduliagrayi*	New Zealand	(Dayaram et al., 2014)
DflaCV-3	KF738875	*Procorduliagrayi*	New Zealand	(Dayaram et al., 2014)
DflaCV-3	KF738876	*Procorduliagrayi*	New Zealand	(Dayaram et al., 2014)
DflaCV-4	KF738877	*Procorduliagrayi*	New Zealand	(Dayaram et al., 2014)
DflaCV-5	KF738878	*Procorduliagrayi*	New Zealand	(Dayaram et al., 2014)
DflaCV-5a	KF738879	*Procorduliagrayi*	New Zealand	(Dayaram et al., 2014)
DflaCV-6	KF738880	*Procorduliagrayi*	New Zealand	(Dayaram et al., 2014)
DflaCV-7	KF738881	*Procorduliagrayi*	New Zealand	(Dayaram et al., 2014)
DflaCV-8	KF738882	*Procorduliagrayi*	New Zealand	(Dayaram et al., 2014)
DflaCV-9	KF738883	*Procorduliagrayi*	New Zealand	(Dayaram et al., 2014)
DfOrV	JX185416	*Diplacodesbipunctata*	Tonga	(Rosario et al., 2012)
DfOrV	JX185417	*Diplacodesbipunctata*	Tonga	(Rosario et al., 2012)
Diporeiasp-CV	KC248416	Diporeia sp.	USA	(Hewson et al., 2013)
EeCV	KC469701	Eel	Hungary	(Doszpoly et al., 2014)
FaGmCV-10	KF371632	*Sturnus vulgaris* feces	New Zealand	(Sikorski et al., 2013d)
FaGmCV-11	KF371631	*Oryctolaguscuniculus* feces	New Zealand	(Sikorski et al., 2013b)
FaGmCV-12	KF371630	*Struthiocamelus* feces	New Zealand	(Sikorski et al., 2013b)
FaGmCV-1a	KF371643	*Ovisaries* feces	New Zealand	(Sikorski et al., 2013b)
FaGmCV-1b	KF371642	*Turdusmerula* feces	New Zealand	(Sikorski et al., 2013b)
FaGmCV-1c	KF371641	*Turdusmerula* feces	New Zealand	(Sikorski et al., 2013b)
FaGmCV-2	KF371640	*Susscrofa* feces	New Zealand	(Sikorski et al., 2013b)
FaGmCV-3	KF371639	*Gerygonealbofrontata* feces	New Zealand	(Sikorski et al., 2013b)
FaGmCV-4	KF371638	*Arctocephalusforsteri* feces	New Zealand	(Sikorski et al., 2013b)
FaGmCV-5	KF371637	*Gerygonealbofrontata* feces	New Zealand	(Sikorski et al., 2013b)
FaGmCV-6	KF371636	*Gerygonealbofrontata* feces	New Zealand	(Sikorski et al., 2013b)
FaGmCV-7	KF371635	*Anasplatyrhynchos* feces	New Zealand	(Sikorski et al., 2013b)
FaGmCV-8	KF371634	*Petroicatraversi* feces	New Zealand	(Sikorski et al., 2013b)
FaGmCV-9	KF371633	*Turdusmerula* feces	New Zealand	(Sikorski et al., 2013b)
FdNV	KC441518	Pink shrimp	USA	(Ng et al., 2013)
FSfaCV	KF246569	*Arctocephalusforsteri* feces	New Zealand	(Sikorski et al., 2013b)
GasCSV	KC172652	*Amphibolacrenata*	New Zealand	(Dayaram et al., 2013)
GOM00012	JX904192	Ocean water	Gulf of Mexico	(Labonté and Suttle, 2013)
GOM00443	JX904231	Ocean water	Gulf of Mexico	(Labonté and Suttle, 2013)
GOM00546	JX904245	Ocean water	Gulf of Mexico	(Labonté and Suttle, 2013)

Acronym	GenBank accession #	Host/environment	Location	Reference
GOM02856	JX904312	Ocean water	Gulf of Mexico	(Labonté and Suttle, 2013)
GOM02962	JX904333	Ocean water	Gulf of Mexico	(Labonté and Suttle, 2013)
GOM03041	JX904344	Ocean water	Gulf of Mexico	(Labonté and Suttle, 2013)
GOM03161	JX904368	Ocean water	Gulf of Mexico	(Labonté and Suttle, 2013)
GOM03193	JX904377	Ocean water	Gulf of Mexico	(Labonté and Suttle, 2013)
hs1	JX559621	Human feces	Brazil	(Castrignano et al., 2013)
hs2	JX559622	Human feces	Brazil	(Castrignano et al., 2013)
LaCopCV	JQ837277	*Acartiatonsa* copepod	USA	(Dunlap et al., 2013)
LaCopCV	JF912805	*Labidoceraaestiva*	USA	(Dunlap et al., 2013)
LM28925	KC248425	Diporeia sp.	USA	(Hewson et al., 2013)
MmFV	JN704610	Melesmeles feces	Netherlands	(van den Brand et al., 2011)
MpaCDV-1	KJ547646	Freshwater Pond	Antarctica	(Zawar-Reza et al., 2014)
MpaCDV-2	KJ547647	Freshwater Pond	Antarctica	(Zawar-Reza et al., 2014)
MpaCDV-3	KJ547648	Freshwater Pond	Antarctica	(Zawar-Reza et al., 2014)
MpaCDV-4	KJ547649	Freshwater Pond	Antarctica	(Zawar-Reza et al., 2014)
MpaCDV-5	KJ547650	Freshwater Pond	Antarctica	(Zawar-Reza et al., 2014)
MpaCDV-6	KJ547651	Freshwater Pond	Antarctica	(Zawar-Reza et al., 2014)
MpaCDV-7	KJ547652	Freshwater Pond	Antarctica	(Zawar-Reza et al., 2014)
MpaCDV-8	KJ547653	Freshwater Pond	Antarctica	(Zawar-Reza et al., 2014)
MS584-5	HQ322117	Picobiliphyte sp.	USA	(Yoon et al., 2011)
MvemV	HQ335086	Mosquito	USA	(Ng et al., 2011)
NephV	JQ898333	Sewage	USA	(Ng et al., 2012)
NimiV	JQ898332	Sewage	USA	(Ng et al., 2012)
PigSCV	JX274036	Porcine feces	New Zealand	(Sikorski et al., 2013a)
PmCV-1	KF481961	Shrimp	Vietnam	(Pham et al., 2014)
PisaCV ANH1	JX305997	Porcine stool	China	(Song et al. 2012)
PisaCV FUJ1	JX305998	Porcine stool	China	(Song et al. 2012)
PisaCV GER2011	JQ023166	Porcine stool	China	(Song et al. 2012)
PisaCV HEN1	JX305991	Porcine stool	China	(Song et al. 2012)
PisaCV HUB1	JX305992	Porcine stool	China	(Song et al. 2012)
PisaCV HUB2	JX305993	Porcine stool	China	(Song et al. 2012)
PisaCV HUN1	JX305995	Porcine stool	China	(Song et al. 2012)
PisaCV HUN2	JX305996	Porcine stool	China	(Song et al. 2012)
PisaCV JIANGX1	JX305994	Porcine stool	China	(Song et al. 2012)
po-circo-like21	JF713716	*Susscrofa* feces	USA	(Shan et al., 2011)
po-circo-like22	JF713717	*Susscrofa* feces	USA	(Shan et al., 2011)
po-circo-like41	JF713718	*Susscrofa* feces	USA	(Shan et al., 2011)
po-circo-like51	JF713719	*Susscrofa* feces	USA	(Shan et al., 2011)
PoSCV2	KC545226	*Susscrofa* feces	USA	(Cheung, 2004)
PoSCV3	KC545226	*Susscrofa* feces	USA	(Cheung, 2004)
PoSCV3	KC545227	*Susscrofa* feces	USA	(Cheung, 2004)
PoSCV3	KC545227	*Susscrofa* feces	USA	(Cheung, 2004)
PoSCV3	KC545228	*Susscrofa* feces	USA	(Cheung, 2004)
PoSCV3	KC545229	*Susscrofa* feces	USA	(Cheung, 2004)
PoSCV3	KC5452309	*Susscrofa* feces	USA	(Cheung, 2004)
PoSCV 3L2T	KC545230	Porcine stool	USA	(Cheung, et al., 2013)
PoSCV-1 DP2	KJ577810	Porcine stool	USA	(Cheung, et al., 2013)
PoSCV-1 DP3	KJ577811	Porcine stool	USA	(Cheung, et al., 2013)
PoSCV-2 TP3	KJ577818	Porcine stool	USA	(Cheung, et al., 2013)
PoSCV-6 XP1	KJ577819	Porcine stool	USA	(Cheung, et al., 2013)
PoSCV-7 EP2-A	KJ577812	Porcine stool	USA	(Cheung, et al., 2013)
PoSCV-7 EP2-B	KJ577813	Porcine stool	USA	(Cheung, et al., 2013)
PoSCV-7 EP3-C	KJ577814	Porcine stool	USA	(Cheung, et al., 2013)
PoSCV-7 EP3-D	KJ577815	Porcine stool	USA	(Cheung, et al., 2013)
PoSCV-8 GP2	KJ577817	Porcine stool	USA	(Cheung, et al., 2013)
PoSCV-9 FP1	KJ577816	Porcine stool	USA	(Cheung, et al., 2013)
PoSCV-Kor J481	KF193403	Porcine stool	USA	(Cheung, et al., 2013)
RodSCV-M-13	JF755410	*Peromyscustruei* feces	USA	(Phan et al., 2011)
RodSCV-M-44	JF755408	*Musmusculus* feces	USA	(Phan et al., 2011)
RodSCV-M-45	JF755409	*Musmusculus* feces	USA	(Phan et al., 2011)
RodSCV-M-53	JF755413	*Microtuspennsylvanicus*	USA	(Phan et al., 2011)
RodSCV-M-89	JF755402	*Microtuspennsylvanicus*	USA	(Phan et al., 2011)
RodSCV-R-15	JF755401	*Neotomacinerea*	USA	(Phan et al., 2011)
RodSCV-V-64	JF755407	*Microtuspennsylvanicus*	USA	(Phan et al., 2011)
RodSCV-V-69	JF755403	*Microtuspennsylvanicus*	USA	(Phan et al., 2011)
RodSCV-V-72	JF755411	*Microtuspennsylvanicus*	USA	(Phan et al., 2011)
RodSCV-V-76	JF755404	*Microtuspennsylvanicus*	USA	(Phan et al., 2011)
RodSCV-V-77	JF755415	*Musmusculus feces*	USA	(Phan et al., 2011)
RodSCV-V-81	JF755412	*Microtuspennsylvanicus*	USA	(Phan et al., 2011)
RodSCV-V-84	JF755413	*Microtuspennsylvanicus*	USA	(Phan et al., 2011)
RodSCV-V-86	JF755416	*Microtuspennsylvanicus*	USA	(Phan et al., 2011)
RodSCV-V-87	JF755406	*Microtuspennsylvanicus*	USA	(Phan et al., 2011)
RodSCV-V-91	JF755417	*Microtuspennsylvanicus*	USA	(Phan et al., 2011)
RodSCV-V-97	JF755414	*Microtuspennsylvanicus*	USA	(Phan et al., 2011)
RW-A	FJ959077	Reclaimed water	USA	(Rosario et al., 2009a)
RW-B	FJ959078	Reclaimed water	USA	(Rosario et al., 2009a)
RW-C	FJ959079	Reclaimed water	USA	(Rosario et al., 2009a)
RW-D	FJ959080	Reclaimed water	USA	(Rosario et al., 2009a)
RW-E	FJ959081	Reclaimed water	USA	(Rosario et al., 2009a)
SaCV-10	KJ547621	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-11	KJ547622	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-12	KJ547623	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-13	KJ547624	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-14	KJ547625	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-15	KM821750	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-16	KM821751	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-17	KM821752	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-18	KM821753	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-19	KM821754	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-2	KJ547626	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-2	KJ547626	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-20	KM821755	Sewage	New Zealand	(Kraberger et al., 2014)

Acronym	GenBank accession #	Host/environment	Location	Reference
SaCV-22	KM821757	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-23	KM821758	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-24	KM821759	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-25	KM821760	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-26	KM821761	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-27	KM821762	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-28	KM821763	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-29	KM821764	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-3	KJ547627	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-3	KJ547627	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-30	KM821765	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-31	KM821766	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-32	KM821767	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-33	KM821768	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-34	KM821769	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-35	KM821770	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-36	KM821748	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-37	KM821749	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-4	KJ547628	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-4	KJ547628	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-5	KJ547629	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-6	KJ547630	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-7	KJ547631	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-8	KJ547632	Sewage	New Zealand	(Kraberger et al., 2014)
SaCV-9	KJ547633	Sewage	New Zealand	(Kraberger et al., 2014)
SaGmV-10a	KJ547644	Sewage	New Zealand	(Kraberger et al., 2014)
SaGmV-10b	KJ547645	Sewage	New Zealand	(Kraberger et al., 2014)
SaGmV-12	KJ547642	Sewage	New Zealand	(Kraberger et al., 2014)
SaGmV-2	KJ547642	Sewage	New Zealand	(Kraberger et al., 2014)
SaGmV-3	KJ547643	Sewage	New Zealand	(Kraberger et al., 2014)
SaGmV-4	KJ547634	Sewage	New Zealand	(Kraberger et al., 2014)
SaGmV-5	KJ547635	Sewage	New Zealand	(Kraberger et al., 2014)
SaGmV-6	KJ547636	Sewage	New Zealand	(Kraberger et al., 2014)
SAR-A	FJ959084	Sargasso Sea	USA	(Rosario et al., 2009a)
SAR-B	FJ959085	Sargasso Sea	USA	(Rosario et al., 2009a)
SDWAPI	HQ335042	Mosquito	USA	(Ng et al., 2011)
ShrimpCDV	KC441518	*Farfantepenaeusduorarum*	USA	(Ng et al., 2013)
SI00003	JX904394	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI00006	JX904395	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI00063	JX904401	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI00078	JX904407	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI00094	JX904412	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI00142	JX904416	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI00197	JX904420	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI00349	JX904427	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI00373	JX904431	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI00441	JX904439	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI00793	JX904469	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI00850	JX904473	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI00898	JX904478	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI01664	JX904518	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI01813	JX904523	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI03513	JX904541	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI03654	JX904548	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI03701	JX904559	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI03705	JX904561	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI03717	JX904562	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI03931	JX904581	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SI04276	JX904605	SaanichInle, Ocean water	Canada	(Labonté and Suttle, 2013)
SiCV-H5	JQ011377	European catfish	Hungary	(Lőrincz et al., 2012)
SiCV-H6	JQ011378	European catfish	Hungary	(Lőrincz et al., 2012)
SOG00160	JX904075	Strait of Georgia, Ocean water	Canada	(Labonté and Suttle, 2013)
SOG00164	JX904076	Strait of Georgia, Ocean water	Canada	(Labonté and Suttle, 2013)
SOG00182	JX904077	Strait of Georgia, Ocean water	Canada	(Labonté and Suttle, 2013)
SOG00781	JX904107	Strait of Georgia, Ocean water	Canada	(Labonté and Suttle, 2013)
SOG03994	JX904139	Strait of Georgia, Ocean water	Canada	(Labonté and Suttle, 2013)
SOG04070	JX904144	Strait of Georgia, Ocean water	Canada	(Labonté and Suttle, 2013)
SOG04106	JX904147	Strait of Georgia, Ocean water	Canada	(Labonté and Suttle, 2013)
SOG04311	JX904151	Strait of Georgia, Ocean water	Canada	(Labonté and Suttle, 2013)
SOG05268	JX904185	Strait of Georgia, Ocean water	Canada	(Labonté and Suttle, 2013)
SsHADV-1	KF268025	River benthic sediments	New Zealand	(Kraberger et al., 2013)
SsHADV-1	GQ365709	*Sclerotiniasclerotiorum*	China	(Yu et al., 2010)
SsHADV-1	KF268026	Benthic sediments	New Zealand	(Kraberger et al., 2013)
SsHADV-1	KF268027	Benthic sediments	New Zealand	(Kraberger et al., 2013)
SsHADV-1	KF268028	Benthic sediments	New Zealand	(Kraberger et al., 2013)
YN-BtCV-1	JF938078	Bat	China	(Ge et al., 2011)
YN-BtCV-2	JF938079	Bat	China	(Ge et al., 2011)
YN-BtCV-3	JF938080	Bat	China	(Ge et al., 2011)
YN-BtCV-4	JF938081	Bat	China	(Ge et al., 2011)
YN-BtCV-5	JF938082	Bat	China	(Ge et al., 2011)

1.2.1 Family *Circoviridae*

The *Circoviridae* family, known to infect vertebrates, include two genera: *Gyrovirus,* which consists of one species, *Chicken anaemia virus* (CAV); and *Circovirus,* which currently recognises 11 species that infect birds and pigs (Biagini, 2012; Li *et al.*, 2010b). However, new taxonomic changes have recently been proposed for *Circoviridae*. This will include the addition of the genus *Cyclovirus*, which will include many viruses that have been discovered through various metagenomic studies and will be assigned based on their phylogenetic relatedness, genome architecture and similarities to other circoviruses (Rosario *et al.*, 2011). It was first thought that circoviruses were non-pathogenic; however, studies have subsequently demonstrated that some strains have detrimental effects on porcine, avian and many other species (Abadie *et al.*, 2001; Allan & Ellis, 2000; Ellis *et al.*, 2003; Niagro *et al.*, 1998). Many of these circoviruses share close sequence identities suggesting possible zoonotic transmission of these viruses between different species, which often leads to an increase in their pathogenicity and host range (Li *et al.*, 2011a).

Circoviruses have compact ssDNA genomes ranging from ~1.7 to ~2.3 kb (Biagini, 2012). They have an ambisense genome organisation with two major open reading frames (ORFs) that encode a Rep and a coat protein (CP). Some have a small intergenic region (SIR) and a long intergenic region (LIR) that contains a stem-loop structure. The stem-loop has a highly conserved nonanucleotide motif (TAGTATTAC) where rolling circle replication (RCR) initiates (Ilyina & Koonin, 1992).

Circoviruses are particularly dependent on actively dividing host cells and utilise the hosts DNA polymerase for its own replication (Biagini, 2012). Once inside the nucleus of the host cell, the ssDNA genome is used as a template for the formation of a complementary strand of DNA, making a double-stranded DNA (dsDNA) intermediate, also known as the replicative form (RF). The RF is then used as a template for the production of mRNA, with ORFs potentially on both the original virion strand (V) and also on the complementary strand (C) (Biagini, 2012).

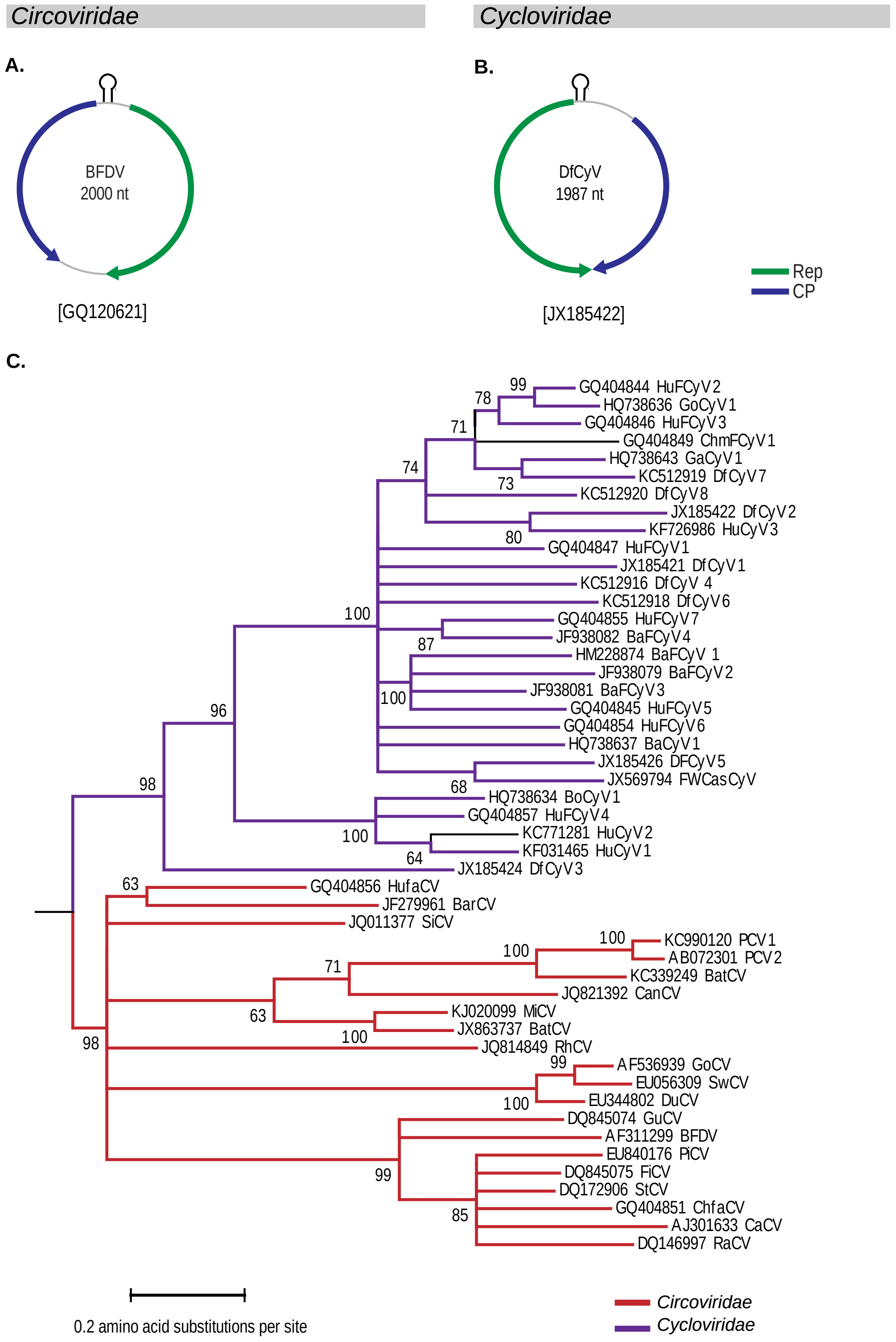


Figure 1.1: Genome organisation of *Circoviridae* and *Cycloviridae.* (**A**) Circovirus: *Beak and feather disease virus* (GQ120621). (**B**) Cyclovirus: *Dragonfly cyclovirus* (JX185422). (**C**) Maximum likelihood phylogenetic tree of the Rep amino acid sequences of representative circoviruses (red) and cycloviruses (purple) inferred using rtREV substitution model with 100 replicate bootstrap and rooted with MSV and TYLCV geminivirus representatives

1.2.1.1 Genus Circovirus

There are currently 12 circovirus species recognised by the ICTV, inlcuding porcine circovirus (PCV) -1 and PCV-2, which infect pigs, whilst the others infect different avian species (*Beak and feather disease virus,* BFDV; *Canary circovirus,* CaCV; *Duck circovirus,* DuCV; *Finch circovirus,* FiCV; *Goose circovirus*, GoCV; *Gull circovirus*, GuCV; *Pigeon circovirus*, PiCV; *Raven circovirus,* RaCV; *Starling circovirus*, StCV; and *Swan circovirus*, SwCV) (Biagini, 2012). Several novel circoviruses have recently been isolated from other animals including dogs (Kapoor *et al.*, 2012), fish (Lőrincz *et al.*, 2011; Lőrincz *et al.*, 2012), bat guano (Li *et al.*, 2010a) and primates, as well as human faecal matter (Blinkova *et al.*, 2010; Li *et al.*, 2010b). This suggests that the diversity and host range of circoviruses may be much larger than previously thought (Table 1.2).

1.2.1.1.1 Porcine circovirus 1 and 2

The most extensively studied circoviruses are PCV-1 and PCV-2 (Allan & Ellis, 2000; Mankertz *et al.*, 2004). Circoviruses in pigs were originally believed to be non-pathogenic (Tischer *et al.*, 1982). This is because the first porcine circovirus (PCV) isolated was a non-pathogenic strain associated with cultured porcine kidneys (PK-15) cells (Tischer *et al.*, 1982). It was subsequently found that many other strains of PCV such as PCV-2 are associated with post-weaning multisystemic wasting syndrome (PMWS) in pigs (Allan *et al.*, 1998; Hamel *et al.*, 1998; Krakowka *et al.*, 2001; Meehan *et al.*, 1998; Morozov *et al.*, 1998).

PCV-1 is thought to be non-pathogenic to porcine whilst a PCV-2 infection can cause a variety of symptoms in swine; including PMWS, porcine dermatitis and nephropathy syndrome (PDNS), reproductive failure, and respiratory disease complex. The effects of PCV-2 therefore have significant economic implications on pig farming (Chae, 2005; Finsterbusch & Mankertz, 2009; Todd *et al.*, 2001a). PMWS was first described in pigs in Canada and infects numerous organs leading to a multitude of symptoms including: progressive weight loss, respiratory signs and jaundice. The disease also results in lesions including granulomatous interstitial pneumonia, lymphadenopathy, and lymphocytic granulomatous hepatitis and nephritis (Allan *et al.*, 1998; Ellis *et al.*, 1998).

The genome size of PCV is ~1.7 kb has two major ORFs; one encodes for a Rep protein and has some conserved motifs. The other ORF shares less similarity with other circoviruses and is thought to encode for a coat protein (Morozov *et al.*, 1998).

1.2.1.1.2 Avian circoviruses

Circoviruses have been discovered in a range of avian species. Many of these viruses are thought to have deleterious consequences and have been linked to deformities of the beak, claws and feathers, immunosuppression, lymphoid depletion and growth retardation (Abadie *et al.*, 2001; Niagro *et al.*, 1998; Todd, 2000, 2004). Circoviruses have been isolated from a range of avian species including: ducks, chickens, starlings, finch, ravens, pigeons, canaries, gulls and swans (Banda *et al.*, 2007; Chen *et al.*, 2006; Hattermann *et al.*, 2003; Johne *et al.*, 2006; Stewart *et al.*, 2006; Todd *et al.*, 2007; Todd *et al.*, 2001b).

BFDV is one of the most extensively studied avian circovirus, it is known to cause Psittacine beak and feather disease (PBFD) in birds from the Psittaciformes order. This virus has been documented to have a huge impact on both wild and captive populations of birds and has a cosmopolitan distribution which is partly a consequence of the international trade of parrots (Todd, 2004). This virus causes varying degrees of disease in from pre-acute to chronic (Doneley, 2003). The disease symptoms often begin with feather abnormalities, followed by symptoms such as depression, lethargy, diarrhoea and deformities in the beaks likely to develop (Huff *et al.*, 1988; Raidal, 1995; Ritchie *et al.*, 1989). Like PCV, the genome of BFDV has an ambisense organisation with two major ORFs encoding for a Rep and CP (Biagini, 2012).

CoCV was first isolated from pigeons in Germany using degenerate primers (Mankertz *et al.*, 2000). The genome of CoCV is ~2 kb and has two major ORFs that are bi-directionally transcribed and contains the conserved nonanucleotide motif “TAGTATTAC”. Phylogenetic analysis shows that the virus falls within the *Circoviridae* family. A further studied examined if CoCV was responsible for disease in pigeons, with pathological findings from this study suggesting that CoCV weakens the immune system of pigeons making them more susceptible to infection and disease (Soike *et al.*, 2001).

GoCV was isolated from bursal homogenate samples taken from geese in Germany (Todd *et al.*, 2001a) and has been identified in geese outside of Europe (Chen *et al.*, 2003). This virus has a ~1.8 kb genome with similar genome organisation to previously described avian circoviruses. GoCV is believed to be associated with increased mortality by affecting growth and development and causes feathering disorders in geese leading to similar symptoms as observed with BFDV. GoCV is able to be detected with both dot blot hybridisation and PCR methods from infected geese samples (Ball *et al.*, 2004).

RaCV is a recently classified avian circovirus (Stewart *et al.*, 2006) reovered from the Australian raven *Corvuscoronoides.* The 1.8 kb genome of RaCV has the same organisation as other circoviruses with two major ORFs that are bi-directionally transcribed encode for both a Rep and CP. Phylogenetic analysis of RaCV showed it shared closest homology to CaCV and PiCV and was more distantly related to other avian circoviuses and PCV (Stewart *et al.*, 2006).

1.2.1.1.3 Bat circovirus

Bats make up one of the largest mammalian populations in the world and act as a reservoir for emerging viral diseases (Calisher *et al.*, 2006). Two species of viruses, Bat circovirus 1 (BatCV-1) and Bat circovirus 2 (BatCV-2), have been isolated from the tissue of bats located in Myanmar and China (He *et al.*, 2013). The genomes of these virus species are between ~1.7-2 kb and have the same genome organisation to other characterised circoviruses with two major ORFs that are bi-directionally transcribed. Both isolated also contain the nonanucleotide motif “TAGTATTAC” in the 5’ intergenic region. Phylogenetic analysis of the Rep protein showed that BatCV-2 (Figure 1.1) grouped with PCV whilst BatCV-1 was more distantly related to PCV but still fell within the *Circoviridae* family (Calisher *et al.*, 2006). Because of the similar genome architecture and phylogenetic relatedness of BatCV-1 and BatCV-2 to circoviruses, it was recently proposed that both these viral isolates be formally classified within the *Circoviridae* family.

1.2.1.1.4 Canine circovirus

Canine circovirus 1 (CaCV-1) is the first circovirus to be isolated from dogs (Kapoor *et al.*, 2012). The virus is genetically most closely related to porcine circoviruses. It has a ~2 kb circular ssDNA genome with a putative stem loop structure with the rolling circle replication nonanucleotide motif TAGTATTAC. It has two major ORFs that are bi-directionally transcribed encoding the Rep and CP (Kapoor *et al.*, 2012).

1.2.1.1.5 Barbel circovirus

Two isolates of *Barbel circovirus* (BarCV) were recently isolated from *Barbus barbus* fish in Hungary (Lőrincz *et al.*, 2011). The genome of both isolates was ~1.9 kb. Phylogenetic analysis of the two genomes and their proteins showed they shared close sequence similarities with the family *Circoviridae* (Figure 1.1). It was recently proposed that BarCV be formally classified within the family *Circoviridae*. BarCV is linked to the high mortality of barbel fish being observed in Europe although the exact significance of the virus is yet to be determined (Lőrincz *et al.*, 2011).

1.2.1.1.6 Silurusglanis circovirus

Silurus glanis circovirus (SiCV) was isolated from individuals of the European catfish species *Silurus glanis* (Lőrincz *et al.*, 2012). These viruses are linked to the high mortality of European cat fish that were found dead during the spawning season in Lake Balaton in Hungary in 2011. Characterisation of the ~1.9 kb genome showed the organisation was similar to that of other circoviruses and also contained the conserved nonanucleotide motif "TAGTATTAC" in the 5' intergenic region. Further phylogenetic analysis illustrated that SiCV falls within the *Circoviridae* family and is distantly related to BarCV (Figure 1.1) (Lőrincz *et al.*, 2012).

1.2.1.1.7 Unclassified circoviruses

Circovirus-like sequences have also been identified from the stools of a diverse range of animals including: chimpanzees, fish, badgers, pine martens, marine copepods, eels cows, rodents, insects, birds, bats and humans (Doszpoly *et al.*, 2014; Dunlap *et al.*, 2013; Ge *et al.*, 2011; Kim *et al.*, 2011; Li *et al.*, 2010b; Lőrincz *et al.*, 2011; Ng *et al.*, 2011b; Phan *et al.*, 2011; van den Brand *et al.*, 2011). Members of new family of *Cycloviridae* were discovered in stools from humans and chimpanzees from around the world. It was hypothesised that these cycloviruses may be present as a result of the consumption of meat products. Although cycloviruses were detected in meat samples from farm animals such as sheep, goats, cows, chickens and camels, they shared low similarity to the strains isolated from the human and chimpanzee stools (Li *et al.*, 2011a; Li *et al.*, 2010b). The new cyclovirus family is considerably different from the circoviruses. Cyclovirus genomes are much smaller than circoviruses at just ~1.7-1.8 kb. The major ORFs encoding for the Rep and CP are also smaller and the 3’ intergenic region between the major ORFs is either just a few base pairs long or completely absent (Li *et al.*, 2010b). The Rep of cycloviruses displays RCR and helicase motifs as well at the nonanucleotide motif “TAGTATTAC”, indicating recent common ancestry with the circoviruses.

Table 1.2: Table showing the current classified species in the genus *Circovirus.*

Acronym	GenBank accession #	host/environemnt	Location	Reference
BFDV	AF071878	Pscittacine birds	USA	(Niagro *et al.*, 1998)
CaCV	AJ301633	Canary	United Kingdom	(Todd *et al.*, 2001b)
DuCV	AY228555	Malard Duck	Germany	(Hattermann *et al.*, 2003)
DuCV	AY394721	Muscovy duck	Taiwan	(Chen *et al.*, 2006)
PiCV	AJ298229	Pigeon	United Kingdom	(Todd *et al.*, 2001a)
GoCV	AJ304456	Goose	United Kingdom	(Todd *et al.*, 2001a)
FiCV	DQ845075	Gouldian finch	United Kingdom	(Todd *et al.*, 2007)
GuCV	DQ845074	Herring gull	United Kingdom	(Todd *et al.*, 2007)
PCV-1	AF071879	Porcine	USA	(Niagro *et al.*, 1998)
StCV	DQ172906	Starling	Spain	(Johne *et al.*, 2006)
SwCV	EU056310	Swan	Germany	(Halami *et al.*, 2008)
PCV-2	AY424401	Porcine	Austria	(Exel, 2003)
BatCV - 1	JX863737	Bat	China/Myanmar	(He *et al.*, 2013)
RhCV	JQ814849	Bat	China	(Wu *et al.*, 2012)
BatCV-2	KC339249	Bat	China/Myanmar	(He *et al.*, 2013)
RaCV	DQ146997	Raven	Australia	(Stewart *et al.*, 2006)
HufaCV	GQ404856	Human stool	Nigeria	(Li *et al.*, 2010b)
ChfaCV	GQ404851	Chimpanzee stool	Middle Africa	(Li *et al.*, 2010b)
BarCV	GU799606	Fish	Hungary	(Lőrincz *et al.*, 2011)
MiCV	KJ020099	Mink	China	(Lian, 2014)
CanCV	KC241982	Dog	USA	(Li *et al.*, 2013)

1.2.1.2 Genus Cyclovirus

With the development of new metagenomic techniques and high-throughput sequencing, many novel ssDNA viruses have been discovered with low sequence similarity to circoviruses. Furthermore, studies using viral metagenomic approaches have uncovered many ssDNA viruses from different environments. Some of these viruses show some similarity to circoviruses, particularly in Rep amino acid sequences; however they differ in genome organisation. It was proposed by Li *et al*. (2010) that these viruses be classified in the new genus *Cyclovirus* within the *Circoviridae* family. The first cyclovirus was isolated from the faecal samples of children using metagenomic approaches (Victoria *et al.*, 2009). There are now many members of the genus Cyclovirus (Table 1.3).

The genome organisation of cycloviruses are between ~1.7-2 kb and have two major bi-directionally transcribed ORFs that encode a putative Rep and CP. The Rep in cycloviruses is in the complimentary sense and the CP in the virion sense , whereas it is *vice versa* for the circoviruses (Figure 1.1). Cycloviruses have a LIR that is usually larger than that of circoviruses; however, the SIR is usually only a few nucleotides or completely absent (Delwart & Li, 2011; Li *et al.*, 2011a; Li *et al.*, 2010b). In the LIR of cycloviruses, just like circoviruses, is a stem-loop structure where rolling circle replication is initiated; this region usually contains a highly conserved nonanucleotide motif (TAATATTAC) which differs slightly to that of circoviruses (TAGTATTAC) (Rosario *et al.*, 2012b).

A recent study identified a novel cyclovirus, Florida woods cockroach–associated cyclovirus GS140 (FWCasCyV-GS140), from *Eurycotis floridana* (Padilla-Rodriguez *et al.*, 2013). FWCasCyV-GS140 had a ~1.7 kb. The FWCasCyV-GS140 genome also has a 5’ intergenic region separating the two ORFs. This region contains the conserved nonanucleotide motif “TAGTATTAC” in the stem-loop structure which differs to other cycloviruses. Phylogenetic analysis of FWCasCyV-GS140 showed it shared similarities to other cyclovirus sequences and had 64% genome wide pairwise identity with a previously described cyclovirus isolated from insectivorous bat guano (Padilla-Rodriguez *et al.*, 2013).

Many recent studied have focused on identifying cycloviruses present in different human samples (de Jong *et al.*, 2014; Li *et al.*, 2010b; Phan *et al.*, 2014; Smits *et al.*, 2013). One study looked at respiratory tract infections of Chilean infants and identified a novel cyclovirus CyCV-ChileNPA1 from nasopharyngeal aspirates (Phan *et al.*, 2014). The genome of CyCV-ChileNPA1 (KF726986, HuCyV3) is ~1.7 kb and has a similar genome organisation to other cycloviruses with an IR region containing a stem-loop structure with the nonanucleotide motif “TAGTATTAC”. Phylogenetic analysis showed it is grouped with other cycloviruses (Figure 1.1). CyCV-ChileNPA1 was isolated from both upper and lower respiratory tracts of children indicating that this virus may play a role in the respiratory illness of these children, although further investigation is needed to access potential hosts and routes of transmission (Phan *et al.*, 2014).

Cycloviruses have also been identified in cerebrospinal fluid (CSF) samples. One of the first studies isolated a novel cyclovirus from the CSF taken from patients in Malawi with unexplained paraplegia (Smits *et al.*, 2013). The virus isolated was named human cyclovirus VS5700009 (KC771281, HuCyV2) and had a similar genome organisation to the previously described human cycloviruses TN18 and TN25 (GQ404857, HuFCyV4) isolated from human stools (Li *et al.*, 2010b). Further phylogenetic analysis of the Rep showed it shared ~75% amino acid identity with these viruses (Figure 1.1) (Li *et al.*, 2010b). In a similar study investigating acute central nervous system (CNS) infections in patients from Vietnam, a novel cyclovirus-Vietnam (CyCV-VN) was isolated from CSF samples (van Doorn *et al.*, 2013). The genome of CyCV-VN was ~1.8 kb and showed similar genome organisation to previously described cycloviruses. Phylogenetic analysis showed that CyCV-VN was divergent to other cycloviruses and again most closely related to cycloviruses TN18 and TN25 (Figure 1.1). A further study looked at the geographic distribution of CyCV-VN across Vietnam, Cambodia, Nepal and The Netherlands, but all screening results were negative (de Jong *et al.*, 2014). The study also conducted sequence comparison and phylogenetic analysis between CyCV-VN and CyCVVS5700009. This suggests that these cycloviruses represent divergent species that are phylogenetically closely related and further demonstrates the diversity within the genus cyclovirus (Figure 1.1) (de Jong *et al.*, 2014). The discovery of many new novel Cyclovirus from these human samples raises questions about their pathogenicity, geographic distribution and potential for zoonosis where further research is needed to address these issues.

Subsequent to the first discovery of cycloviruses, they have since been identified in faecal and tissue samples from bats (Ge *et al.*, 2011; Li *et al.*, 2010a; Li *et al.*, 2011a), tissue from the abdomen of dragonflies (Rosario *et al.*, 2012b; Rosario *et al.*, 2011) , human and chimpanzee faecal matter (Li *et al.*, 2011a) and the tissue of goats, sheep, cows, camels and chickens (Li *et al.*, 2011a).

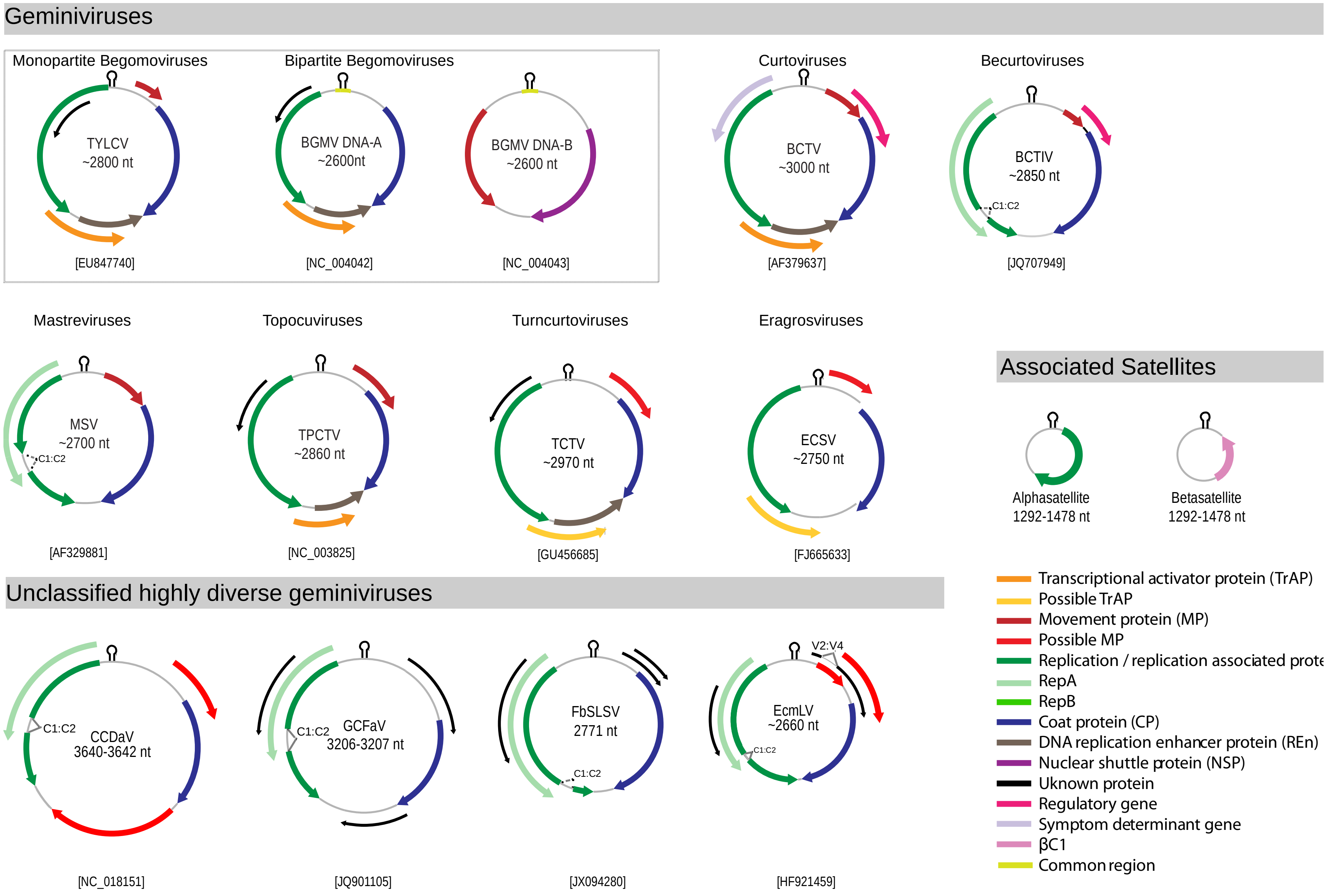
Table 1.3: Table showing putative classified viruses in the genus *Cycloviridae*

Species	Acronym	GenBank Accession #	Isolation source	Reference
Bat cyclovirus -1	BaCyV-1	HQ738637	Bat muscle	(Li *et al.*, 2011a)
Bat faeces associated cyclovirus -1	BaFCyV-1	HM228874	Bat faeces	(Li *et al.*, 2010a)
Bat faeces associated cyclovirus -2	BaFCyV-2	JF938079	Bat faeces	(Li et al., 2010)
Bat faeces associated cyclovirus -3	BaFCyV-3	JF938081	Bat faeces	(Li et al., 2010)
Bat faeces associated cyclovirus -4	BaFCyV-4	JF938082	Bat faeces	(Li et al., 2010)
Bovine cyclovirus -1	BoCyV-1	HQ738634	Cow muscle	(Li et al., 2011)
Bovine cyclovirus -1	BoCyV-1	HQ738635	Goat muscle	(Li et al., 2011)
Chimpanzee faeces associated cyclovirus -1	ChmFCyV-1	GQ404849	Chimpanzee faeces	(Li et al., 2010)
Chimpanzee faeces associated cyclovirus -1	ChmFCyV-1	GQ404850	Chimpanzee faeces	(Li et al., 2010)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638058	*Pantala flavescens*	(Rosario et al., 2011)
Dragonfly Cyclovirus -1	DfCyV-1	JX185420	*Pantala flavescens*	(Rosario *et al.*, 2012b)
Dragonfly Cyclovirus -1	DfCyV-1	JX185421	*Pantala flavescens*	(Rosario et al., 2012)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638065	*Pantala flavescens*	(Rosario et al., 2011)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638066	*Pantala flavescens*	(Rosario et al., 2011)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638067	*Pantala flavescens*	(Rosario et al., 2011)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638069	*Tholymis tillarga*	(Rosario et al., 2011)
Dragonfly Cyclovirus -1	DfCyV-1	JX185419	*Pantala flavescens*	(Rosario et al., 2012)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638068	*Diplacodes bipunctata*	(Rosario *et al.*, 2011)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638049	*Tholymis tillarga*	(Rosario et al., 2011)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638061	*Diplacodes bipunctata*	(Rosario et al., 2011)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638062	*Diplacodes bipunctata*	(Rosario et al., 2011)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638063	*Diplacodes bipunctata*	(Rosario et al., 2011)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638064	*Diplacodes bipunctata*	(Rosario et al., 2011)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638053	*Diplacodes bipunctata*	(Rosario et al., 2011)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638054	*Diplacodes bipunctata*	(Rosario et al., 2011)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638050	*Tholymis tillarga*	(Rosario et al., 2011)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638051	*Tholymis tillarga*	(Rosario et al., 2011)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638052	*Pantala flavescens*	(Rosario et al., 2011)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638055	*Pantala flavescens*	(Rosario et al., 2011)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638059	*Pantala flavescens*	(Rosario et al., 2011)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638056	*Pantala flavescens*	(Rosario et al., 2011)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638060	*Diplacodes bipunctata*	(Rosario et al., 2011)
Dragonfly Cyclovirus -1	DfCyV-1	HQ638057	*Diplacodes bipunctata*	(Rosario et al., 2011)
Dragonfly Cyclovirus -2	DfCyV-2	JX185422	*Pantala flavescens*	(Rosario et al., 2012)
Dragonfly Cyclovirus -2	DfCyV-2	JX185423	*Anax junius*	(Rosario et al., 2012)
Dragonfly Cyclovirus -3	DfCyV-3	JX185424	*Erythemis simplicicollis*	(Rosario et al., 2012)
Bat circovirus China	YN-BtCV-3	JF938080	Bat faeces	(Ge *et al.*, 2011)
Bat circovirus China	YN-BtCV	JN377566	Bat faeces	(Ge et al., 2011)
Dragonfly Cyclovirus -4	DfCyV-4	JX185425	*Somatochlora meridionalis*	(Rosario et al., 2012)
Dragonfly Cyclovirus -4	DfCyV-4	KC512916	*Rhionaeschna multicolor*	(Dayaram *et al.*, 2013b)
Dragonfly Cyclovirus -4	DfCyV-4	KC512917	*Rhionaeschna multicolor*	(Dayaram *et al.*, 2013b)
Dragonfly Cyclovirus -5	DfCyV-5	JX185426	*Erythrodiplax umbrata*	(Rosario et al., 2012)
Dragonfly Cyclovirus -5	DfCyV-5	JX185427	*Erythrodiplax umbrata*	(Rosario et al., 2012)
Dragonfly Cyclovirus -6	DfCyV-6	KC512918	*Rhionaeschna multicolor*	(Dayaram *et al.*, 2013b)
Dragonfly Cyclovirus -7	DfCyV-7	KC512919	*Xanthocnemis zealandica*	(Dayaram *et al.*, 2013b)
Dragonfly Cyclovirus -8	DfCyV-8	KC512920	*Orthetrum sabina*	(Dayaram *et al.*, 2013b)
Florida wood cockroach cyclovirus -1	FWCasCyV-1	JX569794	*Eurycotis floridana*	(Padilla-Rodriguez *et al.*, 2013)
Gallus cyclovirus -1	GaCyV-1	HQ738643	Chicken muscle	(Li et al., 2011)
Gallus cyclovirus -1	GaCyV-1	HQ738644	Chicken muscle	(Li et al., 2011)
Goat cyclovirus -1	GoCyV-1	HQ738636	Goat	(Li et al., 2011)
Human cyclovirus -1	HuCyV-1	KF031471	Human	(Van Tan *et al.*, 2013)
Human cyclovirus -1	HuCyV-1	KF031470	Human	(Van Tan et al., 2013)
Human cyclovirus -1	HuCyV-1	KF031469	Human	(Van Tan et al., 2013)
Human cyclovirus -1	HuCyV-1	KF031467	Human	(Van Tan et al., 2013)
Human cyclovirus -1	HuCyV-1	KF031465	Human	(Van Tan et al., 2013)
Human cyclovirus -1	HuCyV-1	KF031468	Human	(Van Tan et al., 2013)
Human cyclovirus -1	HuCyV-1	KF031466	Human	(Van Tan et al., 2013)
Human cyclovirus -2	HuCyV-2	KC771281	Human	(Smits *et al.*, 2013)
Human cyclovirus -3	HuCyV-3	KF726984	Human	(Phan *et al.*, 2014)
Human cyclovirus -3	HuCyV-3	KF726985	Human	(Phan et al., 2014)
Human cyclovirus -3	HuCyV-3	KF726987	Human	(Phan et al., 2014)
Human cyclovirus -3	HuCyV-3	KF726986	Human	(Phan et al., 2014)
Human faeces associated cyclovirus -1	HuFCyV-1	GQ404847	Human faeces	(Li et al., 2010)
Human faeces associated cyclovirus -2	HuFCyV-2	GQ404844	Human faeces	(Li et al., 2010)
Human faeces associated cyclovirus -3	HuFCyV-3	GQ404846	Human faeces	(Li et al., 2010)
Human faeces associated cyclovirus -3	HuFCyV-3	GQ404848	Human faeces	(Li et al., 2010)
Human faeces associated cyclovirus -4	HuFCyV-4	GQ404858	Human faeces	(Li et al., 2010)
Human faeces associated cyclovirus -4	HuFCyV-4	GQ404857	Human faeces	(Li et al., 2010)
Human faeces associated cyclovirus -5	HuFCyV-5	GQ404845	Human faeces	(Li et al., 2010)
Human faeces associated cyclovirus -6	HuFCyV-6	GQ404854	Human faeces	(Li et al., 2010)
Human faeces associated cyclovirus -8	HuFCyV-7	GQ404855	Human faeces	(Li et al., 2010)

1.2.2 Family *Geminiviridae*

Geminivirids have circular ssDNA genomes of ~2.7-3.6 kb (Loconsole *et al.*, 2012) that are encapsidated in geminate virions. Members of this family infect a range of plant species and are transmitted from plant-to-plant via insect vectors (Markham *et al.*, 1994). These insect vectors vary between the genera of *Geminiviridae*. Mastreviruses, becurtoviruses and turncurtoviruses are vectored by leafhoppers, topocuviruses and curtoviruses by treehoppers and begomoviruses by whiteflies, while the vector for eragroviruses is still unknown. Worldwide, Geminiviruses cause serious crop loss and damage, as many geminiviruses evolve rapidly which often leads to the emergence of new viral strains and species (Mansoor *et al.*, 2003). With the increase in global trade, agriculture and movement of vectors, this in turn has facilitated the spread of these viruses around the world (Fèvre *et al.*, 2006; Gray & Banerjee, 1999; Karesh *et al.*, 2005).

Geminiviruses can either have bipartite genomes, such is the case with begomoviruses which have two twinned icosahedral virions that are packaged in separate virion particles and must both be transmitted together to cause infection. Bipartite begomoviruses are composed of two components: one referred to as DNA A which encodes replication-associated proteins and the other, DNA B that encodes two proteins that are needed for systemic movement (Figure 1.2) (Saeed *et al.*, 2007). Other geminiviruses such as mastreviruses are monopartite (Fauquet *et al.*, 2005). The ICTV currently recognises seven different genera of geminiviruses. These include *Becurtovirus, Begomovirus, Curtovirus, Eragrovirus, Mastrevirus, Topocuvirus* and *Turncurtovirus* (Figure 1.2) (King *et al.*, 2011; Varsani *et al.*, 2014).



Geminiviruses
Monopartite Begomoviruses
Bipartite Begomoviruses
TYLCV
~2800 nt
[EU847740]
BGMV DNA-A
~2600nt
[NC_004042]
BGMV DNA-B
~2600 nt
[NC_004043]
Curtoviruses
BCTV
~3000 nt
[AF379637]
Becurtoviruses
BCTIV
~2850 nt
C1:C2
[JQ707949]
Mastreviruses
MSV
~2700 nt
C1:C2
[AF329881]
Topocuviruses
TPCTV
~2860 nt
[NC_003825]
Turncurtoviruses
TCTV
~2970 nt
[GU456685]
Eragrosviruses
ECSV
~2750 nt
[FJ665633]
Associated Satellites
Alphasatellite
1292-1478 nt
Betasatellite
1292-1478 nt
Unclassified highly diverse geminiviruses
C1:C2
CCDaV
3640-3642 nt
[NC_018151]
C1:C2
GCFaV
3206-3207 nt
[JQ901105]
FbSLSV
2771 nt
C1:C2
[JX094280]
V2:V4
EcmLV
~2660 nt
C1:C2
[HF921459]
Transcriptional activator protein (TrAP)
Possible TrAP
Movement protein (MP)
Possible MP
Replication / replication associated prote
RepA
RepB
Coat protein (CP)
DNA replication enhancer protein (REn)
Nuclear shuttle protein (NSP)
Uknown protein
Regulatory gene
Symptom determinant gene
βC1
Common region

Figure 1.2. Genome organisation of different geminivirus genera. Classified geminiviruses: Monopartite begomoviruses: *Tomato yellow leaf curl virus* (EU847740), Bipartite begomoviruses: *Bean golden mosaic virus DNA-A* (NC_004042) and DNA-B (NC004043), Curtoviruses: *Beet curly top virus* (AF379637), Becurtoviruses: *Beet curly top Iran virus* (JQ707949), Mastreviruses: Maize streak virus (AF329881), Topocuviruses: *Tomato pseudo-curly top virus* (NC003825), Turncurtoviruses: *Turnip curly top virus* (GU456685) and Eragrosviruses: *Eragrostis curvula streak virus* (FJ665633). Associated satellite: Alphasatellite and Betasatellite. Unclassified highly diverse geminiviruses. Citrus chlorotic dwarf associated virus (NC018151), Grapevine red-blotch associated virus (JQ901105), French bean severe leaf curl virus (JX094280) and Euphorbia caput-medusae latent virus (HF921459).

With NGS technology becoming common place this has led to an increase in the number of novel geminivirus-like sequences being discovered. Grapevine red leaf associated virus (GRLaV) was recently isolated from grape vines with grapevine red leaf disease (GRD) (Poojari *et al.*, 2013). Phylogenetic analysis of the viral sequence showed it was only distantly related to other known geminiviruses, but grouped closely with other recently described viruses isolated from grapevines such as Grapevine Cabernet Franc-associated virus and Grapevine red blotch-associated virus (Krenz *et al.*, 2012; Poojari *et al.*, 2013; Sudarshana *et al.*, 2013).

Citrus chlorotic dwarf virus (CCDaV) is another recently described geminivirus and is thought to be the causal agent of citrus chlorotic dwarf disease (CCDD) (Loconsole *et al.*, 2012). The genome of this virus has a different structural organisation to other known geminiviruses highlighting that it is a divergent member of this family (Loconsole *et al.*, 2012). Another highly divergent geminivirus was also recently isolated from *Euphorbia caput-medusae* (Bernardo *et al.*, 2013). The virus Euphorbia caput-medusae latent virus (EcmLV) caused severe symptoms in tomatoes and *Nicotiana benthamiana*. The genome organisation of EcmLV is unique to any of the previously described geminiviruses, although phylogenetic analysis does suggest that its most recent common ancestors are geminiviruses (Bernardo *et al.*, 2013).

Satellite DNA molecules were originally thought to be associated with monopartite begomoviruses. However, recent studies have shown some alphasatellites to be associated with bipartite begomoviruses (Mubin *et al.*, 2010; Romay *et al.*, 2010) as well as a species of mastrevirus (Kumar *et al.*, 2014). These satellites share little or no nucleotide sequence similarity to either the virus or host genome, however, they are thought to aid replication and increase pathogenicity of the virus (Briddon & Stanley, 2006; Idris *et al.*, 2011).

There are two different types of satellites associated with geminiviruses, these are alphasatellites and betasatellites. Betasatellites are ~1.4 kb and have one major ORF encoding for the protein βC1 (Figure 1.2). This protein suppresses silencing, assists with viral movement and up regulates the viral DNA load assisting the helper virus to induce disease symptoms in their hosts and cannot replicate without their helper virus (Kumar *et al.*, 2014; Mubin *et al.*, 2010). Betasatellites also have a stem-loop structure with a nonanucleotide motif which is recognised by a diverse range of begomoviruses (Briddon & Stanley, 2006; Briddon *et al.*, 2003). Alphsatellites are ~1.4 kb and have one major ORF which encodes for a Rep that is similar to those found in nanoviruses. The Rep enables them to replicate independently of the helper virus, which is a virus that often co-infects cells with other viruses providing the necessary enzymes for replication of the viral genome (Malyshenko *et al.*, 1989).

1.2.3 Family *Nanoviridae*

Nanoviruses are found in many regions of the world and are often associated with diseases that cause stunting in plants. The transmission of most nanoviruses occurs via an aphid vector, however, the virus does not replicate within the vector (Franz *et al.*, 1999). Nanoviruses have multicomponent genomes that usually consist of 6-8 components, it is currently assumed that all components are required for infection. The component DNA molecules are circular and usually ~1 kb and are encapsidated in separate virions with each molecule having a common region and a stem-loop. The master Rep of nanoviruses replicates all other genome components by recognising the common region at the origin of replication, which is the same for each component (Timchenko *et al.*, 1999).

Each component of nanoviruses encodes for a single protein with a specific function such as replication, encapsidation, nuclear shuttle protein, movement protein, and many others whose function is yet to be determined (Figure 1.3). Each capsid is icosahedral with a diameter of 18-20 nanometres (Gronenborn, 2004). *Nanoviridae* contain two genera; *Nanovirus*, which includes, *Faba bean necrotic stunt vir*us (FBNSV), *Milk vetch dwarf virus* (MDV), *Pea necrotic yellow dwarf virus* (PNYDV) and *Subterranean clover stunt virus* (SCSV) (Boevink *et al.*, 1995; Grigoras *et al.*, 2014; Inouye *et al.*, 1968; Katul *et al.*, 1995; Mandal *et al.*, 2004); and *Babuvirus,* which includes *Abaca bunchy top virus* (ABTV)*, Banana bunchy top virus* (BBTV) and *Cardamom bushy dwarf virus* (CdBDV) (Burns *et al.*, 1995; Mandal *et al.*, 2004; Savory & Ramakrishnan, 2014; Sharman *et al.*, 2008).

Alphasatellites have also found to be associated with various nanoviruses (Briddon *et al.*, 2012). Several alphsatellites have been found to be associated with FBNSV from Europe which are similar to other alphsatellites found to be associated with MDV and SCSV (Grigoras *et al.*, 2014).

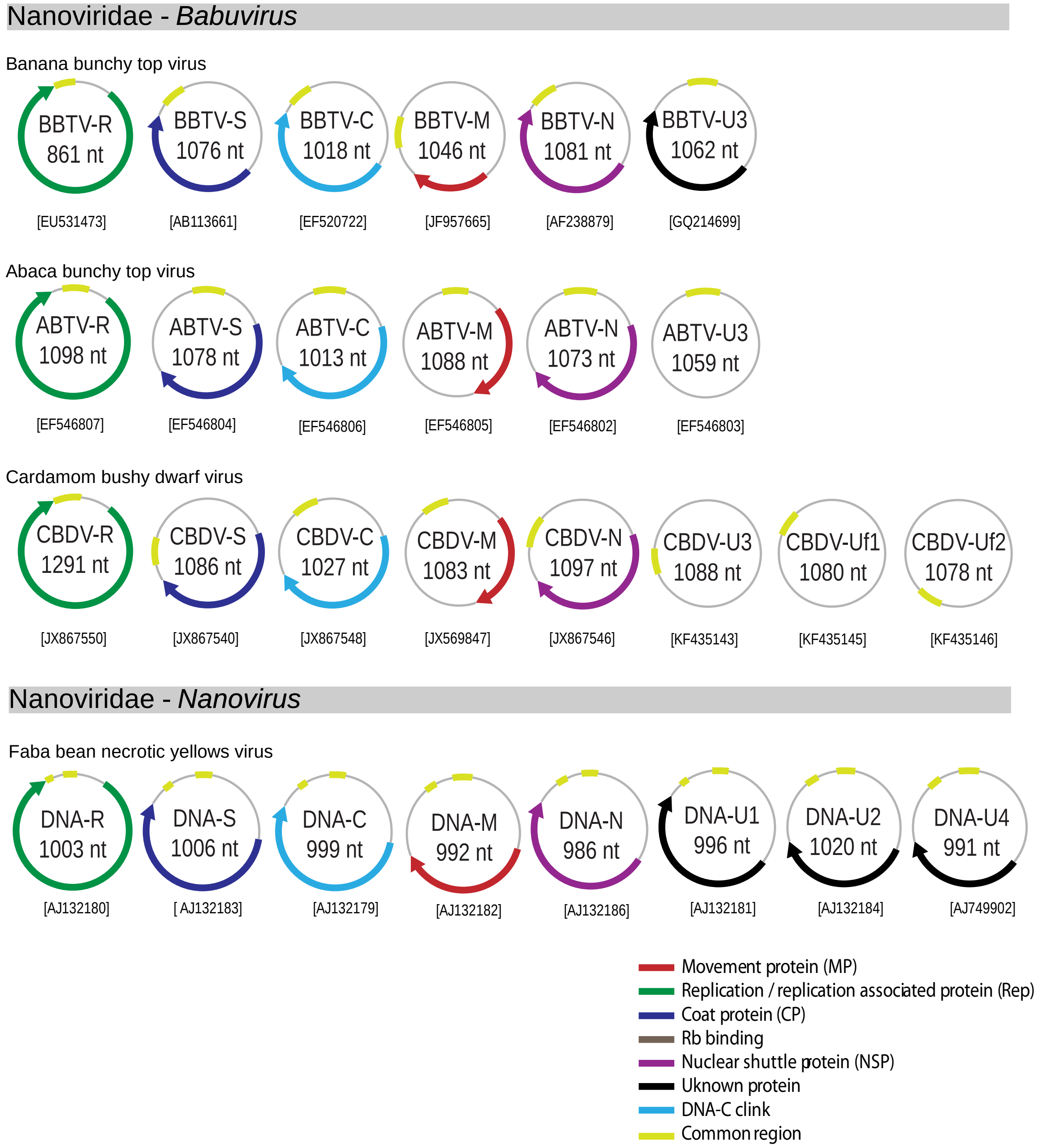


Figure 1.3: Genome organisation and components of the two *Nanoviridae* genera. Babuvirus: *Banana bunch top* virus (BBTV), *Abaca bunchy top virus* (ABTV) and *Cardamom bush dwarf virus* (CdBDV). Nanovirus: *Faba bean necrotic yellows virus* (FBNYV).

1.2.4 Family *Microviridae*

The ICTV currently recognises four genera within the *Microvirdae* family inlcuding: *Microvirus*, *Chlamydiamicrovirus*, *Bdellomicrovirus* and *Spiromicrovirus*. In addition to these genera is the subfamily *Gokushovirinae* (King *et al.*, 2011). There are two different morphologies in *Microviridae*. *Chlamydiamicrovirus*, *Bdellomicrovirus* and *Spiromicrovirus* and more closely related and have slightly smaller genomes to that of Microvirus as they do not have genes encoding for major spike and external scaffolding proteins (Brentlinger *et al.*, 2002; King *et al.*, 2011).

Viruses classified in the Microvirus genus have non-enveloped virions with icosahedral symmetry and have all been isolated from Enterobateriacae, which is thought to be the host. However, it has been suggested that more research is still needed to be conducted into other potential hosts (Fauquet *et al.*, 2005). Microviruses are bacteriophages that have ssDNA genome encoding for over fourteen different proteins, including CPs, Rep and many different structural and binding proteins (Dokland *et al.*, 1999; McKenna *et al.*, 1992). Many of the ORFs in the genome overlap which increases the amount of genetic information encoding for proteins within a small genome (Dokland *et al.*, 1999).

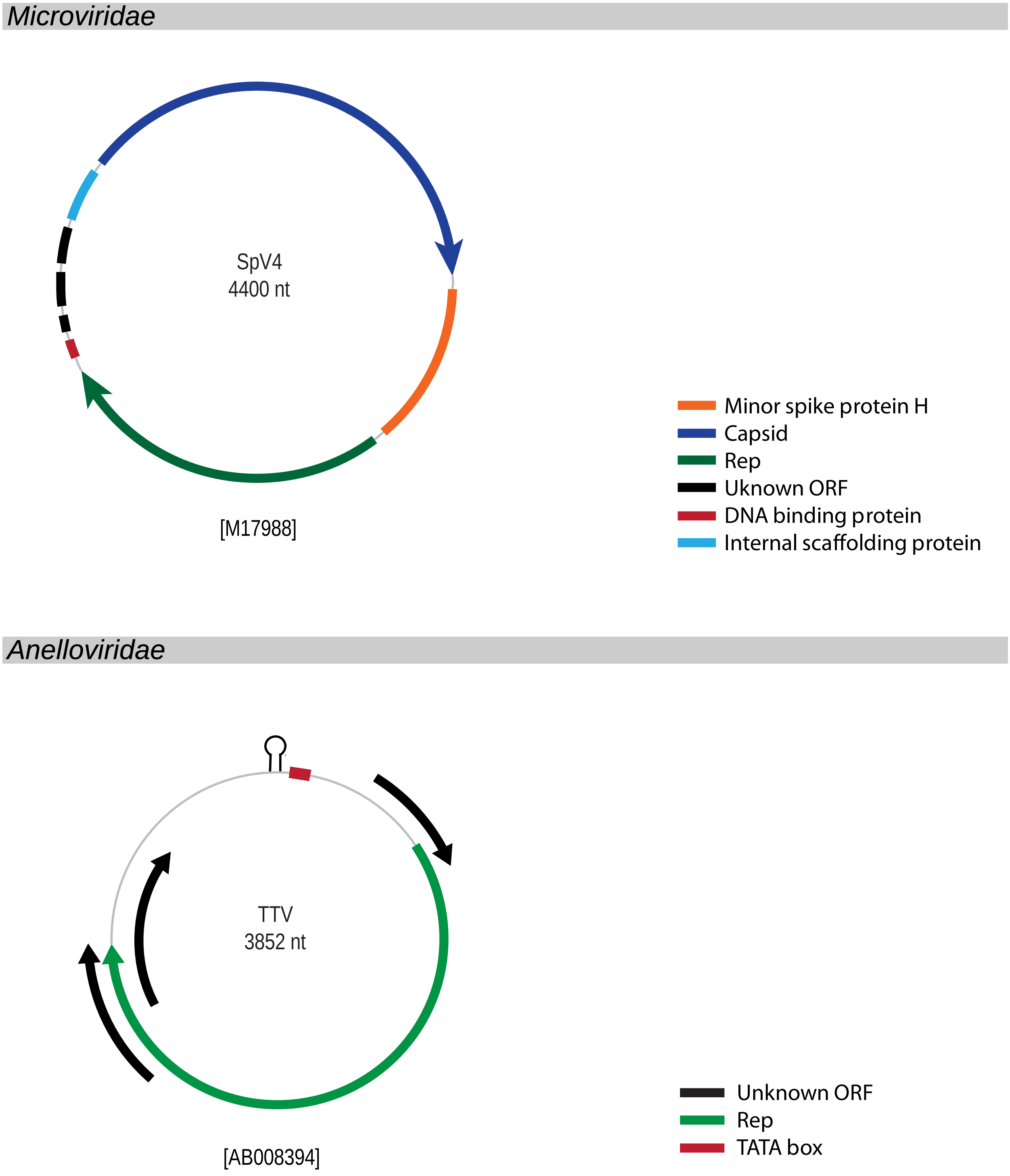


Figure 1.4: Genome organisation of the microvirus Spiroplasma phage 4 (Spv4) (M17988) and anellovirus Torque teno virus (TTV) (AB008394).

1.2.5 Family *Anelloviridae*

Anelloviridae is a family of viruses that infect vertebrates. They have a circular ssDNA genome which exhibits icosahedral symmetry. The *Anelloviridae* family encompasses 11 different genera currently recognised by the ICTV (King *et al.*, 2011). The viral genomes of *Anelloviridae* are negative sense and range in size from ~2.8 to 4 kb (Figure 1.4).

The most studied viruses in the *Anelloviridae* family are *Torque teno virus* (TTV) and *Torque teno mini virus* (TTMV). These viruses have been isolated from humans, primates, tupaias, cats, dogs and farm animals (Bendinelli *et al.*, 2001; Biagini *et al.*, 2001; Okamoto *et al.*, 2001; Okamoto *et al.*, 2002). The genome organisations of TTV and TTMV have two major ORFs and two smaller ORFs that are uni-directionally transcribed with one of the major ORFs encoding for a possible Rep (Okamoto & Mayumi, 2001). Epidemiological studies have shown these viruses to be globally distributed in both rural and urban populations (Prescott *et al.*, 1998). The mode of transmission of the virus is still unclear although it has been isolated from plasma, saliva and faeces suggesting a possible faecal oral route (Gallian *et al.*, 2000). It has also recently been suggested that the *Gyrovirus* genus be reassigned within the *Anelloviridae* family.

1.2.6 Family *Parvoviridae*

The virus family *Parvovirus* includes viruses with small linear ssDNA genomes. The virions of parvoviruses are icosahedral between 18-26 nm in diameter (Cotmorel & Tattersall, 1996; Muzyczka & Berns, 2001). The genome sizes of parvoviruses are ~5 kb and have two major ORFs (Figure 1.5). The first encodes for two non-structural proteins, NS-1 and NS-2, while the second encodes for three coat proteins VP-1 to VP-3 (Lukashov & Goudsmit, 2001). Parvoviruses unlike circular ssDNA viruses replicate using the rolling hairpin replication mechanism. The mechanism works via a 3’ terminal hairpin structure which is where the viral Rep creates a nick. Enzymes then convert viral ssDNA into dsDNA for transcription and replication (Cotmorel & Tattersall, 1996).

The classification of parvoviruses is dependent of their host range and their dependence on other viruses for replication. The viruses in this family replicate in the nuclei of both invertebrate and vertebrate hosts. Based on these principles parvoviruses usually fall under three types: autonomous viruses of vertebrates, helper dependent viruses of vertebrates and autonomous viruses of insects (Lukashov & Goudsmit, 2001; van Regenmortel *et al.*, 2000). This has helped the ICTV classify parvoviruses accordingly. *Parvoviridae* includes two subfamilies, *Parvovirinae* that infect vertebrates and contains eight genera *Parvovirus* which viruses that infect vertebrates, *Erythrovirus*comprises of viruses that have been isolated from rhesus and pig tailed macaques and *Dependovirus* which includes adeno-associated viruses (Lukashov & Goudsmit, 2001; van Regenmortel *et al.*, 2000). The second subfamily is *Densovirinae* that infect invertebrates and contains five genera: *Ambidensovirus, Brevidensovirus, Hepandensovirus, Iteradensovirus*and *Pentyldensovirus* (King *et al.*, 2011).

1.2.7 Family *Inoviridae*

The family *Inoviridae* currently includes two different genera, Inovirus and Plectrovirus. They are a family of filamentous bacteriophages. The virions in this family are rods or filaments that are ~7 nm in diameter with the capsid with helical symmetry having a length between 85-280 nm and have a ssDNA genome ~4.4 - 8.5 kb (Beck & Zink, 1981; Fauquet *et al.*, 2005) (Figure 1.5).

Viruses within this family infect host cells without causing the cells to lyse, these infected cells continue to divide and increase the viral load. In some cases viruses within this family can be lysogenic, using intergrases to integrate their genome in to that of their bacterial hosts as a method of reproduction (Kuo *et al.*, 1987).

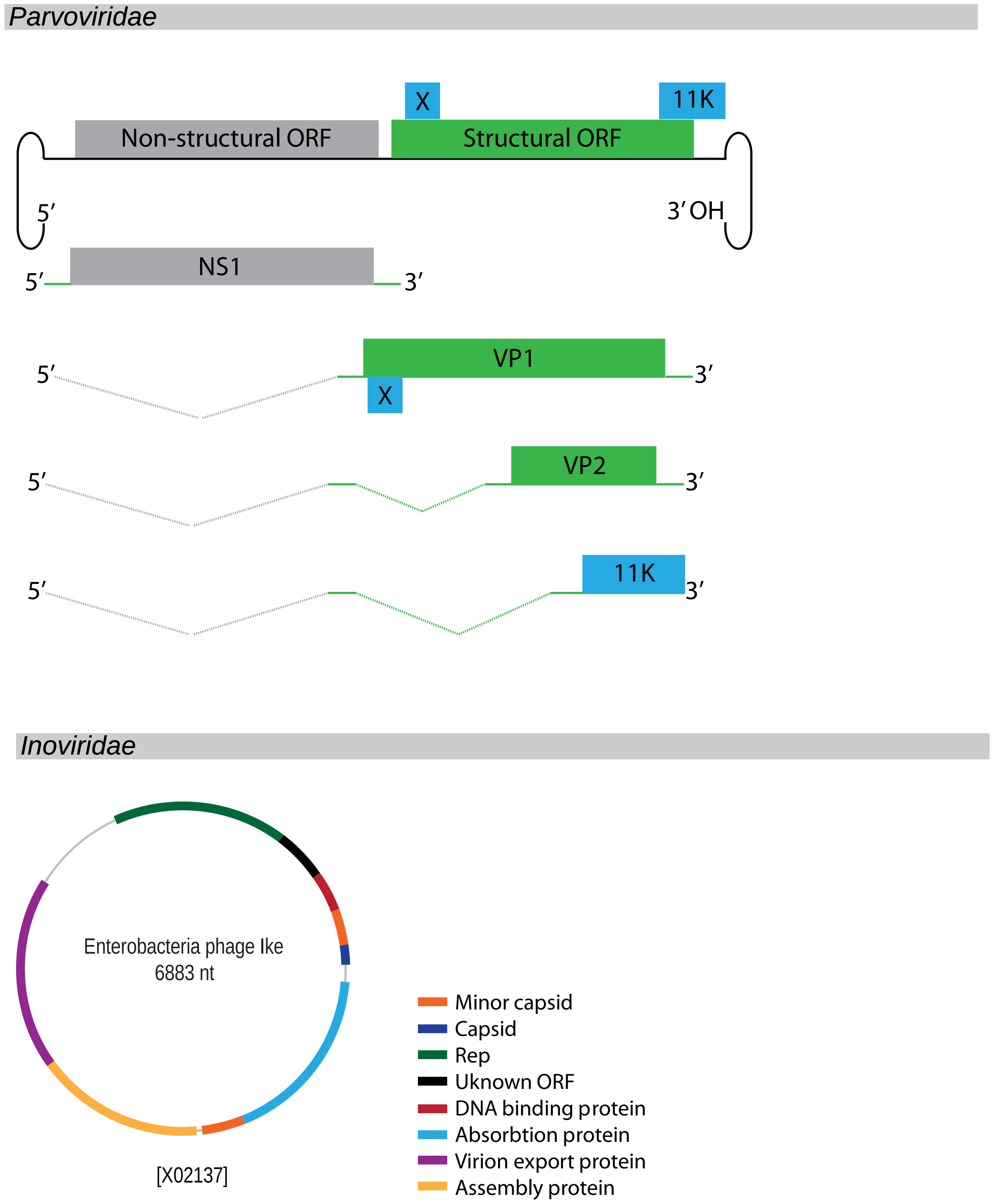


Figure 1.5: Genome organisation of human parvovirus B19 (M13178). VP1: capsid protein, VP2: capsid protein, 11K: host-modulation protein 11K, NS1: non-capsid protein and X: unknown protein. Genome organisation of Enterobacteria phage Ike (X02139), part of the Inovirus genus.

1.2.8 Unclassified ssDNA viruses

The number of unclassified ssDNA viruses has increased remarkably with the advent of metagenomics and NGS. Classifying ssDNA viruses has remained a challenge as there small genome size, varying genome architecture and high mutation rate make inferring ancestry a challenge (Rosario *et al.*, 2012a; Van Regenmortel & Mahy, 2004).

Many of the metagenomic studies to date have investigated viruses in faecal matter from various different animals including rats (Phan *et al.*, 2011), pigs (Sikorski *et al.*, 2013b), cows (Kim *et al.*, 2011), chimpanzees (Blinkova *et al.*, 2010; Li *et al.*, 2010b), bats (Ge *et al.*, 2012; Li *et al.*, 2010a), fur seals (Sikorski *et al.*, 2013a), sea lions (Li *et al.*, 2011b), ostrich (Sikorski *et al.*, 2013d), humans, badgers (van den Brand *et al.*, 2011), foxes (Bodewes *et al.*, 2013), caribou (Ng *et al.*, 2014), rabbits and various bird species (Sikorski *et al.*, 2013d) (Table 1.5).

Other studies have used metagenomics to look at the viruses of tissue and other samples from various animals such as dragonflies (Rosario *et al.*, 2012b), turtles (Ng *et al.*, 2009), shrimp (Ng *et al.*, 2013), mosquitoes (Ng *et al.*, 2011b) and nesting material from birds (Sikorski *et al.*, 2013c) (Table 1.1). Many of these novel CRESS DNA viruses discovered in these studies show some distant similarities to other families of ssDNA viruses such as *Geminiviridae, Nanoviridae*, and *Circoviridae*. Most of the CRESS DNA viral genomes identified have at least one ORF that has similarities to ssDNA Reps, however, most sequences are so divergent that it is likely that they will be assigned to completely new ssDNA viral families.

Other novel CRESS DNA viruses have also be isolated from environmental metagenomic studies. These studies have examined ssDNA viruses within different ecosystems such as ocean sediment (Yoshida *et al.*, 2013), ocean water (Angly *et al.*, 2006; Breitbart *et al.*, 2002; Labonté & Suttle, 2013), soil (Kim *et al.*, 2008), sewage (Ng *et al.*, 2012), reclaimed water (Rosario *et al.*, 2009b), Antarctic lakes (López-Bueno *et al.*, 2009), freshwater lakes (Roux *et al.*, 2012), hot springs (Schoenfeld *et al.*, 2008), aquifers (Smith *et al.*, 2013) and perennial ponds (Fancello *et al.*, 2013) (Table 1.1). There is a vast range in the genome sizes of the novel CRESS DNA viruses being discovered, as well as diverse genome organisations (Figure 1.6).

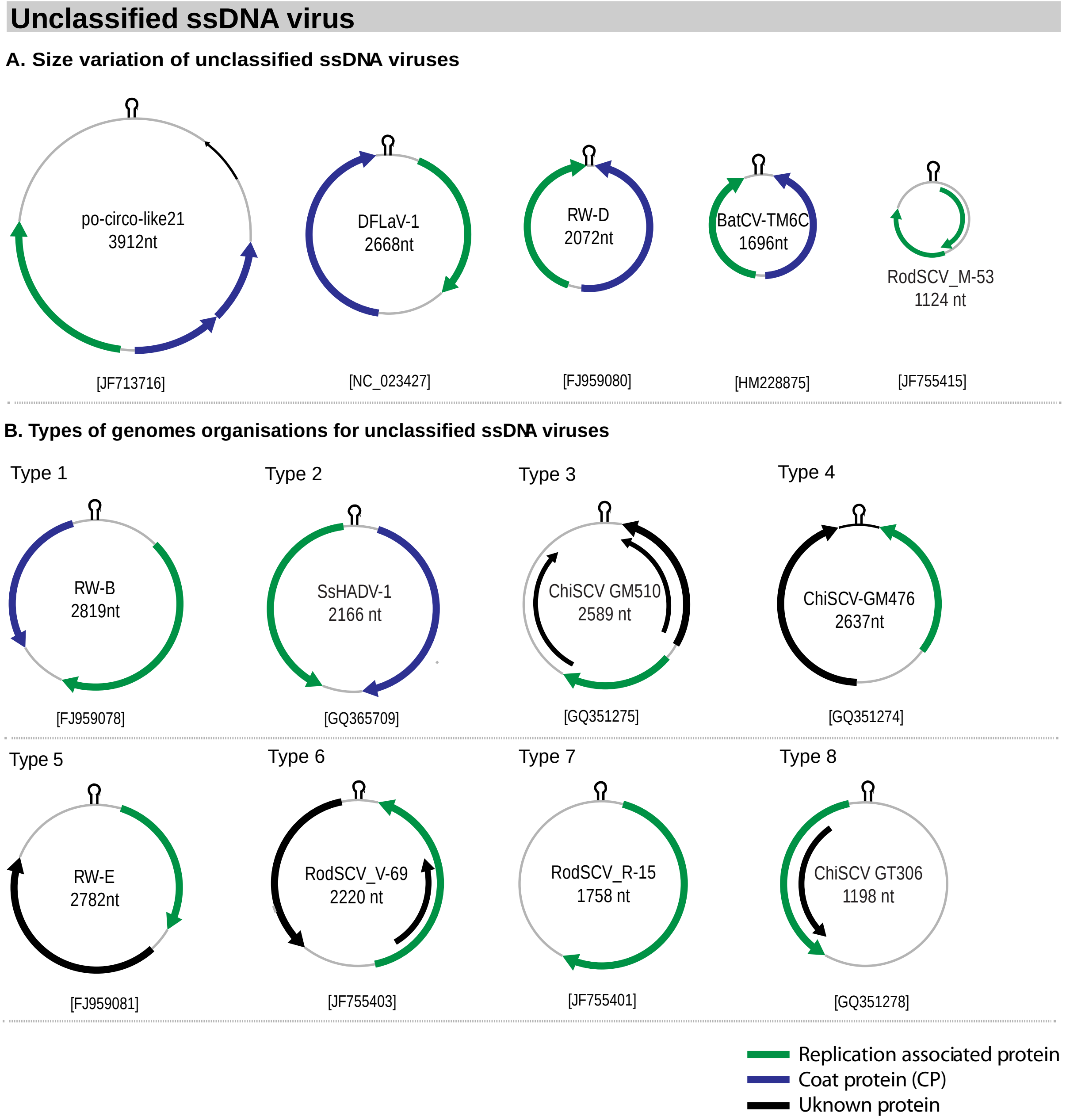


Figure 1.6: (A) Size variation in unclassified ssDNA viruses, po-circo-like virus 21 (JF713716), dragonfly larvae associated circular virus-1 (KF738873), circovirus-like RW-D (FJ959080), bat circovirus TM-6C (HM228875) and rodent stool-associated circular virus M-53 (JF755415) **(B)** Types of genome organisations for unclassified ssDNA viruses. Type 1: circovirus-like RW-B (FJ959078), Type 2: *Sclerotiniasclerotiorum* hypovirulence-associated DNA virus 1 (GQ365709), Type 3: Chimpanzee stool associated circular ssDNA virus GM510 (GQ351275), Type 4: Chimpanzee stool associated circular ssDNA virus GM476 (GQ351274), Type 5: circovirus-like RW-E (FJ959081), Type 6: rodent stool-associated circular virus V-69 (JF755403), Type 7: rodent stool-associated circular virus R-15 (JF755401) and Type 8: Chimpanzee stool associated circular ssDNA virus GT306 (GQ351278). Adapted from Rosario *et al.* (2012b).

1.2.9 Gemycircularviruses

The interaction between viruses and fungi have been documented as early as the 1960s, however, the nature of the interaction between them still remains unclear. Gemycircularviruses are a proposed new genus of viruses, these viruses have ssDNA genomes with two major ORFs that are bi-directionally transcribed, one encodes a Rep which is most closely related to Reps of geminiviruses whilst the other ORF is thought to be a the CP (Figure 1.7). They also have an intergenic region that contains the nonanucleotide motif TAATATTAC on the virion strand. It is also apparent that these viral Rep sequences share similarities with Rep-like sequences in fungal genomes of *Aspergillus fumigatus, Collectorichum higginsianum, Laccaria bicolor, Magmaporthe oryzae, Nectria haematococca, Serpula lacrymans* and *Tuber melanosporm* (Liu *et al.*, 2011).

Sclerotinia sclerotiorum hypovirulence-associated DNA virus 1 (SsHADV-1) was isolated from the fungal species *Sclerotinia sclerotiorum,* sampled in the Hunan Province, China (Figure 1.7) (Yu *et al.*, 2010). This is the first report of a ssDNA virus infecting fungi. Phylogenetic analysis of the Rep showed it clustered with the Reps from the family *Geminiviridae* and also contained the GRS motif, commonly associated with geminiviruses. Analysis of the CP showed it was very divergent and did not cluster with any CP from *Geminiviridae, Nanoviridae* or *Circoviridae* families. Further infection studies and culturing experiments of SsHADV-1 demonstrated that it infects the hyphae of its host extracellularly. It was also found that when the virus infects the fungus is confers hypovirulence to the fungal host which in turn infects plants. Therefore SsHADV-1 is able to effect the pathogenicity of the fungi to the plant. SsHADV-1 and other mycoviruses that confer hypovirulence may be useful in the future as tools for controlling plant disease.

Another study looking at benthic river sediments from two urban rivers in Christchurch, New Zealand identified isolates of SsHADV-1 (Kraberger *et al.*, 2013a). The four genomes recovered from the river sediments shared between ~98.1-99.2% genome wide pairwise identities with SsHADV-1. The genome organisations of these viral isolates were identical to SsHADV-1 (Kraberger *et al.*, 2013a).

The isolation of cassava associated circular DNA virus (CasCV) from cassava leaves infected with fungi *Collectotrichum* and *Plectosphaerella* is a another example of the association of gemycircularviruses with fungi (Dayaram *et al.*, 2012). The genome organisation of CasCV was similar to that of the previously described SsHADV-1, Mosquito VEM virus SDBVL (Ng *et al.*, 2011b) isolated from mosquitoes and Meles meles faecal virus (MmFV) isolated from badger faecal matter (van den Brand *et al.*, 2011) (Figure 1.7). The ~2.2 kb genome had three major ORFs that were bi-directionally transcribed. One encoded a putative CP whilst then other two are believed to encode for a spliced Rep that shared close homology with geminiviruses.

Another gemycircularvirus was recently recovered from the plant *Hypericum japonicum* that was sampled in Vietnam (Du *et al.*, 2014). The virus Hypericum japonicum-associated circular DNA virus (HJasCV) has similar genome organisation to that of CasCV and SsHADV-1, with two major ORFs that are bi-directionally transcribed. One ORF encodes for a putative CP and the other a spliced Rep with an intron. HJasCV shares between ~58-66% nucleotide identity with SsHADV-1, CasCV, MmFV and other novel ssDNA viruses associated with dragonflies. Phylogenetic analysis further illustrated that HJasCV forms a monophyletic clade with SsHADV-1 and other related gemycircularviruses.

A further study isolated a geminivirus-related DNA mycovirus named MmFV whilst looking at the viral flora in European badger faeces from the Netherlands (van den Brand *et al.*, 2011). The viral genome recovered shared close homology and similar genome organisation with previously described SsHADV-1. However, this virus had two intergenic regions, the LIR containing the nonanucleotide motif TAACTTTGT at the apex of the stem-loop structure which differs slight to the motif observed in SsHADV-1 (van den Brand *et al.*, 2011). Phylogenetic analysis of the Rep showed it clustered with other geminiviruses Rep proteins, whilst the putative CP was divergent from any of the major ssDNA viral families. The isolation of a gemycircularvirus from European badger faeces suggest that animal faecal matter is a huge reservoir of viral diversity previously unexplored.

A study of mammal and bird faecal samples identified novel gemycircularviruses (Sikorski *et al.*, 2013d). The sampled faecal matter was taken from a variety of different locations around New Zealand. Fourteen viral genomes were identified to have similarities to gemycircularviruses. The genome organisation was the same as previously described gemycircularviruses (Rosario *et al.*, 2012b), however, many of the isolates had a spliced Rep on the complimentary sense strand with the typical acceptor donor sites identified. Spliced Rep genes are typical of some geminiviruses such as the genus mastrevirus (Dekker *et al.*, 1991; Mullineaux *et al.*, 1990; Wright *et al.*, 1997) and becurtoviruses (Heydarnejad *et al.*, 2013). This study highlights that these gemycircularviruses have a wide distribution across different environments and ecosystems.

A recent study has also reported the first discovery of gemycircularviruses in human tissue samples (Lamberto *et al.*, 2014). This study isolated three novel viral sequences that are distantly related to gemycircularviruses from blood samples taken from healthy cattle and serum from patients suffering with multiple sclerosis (Lamberto *et al.*, 2014). The genome organisation of these three viral sequences were similar to other gemycircularviruses, with a putative spliced Rep on the negative strand and a CP on the positive strand. They also contained the conserved nonanucleotide motif "TAATATTAT" observed in other gemycircularviruses.

.

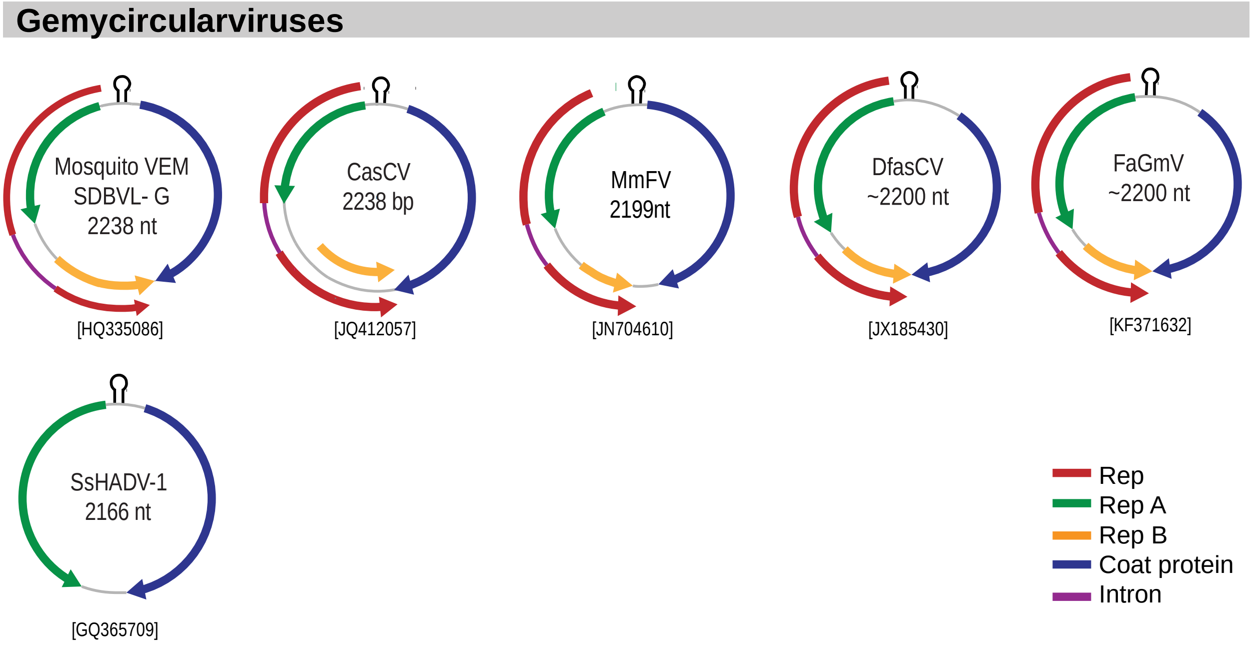


Figure 1.7. Genome organisation of the gemycircularviruses. These include organisations for Mosquito VEM SDBVL-G (Ng *et al.*, 2011b), Cassava associated circular DNA virus (CasCV) (Dayaram *et al.*, 2012), Meles meles fecal virus (MmFV) (van den Brand *et al.*, 2011) and Dragonfly associated circular viruses -1, -2 and -3 (DfasCV-1,-2,-3) (Rosario *et al.*, 2012b)

1.3 Rolling circle replication in ssDNA viruses

RCR mechanism was first described by Gilbert and Dressler (1968) and is responsible for the replication of circular ssDNA of certain bacteriophage, plasmids and viruses (Khan, 1997, 2000; Martin *et al.*, 2011; Rosario *et al.*, 2012a). RCR is a process which enable genomes to replicate when the leading and lagging strands of DNA are unwound (Liu *et al.*, 1996). Although the RCR mechanism has been validated for viral replication in circoviruses, nanoviruses, parvoviruses and geminiviruses (Cheung, 2004; Cotmorel & Tattersall, 1996; Jeske *et al.*, 2001; Timchenko *et al.*, 1999), it is currently assumed that the Reps of novel CRESS ssDNA viruses fulfil a similar function.

Replication of novel CRESS ssDNA viruses via RCR occurs once the ssDNA genome is in the nucleus of the host cell. The ssDNA genome (the virion strand) is used as a template to create the complimentary strand of DNA leading to a double strand intermediate using the host DNA polymerase (Figure 1.8).The dsDNA molecules are then packaged into mini-chromosomes which allows the transcription of the genes (Abouzid *et al.*, 1988; Pilartz & Jeske, 1992). The Rep then binds near the origin of replication either located in the long or short intergenic region (IR). The bound Rep nicks the virion strand between the thymine position 7 and the adenine position 8 of the nonanucleotide motif ‘NNNTATT↓AC’ (Hafner *et al.*, 1997; Koonin & Ilyina, 1992). This creates an open circular form of the virion strand that serves as the template for leading strand synthesis. Stand synthesis continues in cycles until the virion sense strand is displaced by the Rep protein (Figure 1.8). The ssDNA replicates can then either become a monomeric virus whereby Rep catalyses a joining reaction between the end of the synthesized strand, or replication of this strand may continue using host polymerase to again convert it to dsDNA (Cheung, 2004; Heyraud-Nitschke *et al.*, 1995; Rosario *et al.*, 2012a).

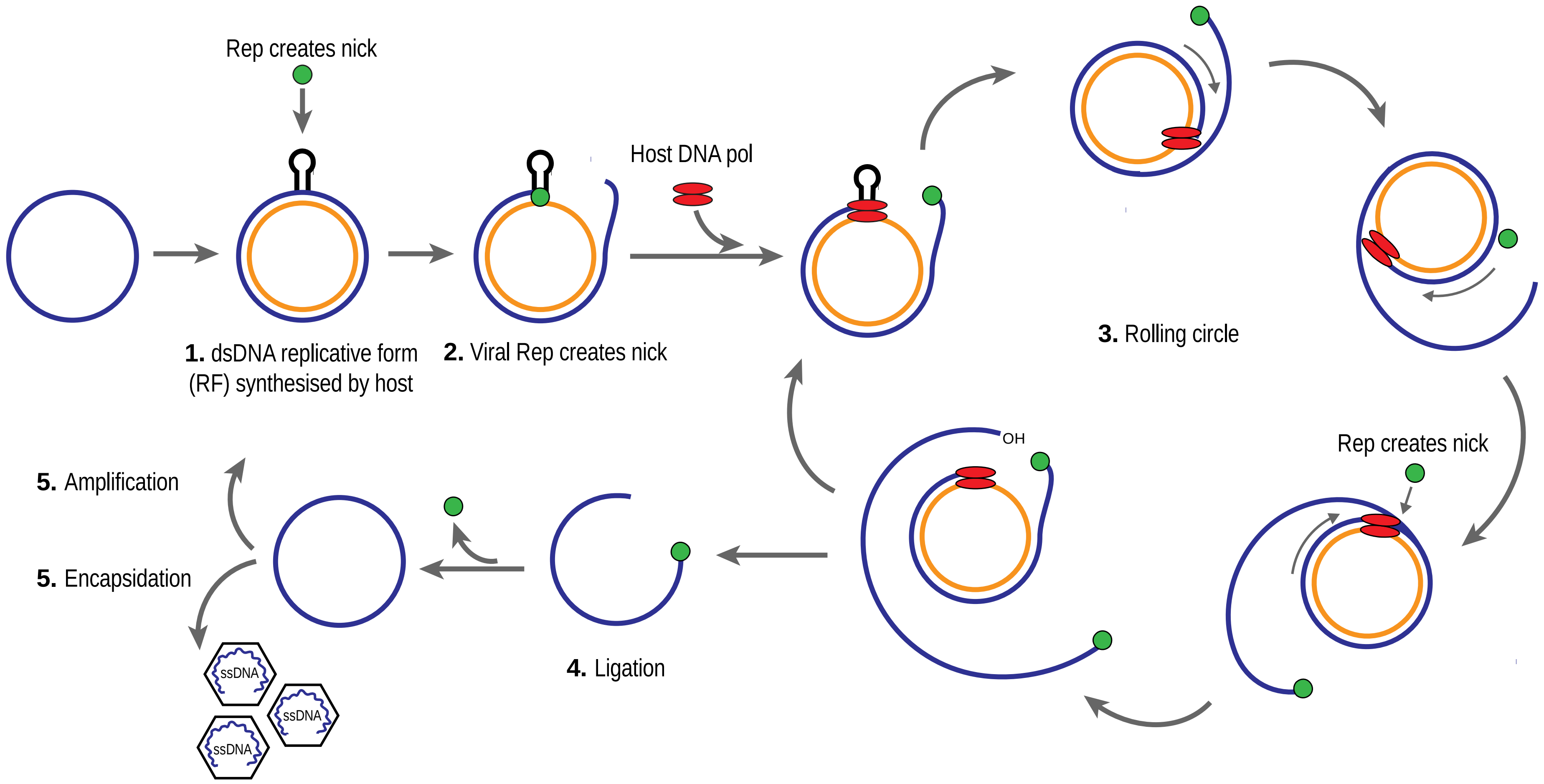


Figure 1.8: Rolling circle replication mechanism in ssDNA viruses. Adapted from Rosario *et al.* (2012a)

1.3.1 Replication-associated protein

The Rep is essential for replication of ssDNA viruses as it initiates amplification of the viral genome. This protein shares close sequence similarity across nanoviruses, circoviruses, geminiviruses, alphasatellites molecules and many novel CRESS DNA viruses. This has been supported by phylogenetic analysis illustrating that the Reps across circular ssDNA viruses are diverse but they share some homology at the protein level, often indicated by the rolling circle replication (RCR) and superfamily 3 (SF3) helicase motifs (Martin *et al.*, 2011).

1.3.2 Motifs and domains

Almost all circular ssDNA viruses that contain a Rep have conserved motifs associated with this protein. These consist of the conserved RCR and SF3helicase motifs. These motifs are conserved across most CRESS DNA viruses (Rosario *et al.*, 2012a).

The N-terminal domain of the Rep contains the RCR motifs and the specificity binding determinants (SPDs). SPDs act as origin recognition elements to allow virus-specific replication by mediating the binding between the Rep and the dsDNA intermediate. The SPDs cluster into two regions near RCR motif one and two of the N-terminal domain (Londoño *et al.*, 2010).The SPDs across geminiviruses, nanoviruses, circoviruses and some novel CRESS DNA viruses are all located at similar positions in the Rep despite their low sequence similarities (Mauricio-Castillo *et al.*, 2014). It is the amino acid residues close to these motifs that are thought to determine the DNA binding properties of the Rep. However, in geminiviruses it has been shown that SPDs can occur outside of this region. The tertiary structures of the SPDs form a beta sheet that are thought to play a role in the high DNA-binding affinity of the Rep. When the SPDs initiate the binding of the Rep to dsDNA sequences that surround the *ori* this is the first step in replication initiation (Londoño *et al.*, 2010).

RCR motif I [F(T/u)(T/L)(T/N)x] (where “u” is a hydrophobic amino acid residue) is highly conserved across most novel CRESS DNA viruses, however, the exact function of this motif is still not understood. It is currently believed that this motif is involved with sequences specific recognition of short repeated nucleotide motifs (iterons) in the intergenic regions (Figure 1.9) (Argüello-Astorga & Ruiz-Medrano, 2001).

RCR motif II is thought to be involved with the protein confirmation and cleavage of the nonanucleotide by acting as a ligand coordinating the binding of metal ions such as Mg^{2+} via two histidine residues (Argüello-Astorga & Ruiz-Medrano, 2001; Koonin & Ilyina, 1992; Nash *et al.*, 2011). It is around RCR motif one and two that the SPDs of a variety of ssDNA viruses have been tentatively mapped to conserved amino acid clusters located in this region (Figure 1.9) (Londoño *et al.*, 2010).

RCR motif III is thought to be the catalytic site of DNA cleavage where a conserved a tyrosine residue covalently binds the DNA during replication initiation (Figure 1.9) (Heyraud-Nitschke *et al.*, 1995).

RCR motif IV is believed to encode a NTP-binding domain commonly found in proteins with helicase, protease and kinase activity. It is thought that this helicase-like activity which displays ATPase activity may play an important role in RCR (Figure 1.9) (Bisaro, 1996; Gutierrez, 1999).

Helicases are a type of enzyme that help to denature or unwind DNA by disrupting of hydrogen bonds between base pairs of double-stranded oligonucleotides using energy derived from ATP hydrolysis. Specifically for ssDNA viruses, helicases are used to unwind the dsDNA intermediate into ssDNA for strand synthesis (Gorbalenya *et al.*, 1990; Ilyina & Koonin, 1992). The helicase domain in ssDNA viruses is characterized by four highly conserved SF3 helicase motifs which are located within a ~100 aa region of the Rep (Gorbalenya *et al.*, 1990). The SF3 helicase family is based on sequence identity between several small DNA and RNA viral Rep helicases and Reps of ssDNA viruses (Gorbalenya *et al.*, 1990; Walker *et al.*, 1982). All these helicases share an NTP binding mechanism.

The structural motifs present in the helicase domain are responsible for NTPase activity which helps the formation of a "P-loop". These motifs include the Walker-A [GxxxxGK(S/T)], Walker-B [hhxh(D/E)(D/E)], and the Walker-C motifs [h(T/S/x)(T/S/x)N]; where "x" represents any amino acid and "h" represents any hydrophobic amino acid residue (Figure 1.9) (Gorbalenya *et al.*, 1990).The Walker-A motif forms part of a P-loop that is thought to act as a deoxyribonucleotide triphosphate (dNTP) binding domain and may also exhibit helicase activity during RCR (Rosario *et al.*, 2012a). The Walker-Band Walker-C motifs are thought to regulate helicase activity through dNTP and the P-loop nucleoside-triphosphate (NTPase) domains (Figure 1.9) (Hickman & Dyda, 2005). Walker-C motif contains a chain of hydrophobic residues followed by an Aspartic acid residue and is found at the C-terminal end of the Rep protein. The fourth motif known as motif B is located between the Walker-B and Walker-C motif. Some of the residues in this motif are involved with ssDNA binding necessary for helicase activity. Other residues are involved with the interaction of ATP with oligomeric nucleotide binding pockets (Koonin & Ilyina, 1992; Story *et al.*, 1992; Yoon-Robarts *et al.*, 2004).

Replication - associated protein

	RCR Motifs					SF3 Helicase Motifs		
	Motif I	Motif II	GRS	Motif III	Oligomerisation 134-180	Walker - A	Walker - B	Motif C
Geminiviruses	**F**L**TY**P I n	**H**L**H**VIV W ILI C L	**R**F**FD**--**FNPNIQ**R**AKS** L -- G	**Y**ME**K**DG ID EE LS		**G**DS**R**T**GKTM** nT I	VI**DD**VD IL SN LS	IF**LCNP** Ls
Circoviruses	**FT**L**NN** W	**H**L**QG**FL ya tm		**YC**S**K**EG G ds A n		**G**PPGC**GKSR** rscs v	VVI**DD**F IM L	III**TSN** LLT VV
Cycloviruses	F**T**L**NN** W W	**H**L**HQG**Y I F		**YC**S**K**SG R En K As		**G**PP**G**S**GKSR** rT T c	VII**DD**F sVL F	**I**FIT**SN** IFe WL
Nanoviruses	**FT**L**N**N i Y F	**H**L**QG**YV I FI V VL		**Y**CS**KE**D aM E		**G**PN**G**GE**GK**S rK nd t D	VIF**D**I IVL V	VIVF**AN** LLL A
Alphasatellite	**FT**LFF INN V	**H**L**Q**G**Y**I I C V		**YC**K**K**DG M EE		**G**PT**GEGK**S**T** sn Ta	IVF**DI** WII C	VFIINI IVLAN HE
Gemycircularviruses	**L**L**TY**A I S	**H**W**H**AFI F CIV	DFF**D**--R**HPNI**EPSA- RII --H KRL-	**Y**AI**K**DG LQ EE		**G**DS**RTGKT**L PT E	VF**DD**I II V	S**IW**I**SN** I C

NB. Bold characters indicate amino acid residue present in all viral sequences for the respective family

Figure 1.9: Conserved rolling circle replication (RCR) motifs and superfamily 3 (SF3) helicase motifs found in geminiviruses, nanoviruses, circoviruses, cycloviruses, gemycircularviruses and alphasatellites replication-associated proteins adapted from Rosario *et al.* (2012b). Conserved residues across all viral sequences within a family are indicated in bold upper case letter, whilst upper case letters indicate higher frequency and lower case indicate low frequency. The amino acid position numbers shown below each motif are derived from representative species from each viral group, and are as follows (top to bottom): *tomato golden mosaic virus* (NC_001507), *porcine circovirus 1* (NC_001792), *cyclovirus PK5222* (GQ404846), and *faba bean necrotic yellows virus* (NC_003560) (Rosario *et al.*, 2012b).

1.3.3 Splicing of the Rep

Splicing of the *rep* to create two ORFs Rep and RepA has been well documented the mastrevirus genus of geminiviruses (Donson *et al.*, 1987; Wright *et al.*, 1997). ssDNA viruses produce two transcripts when replicating, the complimentary sense and the virion sense of the dsDNA molecule. In *Maize streak virus* (MSV; genera: mastrevirus; family: *Geminiviridae*) it has been shown that the splicing of the transcripts of the complimentary sense C1 and C2 ORFs to form a fusion between C1:C2 is necessary in MSV for viral replication. When the complimentary sense 3' transcript is disrupted replication does not occur as functional Rep protein does not form, illustrating that this process is fundamental for MSV replication (Wright *et al.*, 1997). The splicing of the Rep removes the intron to form a fully functioning Rep. RepA is also produced which shares a section of the amino acid sequence of functional Rep. RepA contains the conserved motif LXCXE, which thought to play a role in transactiviating virion-sense expression. RepA does this by binding to the retinoblastoma-related (RBR) protein. The interaction between RepA and RBR generates conditions suitable for supporting viral replication (Liu *et al.*, 1999).

The introns that are spliced out of the complimentary sense can vary in length and are recognized by the flanking donor acceptor sequence sites on either ends. The donor site 5' is recognised by the flanking exon GT while the acceptor site 3' is AT. The introns are often AT rich and can also contain multiple stretches of regions that are T-rich unlike the exons on either side of the intron (Wright *et al.*, 1997).

Although spliced Reps are well documented in mastreviruses putative spliced Reps have also been identified in gemyciruclarviruses viruses (Figure 1.10).

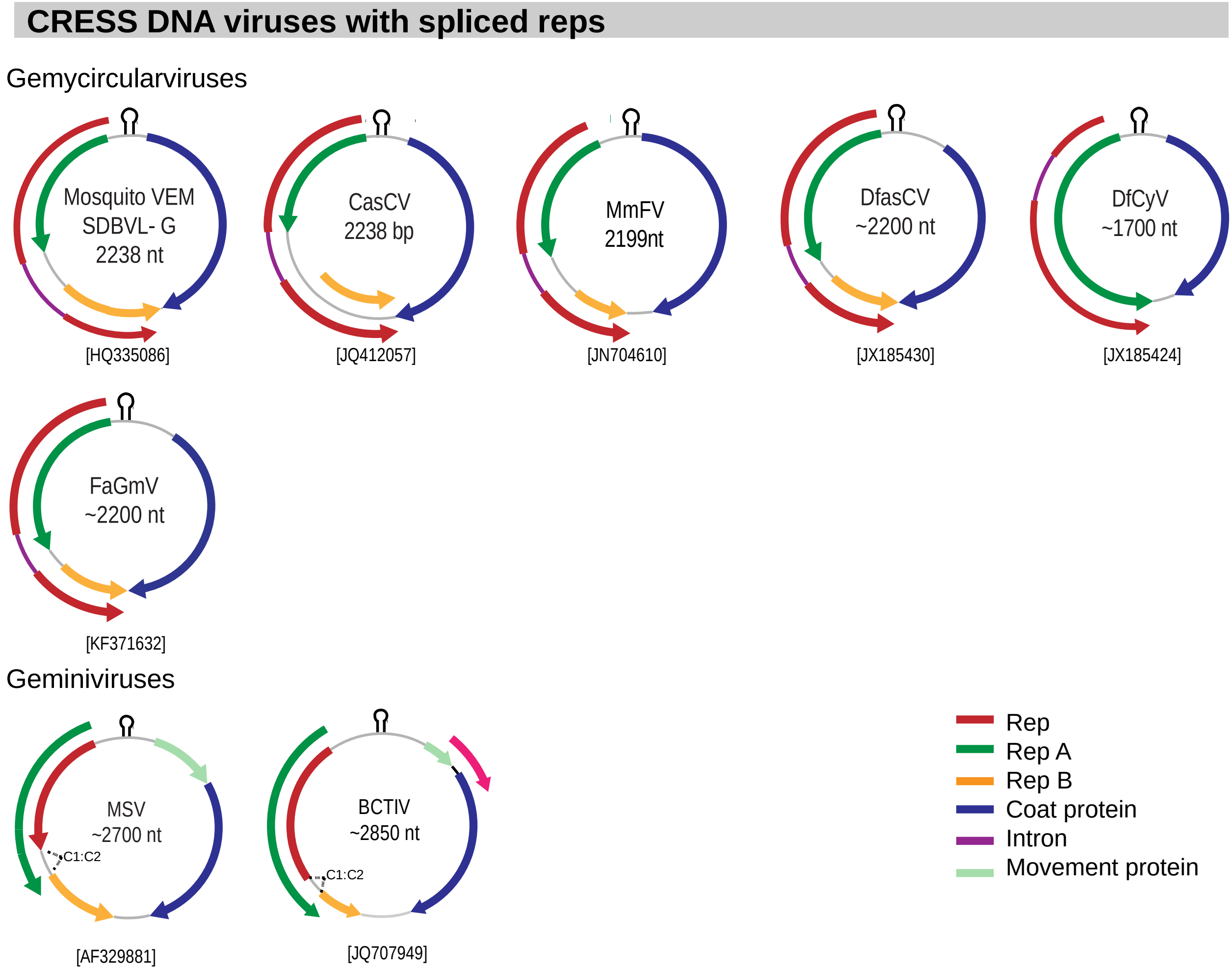


Figure 1.10: CRESS DNA viruses with spliced replication-associated proteins.

1.4 Intergenic region

In geminiviruses and circoviruses the LIR contains transcriptional regulatory elements, promoters, Rep binding sites to initiate replication and on the virion sense strand is the origin of replication. IRs have been extensively studied in nanoviruses and geminiviruses, especially bipartite begomoviruses that contain a common region between the cognate molecules (Argüello-Astorga *et al.*, 1994; Hughes, 2004; Lazarowitz *et al.*, 1992).

The origin of replication is a conserved domain across geminiviruses, nanoviruses, circoviruses, alphasatellites and some novel CRESS DNA viruses and is characterized by a nonanucleotide motif NNNTATTAC located at the apex of a potential stem loop structure known as the N-terminal domain. This stem-loop structure is essential for initiation of replication and termination. The location of the stem-loop within the intergenic region varies amongst CRESS DNA viruses. These structures can be located within either the intergenic region itself or sometimes at the ends of major ORFs encoding for either the Rep or CP. This domain is believed to be of evolutionary significance amongst ssDNA viruses, as high affinity DNA binding specificity determinants (SPDs) have been mapped to two locations in this region near RCR motif one and two (Londoño *et al.*, 2010).

The intergenic region comprises of iterons with are repeated nucleotide motifs located near the stem-loop structure. Iterons vary across ssDNA viruses and play a key role as *cis*-acting elements of viral replication and serve as specific Rep-binding sites (Londoño *et al.*, 2010). This was recently supported by analysis of Rep proteins cycloviruses, which demonstrated that there is a connection between the iteron sequence and the putative Rep SPDs. This was shown in the in the iteron core sequence CGTARC that encodes for Rep proteins displaying a specific SPD region. This SPDs region does differ between cycloviruses suggesting these different SPDs may also have high binding affinity for this iteron and other unique iterons (Dayaram *et al.*, 2013b).

1.5 Coat Protein and unknown ORFs

The CP assembles into macromolecular structures that encapsulates the genetic material of the virus and allows the virus to interact with the host to deliver genetic material. The CP of most ssDNA viruses, apart from novel ssDNA viruses, have been reasonably well studied and are often icosahedral in shape.

The CP among CRESS DNA viruses is highly variable at both amino acid and nucleotide level. However, the major CP in microviruses is highly conserved. The ORFs encoding for the capsid proteins in microviruses are overlapping leading to a densely packages genome (Dokland *et al.*, 1999). Therefore any mutations of insertions or deletions in these ORFs could affect the virion structure and icosahedral symmetry of the capsid (Hafenstein & Fane, 2002). In geminiviruses the coat protein is required for systemic movement especially in viruses with monopartite genomes (Briddon *et al.*, 1989; Pooma *et al.*, 1996).

In circoviruses such as BFDV and PCV and geminiviruses, the CP expression is flanked by regions controlling transcription and polyadenylation signals. The CP contains a high percentage of arginine and lysine residues near the amino terminus. This shows some similarity to protamine block sequence which encodes for channel activity within cells (Niagro *et al.*, 1998; Rosario *et al.*, 2012a).

1.6 Evolution in ssDNA viruses

1.6.1 Genetic drift

Genetic drift is the change in the frequency of an allele in a population due to random sampling and is a key mechanism of evolution. The phylogenetic relatedness in circular ssDNA viruses has been found by analysing the relatively conserved Reps of these viruses. The Reps of circular ssDNA viruses are conserved across most viral families and can often be identified by their RCR and SF3 helicase motifs. Circular ssDNA viruses have nucleotide substitution rates that rival those observed in RNA viruses between 10^{-4} and 10^{-3} substitutions per site per year (Duffy *et al.*, 2008). However, some studies have suggested the "co-divergence hypothesis" which states that all arising mutations and substitutions that appear dominant in populations are purged from the population over a long time period by negative selection and that MSV may have co-diverged with their hosts over millions of years (Wu *et al.*, 2008). However, a long term evolutionary study in MSV and Sugarcane streak Reunion virus (SSRV) showed that neutral genetic drift was the mechanism determining the fate of new mutations in a population (Harkins *et al.*, 2009).

PCV-2 is a virus that is has a recent origin in swine, and is likely to be a result of zoonosis where there was a switch in host from birds to swine (Firth *et al.*, 2009). The substitution rates in PCV-2 (~1.2 x 10^{-3} substitutions/site/year) are similar to those observed in RNA viruses, however, it is possible that the estimations are skewed towards short term mutations due to sampling bias, rather than polymorphisms that will remain fixed long term (Firth *et al.*, 2009).

A further study looking at the movement of BFDV around the world showed that nucleotide substitution rates in BFDV were similar to those in PCV-2 (Harkins *et al.*, 2014). It is thought that these estimates are over inflated due to the short sampling period as some of the nucleotide polymorphisms detected will not ultimately become fixed in the population by genetic drift or selection.

1.6.2 Recombination

One of the ways that recombination occurs in ssDNA viruses is through a host double stranded break point pathway. This occurs during RCR of the virus where a concatemer of dsDNA molecules are formed (Jeske *et al.*, 2001). However, this often triggers host responses that repair these molecules via homology dependent recombination mechanisms (Jeske *et al.*, 2001; Martin *et al.*, 2011; Xu & Price, 2011).

Geminiviruses and circoviruses both use the RCR mechanism to replicate their genomes . It is the RCR mechanism in ssDNA viruses that is thought to play a role in their high recombination rates (Lefeuvre *et al.*, 2009). It has been suggested that this may cause problems with transcription process interfering with replication related complexes which sometimes results in recombination in the complimentary sense genes (Lefeuvre *et al.*, 2009; Owor *et al.*, 2007).

Intra-species recombination is the recombination of genetic material within a species whereas inter-species recombination is between species. Both have been identified in circular ssDNA viruses (Dayaram *et al.*, 2013b; Jeske *et al.*, 2001; Kraberger *et al.*, 2013b; Lefeuvre *et al.*, 2007; Owor *et al.*, 2007; Varsani *et al.*, 2008). In nanoviruses both intra and inter-component recombination events have played a role in driving evolution in the genus *Nanovirus* (Grigoras *et al.*, 2014; Stainton *et al.*, 2012). Majority of the recombination events detected in this study occurred in the non-coding region on the viron strand near the origin of replication, this area is known as a recombination hotspot (Grigoras *et al.*, 2014).

Recombination hotspots are regions of the genome that experience high rates recombination that do not occur by chance, these hotspots are usually located in the intergenic regions as it is believed that ORFs that encode for genes are less tolerable to recombination (Kraberger *et al.*, 2013b; Lefeuvre *et al.*, 2009; Stainton *et al.*, 2012; Stenzel *et al.*, 2014). Recombination events that occur in genes can lead to the misfolding or truncation of proteins.

1.6.3 Reassortment

Reassortment is the exchange of full components among viruses with multicomponent genomes such as nanoviruses and begomoviruses. This phenomenon is known to occur when viruses that are infecting the same cell exchange genetic material (Alberts *et al.*, 1998). This is because individual cells are often infected with many multiple copies of viral components which increases the opportunity for component reassortment (Martin *et al.*, 2011). These genome reassortments allow the opportunity to produce genetically distinct infectious clones that may have traits that allow them to be selected for under different environmental conditions (Unseld *et al.*, 2000). There are examples of this occurring in both begomoviruses and nanoviruses, however, for the genomes to be functional the genes of the reassortment components must be able to interact with other components, this is achieved by each component having a similar common region which enables the components to replicate each other (Chakraborty *et al.*, 2008; Hill *et al.*, 1998; Hu *et al.*, 2007; Stainton *et al.*, 2012; Sung & Coutts, 1995).

1.6.4 ssDNA-viral like sequences in eukaryotic and prokaryotic genomes

Various studies have found that over six different vertebrate species genomes have sequences that are related to those of circoviruses, geminiviruses and nanoviruses (Liu *et al.*, 2011). Horizontal gene transfer and the acquisition of plasmids and viruses into both prokaryote and eukaryote genomes plays an important role in both the evolution of ssDNA viruses and eukaryotes (Liu *et al.*, 2011). In prokaryotes integrated phages account for up to 20% of bacterial genome sequence space and play an important role in their evolution (Canchaya *et al.*, 2003; Casjens, 2003).

In eukaryotes, ssDNA viral-like sequences have been identified in genomes of fungi, animals, plants and protists (Gilbert *et al.*, 2014; Lefeuvre *et al.*, 2011; Liu *et al.*, 2011). These sequences are usually dispersed throughout the non-coding regions of the genome and can be truncated, degraded with stop codons, frame shift mutations and many have insertions and/or deletions. However, in some cases the genes have been found to be inserted into ORF coding regions of the hosts genes or transposons, and have found to be functional in the host (Liu *et al.*, 2011). It is unknown whether these insertions of intact sequences into coding regions represent recent insertions or whether they are highly conserved due to functional constraints (Liu *et al.*, 2011).

There is speculation as to whether many of these viruses have co-evolved with their host eukaryote over a long time period as phylogenetic analysis of the Rep-like sequences from eukaryotes and known viruses demonstrated that higher level eukaryotes are more closely related to circoviruses, geminiviruses and nanoviruses that were infecting higher level eukaryotes (Liu *et al.*, 2011).

Bioinformatic analysis of the C-terminal and N-terminal regions of circovirus Reps from representative viruses, plasmids and bacteria, highlighted that both circovirus and nanovirus like Reps clustered together when the N terminus is analysed; however the C terminal domain circoviruses where grouped with geminivirus-like sequences (Liu *et al.*, 2011). Lui *et al.* (2011) has suggested that as nanoviruses are more closely related to circoviruses in the full Rep phylogenetic tree and N terminal analysis, that it may suggest that the nanoviruses might be the most recent common ancestor of circovirus-like sequences; and the C terminal may be a result of a recombination event between a geminivirus-like Rep or plasmids the ancestral nanovirus like Rep (Liu *et al.*, 2011).

1.6.5 Plasmids

Mobile genetic elements (MGE) occur in all domains of life (Lipps, 2008). MGE are sequences of DNA that may encode for proteins and enzymes that have the ability to move DNA within a genome or between cells (Frost *et al.*, 2005). These elements include plasmids transposons, autonomous retrotransposons and nonautonomous retrotransposons. The movement of MGE between cells can occur via a variety of different mechanisms such as HGT, homologous recombination and integrated conjugative elements.

Plasmids are small pieces of DNA that can replicate independently of the chromosomal DNA within a cell. Plasmids can occur as either covalently closed circular dsDNA molecules or linear dsDNA plasmids and are known to inhabit cells in all three domains of life (Hinnebusch & Tilly, 1993). Comparative genomics has given insights into the relationships of viral proteins with other cellular life forms. This is evident with the similarities between the Rep of DNA viruses and plasmids (Koonin *et al.*, 2006). Phylogenetic analysis of Rep proteins from geminiviruses have suggested that their most recent common ancestor was with Reps encoded on phytoplasmal plasmids (Krupovic *et al.*, 2009b). It has been suggested that their evolutionary relationship evolved from occupying the same ecological niche with both plasmids and geminiviruses replicating in plant and insect host cells. As a tertiary structure of a protein is sometimes more highly conserved than the primary one (Bamford *et al.*, 2005), further analysis using homology based structural modelling showed that the geminivirus CP was the best template of the CP of *Satellite tobacco necrosis virus* (Krupovic *et al.*, 2009b). This has led to speculation that phytoplasmal plasmids gave rise to geminiviruses with the acquisition of a CP encoding gene from ssRNA plant viruses (Krupovic *et al.*, 2009b).

Plasmids have also been reported in many other viral DNA and RNA metagenomic studies (Bench *et al.*, 2007; Breitbart *et al.*, 2003; Breitbart *et al.*, 2002; Kim *et al.*, 2008; Rosario *et al.*, 2009b). Many of the sequences identified in these data sets are associated with ssDNA viruses but share similar properties to elements found in bacteria and eukaryotes (Dayaram *et al.*, 2013a; Liu *et al.*, 2011). This suggests the role of the plasmids between ssDNA viruses, prokaryotes and eukaryotes is not fully understood.

1.7 Determination of novel ssDNA viruses

1.7.1 Primer based amplification

Until recently it was difficult to identity novel and divergent ssDNA viruses using traditional techniques such as sequence specific PCR. Such sequence specific methods utilise specific or degenerate primers, which rely on prior sequence knowledge for specific hybridisation. This limited the discovery of diverse ssDNA viruses, especially when many of these viruses were unable to be cultured in a laboratory setting.

Degenerate primers, designed based on the prior knowledge of a virus sequence, have been successfully used to recover and identify novel ssDNA viruses. They are designed within a conserved region of the viral sequences. Degenerate primers have been used in combination with other techniques such as rolling circle amplification with phi29 DNA polymerase, restriction enzyme digests, cloning, Sanger sequencing and using next-generation sequencing methods to successfully identify novel ssDNA viruses (Li *et al.*, 2010b).

1.7.2 Rolling circle amplification using phi29 DNA polymerase

One of the most valuable tools used towards recovery of highly divergent ssDNA viruses has been the sequence-independent RCA which enriches for circular DNA molecules. The RCA method extends multiple random hexamers annealed to template DNA using bacteriophage *phi*29 DNA polymerase. This results in the synthesis of both strands resulting in multiple replication forks which extend to create a double-stranded DNA product. The use of random hexamer primers eliminates the need for prior sequence knowledge. Bacteriophage phi29 DNA polymerase is isolated from Bacillus subtilis and used to amplify circular DNA templates (Blanco et al., 1989). This polymerase is unique due to its high fidelity proofreading abilities, as well as its ability to perform strand displacement DNA synthesis for more than 70,000 nucleotides without displacing from the template DNA (Blanco et al., 1989). RCA occurs by displacement of the non-template strand producing tandem copies of circular DNA resulting in the exponential amplification of the template DNA. This limitations of this method are that it preferentially amplifies circular ssDNA templates leading to a bias in amplification and can also create chimeras (Lasken & Stockwell, 2007).

However, this is beneficial to discover novel ssDNA viruses with circular genomes as it is used to enrich circular ssDNA before metagenomic sequencing (Edwards & Rohwer, 2005; Haible et al., 2006; Kim et al., 2011; Kim et al., 2008). RCA has been used to create DNA libraries in various studies that have gone on to use NGS technologies such as Roche 454 Pyrosequencing or Illumina sequencing on environmental samples including, fresh water (Fancello et al., 2013; Roux et al., 2012), marine (Angly et al., 2006; Breitbart et al., 2002), Antarctic lakes (López-Bueno et al., 2009), various faecal sources (Kim et al., 2011; Li et al., 2010a; Li et al., 2011b; Li et al., 2010b; Sikorski et al., 2013b; Sikorski et al., 2013d), invertebrates (Ng et al., 2011a; Ng et al., 2011b; Rosario et al., 2012b), sea turtle tissue (Ng et al., 2009) and rice paddy soil (Kim et al., 2008).

1.7.3 Viral purification and nucleic acid enrichment

Viral metagenomics often relies on high quality of the sample preparation which involves extraction, purification and concentration of the viral nucleic acid. As metagenomic studies sequence all DNA from a sample, it is vital that the nucleic acids from eukaryotic and microbial cells are removed before the viral DNA is extracted so the sample only contains intact virions for enrichment. This is often done before lysis of the viral capsid by treating the sample with DNase I which reduces the amount of free nucleic acid. RNase is also used to remove free RNA in a sample; however, as some RNA viruses have RNA in the capsid structure this can often lead to loss of viral particles (Thurber *et al.*, 2009).

There are many different methods that can be used to purify viral particles from samples. Some of the most common ways to do this are tangential-flow filtration (TFF), polyethylene glycol (PEG) precipitation and viral staining (Thurber, 2011). TFF is where the sample (filtrate) is pushed through filters; the pore sizes in the filter allow the viral particles to pass through whilst filtering out any large particles. In order to enrich for virus particles, the filtrate is often mixed with a solution such as SM buffer [0.1 M NaCl, 50 mM Tris/HCl (pH 7.4),10 mM $MgSO_4$], then passed through different filter sizes sequentially beginning with filters with larger pore sizes (Svraka *et al.*, 2010; Thurber, 2011).

Caesium chloride (CsCl) ultracentrifugation can also be used to purify viral particles based on density and is used after TTF and PEG precipitation (Thurber *et al.*, 2009). This method relies on the physical properties of virons. Depending on the density of the virus being targeted this determines the solvent, speed of centrifugation, and types of gradients. Centripetal and diffusive forces create a density gradient that allows that separation of viral particles based on their molecular density (Thurber *et al.*, 2009).

Viral nucleic acid is then extracted. This is often done with commercial kit based methods such as the Roche High Pure Viral Nucleic Acid Kit (Roche) or QIAampMinElute Virus Spin kit (Qiagen) (Svraka *et al.*, 2010; van den Brand *et al.*, 2011). However, the use of these silica binding spin columns for viral particle purification have been questioned after a study found the silica column were contaminated with a parvo-like hybrid virus that was obtained during elution with water through the spin column (Naccache *et al.*, 2013). In depth analysis showed that this virus was present in environmental metagenome libraries created from costal marine water off North America where the diatoms from this water are used to generate the silica matrix used in the columns. This suggests the columns were contaminated during manufacturing (Naccache *et al.*, 2013).

These NGS sequencing technologies rely on the random amplification of all DNA or RNA in a sample to create libraries. Random PCR is a sequence independent amplification technique that uses randomly generated primers to amplify all nucleic acid present in a sample. This technology relies on the amplification of DNA in a sample to create cDNA libraries for NGS. Amplification of the viral nucleic acid is often then needed as the yield of DNA is too little for metagenomic sequencing and PCR based methods. Here methods such as RCA are used to independently amplify circular DNA molecules using the Phi29 DNA polymerase and randomly generated primers. This amplification process increases the total DNA yield and also purifies the DNA but removing inhibitors such an enzymes or PCR inhibitors (Svraka *et al.*, 2010). This amplification creates the DNA libraries necessary for further NGS sequencing. GenomiPhi (GE Healthcare) again uses Phi29 DNA polymerase, but can increase the concentration of linear DNA molecules so is often used for whole genome amplification using isothermal strand displacement.

1.7.4 Sequencing platforms

1.7.4.1 Sanger sequencing

Traditional DNA sequencing techniques were introduced by Sanger *et al.* (1977), which was based on chain-terminating dideoxynucleotides (dNTPs) and deoxynucleosidetriphosphates (ddNTPs) using DNA polymerase. The ddNTPs were originally tagged using radioactive phosphorus but modern Sanger sequencing the ddNTPs are fluorescently labelled to identify each nucleotide and when incorporated causes chain termination. The DNA sample is separated into four different sequencing reactions containing all four deoxynucleotides and the DNA polymerase. To each reaction one of the dideoxynucleotides is added along with the three other standard nucleotides. Template DNA extension then takes place. The resulting DNA fragments are then denatured and run on a gel. The resulting bands on the gel correspond to the appropriate nucleotide in the sequence (Sanger & Coulson, 1975). Using the conventional Sanger method up to 1 kb of nucleotides could be sequenced (Sanger & Coulson, 1975). However, this technique is limited to processing individual samples.

1.8 Next generation sequencing

The introduction of Next Generation Sequencing (NGS) platforms allows larger numbers of samples to be run in parallel (Shokralla *et al.*, 2012). There are two main steps involved in all NGS techniques; the first is the library preparation, which involves amplification and then fragmentation of DNA molecules being sequenced with the ligation of specific oligo adapters at both ends of the DNA; the second is the detection of nucleotides that become incorporated. Unlike traditional Sanger sequencing methods, prior knowledge of the target sequence is not necessary which allows for novel genomes to be targeted (Adams *et al.*, 2009).

The further development and commercialisation of NGS technologies have enabled them to become widely available and cost effective for many researches. This technology has changed the way viral genomes are determined and as a result has led to a better understanding of viral communities within environmental samples. NGS have enabled the study of many diverse viral communities in soil (Kim *et al.*, 2008), the human gut (Breitbart *et al.*, 2003), sea water (Angly *et al.*, 2006), freshwater (López-Bueno *et al.*, 2009; Roux *et al.*, 2012) and many more.

There are several NGS platforms currently available. They all differ slightly in their chemistry, hardware and software but essentially they perform similar functions. Below is an overview on some of the most commonly used platforms for looking at novel ssDNA genomes.

1.8.1 NGS sequencing platforms

1.8.1.1 Roche/454 pyrosequencing

This was the first sequencing platform to be made commercially available. This sequencing platform uses a technique known as sequencing-by-synthesis also known as pyrosequencing (Mardis, 2008a). This sequencing method first involves a cDNA library prep, then the DNA molecules into in the library being sequenced. DNA adaptors are then ligated to the fragmented DNA; these are complimentary to the DNA adaptors on the surface of small enzyme beads with an oil/water emulsion. These beads are then washed across a PicoTiterPlate device with the beads falling into wells on the plate when centrifuged. The plate is then placed in the Genome sequencer FLX Instrument (Mardis, 2008a). During the sequencing run, buffers containing nucleotides are washed across the plate. These nucleotides bind complimentary to the fragmented DNA and generate a light signal that is recorded by a CCD camera (Liu *et al.*, 2011; Mardis, 2008b). This method can produce reads of up to ~1000 base pairs (Zhang *et al.*, 2009). However, the disadvantages to this sequencing method are that it does not accurately sequence homopolymer stretches of DNA (Mardis, 2008a).

1.8.1.2 ABI SOLiDAnalyzer

Life Technologies introduced the commercially available ABI SOLiD platform in 2006. SOLiD stands for Sequencing by Oligonucleotide Ligation and Detection. This metagenomic technique involves a library prep were DNA is sheared into fragments and then specific oligo adaptors are ligated to the ends of these fragments. These adaptors are complimentary to adaptors on small magnetic beads, so the start sequence of every fragment is identical. The adaptor ligated fragments are then amplified by an emulsion PCR and the resulting amplicons attached to the beads are then covalently bound to a glass slide. A universal sequencing primer then hybridises to the adaptor sequences. A set of four fluorescently labelled di-base probes compete to ligate to the sequencing primer. The di-base probes consist of three degenerate bases followed by two specific bases followed by three degenerate bases. In the first sequencing cycle the di-base probes are ligated to the universal primer, the last three degenerate bases of the probe are then cleaved. A fluorescent marker of the two specific bases is then detected. At the end of the cycle, the primer and incorporated oligonucleotides are denatured and washed away. The second cycle then starts ligating the primer to the n-1 position in relation to the first cycle. This is followed by the ligation, cleavage and florescent detection steps determining the nucleotide sequence of another frame. After five successive sequencing cycles the complete sequence of the DNA fragment can be determined. SOLiD unlike 454 pyrosequencing sequencing is very accurate (99.94%) as every nucleotide is sequenced twice during the process making it easy to sequence homopolymers accurately (Mardis, 2008a; Zhang *et al.*, 2011).

1.8.1.3 Illumina/Solexa Genome Analyzer

Illumina sequencing was introduced in 2006 and uses sequencing by synthesis (SBS) to produce reads of ~50-200 bp (Shokralla *et al.*, 2012). The Illumina platform works by combining both SBS with bridge amplification that occurs on the surface of a flow cell which is divided into eight separate lanes. The surface of the flow cell has specific oligos that are complimentary to the adaptors that have been ligated onto both ends of the fragmented DNA during the library prep. The fragmented DNA then becomes hybridised to the complimentary oligos on the flow cell surface during a heating and cooling process. Subsequent bridge amplification of these ligated DNA fragments to the flow cell surface form millions of clusters. The sequencing involves using a polymerase to attach fluorescently labelled nucleotides that have a blocking modification that allow only a single base to be attached at a time to the cluster. After the addition of each nucleotide they are identified via imaging of the fluorescent nucleotide when they are excited. The blocking 3' OH of the nucleotide is then removed and the next nucleotide added (Radford *et al.*, 2012; Shokralla *et al.*, 2012). The advantage of Illumina sequencing method is that the addition of each nucleotide individually means that homopolyer regions of the DNA are accurately sequenced. It also produces more reads per run than when compared with other NGS technologies such as 454 pyrosequencing (Roche Technologies). However, the draw backs to Illumina are the short read lengths with the error rate increasing with longer reads (Zhou *et al.*, 2010).

1.8.1.4 PacBio

A third generation sequencing instrument PacBio RS: a real time single molecule sequencer was recently released by Pacific Biosciences (Koren *et al.*, 2012). The aim of this new technology was to address the problems often incurred through previous NGS platforms such as the need for amplification of source DNA often leading to bias in coverage and amplification artifacts of the DNA (Dohm *et al.*, 2008; Niu *et al.*, 2010). Other problems are the short read lengths generated which often make assemblies more difficult (Kingsford *et al.*, 2010).

The goal behind the PacBio RS system is to produce long sequencing reads in a short period of time that require no amplification beforehand so it reduces compositional bias. The PacBio system relies on the zero-mode waveguide (ZMW) which guides light energy into a volume that is small enough to observe only a single nucleotide of DNA being incorporated by DNA polymerase. Uninterrupted template-directed synthesis of the template DNA incorporates fluorescently labelled dNTPS using DNA polymerase. The fluorescent tags are detected when incorporated and the dye tag of the nucleotide corresponds to one of the four bases. When the dNTP is incorporates, the fluorescent tag is then cleaved off and diffuses out of the observation area of the ZMW (Eid *et al.*, 2009).

Single molecule sequencing is useful for transcriptome and *de novo* genome assembly as it is possible to accurately sequence areas of the genome that are high in complex polymorphisms that often occur across an entire genome (Koren *et al.*, 2012). The limitations to this sequencing method is the high nucleotide error rate which generate reads that are only between 82.1% to 84.6% accurate (Koren *et al.*, 2012).

1.8.2 Bioinformatics of NGS data

As millions of base pairs of data are generated from NGS sequencing, a lot of bioinformatics is need in order to deal with this data load. First the data is subject to a filtering process which removes low quality reads (Liu *et al.*, 2012). The remaining reads then have their adaptor sequences removed and paired end reads are joined. The individual short reads are then assembled using *de novo* software. These types of software use algorithms to find overlapping regions of the different reads eventually forming larger contigs (Radford *et al.*, 2012). All NGS software assembler fall into three categories these are: Overlap/Layout/Consensus (OLC) which rely on an overlap graph, De Bruijn Graph (DBG) which uses some form of K-mer and are useful for data sets with high coverage with short read lengths and Greedy graph algorithms that use either DBG or OLC (Miller *et al.*, 2010). All of these assembly software’s rely on K-mers which is the sequence of K base calls, where K is any positive integer. This means were two reads which are similar and are overlapping they will have shared K-mers (Miller *et al.*, 2010).

1.8.2.1 ABySS

Assembly By Short Sequencing (ABySS) is a software developed to deal with the assembly of data sets with millions of short reads using an algorithm (Simpson *et al.*, 2009). ABySS uses the DBG algorithm with the initial data set produced has all possible substrings of length *K* (K-mers). From this data all read errors are removed and then the contigs are extended until they come to an end with poor coverage. The extended contigs are then assembled by identifying overlapping regions (Simpson *et al.*, 2009). This software is able to quickly and accurately assemble millions of reads and is useful to identify novel genomes where reference sequences are unavailable (Simpson *et al.*, 2009).

1.8.2.2 MIRA

MIRA is a multi-pass DNA sequence data mapper and assembler for sequencing data from Sanger, 454, Illumina, IonTorrent and PacBio and is freely available. MIRA uses OLC algorithm to assemble NGS data. MIRA is conservative read assembler often creating more fragmented transcripts, which makes this assembler better for detecting variation as it produces less chimeric reads (Miller *et al.*, 2010). However, this also makes it harder to complete assemblies as contigs are much shorter (Mundry *et al.*, 2012).

1.8.2.3 Velvet

Velvet is an assembler software programme that uses a DBG algorithm. This assembler software is particularly useful for short paired end reads produced from NGS such as Illumina. This is because the software reduced the graph complexity by implementing a heuristic techniques that read coverage, sequences identity, paired-end read constraints and graph topology (Miller *et al.*, 2010). This enables Velvet software to be used to for assembly of both prokaryote and eukaryote genomes as it is able to remove errors and resolve number of large repeats that often result from paired-end reads (Zerbino & Birney, 2008).

1.8.2.4 SOAP denovo

SOAP denovo program is freely available and again utilises both the OLC and DBG algorithm. SOAP uses pre-set thresholds for K-mer frequencies and then builds the DBGs. SOAP first builds contigs from the reads using the DBG algorithm, it then builds scaffolds using the contig consensus sequences including reads not used in the DGB. It then creates contig graphs and reduces complexity by removing repeats in a similar way to Velvet by bubble smoothing which removes error induced paths. SOAP is useful for assembling NGS data for large genomes as it utilises both algorithm techniques (Miller *et al.*, 2010).

NGS is a powerful tool for identifying novel viral genomes in the environment, having helped uncover the diversity and distribution of many circular ssDNA viruses present in numerous ecosystems. This increased knowledge of their diversity has then contributed to shedding light on the evolutionary histories of these viruses.

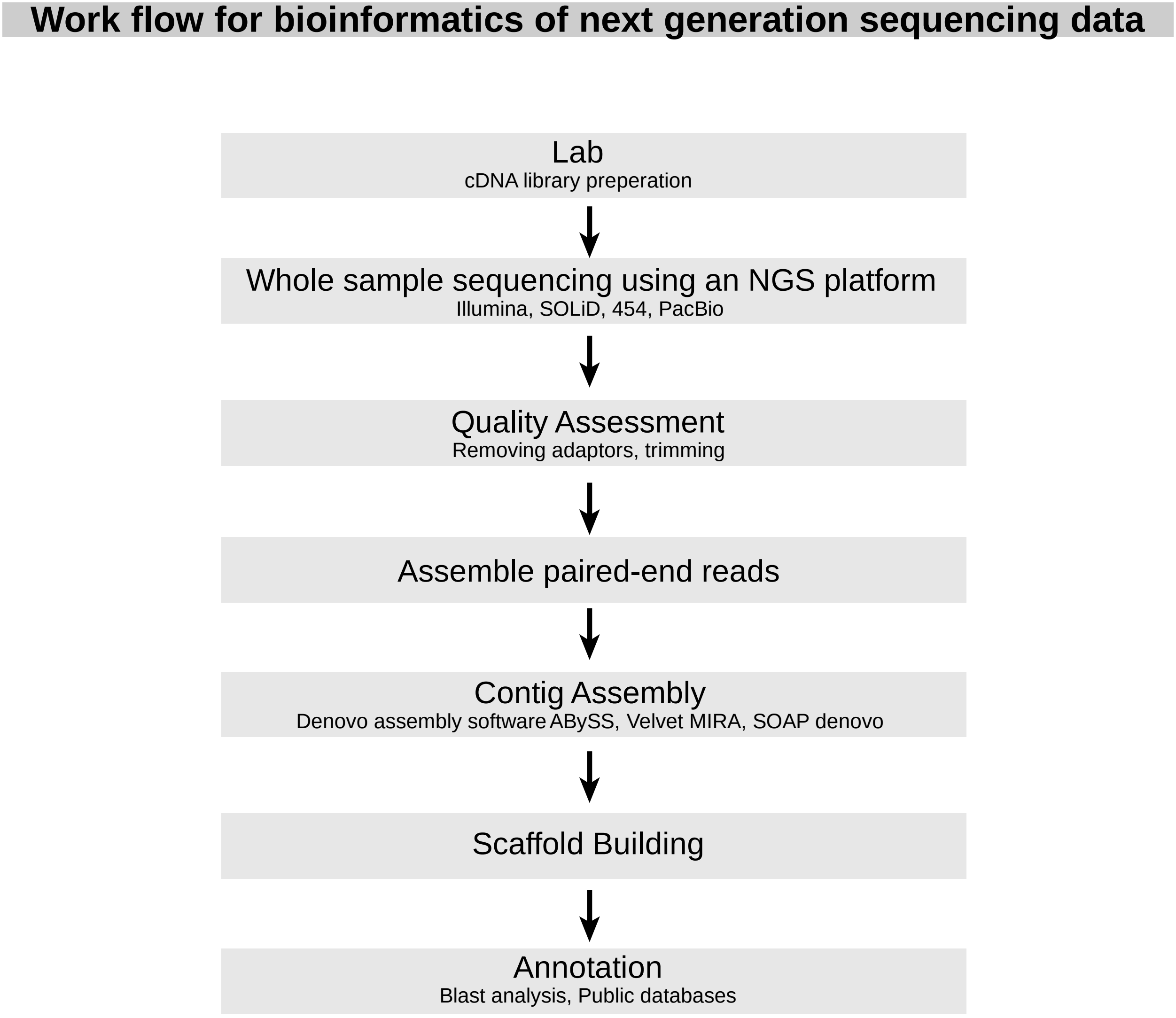

Work flow for bioinformatics of next generation sequencing data
Lab
cDNA library preperation
Whole sample sequencing using an NGS platform
Illumina, SOLiD, 454, PacBio
Quality Assessment
Removing adaptors, trimming
Assemble paired-end reads
Contig Assembly
Denovo assembly software ABySS, Velvet MIRA, SOAP denovo
Scaffold Building
Annotation
Blast analysis, Public databases

Figure 1.11: Workflow of the process of analysing NGS data.

"

1.9 ssDNA viruses in the environment

The diversity of ssDNA viruses is only just beginning to be unravelled by analysis of ssDNA viruses in the environment.

Viruses are thought to be some one of the most abundant entities on Earth and thus are believed to have a significant impact on the ecology. However, many viruses cannot be studied due to lack of appropriate cultureable hosts. With ease of access to new sequencing technologies such as NGS, large genome wide studies have been conducted to identify ssDNA viruses in a range of different environments. This has led to an increase in the number of ssDNA virus genomes being determined and made available in public domains.

1.9.1 Terrestrial soil

Very little is currently known about the virome of soil. It is estimated that the viral load in soil is 3.4 to 4.6 times higher than bacteria; it is thought that they may have a significant impact on the soil microbial ecosystem (Kim *et al.*, 2008). The first environmental study looking at the diversity of ssDNA was carried out in soil from a rice paddy in Daejeon, Korea (Kim *et al.*, 2008). This study used a combination of RCA with phi 29 polymerase and random hexamers to amplify the viral DNA. After the amplification, the viral DNA was sheared using a Hydro Gene instrument into small fragments that were then shotgun cloned and sequenced. The resulting reads were then assembled into larger contigs. Where the contigs overlapped at the beginning and end, these sequences were assumed to be circular ssDNA viruses. From these contigs, back-to-back primers were designed then a PCR was performed to amplify the genomes, these were then cloned and sequenced using traditional Sangar techniques. More than 60% of the viral genomes recovered did not share any significant similarities when compared to the viral databases using BLAST and those that did showed low similarity with Reps of ssDNA viruses. This suggested the extent of ssDNA viral genomes that were identified using the combination of techniques were extremely diverse and highlighted the complexity of viral diversity in soil. It also demonstrated how the combination of techniques can be used to identity novel viruses that were previously unculturable (Kim *et al.*, 2008).

1.9.2 Marine sediments

Viral ecology in marine environments have been rigorously studied, however, most of the current knowledge about viral ecology is limited to the euphotic zone in the ocean (Corinaldesi *et al.*, 2003). As the sea floor takes up over two thirds of the Earth's surface the influence of the virome in the sediment of this unique ecosystem are of extreme interest. A handful of studies have shown that viral infections play a major role in the microbial community in deep sea sediments as high viral loads have been associated with large prokaryotic biomass (Danovaro *et al.*, 2008). Yoshida *et al.* (2013) investigated the virome of the sediment from three locations in the northwest Pacific. Each of the three sites Shimokita, Ogasawara and Mariana; were geographically and geologically diverse locations. Metagenomic techniques were employed using Roche 454 pyrosequencing combined with RCA to examine ssDNA viruses and linker-amplified shotgun library (LASL) for dsDNA (Yoshida *et al.*, 2013). BLASTx analysis of the libraries created from the pyrosequencing revealed that only 24-30% of these reads had similarities to sequences deposited in GenBank. Of these sequence reads, viral hits were only reported for only 10% of the reads from Ogasawara sample and 4% of the Mariana Trench sample, many of these had low hits to geminiviruses, circoviruses, nanoviruses, microviruses and novel CRESS DNA viruses. These viral reads were then examined further using MetaVironline tool (Roux *et al.*, 2011), which enabled the detection of viral marker genes in ssDNA viruses such as the conserved major CP of Microvirdae and the putative Rep of eukaryotic ssDNA viruses. This was one of the first studies that used motifs in sequences to sort and align reads. Phylogenetic trees were then constructed of the sequences with hits to gene markers. This analysis showed that there is huge genetic diversity of the virome in both deep sea and shallow seafloor sediments when looking at both the CP and Rep when compared to previously identified ssDNA viruses (Yoshida *et al.*, 2013).

1.9.3 Diatoms

Diatoms play an important role as primary producers in the oceans. As well as being the primary producer of oxygen in the atmosphere they also account for huge primary production in the oceans (Nelson *et al.*, 1995; Werner, 1977). Viruses play a huge role in the mortality of diatoms. The first viruses to be isolated that was associated with the mortality of diatom was

Rhisoleniasetigera RNA virus (RsRNAV). This was a ssRNA viruses with an 11.2 kb genome that was isolated from was Ariake Sea, Japan (Nagasaki *et al.*, 2004).

Chaetoceros salsugineum DNA virus (CsalDNAV) is an 38 nm icosahedral virus was later isolated from samples from the Shiostuka River in the Ariake Sound in Japan (Nagasaki *et al.*, 2005). The genome of this virus was ~6 kb circular ssDNA as well as a linear segment of 997 nucleotides. Phylogenetic analysis of the putative replication associated protein showed some homology to the Rep of circoviruses (Nagasaki *et al.*, 2005). Other diatom DNA viruses have also been isolated including Chaetoceros lorenzianus DNA virus (ClorDNAV) (Tomaru *et al.*, 2011b) and Chaetoceros tenuissimus DNA virus (CtenDNAV) (Tomaru *et al.*, 2011a). These viruses were isolated from Hiroshima Bay, Japan. Both were between ~34-37 nm in diameter with circular ssDNA genomes between ~5.6–5.8 kb. Both these viruses encode for three major ORFs with at least one encoding for a putative replication associated protein that are related to other ssDNA viruses. All diatom infecting viruses that have been isolated to date cause lysis of different diatom species. The role diatom infecting viruses play in regulation of blooms is still poorly understood although it is believed they play a significant role in host population dynamics (Brussaard, 2004; Brussaard & Martinez, 2008).

A further study by Yoon *et al.* (2011) used whole genome shotgun sequencing from three individual picobiliphte cells isolated from seawater collected from Boothbay Harbor in the Gulf of Maine. The cells recovered from the water were identified using 18SrDNA-based fluorescent *in situ* hybridization probes after whole genome amplification using MDA. The single cell sequencing data from one of the cells was dominated by sequences related to ssDNA viruses (Yoon *et al.*, 2011). From the sequencing data an unknown ssDNA virus was isolated. This virus was ~1.8 kb and had two major ORFs that were bi-directionally transcribed. A BLASTx search against the NCBI RefSeq viral database showed one of the ORFs shared similarities to the Reps of nanoviruses and circoviruses. Further phylogenetic analysis of the putative Rep showed is grouped with other Rep-like sequences isolated other ocean metagenomes. The use of shotgun sequencing of uncultured marine picobiliphytes enable the identification of a novel ssDNA virus and demonstrated the distinct interactions in these cells (Yoon *et al.*, 2011).

1.9.4 Water

Metagenomic sequencing has proved a vital tool in helping elucidate the viral diversity and abundance of different aquatic environments (Edwards & Rohwer, 2005). Despite the ability of viruses to disperse across long ranges there have been studies that have shown there is not much dispersal between marine and freshwater environments (Logares *et al.*, 2009). The first metagenomic study to look at viral communities in aquatic ecosystems was carried out over ten years ago. This showed that there is a vast reservoir of viral diversity that is yet to be explored (Breitbart *et al.*, 2002). To date the aquatic viral metagenome studies carried out have been in a range of unique environments including Antarctic lakes (López-Bueno *et al.*, 2009), fresh water lakes Lake Bourget and Pavin, France (Roux *et al.*, 2012), reclaimed water in the USA (Rosario *et al.*, 2009b), estuarine water from Cheapeake Bay, USA (Rosario *et al.*, 2009a), coastal water of British Columbia and the Sargasso sea (Angly *et al.*, 2006) Yellowstone hot springs, USA (Schoenfeld *et al.*, 2008), confined aquifers in South Australia (Smith *et al.*, 2013) and four perennial pond in central Sahara, Mauritania (Fancello *et al.*, 2013). The viromes from these metagenome studies have been dominated by *Caudovirales*, *picornavirales* with others having a large presence of microviruses and other novel CRESS DNA viruses.

1.9.4.1 Confined aquifers

Confined aquifers are bodies of water that lie below the ground surface and are permanently or semi-permanently separated from ground water by geological barriers (Borchardt *et al.*, 2007). A recent study looked at the viral diversity and abundance from samples collected from a deep confined and unconfined aquifers in South Australia (Smith *et al.*, 2013). Metagenomics techniques were used to construct a viral community profile from the viral metagenomic sequences obtained from both the aquifers. At least 53% of the assembled viral contigswere unclassified viral sequences. Most of the viral sequences present in both aquifers were ssDNA viruses with 72% in the confined and 47% in the unconfined aquifer respectively; however, this could be a result of MDA which preferentially amplifies ssDNA viruses. The lack of dsDNA viruses in these ecosystems is also thought to be a result of the aquifer matrix excluding larger viruses.

The viral metagenomes from both aquifers were compared to other metagenomes from aquatic environments. The confined aquifer was most similar to the viral sequences in the metagenome from reclaimed water in Florida (Rosario *et al.*, 2009b) despite the geographical proximity of the unconfined and confined aquifers. The lack of similarity between these aquifers further suggests that the viruses in the confined aquifer were not from the unconfined aquifer above, indicating that this viral community may be a result of long-term survival rather than being introduced. The similarity in metagenomes is thought to be a result of the extensive filtration within both environments, suggesting that the microbial composition of deep aquifers can be altered. Most of the viruses isolated in the confined aquifer were somewhat similar to sequences of *Circoviridae, Geminiviridae* and *Nanoviridae* which are known to infect plants animals. However; neither plants nor animals have been found in the confined aquifer that suggests the viruses were introduced exogenously and as the confined aquifer is over 1500 years old is suggests the idea of long term survival of these viruses (Smith *et al.*, 2013).

1.9.4.2 Marine

The virome of marine environments is dominated by bacteriophages belonging to the family *Microviridae* (Breitbart *et al.*, 2002). These viruses play a huge role in regulating other microbes in their environment as they kill both heterotrophic and autotrophic microbes (Weinbauer, 2004) and have a huge influence on global biochemical cycles (Fuhrman, 1999). Marine viral species have been shown to be largely dispersed with many being shared between different oceanic regions (Angly *et al.*, 2006). However, the compositions of the marine viromes vary between different geographical locations which are often a result of selective pressure. Cyanophages are a new clade of ssDNA phages were common in a sample from the Sargasso sea, whereas prophage like viruses were most common in the Arctic oceans (Angly *et al.*, 2006). Unlike many of the freshwater viromes, metagenomic studies to date suggest that the marine virome is predominately made up of bacteriophages, however, the dispersal of different phages differs between ecosystems with different environmental conditions creating a selective pressure on certain viral phenotypes (Angly *et al.*, 2006).

Breitbart *et al.* (2002) used shotgun sequencing to clone and sequence two uncultured aquatic viral communities from Mission Bay and Scripps Pier, San Diego, USA (Breitbart *et al.*, 2002). The results from this study again illustrated that most of the sequences recovered were uncharacterised with over 65% of the shotgun sequencing libraries created from the samples showed no similarities to previously described sequences in GenBank. Most of the sequences with significant hits showed similarities to phages; however, the composition of phage families between the two sample sites differed with both sites showing high viral diversity.

Following this study, Labonté and Suttle (2013) used metagenomics to look at viral communities from two different sampling location that included a temperate site the Saanich Inlet in the Strait of Georgia and a subtropical sit in the Gulf of Mexico (Labonté & Suttle, 2013). This study focused on looking at the diversity ssDNA viral communities at both sample sites. The results from this study again illustrated that there is huge viral diversity within both sample sites, with over 608 viral sequences recovered. However, this study illustrated that ssDNA viruses were dominant in these marine viral communities unlike previous marine metagenomic studies were dsDNA phage comprised majority of the virome (Angly *et al.*, 2006; Breitbart *et al.*, 2002). Of these genomes, over 70% of the ssDNA sequences showed no similarities to previously reported viral sequences indicating that many of these viruses may represent new viral families, with many viral sequences having genes that potentially encoded for Reps and CPs. The high viral diversity and abundance of unknown ssDNA viruses indicates that these viruses may infect a large variety of organisms (Labonté & Suttle, 2013).

Overall these studies have shown that the diversity of ssDNA viruses in different marine environments with some indicating that viral diversity, such as genome size and viral abundance can change seasonally.

1.9.4.3 Fresh water Lakes Bourget and Pavin

Roux *et al.* (2012) undertook one of the first studies to examine the idea that fresh-water specific viral clades may exist due to host-specific interactions in these environments. These authors examined this idea by using a viral metagenomic approach to investigate the viral composition and diversity of Lakes Bourget and Pavin in France. This study highlighted that oligotrophic lake (lakes with low species richness) also had lower species richness when examining the virome, when compared with a mesotrophic lake. Thus, indicating that there may be link between low primary productivity and nutrient content in an environment and the species richness of its virome. It also revealed that there is a significant similarity between viromes of related environments despite the geographical distance between the sample locations, indicating that specific viral clades may be found in different aquatic environments such as freshwater, marine and hypersaline.

1.9.4.4 Antarctic lakes

Antarctica is the most remote continent on Earth and has been geographically isolated for millions of years. Because of this, it comprises many microbial ecosystems that have adapted to the harsh conditions such as: low nutrient levels, low moisture levels, low temperatures and extended periods of darkness during the winter. A study by Lopez-Bueno *et al.* (2009) examined the viral diversity of Antarctic Lake Limnopolar near Byers Peninsula and looked into whether there were seasonal changes in the virome composition. The study used Roche 454 pyrosequencing in combination with RCA.

The virome of Antarctic lakes revealed huge genetic diversity spanning many different viral families. Of the 20 million base pairs from the sequencing less than 3% showed any homology to sequences recovered from other viromes sequenced from aquatic ecosystems. Further investigation into these sequences illustrated that that the Antarctic lake virome is dominated by ssDNA viruses that are mostly related to eukaryotic viruses including animal and plant viruses and some dsDNA viruses that infect algae. This contrasts with other aquatic viral metagenomic studies that are commonly dominated by bacteriophages infecting prokaryotes (Angly *et al.*, 2006; Desnues *et al.*, 2008; Rodriguez-Brito *et al.*, 2010). This studied also found that the seasons impact the virome of the Antarctic lakes. In summer there is a shift from smaller ssDNA viruses particles <30 nm in diameter to larger particles >50 nm, including tailed phages. Overall this study has shown that even in areas with harsh conditions there can be huge genetic diversity and richness of ssDNA viruses (López-Bueno *et al.*, 2009).

A further study examined benthic mat samples taken from a freshwater pond located on the McMurdo Ice Shelf in Antarctica, for ssDNA viruses. Eight diverse circular ssDNA viruses were isolated from the samples that had genomes ranging in size from ~1.9-3.1 kb (Zawar-Reza *et al.*, 2014). The genomes organisations of these viruses were either uni or bi-directionally transcribed with all containing at least two major ORFs. Phylogenetic analysis of the putative Reps from these viruses showed they were distantly related to a circular ssDNA viruses isolated from Lake Limnopolar (South Shetland Islands) and shared <35% amino acid pairwise identity with eight Antarctic viral genomes (Zawar-Reza *et al.*, 2014). This study highlights that the knowledge surrounding the diversity and genetic richness of circular ssDNA viruses in Antarctica remains to be further explored.

1.9.4.5 Perennial pond of the Mauritanian Sahara

The Sahara is the largest desert on Earth with the exception of the Polar Regions. Despite this there are still small bodies of water found in rock pools (gueltas). Metagenomics approaches were used to study the viral diversity in desert environments that have many environmental and nutrient limiting factors (Fancello *et al.*, 2013). Roche 454 pyrosequencing combined with RCA were used to produce metagenomic data for four different gueltas. The results showed that the viromes of all four gueltas were dominated by the order *Caudovirales* with myoviruses. In this study ssDNA viruses were almost non-existent with the greatest portion relating to bacteriophages. This demonstrated that unlike the Antarctic lakes, not all extreme environments are dominated by ssDNA viruses.

1.9.4.6 Hypersaline lake

Many of the viral metagenomic studies to date have simply looked at viral diversity within an ecosystem. However, it becomes more complex to try and examine viral population dynamics, genetic composition and diversity within an ecosystem. One study has looked at eight samples collected from the hypersaline Lake Tyrrellin, Victoria, Australia. These samples were analysed for their viral diversity, composition and dynamics of viral assemblages (populations) (Emerson *et al.*, 2013). The hypersaline environment is a great system for studying viral assemblages as the geothermal conditions are very stable and constant. There are also no complex interactions at higher trophic levels because of the extreme hypersaline conditions, as the community structure mainly consists of microbes (Emerson *et al.*, 2013). To examine viral assemblages, this study defined operational taxanomic units (OTU) by looking at genes that encoded protein of similar function. The genes that were chosen formed three clusters each containing many sequence members to define the OTU. The study found a significant correlation between the structure of viral assemblages and environmental factor in Lake Tyrrell such as salinity and potassium concentration which is thought to influence host population dynamics. Whilst viral diversity was found to be high remain relatively constant over time (Emerson *et al.*, 2013).

This study introduced a new way of defining OTUs that can be used for many different analyse to help link the fields of environmental virology and microbial ecology. Additional new techniques for comparing viromes, estimating viral diversity and accessing for different environmental factors may be influencing viral assemblages will also benefit metagenomic studies in the future looking at viral assemblages and diversity.

1.9.4.7 ssDNA viruses in reclaimed water

The pressure on finite resources around the world is increasing with growth in the human population and urbanization. There is an increased need for freshwater supplies around the world and as a result alternative water supplies such as reclaimed water are common practice in many countries in a continuing effort for sustainable water resource management (Levine & Asano, 2004). Florida has been reclaiming water for over 20 years (Young & York, 1996), this water is used in agricultural irrigation, industrial uses and groundwater recharge (Martinez & Clark, 2012). There is little knowledge surrounding the microbial community in reclaimed water. Because of the increase in the use of this water, studies have looked at the spread of viral pathogens through reclaimed water supplies. Most of these studies have discovered many enteric viruses in reclaimed water such as rotaviruses, astroviruses, saproviruses, noroviruses, adenoviruses, reoviruses, enteroviruses and hepatitis A viruses (Arraj *et al.*, 2008; Bofill-Mas *et al.*, 2006; Haramoto *et al.*, 2008; Haramoto *et al.*, 2006; Katayama *et al.*, 2008; Meleg *et al.*, 2008; Morace *et al.*, 2002; Sedmak *et al.*, 2005).

A study carried out in 2009 looked at the viral diversity in reclaimed combining metagenomic and epifluorescent microscopy to examine water from two different sample sites (Rosario *et al.*, 2009b). The virome of reclaimed water from both sample sites was dominated by bacteriophages, however one site was dominant for phages relating to the *Siphoviridae* family and the other was dominated by prophages. This indicates that different reclaimed water consists of different phage communities (Rosario *et al.*, 2009b). Other prominent viruses were eukaryotic ssDNA viruses relating to the families *Nanoviridae*, *Geminiviridae*, *Circoviridae* with over 60% falling under novel CRESS DNA viruses. The persistence of these viruses in reclaimed water suggests they may be resistant to chlorination and may circulate through reclaimed water (Rosario *et al.*, 2009b).

In both the RNA and DNA data generated in this study, many mobile genetic elements relating to plasmids and phages were identified. Over half the sequences originally classified as bacterial were reclassified as plasmids (Rosario *et al.*, 2009b). The sequence data was compared to the ACLAME database. In the DNA data, over 51% of the contigs related to viral proteins, whilst in the RNA data almost all the sequences from both sample sites were related to proteins found in plasmids. The contigs were similar to different types of mobile elements including hypothetical proteins, integrases, transposases, recombinases, replication-associated proteins and many more (Rosario *et al.*, 2009b). Many RNA viruses with plasmid-like properties have been identified from a range of environments including plants, algae, fungi, protozoa and insects. It is therefore likely that the wealth of plasmids detected in the RNA data reflects that the RNA viruses might share many properties with plasmids such as the endornavirus.

1.9.5 Invertebrates

Insects have always been of interest in the study of ssDNA viruses as many rely on insect vectors to be transmitted between different plants. Studies have utilised vector enabled metagenomics (VEM) to investigate the role insects play in the transmission of geminiviruses (Ng *et al.*, 2011a; Ng *et al.*, 2011b; Rosario *et al.*, 2013). VEM involves purifying the virus particles from the sample then shotgun sequencing the viral nucleic acid helping to give a clearer picture of an entire viral community.

Another study used VEM to examine the role that whiteflies play as a vector for geminiviruses in agricultural regions in Florida, USA (Ng *et al.*, 2011a). Using this technique allowed investigators to examine the viral diversity of DNA viruses in whiteflies. The results from this research highlighted how the VEM technique can be used to investigate the viromes of insect vectors, as 79% of the sequences of the viromes shared some identity with previously described plant-infecting viruses, mainly geminiviruses (Ng *et al.*, 2011a).

A follow up study investigated mosquitoes that feed on an extensive range of hosts including humans, primates, birds, other animals and even plant nectar (Ng *et al.*, 2011b). Ng *et al.* (2011) performed metagenomic sequencing on three mosquito samples collected from San Diego, USA to examine the viral diversity within these insects as they were likely to contain both viruses transmitted by the mosquito and other viruses from host reservoirs. The results of this study were remarkable as each of the mosquito samples contained a very unique viral reservoir, with between ~29-81% of the viral contigs has similarities to densovirus-like sequences. Other viruses present were viruses related to animals, plants, bacteria and insects.

Some of these novel viral sequences were verified using specific primers by PCR. Two of these viruses recovered had novel genomes organisations similar to those described by Rosario *et al.* (2012a), further phylogenetic analysis showed these sample grouped with the previously described gemycircularviruses (Ng *et al.*, 2011b; Rosario *et al.*, 2012b).

Dragonflies are top insect predators in their ecosystems and feed on most other winged insects including mosquitoes, moths, whiteflies, aphids and many others (Corbet & Brooks, 2008). Rosario *et al.* (2013) collected six dragonflies from agricultural fields in Puerto Rico to see if geminiviruses could be detected in their mid-gut as they feed on insects which in turn feed on potentially infected plant material. Dragonfly-associated mastrevirus (DfasMV) along with Dragonfly-associated alphasatellite (Dfas-alphasatellite) were recovered from these samples. This study was of particular importance as this was the first report of a mastrevirus in the Caribbean. Analysis of Dfas-alphasatellite showed that it shared closer amino acid identity with alphasatellites associated with old world begomovirus. This suggests that there may have been multiple introductions of alphasatellites into the new world as Dfas-alphasatellite. This study highlighted how top end insect predators can potentially be used to monitor plant related pathogens. However, the limitations to using this top down approach is that the host of the virus is not recovered thus not revealing the source of infection.

Dragonflies are top end predators within their ecosystem and are highly mobile. Rosario *et al.* (2011) looked at circular ssDNA viruses within 12 dragonfly samples from the two islands in the Kingdom of Tonga (Rosario *et al.*, 2011). A combination of RCA and restriction enzyme digests resulted in fragments that were cloned and sequenced. Eleven sub types (>95% nucleotide identity) of dragonfly cyclovirus (DfCyV) were discovered from 12 dragonfly samples. Additionally, a further study by Rosario *et al.* (2012a) looked at isolating CRESS-DNA viruses from dragonfly samples from the Kingdom of Tonga, Florida, USA, Bulgaria, Finland, Hungary, Germany, Finland and Puerto Rico (Rosario *et al.*, 2012b). Seventeen CRESS-DNA viral genomes were recovered from the tissue of eight different dragonfly species. The genomes organisations of these viruses varied, with all having at least two major ORFS one of which showed similarities to Reps of circular ssDNA viruses. Nine of the viral genomes the grouped within the proposed cycloviruses genus also contained the conserved nonanucleotide motif TAGTATTAC at the apex of the stem-loop. These viral genomes had the same bi-directionally transcribed genomes typical of cycloviruses. The Reps of three of the other genomes discovered showed similarities to the Reps of circoviruses, however, their unisense genome organisation was different to circoviruses and cycloviruses. These were names Dragonfly circularisvirus (DfCirV), Dragonfly orbiculatusvirus (DFOrV) and Dragonfly cyclicusvirus (DfCyCIV). Viral isolates were also isolated from three dragonfly species from the Kingdom of Tonga and Florida, USA that showed similarities to the gemycircularviruses (Rosario *et al.*, 2012b; Yu *et al.*, 2010). These three viral isolates from the dragonflies had the same genome organisations and nonanucleotide motifs as SsHADV-1 , these viruses have been proposed to be part of the genus Gemycircularvirus (Rosario *et al.*, 2012b).

A novel cyclovirus was also isolated from *Eurycotis floridana* commonly known as the Florida woods cockroach or palmetto bug (Padilla-Rodriguez *et al.*, 2013). This virus isolate (FWCasCyV-GS140) had the same genome organisation as cycloviruses FWCasCyV-GS140 shared 64% genome-wide identity to a cyclovirus that was isolated from bat faeces.

Many novel ssDNA viruses have been isolated from different species of invertebrates. Invertebrates inhabit a range of different environments and could accumulate ssDNA viruses which is part are dependent on their resource consumption and their dispersal.

1.9.6 Atmosphere

The atmosphere acts as a huge dispersal mechanism for many different chemical and biological particles, in particular viruses. Whon *et al.* (2012) used a metagenomic approach to examine the diversity of airborne viruses in the near-surface atmosphere from three different regions whose land use and geographical location was different. Air samples were collected over a six month period to track changes variation of viruses over space and time. Sample were collected using a connector-linked direct precipitation air sampler that uses different filters to collect particles smaller than 1µm. Viral particles were then selectively amplified using RCA and analysed by 454 pyrosequencing. The virome of the three sampling sites contained 12 different viral families mainly consisted of ssDNA viruses (75.5% to 97.6%). The majority of ssDNA viral sequences from all four viromes were made up of geminivirus-related sequences, further phylogenetic analysis of the putative Rep that showed many sequences had similarities to gemycircularviruses (Whon *et al.*, 2012; Yu *et al.*, 2010). This viral metagenomic study showed that eukaryotic ssDNA viruses dominated the virome which contrasts to results found in other environmental metagenomic studies which showed that ssDNA microphage usually make up the majority of the virome (Angly *et al.*, 2006; Kim *et* *al.*, 2008; Rosario *et al.*, 2009b; Roux *et al.*, 2012). This highlights that the diversity and abundance of viruses does vary between environments.

1.9.7 Sewage

Sewage systems harbour huge viral diversity due to them containing waste from thousands of individuals in one area and often reflect infectious pathogens that are transmitted via the faecal oral route in a population (Pina *et al.*, 2001). As many viruses disperse though these waste water pathways and back into the environment it is important to understand the viral diversity of raw sewage. Cantalupo *et al.* (2011) used metagenomic techniques to investigated the viral diversity of raw sewage from three different locations: Pittsburgh, Pennsylvania, United States; Barcelona, Spain; and Addis Ababa, Ethiopia (Cantalupo *et al.*, 2011). The study used Roche 454 pyrosequencing to sequences DNA samples of untreated waste water from three different continents. The results from this study detected 234 known viruses; however, most of the viral sequences recovered from all samples were of novel viruses. The majority of the sequences detected were bacteriophages, similar to what has been observed in marine metagenomic studies (Angly *et al.*, 2006). Whilst a number of other reads were closely related to eukaryotic viruses with over 90.9% of these being derived from plant viruses. This could be expected as plant viruses have been shown to dominate viral communities in human stools and aquatic environments (Rosario *et al.*, 2009b; Zhang *et al.*, 2005).

Following this, a study by Ng *et al.* (2012) again used the same metagenomic techniques to examine the virome of waste water from a variety of countries including: Maiduguri, Nigeria; San Francisco, California, United States; Bangkok, Thailand; and Kathmandu, Nepal (Ng *et al.*, 2012). Results were similar to the previous study with bacteriophages dominating the virome. Novel ssDNA viruses were detected with similarities to geminiviruses, nimiviruses and baminiviruses (Ng *et al.*, 2011b).

Blinkova *et al.* (2009) examined untreated sewage for 12 cities around the United States using a nested PCR (Blinkova *et al.*, 2009). The nested PCR was based on RNA and DNA viruses recently isolated from stool specimens of South Asian children. The results from this study showed wide viral diversity across viral genera including dioviruses, coasaviruses, circoviruses and bacaviruses. Whilst bocavirus sequences were the most common throughout all samples, nine different circovirus-like sequences were isolated showing a greater distribution of these viruses.

This study demonstrated that there is great viral diversity within sewage system, and the geographic distribution of these viruses is wider than previously recognised (Blinkova *et al.*, 2009).

A recent study investigated treated sewage water which is a combination of CRESS DNA viruses from both terrestrial and aquatic viromes as it is heavily influenced by humans (Kraberger *et al.*, 2014). This study used high throughput sequencing methods, in combination with PCR amplification and Sangar sequencing to investigate the CRESS DNA viral diversity in sewage oxidation pond. In addition to the 50 novel CRESS DNA viral genomes were recovered, 11 small circular molecules that are likely to be subgenomic molecules were also recovered. Some of the genomes recovered were similar to the gemycircularviruses, however, most genomes were very diverse. This study further indicates that the diversity of CRESS DNA viruses is extensive.

1.9.8 Faecal matter

This low impact sampling method is non-invasive and is a top down approach that relies on the concept that animals serve as a large viral reservoir due to their longer life spans and higher position up the food chain, which may provide a broader range for examining viral diversity within an ecosystem. This sampling method has proved to be very effective when looking at ssDNA viral diversity from a range of different faecal samples including that from chimpanzee (Blinkova *et al.*, 2010; Li *et al.*, 2010b), pig (Shan *et al.*, 2011; Sikorski *et al.*, 2013b), bat (Ge *et al.*, 2012; Li *et al.*, 2010a), sea lion (Li *et al.*, 2011b), rodent (Phan *et al.*, 2011), bovine (Kim *et al.*, 2011), badger and pine marten (van den Brand *et al.*, 2011) and various other animals (Sikorski *et al.*, 2013d).

Blinkova *et al*. (2010) used a metagenomic approach coupled with inverse PCR to identify and recover viral sequences from wild Chimpanzee stools collected in Cameroon, Central African Republic, the Democratic Republic of Congo, the Republic of Congo, Tanzania, Uganda and Rwanda. The study uncovered seven novel ssDNA viruses named chimpanzee stool-associated circular viruses (ChiSCVs). The organisation of these viral genomes was unique in that there were two major ORFs that were bi-directionally transcribed towards the stem-loop. The Reps of ChiSCVs contained the major RCR and SF3 helicase motifs present in most ssDNA viruses. However, phylogenetic analysis showed that these viruses grouped separately to previously described viral families nanoviruses, geminiviruses and circoviruses (Blinkova *et al.*, 2010).

Another study again looked at novel viruses isolated from faecal samples from humans and chimpanzees (Li *et al.*, 2010b). This study used degenerate PCR based on viral sequences that were previously isolated from human faecal samples in Pakistan (Victoria *et al.*, 2009). The novel ssDNA viruses recovered has similar genome organisation to circoviruses, however, phylogenetic analysis showed they also grouped into a distinct clade like ChiSCV, this study proposed that this genus be named Cyclovirus which would be part of the Circovirus family (Li *et al.*, 2010b).

Further studies have looked at the virome of faecal matter from high intensity pig farming in North Carolina, USA (Shan *et al.*, 2011). This study used amplified the DNA using random PCR then used Roche 454 pyrosequencing. Only 1% of the reads recovered from the metagenomic sequencing showed similarities to ssDNA viruses. Four novel genomes were characterised using inverse PCR. These genomes showed similarities to circoviruses and were name porcine circovirus-like viruses (po-circo-like viruses). They had three major ORFs that were uni-directionally transcribed, one of which encoded for a putative Rep. The Rep also contained RCR motifs as well as N-terminal and P-loop domains similar to what is observed in other ssDNA viruses (Shan *et al.*, 2011).

In addition, a recent study has also recovered a CRESS DNA viruses named ancient caribou feces associated virus (aCFV) isolated from an ice core sample taken sampled in Greenland (Ng *et al.*, 2014). This ice core sample contained ancient caribou faeces which is estimated to be over 700 years old. It has been suggested that this virus was present in the caribou faeces as a result of diet, as phylogenetic analysis indicated in was closely related to geminiviruses and gemycircularviruses. Further infectivity studies gave a positive result when infecting model plant *Nicotiana benthamiana*. This study demonstrates that CRESS DNA viruses are able to persist in some environments for centuries without degradation.

1.10 Objectives of this research

The main objectives of my research are to first investigate what CRESS DNA viruses are circulating in dragonflies which are top end insect predators, those in both terrestrial environments and the larval form in aquatic environments. In addition, I wish to investigate what CRESS DNA viruses are present in natural bio-concentrators such as molluscs. This will be achieved by combining restriction enzyme disgestion, RCA and Illumina sequencing methods. This will demonstrate the diversity of CRESS DNA viruses that are circulating in both the molluscs bio-concentrators and top end insect predators, if these viruses are species specific, if the viruses occur across different environments and finally if the sample types represent potential surveillance tools for accessing CRESS DNA viral diversity in ecosystems. Finally I wish to look at CRESS DNA viruses circulating in environments, by assessing dragonflies, larvae, molluscs, benthic sediment and water I hope to understand the dynamics of the CRESS DNA viruses in the environment. This information will help to further understand the true diversity of CRESS DNA viruses as well as their evolution and distribution in a given environment.

References

Abadie, J., Nguyen, F., Groizeleau, C., Amenna, N., Fernandez, B., Guereaud, C., Guigand, L., Robart, P., Lefebvre, B. & other authors (2001). Pigeon circovirus infection: pathological observations and suggested pathogenesis. *Avian Pathology* **30**, 149-158.

Abouzid, A. M., Frischmuth, T. & Jeske, H. (1988). A putative replicative form of the Abutilon mosaic virus (gemini group) in a chromatin-like structure. *Molecular and General Genetics MGG* **212**, 252-258.

Adams, I. P., Glover, R. H., Monger, W. A., Mumford, R., Jackeviciene, E., Navalinskiene, M., Samuitiene, M. & Boonham, N. (2009). Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Molecular Plant Pathology* **10**, 537-545.

Alberts, B., Bray, D., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P. & Latchman, D. S. (1998). Essential Cell Biology: An Introduction to the Molecular Biology of the Cell. *Trends in Biochemical Sciences* **23**, 268-268.

Allan, G., McNeilly, F., Kennedy, S., Daft, B., Ellis, J., Haines, D., Meehan, B. & Adair, B. (1998). Isolation of porcine circovirus-like viruses from pigs with a wasting disease in the USA and Europe. *Journal of veterinary diagnostic investigation* **10**, 3.

Allan, G. M. & Ellis, J. A. (2000). Porcine circoviruses: a review. *Journal of veterinary diagnostic investigation* **12**, 3.

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., Chan, A. M., Haynes, M., Kelley, S. & other authors (2006). The marine viromes of four oceanic regions. *PLoS biology* **4**, e368.

Argüello-Astorga, G. & Ruiz-Medrano, R. (2001). 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. *Archives of virology* **146**, 1465-1485.

Argüello-Astorga, G., Guevara-Gonzalez, R., Herrera-Estrella, L. & Rivera-Bustamante, R. (1994). Geminivirus replication origins have a group-specific organization of iterative elements: a model for replication. *Virology* **203**, 90-100.

Arraj, A., Bohatier, J., Aumeran, C., Bailly, J., Laveran, H. & Traore, O. (2008). An epidemiological study of enteric viruses in sewage with molecular characterization by RT-PCR and sequence analysis. *Journal of water and health* **6**, 351-358.

Ball, N., Smyth, J., Weston, J., Borghmans, B., Palya, V., Glávits, R., Ivanics, E., Dan, A. & Todd, D. (2004). Diagnosis of goose circovirus infection in Hungarian geese samples using polymerase chain reaction and dot blot hybridization tests. *Avian Pathology* **33**, 51-58.

Bamford, D. H., Grimes, J. M. & Stuart, D. I. (2005). What does structure tell us about virus evolution? *Current opinion in structural biology* **15**, 655-663.

Banda, A., Galloway-Haskins, R. I., Sandhu, T. S. & Schat, K. A. (2007). Genetic analysis of a duck circovirus detected in commercial Pekin ducks in New York. *Avian diseases* **51**, 90-95.

Beck, E. & Zink, B. (1981). Nucleotide sequence and genome organisation of filamentous bacteriophages f1 and fd. *Gene* **16**, 35-58.

Bench, S. R., Hanson, T. E., Williamson, K. E., Ghosh, D., Radosovich, M., Wang, K. & Wommack, K. E. (2007). Metagenomic characterization of Chesapeake Bay virioplankton. *Applied and environmental microbiology* **73**, 7629-7641.

Bendinelli, M., Pistello, M., Maggi, F., Fornai, C., Freer, G. & Vatteroni, M. L. (2001). Molecular properties, biology, and clinical implications of TT virus, a recently identified widespread infectious agent of humans. *Clinical microbiology reviews* **14**, 98-113.

Bernardo, P., Golden, M., Akram, M., Nadarajan, N., Fernandez, E., Granier, M., Rebelo, A. G., Peterschmitt, M., Martin, D. P. & other authors (2013). Identification and characterisation of a highly divergent geminivirus: evolutionary and taxonomic implications. *Virus Research* **177**, 35-45.

Biagini, P., Gallian, P., Attoui, H., Touinssi, M., Cantaloube, J.-F., de Micco, P. & de Lamballerie, X. (2001). Genetic analysis of full-length genomes and subgenomic sequences of TT virus-like mini virus human isolates. *Journal of General Virology* **82**, 379-383.

Biagini, P., M. Bendinelli, S. Hino, L. Kakkola, A. Mankertz, C. Niel, H. Okamoto, S. Raidal, C. G. Teo, and D. Todd (2012). *Family - Circoviridae, Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*. San Diego: Elsevier.

Bisaro, D. M. (1996). Geminivirus DNA replication. *DNA replication in eukaryotic cells Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY*, 833-854.

Blanco, L., Bernad, A., Lázaro, J. M., Martin, G., Garmendia, C. & Salas, M. (1989). Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *Journal of Biological Chemistry* **264**, 8935-8940.

Blinkova, O., Rosario, K., Li, L., Kapoor, A., Slikas, B., Bernardin, F., Breitbart, M. & Delwart, E. (2009). Frequent detection of highly diverse variants of cardiovirus, cosavirus, bocavirus, and circovirus in sewage samples collected in the United States. *Journal of Clinical Microbiology* **47**, 3507.

Blinkova, O., Victoria, J., Li, Y., Keele, B. F., Sanz, C., Ndjango, J. B. N., Peeters, M., Travis, D., Lonsdorf, E. V. & other authors (2010). Novel circular DNA viruses in stool samples of wild-living chimpanzees. *Journal of General Virology* **91**, 74-86.

Bodewes, R., van der Giessen, J., Haagmans, B. L., Osterhaus, A. D. & Smits, S. L. (2013). Identification of multiple novel viruses, including a parvovirus and a hepevirus, in feces of red foxes. *Journal of virology* **87**, 7758-7764.

Boevink, P., Chu, P. & Keese, P. (1995). Sequence of subterranean clover stunt virus DNA: affinities with the geminiviruses. *Virology* **207**, 354-361.

Bofill-Mas, S., Albinana-Gimenez, N., Clemente-Casares, P., Hundesa, A., Rodriguez-Manzano, J., Allard, A., Calvo, M. & Girones, R. (2006). Quantification and stability of human adenoviruses and polyomavirus JCPyV in wastewater matrices. *Applied and environmental microbiology* **72**, 7894-7896.

Borchardt, M. A., Bradbury, K. R., Gotkowitz, M. B., Cherry, J. A. & Parker, B. L. (2007). Human enteric viruses in groundwater from a confined bedrock aquifer. *Environmental science & technology* **41**, 6606-6612.

Breitbart, M. & Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? *Trends in microbiology* **13**, 278-284.

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J. M., Nulton, J., Salamon, P. & Rohwer, F. (2003). Metagenomic analyses of an uncultured viral community from human feces. *Journal of bacteriology* **185**, 6220-6223.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., Azam, F. & Rohwer, F. (2002). Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences* **99**, 14250-14255.

Brentlinger, K. L., Hafenstein, S., Novak, C. R., Fane, B. A., Borgon, R., McKenna, R. & Agbandje-McKenna, M. (2002). Microviridae, a family divided: isolation, characterization, and genome sequence of φMH2K, a bacteriophage of the obligate intracellular parasitic bacterium *Bdellovibrio bacteriovorus*. *Journal of bacteriology* **184**, 1089-1094.

Briddon, R. & Stanley, J. (2006). Subviral agents associated with plant single-stranded DNA viruses. *Virology* **344**, 198-210.

Briddon, R., Ghabrial, S., Lin, N., Palukaitis, P., Scholthof, K. & Vetten, H. (2012). Satellites and other virus-dependent nucleic acids, pp. 1209-1219: Elsevier Inc.: London, UK.

Briddon, R. W., Watts, J., Markham, P. G. & Stanley, J. (1989). The coat protein of beet curly top virus is essential for infectivity. *Virology* **172**, 628-633.

Briddon, R. W., Bull, S. E., Amin, I., Idris, A. M., Mansoor, S., Bedford, I. D., Dhawan, P., Rishi, N., Siwatch, S. S. & other authors (2003). Diversity of DNA β, a satellite molecule associated with some monopartite begomoviruses. *Virology* **312**, 106-121.

Brussaard, C. (2004). Viral Control of Phytoplankton Populations—a Review. *Journal of Eukaryotic Microbiology* **51**, 125-138.

Brussaard, C. & Martinez, J. M. (2008). Algal bloom viruses. *Plant Viruses* **2**, 1-13.

Burns, T. M., Harding, R. M. & Dale, J. L. (1995). The genome organization of banana bunchy top virus: analysis of six ssDNA components. *Journal of General Virology* **76**, 1471-1482.

Calisher, C. H., Childs, J. E., Field, H. E., Holmes, K. V. & Schountz, T. (2006). Bats: important reservoir hosts of emerging viruses. *Clinical microbiology reviews* **19**, 531-545.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A. & Brüssow, H. (2003). Prophage genomics. *Microbiology and Molecular Biology Reviews* **67**, 238-276.

Cantalupo, P. G., Calgua, B., Zhao, G., Hundesa, A., Wier, A. D., Katz, J. P., Grabe, M., Hendrix, R. W., Girones, R. & other authors (2011). Raw sewage harbors diverse viral populations. *MBio* **2**, e00180-00111.

Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far? *Molecular microbiology* **49**, 277-300.

Chae, C. (2005). A review of porcine circovirus 2-associated syndromes and diseases. *The Veterinary Journal* **169**, 326-336.

Chakraborty, S., Vanitharani, R., Chattopadhyay, B. & Fauquet, C. (2008). Supervirulent pseudorecombination and asymmetric synergism between genomic components of two distinct species of begomovirus associated with severe tomato leaf curl disease in India. *Journal of general virology* **89**, 818-828.

Chen, C.-L., Chang, P.-C., Lee, M.-S., Shien, J.-H., Ou, S.-J. & Shieh, H. (2003). Nucleotide sequences of goose circovirus isolated in Taiwan. *Avian Pathology* **32**, 165-171.

Chen, C. L., Wang, P. X., Lee, M. S., Shien, J. H., Shieh, H. K., Ou, S. J., Chen, C. H. & Chang, P. C. (2006). Development of a polymerase chain reaction procedure for detection and differentiation of duck and goose circovirus. *Avian diseases* **50**, 92-95.

Cheung, A. K. (2004). Detection of template strand switching during initiation and termination of DNA replication of porcine circovirus. *Journal of virology* **78**, 4268-4277.

Corbet, P. & Brooks, S. (2008). *Dragonflies*, vol. 106. Harpercollins Pub Ltd.

Corinaldesi, C., Crevatin, E., Del Negro, P., Marini, M., Russo, A., Fonda-Umani, S. & Danovaro, R. (2003). Large-scale spatial distribution of virioplankton in the Adriatic Sea: testing the trophic state control hypothesis. *Applied and environmental microbiology* **69**, 2664-2673.

Cotmorel, S. F. & Tattersall, P. (1996). Parvovirus DNA replication: DNA replication in eukaryoric cells. New York: Cold Spring Harbor Laboratory Press.

Danovaro, R., Dell'Anno, A., Corinaldesi, C., Magagnini, M., Noble, R., Tamburini, C. & Weinbauer, M. (2008). Major viral impact on the functioning of benthic deep-sea ecosystems. *Nature* **454**, 1084-1087.

Dayaram, A., Goldstien, S., Zawar-Reza, P., Gomez, C., Harding, J. S. & Varsani, A. (2013a). Novel single stranded DNA virus recovered from estuarine Mollusc (Amphibola crenata) whose replication associated protein (Rep) shares similarities with Rep-like sequences of bacterial origin. *Journal of General Virology* **94**, 1104-1110.

Dayaram, A., Opong, A., Jäschke, A., Hadfield, J., Baschiera, M., Dobson, R. C. J., Offei, S. K., Shepherd, D. N., Martin, D. P. & other authors (2012). Molecular characterisation of a novel cassava associated circular ssDNA virus. *Virus Research* **166**, 130-135.

Dayaram, A., Potter, K. A., Moline, A. B., Rosenstein, D. D., Marinov, M., Thomas, J. E., Beitbart, M., Rosario, K., Argüello-Astorga, G. R. & other authors (2013b). High global diversity of cycloviruses amongst dragonflies. *Journal of General Virology* **94**, 1827-1840.

de Jong, M. D., Van Kinh, N., Trung, N. V., Taylor, W., Wertheim, H. F., van der Ende, A., van der Hoek, L., Canuti, M., Crusat, M. & other authors (2014). Limited geographic distribution of the novel cyclovirus CyCV-VN. *Scientific reports* **4**.

Dekker, E. L., Woolston, C. J., Xue, Y., Cox, B. & Mullineaux, P. M. (1991). Transcript mapping reveals different expression strategies for the bicistronic RNAs of the geminivirus *wheat dwarf virus*. *Nucleic Acids Research* **19**, 4075.

Delwart, E. & Li, L. (2011). Rapidly expanding genetic diversity and host range of the *Circoviridae* viral family and other Rep encoding small circular ssDNA genomes. *Virus Research* **1**, 114-121.

Delwart, E. L. (2007). Viral metagenomics. *Reviews in medical virology* **17**, 115-131.

Desnues, C., Rodriguez-Brito, B., Rayhawk, S., Kelley, S., Tran, T., Haynes, M., Liu, H., Furlan, M., Wegley, L. & other authors (2008). Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* **452**, 340-343.

Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research* **36**, e105-e105.

Dokland, T., Bernal, R. A., Burch, A., Pletnev, S., Fane, B. A. & Rossmann, M. G. (1999). The role of scaffolding proteins in the assembly of the small, single-stranded DNA virus φX174. *Journal of Molecular Biology* **288**, 595-608.

Doneley, R. (2003). Acute Beak and Feather Disease in juvenile African Grey parrots-an uncommon presentation of a common disease. *Australian veterinary journal* **81**, 206-207.

Donson, J., Accotto, G. P., Boulton, M. I., Mullineaux, P. M. & Davies, J. W. (1987). The nucleotide sequence of a geminivirus from *Digitaria sanguinalis*. *Virology* **161**, 160-169.

Doszpoly, A., Tarjan, Z. L., Glávits, R., Mueller, T. & Benkő, M. (2014). Full genome sequence of a novel circo-like virus detected in an adult European eel *Anguilla anguilla* showing signs of cauliflower disease. *Diseases of Aquatic Organisms* **109**, 107-115.

Du, Z., Tang, Y., Zhang, S., She, X., Lan, G., Varsani, A. & He, Z. (2014). Identification and molecular characterization of a single-stranded circular DNA virus with similarities to Sclerotinia sclerotiorum hypovirulence-associated DNA virus 1. *Archives of virology* **159**, 1527-1531.

Duffy, S., Shackelton, L. A. & Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics* **9**, 267-276.

Duhaime, M. B. & Sullivan, M. B. (2012). Ocean viruses: Rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology* **434**, 181-186.

Dunlap, D. S., Ng, T. F. F., Rosario, K., Barbosa, J. G., Greco, A. M., Breitbart, M. & Hewson, I. (2013). Molecular and microscopic evidence of viruses in marine copepods. *Proceedings of the National Academy of Sciences* **110**, 1375-1380.

Edwards, R. A. & Rohwer, F. (2005). Viral metagenomics. *Nature Reviews Microbiology* **3**, 504-510.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P. & other authors (2009). Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138.

Ellis, J., Hassard, L., Clark, E., Harding, J., Allan, G., Willson, P., Strokappe, J., Martin, K., McNeilly, F. & other authors (1998). Isolation of circovirus from lesions of pigs with postweaning multisystemic wasting syndrome. *The Canadian veterinary journal* **39**, 44.

Ellis, J., Spinato, M., Yong, C., West, K., McNeilly, F., Meehan, B., Kennedy, S., Clark, E., Krakowka, S. & other authors (2003). Porcine circovirus 2–associated disease in Eurasian wild boar. *Journal of veterinary diagnostic investigation* **15**, 364.

Emerson, J. B., Thomas, B. C., Andrade, K., Heidelberg, K. B. & Banfield, J. F. (2013). New approaches indicate constant viral diversity despite shifts in assemblage structure in an Australian hypersaline lake. *Applied and environmental microbiology* **79**, 6755-6764.

Exel, B., Sipos,W. and Schmoll,F. (2003). Dynamics of porcine circovirus in a PMWS-positive herd. *Medical Clinic for Ruminants and Swine, University of Veterinary Medicine Vienna, Veterinaerplatz 1, Vienna 1210, Austria* **Unpublished**.

Fancello, L., Trape, S., Robert, C., Boyer, M., Popgeorgiev, N., Raoult, D. & Desnues, C. (2013). Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. *The ISME journal* **7**, 359-369.

Fauquet, C. M., Mayo, M., Maniloff, J., Desselberger, U. & Ball, L. A. (2005). *Virus taxonomy: VIIIth report of the International Committee on Taxonomy of Viruses*. Academic Press.

Fenner, F. & Maurin, J. (1976). The classification and nomenclature of viruses. *Archives of virology* **51**, 141-149.

Fèvre, E. M., Bronsvoort, B. M. d. C., Hamilton, K. A. & Cleaveland, S. (2006). Animal movements and the spread of infectious diseases. *Trends in microbiology* **14**, 125-131.

Finsterbusch, T. & Mankertz, A. (2009). Porcine circoviruses—small but powerful. *Virus Research* **143**, 177-183.

Firth, C., Charleston, M. A., Duffy, S., Shapiro, B. & Holmes, E. C. (2009). Insights into the evolutionary history of an emerging livestock pathogen: porcine circovirus 2. *Journal of virology* **83**, 12813-12821.

Franz, A. W., van der Wilk, F., Verbeek, M., Dullemans, A. M. & van den Heuvel, J. F. (1999). Faba Bean Necrotic Yellows Virus (Genus *Nanovirus*) Requires a Helper Factor for Its Aphid Transmission. *Virology* **262**, 210-219.

Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology* **3**, 722-732.

Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature* **399**, 541-548.

Gallian, P., Biagini, P., Zhong, S., Touinssi, M., Yeo, W., Cantaloube, J. F., Attoui, H., de Micco, P., Johnson, P. J. & other authors (2000). TT virus: a study of molecular epidemiology and transmission of genotypes 1, 2 and 3. *Journal of clinical virology* **17**, 43-49.

Ge, X., Li, Y., Yang, X., Zhang, H., Zhou, P., Zhang, Y. & Shi, Z. (2012). Metagenomic Analysis of Viruses from the Bat Fecal Samples Reveals Many Novel Viruses in Insectivorous Bats in China. *Journal of virology* **86**, 4620-4630.

Ge, X., Li, J., Peng, C., Wu, L., Yang, X., Wu, Y., Zhang, Y. & Shi, Z. (2011). Genetic diversity of novel circular ssDNA viruses in bats in China. *Journal of General Virology* **92**, 2646-2653.

Gilbert, C., Meik, J., Dashevsky, D., Card, D., Castoe, T. & Schaack, S. (2014). Endogenous hepadnaviruses, bornaviruses and circoviruses in snakes. *Proceedings of the Royal Society B: Biological Sciences* **281**, 20141122.

Gorbalenya, A. E., Koonin, E. V. & Wolf, Y. I. (1990). A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses. *FEBS letters* **262**, 145-148.

Gray, S. M. & Banerjee, N. (1999). Mechanisms of arthropod transmission of plant and animal viruses. *Microbiology and Molecular Biology Reviews* **63**, 128-148.

Grigoras, I., del Cueto Ginzo, A. I., Martin, D. P., Varsani, A., Romero, J., Mammadov, A. C., Huseynova, I. M., Aliyev, J. A., Kheyr-Pour, A. & other authors (2014). Genome diversity and evidence of recombination and reassortment in nanoviruses from Europe. *Journal of General Virology* **95**, 1178-1191.

Gronenborn, B. (2004). Nanoviruses: genome organisation and protein function. *Veterinary microbiology* **98**, 103-109.

Gutierrez, C. (1999). Geminivirus DNA replication. *Cellular and Molecular Life Sciences* **56**, 313-329.

Hafenstein, S. & Fane, B. A. (2002). φX174 genome-capsid interactions influence the biophysical properties of the virion: evidence for a scaffolding-like function for the genome during the final stages of morphogenesis. *Journal of virology* **76**, 5350-5356.

Hafner, G. J., Stafford, M. R., Wolter, L. C., Harding, R. M. & Dale, J. L. (1997). Nicking and joining activity of banana bunchy top virus replication protein in vitro. *Journal of General Virology* **78**, 1795-1799.

Haible, D., Kober, S. & Jeske, H. (2006). Rolling circle amplification revolutionizes diagnosis and genomics of geminiviruses. *Journal of virological methods* **135**, 9-16.

Halami, M., Nieper, H., Müller, H. & Johne, R. (2008). Detection of a novel circovirus in mute swans *Cygnus olor* by using nested broad-spectrum PCR. *Virus research* **132**, 208-212.

Hamel, A. L., Lin, L. L. & Nayar, G. P. S. (1998). Nucleotide sequence of porcine circovirus associated with postweaning multisystemic wasting syndrome in pigs. *Journal of virology* **72**, 5262-5267.

Haramoto, E., Katayama, H., Phanuwan, C. & Ohgaki, S. (2008). Quantitative detection of sapoviruses in wastewater and river water in Japan. *Letters in applied microbiology* **46**, 408-413.

Haramoto, E., Katayama, H., Oguma, K., Yamashita, H., Tajima, A., Nakajima, H. & Ohgaki, S. (2006). Seasonal profiles of human noroviruses and indicator bacteria in a wastewater treatment plant in Tokyo, Japan. *Water Science & Technology* **54**, 301-308.

Harkins, G. W., Martin, D. P., Christoffels, A. & Varsani, A. (2014). Towards inferring the global movement of *beak and feather disease virus*. *Virology* **450**, 24-33.

Harkins, G. W., Delport, W., Duffy, S., Wood, N., Monjane, A. L., Owor, B. E., Donaldson, L., Saumtally, S., Triton, G. & other authors (2009). Experimental evidence indicating that mastreviruses probably did not co-diverge with their hosts. *Virol J* **6**, 104.

Hattermann, K., Schmitt, C., Soike, D. & Mankertz, A. (2003). Cloning and sequencing of Duck circovirus (DuCV). *Archives of virology* **148**, 2471-2480.

He, B., Li, Z., Yang, F., Zheng, J., Feng, Y., Guo, H., Li, Y., Wang, Y., Su, N. & other authors (2013). Virome profiling of bats from Myanmar by metagenomic analysis of tissue samples reveals more novel mammalian viruses. *PloS one* **8**, e61950.

Heydarnejad, J., Keyvani, N., Razavinejad, S., Massumi, H. & Varsani, A. (2013). Fulfilling Koch's postulates for beet curly top Iran virus and proposal for consideration of new genus in the family *Geminiviridae*. *Archives of virology* **158**, 435-443.

Heyraud-Nitschke, F., Schumacher, S., Laufs, J., Schaefer, S., Schell, J. & Gronenborn, B. (1995). Determination of the origin cleavage and joining domain of geminivirus Rep proteins. *Nucleic Acids Research* **23**, 910.

Hickman, A. B. & Dyda, F. (2005). Binding and unwinding: SF3 viral helicases. *Current opinion in structural biology* **15**, 77-85.

Hill, J. E., Strandberg, J. O., Hiebert, E. & Lazarowitz, S. G. (1998). Asymmetric infectivity of pseudorecombinants of cabbage leaf curl virus and squash leaf curl virus: implications for bipartite geminivirus evolution and movement. *Virology* **250**, 283-292.

Hinnebusch, J. & Tilly, K. (1993). Linear plasmids and chromosomes in bacteria. *Molecular microbiology* **10**, 917-922.

Hu, J.-M., Fu, H.-C., Lin, C.-H., Su, H.-J. & Yeh, H.-H. (2007). Reassortment and concerted evolution in Banana bunchy top virus genomes. *Journal of virology* **81**, 1746-1761.

Huff, D. G., Schmidt, R. E. & Fudge, A. M. (1988). Psittacine beak and feather syndrome in a blue-fronted Amazon (Amazona aestiva). *AAV Today*, 84-86.

Hughes, A. L. (2004). Birth-and-death evolution of protein-coding regions and concerted evolution of non-coding regions in the multi-component genomes of nanoviruses. *Molecular phylogenetics and evolution* **30**, 287-294.

Idris, A. M., Shahid, M. S., Briddon, R. W., Khan, A., Zhu, J.-K. & Brown, J. K. (2011). An unusual alphasatellite associated with monopartite begomoviruses attenuates symptoms and reduces betasatellite accumulation. *Journal of General Virology* **92**, 706-717.

Ilyina, T. V. & Koonin, E. V. (1992). Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. *Nucleic Acids Research* **20**, 3279.

Inouye, T., Inouye, N. & Mitsuhata, K. (1968). Yellow dwarf of pea and broad bean caused by *milk-vetch dwarf virus*. *Ann Phytopathol Soc Jpn* **34**, 28-35.

Jeske, H., Lütgemeier, M. & Preiß, W. (2001). DNA forms indicate rolling circle and recombination-dependent replication of Abutilon mosaic virus. *The EMBO journal* **20**, 6158-6167.

Johne, R., Fernández-de-Luco, D., Höfle, U. & Müller, H. (2006). Genome of a novel circovirus of starlings, amplified by multiply primed rolling-circle amplification. *Journal of General Virology* **87**, 1189.

Kapoor, A., Dubovi, E. J., Henriquez-Rivera, J. A. & Lipkin, W. I. (2012). Complete Genome Sequence of the First Canine Circovirus. *Journal of virology* **86**, 7018-7018.

Karesh, W. B., Cook, R. A., Bennett, E. L. & Newcomb, J. (2005). Wildlife trade and global disease emergence. *Emerg Infect Dis* **11**, 1000-1002.

Katayama, H., Haramoto, E., Oguma, K., Yamashita, H., Tajima, A., Nakajima, H. & Ohgaki, S. (2008). One-year monthly quantitative survey of noroviruses, enteroviruses, and adenoviruses in wastewater collected from six plants in Japan. *Water research* **42**, 1441-1448.

Katul, L., Maiss, E. & Vetten, H. J. (1995). Sequence analysis of a *faba bean necrotic yellows virus* DNA component containing a putative replicase gene. *Journal of General Virology* **76**, 475-479.

Khan, S. A. (1997). Rolling-circle replication of bacterial plasmids. *Microbiology and Molecular Biology Reviews* **61**, 442-455.

Khan, S. A. (2000). Plasmid rolling-circle replication: recent developments. *Molecular microbiology* **37**, 477-484.

Kim, H. K., Park, S. J., Song, D. S., Moon, H. J., Kang, B. K. & Park, B. K. (2011). Identification of a novel single stranded circular DNA virus from bovine stool. *Journal of General Virology*.

Kim, K.-H., Chang, H.-W., Nam, Y.-D., Roh, S. W., Kim, M.-S., Sung, Y., Jeon, C. O., Oh, H.-M. & Bae, J.-W. (2008). Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Applied and environmental microbiology* **74**, 5975-5985.

King, A., Lefkowitz, E., Adams, M. & Carstens, E. (2011). Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses: An Elsevier Title: Burlington.

Kingsbury, D. (1985). Species classification problems in virus taxonomy. *Intervirology* **24**, 62-70.

Kingsford, C., Schatz, M. C. & Pop, M. (2010). Assembly complexity of prokaryotic genomes using short reads. *BMC bioinformatics* **11**, 21.

Koonin, E. V. & Ilyina, T. V. (1992). Geminivirus replication proteins are related to prokaryotic plasmid rolling circle DNA replication initiator proteins. *The Journal of General Virology* **73**, 2763.

Koonin, E. V., Senkevich, T. G. & Dolja, V. V. (2006). The ancient Virus World and evolution of cells. *Biol Direct* **1**, 29.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R. & other authors (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology* **30**, 693-700.

Kraberger, S., Stainton, D., Dayaram, A., Zawar-Reza, P., Gomez, C., Harding, J. S. & Varsani, A. (2013a). Discovery of Sclerotinia sclerotiorum hypovirulence-associated virus-1 in urban river sediments of Heathcote and Styx rivers in Christchurch city, New Zealand. *Genome announcements* **1**, e00559-00513.

Kraberger, S., Argüello-Astorga, G. R., Greenfield, G. L., Galilee, C., Law, D., Martin, D. P. & Varsani, A. (2014). Characterisation of a diverse range of Rep-encoding ssDNA viruses recovered from a sewage treatment oxidation pond. *Infection, Genetics and Evolution* **In Review**.

Kraberger, S., Harkins, G. W., Kumari, S. G., Thomas, J. E., Schwinghamer, M. W., Sharman, M., Collings, D. A., Briddon, R. W., Martin, D. P. & other authors (2013b). Evidence that dicot-infecting mastreviruses are particularly prone to inter-species recombination and have likely been circulating in Australia for longer than in Africa and the Middle East. *Virology* **444**, 282-291.

Krakowka, S., Ellis, J., McNeilly, F., Ringler, S., Rings, D. & Allan, G. (2001). Activation of the immune system is the pivotal event in the production of wasting disease in pigs infected with porcine circovirus-2 (PCV-2). *Veterinary Pathology Online* **38**, 31-42.

Krenz, B., Thompson, J. R., Fuchs, M. & Perry, K. L. (2012). Complete genome sequence of a new circular DNA virus from grapevine. *Journal of virology* **86**, 7715-7715.

Krupovic, M., Ravantti, J. & Bamford, D. (2009a). Geminiviruses: a tale of a plasmid becoming a virus. *BMC Evolutionary Biology* **9**, 112.

Krupovic, M., Ravantti, J. J. & Bamford, D. H. (2009b). Geminiviruses: a tale of a plasmid becoming a virus. *BMC evolutionary biology* **9**, 112.

Kumar, J., Kumar, J., Singh, S. P. & Tuli, R. (2014). Association of satellites with a mastrevirus in natural infection: complexity of Wheat dwarf India virus disease. *Journal of virology* **88**, 7093-7104.

Kuo, T.-T., Lin, Y.-H., Huang, C.-M., Chang, S.-F., Dai, H. & Feng, T.-Y. (1987). The lysogenic cycle of the filamentous phage Cflt from *Xanthomonas campestris*. *Virology* **156**, 305-312.

Labonté, J. M. & Suttle, C. A. (2013). Previously unknown and highly divergent ssDNA viruses populate the oceans. *The ISME journal* **7**, 2169-2177.

Lamberto, I., Gunst, K., Müller, H., zur Hausen, H. & de Villiers, E.-M. (2014). Mycovirus-like DNA virus sequences from cattle serum and human brain and serum samples from multiple sclerosis patients. *Genome announcements* **2**, e00848-00814.

Lasken, R. S. & Stockwell, T. B. (2007). Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC biotechnology* **7**, 19.

Lazarowitz, S. G., Wu, L. C., Rogers, S. G. & Elmer, J. S. (1992). Sequence-specific interaction with the viral AL1 protein identifies a geminivirus DNA replication origin. *The Plant Cell Online* **4**, 799-809.

Lefeuvre, P., Lett, J.-M., Varsani, A. & Martin, D. (2009). Widely conserved recombination patterns among single-stranded DNA viruses. *Journal of virology* **83**, 2697-2707.

Lefeuvre, P., Harkins, G. W., Lett, J.-M., Briddon, R. W., Chase, M. W., Moury, B. & Martin, D. P. (2011). Evolutionary time-scale of the begomoviruses: evidence from integrated sequences in the Nicotiana genome. *PloS one* **6**, e19193.

Lefeuvre, P., Martin, D., Hoareau, M., Naze, F., Delatte, H., Thierry, M., Varsani, A., Becker, N., Reynaud, B. & other authors (2007). Begomovirus 'melting pot'in the south-west Indian Ocean islands: molecular diversity and evolution through recombination. *Journal of General Virology* **88**, 3458-3468.

Legendre, M., Bartoli, J., Shmakova, L., Jeudy, S., Labadie, K., Adrait, A., Lescot, M., Poirot, O., Bertaux, L. & other authors (2014). Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proceedings of the National Academy of Sciences* **111**, 4274-4279.

Levine, A. D. & Asano, T. (2004). Peer reviewed: Recovering sustainable water from wastewater. *Environmental science & technology* **38**, 201A-208A.

Li, L., Victoria, J. G., Wang, C., Jones, M., Fellers, G. M., Kunz, T. H. & Delwart, E. (2010a). Bat guano virome: predominance of dietary viruses from insects and plants plus novel mammalian viruses. *Journal of virology* **84**, 6955-6965.

Li, L., Shan, T., Soji, O. B., Alam, M. M., Kunz, T. H., Zaidi, S. Z. & Delwart, E. (2011a). Possible cross-species transmission of circoviruses and cycloviruses among farm animals. *Journal of General Virology* **92**, 768-772.

Li, L., Shan, T., Wang, C., Côté, C., Kolman, J., Onions, D., Gulland, F. M. & Delwart, E. (2011b). The fecal viral flora of California sea lions. *Journal of virology* **85**, 9909-9917.

Li, L., Kapoor, A., Slikas, B., Bamidele, O. S., Wang, C., Shaukat, S., Masroor, M. A., Wilson, M. L., Ndjango, J. B. N. & other authors (2010b). Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces. *Journal of Virology* **84**, 1674.

Li, L., McGraw, S., Zhu, K., Leutenegger, C. M., Marks, S. L., Kubiski, S., Gaffney, P., Cruz Jr, F. N. D., Wang, C. & other authors (2013). Circovirus in tissues of dogs with vasculitis and hemorrhage. *Emerging infectious diseases* **19**, 534.

Lian, H., Liu,Y., Li,N., Wang,Y., Zhang,S. and Hu,R. (2014). Mink Circovirus: a Novel Infectious Agent of Mink Enteritis in China. *Unpublished*.

Lidmar, J., Mirny, L. & Nelson, D. R. (2003). Virus shapes and buckling transitions in spherical shells. *Physical Review E* **68**, 051910.

Lipps, G. (2008). *Plasmids: current research and future trends*. Horizon Scientific Press.

Liu, D., Daubendiek, S. L., Zillman, M. A., Ryan, K. & Kool, E. T. (1996). Rolling circle DNA synthesis: small circular oligonucleotides as efficient templates for DNA polymerases. *Journal of the American Chemical Society* **118**, 1587-1594.

Liu, H., Fu, Y., Li, B., Yu, X., Xie, J., Cheng, J., Ghabrial, S. A., Li, G., Yi, X. & other authors (2011). Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. *BMC Evolutionary Biology* **11**, 276.

Liu, L., Saunders, K., Thomas, C. L., Davies, J. W. & Stanley, J. (1999). Bean yellow dwarf virus RepA, but not Rep, binds to maize retinoblastoma protein, and the virus tolerates mutations in the consensus binding motif. *Virology* **256**, 270-279.

Liu, Q., Guo, Y., Li, J., Long, J., Zhang, B. & Shyr, Y. (2012). Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. *BMC genomics* **13**, S8.

Loconsole, G., Saldarelli, P., Doddapaneni, H., Savino, V., Martelli, G. P. & Saponari, M. (2012). Identification of a single-stranded DNA virus associated with citrus chlorotic dwarf disease, a new member in the family *Geminiviridae*. *Virology* **432**, 162-172.

Logares, R., Bråte, J., Bertilsson, S., Clasen, J. L., Shalchian-Tabrizi, K. & Rengefors, K. (2009). Infrequent marine–freshwater transitions in the microbial world. *Trends in microbiology* **17**, 414-422.

Londoño, A., Riego-Ruiz, L. & Argüello-Astorga, G. R. (2010). DNA-binding specificity determinants of replication proteins encoded by eukaryotic ssDNA viruses are adjacent to widely separated RCR conserved motifs. *Archives of virology* **155**, 1033-1046.

López-Bueno, A., Tamames, J., Velázquez, D., Moya, A., Quesada, A. & Alcamí, A. (2009). High diversity of the viral community from an Antarctic lake. *Science* **326**, 858-861.

Lőrincz, M., Cságola, A., Farkas, S. L., Székely, C. & Tuboly, T. (2011). First detection and analysis of a fish circovirus. *Journal of General Virology* **92**, 1817-1821.

Lőrincz, M., Dán, Á., Láng, M., Csaba, G., Tóth, Á. G., Székely, C., Cságola, A. & Tuboly, T. (2012). Novel circovirus in European catfish (*Silurus glanis*). *Archives of virology*, 1-4.

Lukashov, V. V. & Goudsmit, J. (2001). Evolutionary relationships among parvoviruses: virus-host coevolution among autonomous primate parvoviruses and links between adeno-associated and avian parvoviruses. *Journal of virology* **75**, 2729-2740.

Malyshenko, S., Kondakova, O., Taliansky, M. & Atabekov, J. (1989). Plant virus transport function: complementation by helper viruses is non-specific. *Journal of general virology* **70**, 2751-2757.

Mandal, B., Mandal, S., Pun, K. & Varma, A. (2004). First report of the association of a nanovirus with foorkey disease of large cardamom in India. *Plant Disease* **88**, 428-428.

Mankertz, A., Hattermann, K., Ehlers, B. & Soike, D. (2000). Cloning and sequencing of columbid circovirus (CoCV), a new circovirus from pigeons. *Archives of virology* **145**, 2469-2479.

Mankertz, A., Çaliskan, R., Hattermann, K., Hillenbrand, B., Kurzendoerfer, P., Mueller, B., Schmitt, C., Steinfeldt, T. & Finsterbusch, T. (2004). Molecular biology of Porcine circovirus: analyses of gene expression and viral replication. *Veterinary microbiology* **98**, 81-88.

Mansoor, S., Briddon, R. W., Zafar, Y. & Stanley, J. (2003). Geminivirus disease complexes: an emerging threat. *Trends in Plant Science* **8**, 128-134.

Mardis, E. R. (2008a). Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9**, 387-402.

Mardis, E. R. (2008b). The impact of next-generation sequencing technology on genetics. *Trends in genetics* **24**, 133-141.

Markham, P. G., Bedford, I. D., Liu, S. & Pinner, M. S. (1994). The transmission of geminiviruses by *Bemisia tabaci*. *Pesticide Science* **42**, 123-128.

Martin, D. P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P. & Varsani, A. (2011). Recombination in eukaryotic single stranded DNA viruses. *Viruses* **3**, 1699-1738.

Martinez, C. J. & Clark, M. W. (2012). Reclaimed water and Florida's water reuse program.

Mauricio-Castillo, J., Torres-Herrera, S., Cárdenas-Conejo, Y., Pastor-Palacios, G., Méndez-Lozano, J. & Argüello-Astorga, G. (2014). A novel begomovirus isolated from sida contains putative cis-and trans-acting replication specificity determinants that have evolved independently in several geographical lineages. *Archives of virology*, 1-12.

McKenna, R., Xia, D., Willingmann, P., Hag, L. L., Krishnaswamy, S., Rossmann, M. G., Olson, N. H., Baker, T. S. & Incardona, N. L. (1992). Atomic structure of single-stranded DNA bacteriophage ΦX174 and its functional implications. *Nature* **355**, 137.

Meehan, B. M., McNeilly, F., Todd, D., Kennedy, S., Jewhurst, V. A., Ellis, J. A., Hassard, L. E., Clark, E. G., Haines, D. M. & other authors (1998). Characterization of novel circovirus DNAs associated with wasting syndromes in pigs. *Journal of General Virology* **79**, 2171-2179.

Meleg, E., Bányai, K., Martella, V., Jiang, B., Kocsis, B., Kisfali, P., Melegh, B. & Szűcs, G. (2008). Detection and quantification of group C rotaviruses in communal sewage. *Applied and environmental microbiology* **74**, 3394-3399.

Miller, J. R., Koren, S. & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics* **95**, 315-327.

Mokili, J. L., Rohwer, F. & Dutilh, B. E. (2012). Metagenomics and future perspectives in virus discovery. *Current Opinion in Virology* **2**, 63-77.

Morace, G., Aulicino, F., Angelozzi, C., Costanzo, L., Donadio, F. & Rapicetta, M. (2002). Microbial quality of wastewater: detection of hepatitis A virus by reverse transcriptase-polymerase chain reaction. *Journal of applied microbiology* **92**, 828-836.

Morozov, I., Sirinarumitr, T., Sorden, S. D., Halbur, P. G., Morgan, M. K., Yoon, K. J. & Paul, P. S. (1998). Detection of a novel strain of porcine circovirus in pigs with postweaning multisystemic wasting syndrome. *Journal of clinical microbiology* **36**, 2535-2541.

Mubin, M., Shahid, M., Tahir, M., Briddon, R. & Mansoor, S. (2010). Characterization of begomovirus components from a weed suggests that begomoviruses may associate with multiple distinct DNA satellites. *Virus genes* **40**, 452-457.

Mullineaux, P. M., Guerineau, F. & Accotto, G. P. (1990). Processing of complementary sense RNAs of Digitaria streak virus in its host and in transgenic tobacco. *Nucleic Acids Research* **18**, 7259.

Mundry, M., Bornberg-Bauer, E., Sammeth, M. & Feulner, P. G. (2012). Evaluating characteristics of de novo assembly software on 454 transcriptome data: a simulation approach. *PloS one* **7**, e31410.

Muzyczka, N. & Berns, K. (2001). Parvoviridae: the viruses and their replication. *Fields virology* **2**, 2327-2359.

Naccache, S. N., Greninger, A. L., Lee, D., Coffey, L. L., Phan, T., Rein-Weston, A., Aronsohn, A., Hackett, J., Delwart, E. L. & other authors (2013). The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *Journal of virology* **87**, 11966-11977.

Nagasaki, K., Tomaru, Y., Katanozaka, N., Shirai, Y., Nishida, K., Itakura, S. & Yamaguchi, M. (2004). Isolation and characterization of a novel single-stranded RNA virus infecting the bloom-forming diatom *Rhizosolenia setigera*. *Applied and environmental microbiology* **70**, 704-711.

Nagasaki, K., Tomaru, Y., Takao, Y., Nishida, K., Shirai, Y., Suzuki, H. & Nagumo, T. (2005). Previously unknown virus infects marine diatom. *Applied and environmental microbiology* **71**, 3528-3535.

Nash, T. E., Dallas, M. B., Reyes, M. I., Buhrman, G. K., Ascencio-Ibanez, J. & Hanley-Bowdoin, L. (2011). Functional Analysis of a Novel Motif Conserved across Geminivirus Rep Proteins. *Journal of Virology* **85**, 1182.

Nelson, D. M., Tréguer, P., Brzezinski, M. A., Leynaert, A. & Quéguiner, B. (1995). Production and dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *Global Biogeochemical Cycles* **9**, 359-372.

Ng, T. F. F., Alavandi, S., Varsani, A., Burghart, S. & Breitbart, M. (2013). Metagenomic identification of a nodavirus and a circular ssDNA virus in semi-purified viral nucleic acids from the hepatopancreas of healthy Farfantepenaeus duorarum shrimp.

Ng, T. F. F., Manire, C., Borrowman, K., Langer, T., Ehrhart, L. & Breitbart, M. (2009). Discovery of a novel single-stranded DNA virus from a sea turtle fibropapilloma by using viral metagenomics. *Journal of virology* **83**, 2500-2509.

Ng, T. F. F., Duffy, S., Polston, J. E., Bixby, E., Vallad, G. E. & Breitbart, M. (2011a). Exploring the Diversity of Plant DNA Viruses and Their Satellites Using Vector-Enabled Metagenomics on Whiteflies. *PloS one* **6**, e19050.

Ng, T. F. F., Marine, R., Wang, C., Simmonds, P., Kapusinszky, B., Bodhidatta, L., Oderinde, B. S., Wommack, K. E. & Delwart, E. (2012). High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. *Journal of virology* **86**, 12161-12175.

Ng, T. F. F., Willner, D. L., Lim, Y. W., Schmieder, R., Chau, B., Nilsson, C., Anthony, S., Ruan, Y., Rohwer, F. & other authors (2011b). Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. *PloS one* **6**, e20579.

Ng, T. F. F., Chen, L.-F., Zhou, Y., Shapiro, B., Stiller, M., Heintzman, P. D., Varsani, A., Kondov, N. O., Wong, W. & other authors (2014). Preservation of viral genomes in 700-y-old caribou feces from a subarctic ice patch. *Proceedings of the National Academy of Sciences*, 201410429.

Niagro, F., Forsthoefel, A., Lawther, R., Kamalanathan, L., Ritchie, B., Latimer, K. & Lukert, P. (1998). Beak and feather disease virus and porcine circovirus genomes: intermediates between the geminiviruses and plant circoviruses. *Archives of virology* **143**, 1723-1744.

Niu, B., Fu, L., Sun, S. & Li, W. (2010). Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC bioinformatics* **11**, 187.

Okamoto, H. & Mayumi, M. (2001). TT virus: virological and genomic characteristics and disease associations. *Journal of gastroenterology* **36**, 519-529.

Okamoto, H., Nishizawa, T., Takahashi, M., Tawara, A., Peng, Y., Kishimoto, J. & Wang, Y. (2001). Genomic and evolutionary characterization of TT virus (TTV) in tupaias and comparison with species-specific TTVs in humans and non-human primates. *Journal of General Virology* **82**, 2041-2050.

Okamoto, H., Takahashi, M., Nishizawa, T., Tawara, A., Fukai, K., Muramatsu, U., Naito, Y. & Yoshikawa, A. (2002). Genomic characterization of TT viruses (TTVs) in pigs, cats and dogs and their relatedness with species-specific TTVs in primates and tupaias. *Journal of General Virology* **83**, 1291-1297.

Owor, B. E., Martin, D. P., Shepherd, D. N., Edema, R., Monjane, A. L., Rybicki, E. P., Thomson, J. A. & Varsani, A. (2007). Genetic analysis of maize streak virus isolates from

Uganda reveals widespread distribution of a recombinant variant. *Journal of General Virology* **88**, 3154-3165.

Padilla-Rodriguez, M., Rosario, K. & Breitbart, M. (2013). Novel cyclovirus discovered in the Florida woods cockroach *Eurycotis floridana* (Walker). *Archives of virology* **158**, 1389-1392.

Phan, T. G., Luchsinger, V., Avendaño, L. F., Deng, X. & Delwart, E. (2014). Cyclovirus in nasopharyngeal aspirates of Chilean children with respiratory infections. *Journal of General Virology* **95**, 922-927.

Phan, T. G., Kapusinszky, B., Wang, C., Rose, R. K., Lipton, H. L. & Delwart, E. L. (2011). The Fecal Viral Flora of Wild Rodents. *PLoS pathogens* **7**, e1002218.

Philippe, N., Legendre, M., Doutre, G., Couté, Y., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L. & other authors (2013). Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* **341**, 281-286.

Pilartz, M. & Jeske, H. (1992). Abutilon mosaic geminivirus double-stranded DNA is packed into minichromosomes. *Virology* **189**, 800-802.

Pina, S., Buti, M., Jardí, R., Clemente-Casares, P., Jofre, J. & Girones, R. (2001). Genetic analysis of hepatitis A virus strains recovered from the environment and from patients with acute hepatitis. *Journal of General Virology* **82**, 2955-2963.

Poojari, S., Alabi, O. J., Fofanov, V. Y. & Naidu, R. A. (2013). A leafhopper-transmissible DNA virus with novel evolutionary lineage in the family geminiviridae implicated in grapevine redleaf disease by next-generation sequencing. *PloS one* **8**, e64194.

Pooma, W., Gillette, W. K., Jeffrey, J. L. & Petty, I. T. D. (1996). Host and viral factors determine the dispensability of coat protein for bipartite geminivirus systemic movement. *Virology* **218**, 264-268.

Prescott, L. E., Simmonds, P., Collaborators & International (1998). Global distribution of transfusion-transmitted virus. *New England Journal of Medicine* **339**, 776-777.

Radford, A. D., Chapman, D., Dixon, L., Chantrey, J., Darby, A. C. & Hall, N. (2012). Application of next-generation sequencing technologies in virology. *Journal of General Virology* **93**, 1853-1868.

Raidal, S. R. (1995). Viral skin diseases of birds. In *Seminars in Avian and Exotic Pet Medicine*, pp. 72-82: Elsevier.

Ritchie, B. W., Niagro, F. D., Lukert, P. D., Latimer, K. S., Steffens III, W. L. & Pritchard, N. (1989). A review of psittacine beak and feather disease: characteristics of the PBFD virus. *Journal of the Association of Avian Veterinarians*, 143-149.

Rodriguez-Brito, B., Li, L., Wegley, L., Furlan, M., Angly, F., Breitbart, M., Buchanan, J., Desnues, C., Dinsdale, E. & other authors (2010). Viral and microbial community dynamics in four aquatic environments. *The ISME journal* **4**, 739-751.

Romay, G., Chirinos, D., Geraud-Pouey, F. & Desbiez, C. (2010). Association of an atypical alphasatellite with a bipartite New World begomovirus. *Archives of virology* **155**, 1843-1847.

Rosario, K. & Breitbart, M. (2011). Exploring the viral world through metagenomics. *Current Opinion in Virology* **1**, 289-297.

Rosario, K., Duffy, S. & Breitbart, M. (2009a). Diverse circovirus-like genome architectures revealed by environmental metagenomics. *Journal of General Virology* **90**, 2418-2424.

Rosario, K., Duffy, S. & Breitbart, M. (2012a). A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Archives of virology* **157**, 1851-1871.

Rosario, K., Nilsson, C., Lim, Y. W., Ruan, Y. & Breitbart, M. (2009b). Metagenomic analysis of viruses in reclaimed water. *Environmental microbiology* **11**, 2806-2820.

Rosario, K., Padilla-Rodriguez, M., Kraberger, S., Stainton, D., Martin, D. P., Breitbart, M. & Varsani, A. (2013). Discovery of a novel mastrevirus and alphasatellite-like circular DNA in dragonflies (Epiprocta) from Puerto Rico. *Virus Research*, 231-237.

Rosario, K., Dayaram, A., Marinov, M., Ware, J., Kraberger, S., Stainton, D., Breitbart, M. & Varsani, A. (2012b). Diverse circular ssDNA viruses discovered in dragonflies (Odonata: Epiprocta). *Journal of General Virology* **93**, 2668-2681.

Rosario, K., Marinov, M., Stainton, D., Kraberger, S., Wiltshire, E. J., Collings, D. A., Walters, M., Martin, D. P., Breitbart, M. & other authors (2011). Dragonfly cyclovirus, a novel single-stranded DNA virus discovered in dragonflies (Odonata: Anisoptera). *Journal of General Virology* **92**, 1302-1308.

Roux, S., Faubladier, M., Mahul, A., Paulhe, N., Bernard, A., Debroas, D. & Enault, F. (2011). Metavir: a web server dedicated to virome analysis. *Bioinformatics* **27**, 3074-3075.

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., Colombet, J., Sime-Ngando, T. & Debroas, D. (2012). Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PloS one* **7**, e33641.

Saeed, M., Zafar, Y., Randles, J. W. & Rezaian, M. A. (2007). A monopartite begomovirus-associated DNA β satellite substitutes for the DNA B of a bipartite begomovirus to permit systemic infection. *Journal of General Virology* **88**, 2881-2889.

Sanger, F. & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* **94**, 441-448.

Sano, E., Carlson, S., Wegley, L. & Rohwer, F. (2004). Movement of viruses between biomes. *Applied and environmental microbiology* **70**, 5842-5846.

Savory, F. & Ramakrishnan, U. (2014). Asymmetric patterns of reassortment and concerted evolution in *Cardamom bushy dwarf virus*. *Infection, Genetics and Evolution* **24**, 15-24.

Schoenfeld, T., Patterson, M., Richardson, P. M., Wommack, K. E., Young, M. & Mead, D. (2008). Assembly of viral metagenomes from Yellowstone hot springs. *Applied and environmental microbiology* **74**, 4164-4174.

Sedmak, G., Bina, D., MacDonald, J. & Couillard, L. (2005). Nine-year study of the occurrence of culturable viruses in source water for two drinking water treatment plants and the influent and effluent of a wastewater treatment plant in Milwaukee, Wisconsin (August 1994 through July 2003). *Applied and environmental microbiology* **71**, 1042-1050.

Shan, T., Li, L., Simmonds, P., Wang, C., Moeser, A. & Delwart, E. (2011). The fecal virome of pigs on a high-density farm. *Journal of virology* **85**, 11697-11708.

Sharman, M., Thomas, J., Skabo, S. & Holton, T. (2008). Abacá bunchy top virus, a new member of the genus Babuvirus (family *Nanoviridae*). *Archives of virology* **153**, 135-147.

Shokralla, S., Spall, J. L., Gibson, J. F. & Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology* **21**, 1794-1805.

Sikorski, A., Dayaram, A. & Varsani, A. (2013a). Identification of a novel circular DNA virus in New Zealand fur seal (*Arctocephalus forsteri*) fecal matter. *Genome announcements* **1**, e00558-00513.

Sikorski, A., Argüello-Astorga, G. R., Dayaram, A., Dobson, R. C. J. & Varsani, A. (2013b). Discovery of a novel circular single-stranded DNA virus from porcine faeces. *Archives of virology* **158**, 283-289.

Sikorski, A., Kearvell, J., Elkington, S., Dayaram, A., Argüello-Astorga, G. R. & Varsani, A. (2013c). Novel ssDNA viruses discovered in yellow-crowned parakeet (*Cyanoramphus auriceps*) nesting material. *Archives of virology*, 1603-1607.

Sikorski, A., Massaro, M., Kraberger, S., Young, L. M., Smalley, D., Martin, D. P. & Varsani, A. (2013d). Novel myco-like DNA viruses discovered in the faecal matter of various animals. *Virus Research* **177**, 209-216.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. & Birol, İ. (2009). ABySS: a parallel assembler for short read sequence data. *Genome research* **19**, 1117-1123.

Sinsheimer, R. L. (1959). A single-stranded deoxyribonucleic acid from bacteriophage [phi] X174+. *Journal of Molecular Biology* **1**, 43-53, IN46.

Smith, R. J., Jeffries, T. C., Roudnew, B., Seymour, J. R., Fitch, A. J., Simons, K. L., Speck, P. G., Newton, K., Brown, M. H. & other authors (2013). Confined aquifers as viral reservoirs. *Environmental microbiology reports* **5**, 725-730.

Smits, S. L., Zijlstra, E., van Hellemond, J. J., Schapendonk, C. M., Bodewes, R., Schürch, A. C., Haagmans, B. L. & Osterhaus, A. D. (2013). Novel cyclovirus in human cerebrospinal fluid, Malawi, 2010–2011. *Emerging infectious diseases* **19**, 1511.

Soike, D., Hattermann, K., Albrecht, K., Segalés, J., Domingo, M., Schmitt, C. & Mankertz, A. (2001). A diagnostic study on columbid circovirus infection. *Avian Pathology* **30**, 605-611.

Stainton, D., Kraberger, S., Walters, M., Wiltshire, E. J., Rosario, K., Lolohea, S., Katoa, I., Tu'amelie, H. F., Aholelei, W. & other authors (2012). Evidence of inter-component recombination, intra-component recombination and reassortment in banana bunchy top virus. *Journal of General Virology* **93**, 1103-1119.

Stenzel, T., Piasecki, T., Chrząstek, K., Julian, L., Muhire, B. M., Golden, M., Martin, D. P. & Varsani, A. (2014). Pigeon circoviruses display patterns of recombination, genomic secondary structure and selection similar to those of beak and feather disease viruses. *Journal of General Virology* **95**, 1338-1351.

Stewart, M. E., Perry, R. & Raidal, S. R. (2006). Identification of a novel circovirus in Australian ravens (*Corvus coronoides*) with feather disease. *Avian Pathology* **35**, 86-92.

Story, R. M., Weber, I. T. & Steitz, T. A. (1992). The structure of the *E. coli* recA protein monomer and polymer.

Sudarshana, M., Gonzalez, A., Dave, A., Wei, A., Smith, R., Anderson, M. & Walker, A. (2013). Grapevine red blotch-associated virus is widespread in California and US vineyards. *Phytopathology* **103**, S2.

Sung, Y. & Coutts, R. H. (1995). Pseudorecombination and complementation between potato yellow mosaic geminivirus and tomato golden mosaic geminivirus. *Journal of General Virology* **76**, 2809-2815.

Suttle, C. A. (2007). Marine viruses—major players in the global ecosystem. *Nature Reviews Microbiology* **5**, 801-812.

Svraka, S., Rosario, K., Duizer, E., van der Avoort, H., Breitbart, M. & Koopmans, M. (2010). Metagenomic sequencing for virus identification in a public-health setting. *Journal of General Virology* **91**, 2846-2856.

Thurber, R. V. (2011). Methods in Viral Metagenomics. *Handbook of Molecular Microbial Ecology II*, 15-24.

Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. (2009). Laboratory procedures to generate viral metagenomes. *Nature protocols* **4**, 470-483.

Timchenko, T., De Kouchkovsky, F., Katul, L., David, C., Vetten, H. J. & Gronenborn, B. (1999). A single rep protein initiates replication of multiple genome components of *faba bean necrotic yellows virus*, a single-stranded DNA virus of plants. *Journal of virology* **73**, 10173-10182.

Tischer, I., Gelderblom, H., Vettermann, W. & Koch, M. (1982). A very small porcine virus with circular single-stranded DNA.

Todd, D. (2000). Circoviruses: immunosuppressive threats to avian species: a review. *Avian Pathology* **29**, 373-394.

Todd, D. (2004). Avian circovirus diseases: lessons for the study of PMWS. *Veterinary microbiology* **98**, 169-174.

Todd, D., Weston, J., Soike, D. & Smyth, J. (2001a). Genome sequence determinations and analyses of novel circoviruses from goose and pigeon. *Virology* **286**, 354-362.

Todd, D., Scott, A., Fringuelli, E., Shivraprasad, H., Gavier-Widen, D. & Smyth, J. (2007). Molecular characterization of novel circoviruses from finch and gull. *Avian Pathology* **36**, 75-81.

Todd, D., Weston, J., Ball, N., Borghmans, B., Smyth, J., Gelmini, L. & Lavazza, A. (2001b). Nucleotide sequence-based identification of a novel circovirus of canaries. *Avian Pathology* **30**, 321-325.

Tomaru, Y., Shirai, Y., Toyoda, K. & Nagasaki, K. (2011a). Isolation and characterisation of a single-stranded DNA virus infecting the marine planktonic diatom *Chaetoceros tenuissimus*. *Aquatic Microbial Ecology* **64**, 175.

Tomaru, Y., Takao, Y., Suzuki, H., Nagumo, T., Koike, K. & Nagasaki, K. (2011b). Isolation and characterization of a single-stranded DNA virus infecting *Chaetoceros lorenzianus* Grunow. *Applied and environmental microbiology* **77**, 5285-5293.

Unseld, S., Ringel, M., Höfer, P., Höhnle, M., Jeske, H., Bedford, I., Markham, P. & Frischmuth, T. (2000). Host range and symptom variation of pseudorecombinant virus produced by two distinct bipartite geminiviruses. *Archives of virology* **145**, 1449-1454.

van den Brand, J. M. A., van Leeuwen, M., Schapendonk, C. M., Simon, J. H., Haagmans, B. L., Osterhaus, A. D. M. E. & Smits, S. L. (2011). Metagenomic analysis of the viral flora of pine marten and European badger feces. *Journal of Virology* **86**, 2360-2365.

van Doorn, H. R., Nghia, H. D. T., Chau, T. T. H., de Vries, M., Canuti, M., Deijs, M., Jebbink, M. F., Baker, S., Bryant, J. E. & other authors (2013). Identification of a new cyclovirus in cerebrospinal fluid of patients with acute central nervous system infections. *MBio* **4**, e00231-00213.

Van Regenmortel, M. H. & Mahy, B. W. (2004). Emerging issues in virus taxonomy. *Emerging infectious diseases* **10**, 8-13.

van Regenmortel, M. H., Fauquet, C., Bishop, D., Carstens, E., Estes, M., Lemon, S., Maniloff, J., Mayo, M., McGeoch, D. & other authors (2000). *Virus taxonomy: classification and nomenclature of viruses. Seventh report of the International Committee on Taxonomy of Viruses*. Academic Press.

Van Tan, L., van Doorn, H. R., Nghia, H. D. T., Chau, T. T. H., de Vries, M., Canuti, M., Deijs, M., Jebbink, M. F., Baker, S. & other authors (2013). Identification of a new cyclovirus in cerebrospinal fluid of patients with acute central nervous system infections. *MBio* **4**, e00231-00213.

Varsani, A., Navas-Castillo, J., Moriones, E., Hernández-Zepeda, C., Idris, A., Brown, J. K., Zerbini, F. M. & Martin, D. P. (2014). Establishment of three new genera in the family *Geminiviridae*: Becurtovirus, Eragrovirus and Turncurtovirus. *Archives of virology*, 1-11.

Varsani, A., Shepherd, D. N., Monjane, A. L., Owor, B. E., Erdmann, J. B., Rybicki, E. P., Peterschmitt, M., Briddon, R. W., Markham, P. G. & other authors (2008). Recombination, decreased host specificity and increased mobility may have driven the emergence of maize streak virus as an agricultural pathogen. *Journal of General Virology* **89**, 2063-2074.

Victoria, J. G., Kapoor, A., Li, L., Blinkova, O., Slikas, B., Wang, C., Naeem, A., Zaidi, S. & Delwart, E. (2009). Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *Journal of virology* **83**, 4642-4651.

Walker, J. E., Saraste, M., Runswick, M. J. & Gay, N. J. (1982). 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. *The EMBO journal* **1**, 945.

Weinbauer, M. G. (2004). Ecology of prokaryotic viruses. *FEMS microbiology reviews* **28**, 127-181.

Werner, D. (1977). Introduction with a note on taxonomy. *Botanical monographs.*

Whon, T. W., Kim, M.-S., Roh, S. W., Shin, N.-R., Lee, H.-W. & Bae, J.-W. (2012). Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere. *Journal of virology* **86**, 8221-8231.

Wright, E. A., Heckel, T., Groenendijk, J., Davies, J. W. & Boulton, M. I. (1997). Splicing features in *maize streak virus* virion and complementary sense gene expression. *The Plant Journal* **12**, 1285-1297.

Wu, B., Melcher, U., Guo, X., Wang, X., Fan, L. & Zhou, G. (2008). Assessment of codivergence of Mastreviruses with their plant hosts. *BMC evolutionary biology* **8**, 335.

Wu, Z., Ren, X., Yang, L., Hu, Y., Yang, J., He, G., Zhang, J., Dong, J., Sun, L. & other authors (2012). Virome analysis for identification of novel mammalian viruses in bat species from Chinese provinces. *Journal of virology* **86**, 10999-11012.

Xu, Y. & Price, B. D. (2011). Chromatin dynamics and the repair of DNA double strand breaks. *Cell Cycle* **10**, 261-267.

Yoon-Robarts, M., Blouin, A. G., Bleker, S., Kleinschmidt, J. A., Aggarwal, A. K., Escalante, C. R. & Linden, R. M. (2004). Residues within the B′ motif are critical for DNA binding by the superfamily 3 helicase Rep40 of adeno-associated virus type 2. *Journal of Biological Chemistry* **279**, 50472-50481.

Yoon, H. S., Price, D. C., Stepanauskas, R., Rajah, V. D., Sieracki, M. E., Wilson, W. H., Yang, E. C., Duffy, S. & Bhattacharya, D. (2011). Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* **332**, 714-717.

Yoshida, M., Takaki, Y., Eitoku, M., Nunoura, T. & Takai, K. (2013). Metagenomic Analysis of Viral Communities in (Hado) Pelagic Sediments. *PloS one* **8**, e57271.

Young, H. W. & York, D. W. (1996). Reclaimed water reuse in Florida and the South Gulf Coast. *Florida Water Resour J*, 32-36.

Yu, X., Li, B., Fu, Y., Jiang, D., Ghabrial, S. A., Li, G., Peng, Y., Xie, J., Cheng, J. & other authors (2010). A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. *Proceedings of the National Academy of Sciences* **107**, 8387.

Zawar-Reza, P., Argüello-Astorga, G. R., Kraberger, S., Julian, L., Stainton, D., Broady, P. A. & Varsani, A. (2014). Diverse small circular single-stranded DNA viruses identified in a freshwater pond on the McMurdo Ice Shelf (Antarctica). *Infection, Genetics and Evolution* **26**, 132-138.

Zerbino, D. R. & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**, 821-829.

Zhang, J., Chiodini, R., Badr, A. & Zhang, G. (2011). The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics* **38**, 95-109.

Zhang, L., Fu, Y., Xie, J., Jiang, D., Li, G. & Yi, X. (2009). A novel virus that infecting hypovirulent strain XG36-1 of plant fungal pathogen *Sclerotinia sclerotiorum*. *Virology journal* **6**, 96.

Zhang, T., Breitbart, M., Lee, W. H., Run, J.-Q., Wei, C. L., Soh, S. W. L., Hibberd, M. L., Liu, E. T., Rohwer, F. & other authors (2005). RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS biology* **4**, e3.

Zhou, X., Ren, L., Li, Y., Zhang, M., Yu, Y. & Yu, J. (2010). The next-generation sequencing technology: a technology review and future perspective. *Science China Life Sciences* **53**, 44-57.

Chapter 2

High global diversity of cycloviruses amongst dragonflies

Contents

This work has been published in Journal of General Virology and is presented in a similar manner to that of the publication:

Dayaram, A., Potter, K., Moline, A.B., Rosenstein, D.D., Marinov, M., Thomas, J.E., Breitbart, M., Rosario, K., Argüello-Astorga, G.R., Varsani, A. (2013) High global diversity of cycloviruses amongst dragonflies. *Journal of General Virology* 94: 1827-1840

2.1 Abstract

Members of the *Circoviridae* family, specifically the Circovirus genus, were thought to infect only vertebrates. However, members of a sister group under the same family, the proposed Cyclovirus genus, have been detected recently in insects. In an effort to explore the diversity of cycloviruses and better understand the evolution of these novel single-stranded DNA (ssDNA) viruses, here we present five cycloviruses isolated from three dragonfly species (*Orthetrum sabina, Xanthocnemis zealandica* and *Aeshna multicolour*) collected in Australia, New Zealand and the United States of America. The genomes of these five viruses share similar genome structure to other cycloviruses, with a circular ~1.7 kb genome and two major bidirectionally transcribed open reading frames (ORFs). The genomic sequence data gathered during this study were combined with all cyclovirus genomes available in public databases to identify conserved motifs and regulatory elements in the intergenic regions, as well as determine diversity and recombinant regions within their genomes. The genomes reported here represent four different cyclovirus species, three of which are novel. Our results confirm that cycloviruses circulate widely in winged insect populations. In eight different cyclovirus species identified in dragonflies to date, some of these exhibit a broad geographical distribution. Recombination analysis revealed both intra- and inter-species recombination events among cycloviruses, including genomes recovered from disparate sources (e.g., goat meat and human faeces) indicating that recombination may play an important role in cyclovirus evolution which is similar to other well-characterised circular ssDNA viruses.

2.2 Introduction

A wealth of diverse single-stranded DNA (ssDNA) viruses are being discovered due to emerging technologies such as metagenomics and next generation sequencing. Although most of these novel ssDNA viruses are similar in encoding only two open reading frames (ORFs), they have high diversity in genome organisation and in the replication-associated protein (Rep) (Blinkova *et al.*, 2010; Ge *et al.*, 2011; Kapoor *et al.*, 2012; Kim *et al.*, 2012; Li *et al.*, 2011; Li *et al.*, 2010; Phan *et al.*, 2011; Rosario *et al.*, 2012b; Rosario *et al.*, 2011; Sikorski *et al.*, 2013; van den Brand *et al.*, 2012), reviewed in Rosario et al. (Rosario *et al.*, 2012a). Within this diverse group of novel ssDNA viruses there is a group of viruses that share significant similarities to the well- characterised circovirus isolates of the *Circoviridae* family. This, in turn, has led to the proposal of a new genus, Cyclovirus, within the *Circoviridae* family. Similar to circoviruses, the cycloviruses encode two open reading frames (ORFs) that are bidirectionally transcribed, which encode the Rep and the capsid protein (CP). Cycloviruses were first discovered in the stool samples from humans in Pakistan, Tunisia and Nigeria, chimpanzees in Central Africa and meat products of farm animals from Nigeria and Pakistan (Li *et al.*, 2010). Subsequently, cycloviruses have been isolated from insectivorous bats and insects, including various dragonfly species and a Florida woods cockroach (Delwart & Li, 2012; Ge *et al.*, 2011; Li *et al.*, 2011; Padilla-Rodriguez *et al.*, 2013; Rosario *et al.*, 2012b; Rosario *et al.*, 2011).

Dragonflies (Odonata: Epiprocta) were the first insects from which circular ssDNA viruses, similar to vertebrate-infecting viruses, were characterised (Rosario *et al.*, 2011). Rosario *et al.* (2011) first report analysed 21 ssDNA viral genomes (~1.7 kb) from three different dragonfly species collected in the Kingdom of Tonga revealing that they represented a single cyclovirus species, and presenting evidence of recombination amongst the reported viral genomes. Subsequently, a study identified diverse cycloviruses in dragonflies from Bulgaria, the United States of America (USA; Florida) and Puerto Rico indicating the widespread distribution of these viruses (Rosario *et al.*, 2012b). Recently, a novel cyclovirus was described from another insect, *Eurycotis floridana* (Florida woods cockroach-associated cyclovirus GS140; FWCasCyV-GS140; JX569794), suggesting that cycloviruses are indeed widespread amongst insect species (Padilla-Rodriguez *et al.*, 2013).

In addition to cycloviruses, many novel ssDNA viruses have been recovered from dragonflies, and thus we have proposed that dragonflies, being top-end insect predators, could be used as ssDNA viral sampling tools in ecosystems by combining their insect-hunting ability with methods that enrich for circular ssDNA viruses (Rosario *et al.*, 2012b). By sampling dragonflies in an agricultural field in Puerto Rico, Rosario *et al.* (2013) identified a novel plant-infecting geminivirus and an associated satellite DNA, including the first mastrevirus and alphasatellite-like molecule, usually associated with begomoviruses, from the Caribbean (Rosario *et al.*, 2013). The above example suggests there may be bioaccumulation of insect-transmitted and insect-infecting viruses in dragonflies. As insects transmit most plant viruses, and are further consumed by dragonflies, it is plausible that dragonflies maybe a natural reservoir for diverse ssDNA viruses.

As a continuing effort to determine the diversity and evolution of cycloviruses amongst insects, in this chapter we characterise at the genome and protein level five new cycloviruses, three of which represent new species. The reported cycloviruses were characterised from single specimens belonging to three different dragonfly species (*Orthetrum sabina, Xanthocnemis zealandica* and *Aeshna multicolour*) collected at Wappa Falls Dam, Brisbane (Australia), Lake Pearson, Canterbury (New Zealand) and the Kachina Wetlands of Arizona (USA). These novel dragonfly cycloviruses, in conjunction with cycloviruses reported in GenBank, were used to perform phylogenetic and recombination analyses among members of the proposed Cyclovirus genus. A classification scheme for cycloviruses is proposed based on the analyses presented here.

2.3 Materials and methods

2.3.1 Viral particle purification and DNA extraction

Adult dragonfly specimens, including *O. sabina (n=1), X. zealandica (n=1) and A. multicolor (n=1)*, were caught using insect nets in Australia (Wappa Falls dam, Queensland), New Zealand (Lake Pearson, Canterbury) and the USA (Kachina Wetlands of Arizona), respectively. The dragonfly samples were preserved either in 95% ethanol or by freezing upon collection. The abdomen from each specimen was dissected and removed from each individual in a sterile environment and then homogenized in SM buffer (0.1 M NaCl, 50 mM Tris/HCl, pH 7.4, 10 mM $MgSO_4$) at a ratio of 1.5 ml SM buffer to 0.5 g of tissue. The cellular debris was pelleted by centrifuging the sample (10,000 *x g* for 10 min). The resulting supernatant was then filtered in a step-wise fashion through syringe filters (Sartorius Stedim Biotech, Germany), first through a 0.45 µm pore size filter and then through a 0.2 µm filter to partially purify virus particles. Viral DNA was then extracted from the filtrate using a High Pure Viral Nucleic Acid Kit (Roche, USA).

2.3.2 Enrichment of circular ssDNA and identification of novel ssDNA viruses

Circular DNA molecules in the viral DNA extracts were enriched by rolling circle amplification using TempliPhi (GE Healthcare) as described previously (Dayaram *et al.*, 2012; Rosario *et al.*, 2013; Rosario *et al.*, 2012b; Rosario *et al.*, 2011; Sikorski *et al.*, 2013). The resulting concatenated DNA was digested with *Bam*HI, *Eco*RI and *Xmn*I in separate reactions yielding between ~1 to 1.7 kb DNA fragments. These fragments were gel purified and cloned into pGEM3ZF (+) (Promega, USA) plasmid restricted with *Eco*RI or into pUC-19 plasmid vector restricted with *Bam*HI or *Sma*I (Fermentas, USA). The resulting clones were sequenced at Macrogen Inc. (South Korea) by primer walking. The genomes of the putative viruses were verified as complete genomes either by designing back-to-back primers followed by PCR amplification of the genome using Kapa HiFi DNA polymerase (Kapa Biosystems, USA), cloning the amplicon and sequencing the recombinant plasmid or by restriction mapping (see Additional Table 2.1 for details).

2.3.3 Viral genome and phylogenetic analysis

The genomic sequences were assembled using DNAMAN (version 5.2.9; Lynnon Biosoft), and preliminary analysis using BLASTx and tBLASTx (Altschul *et al.*, 1990) revealed that they had similarities to cycloviruses. A dataset of all cycloviruses (n=55, including 5 from this study) was generated by downloading all publically available cycloviruses in GenBank on 1st Dec 2012, and all the genomes were linearised at the conserved nonanucleotide motif (TAGTATTAC). The cyclovirus dataset, together with the reverse complemented genomes of Porcine circovirus (PCV-2, AY424401) and Beak and feather disease virus (BFDV, AF071878) to match the orientation of cyclovirus ORFs, were aligned with MUSCLE (Edgar, 2004) with manual editing. The resulting alignment was used to infer a ML phylogenetic tree using PHYML version 3 (Guindon et al., 2010) with GTR+I+G4 as the best substitution model and with approximate likelihood-ratio test (aLRT) branch support (Anisimova & Gascuel, 2006) (Figure 2.1). PCV-2 and BFDV, as representatives of circoviruses, were used to root the tree, and branches with less than 60% aLRT support were collapsed using Mesquite (version 2.75).

Datasets of all the amino acid sequences of the Rep and CP were assembled to determine the phylogenetic relationships and diversity of these proteins within the genomes. The amino acid sequences were aligned with MUSCLE (Edgar, 2004), and the resulting alignments were used to infer ML phylogenetic trees using the LG model of amino acid substitution with aLRT branch support (Figure 2.4) and rooted with PCV-2 and BFDV sequences. Pairwise amino acid identities (p-distance) were calculated using MUSCLE (Edgar, 2004) algorithm implemented in MEGA4.

Genome-wide percentage pairwise identities were calculated using SDTv 1.0 (Muhire *et al.*, 2013). Sequences were aligned using MUSCLE (Edgar, 2004) and pairwise identities calculated ([1-pairwise distance]x100) with pairwise deletion of gaps. In this manner 1485 pairs of pairwise identities were calculated and their distribution plotted to provide insights for the classification of cyclovirus isolates.

2.3.4 Recombination analysis

The cyclovirus full genome dataset was aligned using MUSCLE (Edgar, 2004), and the resulting alignment was used as an input file for recombination analysis using RDP4 (Martin *et al.*, 2010) with the following methods: RDP (Martin & Rybicki, 2000); GENECONV (Padidam *et al.*, 1999); Bootscan (Martin *et al.*, 2005); Maxchi (Smith, 1992); Chimera (Posada & Crandall, 2001); Siscan (Gibbs *et al.*, 2000); and 3Seq (Boni *et al.*, 2007). Only recombination events detected by a minimum of three methods and coupled with clear phylogenetic evidence were considered as putative recombination events.

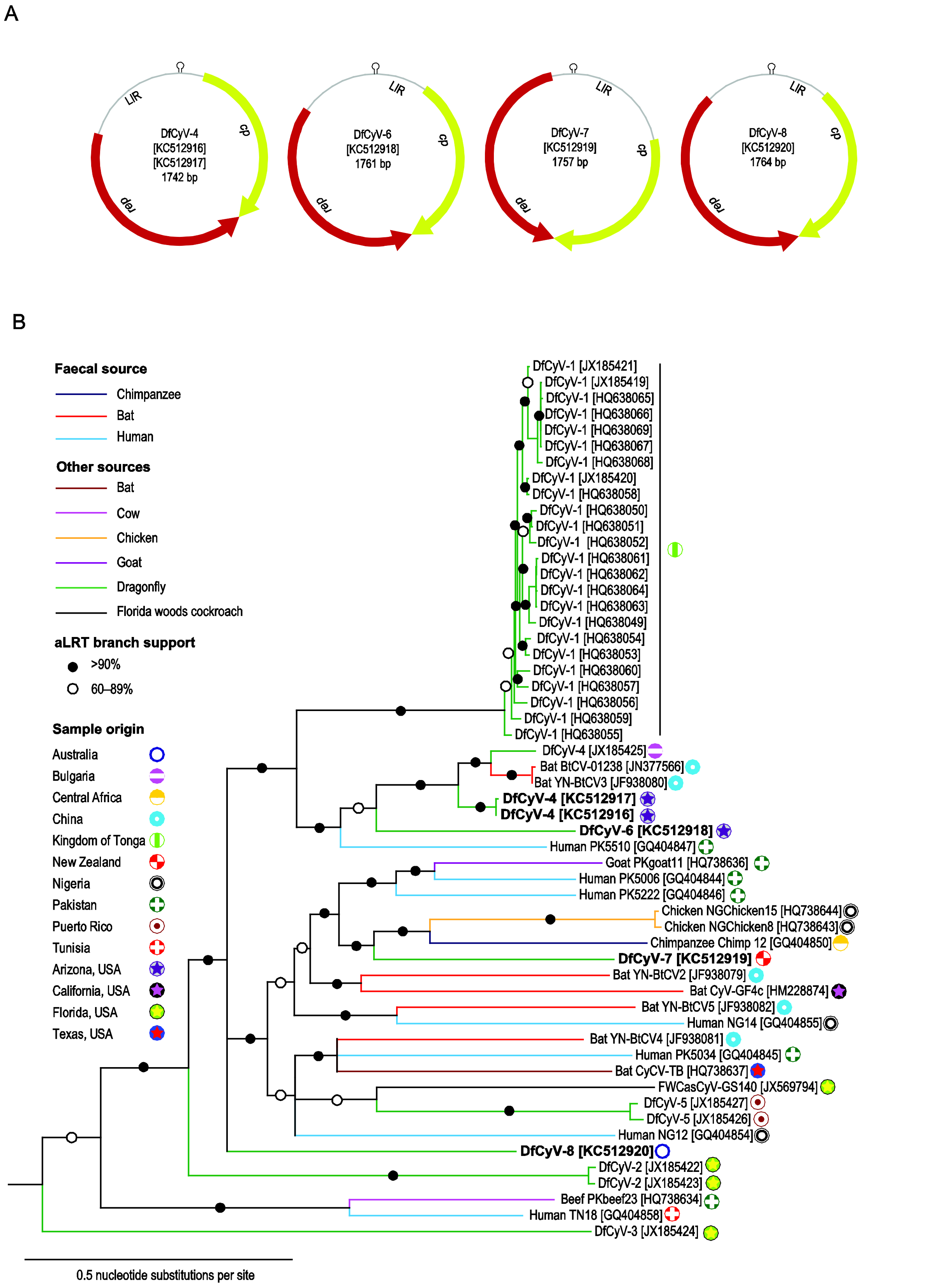


Figure 2.1: **A.** Genome organisation of the cycloviruses recovered from *A. multicolor* (DfCyV-4 and DfCyV-6), *O. sabina* (DfCyV-8) and *X. zealandica* (DfCyV-7). **B.** Maximum likelihood phylogenetic tree showing the relationships of the full genomes of all cyclovirus sequences available in GenBank plus those determined in this study.

2.4 Results and discussion

2.4.1 Viral genome analysis

Five cyclovirus genomes were recovered from three dragonfly species (Figure 2.1; Additional Table 2.2). Three genomes were recovered from *A. multicolor* (collected in USA; 1742-1761 nt), one from *X. zealandica* (New Zealand; 1757 nt) and one from *O. sabina* (Australia; 1764 nt). Other than these five viral genomes, we did not identify any other circular ssDNA viruses within these samples using methods described is this chapter. All of the genomic features identified in the genomes reported here including the sizes and genome organisations are characteristic of cycloviruses rather than circoviruses. Circoviruses are slightly larger (1759-2063 nt) than most cycloviruses (1723-1867 nt) (Li *et al.*, 2010). The Rep and CP are transcribed in opposite directions in both virus genera; however, in cycloviruses the putative origin of replication (ori), which exhibits a conserved nonanucleotide motif at the apex of a stem-loop structure, is not found on the Rep-encoding strand, suggesting that the virion strand of cycloviruses and circoviruses are different (Rosario *et al.*, 2012a).

Circoviruses and cycloviruses both have a long intergenic region (LIR) that contains the ori; however, cycloviruses lack a short intergenic region that occurs at the 3' ends of the Rep- and CP-encoding ORFs in circoviruses. Cycloviruses have two major ORFs are that are separated by a LIR (270-339 nt) that contains the putative ori with the conserved nonanucleotide motif (TAGTATTAC), and the CP-encoding ORF is present on the same strand as the putative ori (Figure 2.1). A BLASTn (Altschul *et al.*, 1990) analysis of the five full genomes showed that they are most closely related to cycloviruses isolated from either chickens, human faecal matter or dragonflies (Additional Table 2.2). The maximum likelihood (ML) phylogenetic trees revealed some level of clustering between closely related isolates and indicate that the previously identified dragonfly cyclovirus (DfCyV-3) is one of the most divergent cycloviruses reported to date (Figure 2.1).

2.4.2 Classification of cycloviruses

In an effort to establish an identity threshold for the classification of cyclovirus species, we investigated the distribution of genome-wide pairwise identities among genomes reported to date. A distribution of pairwise identities of all cycloviruses (1485 pairwise comparisons; Figure 2) indicates that most share between 56-75% and 95-100% pairwise identity. The latter is mainly attributed to DfCyV-1 isolates from the Kingdom of Tonga. A few pairwise identities are found between 86-88% attributed to a DfCyV-4 isolate from Bulgaria (JX185425), two bat cycloviruses from China (JF938080, JN377566) and two DfCyVs identified in this study from the USA (KC512916, KC512917). This indicates the global distribution of highly similar cycloviruses. The current circovirus species demarcation is based on pairwise identities including gaps as a 5th character state and if this was analysed using our methods with pairwise deletion of gaps, this cut off would come down to ~78%. For the purpose of this study, we have demarcated a species level cut-off at 76% genome-wide identity based on the distribution of pairwise identities calculated using SDTv1.0 (Muhire *et al.*, 2013) (Figure 2.2). This criteria is similar to the 75% genome-wide identity demarcation criteria for members of the *Circovirus* genus (Biagini *et al.*, 2012) and could be adopted by other studies for classifying all cyclovirus isolates.

Based on the cyclovirus classification scheme proposed here, we expanded the known geographical range of a previously reported cyclovirus species and identified three novel species. There are two cyclovirus genomes, which share 99.7% pairwise identity, isolated from dragonflies collected in Arizona, USA (Figure 2.2) that share 86% pairwise identity with a species isolated from a dragonfly from Bulgaria (DfCyV-4; JX185425) and, thus, are tentatively named DfCyV-4 [US-DFKWGX-2012; KC512916] and DfCyV-4 [US-DFKWGB-2012; KC512917]. Interestingly, DfCyV-4 seems to be the same species reported from faecal samples collected from insectivorous bats in China (Bat YN-BtCV3; JF938080 and Bat BtCV-01238; JN377566) since they share >85% genome-wide identity. The remaining three genomes reported in this study share <75% pairwise identity with all other cycloviruses, hence they have been tentatively named DfCyV-6 [US-DFKWGX-2012; KC512918], DfCyV-7 [NZ-DFNZ3-2011; KC512919] and DfCyV-8 [AU-DFB007B-2011; KC512920]. DfCyV-6 shares 72% pairwise identity with DfCyV-4 and ~60-72% pairwise identity to all other DfCyVs and cycloviruses. DfCyV-7 and DfCyV-8 share 59-68% to all other cycloviruses, including all the DfCyVs (Figure 2.2).

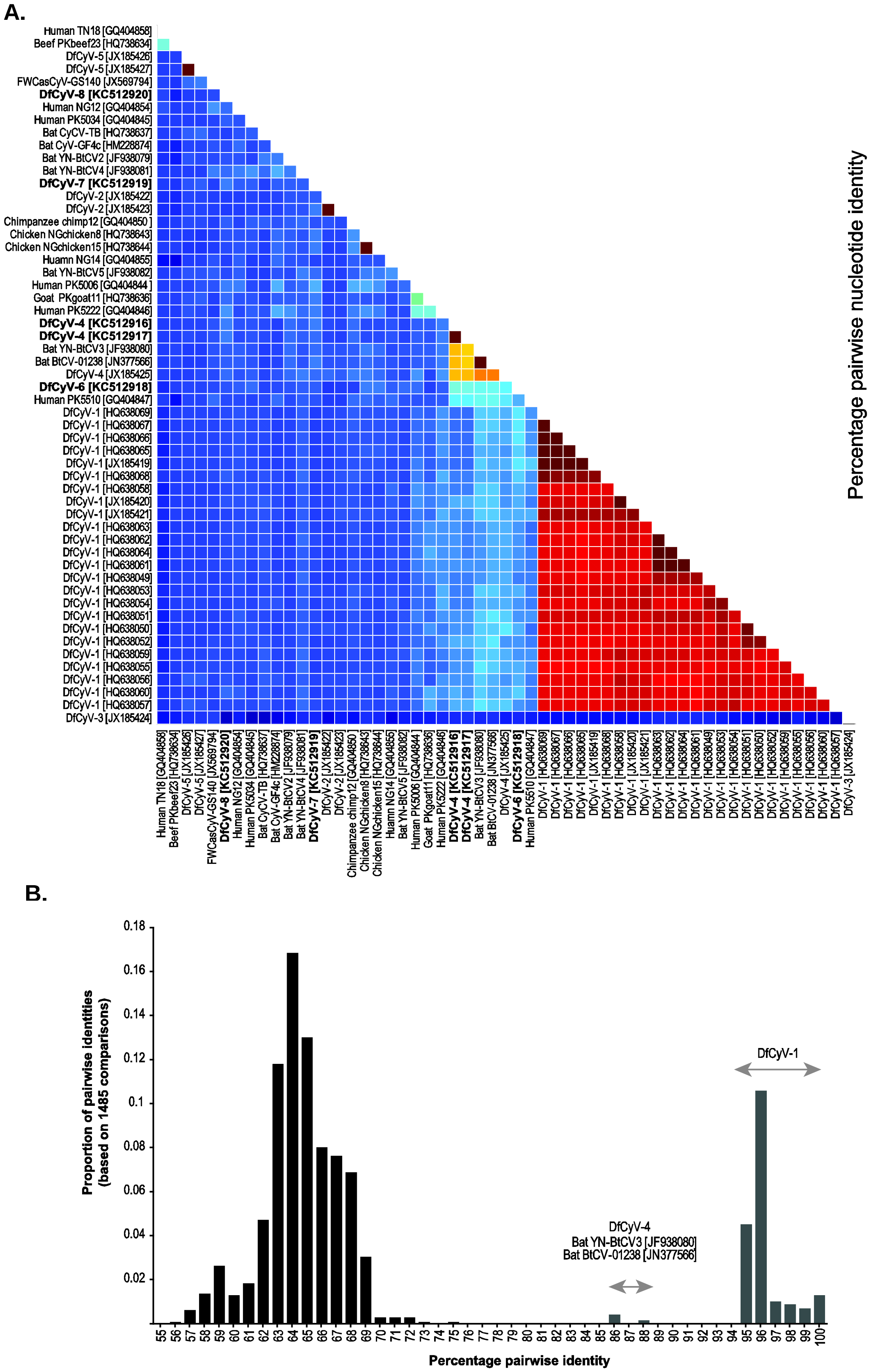
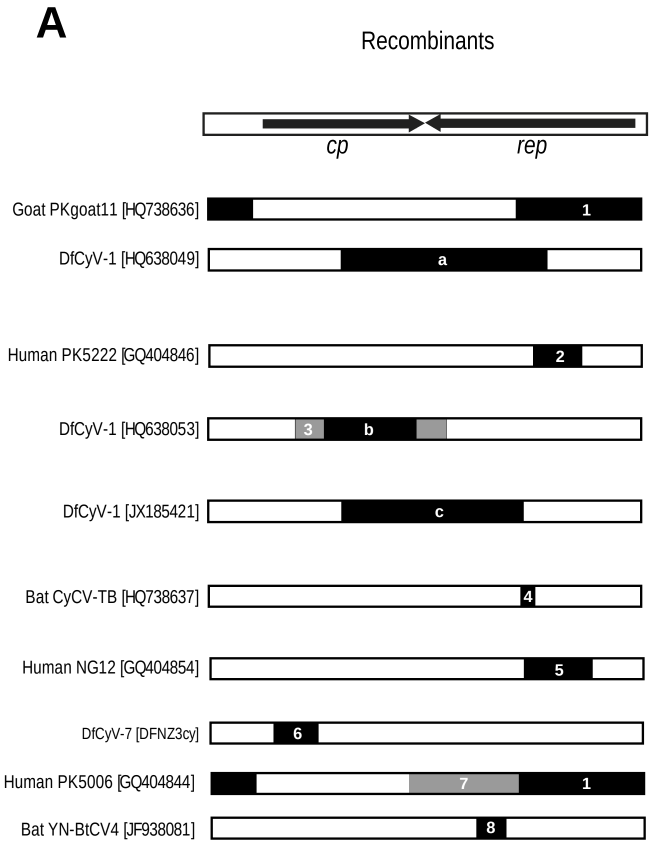


Figure 2.2: (**A**) Pairwise plot of genome-wide pairwise identities (%) of all the cycloviruses. (**B**) Distribution of pairwise identities (1485 comparisons) among cycloviruses.

A recent proposal submitted to the International Committee on Taxonomy of Viruses (ICTV) (assigned code: 2011.011a-bbbV) proposes the creation of two sub-families (cycloviruses and circoviruses) within the *Circoviridae* family, with each ach sub-family being further divided into genera, species and isolates. The proposal also suggests the use of CP amino acid sequences as a possibility for the classification for *Circoviridae*. It proposed that viral isolates that share between 40-75% identity could be classified as species and those that share 20-40% identity be members of different genera. We see some limitations with the use of CP sequences on their own, the primary one being the high sequence divergence found amongst CPs (making it difficult to generate robust credible alignments) and second being recombination. Therefore sequence analysis for taxonomic purposes of segments of genomes is not ideal for viral classification, especially amongst highly recombinant viruses such as most ssDNA viruses. Nonetheless, CP (amino acid identity) based classification of all the currently sequenced cycloviruses would yield 11 genera and 29 species in contrast to the 25 species we identify using full genome analysis. Without a doubt, as more sequence information becomes available the resolution of the analysis will improve and we should be able to establish a robust classification based on full genomes coupled with CP and Rep sequences.

2.4.3 Recombination analysis

Circular ssDNA viruses are known to have mechanistic predispositions for recombination and are notorious for high recombination rates (Cai *et al.*, 2012; Cheung, 2009; Julian *et al.*, 2013; Julian *et al.*, 2012; Lefeuvre *et al.*, 2009; Martin *et al.*, 2011; Massaro *et al.*, 2012; Mu *et al.*, 2012; Varsani *et al.*, 2011). We found evidence of three intra-species recombination events within cycloviruses, all within the highly sampled DfCyV-1 isolates from the Kingdom of Tonga, similar to those described by Rosario et al. (2011). We also found evidence of inter-species recombination in the remaining diverse cycloviruses (Figure 2.3). In most cases we were able to identify potential ancestral sequences from which the recombinant region was likely derived, and these results corroborate with our ML phylogenetic analysis of the Rep and CP (Figure 2.4). We also identified a highly recombinant viral genome, Human PK5006 (GQ404844), in which almost 65% of its genome is recombinant from two different events with recombinant regions derived from Goat PKgoat11 cyclovirus (HQ738636) and Human PK5222 (GQ404846) ancestral viral sequences. Therefore recombination seems to be a prevalent mechanism in cycloviruses.



B

Intra-species recombination

Event	Recombinant region	Potential major parent	Potential minor parent	Detection method	P-value
a	512-1415	DfCyV-1 [HQ638063], [HQ638061], [HQ638062], [HQ638064]	DfCyV-1 [HQ638054], [HQ638053]	GBMCS**T**	$4.52x10^{-14}$
b	403-772	DfCyV-1 [HQ638049], [HQ638054]	DfCyV-1 [HQ638051], [HQ638052], [HQ638050]	GBMCS**T**	$3.25x10^{-12}$
c	527-1296	DfCyV-1 [JX185420], [HQ638058]	DfCyV-1 [HQ638069], [HQ638067], [HQ638065], [HQ638068], [HQ638066], [JX185419]	GBMCS**T**	$6.50x10^{-12}$

Inter-species recombination

Event	Recombinant region	Potential major ancestral parent	Potential minor ancestral parent	Detection method	P-value
1	1387-216	Unknown	Human PK5222 [GQ404846]	RG**B**MCS	$5.24x10^{-14}$
2	1474-1586	Unknown	Chicken NGChicken15 [HQ738644], Chicken NGChicken8 [HQ738643], Chimpanzee Chimp12 [GQ404850]	R**G**BMS	$4.81x10^{-8}$
3	274-789	Unknown	DfCyV-1 [HQ638051], [HQ638050]	GBM**S**T	$1.16x10^{-6}$
4	1376-1482	Bat YN-BtCV4 [JF938081] , Bat CyV-GF4c [HM228874]	Unknown	**R**GBMCS	$1.03x10^{-6}$
5	1301-1548	DfCyV-5 [JX185426], [JX185426]	DfCyV-7 [KC512919], Chicken NGchicken8 [HQ738643], Chicken NGChicken15 [HQ738644]	M**C**S	$5.62x10^{-5}$
6	323-533	Unknown	DfCyV-5 [JX185426], [JX185426]	**R**MC	$7.42x10^{-4}$
7	882-1358†	Unknown	Goat PKgoat11 [HQ738636]	RBMC**S**	$8.68x10^{-9}$
8	1101-1306†	Chicken NGChicken8 [HQ738643]	Unknown	R**B**S	$6.21x10^{-4}$

*RDP (R) GENCONV (G), BOOTSCAN (B), MAXCHI (M), CHIMERA (C), SISCAN (S) and 3SEQ (T)
†= The actual breakpoint position is undetermined.

Figure 2.3: Recombination analysis results showing a schematic of putative recombinant regions (**A**) and details of recombination events detected in cyclovirus genomes using the RDP, GENECONV, BOOTSCAN, MAXCHI, CHIMAERA, SISCAN and 3SEQ methods implemented in the recombination detection software, RDP4 (**B**). Only detection methods with associated p-values <0.05 are shown and the given p-value is for the detection method highlighted in bold.

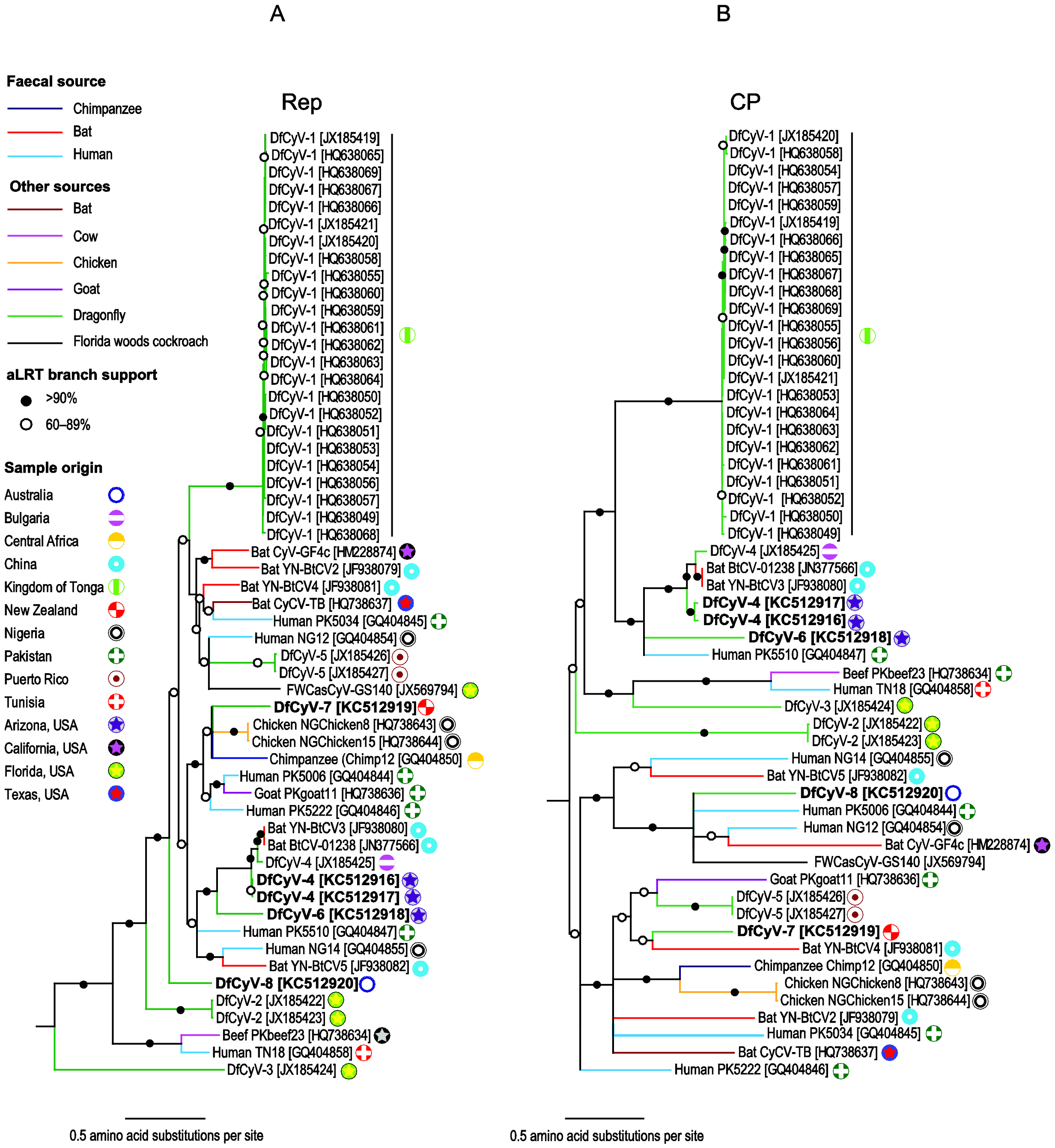


Figure 2.4: Maximum likelihood phylogenetic trees showing the relationships between cyclovirus replication-associated protein (Rep) (**A**) and capsid protein (CP) (**B**) amino acid sequences.

2.4.4 Replication-associated protein and capsid protein analysis

Pairwise amino acid identities were calculated using MUSCLE (Edgar, 2004), and a summary of these are represented in a two-dimensional colour plot in Figure 2.7. It clear from the ML phylogenetic tree (Figure 2.4) and pairwise distance analysis of the Rep and CP amino acid sequences that there is greater diversity within the CP compared to the Rep. For example, with the exception of DfCyV-3 all the cyclovirus Reps share >48% pairwise identity whereas in the case of the CP it is >29%. The Reps of DfCyV-4 and the two related Bat cycloviruses share >93% pairwise identity whereas the CP share >84% pairwise identity. The Reps of DfCyV-4, -5, -6 and -7 share >52%, 72%, 50% and 54% pairwise identity respectively with all other cycloviruses other than with DfCyV-3 (they share ~39%, 47%, 45% and 44% pairwise identity). The CPs of DfCyV-4 and -5 share ~54% pairwise identity, whereas the rest of the DfCyV strains share pairwise identity of ~34-40%. Interestingly, the CP of DfCyV-7 shares 53% pairwise identity with cycloviruses isolated from human faecal samples from Pakistan and Nigeria (PK5006, GQ404844; NG12, GQ404854). BLASTp analysis revealed similar results, and these are summarised in Additional Table 2.3.

Several conserved motifs present in the majority of eukaryotic circular ssDNA virus Reps (reviewed by (Rosario *et al.*, 2012a)) were identified within the putative Reps of all five cycloviruses reported here, including rolling circle replication (RCR) motifs I, II, and III, as well as superfamily 3 (SF3) helicase motifs (Table 2.1). The function of RCR motif I (FTxNN for cycloviruses) is still unknown, however it may be involved in sequence specific recognition of iterons in the LIR (Argüello-Astorga & Ruiz-Medrano, 2001). Details of the iteron sequences found in the LIR are provided in Table 2.1. RCR motif II (HLQGxxNL for cycloviruses) is thought to coordinate the binding of metal ions (Ilyina & Koonin, 1992). RCR motif III (YCSKxGx for cycloviruses) is believed to be a catalytic site where DNA cleavage takes place during replication initiation (Heyraud-Nitschke *et al.*, 1995; Laufs *et al.*, 1995). SF3 helicase motifs are characterised by a highly conserved domain that contains three

Table 2.1: Details of conserved motifs in the replication-associated protein (Rep) amino acid sequences and the iteron nucleic acid sequences found in the cyclovirus genomes. Number of iterons with similar sequences are listed in brackets.

Isolate *	Country	Insect species	RCR Motifs			SF3 Helicase Motifs			Iterons †	
			I	II	III	WalkerA	WalkerB	Motif C		
DfCyV-1 [JX185419]	Kingdom of Tonga	*Pantala flavescens*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [JX185420]	Kingdom of Tonga	*Pantala flavescens*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [JX185421]	Kingdom of Tonga	*Pantala flavescens*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [HQ638065]	Kingdom of Tonga	*Pantala flavescens*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [HQ638066]	Kingdom of Tonga	*Pantala flavescens*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [HQ638067]	Kingdom of Tonga	*Pantala flavescens*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [HQ638068]	Kingdom of Tonga	*Diplacodes bipunctata*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [HQ638069]	Kingdom of Tonga	*Tholymis tillarga*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [HQ638061]	Kingdom of Tonga	*Diplacodes bipunctata*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [HQ638062]	Kingdom of Tonga	*Diplacodes bipunctata*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [HQ638063]	Kingdom of Tonga	*Diplacodes bipunctata*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VNIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [HQ638064]	Kingdom of Tonga	*Diplacodes bipunctata*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [HQ638060]	Kingdom of Tonga	*Diplacodes bipunctata*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [HQ638059]	Kingdom of Tonga	*Pantala flavescens*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [HQ638058]	Kingdom of Tonga	*Pantala flavescens*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [HQ638057]	Kingdom of Tonga	*Tholymis tillarga*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1[HQ638056]	Kingdom of Tonga	*Pantala flavescens*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [HQ638055]	Kingdom of Tonga	*Pantala flavescens*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [HQ638053]	Kingdom of Tonga	*Diplacodes bipunctata*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [HQ638054]	Kingdom of Tonga	*Diplacodes bipunctata*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [HQ638052]	Kingdom of Tonga	*Pantala flavescens*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [HQ638050	Kingdom of Tonga	*Tholymis tillarga*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [HQ638051]	Kingdom of Tonga	*Tholymis tillarga*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-1 [HQ638049]	Kingdom of Tonga	*Tholymis tillarga*	FTWNN	HIQGFCNL	YCSKSGE	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
DfCyV-2 [JX185422]	USA	*Pantala flavescens*	FTVNN	HLQGFANL	YCSKSGE	GPPGSGKS	VIIDDFYGW	ITSN	CGTAAC (3)	CGTAGC (1)
DfCyV-2 [JX185423]	USA	*Anax junius*	FTVNN	HLQGFANL	YCSKSGE	GPPGSGKS	VIIDDFYGW	ITSN	CGTAAC (3)	CGTAGC (1)
DfCyV-3 [JX185424]	USA	*Erythemis simplicicollis*	FTLNN	HLQGFINF	YCTKGGD	GDPGAGKS	VIVDDYYGW	----	CGAGCCC (1)	CGGCCCC(1)
DfCyV-4 [KC512917]	USA	*Aeshna multicolour*	FTWNN	HLQGYANL	YCSKAGE	GLPGTGKS	VIIDDFYGW	ITTN	CGTAGC (3)	
DfCyV-4 [KC512916]	USA	*Aeshna multicolour*	FTWNN	HLQGYANL	YCSKAGE	GLPGTGKS	VIIDDFYGW	ITTN	CGTAGC (3)	
DfCyV-4 [JX185425]	Bulgaria	*Somatochlora meridionalis*	FTWNN	HLQGYANL	YCSKAGE	GLPGTGKS	VIIDDFYGW	ITTN	CGTAGC (3)	
DfCyV-5 [JX185426]	Puerto Rico	*Erythrodiplax umbrata*	FTLNN	HLPGFCNL	YCRKSGT	GPTGSGKS	VIIDDFYGW	ITSN	CGTAAC (2)	
DfCyV-5 [JX185427]	Puerto Rico	*Erythrodiplax umbrata*	FTLNN	HLQGFCNL	YCRKSGT	GPTGSGKS	VIIDDFYGW	ITSN	CGTAAC (2)	
DfCyV-6 [KC512918]	USA	*Aeshna multicolour*	FTLNN	HLQGFCNL	YCSKAGK	GEPGTGKS	VIIDDFYGW	ITSN	CGTAGC (4)	TGTAAC (1)
DfCyV-7 [KC512919]	New Zealand	*Xanthocnemis zealandica*	FTWNN	HLQGFCNL	YCSKSGI	GAPGTGKS	VIIDDFYGW	ITSN	CGTCCAC (3)	
DfCyV-8 [KC512920]	Australia	*Orthetrum Sabina*	FTWNN	HLQGYCNL	YCSKSGE	GPPGSGKS	VIIDDFYGW	ITSN	CGTAGC (2)	CGTAAC (1)
FWCasCyV-GS140 [JX569794]	USA	*Eurycotis floridana*	FTLNN	HLQGFCNL	-------	GPTGTGKS	VIIDDFYGW	ITSN	CGGTACA (2)	
Bat BtCV-01238 [JN377566]	China	-	FTWNN	HLQGYANL	YCSKAGE	GLPGTGKS	VIIDDFYGW	ITSN	CGTAGC (2)	CGTAAC (1)
Bat CyCV-TB [HQ738637]	USA	-	FTWNN	HLQGFCNL	YCKKGNK	GPPGTGKS	VIIDDFYGW	FTSN	CGTATC (1)	CGTAAC (1)
Bat CyV-GF4c [HM228874]	USA	-	FTWNN	HLQGFCNL	YCSKAGD	GEPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
Bat YN-BtCV3 [JF938080]	China	-	FTWNN	HLQGYANL	YCSKAGE	GLPGTGKS	VIIDDFYGW	ITSN	CGTAGC (2)	CGTAAC (1)
Bat YN-BtCV2 [JF938079]	China	-	FTLNN	HLQGFCNL	YCSKAGN	GPPGSGKS	VIIDDFYGW	ITSN	CGTAAC (2)	
Bat YN-BtCV4 [JF938081]	China	-	FTWNN	HLQGFCNL	YCKKSGD	GPPGSGKS	VIVDDFYGW	FTSN	CGTAAC (3)	

Isolate *	Country	Insect species	RCR Motifs			SF3 Helicase Motifs			Iterons †	
			I	II	III	WalkerA	WalkerB	Motif C		
Bat YN-BtCV5 [JF938082]	China	-	FTWNN	HLQGFCNL	YCSKSGD	GPPGTGKS	VVIDDFYGW	ITSN	CGTAAC (3)	
Beef PKbeef23 [HQ738634]	Pakistan	-	WTLNN	HLQGFCNL	YCSKSGE	GPTGAGKS	VIFDDFYGW	FTSN	CGTAAC (2)	
Chicken NGchicken15 [HQ738644]	Nigeria	-	FTWNN	HLQGFCNL	YCSKSGI	GPPGTGKS	VIIDDFYGW	ITSN	CGTAACC (3)	
Chicken NGchicken8 [HQ738643]	Nigeria	-	FTWNN	HLQGFCNL	YCSKSGI	GPPGTGKS	VIIDDFYGW	ITSN	CGTAACC (3)	
Chimpanzee chimp12 [GQ404850]	Central Africa‡	-	FTWNN	HLQGYVNL	YCRKSGI	GPPGSGKS	VIIDDFYGW	FTSN	CGTAACC (2)	
Goat PKgoat11 [HQ738636]	Pakistan	-	FTWNN	HLQGFCNL	YCSKSGI	GPPGSGKS	VIIDDFYGW	ITSN	CGTAACC (2)	
Human NG14 [GQ404855]	Nigeria	-	FTWNN	HLQGFCNL	YCSKTGN	GPPGSGKS	VIIDDFYGW	ITSN	CGTAGC (3)	
Human NG12 [GQ404854]	Nigeria	-	FTLNN	HLQGFCNL	YCSKAGN	GPTGSGKS	VIIDDFYGW	ITSE	CGTAAC (2)	
Human PK5006 [GQ404844]	Pakistan	-	FTWND	HLQGFCNL	YCSKSGI	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (3)	
Human PK5034 [GQ404845]	Pakistan	-	FTWNN	HLQGFCNL	YCKKAGH	GPPGSGKS	VIIDDFYGW	FTSN	CGTAAC (2)	
Human PK5222 [GQ404846]	Pakistan	-	FTWNN	HLQGFCNL	YCSKTGI	GPPGTGKS	VIIDDFYGW	ITSN	CGTAAC (2)	
Human PK5510 [GQ404847]	Pakistan	-	FTWNN	HLQGFCNL	YCSKSGE	GPPGSGKS	VIIDDFYGW	ITSN	YGTAGC (3)	
Human TN18 [GQ404858]	Tunisia	-	WTLNN	HLQGFCNL	YCSKSGE	GPTGSGKS	VIIDDFYGW	ITSE	CGTAAC (3)	CGTAGC (1)

*GenBank accession numbers are shown in square brackets.
†Number of iterons with similar sequences are listed in parentheses.
‡Specimens were collected from Tanzania, Cameroon, Uganda, Rwanda, Central African Republic, Republic of the Congo and the Democratic Republic of the Congo.

motifs within the ~100 amino acid region, reviewed by Rosario et al. (2012a). The SF3 helicase Walker-A motif (GxxGTGKS for cycloviruses) may act as a dNTP binding domain, although it is also thought that this motif may be involved in helicase-like activity in the Rep during RCR. Walker-B and motif C are believed to be involved in controlling helicase activity through the dNTP binding and P-loop NTPase domains.

2.4.5 Sequence analysis of the Large Intergenic Region (LIR)

Analysis of the intergenic region of the five cycloviruses sequenced in this study revealed the presence of short iterative sequences (iterons) located close to the stem-loop element containing the conserved nonanucleotide TAGTATTAC (See Table 2.1 for details of iteron sequences). The iterons are actual or putative Rep-binding sites identified in several groups of circular ssDNA viruses (Arguello-Astorga & Ruiz-Medrano, 2001; Arguello-Astorga *et al.*, 1994; Gutierrez, 1999; Hanley-Bowdoin *et al.*, 2000; Steinfeldt *et al.*, 2001; Timchenko *et al.*, 2000). The DfCyV-7 iterons exhibited a nucleotide core sequence (CGTCCCAC) different from those of the homologous elements in DfCyV-4, DfCyV-6 and DfCV-8, all of which showed iterons with a CGTARC core sequence. In addition, the DfCyV-7 iterative elements exhibited a distinctive arrangement, with one inverted repeat at the 5´ border of the stem-loop element and two direct repeats partially overlapping the right stem arm. In contrast, DfCyV-4, DfCyV-6 and DfCV-8 displayed three direct repeats, although with a different arrangement in all cases (Figure 2.5). Subsequent examination of the intergenic region of all other cycloviruses showed that iterons displaying CGTAAC and/or CGTAGC core sequences are predominant in the proposed Cyclovirus genus (see Table 2.1). Indeed, only two cycloviruses besides DfCyV-7 exhibited iterons with a distinctive sequence, namely, DfCyV-3 and FWCasCyV-GS140, displaying CGRcCCC and CGGTACA core sequences, respectively (Figure 2.5). Given that a recent theoretical analysis of Rep proteins encoded by diverse groups of circular ssDNA viruses mapped the high-affinity DNA binding specificity determinants (SPDs) to two discrete amino acid clusters surrounding the RCR motifs I and II (Londoño *et al.*, 2010), we decided to analyze the cyclovirus Rep proteins to determine whether there is also a correlation between the iteron sequence and the putative SPDs previously identified in other ssDNA viruses. This new analysis revealed that cycloviruses with iterons exhibiting a CGTARC core sequence encode Rep proteins displaying a SPD-region (r)1 (SPD-r1) with a TxR sequence and a SPD-r2 with a PxR motif

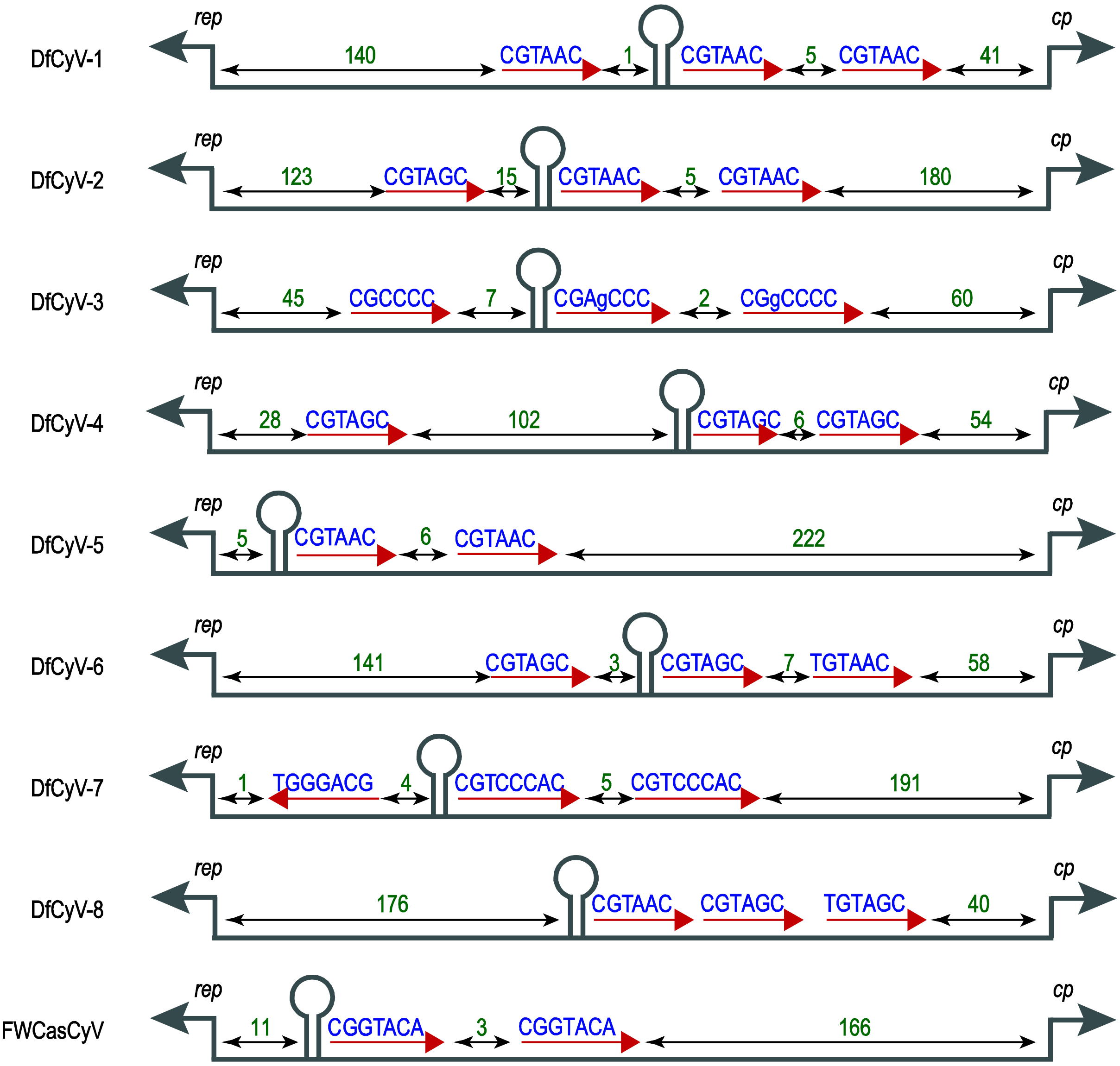
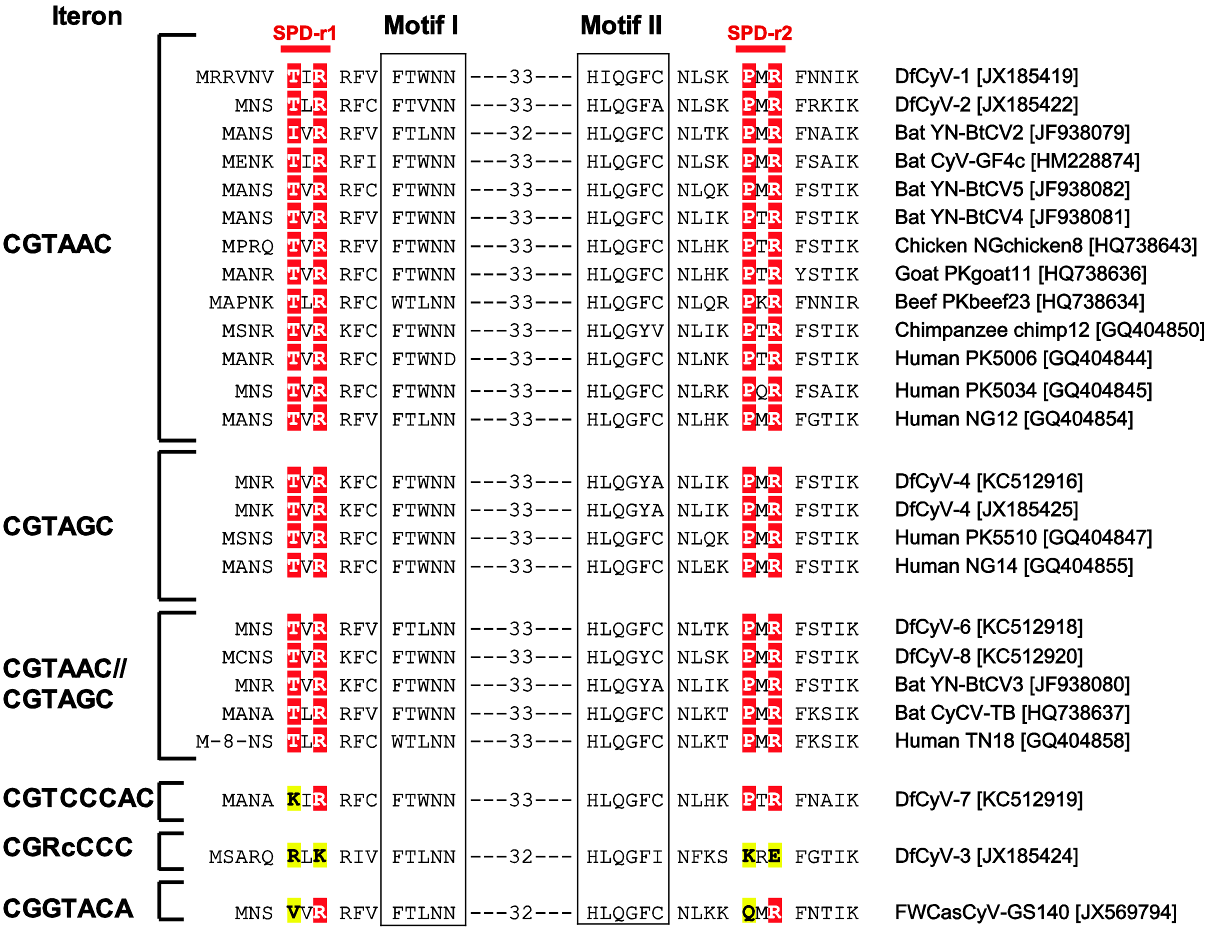


Figure 2.5: Nucleotide sequence and organisation of origin of replication-associated iterative sequences of cycloviruses isolated from insects. Red arrows show the orientation of the iterons with respect to the stem-loop element. Numbers in green denote the nucleotides spanned between a specific iteron and the start codon of the ORF at one end. Lowercase letters in an iterated element indicate a nucleotide that does not match in all the iterons of a virus.

(Figure 2.6). The only exception to this rule was the DfCyV-5 Rep, exhibiting SPDs with **V**xR and **G**xR sequences, respectively (not shown). The concurrent changes in both the SPD-r1 and SPD-r2 elements of DfCyV-5 Rep suggest the existence of different SPD combinations determining similar high-affinity for CGTARC iterons. Interestingly, the Rep proteins encoded by the three cycloviruses (i.e. DfCyV-3, DfCyV-7 and FWCasCyV) possessing distinctive iterons have SPDs that are also unique (Figure 2.6), hence supporting the predicted correlation between Rep SPDs and cognate iterons in ssDNA viruses that replicate by a RCR mechanism.



Iteron		SPD-r1		Motif I		Motif II		SPD-r2		
CGTAAC	MRRVNV	TIR	RFV	FTWNN	---33---	HIQGFC	NLSK	PMR	FNNIK	DfCyV-1 [JX185419]
	MNS	TLR	RFC	FTVNN	---33---	HLQGFA	NLSK	PMR	FRKIK	DfCyV-2 [JX185422]
	MANS	IVR	RFV	FTLNN	---32---	HLQGFC	NLTK	PMR	FNAIK	Bat YN-BtCV2 [JF938079]
	MENK	TIR	RFI	FTWNN	---33---	HLQGFC	NLSK	PMR	FSAIK	Bat CyV-GF4c [HM228874]
	MANS	TVR	RFC	FTWNN	---33---	HLQGFC	NLQK	PMR	FSTIK	Bat YN-BtCV5 [JF938082]
	MANS	TVR	RFV	FTWNN	---33---	HLQGFC	NLIK	PTR	FSTIK	Bat YN-BtCV4 [JF938081]
	MPRQ	TVR	RFV	FTWNN	---33---	HLQGFC	NLHK	PTR	FSTIK	Chicken NGchicken8 [HQ738643]
	MANR	TVR	RFC	FTWNN	---33---	HLQGFC	NLHK	PTR	YSTIK	Goat PKgoat11 [HQ738636]
	MAPNK	TLR	RFC	WTLNN	---33---	HLQGFC	NLQR	PKR	FNNIR	Beef PKbeef23 [HQ738634]
	MSNR	TVR	KFC	FTWNN	---33---	HLQGYV	NLIK	PTR	FSTIK	Chimpanzee chimp12 [GQ404850]
	MANR	TVR	RFC	FTWND	---33---	HLQGFC	NLNK	PTR	FSTIK	Human PK5006 [GQ404844]
	MNS	TVR	RFC	FTWNN	---33---	HLQGFC	NLRK	PQR	FSAIK	Human PK5034 [GQ404845]
	MANS	TVR	RFV	FTLNN	---33---	HLQGFC	NLHK	PMR	FGTIK	Human NG12 [GQ404854]
CGTAGC	MNR	TVR	KFC	FTWNN	---33---	HLQGYA	NLIK	PMR	FSTIK	DfCyV-4 [KC512916]
	MNK	TVR	KFC	FTWNN	---33---	HLQGYA	NLIK	PMR	FSTIK	DfCyV-4 [JX185425]
	MSNS	TVR	RFC	FTWNN	---33---	HLQGFC	NLQK	PMR	FSTIK	Human PK5510 [GQ404847]
	MANS	TVR	RFC	FTWNN	---33---	HLQGFC	NLEK	PMR	FSTIK	Human NG14 [GQ404855]
CGTAAC// CGTAGC	MNS	TVR	RFV	FTLNN	---33---	HLQGFC	NLTK	PMR	FSTIK	DfCyV-6 [KC512918]
	MCNS	TVR	KFC	FTWNN	---33---	HLQGYC	NLSK	PMR	FSTIK	DfCyV-8 [KC512920]
	MNR	TVR	KFC	FTWNN	---33---	HLQGYA	NLIK	PMR	FSTIK	Bat YN-BtCV3 [JF938080]
	MANA	TLR	RFV	FTWNN	---33---	HLQGFC	NLKT	PMR	FKSIK	Bat CyCV-TB [HQ738637]
	M-8-NS	TLR	RFC	WTLNN	---33---	HLQGFC	NLKT	PMR	FKSIK	Human TN18 [GQ404858]
CGTCCCAC	MANA	KIR	RFC	FTWNN	---33---	HLQGFC	NLHK	PTR	FNAIK	DfCyV-7 [KC512919]
CGRcCCC	MSARQ	RLK	RIV	FTLNN	---32---	HLQGFI	NFKS	KRE	FGTIK	DfCyV-3 [JX185424]
CGGTACA	MNS	VVR	RFV	FTLNN	---32---	HLQGFC	NLKK	QMR	FNTIK	FWCasCyV-GS140 [JX569794]

Figure 2.6: Summary of potential DNA-binding SPDs of cyclovirus Rep proteins. Amino acid residues identified as putative SPDs are shaded. These residues cluster into two discrete regions labeled as SPD region 1 (SPD-r1) and SPD region 2 (SPD-r2). The conserved RCR motifs II and II are indicated at the top of the alignments.

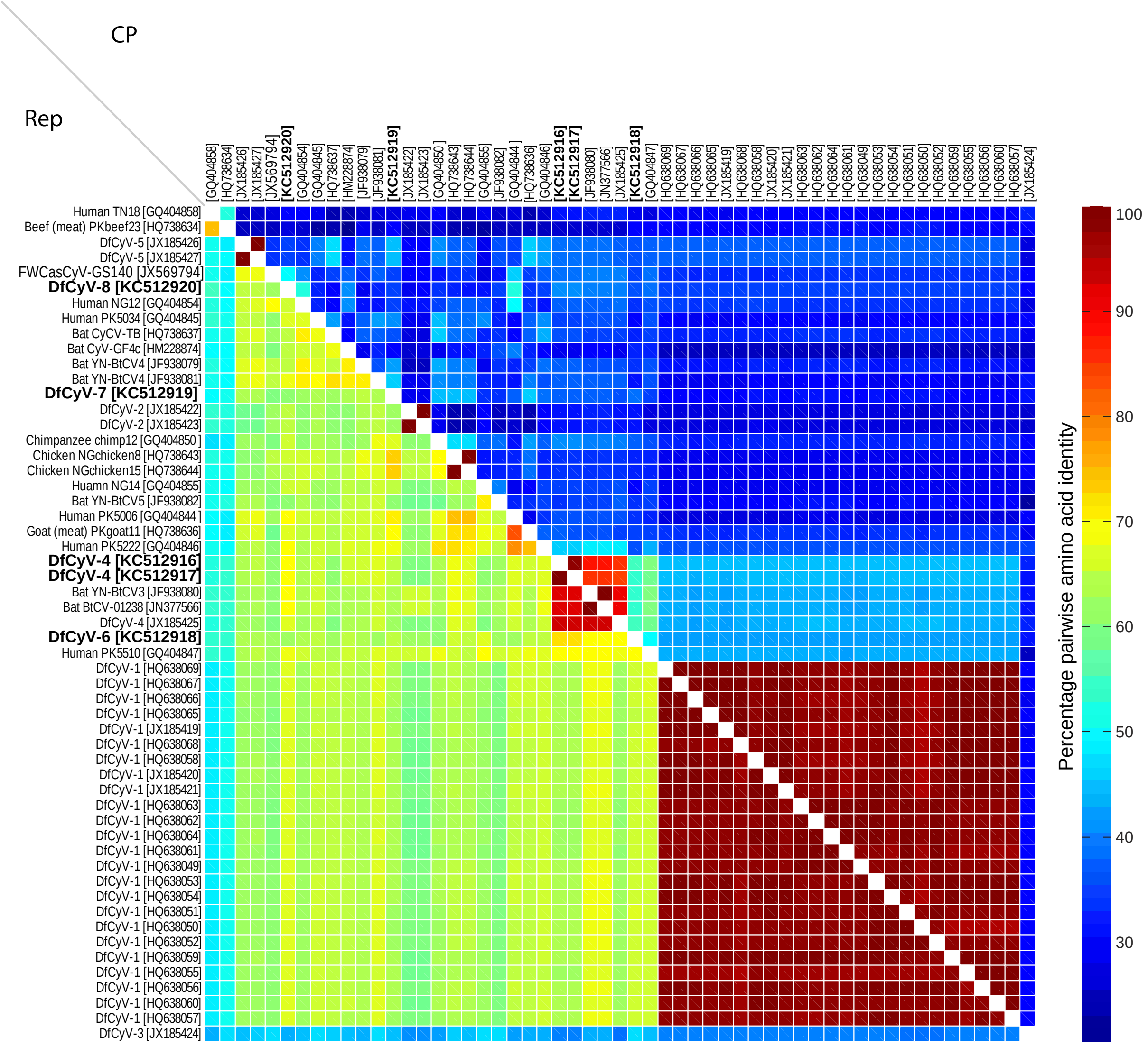


Figure 2.7: Two-dimensional plot of pairwise identities of the cyclovirus replication-associated protein (Rep) and capsid protein (CP) amino acid sequences. isolates from this study shown in bold.

2.5 Concluding remarks

Dragonflies have been shown to harbour a diversity of novel circular ssDNA viruses, in particular cycloviruses (Rosario *et al.*, 2012b). Including the genomes described in this study, eight different cyclovirus species have been discovered in dragonflies. Some of these cycloviruses have a widespread geographical distribution, with the broadest distribution observed for DfCyV-4, which has been identified in Bulgaria and the USA. Therefore cycloviruses clearly circulate widely in winged insect populations. Overall, cycloviruses have been identified in a variety of sample types and organisms, including faeces (bats, humans and chimpanzees), muscle tissues (bats), meat products (cows, chicken, camel, sheep and goats) and insect abdomens (dragonflies and Florida woods cockroach). This contrasts with members of their sister group under the *Circoviridae* family, the *Circovirus* genus, which have only been associated with vertebrates. Although cycloviruses may have a broader host range than circoviruses, we do yet know whether these cycloviruses cause any disease or the definitive host they infect. In order to understand whether these cycloviruses are associated with the insects throughout their life, we are currently investigating the presence of cycloviruses in dragonfly larvae and the water bodies within which these larvae exist (Chapter 3).

Despite the similarities in their Rep, cycloviruses have distinct genomic characteristics from circoviruses. Since currently there are no species demarcation criteria for the Cyclovirus genus, we took advantage of the diversity of cycloviruses that have been recently reported from dragonflies and other sources to propose a classification scheme. Based on our analyses, a species demarcation cut-off of 76% genome-wide pairwise identity should be used. The characterisation of novel insect cycloviruses, whether by sampling viruses present in dragonflies or by directly sampling a variety of insects, will help us better understand evolutionary links within the *Circoviridae* family as well as improve taxonomic classification criteria. Although the cycloviruses recovered from the insect material in this chapter, it remains to be further investigated if this is a result of the diet of the insects or if the viruses are infecting the dragonflies directly.

GenBank accession numbers: KC512916 - KC512920

Additional Table 2.1: Methods used to verify cyclovirus genome sequences reported in this study, including restriction enzyme analysis and PCR

Viral isolate	Initial restriction enzyme	Primers	Second restriction enzyme
DfCyV-8 [KC512920]	*Bam*HI	-	*Eco*RI
DfCyV-4 [KC512917]	*Bam*HI	-	*Xmn*I
DfCyV-4 [KC512916]	*Xmn*I	-	*Bam*HI
DfCyV-7 [KC512919]	*Xmn*I	F: 5’-GTATGGGTACCGGTCCATTA-3’	-
		R: 5’-AAAGTACAGGTGAAAGGAGGG-3’	
DfCyV-6 [KC512918]	*Xmn*I	F: 5’-GTCATATTTTATCCATCCGTAGAAGTC-3’	-
		R: 5’-GAAATGCTCAAAATTATGGACCGC-3'	

Additional Table 2.2: Summary of BLAST analysis results for cyclovirus sequences reported in this study

FULL GENOME BLAST

BLAST Sequence	Hit description	Hit cccession	Max score	Total score	Query coverage	E value	Max ident
DfCyV-4 [KC512916]	replication-associated protein [Dragonfly cyclovirus 4]	AFS65283	607	607	52%	0	93%
DfCyV-4 [KC512917]	replication-associated protein [Dragonfly cyclovirus 4]	AFS65283	607	607	52%	0	93%
DfCyV-6 [KC512918]	replication-associated protein [Dragonfly cyclovirus 4]	AFS65283	448	448	50%	3.00E-151	71%
DfCyV-7 [KC512919]	Rep protein [Cyclovirus NGchicken8/NGA/2009]	ADU77011	434	434	47%	4.00E-146	73%
DfCyV-8 [KC512920]	replication-association protein [Cyclovirus PK5222]	ADD62455	414	414	47%	2.00E-138	70%

REP ORF tBLASTx

Sequence	Hit description	Hit cccession	Max score	Total score	Query coverage	E value	Max ident	AA length	Nucleotide length
DfCyV-4 [KC512916]	replication-associated protein [Dragonfly cyclovirus 4]	AFS65283	606	606	98%	0	93%	310	930
DfCyV-4 [KC512917]	replication-associated protein [Dragonfly cyclovirus 4]	AFS65283	606	606	98%	0	93%	310	930
DfCyV-6 [KC512918]	replication-associated protein [Bat circovirus ZS/Yunnan-China/2009]	AEL28798	409	409	99%	3.00E-141	70%	276	828
DfCyV-7 [KC512919]	Rep protein [Cyclovirus NGchicken8/NGA/2009]	ADU77011	434	434	98%	7.00E-151	73%	283	849
DfCyV-8 [KC512920]	replication-association protein [Cyclovirus PK5222]	ADD62455	414	414	100%	7.00E-143	70%	278	834

CP ORF tBLASTx

Sequence	Hit description	Hit cccession	Max score	Total score	Query coverage	E value	Max ident	AA length	Nucleotide length
DfCyV-4 [KC512916]	putative capsid protein [Bat circovirus ZS/China/2011]	AEL87789	354	354	83%	4.00E-121	88%	223	669
DfCyV-4 [KC512917]	putative capsid protein [Dragonfly cyclovirus 4]	AFS65282	343	343	83%	8.00E-117	84%	223	669
DfCyV-6 [KC512918]	putative capsid protein [Dragonfly cyclovirus 4]	AFS65282	216	216	83%	6.00E-67	54%	224	672
DfCyV-7 [KC512919]	Cap protein [Cyclovirus PKgoat11/PAK/2009]	YP_004152330	173	173	84%	4.00E-50	47%	219	657
DfCyV-8 [KC512920]	capsid protein [Cyclovirus PK5006]	ADD62452	205	205	83%	9.00E-63	54%	220	660

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.

Anisimova, M. & Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic biology* **55**, 539-552.

Argüello-Astorga, G. & Ruiz-Medrano, R. (2001). 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. *Archives of virology* **146**, 1465-1485.

Arguello-Astorga, G. R. & Ruiz-Medrano, R. (2001). 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. *Archives of Virology* **146**, 1465-1485.

Arguello-Astorga, G. R., Guevara-Gonzalez, R. G., Herrera-Estrella, L. R. & Rivera-Bustamante, R. F. (1994). Geminivirus replication origins have a group-specific organization of iterative elements: a model for replication. *Virology* **203**, 90-100.

Biagini, P., Bendinelli, M., Hino, S., Kakkola, L., Mankertz, A., Niel, C., Okamoto, H., Raidal, S., Teo, C. G. & other authors (2012). Circoviridae. In *Virus Taxonomy: Ninth report of the International Committee on Taxonomy of Viruses*, pp. 343-349. Edited by A. M. Q. King, M. J. Adams, E. B. Carstens & E. J. Lefkowitz: Elsevier Academic Press.

Blinkova, O., Victoria, J., Li, Y., Keele, B. F., Sanz, C., Ndjango, J. B. N., Peeters, M., Travis, D., Lonsdorf, E. V. & other authors (2010). Novel circular DNA viruses in stool samples of wild-living chimpanzees. *Journal of General Virology* **91**, 74-86.

Boni, M. F., Posada, D. & Feldman, M. W. (2007). An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* **176**, 1035-1047.

Cai, L., Ni, J., Xia, Y., Zi, Z., Ning, K., Qiu, P., Li, X., Wang, B., Liu, Q. & other authors (2012). Identification of an emerging recombinant cluster in porcine circovirus type 2. *Virus research* **165**, 95-102.

Cheung, A. K. (2009). Homologous recombination within the capsid gene of porcine circovirus type 2 subgroup viruses via natural co-infection. *Archives of Virology* **154**, 531-534.

Dayaram, A., Opong, A., Jäschke, A., Hadfield, J., Baschiera, M., Dobson, R. C. J., Offei, S. K., Shepherd, D. N., Martin, D. P. & other authors (2012). Molecular characterisation of a novel cassava associated circular ssDNA virus. *Virus research* **166**, 130-135.

Delwart, E. & Li, L. (2012). Rapidly expanding genetic diversity and host range of the Circoviridae viral family and other Rep encoding small circular ssDNA genomes. *Virus research* **164**, 114-121.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797.

Ge, X., Li, J., Peng, C., Wu, L., Yang, X., Wu, Y., Zhang, Y. & Shi, Z. (2011). Genetic diversity of novel circular ssDNA viruses in bats in China. *Journal of General Virology* **92**, 2646-2653.

Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. (2000). Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573-582.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology* **59**, 307-321.

Gutierrez, C. (1999). Geminivirus DNA replication. *Cellular and Molecular Life Sciences* **56**, 313-329.

Hanley-Bowdoin, L., Settlage, S. B., Orozco, B. M., Nagar, S. & Robertson, D. (2000). Geminiviruses: models for plant DNA replication, transcription, and cell cycle regulation. *Crit Rev Biochem Mol Biol* **35**, 105-140.

Heyraud-Nitschke, F., Schumacher, S., Laufs, J., Schaefer, S., Schell, J. & Gronenborn, B. (1995). Determination of the origin cleavage and joining domain of geminivirus Rep proteins. *Nucleic Acids Research* **23**, 910.

Ilyina, T. V. & Koonin, E. V. (1992). Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. *Nucleic Acids Research* **20**, 3279.

Julian, L., Piasecki, T., Chrzastek, K., Walters, M., Muhire, B., Harkins, G. W., Martin, D. P. & Varsani, A. (2013). Extensive recombination detected amongst Beak and feather disease virus isolates from breeding facilities in Poland. *Journal of General Virology*.

Julian, L., Lorenzo, A., Chenuet, J. P., Bonzon, M., Marchal, C., Vignon, L., Collings, D. A., Walters, M., Jackson, B. & other authors (2012). Evidence of multiple introductions of beak and feather disease virus into the Pacific islands of Nouvelle-Caledonie (New Caledonia). *Journal of General Virology* **93**, 2466-2472.

Kapoor, A., Dubovi, E. J., Henriquez-Rivera, J. A. & Lipkin, W. I. (2012). Complete Genome Sequence of the First Canine Circovirus. *Journal of virology* **86**, 7018-7018.

Kim, H. K., Park, S. J., Song, D. S., Moon, H. J., Kang, B. K. & Park, B. K. (2012). Identification of a novel single stranded circular DNA virus from bovine stool. *Journal of General Virology* **93**, 635-639.

Laufs, J., Traut, W., Heyraud, F., Matzeit, V., Rogers, S. G., Schell, J. & Gronenborn, B. (1995). In vitro cleavage and joining at the viral origin of replication by the replication initiator protein of tomato yellow leaf curl virus. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 3879-3883.

Lefeuvre, P., Lett, J. M., Varsani, A. & Martin, D. P. (2009). Widely conserved recombination patterns among single-stranded DNA viruses. *Journal of Virology* **83**, 2697-2707.

Li, L., Shan, T., Soji, O. B., Alam, M. M., Kunz, T. H., Zaidi, S. Z. & Delwart, E. (2011). Possible cross-species transmission of circoviruses and cycloviruses among farm animals. *Journal of General Virology* **92**, 768-772.

Li, L., Kapoor, A., Slikas, B., Bamidele, O. S., Wang, C., Shaukat, S., Masroor, M. A., Wilson, M. L., Ndjango, J. B. N. & other authors (2010). Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces. *Journal of Virology* **84**, 1674.

Londoño, A., Riego-Ruiz, L. & Arguello-Astorga, G. R. (2010). DNA-binding specificity determinants of replication proteins encoded by eukaryotic ssDNA viruses are adjacent to widely separated RCR conserved motifs. *Archives of Virology* **155**, 1033-1046.

Martin, D. & Rybicki, E. (2000). RDP: Detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562-563.

Martin, D. P., Posada, D., Crandall, K. A. & Williamson, C. (2005). A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Research and Human Retroviruses* **21**, 98-102.

Martin, D. P., Lemey, P., Lott, M., Moulton, V., Posada, D. & Lefeuvre, P. (2010). RDP3: A flexible and fast computer program for analyzing recombination. *Bioinformatics* **26**, 2462-2463.

Martin, D. P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P. & Varsani, A. (2011). Recombination in eukaryotic single stranded DNA viruses. *Viruses* **3**, 1699-1738.

Massaro, M., Ortiz-Catedral, L., Julian, L., Galbraith, J. A., Kurenbach, B., Kearvell, J., Kemp, J., van Hal, J., Elkington, S. & other authors (2012). Molecular characterisation of beak and feather disease virus (BFDV) in New Zealand and its implications for managing an infectious disease. *Archives of Virology* **157**, 1651-1663.

Mu, C., Yang, Q., Zhang, Y., Zhou, Y., Zhang, J., Martin, D. P., Xia, P. & Cui, B. (2012). Genetic variation and phylogenetic analysis of porcine circovirus type 2 infections in central China. *Virus genes* **45**, 463-473.

Muhire, B., Martin, D. P., Brown, J. K., Navas-Castillo, J., Moriones, E., Zerbini, F. M., Rivera-Bustamante, R., Malathi, V. G., Briddon, R. W. & other authors (2013). A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (Geminiviridae). *Archives of Virology*.

Padidam, M., Sawyer, S. & Fauquet, C. M. (1999). Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218-225.

Padilla-Rodriguez, M., Rosario, K. & Breitbart, M. (2013). Novel cyclovirus discovered in the Florida woods cockroach Eurycotis floridana (Walker). *Archives of Virology*.

Phan, T. G., Kapusinszky, B., Wang, C., Rose, R. K., Lipton, H. L. & Delwart, E. L. (2011). The Fecal Viral Flora of Wild Rodents. *PLoS pathogens* **7**, e1002218.

Posada, D. & Crandall, K. A. (2001). Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 13757-13762.

Rosario, K., Duffy, S. & Breitbart, M. (2012a). A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Archives of virology* **157**, 1851-1871.

Rosario, K., Padilla-Rodriguez, M., Kraberger, S., Stainton, D., Martin, D. P., Breitbart, M. & Varsani, A. (2013). Discovery of a novel mastrevirus and alphasatellite-like circular DNA in dragonflies (Epiprocta) from Puerto Rico. *Virus research* **171**, 231-237.

Rosario, K., Dayaram, A., Marinov, M., Ware, J., Kraberger, S., Stainton, D., Breitbart, M. & Varsani, A. (2012b). Diverse circular single-stranded DNA viruses discovered in dragonflies (Odonata: Epiprocta). *Journal of General Virology* **93**, 2729-2739.

Rosario, K., Marinov, M., Stainton, D., Kraberger, S., Wiltshire, E. J., Collings, D. A., Walters, M., Martin, D. P., Breitbart, M. & other authors (2011). Dragonfly cyclovirus, a novel single-stranded DNA virus discovered in dragonflies (Odonata: Anisoptera). *Journal of General Virology* **92**, 1302-1308.

Sikorski, A., Argüello-Astorga, G. R., Dayaram, A., Dobson, R. C. J. & Varsani, A. (2013). Discovery of a novel circular single-stranded DNA virus from porcine faeces. *Archives of Virology* **158**, 283-289.

Smith, J. M. (1992). Analyzing the mosaic structure of genes. *Journal of Molecular Evolution* **34**, 126-129.

Steinfeldt, T., Finsterbusch, T. & Mankertz, A. (2001). Rep and Rep' protein of porcine circovirus type 1 bind to the origin of replication in vitro. *Virology* **291**, 152-160.

Timchenko, T., Katul, L., Sano, Y., de Kouchkovsky, F., Vetten, H. J. & Gronenborn, B. (2000). The master rep concept in nanovirus replication: identification of missing genome components and potential for natural genetic reassortment. *Virology* **274**, 189-195.

van den Brand, J. M. A., van Leeuwen, M., Schapendonk, C. M., Simon, J. H., Haagmans, B. L., Osterhaus, A. D. M. E. & Smits, S. L. (2012). Metagenomic analysis of the viral flora of pine marten and European badger feces. *Journal of Virology* **86**, 2360-2365.

Varsani, A., Regnard, G. L., Bragg, R., Hitzeroth, II & Rybicki, E. P. (2011). Global genetic diversity and geographical and host-species distribution of beak and feather disease virus isolates. *Journal of General Virology* **92**, 752-767.

Chapter 3

Novel circular DNA viruses identified in *Procordulia grayi* and *Xanthocnemis zealandica* larvae using metagenomic approaches

Contents

This work has been published in Infection, Genetics and Evolution and is presented in a similar manner to that of the publication:

Dayaram, A., Galatowitsch, M., Harding, J.S., Argüello-Astorga, G.R., Varsani, A. (2014) Novel circular DNA viruses identified in *Procordulia grayi* and *Xanthocnemis zealandica* larvae. *Infection, Genetics and Evolution* 22: 134-141

3.1 Abstract

Recent advances in sequencing and metagenomics have enabled the discovery of many novel single stranded DNA (ssDNA) viruses from various environments. We have previously demonstrated that adult dragonflies, as predatory insects, are useful indicators of ssDNA viruses in terrestrial ecosystems. Here we recover and characterise 13 viral genomes which represent 10 novel and diverse circular replication associated protein (Rep) - encoding single stranded (CRESS) DNA viruses (1628 - 2668 nt) from *Procordulia grayi* and *Xanthocnemis zealandica* dragonfly larvae collected from four high-country lakes in the South Island of New Zealand. The dragonfly larvae-associated CRESS DNA viruses have diverse genome architectures, however, they all encode two major open reading frames (ORFs) three of which were bidirectional and four with unidirectional arrangements. The 13 viral genomes have a conserved NAGTATTAC nonanucleotide motif and, within their predicted Rep proteins, we identified the rolling circle replication (RCR) motifs 1, 2 and 3, as well as superfamily 3 (SF3) helicase motifs. Maximum likelihood phylogenetic and pairwise identity analysis of the Rep amino acid sequences reveal that the dragonfly larvae novel CRESS DNA viruses share <63% pairwise amino acid identity to the Reps of other CRESS DNA viruses whose complete genomes have been determined and available in public databases indicating that these are novel viruses. Our data indicate that CRESS DNA viruses are circulating in larval dragonfly populations, however, we are unable to ascertain whether these viruses are infecting the larvae directly or are transient within dragonflies via their diet.

3.2 Introduction

While chapter 2 describes the isolation of novel CRESS DNA viruses in adult dragonflies in terrestrial environments, chapter 3 extends on this work by examining CRESS DNA viruses in dragonfly and damselfly larvae in aquatic environments using similar techniques.

Metagenomic approaches, coupled with new sequencing technologies, have allowed the exploration of novel and known viral communities in various ecosystems. Novel small circular DNA viruses with diverse genome architectures have been recovered from a variety of sources including environmental samples (soil, water, river sediments), invertebrates and birds, as well as faecal sources. The majority of these viruses are yet to be properly classified within the viral taxonomic structure, given that in most cases the pathway of transmission and hosts species have not yet been identified. Nonetheless, environmental sampling has enabled the exploration of the viral sequence space, enabling identification and the occurrence of viruses in natural ecosystems, beyond what has been discovered as a result of pathogenicity toward humans, plants and animals.

Over the last couple of years our various studies have recovered a variety of ssDNA viruses from adult dragonfly species (*Anax junius, Diplacodes bipunctata*, *Erythemis fusca*, *Erythemis simplicicollis*, *Erythemis vesiculosa*, *Erythrodiplax umbrata*, *Myathiria Marcella*, *Orthetrum Sabina*, *Pantala flavescens*, *Rhionaeschna multicolour*, *Somatochlora meridionalis*, *Tholymis tillarga* and *Xanthocnemis zealandica*) from Kingdom of Tonga, Australia, USA, Bulgaria, Puerto Rico, and New Zealand (Dayaram *et al.*, 2013; Rosario *et al.*, 2013; Rosario *et al.*, 2012b; Rosario *et al.*, 2011). Interestingly, most of the viruses recovered from these dragonfly species have been identified as cycloviruses. However, a few gemycircularviruses, mastreviruses, diverse circular replication associated protein (Rep) - encoding single stranded (CRESS) DNA viruses and a microphage were also recovered from adult dragonflies. In addition to dragonflies, Padilla-Rodriguez *et al*. (2013) recently recovered a novel cyclovirus from *Eurycotis floridana* (Florida woods cockroach) while Pham et al. (Pham *et al.*, 2013a; Pham *et al.*, 2013b) have recently recovered the genomes of circular ssDNA volvoviruses from *Acheta domesticus* and *Gryllus assimilis*.

The isolation and characterisation of these novel viruses suggests that these circular ssDNA viruses may be commonplace in terrestrial insect and invertebrate species, especially top-end predators. It has previously been suggested that predatory insects may be acting as bio-magnifiers in their given environment (see Chapter 2). In addition, a recent study identified plant-infecting geminiviruses and an associated DNA satellite (associated with begomoviruses) in dragonflies sampled at an agricultural field in Puerto Rico (Rosario *et al.*, 2013). This demonstrates that top-end insect predators may act as viral reservoirs for ssDNA viruses and could be used as a ssDNA viral sampling tool in ecosystems for virus surveillance (Rosario *et al.*, 2012b).

Dragonfly larvae are well documented predators in aquatic environments with prey including aquatic invertebrates, tadpoles, small fish, and other dragonfly larvae (Corbet & Brooks, 2008). Hence, just like adult dragonflies in the terrestrial environment, their larvae could be ideal indicators of viruses in aquatic environments. Over the last five years numerous aquatic environments (e.g. freshwater lakes, hypersaline lakes, reclaimed water, and oceans) have been analysed using metagenomic approaches and in all cases a high diversity of ssDNA viruses have been identified (Culley *et al.*, 2006; Emerson *et al.*, 2012; Hewson *et al.*, 2012; Labonte & Suttle, 2013; López-Bueno *et al.*, 2009; Rosario *et al.*, 2009a; Rosario *et al.*, 2009b; Roux *et al.*, 2012).

In order to determine whether ssDNA viruses could be detected in dragonfly larvae, we undertook a pilot study in the South Island of New Zealand where commonly found species of Odonata larvae include *Procordulia grayi* (Selys) and *Xanthocnemis zealandica* (McLachlan).

3.3 Materials and methods

3.3.1 Sample collection

Dragonfly and damselfly larvae samples were collected from four high-country lakes with minimal human impact: Lake Donne (43°36'30.34"S, 171° 6'56.73"E; 663m alt.), Lake Grasmere (43° 3'48.28"S, 171°46'30.33"E; 584m alt.), Lake Sarah (43° 2'57.11"S, 171°46'35.04"E; 577m alt.) and Lake Hawdon (43° 6'12.83"S, 171°50'57.58"E; 576m alt.). Late instar larvae were collected in the austral summer (February and December 2012) using 1-meter D-net sweeps (1 mm mesh) along the shallow lake bottoms and emergent aquatic vegetation. Repeated sweeps in these habitats were undertaken until 10-20 larvae of each species were collected. For virus analyses, live larvae were sorted in the field and kept in their lake water for transport back to the laboratory. Other larvae were collected and preserved in 80% ethanol for gut analyses. Species identifications were confirmed in the laboratory using identification keys outlined for New Zealand Odonata species (Rowe, 1987).

3.3.2 Viral purification and DNA extraction

Larvae of each species within each lake were homogenized in SM buffer [0.1 M NaCl, 50 mM Tris/HCl (pH 7.4), 10mM $MgSO_4$] at a ratio of 5 ml SM buffer to 2.5 g of larvae tissue. The homogenate was processed as previously described (Dayaram *et al.*, 2013; Rosario *et al.*, 2012b). In brief, the homogenate was pelleted and the supernatant was then filtered sequentially through a 0.45 µm pore size syringe filter (Sartorius Stedim Biotech, Germany) followed by a 0.2 µm filter and finally viral DNA was then extracted from the filtrate using the High Pure Viral Nucleic Acid kit (Roche, USA).

3.3.3 Enrichment of circular DNA, viral identification and cloning of viral genomes

The viral nucleic acid was enriched for circular DNA using rolling circle amplification (RCA) using Illustra TempliPhi Amplification kit (GE Healthcare, USA). The RCA concatenated DNA from each sample group was digested with either *Bam*H1, *Sma*1, *Eco*R1 or *Xmn*1 in separate reactions resulting in DNA fragments between ~1.7-2.6 KB. The resulting fragments were gel purified and then cloned into *Eco*R1 pGEM3ZF (+) (Promega, USA) cut plasmid or into pUC-19 plasmid vector cut with *Bam*H1 or *Sma*1 (Fermentas, USA). The clones were sequenced by primer walking at Macrogen Inc. (South Korea). The putative genomes were verified either by restriction mapping or by designing back-to-back primers followed by PCR amplification of the genomes using Kapa HiFi HotStart DNA polymerase (Kapa Biosystems, USA), then cloning the amplicon into pJET1.2 (ThermoFisher, USA) and sequencing the recombinant plasmid.

A next-generation sequencing approach was also implemented to explore the viral diversity amongst the larvae. All enriched RCA products were grouped together and sequenced on an Illumina HiSeq 2000 (Illumina) platform at the Beijing Genomics Institute (Hong Kong). The resulting paired end reads were then assembled using ABySS V1.3.5 (Simpson *et al.*, 2009) with kmer = 64. A full BLASTx (Altschul *et al.*, 1990) analysis was performed on the assembled contigs using KoriBlast v4.1 (Korilog SARL, France). A BLASTx analysis was carried out in contigs >750nt to identify credible hits to potential ssDNA viral proteins. To recover the full genomes for these sequences, back-to-back primers were designed (Table 3.1) and the full genomes were amplified using Kapa HiFi HotStart polymerase (Kapa Biosystems USA). The amplicons were gel purified, cloned into pJET1.2 (ThermoFisher, USA) and the clones were sequenced at Macrogen Inc. (South Korea) by primer walking.

3.3.4 Viral genome and phylogenetic analysis

All genomes were assembled using DNAMAN (version 5.2.9; Lynnon Biosoft) and preliminary genome analysis was carried out using BLASTx and tBLASTx (Altschul *et al.*, 1990). Maximum likelihood (ML) phylogenetic analysis was performed on all the Rep sequences of ssDNA viruses available in public databases derived from complete CRESS DNA viruses together with those recovered in this chapter. The amino acid sequences of Rep proteins were aligned using T-coffee (Notredame *et al.*, 2000), refined using MUSCLE (Edgar, 2004) and edited manually. This alignment was then used to infer a ML phylogenetic tree using PHYML (Guindon *et al.*, 2010) using WAG+G model of amino acid substitution, chosen as best fit model using ProtTest 3 (Darriba *et al.*, 2011) with approximate likelihood-ratio test for branches (Anisimova & Gascuel, 2006). All branches with less that 80% branch support were collapsed using Mesquite v2.75 (http://mesquiteproject.org). Pairwise identities of the Rep sequences were calculated using SDT v1.0 (Muhire *et al.*, 2013).

Table 3.1: Detail of primer sequences and restriction enzymes to recover complete dragonfly larvae associated CRESS DNA viral genomes

	Initial restriction enzyme	Second restriction enzyme	Forward primer	Reverse primer
DflaCV-1 [NZ-PG11-LD]	*Sma*1	*Hind*III	-	-
DflaCV-2 [NZ-PG8-LS]			GACTGGCACTGGGAAATCTAGAACT	GAACCCCAGAACACATAGCAGG
DflaCV-3 [NZ-PG1-LG]			GACGAGCCATATGGTCCTTGG	CAGATGGTTTGGATGGTGTTTGTGG
DflaCV-3 [NZ-PG4-LH]			GACGAGCCATATGGTCCTTGG	CAGATGGTTTGGATGGTGTTTGTGG
DflaCV-4 [NZPG3-LG]			AAGAGGCGAATGGTGGGAT	GGCTTGGGATAGGGATTCT
DflaCV-5 [NZ-PG2-LG]	*Eco*RI		CTAACCCACAAATGAATAATGTTGCC	GGTAATGGATGATGTGATAGATGG
DflaCV-5 [NZ-PG7-LS]	*Xmn*I		GTGGTGGGTCATGAGATCTG	GATGTACTTGCAGGCGATGTC
DflaCV-6 [NZ-PG9-LD]			AATGGTGCACGGATAGGATTG	TTTGTCTTACGGATCCAACG
DflaCV-7 [NZ-PG5-LH]			CAGGGGTATCTGGAGCTGAA	CCAATGCTCCCTCCCCG
DflaCV-8 [NZ-PG5-LS]			GAGAGCAGAGGGACATAGTACGG	CGGGTAGGGATAGTACGGGGTA
DflaCV-9 [NZ-PG10-LD]	*Xmn*I	*Eco*RV	-	-
DflaCV-10 [NZ-XZ2-LS]	*Xmn*I		GGGGTATCTAGTGATTGTCTTTC	GGTGGTAAACTGCTGATCTCC
DflaCV-10 [NZ-XZ1-LH]	*Pst*I		GCTAGTGATGTCAACAATCTTCCCG	ATTCATTGAGAGACAGTCGGTATAGATG

3.4 Results and discussion

3.4.1 Novel viral genome analysis

Two odonata larval species were collected from the four lakes sampled. From these, thirteen novel ssDNA viruses were recovered. Three genomes were isolated from Lake Donne (all from *P. grayi*), four from Lake Sarah (*P. grayi,* n=3; *X. zealandica,* n=1), three from Lake Grasmere (all from *P. grayi*) and three from Lake Hawdon (*P. grayi,* n=2*; X. zealandica,* n=1). Overall, eleven viral genomes were recovered from *P. grayi* and only two from *X. zealandica.* Taking genome-wide pairwise identities into account, we were able to distinguish 10 unique viral species for which we propose a simple nomenclature, Dragonfly larvae-associated circular virus (DflaCV) 1 through to 10. Two DflaCV-3 isolates (isolate NZ-PG1-LG and NZ-PG4-LH) both isolated from *P. grayi* (Lake Grasmere and Lake Hawdon) shared 99.3% genome wide nucleotide identity; two DflaCV-5 isolates (isolates NZ-PG2-LG and NZ-PG7-LS) from *P. grayi* (Lake Grasmere and Lake Sarah) shared 99.8% genome wide nucleotide identity; whereas DflaCV-10 isolates (isolates NZ-XZ2-LS and NZ-XZ2-LH), from *X. zealandica* larvae (Lake Hawdon and Lake Sarah) shared 84% genome wide pairwise nucleotide identity. The genomes of the isolated DflaCVs range in size from 1628 to 2688 nt and have a variable genome organisation, however, they all have only two major open reading frames (ORFs) (Figure 3.1). In eleven of the DflaCV genomes recovered, the ORFs are bidirectionally organised whereas in the remaining two, the ORFs are unidirectionally organised in the viron sense. Of the two major ORFs in the CRESS DNA viruses, one ORF encodes the Rep and the other a putative capsid protein (CP). It is worth noting that BLASTp analysis of the putative capsid protein did not provide any credible hits (E-value <0.01) to homologues in the public databases.

3.4.2 Sequence analysis of the intergenic region

In all the DflaCV genomes, we identified a putative stem-loop structure with the conserved nonanucleotide NAGTATTAC motif (Table 3.2; Figure 3.1). As observed in other ssDNA viruses, this motif is where Rep cleaves the DNA positive strand to initiate the virus replication by a rolling-circle (RC) mechanism (Gutierrez *et al.*, 2004).

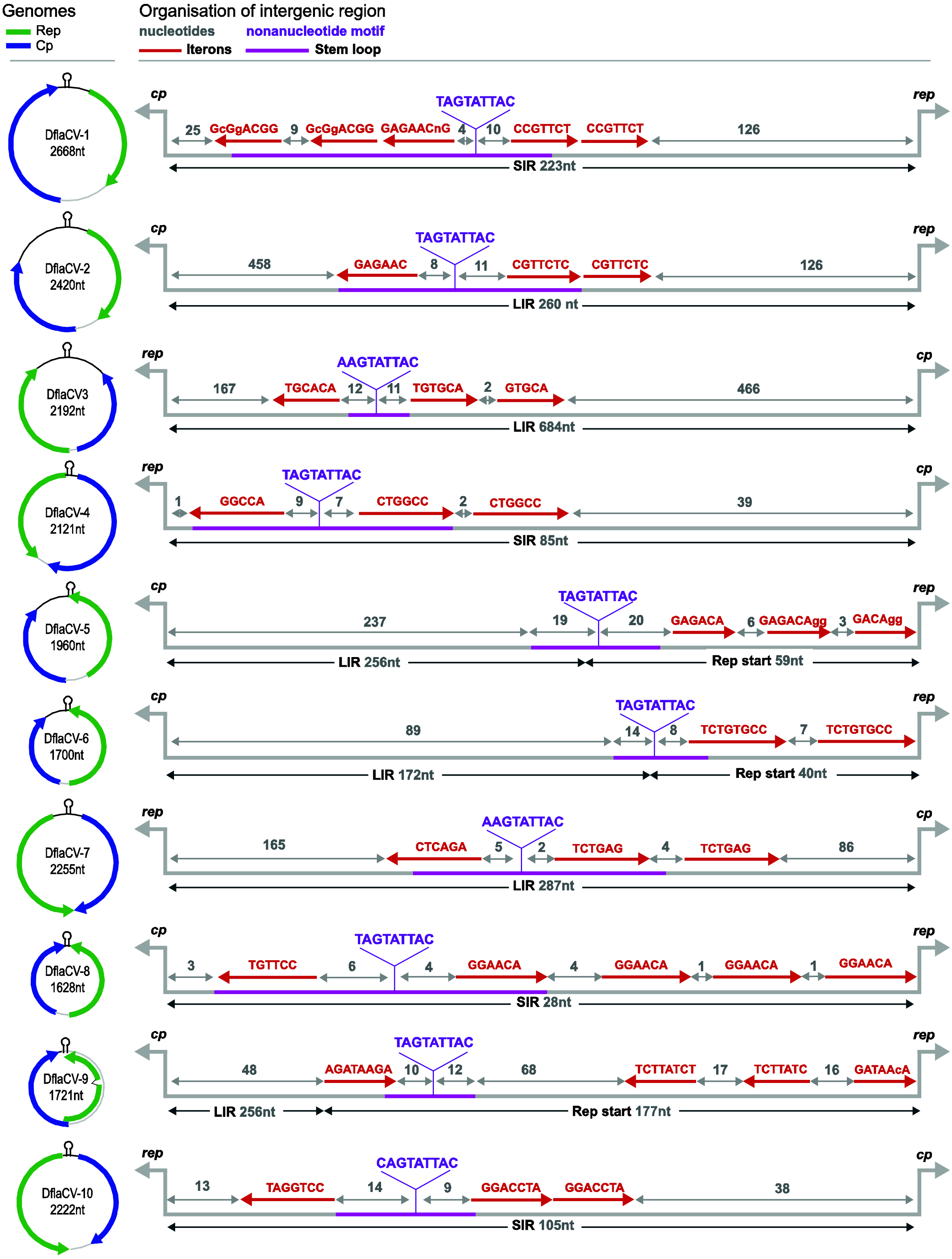


Figure 3.1. Genome organisation of dragonfly larvae-associated CRESS DNA viruses indicating the organisation of their intergenic regions. LIR – long intergenic region; SIR – short intergenic region.

Close to the former element we identified the repeated sequence motifs (iterons) that are assumed to function as specific Rep-binding sites (Figure 3.1). The position of the conserved stem-loop element within the genome of the DflaCVs was variable: in some viruses, the putative *Ori* was located in the region between the 5´ ends of the *rep* and *cp* ORFs (i.e. DflaCV-4, -7, and -10), others between the 3´ends of those genes (i.e. DflaCV-3 and -8), and yet others between the 3´end of *cp* and the 5´end of *rep* (i.e. DflaCV-1 and -2). A fourth group of viruses harbour the *Ori* within the rep gene segment encoding the Rep C-terminus (i.e. DflaCV-5, -6, and -9). The DflaCVs were also diverse in the sequence and arrangement of their iterons. In fact, with the sole exceptions of DflaCV-1 and -2, all the novel viral species exhibited iterons with different core sequences (Figure 3.1 and 3.2).

3.4.3 Replication associated protein analysis

The Reps of eukaryotic circular ssDNA viruses have been found to have several highly conserved motifs, reviewed in Rosario *et al*. (2012a). Most of these motifs were identified in the Rep sequences of the 13 CRESS DNA viral isolates (Table 3.2), including rolling circle replication (RCR) motifs 1, 2 and 3, as well as superfamily 3 (SF3) helicase motifs. The SF3 helicase Walker A motif is found to be conserved across most circular ssDNA viruses and may act as a deoxyribonucleotide triphosphate (dNTP) binding domain (Hickman & Dyda, 2005), but it also may be involved in the Reps helicase activity during rolling circle replication (RCR) (Londoño *et al.*, 2010). We note that in two viral genomes of DflaCV-5 (isolated from samples NZ-LG-13X and LS-16E) which are 99.8% similar, the Walker A motif is not present (Table 3.2). Additionally, in the genome of DflaCV-9 we identified an intron sequence (38 nt) within the rep-coding region.

The DNA-binding specificity determinants (SPDs) of the Reps of a variety of circular ssDNA viruses have been tentatively mapped to discrete amino acid clusters (i.e., SPD-r1 and –r2) in the vicinity of the RCR motifs 1 and 2 (Londono *et al.*, 2010). Recently, a comprehensive analysis of the proteins encoded by the members of the *Cyclovirus* genus (including seven species isolated from adult dragonflies) supported the hypothesis of a possible connection between the iteron core sequence and the putative Rep SPDs (Chapter 2). Accordingly, we decided to systematically analyse the Rep proteins encoded by DflaCVs to verify the existence of such a correlation between the putative Rep-binding sites and SPDs.

Table 3.2: Conserved motifs identified in the dragonfly larvae-associated CRESS DNA viral genomes

						RCR motifs			SF3 Helicase Motifs		
Virus isolate	Accession #	Recovery	Host species	Sample site	Nonanucleotide motif	I	II	III	Walker A	Walker B	Motif C
DflaCV-1 [NZ-PG11-LD]	KF738873	RR	*Procordulia grayi*	Lake Donne	TAGTATTAC	LTIPH	HWQILV	YVWKEDTR	GPTG-TGKS	VIVDEFRGA	FTSN
DflaCV-2 [NZ-PG8-LS]	KF738874	I	*Procordulia grayi*	Lake Sarah	TAGTATTAC	LITPR	HWQVMV	YVWKEETR	GSTG-TGKS	VVIDEFRGG	ITSN
DflaCV-3 [NZ-PG1-LG]	KF738875	I	*Procordulia grayi*	Lake Grassmere	AAGTATTAC	FTLNN	HLQGFV	YCSKDGDV	RSAG-TGKS	VVMDDMDPD	VTSQ
DflaCV-3 [NZ-PG4-LH]	KF738876	I	*Procordulia grayi*	Lake Hawdon	AAGTATTAC	FTLNN	HLQGFV	YCSKDGDV	GAAG-TGKS	VVMDDMDPD	VTSQ
DflaCV-4 [NZPG3-LG]	KF738877	I	*Procordulia grayi*	Lake Grassmere	TAGTATTAC	FTINN	HLQGFA	YCKKEGEF	GPTG-TGKF	VAIEEFEPR	YAAW
DflaCV-5 [NZ-PG2-LG]	KF738878	RR	*Procordulia grayi*	Lake Grassmere	TAGTATTAC	FTAFV	HWQGYV	YCKKDMKF	---------	IIVRHAKSY	ICVE
DflaCV-5 [NZ-PG7-LS]	KF738879	RR	*Procordulia grayi*	Lake Sarah	TAGTATTAC	FTAFV	HWQGYV	YCKKDMKF	---------	IIVRHAKSY	ICVE
DflaCV-6 [NZ-PG9-LD]	KF738880	I	*Procordulia grayi*	Lake Donne	TAGTATTAC	FTLFH	HLQGYL	YCKKEGDW	GSTG-VGKT	IILDDFHWD	ITCE
DflaCV-7 [NZ-PG5-LH]	KF738881	I	*Procordulia grayi*	Lake Hawdon	AAGTATTAC	FTLFE	HWQGYL	YCSKDGNW	GPTG-CGKS	VILDELRAD	ITSP
DflaCV-8 [NZ-PG5-LS]	KF738882	I	*Procordulia grayi*	Lake Sarah	TAGTATTAC	ITLFG	HSHMYL	YCWSEGDH	GPTG-TGKT	LANAPSPPD	YCSM
DflaCV-9 [NZ-PG10-LD]	KF738883	RR	*Procordulia grayi*	Lake Donne	TAGTATTAC	FTLNN	HLQGYI	YCKKENNF	GNIT-TYTG	IVIDDFRAS	ITSI
DflaCV-10 [NZ-XZ2-LS]	KF738884	RR	*Xanthocnemis zealandica*	Lake Sarah	CAGTATTAC	FTLNN	HLQGFI	YCGKDADV	DSVGNTGKT	VVFDLSRSQ	VFAN
DflaCV-10 [NZ-XZ1-LH]	KF738885	RR	*Xanthocnemis zealandica*	Lake Hawdon	CAGTATTAC	FTLNN	HLQGFI	YCGKDSDV	DSVGNTGKT	VVFDLSRSQ	VFAN

RR - RCA / Restriction enzyme-mediated recovered viral genomes

I - Illumina next generation-informed recovered viral genomes

Iterons		SPD-r1		Motif I		Motif II		SPD-r2		
GTTCTC	M-16-RR	QGI	FWM	LTIPH	---26---	HWQILV	AMST	KSS	LAQVK	**DflaCV-1**
	MARR	QGI	FWL	LTIPR	---27---	HWQVMV	AFTK	KIG	LSGVK	**DflaCV-2**
	MIRR	QGV	YWL	LTIPE	---26---	HWQVLV	VLRR	KGS	LSTIT	RW-E [FJ959081]
	MRR	QGI	FWM	LTIPH	---26---	HWQLLV	GFAK	KVS	LRRVR	18-LDMD [KF133825]
TGTGYA	MANR	RSR	GWC	FTLNN	---29---	HLQGFV	YFPN	AKT	FNGAK	**DflaCV-3**
	MS	RNR	NYV	FTLNN	---30---	HLQGYI	CFPN	AKT	ISAVR	DipCV [KC248416]
	MNK	RIR	NVC	FTGFN	---27---	HWQGYV	ELKN	AKT	FSAIK	RW-C [FJ959079]
	MS	RKR	DYC	FTDFV	---27---	HLQGYI	YFKN	AKT	FSAVR	SI03931 [JX904581]
CTGGCCa	MVGTTQ	RSR	GWV	FTINN	---31---	HLQGFA	YFKQ	RIS	FNGIR	**DflaCV-4**
	MPVE	RAR	GWC	FTLNN	---31---	HWQGFC	YFRE	RKS	FLQVK	d33403 [JN377561]*
GAGACA	M-18-DQ	RGR	CWC	FTAFV	---25---	HWQGYV	EFGT	QRR	PTGVQ	**DflaCV-5**
CTGTKCC	MAC	RSR	GWA	FTLFH	---30---	HLQGYL	YYAN	PVR	FETVK	**DflaCV-6**
	MS	RLR	NVC	FTLYN	---37---	HLQGYL	ECTK	PVR	FGTLK	15-LDMD [KF133822]
TCTGAG	MALILAK	RVR	RVC	FTLFE	---27---	HWQGYL	ELKS	QML	WTTVK	**DflaCV-7**
	MP	RFR	NIC	FTHYS	---33---	HLQGYM	EFSK	QLS	RKKIK	MPSH01427-GM3 [BAKC01000035]*
YGGAACW	MAN	PNT	CRW	MITLF	---27---	HSHMYL	RFDK	KCR	MSTLK	**DflaCV-8**
	MGS	PAI	GWC	FTLNN	---29---	HLQGYL	ELHK	KKR	FKQVK	DfCycIV [JX185418]
	M-27-VS	PAL	RWC	FTLNN	---32---	HLQGFL	RFKT	KRR	PKAIT	MPSH06489-GM1 [BAKC01000099]*
TCTTATC	MGH	QVK	RWC	FTLNN	---30---	HLQGYI	ELKK	KSS	LNQIK	**DflaCV-9**
GGACCTA	M-5-GS	QAR	RWC	FTLNN	---33---	HLQGFI	HLRK	RLR	LACVK	**DflaV-10**

*metagenomic contig derived partials

Figure 3.2: Summary of potential DNA-binding SPDs, iterons and conserved RCR motif I and II of DflaCVs and other CRESS DNA homologues. Putative SPD residues are highlighted in red and these cluster in two distinct regions labelled as SPD-r1 and SPD-r2. Metagenomic contig derived partials are not full viral genomes but contain the full Rep.

Table 3.3: Percentage pairwise identities of the Rep amino acid sequences of dragonfly larvae associated CRESS DNA viruses with those encoded by other complete CRESS DNA virus genomes (five highest identities). Ranging percentage identities relate to all isolates of a given virus.

DflaCV ID	CRESS DNA virus	Genbank accession #	% pairwise amino acid identity
DflaCV-1	DflaCV-2	KF738874	59.6
	RW-E	FJ959081	59.1
	18-LDMD	KF133825	58.2
	12-LDMD	KF133819	57.3
	SI00898	JX904478	53.8
DflaCV-2	SI00898	JX904478	62.6
	RW-E	FJ959081	61.9
	18-LDMD	KF133825	60.5
	DflaCV-1	KF738873	59.6
	12-LDMD	KF133819	59.6
DflaCV-3	Diporeia_spCV	KC248416	50.0 - 50.4
	CB-A	FJ959082	43.9 - 44.3
	19-LDMD	KF133826	43.9 - 44.2
	SOG00182	JX904077	43.3 - 44.1
	SOG05268	JX904185	43.1 - 43.8
DflaCV-4	RodSCV-V-77	JF755415	47.7
	19-LDMD	KF133826	43.8
	GOM03041	JX904344	40.9
	SI03717	JX904562	39.8
	CB-A	FJ959082	38.3
DflaCV-5	RodSCV-M-45	JF755409	33.5
	RodSCV-V-69	JF755403	33.2
	CB-A	FJ959082	32.5
	SI00373	JX904431	32.0
	SI03701	JX904559	31.8
DflaCV-6	StCV	DQ172906	39.2
	DfCyV 3	JX185424	38.4
	AtCopCV	JQ837277	38.2
	FWCasCyV	JX569794	38.1
	NG12-CyV	GQ404854	37.6
DflaCV-7	RW-C	FJ959079	43.7
	SI00003	JX904394	42.8
	5-LDMD	KF133812	36.4
	20-LDMD	KF133827	36.4
	SI00349	JX904427	35.8
DflaCV-8	ShrimpCDV	KC441518	33.2
	DflaCV-7	KF738881	32.2
	RW-A	FJ959077	31.8
	9-LDMD	KF133816	31.8
	CanaryCV	AJ301633	31.7
DflaCV-9	SDWAPI	HQ335042	43.7
	BatCV-TM6C	HM228875	42.9
	DfCyclV	JX185418	40.4
	13-LDMD	KF133820	39.9
	SI00850	JX904473	39.4
DflaCV-10	CynNCXV	JX908739	44.3 - 44.6
	SI00142	JX904416	39.7 - 42.6
	SI03654	JX904548	39.3 - 40.5
	MS584-5	HQ322117	38.6 - 39.2
	SOG04070	JX904144	38.4 - 39.3

The results of this new analysis are summarized in Figure 3.2, where the N-terminal domain of all DflaCV Reps are aligned using the conserved RCR motifs as references.By means of a heuristic approach the SPD regions together with the adjacent RCR motifs were utilised to retrieve the sequences of circular ssDNA viruses with similar iterons from public databases (Arguello-Astorga, unpublished). From the results we were able to identify several viruses from aquatic ecosystems or animals that display similar iterons and putative SPDs to those of DflaCV-1, -2, -3, -4, -6, and -7 (Figure 3.2).

The Reps of DflaCV-1 and DflaCV-2 share 59.6% pairwise amino acid identity and 53.8 – 62.6% pairwise amino acid identity with Reps of RW-E, 18-LDMD, 12-LDMD, SI00898 (Table 3.3; see Table 3.4 for GenBank accession numbers). The Reps of the two DflaCV-3 genomes share 98.6% pairwise amino acid identity and ~50% pairwise amino acid identity to *Diporeia sp*-associated circular virus (KC248416). DflaCV-4 shares >40% pairwise amino acid identity with rodent stool isolate RodSCV-V-77 (JF755415), as does DflaCV-7 with reclaimed water viral genome RW-C (FJ959079), DflaCV-10 with *Cyanoramphus auriceps* nest material viral genome CynNCXV (JX908739) and DflaCV-9 with viral genome SDWAPI from mosquitos (HQ335042). The ML phylogenetic analysis (Figure 3.3) of the Rep sequences corroborates with our pairwise-identity based analysis suggesting that all ten DflaCV species described in this chapter are all novel CRESS DNA viruses.

Based on the Rep sequence analysis (Figure 3.3, Table 3.3) we can putatively assign DflaCV-1 and -2 isolates to the same group as RW-E, CB-B, RodSCV-M-44, YN-BTCV-1, 12-LDMD, 18-LDMD and SI00898 (see Table 3.4 for accession numbers) given that their Reps share >53% amino acid pairwise identity. Interestingly, the genomes of DflaCV-1, DflaCV-2, RW-E, 12-LDMD, 18-LDMD and SI00898 have a unidirectional genome organisation (see Figure 3.1). The associations to these viruses based on the Rep analysis and the similar genome organisation may suggest evolutionary relationships between these viruses.

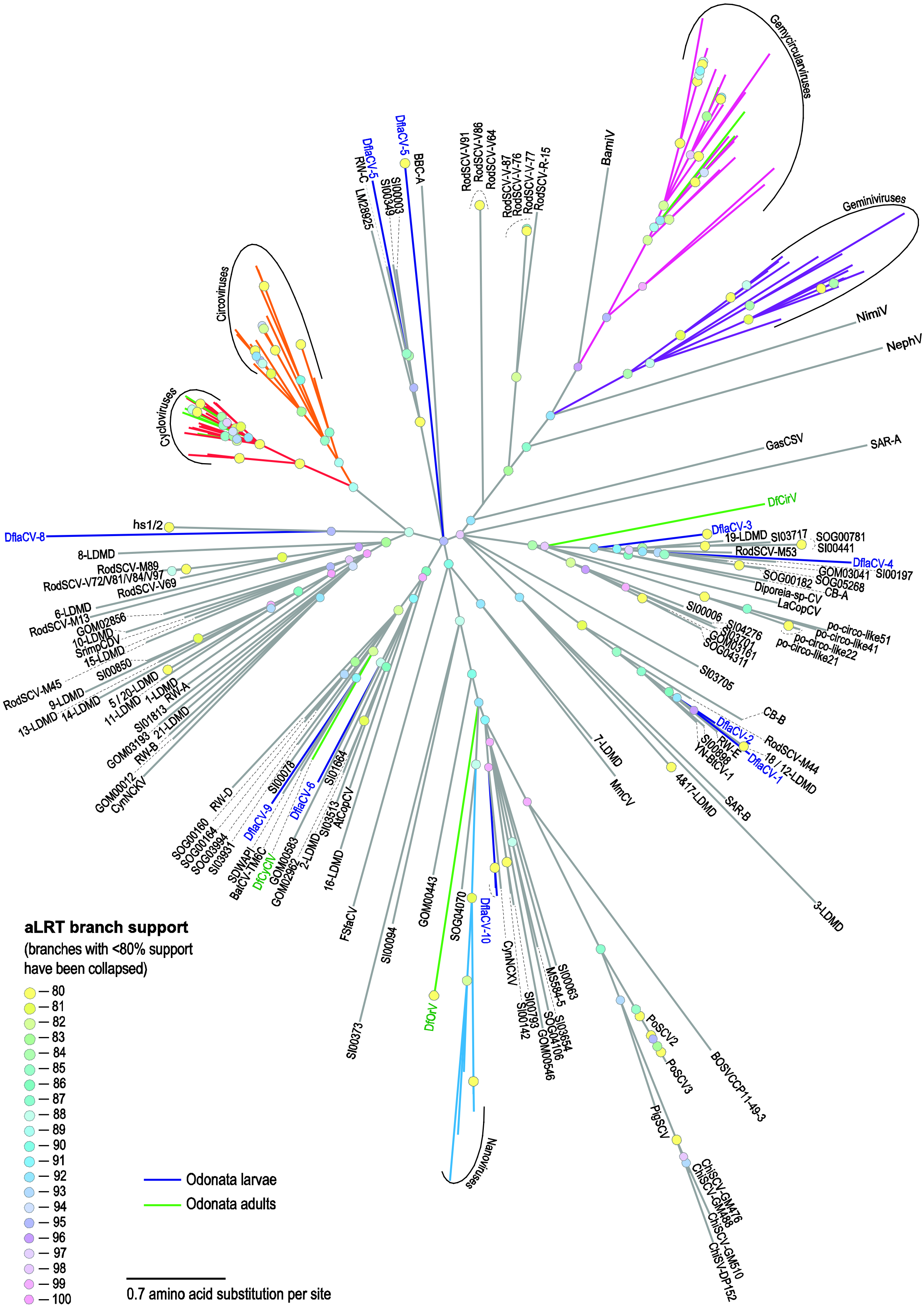


Figure 3.3: Unrooted maximum likelihood phylogenetic tree of Rep sequences encoded by complete CRESS DNA viral genomes. See Table 3.4 for GenBank accession numbers associated with acronyms used in the figure

Table 3.4: List of complete CRESS DNA virus genomes and their accession number

Acronym	Genbank accession #	Acronym	Genbank accession #	Acronym	Genbank accession #
10-LDMD	KF133817	GOM00231	JX904212	RodSCV-V-91	JF755417
11-LDMD	KF133818	GOM00443	JX904231	RodSCV-V-97	JF755414
12-LDMD	KF133819	GOM00546	JX904245	RW-A	FJ959077
13-LDMD	KF133820	GOM00583	JX904250	RW-B	FJ959078
14-LDMD	KF133821	GOM02856	JX904312	RW-C	FJ959079
15-LDMD	KF133822	GOM02962	JX904333	RW-D	FJ959080
16-LDMD	KF133823	GOM03041	JX904344	RW-E	FJ959081
17-LDMD	KF133824	GOM03161	JX904368	SAR-A	FJ959084
18-LDMD	KF133825	GOM03193	JX904377	SAR-B	FJ959085
19-LDMD	KF133826	hs1	JX559621	ShrimpCDV	KC441518
1-LDMD	KF133807	hs2	JX559622	SI00003	JX904394
20-LDMD	KF133827	LaCopCV	JF912805	SI00006	JX904395
21-LDMD	KF133828	LM28925	KC248425	SI00063	JX904401
2-LDMD	KF133808	MmCV	JQ085285	SI00078	JX904407
3-LDMD	KF133810	MS584-5	HQ322117	SI00094	JX904412
4-LDMD	KF133811	NephV	JQ898333	SI00142	JX904416
5-LDMD	KF133812	NimiV	JQ898332	SI00197	JX904420
6-LDMD	KF133813	PigSCV	JX274036	SI00349	JX904427
7-LDMD	KF133814	po-circo-like21	JF713716	SI00373	JX904431
8-LDMD	KF133815	po-circo-like22	JF713717	SI00441	JX904439
9-LDMD	KF133816	po-circo-like41	JF713718	SI00793	JX904469
AtCopCV	JQ837277	po-circo-like51	JF713719	SI00850	JX904473
BamiV	JQ898331	PoSCV2	KC545226	SI00898	JX904478
BatCV-TM6C	HM228875	PoSCV3	KC545227	SI01664	JX904518
BBC-A	FJ959086	PoSCV3	KC545228	SI01813	JX904523
BOSVCCP11-49-3	JN634851	PoSCV3	KC545229	SI03513	JX904541
CB-A	FJ959082	PoSCV3	KC545230	SI03654	JX904548
CB-B	FJ959083	PSMV	JF905486	SI03701	JX904559
ChiSCV-DP152	GQ351272	RhFeCV	JQ814849	SI03705	JX904561
ChiSCV-GM476	GQ351274	RodSCV-M-13	JF755410	SI03717	JX904562
ChiSCV-GM488	GQ351276	RodSCV-M-44	JF755408	SI03931	JX904581
ChiSCV-GM510	GQ351275	RodSCV-M-45	JF755409	SI04276	JX904605
CynNCKV	JX908740	RodSCV-M-53	JF755414	SOG00160	JX904075
CynNCXV	JX908739	RodSCV-M-89	JF755402	SOG00164	JX904076
DfaCV-1	JX185430	RodSCV-R-15	JF755401	SOG00182	JX904077
DfaCV-2	JX185429	RodSCV-V-64	JF755407	SOG03994	JX904139
DfCirV	JX185415	RodSCV-V-69	JF755403	SOG04070	JX904144
DfCyClV	JX185418	RodSCV-V-72	JF755411	SOG04106	JX904147
DfOrV	JX185416	RodSCV-V-76	JF755404	SOG04311	JX904151
DfOrV	JX185417	RodSCV-M-53	JF755415	SOG05268	JX904185
Diporeia sp-CV	KC248416	RodSCV-V-81	JF755412	YN-BtCV-1	JF938078
FSfaCV	KF246569	RodSCV-V-84	JF755413	SOG00781	JX904107
GasCSV	KC172652	RodSCV-V-86	JF755416		
GOM00012	JX904192	RodSCV-V-87	JF755406		

3.5 Concluding remarks

Both Odonata species sampled in this study are relatively long-lived benthic invertebrate predators (*X. zealandica,* 1-2 years; *P. grayi,* 2-4 years larval development) and feed on a variety of invertebrate prey. Their diet has been previously determined by examining gut contents, field observations, and published accounts of gut/faecal contents (Crumpton, 1979; Rowe, 1987). *X. zealandica* feed on predominately on non-biting midge larvae (Chironomidae) and Oligochaeta worms; however, they also feed on Crustacea (*Cladocera, Ostracoda*), other damselflies (*Austrolestes colensonis*), and are cannibalistic. *P. grayi* may be the top lake invertebrate predator feeding on the same diet as *X. zealandica* but also including larger prey: water true bugs (*Notonectidae, Corixidae*), beetles (Dytiscidae), caddisflies (*Leptoceridae, Hydroptilidae*) as well as displaying cannibalism. Therefore the CRESS DNA viruses we have recovered from the *X. zealandica* and *P. grayi* could have resulted from any of these dietary sources, the larvae themselves or any microorganism present in the lakes. We are currently testing some of these other aquatic invertebrates to determine whether these viruses are shared through the food web.

It is interesting to note that we have found a large diversity of cycloviruses in adult dragonflies and damselflies including *X. zealandica* (Dayaram *et al.*, 2013; Rosario *et al.*, 2013; Rosario *et al.*, 2012b; Rosario *et al.*, 2011), however, we did not detect or recover any cycloviruses from odonata larvae in this study. To date, cyclovirus genomes have been recovered from adult dragonflies, Florida wood cockroach, faecal matter, bats, animal meats and human cerebrospinal fluid (Dayaram *et al.*, 2013; Delwart & Li, 2012; Ge *et al.*, 2011; Li *et al.*, 2010; Li *et al.*, 2011; Padilla-Rodriguez *et al.*, 2013; Rosario *et al.*, 2012b; Rosario *et al.*, 2011; Smits *et al.*, 2013; Tan *et al.*, 2013), hence it is highly likely that they are associated with (and infect) terrestrial organisms. Nonetheless, our discovery of novel CRESS DNA viruses in the larvae of dragonflies clearly indicates that, like the adults, they are a useful tool for viral surveillance in ecosystems.

GenBank accession numbers: KF738873 - KF738885

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.

Anisimova, M. & Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic biology* **55**, 539-552.

Corbet, P. & Brooks, S. (2008). *Dragonflies*, vol. 106. Harpercollins Pub Ltd.

Crumpton, J. (1979). Aspects of the Biology of Xanthocnemis-Zealandica and Austrolestes-Colensonis (Odonata, Zygoptera) at 3 Ponds in the South-Island, New-Zealand. *New Zeal J Zool* **6**, 285-297.

Culley, A. I., Lang, A. S. & Suttle, C. A. (2006). Metagenomic analysis of coastal RNA virus communities. *Science* **312**, 1795-1798.

Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164-1165.

Dayaram, A., Potter, K. A., Moline, A. B., Rosenstein, D. D., Marinov, M., Thomas, J. E., Beitbart, M., Rosario, K., Argüello-Astorga, G. R. & other authors (2013). High global diversity of cycloviruses amongst dragonflies. *Journal of General Virology* **94**, 1827-1840.

Delwart, E. & Li, L. L. (2012). Rapidly expanding genetic diversity and host range of the Circoviridae viral family and other Rep encoding small circular ssDNA genomes. *Virus Research* **164**, 114-121.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797.

Emerson, J. B., Thomas, B. C., Andrade, K., Allen, E. E., Heidelberg, K. B. & Banfield, J. F. (2012). Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. *Applied and environmental microbiology* **78**, 6309-6320.

Ge, X. Y., Li, J. L., Peng, C., Wu, L. J., Yang, X. L., Wu, Y. Q., Zhang, Y. Z. & Shi, Z. L. (2011). Genetic diversity of novel circular ssDNA viruses in bats in China. *Journal of General Virology* **92**, 2646-2653.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology* **59**, 307-321.

Gutierrez, C., Ramirez-Parra, E., Mar Castellano, M., Sanz-Burgos, A. P., Luque, A. & Missich, R. (2004). Geminivirus DNA replication and cell cycle interactions. *Veterinary microbiology* **98**, 111-119.

Hewson, I., Barbosa, J. G., Brown, J. M., Donelan, R. P., Eaglesham, J. B., Eggleston, E. M. & LaBarre, B. A. (2012). Temporal dynamics and decay of putatively allochthonous and autochthonous viral genotypes in contrasting freshwater lakes. *Applied and environmental microbiology* **78**, 6583-6591.

Hickman, A. B. & Dyda, F. (2005). Binding and unwinding: SF3 viral helicases. *Current opinion in structural biology* **15**, 77-85.

Labonte, J. M. & Suttle, C. A. (2013). Previously unknown and highly divergent ssDNA viruses populate the oceans. *The ISME journal* **7**, 2169-2177.

Li, L., Kapoor, A., Slikas, B., Bamidele, O. S., Wang, C., Shaukat, S., Masroor, M. A., Wilson, M. L., Ndjango, J. B. N. & other authors (2010). Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces. *Journal of Virology* **84**, 1674.

Li, L. L., Shan, T. L., Soji, O. B., Alam, M. M., Kunz, T. H., Zaidi, S. Z. & Delwart, E. (2011). Possible cross-species transmission of circoviruses and cycloviruses among farm animals. *Journal of General Virology* **92**, 768-772.

Londono, A., Riego-Ruiz, L. & Arguello-Astorga, G. R. (2010). DNA-binding specificity determinants of replication proteins encoded by eukaryotic ssDNA viruses are adjacent to widely separated RCR conserved motifs. *Archives of Virology* **155**, 1033-1046.

Londoño, A., Riego-Ruiz, L. & Argüello-Astorga, G. R. (2010). DNA-binding specificity determinants of replication proteins encoded by eukaryotic ssDNA viruses are adjacent to widely separated RCR conserved motifs. *Archives of virology* **155**, 1033-1046.

López-Bueno, A., Tamames, J., Velázquez, D., Moya, A., Quesada, A. & Alcamí, A. (2009). High diversity of the viral community from an Antarctic lake. *Science* **326**, 858-861.

Muhire, B., Martin, D. P., Brown, J. K., Navas-Castillo, J., Moriones, E., Zerbini, F. M., Rivera-Bustamante, R., Malathi, V. G., Briddon, R. W. & other authors (2013). A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). *Archives of Virology* **158**, 1411-1424.

Notredame, C., Higgins, D. G. & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **302**, 205-217.

Padilla-Rodriguez, M., Rosario, K. & Breitbart, M. (2013). Novel cyclovirus discovered in the Florida woods cockroach Eurycotis floridana (Walker). *Archives of virology* **158**, 1389-1392.

Pham, H. T., Bergoin, M. & Tijssen, P. (2013a). Acheta domesticus Volvovirus, a Novel Single-Stranded Circular DNA Virus of the House Cricket. *Genome announcements* **1**, e0007913.

Pham, H. T., Iwao, H., Bergoin, M. & Tijssen, P. (2013b). New Volvovirus Isolates from Acheta domesticus (Japan) and Gryllus assimilis (United States). *Genome announcements* **1**, e00328-00313.

Rosario, K., Duffy, S. & Breitbart, M. (2009a). Diverse circovirus-like genome architectures revealed by environmental metagenomics. *Journal of General Virology* **90**, 2418-2424.

Rosario, K., Duffy, S. & Breitbart, M. (2012a). A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Archives of virology* **185**, 1851-1871.

Rosario, K., Nilsson, C., Lim, Y. W., Ruan, Y. & Breitbart, M. (2009b). Metagenomic analysis of viruses in reclaimed water. *Environmental Microbiology* **11**, 2806-2820.

Rosario, K., Padilla-Rodriguez, M., Kraberger, S., Stainton, D., Martin, D. P., Breitbart, M. & Varsani, A. (2013). Discovery of a novel mastrevirus and alphasatellite-like circular DNA in dragonflies (Epiprocta) from Puerto Rico. *Virus Research* **171**, 231-237.

Rosario, K., Dayaram, A., Marinov, M., Ware, J., Kraberger, S., Stainton, D., Breitbart, M. & Varsani, A. (2012b). Diverse circular ssDNA viruses discovered in dragonflies (Odonata: Epiprocta). *Journal of General Virology* **93**, 2668-2681.

Rosario, K., Marinov, M., Stainton, D., Kraberger, S., Wiltshire, E. J., Collings, D. A., Walters, M., Martin, D. P., Breitbart, M. & other authors (2011). Dragonfly cyclovirus, a novel single-stranded DNA virus discovered in dragonflies (Odonata: Anisoptera). *Journal of General Virology* **92**, 1302-1308.

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., Colombet, J., Sime-Ngando, T. & Debroas, D. (2012). Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PloS one* **7**, e33641.

Rowe, R. (1987). *The Dragonflies of New Zealand.* Auckland, New Zealand: Auckland University Press.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. & Birol, İ. (2009). ABySS: a parallel assembler for short read sequence data. *Genome research* **19**, 1117-1123.

Smits, S. L., Zijlstra, E. E., van Hellemond, J. J., Schapendonk, C. M., Bodewes, R., Schurch, A. C., Haagmans, B. L. & Osterhaus, A. D. (2013). Novel cyclovirus in human cerebrospinal fluid, Malawi, 2010-2011. *Emerging infectious diseases* **19**.

Tan, L. V., van Doorn, H. R., Nghia, H. D. T., Chau, T. T. H., Tu, L. T. P., de Vries, M., Canuti, M., Deijs, M., Jebbink, M. F. & other authors (2013). Identification of a New Cyclovirus in Cerebrospinal Fluid of Patients with Acute Central Nervous System Infections. *Mbio* **4**.

Chapter 4

Identification of diverse circular Rep-encoding DNA viruses in adult dragonflies and damselflies (Insecta: Odonata) of Arizona and Oklahoma, USA

Contents

This work has been published in Infection, Genetics and Evolution and is presented in a similar manner to that of the publication:

Dayaram, A., Potter, K., Pailes, R., Moline, A. B., Marinov, M., Rosenstein, D. D. & Varsani, A. (2014) Identification of diverse circular Rep-encoding DNA viruses in dragonflies and damselflies of Arizona and Oklahoma. *Infection, Genetics and Evolution* (In press).

4.1 Abstract

Next generation sequencing and metagenomic approaches are commonly used for the identification of circular replication-associated protein (Rep) - encoding single-stranded (CRESS) DNA viruses circulating in various environments. These approaches have enabled the discovery of some CRESS DNA viruses associated with insects. In this study we identified and recovered 31 viral genomes which represent 24 distinct CRESS DNA viruses from seven dragonfly species (*Rhionaeschna multicolor*, *Erythemis simplicicollis*, *Erythrodiplax fusca*, *Libellula quadrimaculata*, *Libellula saturata*, *Pachydiplax longipennis, and Pantala hymenaea*) and two damselfly species (*Ischnura posita*, *Ischnura ramburii*) sampled in various locations in the states of Arizona and Oklahoma of the United States of America (USA). We also identified Sclerotinia sclerotiorum hypovirulence-associated DNA virus 1 (SsHADV-1) in *Pantala hymenaea*, *Erythemis simplicicollis*, and *Ischnura ramburii* sampled in Oklahoma, which is the first report of SsHADV-1 in the New World. The genome architectures of the CRESS DNA viruses recovered vary, but they all have at least two major open reading frames (ORFs) that have either a bidirectional or unidirectional arrangement. Four of the viral genomes recovered, in addition to the three isolates of SsHADV-1, show similarities to viruses of the gemycircularvirus group. Analysis of the Rep encoded by the remaining 24 viral genomes reveals that these are highly diverse and allude to the fact that they represent novel CRESS DNA viruses.

4.2 Introduction

Chapter 4 introduces metagenomic techniques for isolating circular replication associated protein (Rep) - encoding ssDNA (CRESS DNA) viruses from adult dragonfly and damselfly samples. The phenomenal diversity of single stranded DNA (ssDNA) viruses is starting to emerge as a result of the low cost and increased use of next generation sequencing-based metagenomic approaches. This technology has facilitated the rapid exploration and discovery of ssDNA viruses in a range of environmental samples (soil, water, sediment, faecal, and atmospheric), and in vertebrate and invertebrate specimens. These studies have helped to provide insights into the diversity of ssDNA viruses. Many of the novel CRESS DNA viruses being discovered show little similarity to the currently described viral families and have different genome architectures, and in many cases the host and transmission pathways are yet to be identified.

Most of the novel CRESS DNA viruses discovered have two major open reading frames (ORFs); one of these ORFs typically displays some degree of sequence similarity to previously described Reps from circoviruses, nanoviruses and geminiviruses. However, as a result of the development of a relatively large dataset of CRESS DNA viruses that have specific similarities to circoviruses, a proposal for a new genus (cyclovirus) in the family *Circoviridae* is being considered by the International Committee for the Taxonomy of Viruses (ICTV). Both cycloviruses and circoviruses encode two major ORFs that are bi-directionally transcribed; however, in the cycloviruses the capsid protein (CP) is encoded in the virion sense and the Rep in the complementary sense. The first cycloviruses to be described were isolated from human and chimpanzee stool samples from Pakistan, Tunisia and Nigeria (Li *et al.*, 2010). Subsequently, viruses that fall within this proposed genus have been isolated from a variety of samples including stools from a range of different animals (Ge *et al.*, 2012; Li *et al.*, 2011; Li *et al.*, 2010), human tissue (Phan *et al.*, 2014; Smits *et al.*, 2013; Van Tan *et al.*, 2013), and insect tissue including cockroaches (Padilla-Rodriguez *et al.*, 2013), dragonflies, and damselflies (Dayaram *et al.*, 2013; Rosario *et al.*, 2012b; Rosario *et al.*, 2011).

Insects are of interest in ssDNA virus research because many viruses rely on insect vectors for transmission. Many studies are using metagenomics to identify ssDNA viruses in various insect species and determine the roles these insects play in virus transmission (Ng *et al.*, 2011a; Rosario *et al.*, 2014; Rosario *et al.*, 2011). Among insects that have been investigated, dragonflies are turning out to be useful viral sampling tools in an ecosystem, because they are mobile, top-end predators that can accumulate many different viruses, including pathogenic plant-infecting viruses (Ng *et al.*, 2011b; Rosario & Breitbart, 2011; Rosario *et al.*, 2014; Rosario *et al.*, 2013; Rosario *et al.*, 2012b). For example, viruses and satellite molecules that infect plants, such as alphasatellite DNA molecules and mastreviruses (family *Geminiviridae*) which are transmitted by insects, were isolated from dragonfly samples collected in Puerto Rico (Rosario *et al.*, 2013).

Circular ssDNA viruses other than geminiviruses and nanoviruses were unknown in insects until the identification of a cyclovirus in dragonflies (Odonata:Anisoptera) from the Kingdom of Tonga (Rosario *et al.*, 2011). Subsequently, cycloviruses were found in adult dragonflies collected in Bulgaria; Florida, USA; and Puerto Rico. Other novel CRESS DNA viruses were also isolated from these samples, including viruses that fall within a proposed new group called gemycircularvirus (Dayaram *et al.*, 2012; Du *et al.*, 2014; Kraberger *et al.*, 2013; Ng *et al.*, 2011a; Rosario *et al.*, 2012b; Sikorski *et al.*, 2013b; van den Brand *et al.*, 2011). The Reps of these gemycircularviruses are most closely related to those of plant-infecting geminiviruses, however, they have slightly smaller genomes and no movement protein genes have been identified in their genomes. A subsequent study on larvae of dragonflies *Procordulia grayi* (Selys, 1871) and damselflies *Xanthocnemis zealandica* (McLachlan, 1873) from high country lakes in the South Island of New Zealand identified 13 viral genomes with differing genome architectures, 10 of which were novel CRESS DNA viruses (Chapter 2).

We continue to investigate the presence of other novel CRESS DNA viruses that may be circulating in dragonfly and damselfly populations. In this study we characterise 31 CRESS DNA viruses from nine different dragonfly and damselfly species collected from seven different locations from the states of Oklahoma and Arizona, USA. Many of the viral genomes recovered are extremely divergent from previously identified novel CRESS DNA viruses, adding to the rapidly increasing CRESS DNA virus database.

4.3 Materials and methods

4.3.1 Sample collection, viral particle purification and circular DNA enrichment

Specimens of *Rhionaeschna multicolor* (Hagen, 1861), *Erythemis simplicicollis* (Say, 1840), *Erythrodiplax fusca* (Rambur, 1842), *Libellula quadrimaculata* (Linnaeus, 1780), *Libellula saturata* (Uhler, 1857), *Pachydiplax longipennis* (Burmeister, 1839) and *Pantala hymenaea* (Say, 1840) and six adult damselfly specimens of *Ischnura posita* (Hagen, 1861) and *Ischnura ramburii* (Selys and Sagra, 1857) were collected using insect nets from four different locations in Oklahoma, and three different locations in Arizona, USA (Table 4.1).

The dragonfly and damselfly specimens were preserved in 95% ethanol after collection, and subsequently air dried. Dragonflies and damselflies were then grouped according to location and species.

The abdomens from each specimen in the assigned species groups were then dissected in a sterile environment. The abdomens for each species group were combined and manually homogenised in SM buffer (0.1 M NaCl, 50 mM Tris/HCl, pH 7.4, 10 mM $MgSO_4$) at a ratio of 2ml SM buffer to 1g of tissue as previously described (Rosario *et al.*, 2011). The homogenate was then centrifuged (10,000 *x g* for 10 min) in order to pellet the remaining tissue debris. The supernatant was sequentially filtered first through a 0.45 µm and then a 0.2 µm pore size syringe filter (Sartorius Stedim Biotech, Germany). Viral DNA was then extracted from the filtrate using a High Pure Viral Nucleic Acid Kit (Roche, USA).

Circular DNA in the grouped viral DNA extract was then enriched by rolling circle amplification using TempliPhi (GE Healthcare) as described previously (Dayaram *et al.*, 2014; Dayaram *et al.*, 2012; Dayaram *et al.*, 2013; Rosario *et al.*, 2012b; Rosario *et al.*, 2011; Sikorski *et al.*, 2013a).

4.3.2 Illumina sequencing and NGS data assembly

All enriched RCA products from Arizona and Oklahoma dragonfly and damselfly specimens were grouped together and sequenced on an Illumina HiSeq 2000 (Illumina, USA) platform at the Beijing Genomics Institute (Hong Kong).

Table 4.1: Sampling locations, specimens, viral isolates recovered and conserved motifs identified in the Rep and the putative nonanucleotide sequences.

					Rep motifs					
Accession #	Isolate	Sampling location in USA	Insect species	Putative nonanucleotide sequence	I	II	III	Walker A	Walker B	Motif C
KM598383	SsHADV-1 [US-549DFS-12]	Sutton, Oklahoma	*Erythemis simplicicollis*	TAATATTAT	VLLTY	HLHCFAEF	YAIKDGDVI	GPSQTG-KT	VFDDIRGG	IWCSN
KM598382	SsHADV-1 [US-549LB-12]	Lows Pond, Oklahoma	*Ischnura ramburii*	TAATATTAT	VLLTY	HLHCFAEF	YAIKDGDVI	GPSQTG-KT	VFDDIRGG	IWCSN
KM598384	SsHADV-1 [US-549SR-12]	South Canadian River, Oklahoma	*Pantala hymenaea*	TAATATTAT	VLLTY	HLHCFAEF	YAIKDGDVI	GPSQTG-KT	VFDDIRGG	IWCSN
KM598385	DfasCV-4 [US-260BC-12]	Bishop Creek, Oklahoma	*Ischnura posita*	TAATATTAT	FLITY	HLHVFVDF	YAIKDGEVV	GVHGSG-KT	YFDDIRGG	IWISN
KM598386	DfasCV-4 [US-260SR1-12]	South Canadian River, Oklahoma	*Pantala hymenaea*	TAATATTAT	FLITY	HLHVFVDF	YAIKDGEVV	GVHGSG-KT	YFDDIRGG	IWISN
KM598388	DfasCV-5 [US-1634LM2-12]	Lower Lake Mary, Arizona	*Libellula saturata*	TAATATTAT	VLLTY	HFHVFVDF	YAAKDGDIV	GPTQYG-QS	VFDDWKGG	IWLCN
KM598387	DfasCV-5 [US-1642KW-12]	Kachina Wetlands, Arizona	*Rhionaeschna multicolor*	TAATATTAT	VLLTY	HFHVFVDF	YAAKDGDIV	GPTQYG-KS	VFDDWKGG	IWLCN
KM598393	OdasCV-1 [US-504LB-12]	Lows Pond, Oklahoma	*Ischnura ramburii*	TAGTATTAC	WLLTL	HIQGYIHT	YAGKEETSV	---------	ILEENPDL	RQYEN
KM598399	OdasCV-2 [US-364BC-12]	Bishop Creek, Oklahoma	*Ischnura posita*	TAATATTAT	FLLTY	HLHSACHY	YCRKHGNFI	GPSNSG-KT	WIDEYKGC	VILSN
KM598407	OdasCV-3 [US-221LB1-12]	Lows Pond, Oklahoma	*Ischnura ramburii*	CAGTATTAC	WCFTI	HLQGFVKT	YCKKDGDWE	GDPGVG-KS	HFEDVQKG	YISSN
KM598408	OdasCV-4 [US-517BC-12]	Bishop Creek, Oklahoma	*Ischnura posita*	TAGTATTAC	YIFTI	HIQGAIRF	YCTKLETRE	GPTGTG-KT	LLDEFYGW	WITSN
KM598410	OdasCV-5 [US-1683LM1-12]	Lower Lake Mary, Arizona	*Libellula quadrimaculata*	TAATGGTTG	PYYTW	HFQIRVVF	YVEKEGHFW	DRYGNSGKT*	TNDEPDRN	IKWDS
KM598389	OdasCV-6 [US-1642LM1-12]	Lower Lake Mary, Arizona	*Libellula quadrimaculata*	TAATATTAC	VFLTY	HIHAYAAW	YCGKHDTEA	GASKLG-KT	VLDDFNIK	IWLCN
KM598390	OdasCV-7 [US-1706LM1-12]	Lower Lake Mary, Arizona	*Libellula quadrimaculata*	TATTTATAG	VFLTY	HFHAVLCF	YVKKDGDFL	GPSGCG-KT	IFDDMTFN	VFTCN
KM598391	OdasCV-8 [US-1739LM1-12]	Lower Lake Mary, Arizona	*Libellula quadrimaculata*	TAATATTAC	FFLTY	HLHVYMQL	YVMKDGNFK	GIPDTG-KS	IYNDPQTV	IFICN
KM598392	OdasCV-9 [US-466DFS-12]	Sutton, Oklahoma	*Erythemis simplicicollis*	TAATATTAC	YVLTI	HKQAYIVL	YISKKNTQQ	GCTGTG-KS	IIDDLNPD	IITTR
KM598412	OdasCV-10 [US-1675LM1-12]	Lower Lake Mary, Arizona	*Libellula quadrimaculata*	TTCTATTAC	VCFTF	HLQGYIQC	YCKKDGIWQ	EDKGNVGKS*	TNEEYLDG	VFVFA
KM598394	OdasCV-11 [US-341DFS-12]	Sutton, Oklahoma	*Erythemis simplicicollis*	TAGTATTC	WCFTL	HLQGAIYY	YCSKEKVLV	GAPRTG-KT	LIDDFDPQ	ILTSN
KM598395	OdasCV-12 [US-1518LM1-12]	Lower Lake Mary, Arizona	*Libellula quadrimaculata*	CAGTATTAC	WLGTI	HWQVFVIT	YVWKEATRA	GPTGTG-KS	VIDEFRGG	WITSN
KM598396	OdasCV-13 [US-1591LM1-12]	Lower Lake Mary, Arizona	*Libellula quadrimaculata*	TAGTATTAC	YCFTS	HLQGFVAF	YCKKDGDYK	GPPRSG-KD	HLSDMDKN	VVTSN
KM598397	OdasCV-14 [US-1577SC3-12]	South Canadian River, Oklahoma	*Erythrodiplax fusca*	TAATATTAC	WFITI	HVHALVIF	YVTKDGRII	GPPGTG-KS	LMEDLDPK	IVTSH
KM598398	OdasCV-15 [US-1640LM1-12]	Lower Lake Mary, Arizona	*Libellula quadrimaculata*	TAATATTAC	FFLTY	HLHALVTY	YISKEDIEP	GPSGIG-KT	IFDDMSFQ	IFLSN
KM598411	OdasCV-16 [US-1614LM1-12]	Lower Lake Mary, Arizona	*Libellula quadrimaculata*	TACTATTAC	WCFTW	HLQGYFEF	YCSKSGNYF	DLKGQNGKT*	IINLIYNN	GLISN
KM598400	OdasCV-17 [US-1619LM1-12]	Lower Lake Mary, Arizona	*Libellula quadrimaculata*	TAGTATTAC	YQFTL	HMHVYAHF	YIRKGGDII	GGTGKG-KT	VIEEFRPS	IICSI
KM598401	OdasCV-18 [US-1735LM1-12]	Lower Lake Mary, Arizona	*Libellula quadrimaculata*	CTATATTAC	FCFTK	HLQGYFEF	YCSKDGSFF	GPTGSG-KS	IMDDFRPS	IYVTT
KM598402	OdasCV-18 [US-1736LM1-12]	Lower Lake Mary, Arizona	*Libellula quadrimaculata*	CTATATTAC	FCFTK	HLQGYFEF	YCSKDGSFF	GPTGSG-KS	IMDDFRPS	IYVTT
KM598403	OdasCV-18 [US-1736LM2-12]	Lower Lake Mary, Arizona	*Libellula saturata*	CTCTATTAC	FCFTR	HLQGYFEF	YCSKDGSFF	GPTGSG-KS	IMDDFRPS	IYVTT
KM598404	OdasCV-19 [US-1594LM1-12]	Lower Lake Mary, Arizona	*Libellula quadrimaculata*	TAATATAGC	FVFTS	HLQGYCEL	YCQKEGNWK	GDSGSG-KS	IIDDFRDS	IITSL
KM598405	OdasCV-19 [US-1604SC1-12]	Stage cross Rd, Arizona	*Pachydiplax longipennis*	TAATATAGC	FVFTS	HLQGYCEL	YCQKEGNWK	GDSGSG-KS	IIDDFRDS	IITSL
KM598406	OdasCV-20 [US-718DFS-12]	Sutton, Oklahoma	*Erythemis simplicicollis*	TAGTATTAC	WCFTL	HLQGYLYW	YCSKDKIIA	GPSGTG-KT	IIDDFRGS	SITSI
KM598409	OdasCV-21 [US-1679SC3-12]	Stage Cross Rd, Arizona	*Erythrodiplax fusca*	TATTACCTT	YIGTI	HYQFCMDC	YCRKTGDYR	DTEGKAGKS*	IVDIPRDQ	VFTNH

*Putative Walker A motif

Table 4.2: Details of primer sequences used to recover viral genomes in this study.

CRESS DNA virus	Forward primer	Reverse primer
SsHADV-1 [US-549DFS-12]	GATTGTCTGAGGCATGTTGTTGACGTTCTC	AATCGCCAACGTAACATTTTGTTTAAGGGGGC
SsHADV-1 [US-549LB-12]	GATTGTCTGAGGCATGTTGTTGACGTTCTC	AATCGCCAACGTAACATTTTGTTTAAGGGGGC
SsHADV-1 [US-549SR-12]	GATTGTCTGAGGCATGTTGTTGACGTTCTC	AATCGCCAACGTAACATTTTGTTTAAGGGGGC
DfasCV-4 [US-260BC-12]	CTGTGTCCGAAAGTCCTGGGCGAAAC	AATGCCAAGCTTCTGCACCAATGCACGAC
DfasCV-4 [US-260SR1-12]	CTGTGTCCGAAAGTCCTGGGCGAAAC	AATGCCAAGCTTCTGCACCAATGCACGAC
DfasCV-5 [US-1634LM2-12]	TGCGCCGGCTTCGTCGTCATCATAAAC	GCTGAAGTACCGGCATACTACAGTACAGATG
DfasCV-5 [US-1642KW-12]	TGCGCCGGCTTCGTCGTCATCATAAAC	GCTGAAGTACCGGCATACTACAGTACAGATG
OdasCV-1 [US-504LB-12]	GGTAAAGAGATTCAAGTTCCAGGACGAGATAGG	CATTTGGGCAGATTCTTGCACCGTTCCC
OdasCV-2 [US-364BC-12]	GCGTTTACAGAGCCTTGGGGAAATCAAGTG	GCGAGTACTTCTTCCTTAGAAGCTTGGCATTG
OdasCV-3 [US-221LB1-12]	ATGTCGGACTCCCATGTATACACATGCTCG	GGGAAACTTTTTTGACGGCTATCGCGGTCA
OdasCV-4 [US-517BC-12]	GGTCAGGTTCCGCACTGCAAGTCTG	TAGTCGGTCAGGATTCCTATAAGTCAATTAGGATCATG
OdasCV-5 [US-1683LM1-12]	CAGACGGGGAAGATAGTTGATAGTCTGCG	TTCATGGGCGTGTGTCTGCTCCCTC
OdasCV-6 [US-1642LM1-12]	TCCACGTCAAAGCATCCAGCTCCGAC	CGGACACCATCCTAACATACAGAAGCCG
OdasCV-7 [US-1706LM1-12]	GCGCATTAGTTCCATACTCCAATCTAAACTTCTTTAG	CTGCGCAGTTACAGGATCAGTATAATCTGAATTATTGTC
OdasCV-8 [US-1739LM1-12]	GAAGAAGGCTTTGCCTGTAGCTCTGTATCC	CTTACCTACCCAAAATGCGATATCAAACGCGATG
OdasCV-9 [US-466DFS-12]	GCAGTGTAACAAATGGTGGAATGATTACGATGGAG	TTGATATAGGCATCTGGGTACAATTGCCTCGC
OdasCV-10 [US-1675LM1-12]	CTGCCGACCACTTCTTCATGTTCGGC	AGATCGTTATAACATCATCGATCTTAACCGTCCTTG
OdasCV-11 [US-341DFS-12]	GGTAACTGTCAACAGAGTAGAGAATATTGCTCTAAAG	TCTCATAGCCTCCCAGTGGATTCTAGG
OdasCV-12 [US-1518LM1-12]	GGCCTCTTTCCATACATATTCTTCGGCG	ACTAGAGCCGGAGATCAATTCCAGTTTGGA
OdasCV-13 [US-1591LM1-12]	CCGGTACGTCTTCCCCCTTCGGT	ACATGGCGCCGAAAGCGGGGC
OdasCV-14 [US-1577SC3-12]	AAACTTTGGCAAAAGGCGCTTACTTCCTC	CAACTCATCGTTACCTCAAACTATTCCCTTAG
OdasCV-15 [US-1640LM1-12]	GTTCTATCTGAGCTGTTCTTGGCAGATGTTGG	TAGTGGACCGTTACCATCCTCAGCAGATTC
OdasCV-16 [US-1614LM1-12]	CGGCATTCTTGGTTGTGAGGTAGTACTG	CCATGGTCTATAATCTTGGTAGATCTGAAGATCTAG
OdasCV-17 [US-1619LM1-12]	GAAATTGGCGATGAACCTCACCAAGGTGC	ATCAATGATATCACCACCCTTACGGATGTAAGC
OdasCV-18 [US-1735LM1-12]	GCGTATGTGAGTTAATCAAATCTGGCGCAAC	TGTCGAGGTCAGTACGTTTACCCTGAC
OdasCV-18 [US-1736LM1-12]	CCTTCATGTAGGAAGATGGGTACTGATCCC	AATGCAGCCACAAGTGGTGGGATGGTTAC
OdasCV-18 [US-1736LM2-12]	CCTTCATGTAGGAAGATGGGTACTGATCCC	AATGCAGCCACAAGTGGTGGGATGGTTAC
OdasCV-19 [US-1594LM1-12]	CAGTATCGCCCTTAGCTTGCGCTTC	CGTATCAGATTCTCCGCCGGATAGAC
OdasCV-19 [US-1604SC1-12]	TTATAGTACAGTTGGTAGCCTCGACCTTCAG	AACAAGAGGCAATTGGTGGGACAATTACAAAGG
OdasCV-20 [US-718DFS-12]	GGAAGCAGCATAGCATTCACAGATCTTCTCAG	ACGGAAATCGTCGATGATAACTTCTTTATGCCCG
OdasCV-21 [US-1679SC3-12]	CCTGTCTATTAACACCAGTTTCCCCAGCC	AAAACGGTATGACCAACGGACACCCTTACAACTG

The paired end reads were then *de novo* assembled using ABySS V1.3.5 (Simpson *et al.*, 2009) with kmer = 64. A BLASTx (Altschul *et al.*, 1990) analysis was performed on the *de novo* assembled contigs >500 nts using KoriBlast v4.1 (Korilog SARL, Bioinformatics Solutions France) to identify contigs with ORFs encoding known viral-like proteins.

4.3.3 Viral genome verification

Back-to-back (abutting) primers were designed (Table 4.2) based on the *de novo* assembled contig sequences with viral protein BLASTx hits to those encoded by CRESS DNA viruses and known circular ssDNA viruses in order to recover and verify the full genomes. The viral genome sequences were amplified with specific primers (Table 4.2) using Kapa HiFi HotStart polymerase (Kapa Biosystems, USA). The amplicons were resolved on agarose gels then gel purified and cloned into pJET1.2 vector (Thermo Fisher, USA). The resulting plasmid clones were Sanger sequenced at Macrogen Inc. (South Korea) by primer walking.

The Sanger sequencing derived contigs were assembled using DNA Baser Sequence Assembler (Version 4.16; Heracle BioSoft S.R.L., Romania). Preliminary analysis of the sequences was carried out using BLASTx and tBLASTx (Altschul *et al.*, 1990). The full viral genomes were then annotated. Major ORFs encoding for putative Rep and CP were identified as well as stem-loop structures with conserved nonanucleotide motifs.

4.3.4 Analysis of recovered genomes

A data set of Rep protein sequences from all characterised ssDNA viruses available on GenBank (downloaded 20th August 2014; Additional Table 4.1 and 4.2) was created for analysis. These Rep sequences were aligned using MUSCLE (Edgar, 2004) with manual editing. The alignment was then used to create a phylogenetic tree using a JTT + CAT model with approximate likelihood-ratio test (aLRT) branch support (Anisimova & Gascuel, 2006) using FastTree version 2.1.7 (Price *et al.*, 2010). Branches with less than 80% aLRT support were collapsed using Mesquite (version 2.75) and the Rep phylogenetic tree was mid-point rooted.

A subset of the sequences (n=7) had similarities to gemycircularviruses (Rosario *et al.*, 2012b; Sikorski *et al.*, 2013b). The Reps of these and those encoded by other gemycircularvirus sequences together with those of representative geminiviruses were aligned using MUSCLE (Edgar, 2004) and analyzed using ProTest (Darriba *et al.*, 2011) to determine the best fit model of substitution. The maximum likelihood phylogenetic tree was inferred using PhyML version 3.0 (Guindon *et al.*, 2010) using the rtREV+G model with aLRT (Anisimova & Gascuel, 2006) for the gemycircularvirus-like sequences. The resulting tree was rooted with geminivirus Rep sequences.

The Reps and CPs of viral sequences that are similar to pig-stool-associated circular ssDNA virus (PisaCV), bovine stool associated circular virus (BoSCV), porcine-stool-associated circular virus (PoSCV), turkey-stool-associated circular virus (TuSCV) and chimpanzee-stool-associated circular virus (ChiSV) were aligned using MUSCLE (Edgar, 2004). For the purpose of this study we have named these chipoviruses (chimpanzee and porcine viruses; though this group includes all of the above). ProTest was used to determine the best fit model and the phylogenetic trees were inferred using PhyML version 3.0 (Guindon *et al.*, 2010) for both the Rep and CP protein datasets (Guindon *et al.*, 2010). The Rep tree was created using the WAG+G+I model with aLRT branch support and was rooted with the Rep sequence of McMurdo Ice Shelf pond-associated circular virus 8 (MpaCDV-8) (KJ547653; (Zawar-Reza *et al.*, 2014)). The CP tree was inferred using the reREV+G model with aLRT branch support and was midpoint rooted.

All pairwise identities were calculated using SDT v1.2 (Muhire *et al.*, 2014). A further BLASTp (Altschul *et al.*, 1990) analysis was carried out on all Rep-like proteins from the recovered viral genomes against the NCBI non-redundant protein database.

4.4 Results and Discussion

4.4.1 Recovery and characterisation of viral genomes

All the recovered circular DNA viral genomes (n=31) have two major ORFs, however, most of them had differing genome architecture (Figure 4.1). The major ORFs were either uni-directionally or bi-directionally oriented. The size of the viral genomes ranged from ~1,600 nts to 3,200 nts (Figure 4.1). All of the genomes have at least one intergenic region (IR), with 23 of the genomes having both small (SIR) and large intergenic regions (LIR). In all of the genomes the putative stem loop structures were identified including the region tentatively identified as the *ori*. The putative nonanucleotide motifs identified for each genome are listed in Table 4.1.

Of the 31 recovered CRESS DNA viruses, three are Sclerotinia sclerotiorum hypovirulence-associated DNA virus 1 (SsHADV-1). These were recovered from *E. simplicicollis, P. hymenaea* and *I. ramburii*, all sampled in Oklahoma. They share >99% genome-wide identity with previously described SsHADV-1 isolates recovered from plant pathogenic fungi *Sclerotinia sclerotiorum* in China (Yu *et al.*, 2010) and benthic river sediments from New Zealand (Kraberger *et al.*, 2013). An additional four viral genomes have similarities to gemycircularviruses and represent two distinct viruses. We have tentatively named these Dragonfly associated circular virus -4 and -5 (DfasCV-4; DfasCV-5). DfasCV-4 (KM598385) and DfasCV-4 (KM598386) were recovered from species *I. posita* and *P. hymenaea* sampled at Bishop Creek and the South Canadian River in Oklahoma respectively. Their genomes share >99.5% genome-wide pairwise identity. DfasCV-5 (KM598388) and DfasCV-5 (KM598387) were recovered from *L. saturata* and *R. multicolour* sampled at Lower Lake Mary and Kachina Wetlands in Arizona respectively; these two genomes share >99.1% genome-wide pairwise identity.

The remaining 24 CRESS DNA viral genomes represented 21 distinct CRESS DNA viruses based on their Rep analysis, and we have tentatively named these as Odonata associated circular DNA virus 1 to 21 (OdasCV-1 to -21). These were recovered from six dragonfly species (*R. multicolor*, *E. simplicicollis*, *E. fusca*, *L. quadrimaculata*, *L. saturata*, *P. longipennis*) and two damselfly species (*I. posita*, *I. ramburii*) (Table 4.1).

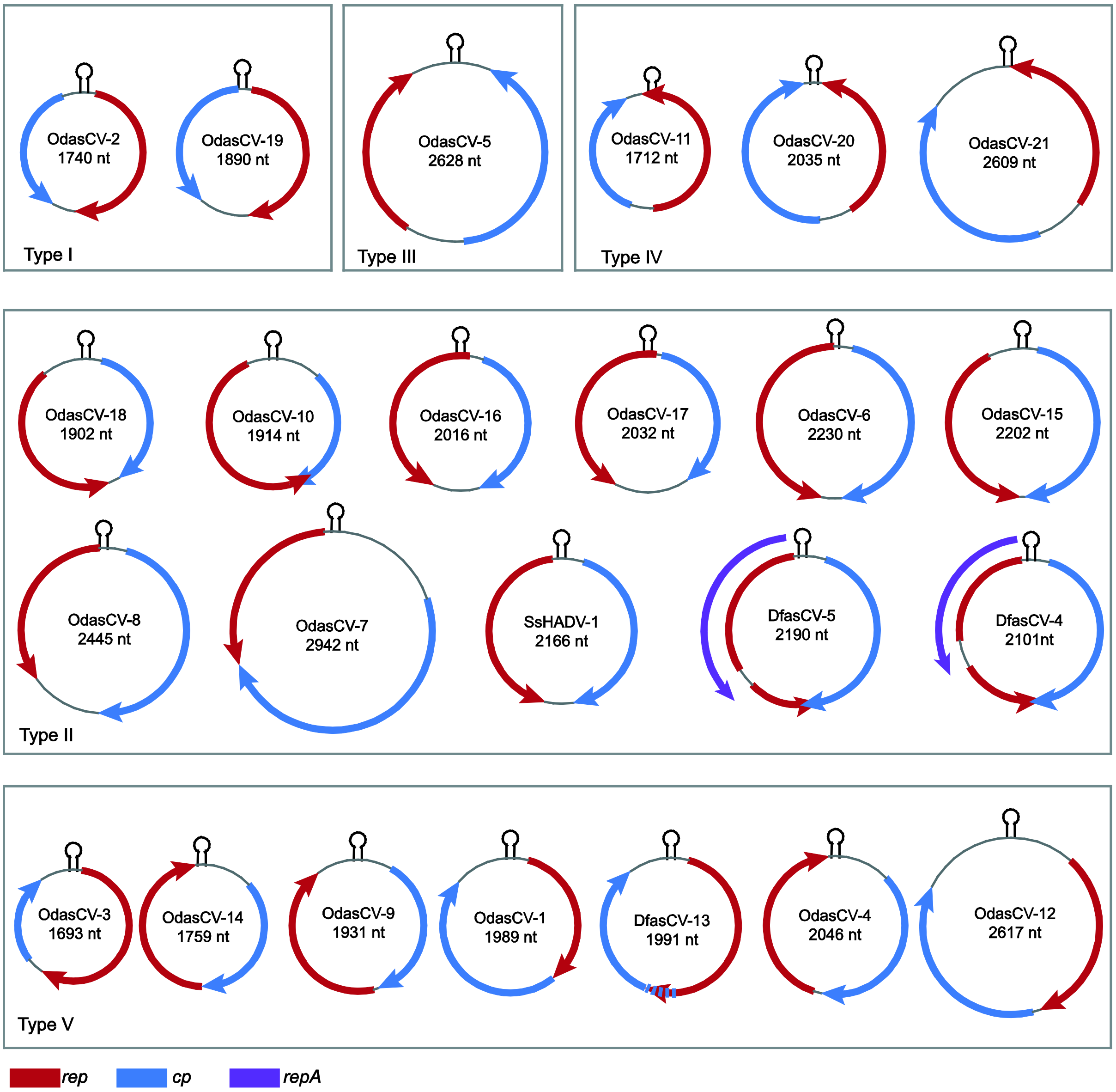


Figure 4.1: Genome organisations of the CRESS DNA viruses recovered from nine different Odonata (dragonfly and damselfly) species.

OdasCV-18 has three isolates (OdasCV-18 [KM598401], OdasCV-18 [KM598402], OdasCV-18 [KM598403]) all sampled at Lower Lake Mary from *Libellula* sp. The two OdasCV-18 isolates recovered from *L. quadrimaculata* share 99% genome-wide pairwise identity, however, these two share 82% genome-wide pairwise identity with the isolate from *L. saturata*. This is the second example of viruses recovered from dragonflies that could be the same viral species with a marked genomic diversity of 18%. The first example was isolates of Dragonfly cyclovirus - 4 (DfCyV-4; n=3) recovered from *Rhionaeschna multicolour* and *Somatochlora meridionalis* (Nielsen 1935) from Bulgaria and USA respectively, which have a genomic diversity of 15% (Rosario *et al.*, 2012b). On the other hand DfCyV-1 (n=24) recovered from *P. flavescens*, *Diplacodes bipunctata* and *Tholymis tillarga* sampled in the Kingdom of Tonga (Rosario *et al.*, 2012b; Rosario *et al.*, 2011) have a genomic diversity of 5%.

We identified 12 distinct CRESS DNA viruses (OdasCV-5,-6, -7, -8, -10, -12, -13, -15, -17, -18,-19) in *L. quadrimaculate* and *L. saturate*, all sampled at Lower Lake Mary in Arizona (Table 4.1), while four distinct CRESS DNA viruses (SsHADV-1; OdasCV-9, -11, -20) were recovered from the dragonfly species *E. simplicicollis* sampled in Sutton in Oklahoma (Table 4.1).

4.4.2 Analysis of the replication associated protein

All CRESS DNA viruses identified to date have some degree of conservation in their Reps (Rosario *et al.*, 2012a). Rolling circle replication (RCR) motifs I, II and III as well as the superfamily (SF3) helicase motifs which include the Walker-A, Walker-B and motif C were identified in all but one of the putative Reps encoded by the viral genomes recovered in this study (Table 4.1). In the Rep of OdasCV-1 we failed to identify the putative Walker A motif (Table 4.1). The exact function of RCR motif I is yet to be determined; however, it is currently believed to be involved with recognition of tandem sequence repeats known as iterons that are located in the IR (Argüello-Astorga & Ruiz-Medrano, 2001). RCR motif II has two histidine residues that are important for coordinating the binding of metal ions. Both RCR motif II and RCR motif III are involved during replication initiation and are the catalytic site of DNA cleavage (Heyraud-Nitschke *et al.*, 1995; Steinfeldt *et al.*, 2006).

Table 4.3: Summary of the BLASTp analysis of the Rep proteins.

Sequence	Accession #	CRESS DNA virus hit	% pairwise Rep amino acid identity	E-value	GenBank reference #
OdasCV-12	KM598395	SaCV-6	58%	7×10^{-105}	AIF34816
OdasCV-14	KM598397	DfLaCV-3	38%	2×10^{-41}	AHH31467
OdasCV-13	KM598396	DfCirV	44%	2×10^{-90}	YP_009021241
OdasCV-19	KM598404	DfCirV	47%	6×10^{-69}	YP_009021245
OdasCV-19	KM598405	DfCirV	47%	6×10^{-69}	YP_009021246
OdasCV-16	KM598411	SI00142	34%	2×10^{-34}	AGA18387
OdasCV-17	KM598400	TM-6C	31%	5×10^{-15}	ADI48253
DfasCV-5	KM598387	FaGmV-1b	42%	4×10^{-76}	AGU67671
OdasCV-15	KM598398	MSV	46%	6×10^{-8}	ACZ04195
DfasCV-5	KM598388	FaGmV-1b	42%	3×10^{-77}	AGU67673
OdasCV-6	KM598389	SaCV-3	48%	3×10^{-94}	AIF34810
OdasCV-10	KM598412	SOG04070	35%	8×10^{-45}	AGA18265
OdasCV-21	KM598409	BoSCV	48%	1×10^{-77}	AEW47007
OdasCV-5	KM598410	PigSCV	34%	9×10^{-21}	AFV77589
OdasCV-7	KM598390	NepaV	34%	5×10^{-49}	YP_009021041
OdasCV-18	KM598401	YN-BtCV-2	42%	1×10^{-52}	AEL87786
OdasCV-18	KM598402	YN-BtCV-2	42%	2×10^{-51}	AEL87786
OdasCV-18	KM598403	YN-BtCV-2	39%	2×10^{-51}	AEL87786
OdasCV-8	KM598391	EuMV	36%	4×10^{-15}	ACJ02762
OdasCV-3	KM598407	LDMD-11	41%	4×10^{-46}	AGS36210
DfasCV-4	KM598385	DfasCV-2	73%	2×10^{-165}	YP_009021856
DfasCV-4	KM598386	DfasCV-2	72%	1×10^{-164}	YP_009021856
OdasCV-11	KM598394	SOG00160	37%	2×10^{-43}	AGA18245
OdasCV-2	KM598399	SaCV-4	27%	9×10^{-16}	AIF34812
OdasCV-9	KM598392	SI00197	40%	9×10^{-50}	AGA18388
OdasCV-1	KM598393	GuCV	42%	1×10^{-9}	AFJ93342
OdasCV-4	KM598408	BFDV	38%	3×10^{-51}	AFM55150
SsHADV-1	KM598382	SsHADV-1	100%	0	AGP05335
SsHADV-1	KM598383	SsHADV-1	100%	0	AGP05335
SsHADV-1	KM598384	SsHADV-1	100%	0	AGP05335
OdasCV-20	KM598406	SOG03994	53%	5×10^{-85}	AGA18263

The SF3 helicase motifs all share a nucleotidetriphosphate (NTP) binding mechanism recognised by small DNA and RNA viral Rep helicases (Gorbalenya *et al.*, 1990; Walker *et al.*, 1982). The Walker-A motif is a structural motif that forms part of a P-loop that is thought to act as a deoxyribonucleotide triphosphate (dNTP) binding domain and may also exhibit helicase activity during RCR (Rosario *et al.*, 2012a), whilst the Walker-B and motif C are thought to regulate helicase activity through the dNTP and P-loop nucleoside-triphosphate (NTPase) domains (Hickman & Dyda, 2005).

The maximum likelihood phylogenetic tree of the Rep sequences reveals that these are diverse CRESS DNA viruses (Figure 4.2). It is striking to note that a significant number of the novel CRESS DNA viruses identified to date are from Odonata samples (Figures 2, 3, 4). The Reps of DfasCV-4 share ~ 51% and 40% pairwise identity with the Reps of SsHADV-1 and DfasCV-5 respectively. The Reps of OdasCV-18 share >91% pairwise identity. A BLASTp analysis of the Reps OdasCV reveals that they share ~27-58% pairwise identity with Reps of other CRESS DNA viral sequences available in public databases (Table 4.3). This, together with the Rep phylogenetic analysis, clearly demonstrates the extent of the diversity of the CRESS DNA viruses.

4.4.3 Gemycircularviruses

Seven of the viral genomes recovered from six different dragonfly and damselfly species (*I. posita*, *R. multicolor, L. saturata, I. ramburii, P. hymenaea* and *E. simplicicollis)* collected in Arizona and Oklahoma (Table 4.1) showed some similarity to previously described gemycircularviruses (Figure 4.3). The first gemycircularvirus to be described was SsHADV-1, which is a DNA mycovirus that infects *S. sclerotiorum* and confers hypovirulence (Yu *et al.*, 2010). This virus was subsequently identified in benthic sediments in New Zealand (Kraberger *et al.*, 2013). We have recovered three isolates of SsHADV-1 from *E. simplicicollis, I. ramburii*, *P. hymenaea* sampled in Oklahoma in this study, which is the first identification of SsHADV-1 in the USA. Other recently isolated gemycircularviruses include: Cassava associated circular virus (CasCV) and Hypericum japonicum-associated circular DNA virus (HJasCV) recovered from plant samples (Dayaram *et al.*, 2012; Du *et al.*, 2014); mosquito VEM SDBVL-G (MvemV) and DfasCV-1, -2 and -3 recovered from mosquitoes and dragonflies (Ng *et al.*, 2009; Rosario *et al.*, 2012b); faecal associated gemycircularviruses (FaGmCVs) (Sikorski *et al.*, 2013b) and Meles meles fecal virus (MmFV) from animal faecal

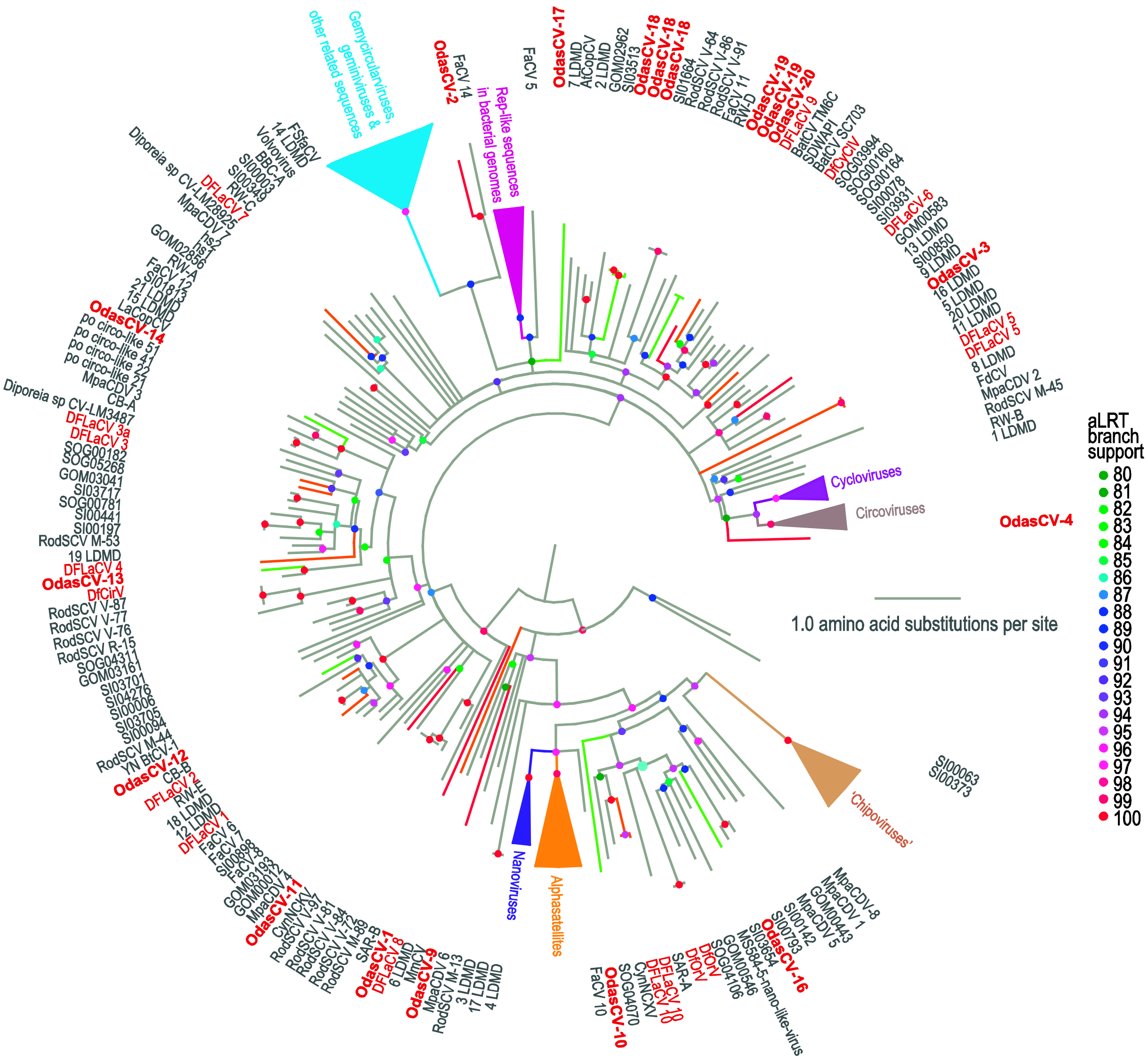


Figure 4.2: Maximum likelihood phylogenetic tree of the Rep sequences with aLRT branch support and mid-point rooted. All CRESS DNA viruses recovered from Odonata larvae and adults are highlighted in red and those from this study are in bold font. For the purpose of this study chipoviruses (**chi**mpanzee and **po**rcine **vi**ruses) are PisaCV, BoSVC, PoSCV, TuSCV and ChiSV. See Additional Table 4.3 for GenBank accession numbers associated with the acronyms used in the figure.

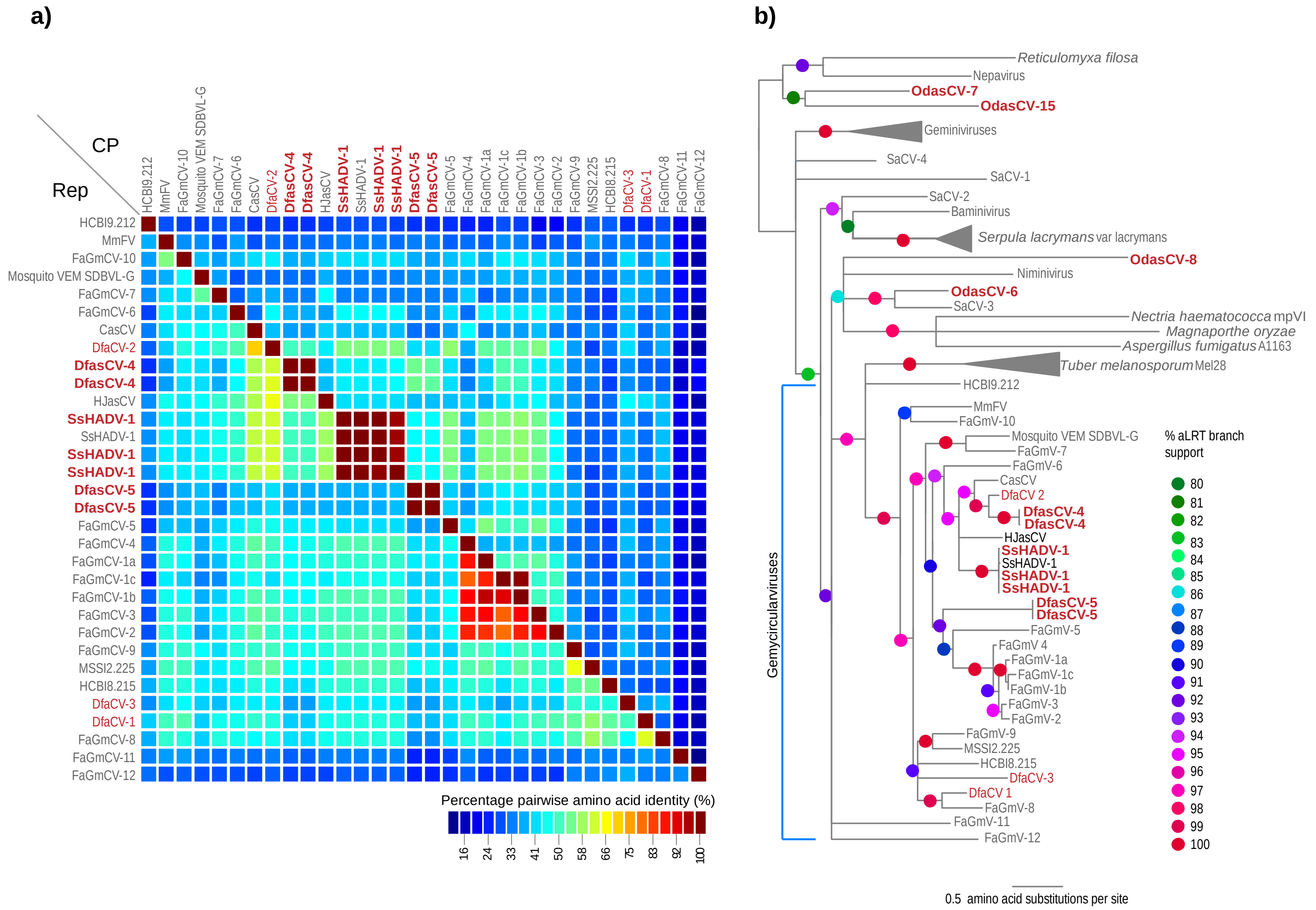


Figure 4.3: (a) Pairwise identity plot of amino acid of the Reps and CPs of gemycircularviruses (b) Maximum likelihood phylogenetic tree with aLRT branch support of the Rep sequences of gemycircularviruses and those closely related. All CRESS DNA viruses recovered from Odonata larvae and adults are highlighted in red and those from this study are in bold font.

matter (van den Brand *et al.*, 2011); and viral isolates HCB19.212, HCB18.215 and MSS12.225 from bovine serum (Lamberto *et al.*, 2014). Despite the broad sample range from gemycircularviruses have been recovered, SsHADV-1 is the only one for which a host (*S. sclerotiorum*) is known. The identification of Rep-like sequences most closely related to gemycircularviruses in fungal genomes provides some support that gemycircularviruses may in fact be mycoviruses. There have been previous reports of interactions of fungi with both insects and faecal matter, suggesting there could be an association (Aanen *et al.*, 2002; Davies *et al.*, 1993; Dowd, 1992; Hajek & St. Leger, 1994). All the viruses currently grouped as gemycircularviruses (all ~2200 nts) are bi-directionally transcribed and encode Reps in the complementary sense and CPs in the virion sense. It is worth noting that the gemycircularvirus Reps are most closely related to those of geminiviruses and all have an intron in their Reps with the exception of SsHADV-1. The Rep of gemycircularviruses also have a geminivirus-like Rep sequence motif (GRS; (Nash *et al.*, 2011)). Further, in general, the gemycircularviruses, including those identified in this study, have a highly conserved nonanucleotide motif ‘TAATATTAN’.

Analysis of the pairwise amino acid identity of both the Reps and CPs of gemycircularvirus-like sequences reveals that they share >60% identity. The Reps of the two DfaCV-4 isolates share ~73% pairwise identity with DfaCV-2, whereas the CPs shares ~51% pairwise identity (Figure 4.3). The Reps of DfaCV-4 cluster with CasCV and DfasCV-2. On the other hand DfasCV-5 is basal to a cluster of Reps of FaGmCV 1-5 (Figure 4.3). Overall the Rep is far more conserved than the CP in the gemycircularviruses. The Reps of OdasCV-6, -7, -8 and -15 are most closely related to gemycircularviruses, geminiviruses, and a few viruses recently recovered from faecal samples (Ng *et al.*, 2011a; Ng *et al.*, 2012).

4.4.4 CRESS DNA viruses with similarities to chipoviruses (PisaCV, BoSVC, PoSCV, TuSCV and ChiSV)

Two diverse CRESS DNA viruses, OdasCV-5 and -21, were recovered from *L. quadrimaculata* and *E. fusca* respectively in Arizona (Table 4.1). Phylogenetic analysis of both the Rep and CP of OdasCV-5 and -21 showed they cluster with those of chimpanzee-, bovine-, turkey- and porcine-stool-associated circular viruses (Figure 4.4), which for this study we have named chipoviruses. All chipoviruses have genomes of ~2500 nts. The

PisaCV -FUJ1, HUN1, GER2011, ANH1, HUN2, HUN1, JIANGX1, HEN1 and HUB2, all from porcine faecal samples, have uni-directionally organised ORFs, whereas the rest have bi-directionally organised ORFs (Additional Table 4.3). The genome organisations of these viruses also contain both SIR and LIR with some level of conservation observed in the nonanucleotide motif of the ori (Figure 4.1, Table 4.3).

OdaCV-5 and -21 share ~30% CP and 27% Rep pairwise identity. The Rep of OdaCV-5 shares ~32% pairwise identity with the Reps encoded by the unidirectional PisaCVs, however, the CP shares ~45% pairwise identity to CPs of the PisaSC-7 isolates. The CP and Rep of OdaCV-21 shares ~50% and 48% pairwise identity to the CP and Rep of BoSCV respectively. The phylogenetic analysis of the Reps and CPs of chipoviruses show clear clustering of the unidirectional PisaCV. The Reps of these unidirectional PisaCVs share >96% pairwise identity with each other, and ~77% to the Rep of SaCV-9 (Figure 4.4). Among the chipoviruses the Rep is more conserved than the CP; however, overall, conservation is seen in the CP of chipoviruses such that reasonably good protein sequence alignments allow generation of CP phylogeny with some confidence. Nonetheless, the hosts of these chipoviruses still need to be identified. The discovery of these viruses in various faecal sources suggests they infect the animals of the faecal source; however, the discovery of these viruses in insects suggests that these viruses may be associated with insects, which are a small portion of the chimpanzee diet (Blinkova *et al.*, 2010).

4.5 Concluding remarks

Thirty-one novel CRESS DNA viruses were isolated from a range of Odonata species collected from different sample sites in Arizona and Oklahoma, USA. The viral genomes recovered in this study have diverse genomes and genome architectures and include a range of CRESS DNA viruses that do not group with well-characterised ssDNA viral families. The addition of nine new isolates of gemycircularviruses from this study highlights that these related viruses are relatively common in nature. Gemycircularviruses have been isolated from a range of different sources suggesting that they are present in many environments. The hosts of most of the gemycircularviruses isolated to date, with the exception of SsHADV-1, are yet to be determined.

The discovery of two viral genomes that clustered closely with chipoviruses further indicates that this is an expanding group of viruses that are highly conserved within both the Rep and CP proteins. As for gemycircularviruses, the hosts of these viruses are also yet to be determined. The close association with animal faecal matter suggests they either infect the animal of the faecal source or result from a diet of infected plant and animal material.

The additional discovery of novel CRESS DNA viruses supports the concept that adult dragonflies can be used as a valuable tool for the identification of CRESS DNA viruses circulating in ecosystems.

GenBank accession numbers: KM598382 - KM598412

Additional Table 4.1: Details of sample groupings.

Location	Host	Group	Number of individuals
Bishop Creek, Oklahoma, USA	*Ischnura posita*	Group1	2 individuals
Kachima Wetlands, Arizona, USA	*Aeshna multicolor*	Group 2	1 individual
Lower Lake Mary, Arizona, USA	*Libellula quadrimaculate*	Group 3	10 individuals
Lower Lake Mary, Arizona, USA	*Libellula saturata*	Group 4	1 individual
Lows Pond, Oklahoma, USA	*Ischnura ramburii*	Group 5	2 individuals
Lows Pond, Oklahoma, USA	*Ischnura ramburii*	Group 5	2 individuals
South Canadian River, Oklahoma, USA	*Erythrodiplax fusca*	Group 6	1 individual
South Canadian River, Oklahoma, USA	*Pantala hymenaea*	Group 7	4 individuals
Stage Cross Rd, Arizona, USA	*Erythrodiplax fusca*	Group 8	1 individual
Stage Cross Rd, Arizona, USA	*Pachydiplax longipennis*	Group 9	3 individuals
Sutton, Oklahoma, USA	*Erythemis simplicicollis*	Group 10	3 individuals

Additional Table 4.2: List of CRESS DNA viral replication-associated proteins in Figure 4.1 and their accession number.

Acronym	Accession #	Acronym	Accession #	Acronym	Accession #
10-LDMD	KF133817	FaCV-6	KJ547630	RodSCV M-53	JF755415
11-LDMD	KF133818	FaCV-7	KJ547631	RodSCV M-89	JF755402
12-LDMD	KF133819	FaCV-8	KJ547632	RodSCV R-15	JF755401
13-LDMD	KF133820	FaCV-9	KJ547633	RodSCV V-64	JF755407
14-LDMD	KF246569	FaGmCV-10	KF371632	RodSCV V-69	JF755403
15-LDMD	KF133822	FaGmCV-11	KF371631	RodSCV V-72	JF755411
16-LDMD	KF133823	FaGmCV-12	KF371630	RodSCV V-76	JF755404
17-LDMD	KF133824	FaGmCV-1a	KF371643	RodSCV V-77	JF755405
18-LDMD	KF133825	FaGmCV-1b	KF371642	RodSCV V-84	JF755413
19-LDMD	KF133826	FaGmCV-1c	KF371641	RodSCV V-86	JF755416
1-LDMD	KF133807	FaGmCV-2	KF371640	RodSCV V-87	JF755406
20-LDMD	KF133827	FaGmCV-3	KF371639	RodSCV V-91	JF755417
21-LDMD	KF133828	FaGmCV-4	KF371638	RodSCV V-97	JF755414
2-LDMD	KF133808	FaGmCV-5	KF371637	RW-A	FJ959077
3-LDMD	KF133810	FaGmCV-6	KF371636	RW-B	FJ959078
4-LDMD	KF133811	FaGmCV-7	KF371635	RW-C	FJ959079
5-LDMD	KF133812	FaGmCV-8	KF371634	RW-D	FJ959080
6-LDMD	KF133813	FaGmCV-9	KF371633	RW-E	FJ959081
7-LDMD	KF133814	FdCV	KC441518	SAR-A	FJ959084
8-LDMD	KF133815	FSfaCV	JQ898332	SAR-B	FJ959085
9-LDMD	KF133816	GOM00012	JX904192	SDWAPI	HQ335042
BasCV-2	KM510191	GOM00182	JX904206	*Serpula lacrymans* var lacrymans S7 3	EGN98653
BasCV-3	KM510192	GOM00443	JX904231	*Serpula lacrymans* var lacrymans S7 3	EGN95344
batCV-SC703	JN857329	GOM00546	JX904245	*Serpula lacrymans* var lacrymans S7 9	EGO24257
BatCV-TM6C	HM228875	GOM00583	JX904250	*Serpula lacrymans* var lacrymans S7 9	EGO20879
BBC-A	FJ959086	GOM02856	JX904312	SI00003	JX904394
BOSVCCP11-49-3	JN634851	GOM02962	JX904333	SI00006	JX904395
CasCV	JQ412057	GOM03041	JX904344	SI00063	JX904401
CB-A	FJ959082	GOM03161	JX904368	SI00078	JX904407
CB-B	FJ959083	GOM03193	JX904377	SI00094	JX904412
ChiSCV-DP152	GQ351272	HCBI8.215	LK931483	SI00142	JX904416
ChiSCV-GM415	GQ351277	HCBI9.212	LK931484	SI00197	JX904420
ChiSCV-GM476	GQ351274	HJasCV	KF413620	SI00349	JX904427
ChiSCV-GM488	GQ351276	hs1	JX559621	SI00373	JX904431
ChiSCV-GM495	GQ351273	hs2	JX559622	SI00441	JX904439
ChiSCV-GM510	GQ351275	LaCopCV	JF912805	SI00793	JX904469
ChiSV-GT306	GQ351278	MmCV	JQ085285	SI00850	JX904473
CynNCKV	JX908740	MpaCDV-1	KJ547646	SI00898	JX904478
CynNCXV	JX908739	MpaCDV-3	KJ547648	SI01664	JX904518
DfaCV-1	JX185430	MpaCDV-4	KJ547649	SI01813	JX904523
DfaCV-2	JX185429	MpaCDV-5	KJ547650	SI03513	JX904541
DfaCV-3	JX185428	MpaCDV-6	KJ547651	SI03654	JX904548
DfCirV	JX185415	MpaCDV-7	KJ547652	SI03701	JX904559
DfCyClV	JX185418	MpaCDV-8	KJ547653	SI03705	JX904561
DFLaCV-1	KF738873	MSSI2.225	LK931485	SI03717	JX904562
DFLaCV-10	KF738884	Nepavirus	KJ547625	SI03931	JX904581
DFLaCV-10a	KF738885	PigSCV	JX274036	SI04276	JX904605
DFLaCV-2	KF738874	PisaCV-ANH1	JX305997	SOG00160	JX904075
DFLaCV-3	KF738875	PisaCV-FUJ1	JX305998	SOG00164	JX904076
DFLaCV-3a	KF738876	PisaCV-GER2011	JQ023166	SOG00182	JX904077
DFLaCV-4	KF738877	PisaCV-HEN1	JX305991	SOG00781	JX904107
DFLaCV-5	KF738878	PisaCV-HUB1	JX305992	SOG03994	JX904139
DFLaCV-5a	KF738879	PisaCV-HUB2	JX305993	SOG04070	JX904144
DFLaCV-6	KF738880	PisaCV-HUN1	JX305995	SOG04106	JX904147
DFLaCV-7	KF738881	PisaCV-HUN2	JX305996	SOG04311	JX904151
DFLaCV-8	KF738882	PisaCV-JIANGX1	JX305994	SOG05268	JX904185
DFLaCV-9	KF738883	po-circo-like21	JF713716	SsHADV-1	GQ365709
DfOrV	JX185416	po-circo-like22	JF713717	SsHADV-1	KF268025
DfOrV	JX185417	po-circo-like41	JF713718	SsHADV-1	KF268026
Diporeia sp CV-LM28925	KC248425	po-circo-like51	JF713719	SsHADV-1	KF268027
Diporeia sp CV-LM3487	KC248416	PoSCV2-f	KC545226	TuSCV	KF880727
FaCV-10	KJ547621	PoSCV3-3L7	KC545227	Volvovirus-IAF	KF133821
FaCV-11	KJ547622	PoSCV3-4L13	KC545228	YN-BtCV-1	JF938078
FaCV-12	KJ547623	PoSCV3-4L5	KC545229	*Tuber melanosporum* Mel28	295499105
FaCV-2	KJ547626	PoSCV3-L2T	KC545230	*Tuber melanosporum* Mel28	295503907
FaCV-3	KJ547627	RodSCV M-13	JF755410	*Tuber melanosporum* Mel28	295507048
FaCV-4	KJ547628	RodSCV M-44	JF755408	*Tuber melanosporum* Mel28	295507212
FaCV-5	KJ547629	RodSCV M-45	JF755409		

Additional Table 4.3: GenBank accession number, genome organisation, nonanucleotide sequence and genome length of chipoviruses.

Sequence	Accession #	Genome organisation	Nonanucleotide motif	Length (nts)
OdasCV-21	KM598409	Bi-directional	TATTACCTT	2609
OdasCV-5	KM598410	Bi-directional	TAATGGTTG	2625
ChiSCV-DP152	GQ351272	Bi-directional	AATAATTAC	2609
ChiSCV-GM495	GQ351273	Bi-directional	AATAGTTAC	2640
ChiSCV-GM476	GQ351274	Bi-directional	AATAGTTAC	2637
ChiSCV-GM488	GQ351276	Bi-directional	AATAGTTAC	2638
ChiSCV-GM415	GQ351277	Bi-directional	AATAGTTAC	2639
ChiSCV-GM510	GQ351275	Bi-directional	AATAGTTAC	2589
BOSVCCP11 49 3	JN634851	Bi-directional	CAGTATTAC	2600
PisaCV-GER2011	JQ023166	Uni-directional	CAGGTCATT	2459
PigSCV	JX274036	Bi-directional	TAGATTACC	2589
PisaCV-HEN1	JX305991	Uni-directional	TAGGTCATT	2460
PisaCV-HUB1	JX305992	Uni-directional	TAGGTCATT	2460
PisaCV-HUB2	JX305993	Uni-directional	TAGGTCATT	2460
PisaCV-JIANGX1	JX305994	Uni-directional	TAGGTCATT	2460
PisaCV-HUN1	JX305995	Uni-directional	TAGGTCATT	2460
PisaCV-HUN2	JX305996	Uni-directional	TAGGTCATT	2460
PisaCV-ANH1	JX305997	Uni-directional	TAGGTCATT	2460
PisaCV-FUJ1	JX305998	Uni-directional	CAGGTCATT	2460
PoSCV-2 f	KC545226	Bi-directional	TAGTATTAC	2539
PoSCV-3 3L7	KC545227	Bi-directional	TAGTATTAC	2495
PoSCV-3 4L13	KC545228	Bi-directional	TAGTATTAC	2495
PoSCV-3 4L5	KC545229	Bi-directional	TAGTATTAC	2494
PoSCV-3 L2T	KC545230	Bi-directional	TAGTATTAC	2502
PoSCV-Kor J481	KF193403	Bi-directional	TAGATTACC	2589
TuSCV	KF880727	Bi-directional	TAGTGTTAC	2479
SaCV-9	KJ547633	Uni-directional	ATGCTACCC	2423
PoSCV-1 DP2	KJ577810	Bi-directional	TAGTGTTAC	2403
PoSCV-1 DP3	KJ577811	Bi-directional	TAGTGTTAC	2590
PoSCV-7 EP2-A	KJ577812	Bi-directional	TAGTGTTAC	2601
PoSCV-7 EP2-B	KJ577813	Bi-directional	TAGTGTTAC	2601
PoSCV-7 EP3-C	KJ577814	Bi-directional	TAGTGTTAC	2631
PoSCV-7 EP3-D	KJ577815	Bi-directional	TAGTGTTAC	2596
PoSCV-8 GP2	KJ577817	Bi-directional	CAGTGTTAC	2477
PoSCV-6 XP1	KJ577819	Bi-directional	CAGTGTTAC	2603
PoSCV-9 FP1	KJ577816	Bi-directional	TAGTATTAC	2564
PoSCV-2 TP3	KJ577818	Bi-directional	TAGTATTAC	2487

References

Aanen, D. K., Eggleton, P., Rouland-Lefevre, C., Guldberg-Frøslev, T., Rosendahl, S. & Boomsma, J. J. (2002). The evolution of fungus-growing termites and their mutualistic fungal symbionts. *Proceedings of the National Academy of Sciences* **99**, 14887-14892.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.

Anisimova, M. & Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic biology* **55**, 539-552.

Argüello-Astorga, G. & Ruiz-Medrano, R. (2001). 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. *Archives of virology* **146**, 1465-1485.

Blinkova, O., Victoria, J., Li, Y., Keele, B. F., Sanz, C., Ndjango, J. B. N., Peeters, M., Travis, D., Lonsdorf, E. V. & other authors (2010). Novel circular DNA viruses in stool samples of wild-living chimpanzees. *Journal of General Virology* **91**, 74-86.

Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164-1165.

Davies, D. R., Theodorou, M. K., Lawrence, M. I. & Trinci, A. P. (1993). Distribution of anaerobic fungi in the digestive tract of cattle and their survival in faeces. *Journal of general microbiology* **139**, 1395-1400.

Dayaram, A., Galatowitsch, M., Harding, J. S., Argüello-Astorga, G. R. & Varsani, A. (2014). Novel circular DNA viruses identifiedin *Procordulia grayi* and *Xanthocnemis zealandica* larvae using metagenomic approaches. *Infection, Genetics and Evolution*, 134-141.

Dayaram, A., Opong, A., Jäschke, A., Hadfield, J., Baschiera, M., Dobson, R. C. J., Offei, S. K., Shepherd, D. N., Martin, D. P. & other authors (2012). Molecular characterisation of a novel cassava associated circular ssDNA virus. *Virus Research* **166**, 130-135.

Dayaram, A., Potter, K. A., Moline, A. B., Rosenstein, D. D., Marinov, M., Thomas, J. E., Beitbart, M., Rosario, K., Argüello-Astorga, G. R. & other authors (2013). High global diversity of cycloviruses amongst dragonflies. *Journal of General Virology* **94**, 1827-1840.

Dowd, P. F. (1992). Insect fungal symbionts: a promising source of detoxifying enzymes. *Journal of industrial microbiology* **9**, 149-161.

Du, Z., Tang, Y., Zhang, S., She, X., Lan, G., Varsani, A. & He, Z. (2014). Identification and molecular characterization of a single-stranded circular DNA virus with similarities to Sclerotinia sclerotiorum hypovirulence-associated DNA virus 1. *Archives of virology* **159**, 1527-1531.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797.

Ge, X., Li, Y., Yang, X., Zhang, H., Zhou, P., Zhang, Y. & Shi, Z. (2012). Metagenomic Analysis of Viruses from the Bat Fecal Samples Reveals Many Novel Viruses in Insectivorous Bats in China. *Journal of virology* **86**, 4620-4630.

Gorbalenya, A. E., Koonin, E. V. & Wolf, Y. I. (1990). A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses. *FEBS letters* **262**, 145-148.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology* **59**, 307-321.

Hajek, A. & St. Leger, R. (1994). Interactions between fungal pathogens and insect hosts. *Annual review of entomology* **39**, 293-322.

Heyraud-Nitschke, F., Schumacher, S., Laufs, J., Schaefer, S., Schell, J. & Gronenborn, B. (1995). Determination of the origin cleavage and joining domain of geminivirus Rep proteins. *Nucleic Acids Research* **23**, 910.

Hickman, A. B. & Dyda, F. (2005). Binding and unwinding: SF3 viral helicases. *Current opinion in structural biology* **15**, 77-85.

Kraberger, S., Stainton, D., Dayaram, A., Zawar-Reza, P., Gomez, C., Harding, J. S. & Varsani, A. (2013). Discovery of Sclerotinia sclerotiorum hypovirulence-associated virus-1 in urban river sediments of Heathcote and Styx rivers in Christchurch city, New Zealand. *Genome announcements* **1**, e00559-00513.

Lamberto, I., Gunst, K., Müller, H., zur Hausen, H. & de Villiers, E.-M. (2014). Mycovirus-like DNA virus sequences from cattle serum and human brain and serum samples from multiple sclerosis patients. *Genome announcements* **2**, e00848-00814.

Li, L., Shan, T., Wang, C., Côté, C., Kolman, J., Onions, D., Gulland, F. M. & Delwart, E. (2011). The fecal viral flora of California sea lions. *Journal of virology* **85**, 9909-9917.

Li, L., Kapoor, A., Slikas, B., Bamidele, O. S., Wang, C., Shaukat, S., Masroor, M. A., Wilson, M. L., Ndjango, J. B. N. & other authors (2010). Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces. *Journal of Virology* **84**, 1674.

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

Nash, T. E., Dallas, M. B., Reyes, M. I., Buhrman, G. K., Ascencio-Ibanez, J. & Hanley-Bowdoin, L. (2011). Functional Analysis of a Novel Motif Conserved across Geminivirus Rep Proteins. *Journal of Virology* **85**, 1182.

Ng, T. F. F., Manire, C., Borrowman, K., Langer, T., Ehrhart, L. & Breitbart, M. (2009). Discovery of a novel single-stranded DNA virus from a sea turtle fibropapilloma by using viral metagenomics. *Journal of virology* **83**, 2500-2509.

Ng, T. F. F., Duffy, S., Polston, J. E., Bixby, E., Vallad, G. E. & Breitbart, M. (2011a). Exploring the Diversity of Plant DNA Viruses and Their Satellites Using Vector-Enabled Metagenomics on Whiteflies. *PloS one* **6**, e19050.

Ng, T. F. F., Marine, R., Wang, C., Simmonds, P., Kapusinszky, B., Bodhidatta, L., Oderinde, B. S., Wommack, K. E. & Delwart, E. (2012). High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. *Journal of virology* **86**, 12161-12175.

Ng, T. F. F., Willner, D. L., Lim, Y. W., Schmieder, R., Chau, B., Nilsson, C., Anthony, S., Ruan, Y., Rohwer, F. & other authors (2011b). Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. *PloS one* **6**, e20579.

Padilla-Rodriguez, M., Rosario, K. & Breitbart, M. (2013). Novel cyclovirus discovered in the Florida woods cockroach *Eurycotis floridana* (Walker). *Archives of virology* **158**, 1389-1392.

Phan, T. G., Luchsinger, V., Avendaño, L. F., Deng, X. & Delwart, E. (2014). Cyclovirus in nasopharyngeal aspirates of Chilean children with respiratory infections. *Journal of General Virology* **95**, 922-927.

Price, M. N., Dehal, P. S. & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one* **5**, e9490.

Rosario, K. & Breitbart, M. (2011). Exploring the viral world through metagenomics. *Current Opinion in Virology* **1**, 289-297.

Rosario, K., Duffy, S. & Breitbart, M. (2012a). A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Archives of virology* **157**, 1851-1871.

Rosario, K., Capobianco, H., Ng, T. F. F., Breitbart, M. & Polston, J. E. (2014). RNA Viral Metagenome of Whiteflies Leads to the Discovery and Characterization of a Whitefly-Transmitted Carlavirus in North America. *PloS one* **9**, e86748.

Rosario, K., Padilla-Rodriguez, M., Kraberger, S., Stainton, D., Martin, D. P., Breitbart, M. & Varsani, A. (2013). Discovery of a novel mastrevirus and alphasatellite-like circular DNA in dragonflies (Epiprocta) from Puerto Rico. *Virus Research*, 231-237.

Rosario, K., Dayaram, A., Marinov, M., Ware, J., Kraberger, S., Stainton, D., Breitbart, M. & Varsani, A. (2012b). Diverse circular ssDNA viruses discovered in dragonflies (Odonata: Epiprocta). *Journal of General Virology* **93**, 2668-2681.

Rosario, K., Marinov, M., Stainton, D., Kraberger, S., Wiltshire, E. J., Collings, D. A., Walters, M., Martin, D. P., Breitbart, M. & other authors (2011). Dragonfly cyclovirus, a novel single-stranded DNA virus discovered in dragonflies (Odonata: Anisoptera). *Journal of General Virology* **92**, 1302-1308.

Sikorski, A., Argüello-Astorga, G. R., Dayaram, A., Dobson, R. C. J. & Varsani, A. (2013a). Discovery of a novel circular single-stranded DNA virus from porcine faeces. *Archives of virology* **158**, 283-289.

Sikorski, A., Massaro, M., Kraberger, S., Young, L. M., Smalley, D., Martin, D. P. & Varsani, A. (2013b). Novel myco-like DNA viruses discovered in the faecal matter of various animals. *Virus Research* **177**, 209-216.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. & Birol, İ. (2009). ABySS: a parallel assembler for short read sequence data. *Genome research* **19**, 1117-1123.

Smits, S. L., Zijlstra, E., van Hellemond, J. J., Schapendonk, C. M., Bodewes, R., Schürch, A. C., Haagmans, B. L. & Osterhaus, A. D. (2013). Novel cyclovirus in human cerebrospinal fluid, Malawi, 2010–2011. *Emerging infectious diseases* **19**, 1511.

Steinfeldt, T., Finsterbusch, T. & Mankertz, A. (2006). Demonstration of nicking/joining activity at the origin of DNA replication associated with the rep and rep′ proteins of porcine circovirus type 1. *Journal of virology* **80**, 6225-6234.

van den Brand, J. M. A., van Leeuwen, M., Schapendonk, C. M., Simon, J. H., Haagmans, B. L., Osterhaus, A. D. M. E. & Smits, S. L. (2011). Metagenomic analysis of the viral flora of pine marten and European badger feces. *Journal of Virology* **86**, 2360-2365.

Van Tan, L., van Doorn, H. R., Nghia, H. D. T., Chau, T. T. H., de Vries, M., Canuti, M., Deijs, M., Jebbink, M. F., Baker, S. & other authors (2013). Identification of a new cyclovirus in cerebrospinal fluid of patients with acute central nervous system infections. *MBio* **4**, e00231-00213.

Walker, J. E., Saraste, M., Runswick, M. J. & Gay, N. J. (1982). 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. *The EMBO journal* **1**, 945.

Yu, X., Li, B., Fu, Y., Jiang, D., Ghabrial, S. A., Li, G., Peng, Y., Xie, J., Cheng, J. & other authors (2010). A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. *Proceedings of the National Academy of Sciences* **107**, 8387.

Zawar-Reza, P., Argüello-Astorga, G. R., Kraberger, S., Julian, L., Stainton, D., Broady, P. A. & Varsani, A. (2014). Diverse small circular single-stranded DNA viruses identified in a freshwater pond on the McMurdo Ice Shelf (Antarctica). *Infection, Genetics and Evolution* **26**, 132-138.

Chapter 5

Novel single-stranded DNA virus recovered from estuarine mollusc (*Amphibola crenata)* whose replication-associated protein (Rep) shares similarities with Rep-like sequences of bacterial origin

Contents

This work has been published in Journal of General Virology and is presented in a similar manner to that of the publication:

Dayaram, A., Goldstien, S., Zawar-Reza, P., Gomez, C., Harding, J. S. & Varsani, A. (2013) Novel single stranded DNA virus recovered from estuarine Mollusc (Amphibola crenata) whose replication associated protein (Rep) shares similarities with Rep-like sequences of bacterial origin. *Journal of General Virology* 94, 1104-1110.

5.1 Abstract

Over the last couple of years, highly diverse novel single-stranded DNA (ssDNA) viruses have been discovered. Here we present the first ssDNA virus, Gastropod associated circular ssDNA virus (GaCSV), recovered from the mollusc *Amphibola crenata* (Martyn, 1784), which grazes microorganisms and organic detritus on the surface of tidal mudflats. The GaCSV (2351 nt) genome contains two large bidirectionally transcribed open-reading frames (ORFs). The smaller ORF (874 nt) has similarities to viral replication-associated protein (Rep) sequences of some bacteria and circoviruses, whereas the larger ORF (955 nt) does not share similarities to any sequences in public databases and we presume it potentially encodes the capsid protein (CP). Phylogenetic analysis shows that the Rep of GaCSV clusters with Rep-like sequences of bacterial origin highlighting the potential role of ssDNA viruses in horizontal gene transfer. The occurrence of unknown viruses in organisms associated with human pollution is a relatively unexplored field.

5.2 Introduction

Chapter 5 further explores single stranded DNA (ssDNA) viral diversity using mollusc samples as a potential surveillance tool for the identification of novel circular Rep-encoding single-stranded (CRESS) DNA viruses. The extent of ssDNA virus diversity in ecosystems has been grossly underestimated. However, over the past four years, through various metagenomics and new molecular methods, a wealth of previously unknown ssDNA viruses have been discovered and subsequently characterised. In particular, the combination of next-generation sequencing technologies, shotgun sequencing and sequence independent rolling circle amplification (RCA) have played a key role in the identification of novel ssDNA viruses. Non-invasive faecal sampling of chimpanzees (Blinkova *et al.*, 2010; Yu *et al.*, 2010a), pigs (Allan & Ellis, 2000; Shan *et al.*, 2011; Sikorski *et al.*, 2013; Yu *et al.*, 2010a), badgers, pine martens (van den Brand et al., 2011), rodents (Phan et al., 2011), dogs (Kapoor et al., 2012) and cows (Kim et al., 2011) has resulted in the identification of novel ssDNA viruses. Several research groups have extended their ssDNA virus metagenomic endeavours to sampling including insects (Ng *et al.*, 2011; Rosario *et al.*, 2012b; Rosario *et al.*, 2011), using a vector enabled metagenomic approach.

In this chapter we describe a novel circular ssDNA virus recovered from the estuarine mollusc *Amphibola crenata* (Martyn, 1784) (family, Amphibolidae), collected at the hydrologically low-energy estuarine wetlands at the confluence between the Heathcote River and the Avon-Heathcote estuary. *A. crenata* are deposit feeders that graze on microorganisms and organic detritus on the surface of tidal mudflats (Juniper, 1986). Molluscs, known for their bio-magnifying capabilities are often used as indicator organisms in bioaccumulation studies to monitor pollutants such as heavy metals (Baršienė *et al.*, 2002; Gunkel & Streit, 1980; Walsh *et al.*, 1994). In addition, most viruses that have been identified in molluscs are a threat to the aquaculture industry, as most currently known mollusc viruses are pathogenic (Cheng *et al.*, 2005; Comps, 1988; Farley, 1976; Meyers *et al.*, 2009a; Morley, 2010). It is important to note that between September 2010 and July 2011, sediment and sewage were deposited into the Avon-Heathcote Estuary (Christchurch, New Zealand) as a result of >8000 earthquakes which damaged sewage and storm water infrastructure in the surrounding city of Christchurch. Any relationship between the occurrence of this virus and human faecal pollution is currently unknown.

5.3 Materials and methods

5.3.1 Processing, viral particle purification and DNA extraction

Between September 2010 and July 2011, twenty *Amphibola crenata* were collected from estuarine wetlands at the confluence of the Heathcote River (Lat: -43.557778; Lon: 172.705833) and the Avon-Heathcote Estuary (Christchurch, New Zealand). The *A. crenata* were de-shelled and homogenised together in SM buffer (0.1 M NaCl, 50 mM Tris/HCl, pH 7.4, 10 mM $MgSO_4$) at a ratio of 10 ml buffer to 5 g of tissue sample. The homogenate was centrifuged (10,000 rpm for 10 min) to pellet cellular debris. The supernatant was subsequently filtered through a 0.45µm syringe filter (Sartorius Stedim Biotech, Germany) and the resulting filtrate was subjected to a second filtration using a 0.2µm syringe filter (Sartorius Stedim Biotech, Germany). Viral DNA was then extracted from the filtrate using a High Pure Viral Nucleic Acid Kit (Roche, USA).

5.3.2 Enrichment of circular DNA

Circular viral DNA was enriched non-specifically by rolling circle amplification (RCA) using the Illustra TempliPhi Amplification Kit (GE Healthcare, USA) as described previously by (Dayaram *et al.*, 2012; Rosario *et al.*, 2012b; Rosario *et al.*, 2012c; Rosario *et al.*, 2011; Sikorski *et al.*, 2013). The concatenated rolling circle amplified product was then restricted with *Bam*H1 yielding a ~2.3 kb DNA fragment. This fragment was gel purified and cloned into *Bam*H1-restricted pUC19 plasmid (Fermentas, USA) and sequenced by Macrogen Inc. (South Korea) using primer walking. Preliminary analysis of the ~2.3 kb DNA sequence was carried out using BLASTx and tBLASTx (Altschul *et al.*, 1990).

5.3.3 Genome verification

In order to confirm that we had obtained a complete genome of the novel ssDNA virus, back-to-back primers GNZ-F 5'-CAC CGC ACC TAC AGG G-3'; GNZ-R 5'- TAC CCT TCT TGC CCA CTT C-3' (targeted in the Rep) were designed to amplify the full circular genome. The novel viral genome from the RCA enriched material (1µl) was used as a template for PCR amplification using primer pair GNZ-F/ GNZ-R with Kapa HiFi™ Hotstart polymerase

(Kapa Biosystems, USA) with the following protocol: initial denaturation at 95°C for 2 mins followed by 25 cycles of 98°C for 20 sec, 50°C for 30 sec, 72°C for 2 mins, a final elongation at 72°C for 5 min and a final renaturation at 4°C for 10 mins. The resulting ~2.3kb amplicon was gel purified and cloned into pJET1.2 (Fermentas, USA) plasmid and sequenced at Macrogen Inc. (South Korea) by primer walking. The resulting sequence contigs were then assembled using DNAMAN (version 5.2.9; Lynnon Biosoft).

5.4 Results and discussion

5.4.1 Viral genome analysis

The assembled complete genome of the ssDNA virus was found to be 2351 nt, exactly the same as that obtained by *Bam*HI restriction of the RCA product. The novel ssDNA viral genome has two large open reading frames (ORFs) and two smaller ORFs, two in the virion sense and two in the complementary sense (Figure 5.1). The two large open reading frames , which potentially are bi-directionally transcribed, are similar to the genome organisations of circoviruses, cycloviruses and novel CRESS-DNA viruses (Delwart & Li, 2012; Rosario *et al.*, 2012a; Rosario *et al.*, 2012c; Rosario *et al.*, 2011; Varsani *et al.*, 2011). Two intergenic regions were also identified; a long intergenic region (LIR) at 461 nucleotides and a short intergenic region (SIR) at 61 nucleotides in length (Figure 5.1). Within the LIR we identified the conserved nonanucleotide motif CAGTATTAC in the stem loop element (Figure 5.1). A BlastX (Altschul et al., 1990) analysis of the full genome indicates that the viral isolate was most closely related to the Reps found in certain bacteria (Figure 5.1 and Table 5.2), sharing 35% amino acid identity and 36% coverage (E-value $4x10^{-40}$), whilst only sharing 30% identity over 36% coverage (E-value $1x10^{-23}$). with circovirus BFDV.

5.4.2 ORF analysis

A NCBI BlastP of largest ORF (960 nt) putatively transcribed in the virion sense showed no homology to any sequences in the NCBI non-redundant protein database. However, BLASTp analysis of the putative protein encoded by the second largest ORF (876 nt) transcribed in the complementary sense DNA revealed that it shared some sequence similarity with the viral Reps of *Clostridium saccharolyticum* WM1 BlastP (GenBank accession number, NC_014376)**;** 97% coverage, maximum identity of 35%, E-value $4x10^{-43}$), amongst other bacteria including *Ruminococcus bromii* L2-63 (*Erysipelothrix rhusiopathiae* str. Fujisaw, *Clostridium citroniae* WAL-17108, *Roseburia inulinivorans* DSM 16841, *Eggerthella sp.* YY7918, *Oscillibacter valericigenes* Sjm18-20 and *Lachnospiraceae bacterium* (Figure 5.2). The BLASTp analysis also revealed some level of similarity to the Reps of BFDV (97% coverage, 30% identity, E-value $2x10^{-25}$). Hence this ORF, transcribed in the complementary sense, is putatively a Rep gene and we presume that the large ORF in the virion sense transcribes the capsid protein (CP).

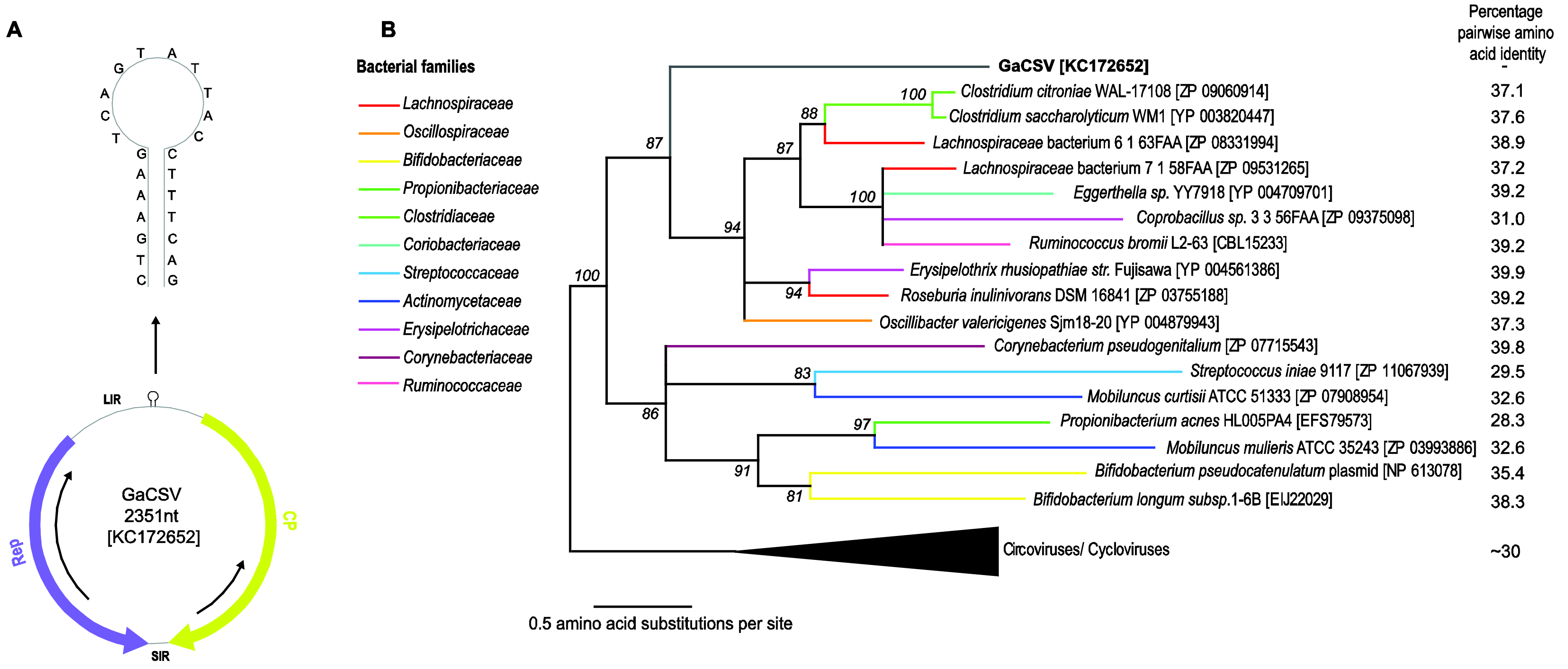

A
T A T G T A C A T C A C
G C
A T
A T
A T
G C
T A
C G
LIR
SIR
GaCSV
2351nt
[KC172652]
Rep
CP
B
Bacterial families
Lachnospiraceae
Oscillospiraceae
Bifidobacteriaceae
Propionibacteriaceae
Clostridiaceae
Coriobacteriaceae
Streptococcaceae
Actinomycetaceae
Erysipelotrichaceae
Corynebacteriaceae
Ruminococcaceae
Percentage pairwise amino acid identity
GaCSV [KC172652]
Clostridium citroniae WAL-17108 [ZP 09060914] 37.1
Clostridium saccharolyticum WM1 [YP 003820447] 37.6
Lachnospiraceae bacterium 6 1 63FAA [ZP 08331994] 38.9
Lachnospiraceae bacterium 7 1 58FAA [ZP 09531265] 37.2
Eggerthella sp. YY7918 [YP 004709701] 39.2
Coprobacillus sp. 3 3 56FAA [ZP 09375098] 31.0
Ruminococcus bromii L2-63 [CBL15233] 39.2
Erysipelothrix rhusiopathiae str. Fujisawa [YP 004561386] 39.9
Roseburia inulinivorans DSM 16841 [ZP 03755188] 39.2
Oscillibacter valericigenes Sjm18-20 [YP 004879943] 37.3
Corynebacterium pseudogenitalium [ZP 07715543] 39.8
Streptococcus iniae 9117 [ZP 11067939] 29.5
Mobiluncus curtisii ATCC 51333 [ZP 07908954] 32.6
Propionibacterium acnes HL005PA4 [EFS79573] 28.3
Mobiluncus mulieris ATCC 35243 [ZP 03993886] 32.6
Bifidobacterium pseudocatenulatum plasmid [NP 613078] 35.4
Bifidobacterium longum subsp.1-6B [EIJ22029] 38.3
Circoviruses/ Cycloviruses ~30
100
88
87
100
87
94
94
100
83
86
97
91
81
100
0.5 amino acid substitutions per site

Figure 5.1: (A) Genome organisation of GaCSV with stem loop element. (B) Maximum likelihood phylogenetic tree of the Rep protein of GaCSV and bacterial Rep-like sequences with aLRT branch support. Column on the right indicates amino acid pairwise identities of the Rep-like sequences to the Rep of GaCSV based pairwise alignment (using MUSCLE) of pairs of sequences with pairwise deletion of gaps.

Based on the low virus level of similarity of the Rep to other ssDNA viruses, we tentatively propose to name the novel isolate as Gastropod associated circular ssDNA virus (GaCSV). We found no significant hits in public databases for the two smaller ORFs.

The mechanisms of Reps of viruses and plasmids that replicate via rolling circle replication (RCR) method are highly conserved (Ilyina & Koonin, 1992; Koonin & Ilyina, 1992). As a result, conserved RCR and superfamily 3 (SF3) helicase motifs are found in all Reps of circular ssDNA viruses. Within the Rep of GaCSV (Table 5.1) we identified the three RCR conserved motifs (I: CLTLNN; II: HLQGFV; III: YIGLLNEG) and the three conserved SF3 helicase motifis (Walker-A: GPTGTGKT; Walker-B: VFEEF; Motif-C: ILSN) reviewed in Rosario et al. (Rosario *et al.*, 2012a). We also identified most of these conserved motifs in the Rep-like sequences of bacterial origin (Table 5.1). A pairwise comparison of the Rep-like proteins from the various bacterial species identified via BlastP analysis is presented in Figure 5.1. The comparison shows that GaCSV shares between 28.3 – 39.9% pairwise identity with the Rep-like sequences of bacterial origin and ~30% with BFDV (GenBank accession # AF071878).

5.4.3 Phylogenetic analysis of the Rep

A maximum phylogenetic tree was constructed using PHYML (version 3) (Guindon et al., 2010) with approximate likelihood-ratio test (aLRT) branch support (Anisimova & Gascuel, 2006), and LG model of the MUSCLE (Edgar, 2004) aligned Rep amino acid sequences found in bacteria, GaCSV and CRESS-DNA viruses. Branches with less than 60% aLRT support were collapsed using Mesquite (version 2.75). The ML phylogenetic analysis supports the BLASTp analysis; the Rep of GaCSV clusters with all the Rep-like sequences found in the bacterial species outlined in Figure 5.1. Recently our research group identified a suite of novel ssDNA viruses (Dayaram *et al.*, 2012; Martin *et al.*, 2011; Rosario *et al.*, 2012c) that share similarities to *Sclerotinia sclerotiorum* hypovirulence-associated DNA virus 1 (Yu *et al.*, 2010b) and Rep-like sequences in phytoplasma plasmids, fungi and tubers. Liu *et al.* (2011) recently reported on their discovery of frequent transfer of Rep-like sequences of geminiviruses, nanoviruses and circoviruses to a broad range of eukaryotic species including animals, plants, fungi and protists complementing the earlier findings of Gibbs *et al.* (2006).

Table 5.1: Details of conserved motifs identified in the Reps of GaCSV, BFDV (a representative circovirus) and Rep-like sequences of bacterial origin.

Species strains	**Bacterial / viral family**	**RCR Motifs**			**SF3 Helicase Motifs**		
		I	**II**	**III**	**Walker-A**	**Walker-B**	**Motif C**
Gastropod associated circular ssDNA virus (GaCSV)	Unknown	CLTLNN	HLQGFV	YIGLLNEG	GPTGTGKT	VFEEF	ILSN
Beak and feather disease virus (BFDV)	*Circoviridae*	CFTLNN	HLQGYF	YCSKEGDV	GPPGCGKS	ILDDF	ITSN
Lachnospiraceae bacterium 7 1 58FAA	*Lachnospiraceae*	ALVINN	HTHLFF	YIRKDGQW	GASGAGKT	VFEEF	ITSN
Lachnospiraceae bacterium 6 1 63FAA	*Lachnospiraceae*	QVTINN	HTHIYV	YICKRGKW	GATGTGKT	AFDEF	IISN
Roseburia inulinivorans DSM 16841	*Lachnospiraceae*	-	HTHVYI	YIFKEGEK	GDTGSGKT	MFEEF	ILSN
Coprobacillus sp. 3 3 56FAA	*Erysipelotrichaceae*	NLTINN	HTHIFI	YVSKTGKW	GESGVGKT	LLDNF	LLSV
Erysipelothrix rhusiopathiae str. Fujisawa	*Erysipelotrichaceae*	QITINN	HTHIYL	YIFKLGKW	GDTETGKT	CFEEF	IVSN
Clostridium citroniae WAL-17108	*Clostridiaceae*	NCVFNN	HTHLFV	YIRKDGKW	GATGYGKS	VFEEF	IVSN
Clostridium saccharolyticum WM1	*Clostridiaceae*	QLTFNN	HTHLFI	YLRKEGKW	GATGYGKS	VFEEF	IVSN
Bifidobacterium pseudocatenulatum plasmid	*Bifidobacteriaceae*	LLTIRR	HYQIFA	YCSKEKTR	GETGVGKT	LLDEF	VVSN
Bifidobacterium longum subsp. longum 1-6B	*Bifidobacteriaceae*	-	-	-	GQTGVGKT	LMDEF	IVSN
Mobiluncus mulieris ATCC 35243	*Actinomycetaceae*	MLTIPA	HWQVFV	YVTKLDTA	GPPGAGKS	ILEDF	IISN
Mobiluncus curtisii ATCC 51333	*Actinomycetaceae*	-	-	-	GAPGVGKT	VLDEY	ILSN
Ruminococcus bromii L2-63	*Ruminococcaceae*	LLTINN	HMHIFI	YLRKEGKW	GKSGTGKT	VFEEF	ITSN
Eggerthella sp. YY7918	*Coriobacteriaceae*	FFTFNN	HTHGFI	YVAKENKF	GETGTGKT	VFEEF	ITSN
Oscillibacter valericigenes Sjm18-20	*Oscillospiraceae*	FVTINN	HIHIYI	YVCKSGKW	GTTGAGKT	LFDEF	IVSN
Corynebacterium pseudogenitalium	*Corynebacteriaceae*	-	-	-	GATGAGKT	ILDEF	LLSN
Propionibacterium acnes HL005PA4	*Propionibacteriaceae*	MLTLPE	HFQIYV	YCTKSDTR	GPSRVGKT	LLDEF	IVSN
Streptococcus iniae 9117	*Streptococcaceae*	CCTLNN	HSHLIL	YLNKSGHK	GSSGTGKS	FMDEF	ISSI

In fact Gibbs *et al.* (2006) highlighted the viral Rep-like element (NP_613078) found in the plasmid p4M of *Bifidobacterium pseudocatenulatum* strain VMKB4M, which shares ~27% identity to the Rep of GaCSV. Based on our analysis it is evident that GaCSV is ancestral to ssDNA viruses that have integrated into the genomes of *Ruminococcus bromii* L2-63, *Coprobacillus sp.* 3 3 56FAA, *Lachnospiraceae* bacterium 6 1 63FAA, *Lachnospiraceae* bacterium 7 1 58FAA, *Eggerthella sp.* YY791, *Roseburia inulinivorans* DSM 16841, *Clostridium citroniae* WAL-17108 and *Clostridium saccharolyticum* WM1. The identification and characterisation of GaCSV and its relatively high similarity to the Rep-like elements found in various bacteria does suggest that the diversity of ssDNA viruses and the extent of their role in horizontal gene transfer are largely underestimated. The identification of transposable elements of Geminivirus-like origin in fungi and parvovirus-like origin in animals prompted Liu *et al.* (2011) to propose that eukaryotic transposons could have been derived from ssDNA viruses.

Table 5.2: GaCSV full genome BLASTx results

Accession	Description	Coverage	max ID	E-value
YP_003820447	Viral replication-associated protein [Clostridium saccharolyticum WM1]	36%	35%	$4x10^{-40}$
ZP_08331994	hypothetical protein HMPREF0992_00918 [Lachnospiraceae bacterium 6_1_63FAA]	35%	34%	$2x10^{-39}$
CBL15233	Putative viral replication protein [Ruminococcus bromii L2-63]	36%	36%	$1x10^{-37}$
YP_004561386	putative replication protein [Erysipelothrix rhusiopathiae str. Fujisawa]	35%	35%	$2x10^{-37}$
ZP_09060914	hypothetical protein HMPREF9469_03951 [Clostridium citroniae WAL-17108]	36%	34%	$3x10^{-37}$
ZP_03755188	hypothetical protein ROSEINA2194_03627 [Roseburia inulinivorans DSM 16841]	31%	34%	$1x10^{-36}$
YP_004709701	hypothetical protein EGYY_00170 [Eggerthella sp. YY7918]	35%	36%	$3x10^{-35}$
YP_004879943	hypothetical protein OBV_02390 [Oscillibacter valericigenes Sjm18-20]	35%	33%	$5x10^{-32}$
ZP_09531265	hypothetical protein HMPREF0995_02101 [Lachnospiraceae bacterium 7_1_58FAA]	36%	32%	$5x10^{-31}$
AFM55150	replication associated protein [Beak and feather disease virus]	36%	30%	$1x10^{-23}$

Table 5.3: GaCSV Rep sequence BLASTp results

Accession	Description	Coverage	max ID	E-value
YP_003820447	Viral replication-associated protein [Clostridium saccharolyticum WM1]	97%	35%	$4x10^{-43}$
ZP_08331994	hypothetical protein HMPREF0992_00918 [Lachnospiraceae bacterium 6_1_63FAA]	95%	34%	$2x10^{-42}$
CBL15233	Putative viral replication protein [Ruminococcus bromii L2-63]	97%	36%	$2x10^{-40}$
YP_004561386	putative replication protein [Erysipelothrix rhusiopathiae str. Fujisawa]	96%	36%	$3x10^{-40}$
ZP_09060914	hypothetical protein HMPREF9469_03951 [Clostridium citroniae WAL-17108]	97%	34%	$4x10^{-40}$
ZP_03755188	hypothetical protein ROSEINA2194_03627 [Roseburia inulinivorans DSM 16841]	85%	34%	$2x10^{-39}$
YP_004709701	hypothetical protein EGYY_00170 [Eggerthella sp. YY7918]	96%	36%	$8x10^{-38}$
YP_004879943	hypothetical protein OBV_02390 [Oscillibacter valericigenes Sjm18-20]	96%	33%	$2x10^{-34}$
ZP_09531265	hypothetical protein HMPREF0995_02101 [Lachnospiraceae bacterium 7_1_58FAA]	97%	32%	$5x10^{-33}$
AFM55150	replication associated protein [Beak and feather disease virus]	97%	30%	$2x10^{-25}$

5.5 Concluding remarks

As early as 1969 there has been documentation of viruses detected (none of them ssDNA viruses) in and possibly infecting molluscus globally (Farley, 1969). To date, the majority have been detected in bivalve molluscs, which are filter feeders, and are likely to bioconcentrate viruses in their environments. It is not surprising that a large proportion of bivalve mollusc virology (such as in oysters, clams, mussels and scallops) has predominantly focused on bioconcentrated and bioaccumulated human-infecting viruses and viruses that have been detrimental to aquaculture of bivalves used in the food industry. Further, the lack of mollusc cell lines has also played a significant role in characterising viruses infecting molluscs. The viruses so far characterised from molluscs include members of the *Birnaviridae*, *Malacoherpesviridae*, *Papovariridae*, *Picornaviridae*, *Reoviridae* and *Togaviridae* (Cheng *et al.*, 2005; Comps, 1980, 1988; Cutrin *et al.*, 2000; Elston, 1997; Elston & Wilkinson, 1985; Farley, 1976; Farley *et al.*, 1972; Meyers *et al.*, 2009b; Morley, 2010; Munn, 2006; Oprandy *et al.*, 1981; Renault & Novoa, 2004; Romalde *et al.*, 2002; Savin *et al.*, 2010).

We are not aware of any previous reports of ssDNA viruses in molluscs. However, Jones *et al.* (1996) isolated non-enveloped~25nm virus-like particles from *Perna Canaliculus* (New Zealand green-lipped mussel or spat) and *Mytilus galloprovincialis* (Mediterranean mussel) from the Marlborough Sounds of New Zealand. They also observed some larger enveloped particles in their caesium chloride purifications with a few viral particles. Nonetheless, in their study the high mortality (50-100%, January-April 1994) of *P. canaliculus* was attributed to the ~25nm viral particles. Most of the well-studied ssDNA viruses (i.e. circoviruses) with genome architectures similar to GaCSV are ~ 20nm in diameter therefore the virus particles purified by Jones *et al.* (1996) could potentially have been ssDNA viruses. At this point in time we are unable to verify whether GaCSV is able to infect and cause disease in *Amphibola crenata* (Martyn, 1784).

In summary, we have reported the first case of a ssDNA virus (GaCSV) recovered from molluscs (*A. crenata* Martyn, 1784) in an estuarine environment. *A. crenata* are abundant in estuaries of New Zealand, including the Avon-Heathcote estuary. The occurrence of this virus in this species in an environment which has been subjected to large-scale human pollution may or may not be a coincidence and warrants further investigation. Similarly we are currently unsure if this virus occurs in other marine organisms in this estuary.

Phylogenetic analysis of the Rep sequence shows that GaCSV clusters with all the Rep-like sequences found in some bacterial species and plasmids, strongly suggesting an ancestral transfer of Rep-like sequences from ssDNA viral genomes into bacteria, and probably eukaryotes. Therefore the role of ssDNA virions and their ubiquity needs to be further researched to understand their ecological significance. Furthermore, it is possible that GaCSV-like virions might have been responsible for high mortality amongst at least one species of mollusc in New Zealand.

GenBank accession number: KC172652

References

Allan, G. M. & Ellis, J. A. (2000). Porcine circoviruses: a review. *Journal of veterinary diagnostic investigation* **12**, 3.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.

Anisimova, M. & Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic biology* **55**, 539-552.

Baršienė, J., Bučinskienė, R. & Jokšas, K. (2002). Cytogenetic damage and heavy metal bioaccumulation in molluscs inhabiting different sites of the Neris River. *Ekologija* **2**, 52-57.

Blinkova, O., Victoria, J., Li, Y., Keele, B. F., Sanz, C., Ndjango, J. B. N., Peeters, M., Travis, D., Lonsdorf, E. V. & other authors (2010). Novel circular DNA viruses in stool samples of wild-living chimpanzees. *Journal of General Virology* **91**, 74-86.

Cheng, P. K. C., Wong, D. K. K., Chung, T. W. H. & Lim, W. W. L. (2005). Norovirus contamination found in oysters worldwide. *Journal of medical virology* **76**, 593-597.

Comps, M. (1980). Fluorescence study of the branchiae disease virus of Portuguese oyster Crassotrea angulata Lmk. *Science et Peche*.

Comps, M. (1988). Epizootic diseases of oysters associated with viral infections. *American Fisheries Society Special Publication* **18**, 23-37.

Cutrin, J. M., Olveira, J. G., Barja, J. L. & Dopazo, C. P. (2000). Diversity of infectious pancreatic necrosis virus strains isolated from fish, shellfish, and other reservoirs in northwestern Spain. *Applied and Environmental Microbiology* **66**, 839-843.

Dayaram, A., Opong, A., Jäschke, A., Hadfield, J., Baschiera, M., Dobson, R. C. J., Offei, S. K., Shepherd, D. N., Martin, D. P. & other authors (2012). Molecular characterisation of a novel cassava associated circular ssDNA virus. *Virus Research* **166**, 130-135.

Delwart, E. & Li, L. (2012). Rapidly expanding genetic diversity and host range of the Circoviridae viral family and other Rep encoding small circular ssDNA genomes. *Virus Research* **164**, 114-121.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797.

Elston, R. (1997). Bivalve mollusc viruses. *World Journal of Microbiology & Biotechnology* **13**, 393-403.

Elston, R. & Wilkinson, M. (1985). Pathology, management and diagnosis of oyster velar virus disease (OVVD). *Aquaculture* **48**, 189-210.

Farley, C. (1969). Probable neoplastic disease of the hematopoietic system in oysters, Crassostrea virginica and Crassostrea gigas. *National Cancer Institute Monograph* **31**, 541.

Farley, C. (1976). Ultrastructural observations on epizootic neoplasia and lytic virus infection in bivalve mollusks. *Progress in experimental tumor research* **20**, 283.

Farley, C. A., Banfield, W. G., Kasnic Jr, G. & Foster, W. S. (1972). Oyster herpes-type virus. *Science* **178**, 759-760.

Gibbs, M. J., Smeianov, V. V., Steele, J. L., Upcroft, P. & Efimov, B. A. (2006). Two families of Rep-like genes that probably originated by interspecies recombination are represented in viral, plasmid, bacterial, and parasitic protozoan genomes. *Molecular Biology and Evolution* **23**, 1097-1100.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology* **59**, 307-321.

Gunkel, G. & Streit, B. (1980). Mechanisms of bioaccumulation of a herbicide (atrazine, s-triazine) in a freshwater mollusc *Ancylus fluviatilis* (müll) and a fish *Coregonus fera* (jurine). *Water Research* **14**, 1573-1584.

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

Jones, J., Scotti, P., Dearing, S. & Wesney, B. (1996). Virus-like particles associated with marine mussel mortalities in New Zealand. *Diseases of Aquatic Organisms* **25**, 143-149.

Juniper, K. (1986). Deposit feeding strategy of Amphibola crenata: feeding behaviour, selective feeding and digestion. *Mauri ora* **13**, 103-115.

Kapoor, A., Dubovi, E. J., Henriquez-Rivera, J. A. & Lipkin, W. I. (2012). Complete Genome Sequence of the First Canine Circovirus. *Journal of virology* **86**, 7018-7018.

Kim, H. K., Park, S. J., Song, D. S., Moon, H. J., Kang, B. K. & Park, B. K. (2011). Identification of a novel single stranded circular DNA virus from bovine stool. *Journal of General Virology*.

Koonin, E. V. & Ilyina, T. V. (1992). Geminivirus replication proteins are related to prokaryotic plasmid rolling circle DNA replication initiator proteins. *J Gen Virol* **73 (Pt 10)**, 2763-2766.

Liu, H., Fu, Y., Li, B., Yu, X., Xie, J., Cheng, J., Ghabrial, S. A., Li, G., Yi, X. & other authors (2011). Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. *BMC evolutionary biology* **11**, 276.

Martin, D. P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P. & Varsani, A. (2011). Recombination in Eukaryotic Single Stranded DNA Viruses. *Viruses-Basel* **3**, 1699-1738.

Meyers, T. R., Burton, T., Evans, W. & Starkey, N. (2009a). Detection of viruses and virus-like particles in four species of wild and farmed bivalve molluscs in Alaska, USA, from 1987 to 2009. *Diseases of aquatic organisms* **88**, 1-12.

Meyers, T. R., Burton, T., Evans, W. & Starkey, N. (2009b). Detection of viruses and virus-like particles in four species of wild and farmed bivalve molluscs in Alaska, USA, from 1987 to 2009. *Diseases of Aquatic Organisms* **88**, 1-12.

Morley, N. J. (2010). Interactive effects of infectious diseases and pollution in aquatic molluscs. *Aquatic Toxicology* **96**, 27-36.

Munn, C. B. (2006). Viruses as pathogens of marine organisms - from bacteria to whales. *Journal of the Marine Biological Association of the United Kingdom* **86**, 453-467.

Ng, T. F. F., Willner, D. L., Lim, Y. W., Schmieder, R., Chau, B., Nilsson, C., Anthony, S., Ruan, Y., Rohwer, F. & other authors (2011). Broad Surveys of DNA Viral Diversity Obtained through Viral Metagenomics of Mosquitoes.

Oprandy, J. J., Chang, P. W., Pronovost, A. D., Cooper, K. R., Brown, R. S. & Yates, V. J. (1981). Isolation of a viral agent causing hematopoietic neoplasia in the soft-shell clam, *Mya arenaria*. *Journal of Invertebrate Pathology* **38**, 45-51.

Phan, T. G., Kapusinszky, B., Wang, C., Rose, R. K., Lipton, H. L. & Delwart, E. L. (2011). The Fecal Viral Flora of Wild Rodents. *PLoS pathogens* **7**, e1002218.

Renault, T. & Novoa, B. (2004). Viruses infecting bivalve molluscs. *Aquatic Living Resources* **17**, 397-409.

Romalde, J., Area, E., Sánchez, G., Ribao, C., Torrado, I., Abad, X., Pinto, R., Barja, J. & Bosch, A. (2002). Prevalence of enterovirus and hepatitis A virus in bivalve molluscs from Galicia (NW Spain): inadequacy of the EU standards of microbiological quality. *International journal of food microbiology* **74**, 119-130.

Rosario, K., Duffy, S. & Breitbart, M. (2012a). A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Archives of virology* **157**, 1851-1871.

Rosario, K., Padilla-Rodriguez, M., Kraberger, S., Stainton, D., Martin, D. P., Breitbart, M. & Varsani, A. (2012b). Discovery of a novel mastrevirus and alphasatellite-like circular DNA in dragonflies (Epiprocta) from Puerto Rico. *Virus Research*.

Rosario, K., Dayaram, A., Marinov, M., Ware, J., Kraberger, S., Stainton, D., Breitbart, M. & Varsani, A. (2012c). Diverse circular single-stranded DNA viruses discovered in dragonflies (Odonata: Epiprocta). *Journal of General Virology* **93**, 2668-2681.

Rosario, K., Marinov, M., Stainton, D., Kraberger, S., Wiltshire, E. J., Collings, D. A., Walters, M., Martin, D. P., Breitbart, M. & other authors (2011). Dragonfly cyclovirus, a novel single-stranded DNA virus discovered in dragonflies (Odonata: Anisoptera). *Journal of General Virology* **92**, 1302-1308.

Savin, K. W., Cocks, B. G., Wong, F., Sawbridge, T., Cogan, N., Savage, D. & Warner, S. (2010). A neurotropic herpesvirus infecting the gastropod, abalone, shares ancestry with oyster herpesvirus and a herpesvirus associated with the amphioxus genome. *Virology Journal* **7**, 308.

Shan, T., Li, L., Simmonds, P., Wang, C., Moeser, A. & Delwart, E. (2011). The fecal virome of pigs on a high-density farm. *Journal of virology* **85**, 11697-11708.

Sikorski, A., Argüello-Astorga, G. R., Dayaram, A., Dobson, R. C. J. & Varsani, A. (2013). Discovery of a novel circular single-stranded DNA virus from porcine faeces. *Archives of virology*.

van den Brand, J. M. A., van Leeuwen, M., Schapendonk, C. M., Simon, J. H., Haagmans, B. L., Osterhaus, A. D. M. E. & Smits, S. L. (2011). Metagenomic analysis of the viral flora of pine marten and European badger feces. *Journal of virology*.

Varsani, A., Regnard, G. L., Bragg, R., Hitzeroth, I. I. & Rybicki, E. P. (2011). Global genetic diversity and geographical and host-species distribution of beak and feather disease virus isolates. *Journal of General Virology* **92**, 752-767.

Walsh, K., Dunstan, R., Murdoch, R., Conroy, B., Roberts, T. & Lake, P. (1994). Bioaccumulation of pollutants and changes in population parameters in the gastropod mollusc *Austrocochlea constricta*. *Archives of Environmental Contamination and Toxicology* **26**, 367-373.

Yu, X., Li, B., Fu, Y., Jiang, D., Ghabrial, S. A., Li, G., Peng, Y., Xie, J., Cheng, J. & other authors (2010a). A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. *Proceedings of the National Academy of Sciences* **107**, 8387.

Yu, X., Li, B., Fu, Y. P., Jiang, D. H., Ghabrial, S. A., Li, G. Q., Peng, Y. L., Xie, J. T., Cheng, J. S. & other authors (2010b). A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 8387-8392.

Chapter 6

Diverse circular Rep-encoding ssDNA viruses circulating amongst molluscs at the Avon-Heathcote estuary in Christchurch, New Zealand

Contents

This work is in review with Infection, Genetics and Evolution and is presented in a similar manner to that of the publication:

Dayaram, A., Goldstien, S., Argüello-Astorga, G. R., Zawar-Reza, P., Gomez, C., Harding, J. S. & Varsani, A. (2014) Diverse circular Rep encoding ssDNA viruses circulating amongst molluscs at the Avon-Heathcote estuary in Christchurch, New Zealand. *Infection, Genetics and Evolution* (Accepted March 2015)

6.1 Abstract

Our understanding of the diversity and abundance of circular replication-associated protein (Rep) - encoding single-stranded (CRESS) DNA viruses has increased considerably over the last few years due. This can be attributed to a combination of modern sequencing technologies and new molecular tools. Studies have used these techniques to identify and recover CRESS DNA viruses from a range of different marine organisms, including copepods, shrimp and molluscs. In our study we identify 79 novel CRESS DNA viruses recovered from three mollusc species (*Austrovenus stutchburyi*, *Paphies subtriangulata* and *Amphibola crenata*) and benthic sediment from the Avon-Heathcote estuary in Christchurch, New Zealand. The recovered viral genomes display varied architectures, with all encoding at least two major ORFs arranged in a unidirectional or bidirectional organisation. Analysis of the Reps of these viral genomes indicates they are all highly diverse; the maximum amino acid identity of one Rep sequence sharing 65% with the Rep of Gastropod associated circular ssDNA virus (GaCSV). This study adds significantly to the wealth of CRESS DNA viruses recovered from freshwater and marine environments and extends our knowledge of the distribution of these viruses.

6.2 Introduction

In Chapter 5, a novel single-stranded DNA (ssDNA) virus was isolated from the estuarine mollusc *Amphibola crenata*. The replication-associated protein (Rep) of this virus showed similarities to Rep-like elements identified in bacterial genomes. Chapter 6 continues to examine the diversity of circular Rep encoding single stranded (CRESS) DNA viruses in three molluscs species as well as benthic sediment from an estuary.

A large number of novel CRESS viruses have been discovered in the last five years in various ecosystems. This is primarily due to new sequence-independent amplification methods and the use of next generation sequencing (NGS) approaches. Many studies have used a combination of rolling circle amplification (RCA), shotgun sequencing and NGS to recover CRESS DNA viruses from a range of environments including deep-sea vents (Yoshida *et al.*, 2013), Antarctic lakes (López-Bueno *et al.*, 2009), reclaimed water (Rosario *et al.*, 2009b), rice paddy soil (Kim *et al.*, 2008) and ocean water (Angly *et al.*, 2006; Labonté & Suttle, 2013; Rosario & Breitbart, 2011; Rosario *et al.*, 2009a). Given that viruses are the most abundant entities in most environments and they play an important role in regulating the structure of microbial communities (Danovaro *et al.*, 2011; Suttle, 2007), it is not surprising that diverse CRESS DNA viruses are abundant. This raises question of their role and 'flow' within ecosystems.

Herpes-like viruses were first identified in adult *Crassostrea virginica* (Farley, 1976b; Farley *et al.*, 1972) and subsequently herpes-like viruses have been associated with high mortalities of *C. gigas* hatchery-reared larvae (Hine *et al.*, 1992; Nicolas *et al.*, 1992). Experimental studies have demonstrated the transmission of these herpes-like viruses between different species of molluscs and their larvae (Le Deuff *et al.*, 1994; LeDeuff *et al.*, 1996). Viruses belonging to the families *Birnavirida*e, *Picornaviridae*, *Togaviridae*, *Reoviridae* and *Retrovirdae* have also been found to infect molluscs, reviewed by Renault and Novoa (2004).

In general, molluscs have inefficient metabolic systems and are herbivorous filter feeders. As a result, they tend to concentrate entities present in their aquatic environment. For this reason they can be used as indicator organisms in bioaccumulation and bio-concentration studies to monitor pollutants in ecosystems (Baršienė *et al.*, 2002; Gunkel & Streit, 1980; van der Oost *et al.*, 1988; Walsh *et al.*, 1994). The concept of bio-concentration has also been applied to areas of viral research, particularly for commercial bivalve molluscs such as oysters, clams, mussels and scallops to monitor viruses that are pathogenic to humans (Cheng *et al.*, 2005; Comps, 1988; Elston, 1997; Farley, 1976a; Meyers *et al.*, 2009; Morley, 2010).

Recently the first CRESS DNA virus associated with the gastropod *Amphibola crenata* was described in Christchurch, New Zealand (Chapter 5). Gastropod-associated circular ssDNA virus (GaCSV) has a 2351 nt genome with two major open reading frames (ORFs) that are bidirectionally transcribed. Phylogenetic analysis of the Rep of GaCSV shows that it is similar to Rep-like sequences of bacterial origin. At the same sample site a *Starling circovirus* (StCV) was also identified in *A. crenata* (Dayaram *et al.*, 2013a). Prior to this study there had been no reports of starling circoviruses outside of Europe. Starlings have been seen foraging around the site where *A. crenata* were sampled and it is highly likely that the source of the StCV was starling faecal matter. This shows a unique example that molluscs may be ‘natural tools’ for viral surveillance in ecosystems.

Given that most studies have looked at specific samples within ecosystems for assessing viral diversity, we sampled *Austrolvenus stutchburyi* (cockles), *Paphies subtriangulata* (tuatua) and *Amphibola crenata* (gastropods) as potential ‘natural surveillance tools for viruses. We also sampled benthic sediment that is the common substrate in the ecosystem where these three mollusc species inhabit. Our objective was to identify CRESS DNA viruses circulating in the Avon-Heathcote estuary and to determine whether the mollusc species sampled in this study could be used for future viral surveillance work.

6.3 Materials and methods

6.3.1 Sample collection and processing

Samples were collected at three different sites in the Avon-Heathcote estuary on the 4th of July 2012. The estuary is formed by the confluence of two rivers (the Avon and Heathcote), which flow through the city of Christchurch (approx. population 340 000 people). The two rivers have > 90 small tributary streams that drain from the urban area. Fifty five individual *A. stutchburyi* were collected from the Heathcote Bridge (near the mouth of the Heathcote River, (43.5578 S, 172.70588 E), 40 individuals from the Causeway (43.5564 S, 172.7289 E), 40 individuals from Beachville Street (43.5567 S, 172.7350 E) and 40 individuals from PP the Yacht club (43.5661 S, 172.7469 E). Forty *P. subtriangulata* were collected from PP Yacht club (43.5661 S, 172.7469 E) and 50 *A. crenata* were collected from Heathcote Bridge (43.5578 S, 172.7058 E). Samples were refrigerated upon collection, then were removed from the shells and washed in sterile distilled water prior to processing. Whole soft bodies were then pooled according to species for each of the sites sampled.

6.3.2 DNA extraction, viral purification and circular DNA enrichment

Each pooled species and the benthic sediment were homogenized in SM buffer [0.1 M NaCl, 50 mM Tris/HCl (pH 7.4), 10 mM $MgSO_4$] at a ratio of 10 ml buffer per 5 g of tissue. The homogenate was then centrifuged (10,000 *x g* for 10 min) in order to pellet the cellular debris and the supernatant was filtered first through a 0.45 µm, followed by a 0.2 µm syringe filter (Sartorius Stedim Biotech, Germany). Viral DNA was then recovered using the High Pure Viral Nucleic Acid kit (Roche, USA) from the filtrate. In order to enrich circular DNA, we used TempliPhi 2000 (GE Healthcare, USA) as previously described (Dayaram *et al.*, 2014; Dayaram *et al.*, 2013b; Rosario *et al.*, 2012b, c; Rosario *et al.*, 2011; Sikorski *et al.*, 2013a).

6.3.3 Illumina HiSeq 2000 sequencing and NGS data assembly

The RCA products were then pooled by species and sequenced on an Illumina HiSeq 2000 (Illumina, USA) platform at the Beijing Genomics Institute (Hong Kong). The paired end reads were *de novo* assembled using ABySS V1.3.5 (Simpson *et al.*, 2009) with kmer = 64. A

BLASTx analysis was performed on the *de novo* assembled contigs >1500 nts using KoriBlast v 4.1 (Korilog SARL, Bioinformatics Solutions, France) in order to identify contigs encoding known viral-like proteins.

6.3.4 Recovery of viral genomes

Back-to-back primers were then designed (Table 6.1) based on the BLASTx hits of *de novo* assembled contigs which had similarities to CRESS DNA viral-like sequences in order to recover the full genomes from the three mollusc species and the benthic sediment samples. The genomes were recovered by PCR amplification with the specific back-to-back primer pairs using Kapa HiFi HotStart polymerase (Kapa Biosystems, USA). The resulting amplicons were gel purified, cloned into pJET1.2 plasmid (Thermo Fisher, USA) and Sanger sequenced at Macrogen Inc. (South Korea) by primer walking.

Sanger sequencing reads were assembled using DNA Baser Sequence Assembler (version 4.16; Heracle Biosoft S.R.L., Romania). An initial comparison of the genomes sequences was then carried out using BLASTx and tBLASTx (Altschul *et al.*, 1990). The viral genomes were then annotated and the major open reading frames (ORFs), conserved motifs and stem-loop structures were identified using the Sfold server (http://sfold.wadsworth.org/cgi-bin/index.pl).

6.3.5 Analysis of the major ORFs

The Reps from all CRESS DNA viruses available in GenBank (downloaded 20th August 2014) were used to create a dataset for analysis. These, together with the Reps of viruses determined in this study, were aligned using MUSCLE (Edgar, 2004) with manual editing where necessary. The resulting alignment was used to infer a maximum likelihood phylogenetic tree using a JTT + CAT model with aLRT branch (Anisimova & Gascuel, 2006) support using FastTree version 2.1.7 (Price *et al.*, 2010). The Rep maximum likelihood phylogenetic tree was mid-point rooted and branches with less than 80% aLRT support were collapsed using Mesquite (version 2.75) (Maddison & Maddison, 2011).

Table 6.1: Details of primer sequences used to recover viral genomes in this study.

Sample	Forward primer	Reverse primer
AHEaCV-1	GAAGTTTGTCCCACTACGG	GTCCCCCCATGCCAC
AHEaCV-2	ATTAGTATTACCCCTTTTGACTTCTGT	GACCCTTTTGACTATACCTCG
AHEaCV-3	GGGGTCGAGGTTTGAAGTAAT	CGAGAATGGTATGCTGATCTAG
AHEaCV-4	GAAGGCATAGCGGTCTATC	CGTGTAGAATGTAAAGGAAGTTC
AHEaCV-5	TCATCAATCGCACCCATCGTT	AATCAGGTCCAACAACATAGGG
AHEaCV-6	GGTGTCTAATGTGGTGTAGGG	GGTCAATGGTGGGAAGGATATG
AHEaCV-7	CTACCCACTAATGCGCAATCAA	ACCTCTACCCCATCGCG
AHEaCV-8	ACTAGTAAAACAGAGGCGACGAAGG	CGGTTCAAGATAGTTAGGGGGC
AHEaCV-9	CATGAATGCGTATCTGTCCAAC	GTTGAATGTAAAGGCGGTAGC
AHEaCV-10	GAGTGTACGGGTTCTTCTTGG	GCAAGAGGAAGGCGAGCAA
AHEaCV-11	CCTCCCCCAAGAAGTACAAC	CCATTTGAGCTAGTGATAAGAAAAGG
AHEaCV-12	TCATCCTCTAATGCTACCTGCTCGG	CCGTCCACCCAGTTCAGCATGTCC
AHEaCV-13	AAGGAGCAATGGGTATCGATC	GAAACGAAAGGAGGGCAATG
AHEaCV-14	GTGCCATTAGATATGCTGCCG	GTCCGGCACCAACGACAG
AHEaCV-15	GCTCTCCAATCCTCTTATCCTAAC	TGTGGTGTTGTGCATGTTTCGG
AHEaCV-16	AGGTTTCGTTCGAGGCTGGTG	GGAGAGGATCTGCGGGAAG
AHEaCV-17	TTCCTACAGCGCATATTGAAATAATG	ATAGCTTCTTCCATCCTGTCAG
AHEaCV-18	GAAATACCATCCACTTGCTTACTC	CAATATGGGGTGATCAATTAAGTTTTAC
AHEaCV-19	GATATAACAGGGGGATCGAAAAAC	GAATAAAAGTCGTAGGGAACTCC
AHEaCV-20	CCATTATAACCATCCCACCATTTAC	GCAAGACGTTGTTGTATTTGATGATATA
AHEaCV-21	TCATAATGATAATACTTCGGACAGC	CGACCTTGAAATAGAGCAACG
AHEaCV-22	TCCCTACTTAATAGTGTTTCATCCG	CAGATGGAGAAAGTGGGAAATC
AHEaCV-23	GACTTTAAATCAGAAAAATGGATGGATCT	GATATGTTCTGTAGGGTTGTTTAAGG
AHEaCV-24	CAGTATTACCCCACTCGAACTTG	TACCCACTCTTTTAATTTGATGCGCG
AHEaCV-25	CAAGAATGGTAGTGAAAGGCG	TAGTATCATCTCCATTCGCAGAT
AHEaCV-26	CTGGGTCGATGTTACTGGTAAT	CTAAATGGTATGAGGGAAAAGACG
AHEaCV-27	CTATATGGCAGGGAAACGTGTC	CAGGTACTTCATGTGATCGGG
AHEaCV-28	AGTGGCAGGGAATGAACGTTG	GTTTAGCACATTCCCTGCCACT
AHEaCV-29	CCAAAAGGAGAAAGGAGATACTGG	AACACGAAACTCTTAGCAACACCC

Rep-like sequences found in bacterial genomes were downloaded from GenBank on the 5th of October 2014 and a small dataset was created with the Rep of GasCV (KC172652) and that of a virus identified in this study that was closely related to it. These Rep sequences were aligned using MUSCLE (Edgar, 2004). ProtTest (Darriba *et al.*, 2011) was used to determine the model of best fit and the phylogenetic tree was generated using PhyML version 3.0 (Guindon *et al.*, 2010). The Rep tree was created using RtREV+G+F amino acid substitution model with aLRT branch support and was rooted with circovirus sequences, namely *Beak and feather disease virus* (BFDV; AF071878), *Porcine circovirus* (PCV1; AF012107), *Duck circovirus* (DuCV; DQ100076) and *Raven circovirus* (RaCV; DQ146997).

The Reps and coat proteins (CP) of ten viruses from this study with similarities to pig-stool-associated circular ssDNA virus (PisaCV), bovine-stool-associated circular virus (BoSVC), porcine-stool-associated circular virus (PoSCV), turkey-stool-associated circular virus (TuSCV), chimpanzee-stool-associated circular virus (ChiSV) and Odonata-associated circular virus 5 and 21 (OdasCV-5, -21) were assembled into datasets and aligned using MUSCLE (Edgar, 2004). For the purpose of this study we have named these chipoviruses (chimpanzee and porcine faeces associated DNA viruses; although this group includes all of the above).

We used ProtTest (Darriba *et al.*, 2011) to determine the best-fit model of substitution and maximum likelihood phylogenetic trees (Rep and CP) were inferred using PhyML version 3.0 (Guindon *et al.*, 2010). The Rep likelihood phylogenetic tree was inferred using the RtREV+G+I model whereas the CP tree was inferred using the WAG+G model. For both trees, branches with aLRT branch support <80% were collapsed using Mesquite (version 2.75). The Rep sequence of McMurdo Ice Shelf pond-associated circular virus 8 (MpaCDV-8; KJ547653) (Zawar-Reza *et al.*, 2014) was used to root the Rep phylogenetic tree whereas the CP phylogenetic tree was midpoint rooted.

All pairwise identities were calculated using SDT v1.2 (Muhire *et al.*, 2014). A BLASTp (Altschul *et al.*, 1990) analysis was carried out comparing the NCBI non-redundant protein database to the putative Reps from all the viral genomes recovered.

6.3.6 Viral distribution analysis

Statistical analysis was performed using the Pearson's chi-squared test to determine whether the viral distribution across samples differs from the expected frequencies. The expected frequency was calculated to be 19.75 viruses per sample and was then compared to the number of viruses recovered from each sample type (*A. stutchburyi*, *P. subtriangulata*, *A. crenata* and benthic sediment). The threshold for significance was set at P<0.01.

6.4 Results and Discussion

6.4.1 Characterisation of viral genomes

In our *de novo* assembled sequences, we identified sequences that encode protein that have similarities to CRESS-DNA viruses, gokushoviruses, chlamydia pneumoniae phages, enterobacteria phages and recently characterised bacilladnaviruses. However, for the purpose of this study we focused only on CRESS-DNA viruses. A total of seventy-nine circular DNA viral sequences (assumed to be complete genomes) were recovered from *A. stutchburyi* (n=37), *P. subtriangulata* (n=16), *A. crenata* (n=14) and benthic sediment (n=12) from four different sites in the Avon-Heathcote estuary (Christchurch, New Zealand) (Figure 6.2 A). Of these, twenty-nine represented novel CRESS DNA viruses and we have tentatively named them as Avon-Heathcote estuary-associated circular DNA virus 1 through to 29 (AHEaCV-1 to -29) and have been grouped into six different genome types defined by Rosario *et al.* (2012a). The CRESS DNA viral sequences identified in this study range in size from 1741 to 2817 nt (Figure 6.1). All genomes have two ORFs that were either uni-directionally or bi-directionally orientated. BLASTp analysis carried out on the major ORFs indicated that they either encoded for a Rep or putative CP. Only AHEaCV-21 was found to have a putative spliced Rep with an 86 nt intron; introns in the Rep are common in certain members of the *Geminiviridae* family (e.g. mastreviruses) (Donson *et al.*, 1987; Wright *et al.*, 1997) and have also been identified in various CRESS DNA viruses (Ng *et al.*, 2011; Rosario *et al.*, 2012b, c; Sikorski *et al.*, 2013b; van den Brand *et al.*, 2011). In the AHEaCVs we identified at least one intergenic region (IR) and 66 AHEaCVs have both small intergenic regions (SIR) and large intergenic regions (LIR) (Figure 6.1). The intergenic regions of all AHEaCVs contain a stem-loop structure which is the origin of replication (*ori*), a common feature found in all CRESS DNA viruses containing a conserved nonanucleotide motif (Table 6.2).

The pairwise identities between the Reps of the recovered viral isolates with other CRESS DNA viral reps are provided in Figure 6.2 B. The Reps of AHEaCV-29 and AHEaCV-8 that were recovered from *P. subtriangulata*, *A. stutchburyi*, *A. crenata* and benthic sediment shares 63% pairwise identity whilst the Reps from the other viral genomes recovered shared between <52% pairwise identity. The diversity of these viruses was further elucidated in the phylogenetic analysis of the AHEaCV Reps with other CRESS DNA viral sequences available in public databases (Figure 6.3). Together, both analyses suggest a high diversity of CRESS DNA viruses were recovered from this study.

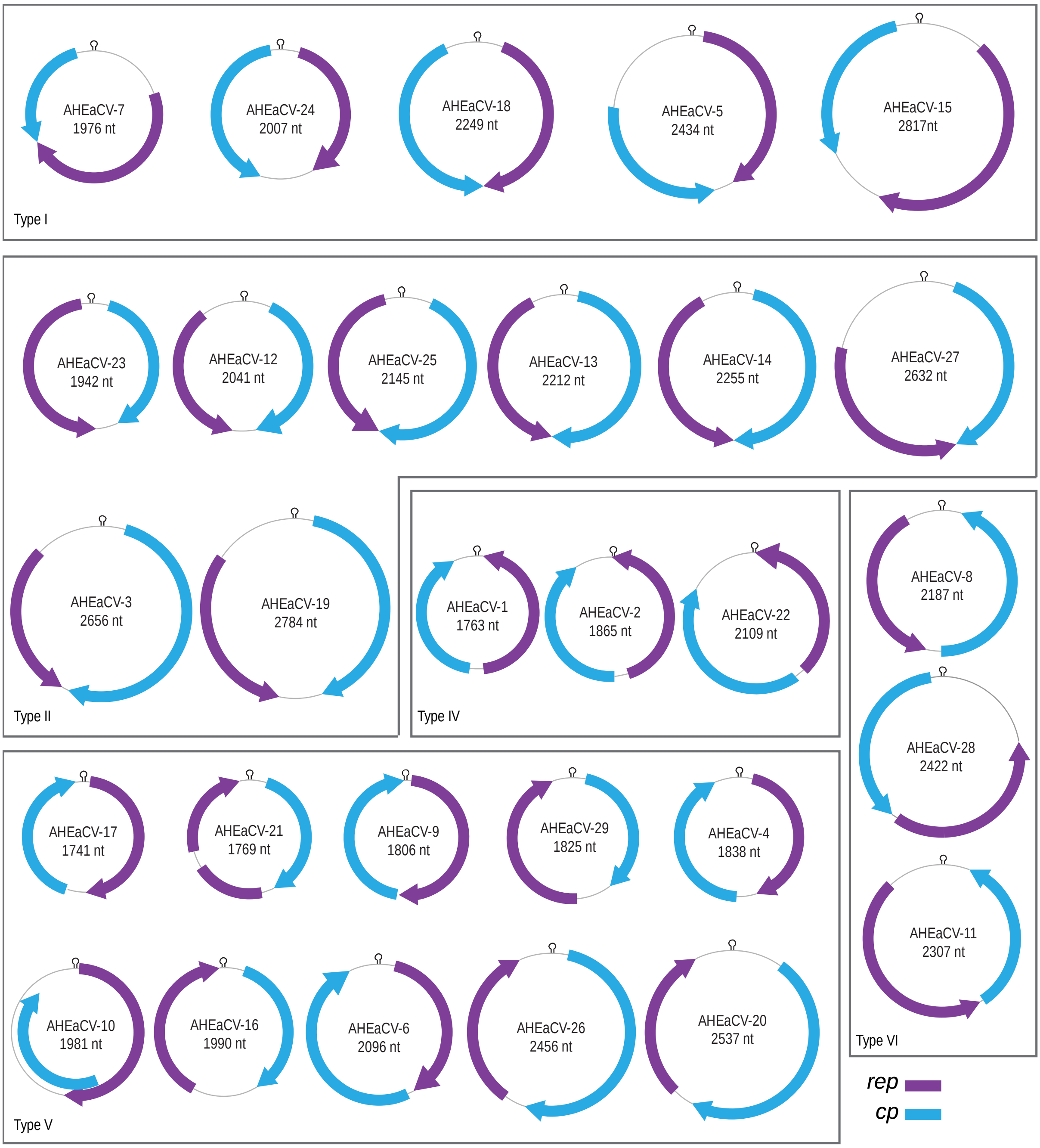


Figure 6.1: Genome organisations of the CRESS DNA viruses recovered from three different mollusc species and benthic sediment.

Table 6.2: Sampling locations, specimens, viral isolates recovered and conserved motifs identified in the Rep and the putative nonanucleotide sequences.

			Nonanucleotide	RCR Motifs			SF3 Helicase Motifs		
Viral Isolate	Accession #	Sample source	motif	I	II	III	WalkerA	WalkerB	Motif C
AHEaCV-1 [NZ-2981C1-2012]	KM874290	*A. stutchburyi*	CAGTATTAC	LGFNF	HKQFWCMF	YCKKEGSY	GEPGSGKS	VLLEDFDF	ITSN
AHEaCV-1 [NZ-2981C2-2012]	KM874291	*A. stutchburyi*	CAGTATTAC	LGFNF	HKQFWCMF	YCKKEGSY	GEPGSGKS	VLLEDFDF	ITSN
AHEaCV-2 [NZ-3024C1-2012]	KM874292	*A. stutchburyi*	TAGTATTAC	ITENN	HLQIYLEF	YCSKDQDV	GASGSGKT	VLLDDLRG	ITSP
AHEaCV-2 [NZ-3024C2-2012]	KM874293	*A. stutchburyi*	TAGTATTAC	ITENN	HLQIYLEF	YCSKDQDV	GASGSGKT	VLLDDLRG	ITSP
AHEaCV-2 [NZ-3024C3-2012]	KM874294	*A. stutchburyi*	TAGTATTAC	ITENN	HLQIYLEF	YCSKDQDV	GASGSGKT	VLLDDLRG	ITSP
AHEaCV-3 [NZ-3030C2-2012]	KM874295	*A. stutchburyi*	TAGTATTAC	LTIRH	HWQLVVHF	YVWKEDSA	GVTGSGKS	VVIDEFRG	ITSN
AHEaCV-3 [NZ-3030C3-2012]	KM874296	*A. stutchburyi*	TAGTATTAC	LTIRH	HWQLVVHF	YVWKEDSA	GVTGSGKS	VVIDEFRG	ITSN
AHEaCV-3 [NZ-3887G-2012]	KM874297	*A. crenata*	TAGTATTAC	LTIRH	HWQLVVHF	YVWKEDTA	GVTGSGKS	VVIDEFRG	ITSN
AHEaCV-4 [NZ-3049C1-2012]	KM874298	*A. stutchburyi*	TAGTATTAC	MTIHN	HLQVYAYF	YCSKEDVL	GESRSGKT	VLIDDIRK	FPPV
AHEaCV-4 [NZ-3049C2-2012]	KM874299	*A. stutchburyi*	TAGTATTAC	MTIHN	HLQVYAYF	YCSKEDVL	GESRSGKT	VLIDDIRK	FPPV
AHEaCV-4 [NZ-3049C3-2012]	KM874300	*A. stutchburyi*	TAGTATTAC	MTIHN	HLQVYAYF	YCSKEDVL	GESRSGKT	VLIDDIRK	FPPV
AHEaCV-5 [NZ-3091C2-2012]	KM874301	*A. stutchburyi*	CAGTATTAC	FTLNN	HYQGYCEL	YCKKDDTR	GPTGLGKT	VLLDDFSG	VTTN
AHEaCV-5 [NZ-3091C3-2012]	KM874302	*A. stutchburyi*	CAGTATTAC	FTLNN	HYQGYCEL	YCKKDDTR	GPTGLGKT	VLLDDFSG	VTTN
AHEaCV-5 [NZ-4834GA-2012]	KM874303	*A. crenata*	CAGTATTAC	FTLNN	HYQGYCEL	YCKKDDTR	GPTGLGKT	VLLDDFSG	VTTN
AHEaCV-6 [NZ-2194TU-2012]	KM874307	*P. subtriangulata*	TAGTATTAC	FTLHN	HLQGYIQF	YCSKDGDM	GPTGTGKT	VTIDDMRG	ITSA
AHEaCV-6 [NZ-2974SG-2012]	KM874306	Benthic sediment	TAGTATTAC	FTLHN	HLQGYIQF	YCSKDGDM	GPTGTGKT	VIIDDMRG	ITSA
AHEaCV-6 [NZ-3103C1-2012]	KM874304	*A. stutchburyi*	TAGTATTAC	FTLHN	HLQGYIQF	YCSKDGDM	GPTGTGKT	VIIDDMRG	ITSA
AHEaCV-6 [NZ-3103C3-2012]	KM874305	*A. stutchburyi*	TAGTATTAC	FTLHN	HLQGYIQF	YCSKDGDM	GPTGTGKT	VIIDDMRG	ITSA
AHEaCV-6 [NZ-4645GA-2012]	KM874308	*A. crenata*	TAGTATTAC	FTLHN	HLQGYIQF	YCSKDGDM	GPTGTGKT	VIIDDMRG	ITSA
AHEaCV-7 [NZ-3107C3-2012]	KM874309	*A. stutchburyi*	TAGTATTAC	FTNER	HLQFYCEF	YCSKEDTR	GEPGTGKT	LLIDDYYG	ITSN
AHEaCV-8 [NZ-2216TU-2012]	KM874312	*P. subtriangulata*	CCTACTTAC	FRYNA	HYQGRMSL	YVIKEDTR	PIGDLGKS	YIVDMPRG	VFTN
AHEaCV-8 [NZ-3072SG-2012]	KM874314	Benthic sediment	CCTACTTAC	FRYNA	HYQGRMSL	YVIKEDTR	PIGDLGKS	YIVDMPRG	VFTN
AHEaCV-8 [NZ-3159C1-2012]	KM874310	*A. stutchburyi*	CCTACTTAC	FRYNA	HYQGRMSL	YVIKEDTR	PIGDLGKS	YIVDMPRG	VFTN
AHEaCV-8 [NZ-3159C3-2012]	KM874311	*A. stutchburyi*	CCTACTTAC	FRYNA	HYQGRMSL	YVIKEDTR	PIGDLGKS	YIVDMPRG	VFTN
AHEaCV-8 [NZ-4738GA-2012]	KM874313	*A. crenata*	CCTACTTAC	FRYNA	HYQGRMSL	YVIKEDTR	PIGDLGKS	YIVDMPRG	VFTN
AHEaCV-9 [NZ-3131SG-2012]	KM874318	*Benthic sediment*	TAGTATTAC	FTLNN	HLQGCIIF	YCKKDGDV	GATGTGKS	VIIDDMRK	VFTN
AHEaCV-9 [NZ-3171C1-2012]	KM874315	*A. stutchburyi*	TAGTATTAC	FTLNN	HLQGCIIF	YCKKDGDV	GATGTGKS	VIIDDMRK	VFTN
AHEaCV-9 [NZ-3171C3-2012]	KM874316	*A. stutchburyi*	TAGTATTAC	FTLNN	HLQGCIIF	YCKKDGDV	GATGTGKS	VIIDDMRK	VFTN
AHEaCV-9 [NZ-4424GA-2012]	KM874317	*A. crenata*	TAGTATTAC	FTLNN	HLQGCIIF	YCKKDGDV	GATGTGKS	VIIDDMRK	VFTN
AHEaCV-10 [NZ-2599SG-2012]	KM874322	*Benthic sediment*	TAGTATTAC	FTEFK	HYQGFAYS	YCSKEGSL	GAAGTGKT	MLFDDVEI	FTSN
AHEaCV-10 [NZ-3241C2-2012]	KM874319	*A. stutchburyi*	TAGTATTAC	FTEFK	HYQGFAYS	YCSKEGNL	GPSGTGKT	MLFDDVEI	FTSN
AHEaCV-10 [NZ-3241C3-2012]	KM874320	*A. stutchburyi*	TAGTATTAC	FTEFK	HYQGFAYS	YCSKEGSL	GAAGTGKT	MLFDDVEI	FTSN
AHEaCV-10 [NZ-4985GA-2012]	KM874321	*A. crenata*	TAGTATTAC	FTEFK	HYQGFAYS	YCSKEGSL	GAAGTGKT	MLFDDVEI	FTSN
AHEaCV-11 [NZ-2256TU-2012]	KM874327	*P. subtriangulata*	TATTATTAC	FTINN	HLQGYIRW	YCKKEEGV	GPSGIGKT	IFMDEFRW	ICTN
AHEaCV-11 [NZ-3130SG-2012]	KM874326	Benthic sediment	TATTATTAC	FTINN	HLQGYIRW	YCKKEEGV	GPSGIGKT	IFMDEFRW	ICTN
AHEaCV-11 [NZ-3371C1-2012]	KM874323	*A. stutchburyi*	TATTATTAC	FTINN	HLQGYIRW	YCKKEEGV	GPSGIGKT	IFMDEFRW	ICTN
AHEaCV-11 [NZ-3371C2-2012]	KM874324	*A. stutchburyi*	TATTATTAC	FTINN	HLQGYIRW	YCKKEEDV	GPSGIGKT	IFMDEFRW	ICTN
AHEaCV-11 [NZ-3371C3-2012]	KM874325	*A. stutchburyi*	TATTATTAC	FTINN	HLQGYIRW	YCKKEEDV	GPSGIGKT	IFMDEFRW	ICTN
AHEaCV-12 [NZ-3316C1-2012]	KM874328	*A. stutchburyi*	CAGTATTAC	LTLNN	HLQGFVHL	YIMEDVEG	GPTGCGKT	IVFEEFRS	STIS
AHEaCV-13 [NZ-1986SG-2012]	KM874330	*Benthic sediment*	TAGTATTAC	LTLNN	HLQGYIET	YCLKEDSR	GPTGTGKS	IIIDEFYG	ITSN
AHEaCV-13 [NZ-3331CO-2012]	KM874329	*A. stutchburyi*	TAGTATTAC	LTLNN	HLQGYIET	YCLKEDSR	GPTGTGKS	IIIDEFYG	ITSN
AHEaCV-13 [NZ-4754GA-2012]	KM874331	*A. crenata*	TAGTATTAC	LTLNN	HLQGYIET	YCLKEDSR	GPTGTGKS	IIIDEFYG	ITSN
AHEaCV-14 [NZ-2438TU-2012]	KM874335	*P. subtriangulata*	TAGTATTAC	FTINN	HLQGYLEI	YCCKEDSR	GPTGTGKS	VIIDEYYG	FTTN
AHEaCV-14 [NZ-3341SG-2012]	KM874334	Benthic sediment	TAGTATTAC	FTINN	HLQGYLEI	YCCKEDSR	GPTGTGKS	VIIDEYYG	FTTN
AHEaCV-14 [NZ-3348C3-2012]	KM874332	*A. stutchburyi*	TAGTATTAC	FTINN	HLQGYLEI	YCCKEDSR	GPTGTGKS	VIIDEYYG	FTTN
AHEaCV-14 [NZ-4781GA-2012]	KM874333	*A. crenata*	TAGTATTAC	FTINN	HLQGYLEI	YCCKEDSR	GPTGTGKS	VIIDEYYG	FTTN
AHEaCV-15 [NZ-2320TU-2012]	KM874338	*P. subtriangulata*	TAGTATTAC	FTLNN	HLQGAVVI	YCTKQDTD	GETGVGKT	AIFDDLRT	VTAP
AHEaCV-15 [NZ-2957SG-2012]	KM874339	Benthic sediment	TAGTATTAC	FTLNN	HLQGAVVI	YCTKQDTD	GETGVGKT	AIFDDLRT	VTAP
AHEaCV-15 [NZ-3424C1-2012]	KM874336	*A. stutchburyi*	TAGTATTAC	FTLNN	HLQGAVVI	YCTKQDTD	GETGVGKT	AIFDDLRT	VTAP
AHEaCV-15 [NZ-3424C2-2012]	KM874337	*A. stutchburyi*	TAGTATTAC	FTLNN	HLQGAVVI	YCTKQDTD	GETGVGKT	AIFDDLRT	VTAP
AHEaCV-16 [NZ-2991SG-2012]	KM874342	Benthic sediment	TAGTATTAC	YTLNE	HLQGFVSF	YCTKGDDL	GPTGGGKS	VVFDDPDL	VASN
AHEaCV-16 [NZ-3310C1-2012]	KM874340	*A. stutchburyi*	TAGTATTAC	YTLHE	HLQGFVSF	YCTKGDDL	GPTGGGKS	VVFDDPDL	VASN
AHEaCV-16 [NZ-3310C3-2012]	KM874341	*A. stutchburyi*	TAGTATTAC	FTINN	HLQGFVSF	YCTKGDDL	GPTGGGKS	VVFDDPDL	VASN
AHEaCV-17 [NZ-2032CO-2012]	KM874345	*A. stutchburyi*	TAGTATTAC	LTNFN	HFQCFAQA	YCSKEGKL	GPAGTGKT	ILFDDVEA	FTSN
AHEaCV-17 [NZ-2032GA-2012]	KM874344	*A. crenata*	TAGTATTAC	LTNFN	HFQCFAQA	YCSKEGKL	GPAGTGKT	ILFDDVEA	FTSN
AHEaCV-17 [NZ-2032TU-2012]	KM874343	*P. subtriangulata*	TAGTATTAC	LTNFN	HFQCFAQA	YCSKEGKL	GPAGTGKT	ILFDDVEA	FTSN
AHEaCV-18 [NZ-4778GA-2012]	KM874346	*A. crenata*	TATGATTAC	FTINN	HLQGYLEL	YCLKTISL	GPTGTGKS	VIIDEFYG	ITSN
AHEaCV-19 [NZ-4942GA-2012]	KM874347	*A. crenata*	TAGTATTAC	FTLFP	HLQGYVEY	YCEKEDSF	GDAGTGKT	LIIDDFYG	ITSN
AHEaCV-20 [NZ-2283TU-2012]	KM874349	*P. subtriangulata*	TAGTATTAC	VTIYF	HFQGYIEF	YCKKHETR	GQTGTGKS	VVFDDIRG	STSI
AHEaCV-20 [NZ-4957GA-2012]	KM874348	*A. crenata*	TAGTATTAC	VTIYF	HFQGYIEF	YCKKHETR	GQTGTGKS	VVFDDIRG	STSI
AHEaCV-21 [NZ-2050TU-2012]	KM874350	*P. subtriangulata*	TAGTATTAC	FTINN	HYQAYIEL	YCFKEDLN	GPTASGKT	LLFDDVHN	FTTN
AHEaCV-22 [NZ-2138TU-2012]	KM874351	*P. subtriangulata*	TAGTATTAC	FTLKK	HYQGRISL	YTTKEDTR	TEGNIGKS	FIIDMPRA	VFTN
AHEaCV-22 [NZ-3009SG-2012]	KM874352	Benthic sedime*nt*	TAGTATTAC	FTLKK	HYQGRISL	YTTKEDTR	TEGNIGKS	FIIDMPRA	VFTN
AHEaCV-23 [NZ-2161TU-2012]	KM874353	*P. subtriangulata*	TAGTATTAC	FTLNN	HFQGYLEM	YVLKTMDP	GPTGTGKS	IVLDEFYG	ITTN
AHEaCV-24 [NZ-2183TU-2913	KM874354	*P. subtriangulata*	CAGTATTAC	FTLNN	HLQGFAIF	YCKKDGDF	ETGGSGKS	FLFNVPRT	VFSN
AHEaCV-25 [NZ-1935SG-2012]	KM874357	Benthic sediment	TAATATTAC	FLTYP	HIHAAVMF	YCKKGDNW	GPSGIGKT	IIFDDMSF	FTAN
AHEaCV-25 [NZ-2250TU-2012]	KM874356	*P. subtriangulata*	TAATATTAC	FLTYP	HIHAAVMF	YCKKGDNW	GPSGIGKT	IIFDDMSF	FTAN
AHEaCV-25 [NZ-2942CO-2012]	KM874355	*A. stutchburyi*	TAATATTAC	FLTYP	HIHAAVMF	YCKKGDNW	GPSGIGKT	IIFDDMSF	FTAN
AHEaCV-25 [NZ-3789GA-2012]	KM874358	*A. crenata*	TAATATTAC	FLTYP	HIHAAVMF	YCKKGDNW	GPSGIGKT	IIFDDMSF	FTAN
AHEaCV-26 [NZ-2311TU-2012]	KM874359	*P. subtriangulata*	TAGTATTAC	FIITV	HWQIICAF	YCHKEETR	GQTGTGKT	VVIDEFTG	ITSN
AHEaCV-27 [NZ-2332TU-2012]	KM874360	*P. subtriangulata*	AAGTATTAC	FTAFD	HLQGYVQF	YCTKTRVN	TKGNTGKS	VCFDLSRT	VFAN
AHEaCV-27 [NZ-3061CO-2012]	KM874361	*A. stutchburyi*	AAGTATTAC	FTAFD	HLQGYVQF	YCTKTRVN	TKGNTGKS	VCFDLSRT	VFAN

		Sample	Nonanucleotide	RCR Motifs			SF3 Helicase Motifs		
Viral Isolate	Accession #	source	motif	I	II	III	WalkerA	WalkerB	Motif C
AHEaCV-28 [NZ-2426SG-2012]	KM874364	Benthic sediment	TTGTATTAC	LGFNF	HKQFWCMF	YCKKEGSY	GEPGSGKS	VLLEDFDF	ITSN
AHEaCV-28 [NZ-3281C3-2012]	KM874362	*A. stutchburyi*	TTGTATTAC	LGFNF	HKQFWCMF	YCKKEGSY	GEPGSGKS	VLLEDFDF	ITSN
AHEaCV-28 [NZ-3281GA-2012]	KM874363	*A. crenata*	TTGTATTAC	LGFNF	HKQFWCMF	YCKKEGSY	GEPGSGKS	VLLEDFDF	ITSN
AHEaCV-29 [NZ-1590TU-2012]	KM874368	*P. subtriangulata*	TAATATTAG	FRYNA	HYQGRLRL	YFQPTVGS	TTGNCGKS	YVVDMPRG	VFTN
AHEaCV-29 [NZ-3425C1-2012]	KM874366	*A. stutchburyi*	TAATATTAG	FRYNA	HYQGRLRL	YFQPTVGS	TTGNCGKS	YVVDMPRG	VFTN
AHEaCV-29 [NZ-3425C3-2012]	KM874367	*A. stutchburyi*	TAATATTAG	FRYNA	HYQGRLRL	YFQPTVGS	TTGNCGKS	YVVDMPRG	VFTN

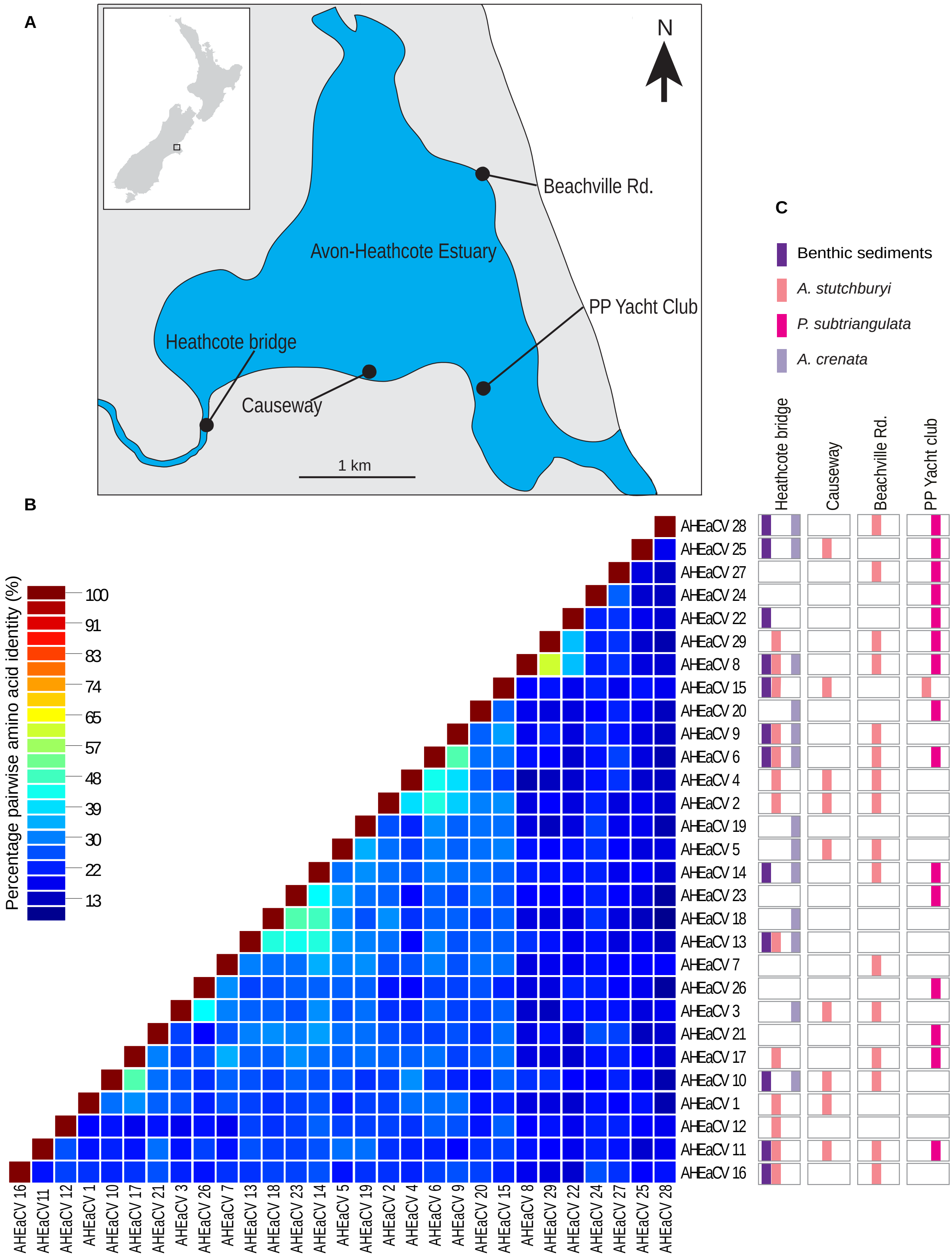


Figure 6.2: (A) Map of the four sample sites at the Avon-Heathcote estuary (B) Pairwise identity plot of amino acid of the Reps of AHEaCVs (C) Isolation source of recovered isolates and location.

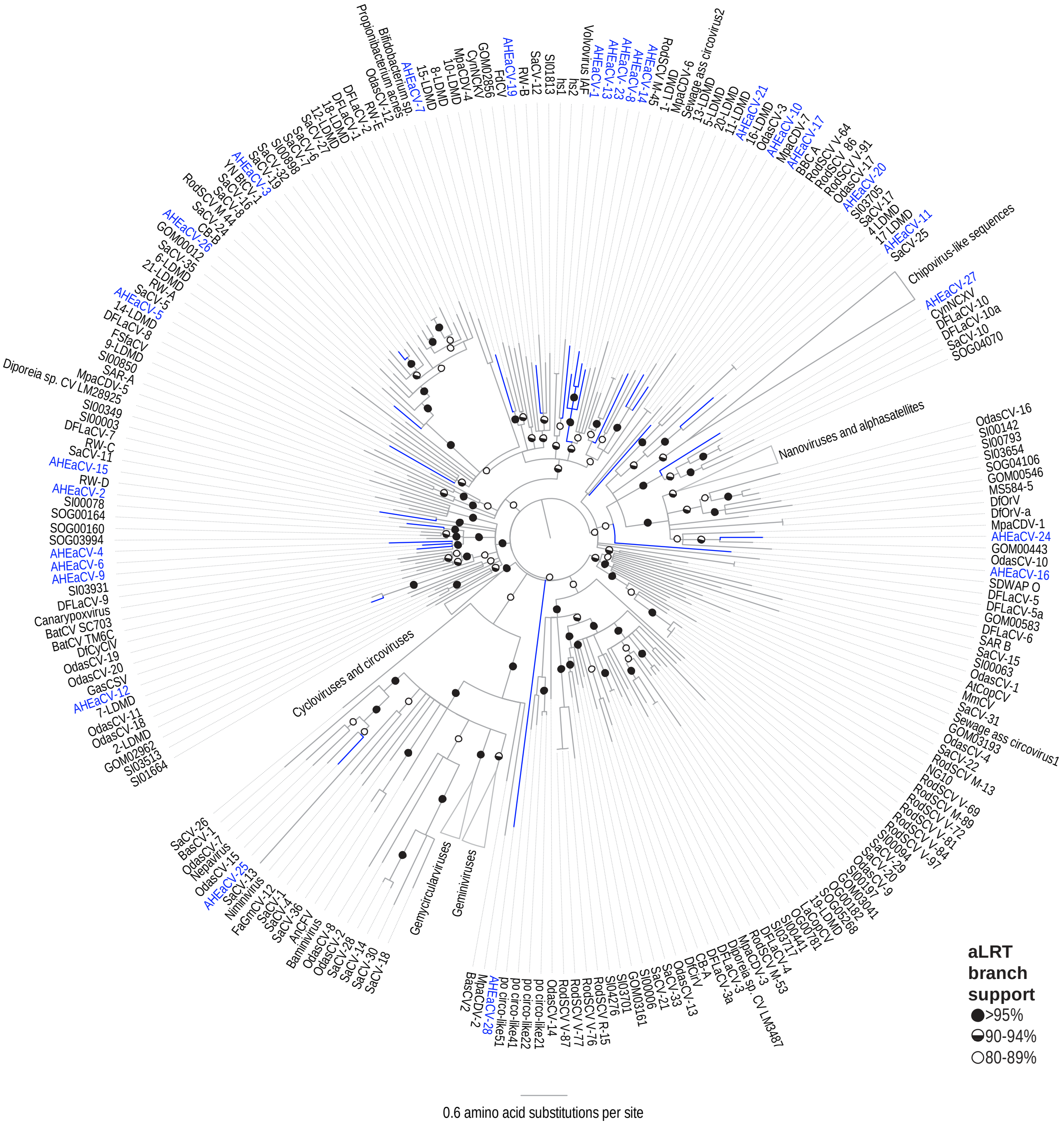


Figure 6.3: Maximum likelihood phylogenetic tree with aLRT branch support of the Rep sequences of all known CRESS DNA viruses. See Table 6.5 for GenBank accession numbers associated with the acronyms used in the figure.

6.4.2 Analysis of the Reps of AHEaCVs

All CRESS DNA viruses have conserved domains in their Reps, including the conserved rolling circle replication (RCR) and superfamily three (SF3) helicase motifs (Rosario *et al.*, 2012b). These motifs were identified in the Reps of all the 79 CRESS DNA viruses recovered in this study (Table 6.2). RCR motifs are located in the N-terminal domain of the Rep, RCR motif I is conserved across most CRESS DNA viruses and is thought to be involved in recognition of iteron sequences in the LIR (Argüello-Astorga & Ruiz-Medrano, 2001). RCR motif II initiates replication by the binding of metal ions via two histidine residues (Heyraud-Nitschke *et al.*, 1995), whereas both RCR II and III initiate DNA cleavage via a conserved tyrosine residue with the binding of the Rep (Steinfeldt *et al.*, 2006). The function of Walker-A is thought to be a nucleotide-triphosphate (dNTP) binding domain, but could potentially be involved with helicase activity of the Rep during RCR. The Walker-B and motif C are involved with controlling the helicase activity which interacts with dNTP and P-loop nucleoside-triphosphate hydrolase (NTPase) binding domains (Hickman & Dyda, 2005)

BLASTp analysis of all the Rep proteins against the NCBI non-redundant protein database showed that the Rep of AHEaCV-12 (recovered from species *A. stutchburyi*) shared the highest amino acid identity to known Rep sequences of all our recovered isolates, with 65% identity to GaCSV (Table 6.3). The Reps of AHEaCV-3 (recovered from species *A. stutchburyi*) shared 60% identity with SI00898 (JX904478) a marine virus isolated from the Saanich Inlet in British Columbia and SaCV-8 (KJ547632) sewage associated virus from New Zealand (Table 6.3). The Reps encoded by the remaining AHEaCV genomes share 30-52% amino acid identity with Reps to those of other CRESS DNA viral sequences (Table 6.3). A summary of the pairwise identities of the Reps of AHEaCVs with all CRESS DNA viruses encoded Reps are provided in Table 6.4.

6.4.3 Analysis of AHEaCV-12 with Rep-like sequences isolated from bacterial genomes

The Rep of AHEaCV-12 (recovered from *A. stutchburyi*) has high similarity to the Rep of GaCSV, which was isolated from *A. crenata* (Dayaram *et al.*, 2013b), sharing 75% pairwise identity at the amino acid level (Figure 6.4 and 6.5). As shown in Dayaram *et al.* (2013b), the Rep encoded by GaCSV, and now that of AHEaCV-12, are most closely related to Rep encoding sequences found in certain bacterial genomes (Figure 6.5) sharing 16-37% pairwise amino acid identity with these bacterial sequences (Figure 6.5). Phylogenetic analysis of the Reps of these two viruses shows that they are nested with Rep-like sequences found in bacterial genomes (Figure 6.4).

Rep-like sequences in bacteria were first identified by Gibbs *et al.* (2006) in six different bacterial genera (*Ruminococcus*, *Coprobacillus*, *Lachnospiraceae, Eggerthella*, *Roseburia,* and *Clostridium*). In addition, viral Rep-like sequences have also been identified in many different eukaryotic species including plants, animals, fungi and protists (Liu *et al.*, 2011). The presence of viral Rep-encoding sequences in these species suggests they have been involved in recombination events involving the horizontal transfer of genetic material from viruses and plasmids to bacteria and eukaryotes (Gibbs *et al.*, 2006).

The isolation of the second CRESS DNA virus from molluscs whose Rep clusters with Rep-like sequences of bacterial origin suggests that there could be or may have been some association between these bacteria and ancestral AHEaCV-12 and GaSCV like viruses or viral sequences. Given that none of the known prokaryotic ssDNA viruses encode replication proteins with a SF3 helicase domain suggests that AHEaCV-12 and GaSCV may represent a novel group of prokaryotic ssDNA viruses. However, taking into consideration the suggestion by Krupovic (2013), that some ssDNA viruses may have evolved from prokaryotic plasmids, it is more likely that AHEaCV-12 and GaSCV may represent a new group of eukaryotic ssDNA viruses that have recently emerged from bacterial Rep-encoding elements.

Table 6.3: Summary of the top hit from the BLASTp analysis of the putative Rep proteins of AHEaCVs.

Isolate	Accession #	Top BlastP hit	Accession #	% Identity	E-value
AHEaCV-1	KM874290	CLV-BR hs2	JX559622	33%	2 x 10^{-32}
AHEaCV-1	KM874291	CLV-BR hs2	JX559622	33%	2 x 10^{-32}
AHEaCV-2	KM874292	SOG00160	JX904075	47%	1 x 10^{-74}
AHEaCV-2	KM874293	SOG00160	JX904075	47%	1 x 10^{-74}
AHEaCV-2	KM874294	SOG00160	JX904075	47%	1 x 10^{-74}
AHEaCV-3	KM874295	SI00898	JX904478	60%	5 x 10^{-113}
AHEaCV-3	KM874296	SI00898	JX904478	60%	5 x 10^{-113}
AHEaCV-3	KM874297	SaCV-8	KJ547632	61%	1 x 10^{-108}
AHEaCV-4	KM874298	SOG03994	JX904139	47%	9 x 10^{-68}
AHEaCV-4	KM874299	SOG03994	JX904139	47%	9 x 10^{-68}
AHEaCV-4	KM874300	SOG03994	JX904139	47%	9 x 10^{-68}
AHEaCV-5	KM874301	LDMD-21	KF133828	40%	6 x 10^{-51}
AHEaCV-5	KM874302	LDMD-21	KF133828	40%	6 x 10^{-51}
AHEaCV-5	KM874303	LDMD-21	KF133828	40%	6 x 10^{-51}
AHEaCV-6	KM874304	SOG03994	JX904139	50%	4 x 10^{-89}
AHEaCV-6	KM874305	SOG03994	JX904139	50%	4 x 10^{-89}
AHEaCV-6	KM874306	SOG03994	JX904139	50%	4 x 10^{-89}
AHEaCV-6	KM874307	SOG03994	JX904139	50%	4 x 10^{-89}
AHEaCV-6	KM874308	SOG03994	JX904139	50%	4 x 10^{-89}
AHEaCV-7	KM874309	LDMD-15	KF133822	41%	2 x 10^{-61}
AHEaCV-8	KM874310	MPSH06775-GM2	GAC77869	37%	3 x 10^{-48}
AHEaCV-8	KM874311	MPSH06775-GM2	GAC77869	37%	3 x 10^{-48}
AHEaCV-8	KM874312	MPSH06775-GM2	GAC77869	37%	3 x 10^{-48}
AHEaCV-8	KM874313	MPSH06775-GM2	GAC77869	37%	3 x 10^{-48}
AHEaCV-8	KM874314	MPSH06775-GM2	GAC77869	37%	3 x 10-^{48}
AHEaCV-9	KM874315	SOG03994	JX904139	46%	3 x 10^{-74}
AHEaCV-9	KM874316	SOG03994	JX904139	46%	3 x 10^{-74}
AHEaCV-9	KM874317	SOG03994	JX904139	46%	3 x 10^{-74}
AHEaCV-9	KM874318	SOG03994	JX904139	46%	3 x 10^{-74}
AHEaCV-10	KM874319	MpaCDV-7	KJ547652	44%	2 x 10^{-69}
AHEaCV-10	KM874320	MpaCDV-7	KJ547652	44%	1 x 10^{-70}
AHEaCV-10	KM874321	MpaCDV-7	KJ547652	44%	2 x 10^{-69}
AHEaCV-10	KM874322	MpaCDV-7	KJ547652	44%	2 x 10^{-69}
AHEaCV-11	KM874323	PCV-2	JX982226	30%	6 x 10^{-34}
AHEaCV-11	KM874324	PCV-2	JX982226	30%	6 x 10^{-34}
AHEaCV-11	KM874325	PCV-2	JX982226	30%	6 x 10^{-34}
AHEaCV-11	KM874326	PCV-2	JX982226	30%	6 x 10^{-34}
AHEaCV-11	KM874327	PCV-2	JX982226	30%	6 x 10^{-34}
AHEaCV-12	KM874328	GasCV	KC172652	75%	7 x 10^{-130}
AHEaCV-13	KM874329	CanCV	KC241983	43%	2 x 10^{-60}
AHEaCV-13	KM874330	CanCV	KC241983	43%	2 x 10^{-60}
AHEaCV-13	KM874331	CanCV	KC241983	43%	2 x 10^{-60}
AHEaCV-14	KM874332	CanCV	KC241983	35%	7 x 10^{-54}
AHEaCV-14	KM874333	CanCV	KC241983	35%	7 x 10^{-54}
AHEaCV-14	KM874334	CanCV	KC241983	35%	7 x 10^{-54}
AHEaCV-14	KM874335	CanCV	KC241983	35%	7 x 10^{-54}
AHEaCV-15	KM874336	SaCV-11	KJ547622	47%	2 x 10^{-75}
AHEaCV-15	KM874337	SaCV-11	KJ547622	47%	2 x 10^{-75}
AHEaCV-15	KM874338	SaCV-11	KJ547622	48%	1 x 10^{-76}
AHEaCV-15	KM874339	SaCV-11	KJ547622	47%	2 x 10^{-75}
AHEaCV-16	KM874340	RaCV	DQ146997	35%	6 x 10^{-31}
AHEaCV-16	KM874341	RaCV	DQ146997	34%	4 x 10^{-29}
AHEaCV-16	KM874342	RaCV	DQ146997	35%	6 x 10^{-31}
AHEaCV-17	KM874343	MpaCDV-7	KJ547652	52%	3 x 10^{-85}
AHEaCV-17	KM874344	MpaCDV-7	KJ547652	52%	3 x 10^{-85}
AHEaCV-17	KM874345	MpaCDV-7	KJ547652	52%	3 x 10^{-85}
AHEaCV-18	KM874346	FiCV	DQ845075	36%	2 x 10^{-53}
AHEaCV-19	KM874347	SaCV-12	KJ547623	43%	4 x 10^{-54}
AHEaCV-20	KM874348	RodSCV V-69	JF755403	35%	2 x 10^{-38}
AHEaCV-20	KM874349	RodSCV V-69	JF755403	35%	2 x 10^{-38}
AHEaCV-21	KM874350	MPSH05440-GM1	GAC77831	46%	2 x 10^{-74}
AHEaCV-22	KM874351	MPSH06775-GM2	GAC77869	51%	1 x 10^{-76}
AHEaCV-22	KM874352	MPSH06775-GM2	GAC77869	51%	1 x 10^{-76}
AHEaCV-23	KM874353	BFDV	HM748921	38%	2 x 10^{-53}
AHEaCV-24	KM874354	GOM00443	JX904231	44%	3 x 10^{-83}
AHEaCV-25	KM874355	Nepavirus	JQ898333	36%	2 x 10^{-46}
AHEaCV-25	KM874356	Nepavirus	JQ898333	36%	2 x 10^{-46}
AHEaCV-25	KM874357	Nepavirus	JQ898333	36%	2 x 10^{-46}
AHEaCV-25	KM874358	Nepavirus	JQ898333	36%	2 x 10-^{46}
AHEaCV-26	KM874359	SaCV-7	KJ547631	44%	3 x 10^{-76}
AHEaCV-27	KM874360	DfLaCV-10	KF738884	37%	1 x 10^{-47}
AHEaCV-27	KM874361	DfLaCV-10	KF738884	37%	1 x 10^{-47}
AHEaCV-28	KM874362	CLV-BR hs2	JX559622	33%	2 x 10^{-32}
AHEaCV-28	KM874363	CLV-BR hs2	JX559622	33%	2 x 10^{-32}
AHEaCV-28	KM874364	CLV-BR hs2	JX559622	33%	2 x 10^{-32}
AHEaCV-28	KM874365	CLV-BR hs2	JX559622	33%	2 x 10^{-32}
AHEaCV-29	KM874366	MPSH06775-GM2	GAC77869	41%	1 X 10^{-48}
AHEaCV-29	KM874367	MPSH06775-GM2	GAC77869	41%	1 X 10^{-48}
AHEaCV-29	KM874368	MPSH06775-GM2	GAC77869	41%	1 X 10^{-48}

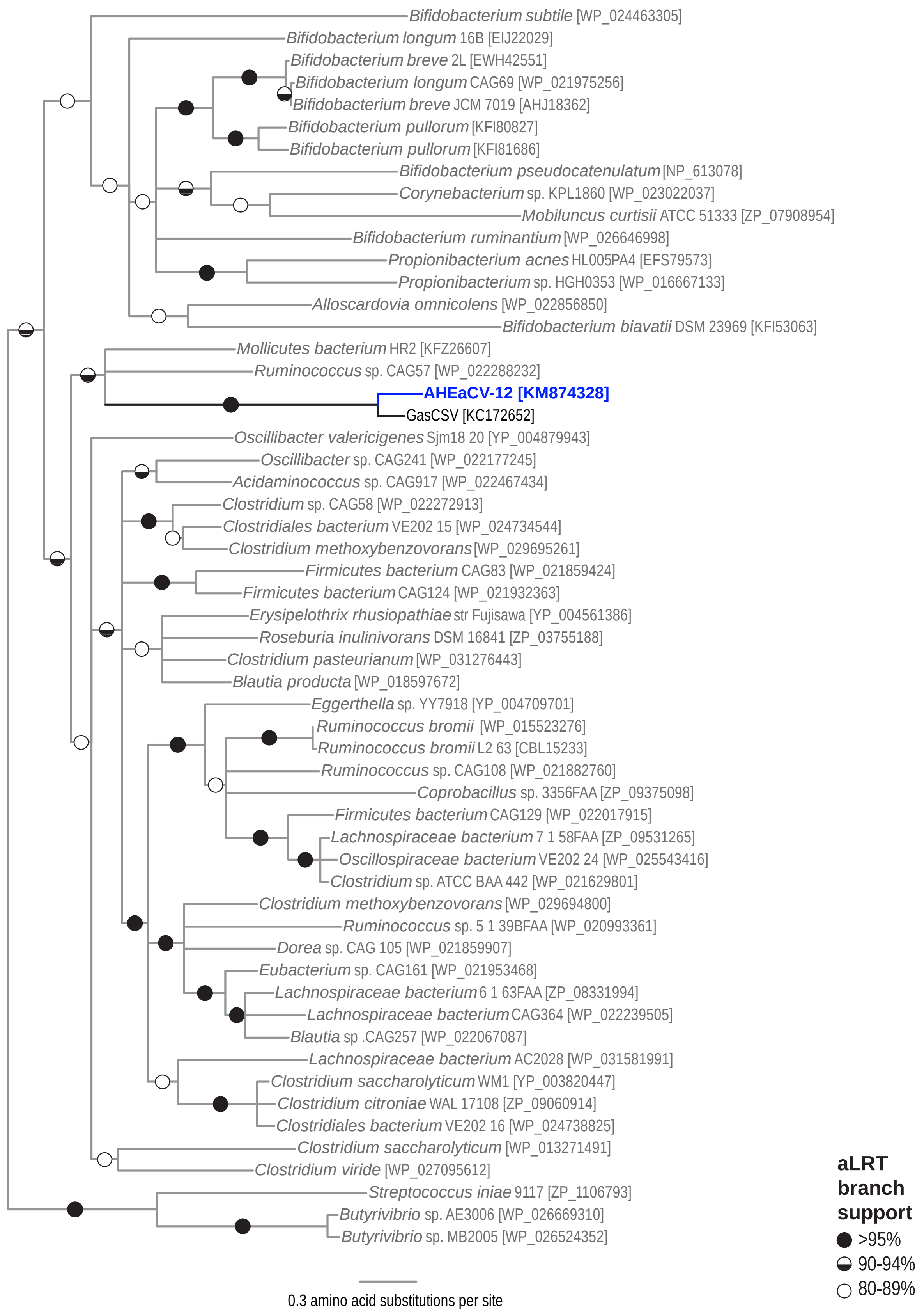

Bifidobacterium subtile [WP_024463305]
Bifidobacterium longum 16B [EIJ22029]
Bifidobacterium breve 2L [EWH42551]
Bifidobacterium longum CAG69 [WP_021975256]
Bifidobacterium breve JCM 7019 [AHJ18362]
Bifidobacterium pullorum [KFI80827]
Bifidobacterium pullorum [KFI81686]
Bifidobacterium pseudocatenulatum [NP_613078]
Corynebacterium sp. KPL1860 [WP_023022037]
Mobiluncus curtisii ATCC 51333 [ZP_07908954]
Bifidobacterium ruminantium [WP_026646998]
Propionibacterium acnes HL005PA4 [EFS79573]
Propionibacterium sp. HGH0353 [WP_016667133]
Alloscardovia omnicolens [WP_022856850]
Bifidobacterium biavatii DSM 23969 [KFI53063]
Mollicutes bacterium HR2 [KFZ26607]
Ruminococcus sp. CAG57 [WP_022288232]
AHEaCV-12 [KM874328]
GasCSV [KC172652]
Oscillibacter valericigenes Sjm18 20 [YP_004879943]
Oscillibacter sp. CAG241 [WP_022177245]
Acidaminococcus sp. CAG917 [WP_022467434]
Clostridium sp. CAG58 [WP_022272913]
Clostridiales bacterium VE202 15 [WP_024734544]
Clostridium methoxybenzovorans [WP_029695261]
Firmicutes bacterium CAG83 [WP_021859424]
Firmicutes bacterium CAG124 [WP_021932363]
Erysipelothrix rhusiopathiae str Fujisawa [YP_004561386]
Roseburia inulinivorans DSM 16841 [ZP_03755188]
Clostridium pasteurianum [WP_031276443]
Blautia producta [WP_018597672]
Eggerthella sp. YY7918 [YP_004709701]
Ruminococcus bromii [WP_015523276]
Ruminococcus bromii L2 63 [CBL15233]
Ruminococcus sp. CAG108 [WP_021882760]
Coprobacillus sp. 3356FAA [ZP_09375098]
Firmicutes bacterium CAG129 [WP_022017915]
Lachnospiraceae bacterium 7 1 58FAA [ZP_09531265]
Oscillospiraceae bacterium VE202 24 [WP_025543416]
Clostridium sp. ATCC BAA 442 [WP_021629801]
Clostridium methoxybenzovorans [WP_029694800]
Ruminococcus sp. 5 1 39BFAA [WP_020993361]
Dorea sp. CAG 105 [WP_021859907]
Eubacterium sp. CAG161 [WP_021953468]
Lachnospiraceae bacterium 6 1 63FAA [ZP_08331994]
Lachnospiraceae bacterium CAG364 [WP_022239505]
Blautia sp .CAG257 [WP_022067087]
Lachnospiraceae bacterium AC2028 [WP_031581991]
Clostridium saccharolyticum WM1 [YP_003820447]
Clostridium citroniae WAL 17108 [ZP_09060914]
Clostridiales bacterium VE202 16 [WP_024738825]
Clostridium saccharolyticum [WP_013271491]
Clostridium viride [WP_027095612]
Streptococcus iniae 9117 [ZP_1106793]
Butyrivibrio sp. AE3006 [WP_026669310]
Butyrivibrio sp. MB2005 [WP_026524352]
0.3 amino acid substitutions per site
aLRT
branch
support
● >95%
◒ 90-94%
○ 80-89%

Figure 6.4: Maximum likelihood phylogenetic tree with aLRT branch support of Rep-like sequences isolated from bacterial genomes including GasCSV and AHEaCV-12

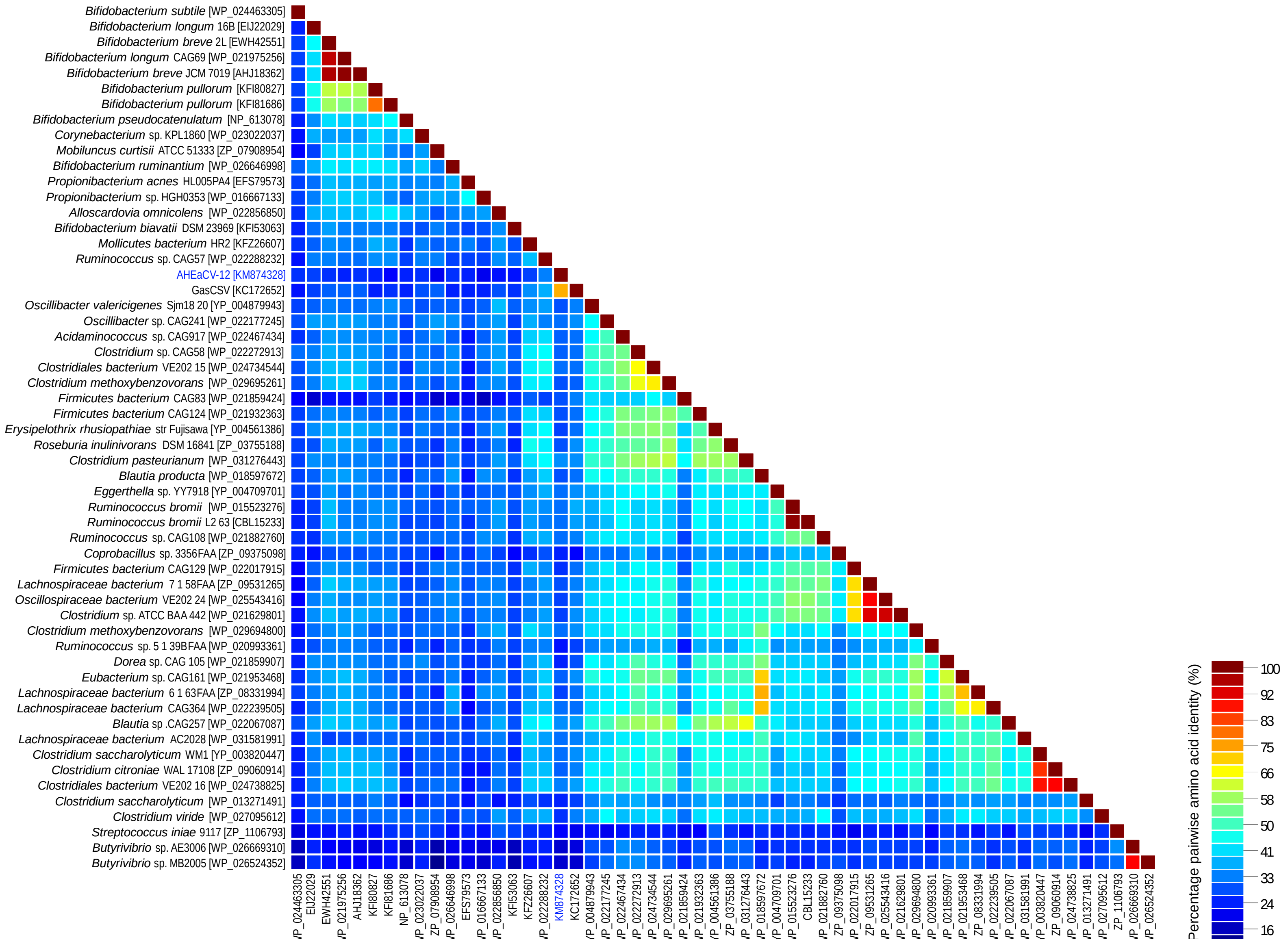


Figure 6.5: Percentage pairwise identity plot of amino acid sequences of the Rep-like sequences isolated from bacterial genomes.

Table 6.4: Top five amino acid pairwise identities of the Reps of AHEaCVs with other CRESS DNA viral Reps.

Sequence query	Sequence	Identity score	Sequence query	Sequence hit	Identity score
AHEaCV-1 [KM874290]	RodSCV-M89 [JF755402]	36.68%	AHEaCV-16 [KM874340]	RaCV [DQ146997]	36.09%
AHEaCV-1 [KM874290]	hs2 [JX559622]	35.71%	AHEaCV-16 [KM874340]	Sewage ass circo [ACY68124]	36.00%
AHEaCV-1 [KM874290]	AHEaCV-17 [KM874343]	35.32%	AHEaCV-16 [KM874340]	CoCV [KF738868]	34.96%
AHEaCV-1 [KM874290]	hs1 [JX559621]	34.96%	AHEaCV-16 [KM874340]	DfCyV-3 [JX185424]	34.40%
AHEaCV-1 [KM874290]	CanCV [KC241982]	34.75%	AHEaCV-16 [KM874340]	FinchCV [DQ845075]	34.34%
AHEaCV-2 [KM874292]	AHEaCV-6 [KM874304]	48.09%	AHEaCV-17 [KM874343]	AHEaCV-10 [KM874320]	52.54%
AHEaCV-2 [KM874292]	SI00078 [JX904407]	46.30%	AHEaCV-17 [KM874343]	MpaCDV-7 [KJ547652]	50.00%
AHEaCV-2 [KM874292]	SOG00160 [JX904075]	46.13%	AHEaCV-17 [KM874343]	AHEaCV-7 [KM874309]	37.74%
AHEaCV-2 [KM874292]	OdasCV-20 [KM598409]	45.90%	AHEaCV-17 [KM874343]	CanineCV [JQ821392]	36.82%
AHEaCV-2 [KM874292]	SI03931 [JX904581]	45.56%	AHEaCV-17 [KM874343]	CanineCV [AFK82575]	36.82%
AHEaCV-3 [KM874295]	SaCV-19 [KM821754]	78.41%	AHEaCV-18 [KM874346]	AHEaCV-23 [KM874353]	52.26%
AHEaCV-3 [KM874295]	SaCV-32 [KM821767]	61.36%	AHEaCV-18 [KM874346]	AHEaCV-14 [KM874332]	50.32%
AHEaCV-3 [KM874295]	SaCV-16 [KM821751]	61.22%	AHEaCV-18 [KM874346]	AHEaCV-13 [KM874329]	47.92%
AHEaCV-3 [KM874295]	SaCV-8 [KJ547632]	59.92%	AHEaCV-18 [KM874346]	CanCV [KC241982]	42.96%
AHEaCV-3 [KM874295]	SI00898 [JX904478]	59.32%	AHEaCV-18 [KM874346]	DuCV [AY228555]	42.60%
AHEaCV-4 [KM874298]	AHEaCV-6 [KM874304]	46.15%	AHEaCV-19 [KM874347]	SaCV-12 [KJ547623]	44.44%
AHEaCV-4 [KM874298]	SI03931 [JX904581]	43.09%	AHEaCV-19 [KM874347]	Sewage ass circo [GQ243674]	41.61%
AHEaCV-4 [KM874298]	AHEaCV-9 [KM874315]	42.91%	AHEaCV-19 [KM874347]	GOM00012 [JX904192]	41.07%
AHEaCV-4 [KM874298]	SOG03994 [JX904139]	42.58%	AHEaCV-19 [KM874347]	RodSCV-M45 [JF755409]	40.10%
AHEaCV-4 [KM874298]	AHEaCV-2 [KM874292]	42.46%	AHEaCV-19 [KM874347]	DfCyV-4 [KC512917]	40.00%
AHEaCV-5 [KM874301]	13-LDMD [KF133820]	43.84%	AHEaCV-20 [KM874348]	Sewage ass circo [GQ243677]	40.27%
AHEaCV-5 [KM874301]	SI00850 [JX904473]	42.97%	AHEaCV-20 [KM874348]	SaCV-12 [KJ547623]	40.25%
AHEaCV-5 [KM874301]	21-LDMD [KF133828]	39.54%	AHEaCV-20 [KM874348]	Human cyclovirus VS5700009 [KC771281]	38.72%
AHEaCV-5 [KM874301]	Sewage ass circo [GQ243677]	38.71%	AHEaCV-20 [KM874348]	CyCV-VN [KF031470]	38.46%
AHEaCV-5 [KM874301]	5-LDMD [KF133812]	38.58%	AHEaCV-20 [KM874348]	CyCV-VN [KF031471]	38.46%
AHEaCV-6 [KM874304]	SI03931 [JX904581]	52.31%	AHEaCV-21 [KM874350]	5-LDMD [KF133812]	39.61%
AHEaCV-6 [KM874304]	AHEaCV-9 [KM874315]	51.89%	AHEaCV-21 [KM874350]	20-LDMD [KF133827]	39.61%
AHEaCV-6 [KM874304]	SOG03994 [JX904139]	50.57%	AHEaCV-21 [KM874350]	PK5034 [GQ404845]	38.58%
AHEaCV-6 [KM874304]	SOG00160 [JX904075]	49.43%	AHEaCV-21 [KM874350]	Sewage ass circo [GQ243677]	38.41%
AHEaCV-6 [KM874304]	SOG00164 [JX904076]	48.47%	AHEaCV-21 [KM874350]	13-LDMD [KF133820]	37.60%
AHEaCV-7 [KM874309]	15-LDMD [KF133822]	43.01%	AHEaCV-22 [KM874351]	AHEaCV-8 [KM874310]	39.41%
AHEaCV-7 [KM874309]	Sewage ass circo [GQ243677]	38.31%	AHEaCV-22 [KM874351]	AHEaCV-29 [KM874367]	38.93%
AHEaCV-7 [KM874309]	FdCV [KC441518]	38.30%	AHEaCV-22 [KM874351]	SI00793 [JX904469]	32.77%
AHEaCV-7 [KM874309]	GOM02856 [JX904312]	38.03%	AHEaCV-22 [KM874351]	MeCMValpha1 [HM163578]	31.54%
AHEaCV-7 [KM874309]	AHEaCV-14 [KM874332]	37.76%	AHEaCV-22 [KM874351]	SI00142 [JX904416]	31.49%
AHEaCV-8 [KM874310]	AHEaCV-29 [KM874367]	63.43%	AHEaCV-23 [KM874353]	AHEaCV-18 [KM874346]	52.26%
AHEaCV-8 [KM874310]	AHEaCV-22 [KM874351]	39.41%	AHEaCV-23 [KM874353]	AHEaCV-13 [KM874329]	46.60%
AHEaCV-8 [KM874310]	ChiSV-GT306 [GQ351278]	36.60%	AHEaCV-23 [KM874353]	AHEaCV-14 [KM874332]	44.44%
AHEaCV-8 [KM874310]	TuSCV [KF880727]	33.91%	AHEaCV-23 [KM874353]	Sewage ass circo [GQ243677]	42.58%
AHEaCV-8 [KM874310]	MpaCDV-8 [KJ547653]	33.46%	AHEaCV-23 [KM874353]	BFDV [ADY62621]	41.67%
AHEaCV-9 [KM874315]	AHEaCV-6 [KM874304]	51.89%	AHEaCV-24 [KM874354]	GOM00443 [JX904231]	44.26%
AHEaCV-9 [KM874315]	SI03931 [JX904581]	50.39%	AHEaCV-24 [KM874354]	SOG04070 [JX904144]	39.24%
AHEaCV-9 [KM874315]	SOG00164 [JX904076]	48.67%	AHEaCV-24 [KM874354]	9-LDMD [KF133816]	37.73%
AHEaCV-9 [KM874315]	SOG03994 [JX904139]	46.59%	AHEaCV-24 [KM874354]	SaCV-21 [KM821756]	37.55%
AHEaCV-9 [KM874315]	OdasCV-20 [KM598409]	46.15%	AHEaCV-24 [KM874354]	MpaCDV-1 [KJ547646]	37.07%
AHEaCV-10 [KM874320]	AHEaCV-17 [KM874343]	52.54%	AHEaCV-25 [KM874355]	OdasCV-15 [KM598398]	42.75%
AHEaCV-10 [KM874320]	MpaCDV-7 [KJ547652]	43.97%	AHEaCV-25 [KM874355]	Nepavirus [JQ898333]	36.33%
AHEaCV-10 [KM874320]	15-LDMD [KF133822]	35.79%	AHEaCV-25 [KM874355]	OdasCV-7 [KM598390]	35.11%
AHEaCV-10 [KM874320]	AHEaCV-4 [KM874298]	35.25%	AHEaCV-25 [KM874355]	BasCV-1 [KM510189]	32.08%
AHEaCV-10 [KM874320]	PKbeef23 [HQ738634]	34.23%	AHEaCV-25 [KM874355]	MpaCDV-2 [KJ547647]	31.75%
AHEaCV-11 [KM874323]	SaCV-25 [KM821760]	37.85%	AHEaCV-26 [KM874359]	SaCV-19 [KM821754]	46.21%
AHEaCV-11 [KM874323]	9-LDMD [KF133816]	37.50%	AHEaCV-26 [KM874359]	YN-BtCV-1 [JF938078]	45.86%
AHEaCV-11 [KM874323]	NG10 [GQ404895]	34.78%	AHEaCV-26 [KM874359]	SaCV-16 [KM821751]	45.86%
AHEaCV-11 [KM874323]	Sewage ass circo [ACY68124]	34.42%	AHEaCV-26 [KM874359]	DFLaCV-2 [KF738874]	45.66%
AHEaCV-11 [KM874323]	SgCV [JQ011377]	34.16%	AHEaCV-26 [KM874359]	SaCV-24 [KM821759]	45.19%
AHEaCV-12 [KM874328]	GasCSV [KC172652]	74.90%	AHEaCV-27 [KM874360]	SOG04070 [JX904144]	40.70%
AHEaCV-12 [KM874328]	SgCV [JQ011377]	36.02%	AHEaCV-27 [KM874360]	SI00793 [JX904469]	38.30%
AHEaCV-12 [KM874328]	DFLaCV-10a [KF738885]	35.15%	AHEaCV-27 [KM874360]	SI00142 [JX904416]	36.64%
AHEaCV-12 [KM874328]	DFLaCV-10 [KF738884]	34.73%	AHEaCV-27 [KM874360]	MVDValpha [AB000921]	35.77%
AHEaCV-12 [KM874328]	NG13 [GQ404856]	34.47%	AHEaCV-27 [KM874360]	GOM00443 [JX904231]	35.31%
AHEaCV-13 [KM874329]	AHEaCV-18 [KM874346]	47.92%	AHEaCV-28 [KM874362]	Sewage ass circo [GQ243677]	27.56%
AHEaCV-13 [KM874329]	AHEaCV-14 [KM874332]	47.22%	AHEaCV-28 [KM874362]	RodSCV-M45 [JF755409]	25.79%
AHEaCV-13 [KM874329]	AHEaCV-23 [KM874353]	46.60%	AHEaCV-28 [KM874362]	SaGmV-5 [KJ547635]	25.57%
AHEaCV-13 [KM874329]	Sewage ass circo [GQ243677]	44.52%	AHEaCV-28 [KM874362]	DFLaCV-3a [KF738876]	24.31%
AHEaCV-13 [KM874329]	CanCV [KC241982]	41.99%	AHEaCV-28 [KM874362]	AHEaCV-16 [KM874340]	24.21%
AHEaCV-14 [KM874332]	AHEaCV-18 [KM874346]	50.32%	AHEaCV-29 [KM874367]	AHEaCV-8 [KM874310]	63.43%
AHEaCV-14 [KM874332]	AHEaCV-13 [KM874329]	47.22%	AHEaCV-29 [KM874367]	AHEaCV-22 [KM874351]	38.93%
AHEaCV-14 [KM874332]	AHEaCV-23 [KM874353]	44.44%	AHEaCV-29 [KM874367]	ChiSV-GT306 [GQ351278]	33.33%
AHEaCV-14 [KM874332]	15-LDMD [KF133822]	41.95%	AHEaCV-29 [KM874367]	TuSCV [KF880727]	32.46%
AHEaCV-14 [KM874332]	Sewage ass circo [GQ243677]	41.94%	AHEaCV-29 [KM874367]	PigSCV [JX274036]	32.44%
AHEaCV-15 [KM874336]	SaCV-11 [KJ547622]	42.78%			
AHEaCV-15 [KM874336]	2-LDMD [KF133808]	39.67%			
AHEaCV-15 [KM874336]	SOG00160 [JX904075]	39.27%			
AHEaCV-15 [KM874336]	AtCopCV [JQ837277]	39.17%			
AHEaCV-15 [KM874336]	SI00850 [JX904473]	39.10%			

6.4.4 CRESS DNA viruses with similarities to chipoviruses

Of the 79 CRESS DNA viruses we recovered, 10 of the viral sequences, AHEaCV-8 (n=5), AHEaCV-29 (n=3) and AHEaCV-22 (n=2), were found to be most similar but distantly related to a group of viruses that we labelled as chipoviruses (n=47; Figure 6.6). These 47 chipoviruses have been recovered from chimpanzee, bovine, turkey and porcine faecal in addition to dragonflies (*Libellula quadrimaculata* and *Erythrodiplax fusca*) (Chapter 5). The genomes of chipoviruses are ~2500 nt and have two major, bi-directionally transcribed ORFs encoding a Rep and CP (with the exception of PisCV -FUJ1, HUN1, GER2011, ANH1, HUN2, HUN1, JIANGX1, HEN1 and HUB2 where the ORFs are uni-directionally transcribed).

The five AHEaCV-8 isolates share between 98-99% CP and 100% Rep pairwise identity (Figure 6.6). AHEaCV-8 isolates and three AHEaCV-29 isolates share 62-63% Rep and ~27-28% CP pairwise identity. AHEaCV-8 only shares ~14-22% CP and 25-39% Rep pairwise identity with other chipoviruses. The three isolates of AVEaCV-29 share >93% Rep and 95-99% CP pairwise amino acid identity (Figure 6.6) and were from *A. stutchburyi* and *P. subtriangulata*. The Reps that were most closely related (~97%) were recovered from different mollusc species. When compared to other chipoviruses, AHEaCV-29 only shared 13-28% CP and 22-39% Rep pairwise identity (Figure 6.6). The two AHEaCV-22 isolates share 100% Rep and CP pairwise identity. AHEaCV-22 shares lower amino acid identity in the CP with other chipovirus sequences (~14-23% amino acid identity) whilst sharing 38-39% amino acid identity in the Rep with isolates of AHEaCV-8 and AHEaCV-29. AHEaCV-22 shares 31% identity in the Rep with Turkey stool-associated circular virus (TuSCV) despite only sharing 17% amino acid identity in the CP.

All Reps sequences of AHEaCV-8, AHEaCV-22 and AHEaCV-28 group together in the Rep-based phylogenetic tree (Figure 6.6) and appear to be distantly related to other chipoviruses sequences. This differs slightly in the CP phylogenetic tree where AVEaCV-22 groups with the chipovirus CPs and AHEaCV-8 and AHEaCV-29 CPs fall basal to this group (Figure 6.6). Taking into account the low pairwise identity to chipoviruses for both the Rep and CP, the smaller genome sizes (~1800 - 2100 nt) and the difference in genome architecture, this suggests that these viral isolates are only distantly related to chipoviruses. The hosts of the viruses are still unknown.

As bivalve molluscs concentrate suspended matter from their surrounding environment, it is possible that these viruses are concentrated in molluscs from faecal matter present in the estuary (Dayaram *et al.*, 2013a).

6.4.5 Viral associations between *P. subtriangulata*, *A. crenata*, *A. stutchburyi* and benthic sediment

When we compared the occurrence of viruses between the three mollusc species and the benthic sediment. We found four viruses, AHEaCV-6 (n=5), -8 (n=5),-14 (n=4), -25 (n=4) and -28 (n=4) across all sample types, with 28% of all the viruses recovered from all four sample types (Figure 6.2). The highest number of viruses (n=37; 46.8%) were found in *A. stutchburyi* whereas in *P. subtriangulata* and *A. crenata* we recovered 16 (20.3%) and 14 (17.7%) of the viruses respectively. From the benthic sediment we were able to recover 12 (15.2%) of the viruses. Therefore, since these viruses have been isolated from a range of different mollusc species, there is a possibility that they are the hosts for many of these recovered viruses. The results from the Pearson's chi-squared test are significant (P=<0.001, Pearson's chi-squared test) indicating the distribution of our viruses does not reflect what we would expect to see with more viruses being isolated from *A. stutchburyi* than the other mollusc species or benthic sediment.

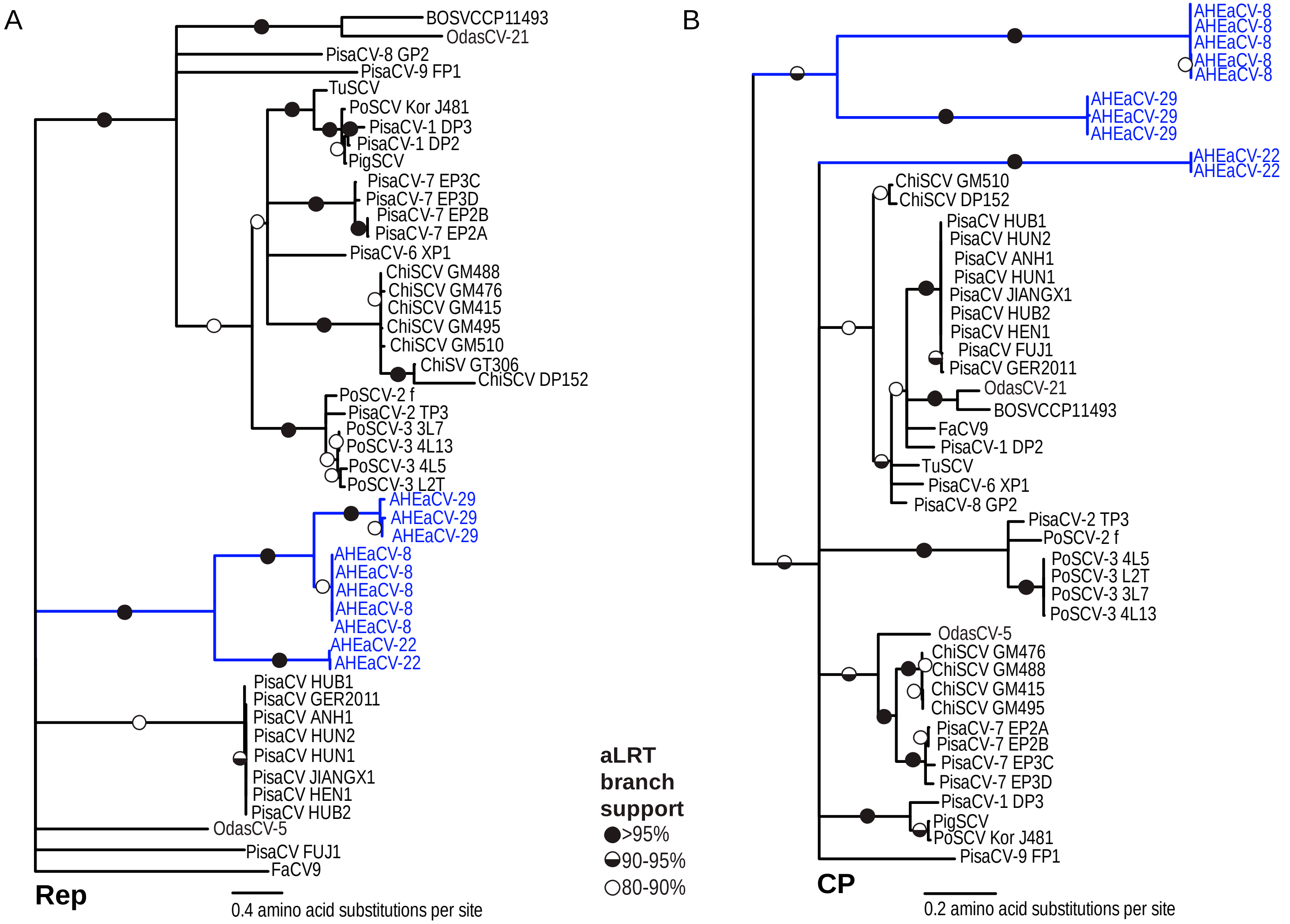


Figure 6.6: Maximum likelihood phylogenetic trees with aLRT branch support. Viral proteins from this study shown in blue. (A) Replication-associated protein of chipoviruses (B) Capsid protein of chipoviruses (figure continued).

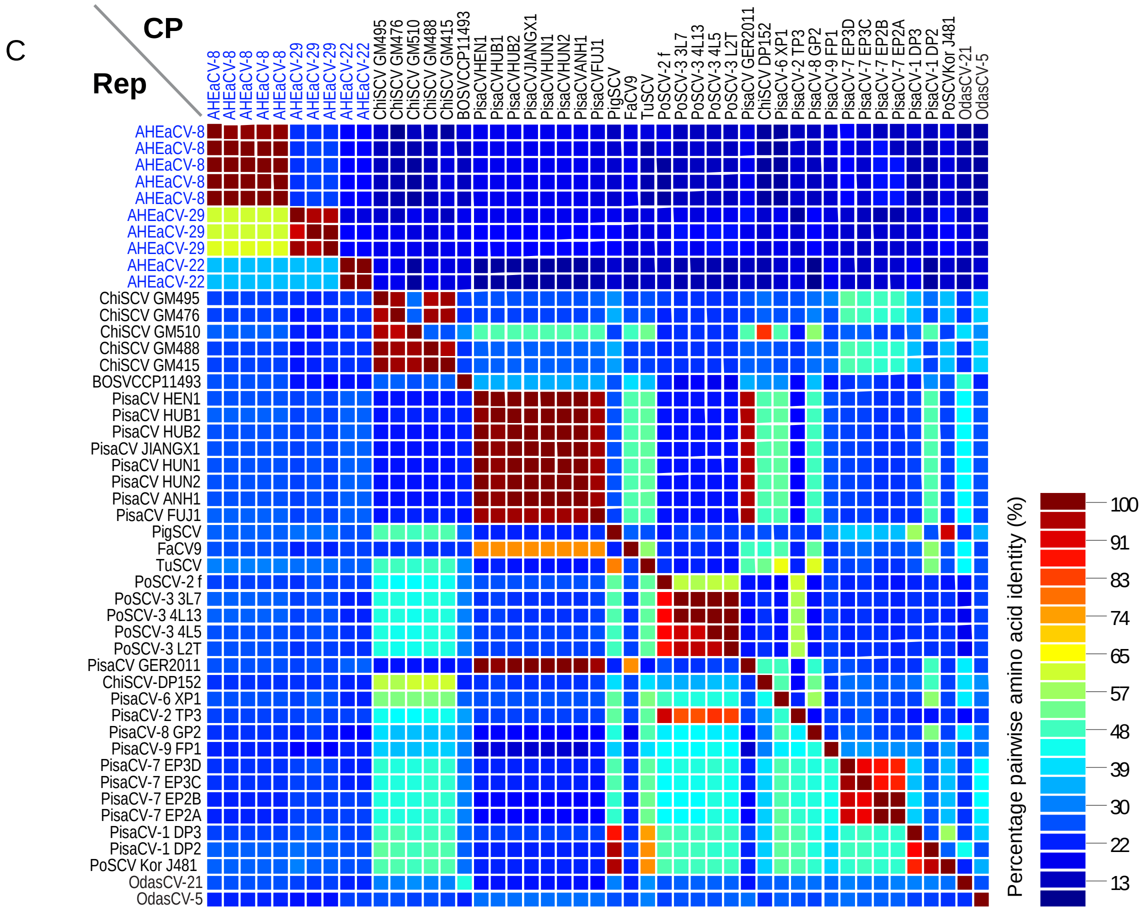


Figure 6.6: (C) Percentage pairwise identity plot of amino acid sequences of the Reps and CPs encoded by chipoviruses. Isolates from this study as shown in blue. See Table 6.5 for GenBank accession numbers associated with the acronyms used in the figure.

Table 6.5: List of CRESS DNA viral replication-associated proteins and their accession number in Figure 6.3 and Figure 6.6.

Acronym	Genbank #	Acronym	Genbank #	Acronym	Genbank #	Acronym	Genbank #
10-LDMD	KF133817	DfCyV-8	KC512920	OdasCV-17	KM598400	SaCV-3	KJ547627
11-LDMD	KF133818	DFLaCV-1	KF738873	OdasCV-18	KM598401	SaCV-3	KJ547627
12-LDMD	KF133819	DFLaCV-10	KF738884	OdasCV-19	KM598404	SaCV-30	KM821765
13-LDMD	KF133820	DFLaCV-10a	KF738885	OdasCV-2	KM598399	SaCV-31	KM821766
14-LDMD	KF133821	DFLaCV-2	KF738874	OdasCV-20	KM598406	SaCV-32	KM821767
15-LDMD	KF133822	DFLaCV-3	KF738875	OdasCV-21	KM598409	SaCV-33	KM821768
16-LDMD	KF133823	DFLaCV-3a	KF738876	OdasCV-3	KM598407	SaCV-34	KM821769
17-LDMD	KF133824	DFLaCV-4	KF738877	OdasCV-4	KM598408	SaCV-35	KM821770
18-LDMD	KF133825	DFLaCV-5	KF738878	OdasCV-5	KM598410	SaCV-36	KM821748
19LDMD	KF133826	DFLaCV-5a	KF738879	OdasCV-7	KM598390	SaCV-37	KM821749
1-LDMD	KF133807	DFLaCV-6	KF738880	OdasCV-8	KM598391	SaCV-4	KJ547628
20-LDMD	KF133827	DFLaCV-7	KF738881	OdasCV-9	KM598392	SaCV-4	KJ547628
21-LDMD	KF133828	DFLaCV-8	KF738882	ODV	AM296025	SaCV-5	KJ547629
2-LDMD	KF133808	DFLaCV-9	KF738883	PanSV-A	L39638	SaCV-6	KJ547630
3-LDMD	KF133810	DfOrV	JX185416	PCV1	AF012107	SaCV-7	KJ547631
4-LDMD	KF133811	DfOrV	JX185417	PCV1-2	FJ655418	SaCV-8	KJ547632
5-LDMD	KF133812	Diporeia sp CV LM28925	KC248425	PCV2	AY424401	SaCV-9	KJ547633
6-LDMD	KF133813	Diporeia sp CV LM3487	KC248416	PeCTV	EF501977	SaGmV-10a	KJ547644
7-LDMD	KF133814	DoYMV	AM157413	PeYDV	EU921828	SaGmV-10b	KJ547645
8-LDMD	KF133815	DSV	M23022	PigSCV	JX274036	SaGmV-12	KJ547642
9-LDMD	KF133816	DuCV	AY228555	PisaCV ANH1	JX305997	SaGmV-2	KJ547642
ABTV	EF546807	DuCV	DQ100076	PisaCV FUJ1	JX305998	SaGmV-3	KJ547643
ACMV	GQ204107	EcmlV	HF921477	PisaCV GER2011	JQ023166	SaGmV-4	KJ547634
ACMV	J02057	ECSV	FJ66563	PisaCV HEN1	JX305991	SaGmV-5	KJ547635
AnCFV	KJ938716	ECSV	FJ66563	PisaCV HUB1	JX305992	SaGmV-6	KJ547636
AnCFV	KJ938716	EMSV	JF508490	PisaCV HUB2	JX305993	SAR-A	FJ959084
AtCopCV	JQ837277	ESV	EU244915	PisaCV HUN1	JX305995	SAR-B	FJ959085
AYVSalpha	AJ416153	FaGmCV-10	KF371632	PisaCV HUN2	JX305996	SCSV	AAA68022
Baminivirus	KJ938716	FaGmCV-11	KF371631	PisaCV JIANGX1	JX305994	SCSV	U16735
Baminivirus	JQ898331	FaGmCV-12	KF371630	PK5006	GQ404844	SCTAV	HQ443515
BarCV	GU799606	FaGmCV-1a	KF371643	PK5034	GQ404845	SDWAP	HQ335074
BarCV	JF279961	FaGmCV-1b	KF371642	PK52222	GQ404846	SDWAPI	HQ335042
BasCV-1	KJ938716	FaGmCV-1c	KF371641	PK5510	GQ404847	Sewage as circo	GQ243674
BasCV-2	KM510191	FaGmCV-2	KF371640	PK6197	GQ404848	Sewage as circo	GQ243677
BatCV SC703	JN857329	FaGmCV-3	KF371639	PKbeef23	HQ738634	SgCV	JQ011377
BatCV TM6C	HM228875	FaGmCV-4	KF371638	PKgoat11	HQ738636	SI00003	JX904394
batCyVGF4c	HM228874	FaGmCV-5	KF371637	PKgoat21	HQ738635	SI00006	JX904395
BBC-A	FJ959086	FaGmCV-6	KF371636	PNYDV	GU553134	SI00063	JX904401
BBTV	EU531473	FaGmCV-7	KF371635	pocircolike21	JF713716	SI00078	JX904407
BBTV	EU531473	FaGmCV-8	KF371634	pocircolike22	JF713717	SI00094	JX904412
BBTV-SAT	L32166	FaGmCV-9	KF371633	pocircolike41	JF713718	SI00142	JX904416
BCSMV	HQ113104	FaGmV-13	KJ938717	pocircolike51	JF713719	SI00197	JX904420
BCTIV	EU273818	FBNSV	GQ150778	PoSCV 2	KC545226	SI00349	JX904427
BCTIV	EU273818	FBNYV	AJ132187	PoSCV 33L7	KC545227	SI00373	JX904431
BCTIV	JX082259	FbSLCV-2	JX094281	PoSCV 34L13	KC545228	SI00441	JX904439
BCTV	M24597	FdCV	KC441518	PoSCV 34L5	KC545229	SI00793	JX904469
BDV	AM922261	FiCV	DQ845075	PoSCV 3L2T	KC5452309	SI00850	JX904473
BeYDV	AM849096	FSfaCV	KF246569	PoSCV-1 DP2	KJ577810	SI00898	JX904478
BFDV	AF071878	FWCasCyV	JX569794	PoSCV-1 DP3	KJ577811	SI01664	JX904518
BGYMV	D00201	GasCSV	KC172652	PoSCV-2 TP3	KJ577818	SI01813	JX904523
BMCTV	AY134867	GCFaV	JQ901105	PoSCV-6 XP1	KJ577819	SI03513	JX904541
BOSVCCP11493	JN634851	GoCV	DQ192280	PoSCV-7 EP2-A	KJ577812	SI03654	JX904548
BSCTV	X97203	GOM00012	JX904192	PoSCV-7 EP2-B	KJ577813	SI03701	JX904559
BtCV	JN377566	GOM00443	JX904231	PoSCV-7 EP3-C	KJ577814	SI03705	JX904561
CaCV	AJ301633	GOM00546	JX904245	PoSCV-7 EP3-D	KJ577815	SI03717	JX904562
Canarypoxvirus	NP955176	GOM00583	JX904250	PoSCV-8 GP2	KJ577817	SI03931	JX904581
CanCV	JQ821392	GOM02856	JX904312	PoSCV-9 FP1	KJ577816	SI04276	JX904605
CB-A	FJ959082	GOM02962	JX904333	PoSCV-Kor J481	KF193403	SiYVValpha	DQ641718
CB-B	FJ959083	GOM03041	JX904344	Propionibacterium	EFS79573	SOG00160	JX904075
CCDaV	JQ920490	GOM03161	JX904368	PSMV	JF905486	SOG00164	JX904076
CFDV	M29963	GOM03193	JX904377	RaCV	DQ146997	SOG00182	JX904077
CGMV	AF029217	GuCV	DQ845074	RhFeCV	JQ814849	SOG00781	JX904107
ChCDV	AM850136	GuCV	JQ685854	RodSCV M 13	JF755410	SOG03994	JX904139
Chimp11	GQ404849	HJasCV	KF413620	RodSCV M 44	JF755408	SOG04070	JX904144
Chimp12	GQ404850	HrCTV	U49907	RodSCV M 45	JF755409	SOG04106	JX904147
chimp17	GQ404851	hs1	JX559621	RodSCV M 53	JF755415	SOG05268	JX904185
ChiSCV DP152	GQ351272	hs2	JX559622	RodSCV M 89	JF755402	SpCTV	AY548948
ChiSCV GM415	GQ351277	HuCyV-5841A	KF726986	RodSCV R 15	JF755401	SPLCV	AF104036
ChiSCV GM476	GQ351274	HuCyV-7046A	KF726987	RodSCV V 64	JF755407	SSCTV	GU734126
ChiSCV GM488	GQ351276	HuCyV-7078A	KF726984	RodSCV V 69	JF755403	SSEV	AF239159
ChiSCV GM495	GQ351273	HuCyV-7081A	KF726985	RodSCV V 72	JF755411	SsHADV-1	KF268025
ChiSCV GM510	GQ351275	LaCopCV	JF912805	RodSCV V 76	JF755404	SsHADV-1	KF268026
ChiSCV GT306	GQ351278	MaMPRV	AY044133	RodSCV V 77	JF755405	SsHADV-1	KF268027
CLCRV	AM501481	MCMValpha	HM163578	RodSCV V 81	JF755412	SsHADV-1	GQ365709
ClGMV	DQ641692	MiCVDL-13	KJ020099	RodSCV V 84	JF755413	SsHADV-1	KF268028
CoCV	KF738868	MiSV	D00800	RodSCV V 86	JF755416	SSMV-1	JQ948051
CoGMV	EU636712	MmCV	JQ085285	RodSCV V 87	JF755406	SSMV2	JQ948052

Acronym	Genbank #	Acronym	Genbank #	Acronym	Genbank #	Acronym	Genbank #
CpCDV	AM933135	MmFV	JN704610	RodSCV V 91	JF755417	SSRV	AF072672
CpCV	GU256530	Mosquito VEM SDBVL-G	HQ335086	RodSCV V 97	JF755414	SSV-A	M82918
CpRV	GU256532	MpaCDV-1	KJ547646	RW-A	FJ959077	StCV	KC846095
CpYV	JN989439	MpaCDV-2	KJ547647	RW-B	FJ959078	TbCSalpha	HQ407396
CSMV	M20021	MpaCDV-3	KJ547648	RW-C	FJ959079	TCTV	GU456685
CyCV	EU056309	MpaCDV-4	KJ547649	RW-D	FJ959080	TGMV	K02029
CyCV-TB	HQ738637	MpaCDV-5	KJ547650	RW-E	FJ959081	TLCNDalpha	JQ041697
CynNCKV	JX908740	MpaCDV-6	KJ547651	SaCV-1	KJ547620	TLCNDV	U1501
CynNCXV	JX908739	MpaCDV-7	KJ547652	SaCV-10	KJ547621	TLCSV	AY044137
CyVN-cs1	KF031471	MpaCDV-8	KJ547653	SaCV-11	KJ547622	TLCYV	AJ512761
CyVN-hcf	KF031466	MS584-5	HQ322117	SaCV-12	KJ547623	TN18	GQ404858
CyVN-hcf1	KF031465	MSRV	JQ624880	SaCV-13	KJ547624	TN25	GQ404857
CyVN-hcf3	KF031467	MSV	AF329881	SaCV-14	KJ547625	TpCTV	X84735
CyVN-hcf4	KF031468	MVDV	AB000920	SaCV-15	KM821750	TSLCV	AF130415
CyVN-hcf5	KF031469	MYMCalpha	FN675297	SaCV-16	KM821751	TuSCV	KF880727
CyVN-ps1	KF031470	Nepavirus	JQ898333	SaCV-17	KM821752	TYDV	M81103
DfaCV-1	JX185430	NG10	GQ404895	SaCV-18	KM821753	TYDV-A	M81103
DfasMV	JX458740	NG12	GQ404854	SaCV-19	KM821754	TYLCSV	X61153
DfCirV	JX185415	NG13	GQ404856	SaCV-2	KJ547626	TYLCV	AJ512761
DfCyClV	JX185418	NG14	GQ404855	SaCV-2	KJ547626	TYLCV	AF311734
DfCyV-1	HQ638049	NGchicken15	HQ738644	SaCV-20	KM821755	USV	EU445697
DfCyV-2	JX185423	NGchicken8	HQ738643	SaCV-21	KM821756	Volvovirus-IAF	KC543331
DfCyV-3	JX185424	Niminivirus	KJ938716	SaCV-22	KM821757	VS5700009	KC771281
DfCyV-3	JX185428	OdasCV-1	KM598393	SaCV-23	KM821758	WDIV	JQ361910
DfCyV-4	JX185425	OdasCV-10	KM598412	SaCV-24	KM821759	WDV	X02869
DfCyV-4	KC512917	OdasCV-11	KM598394	SaCV-25	KM821760	YNBtCV-1	JF938078
DFCyV-5	JX185426	OdasCV-12	KM598395	SaCV-26	KM821761	YNBtCV-2	JF938079
DFCyV-5	JX185427	OdasCV-13	KM598396	SaCV-27	KM821762	YNBtCV-3	JF938080
DfCyV-6	KC512918	OdasCV-14	KM598397	SaCV-28	KM821763	YNBtCV-4	JF938081
DfCyV-7	KC512919	OdasCV-16	KM598411	SaCV-29	KM821764	YNBtCV-5	JF938082

6.5 Concluding remarks

Overall, 79 novel CRESS DNA viruses were recovered from three mollusc species and the benthic sediment of the Avon-Heathcote estuary in New Zealand. These 79 isolates represent 29 putative new viral species. All the viral sequences showed some level of conservation of the Rep with the RCR and SF3 helicase motifs present in all Rep sequences. However, further phylogenetic analysis demonstrated that these viral sequences are highly divergent CRESS DNA viruses. The identification of these viruses in known bioaccumulators such as molluscs suggests they may be concentrating CRESS DNA viruses, and may therefore be useful sampling tools to explore CRESS DNA diversity in different ecosystems. Due to the tidal nature of estuaries, the sample site used in this study (the Avon-Heathcote estuary) offers unique insight into CRESS DNA viruses that are circulating in both freshwater and marine ecosystems.

Eleven of the viral isolates encode proteins most closely related to chipovirus sequences. However, further phylogenetic analysis and pairwise amino acid identities of both the Rep and CP illustrated that these viruses are only distantly related to the previously described chipoviruses. As molluscs are natural concentrators and are often used as indicators of heavy metal contamination due to their ability to bio-accumulate (Berandah et al., 2010), the presence of these viruses could potentially be a result of the same mechanisms, molluscs potentially concentrating animal faecal matter in the estuary ecosystem.

The identification of a second CRESS DNA virus that shows similarities to the previously described GaCSV and Rep-like sequences found in bacterial genomes indicates that these types of viruses are perhaps more prevalent in nature than previously thought. This is further supported by recent studies which discovered novel CRESS DNA viruses in *Acartia tonsa* (marine copepod), *Labidocera aestiva* (worms), and *Farfantepenaeus duorarum* (shrimp) tissue samples (Dunlap *et al.*, 2013; Ng *et al.*, 2013). In addition, a diverse group of CRESS DNA viruses have been identified in aquatic environments (Hewson *et al.*, 2013; Labonté & Suttle, 2013; McDaniel *et al.*, 2014; Yoon *et al.*, 2011).

This study has generated baseline data on CRESS DNA viruses in molluscs and benthic sediments in the Avon-Heathcote estuarine environment. It has also shown that concentrators in ecosystems can be extremely useful in viral surveillance. However, it is clear that it is difficult to build virus – host interactions of novel viruses circulating in ecosystems. Towards achieving such a goal, we have collected preliminary data on viruses circulating in the Avon-Heathcote estuary. The primers used to recover the full genomes could be used as ‘probes’ for a much larger scale study involving larger sample sizes (in the order of 1000s) to determine virus-host interaction networks and flow of these viruses in estuarine ecosystems.

GenBank accession numbers: KM874368 - KM874290

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., Chan, A. M., Haynes, M., Kelley, S. & other authors (2006). The marine viromes of four oceanic regions. *PLoS biology* **4**, e368.

Anisimova, M. & Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic biology* **55**, 539-552.

Argüello-Astorga, G. & Ruiz-Medrano, R. (2001). 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. *Archives of virology* **146**, 1465-1485.

Baršienė, J., Bučinskienė, R. & Jokšas, K. (2002). Cytogenetic damage and heavy metal bioaccumulation in molluscs inhabiting different sites of the Neris River. *Ekologija* **2**, 52-57.

Berandah, F. E., Kong, Y. C. & Ismail, A. (2010). Bioaccumulation and distribution of heavy metals (Cd, Cu, Fe, Ni, Pb and Zn) in the different tissues of Chicoreus capucinus lamarck (Mollusca: Muricidae) collected from Sungai Janggut, Kuala Langat, Malaysia. *Environ Asia* **3**, 65-71.

Cheng, P. K. C., Wong, D. K. K., Chung, T. W. H. & Lim, W. W. L. (2005). Norovirus contamination found in oysters worldwide. *Journal of medical virology* **76**, 593-597.

Comps, M. (1988). Epizootic diseases of oysters associated with viral infections. *American Fisheries Society Special Publication* **18**, 23-37.

Danovaro, R., Corinaldesi, C., Dell'Anno, A., Fuhrman, J. A., Middelburg, J. J., Noble, R. T. & Suttle, C. A. (2011). Marine viruses and global climate change. *FEMS microbiology reviews* **35**, 993-1034.

Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164-1165.

Dayaram, A., Galatowitsch, M., Harding, J. S., Argüello-Astorga, G. R. & Varsani, A. (2014). Novel circular DNA viruses identifiedin *Procordulia grayi* and *Xanthocnemis zealandica* larvae using metagenomic approaches. *Infection, Genetics and Evolution*, 134-141.

Dayaram, A., Goldstien, S., Zawar-Reza, P., Gomez, C., Harding, J. S. & Varsani, A. (2013a). Identification of starling circovirus in an estuarine mollusc (*Amphibola crenata*) in New Zealand using metagenomic approaches. *Genome announcements* **1**, e00278-00213.

Dayaram, A., Potter, K. A., Moline, A. B., Rosenstein, D. D., Marinov, M., Thomas, J. E., Breitbart, M., Rosario, K., Argüello-Astorga, G. R. & other authors (2013b). High global diversity of cycloviruses amongst dragonflies. *Journal of General Virology* **94**, 1827-1840.

Donson, J., Accotto, G. P., Boulton, M. I., Mullineaux, P. M. & Davies, J. W. (1987). The nucleotide sequence of a geminivirus from< i> Digitaria sanguinalis</i>. *Virology* **161**, 160-169.

Dunlap, D. S., Ng, T. F. F., Rosario, K., Barbosa, J. G., Greco, A. M., Breitbart, M. & Hewson, I. (2013). Molecular and microscopic evidence of viruses in marine copepods. *Proceedings of the National Academy of Sciences* **110**, 1375-1380.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797.

Elston, R. (1997). Bivalve mollusc viruses. *World Journal of Microbiology and Biotechnology* **13**, 393-403.

Farley, C. (1976a). Ultrastructural observations on epizootic neoplasia and lytic virus infection in bivalve mollusks. *Progress in experimental tumor research* **20**, 283.

Farley, C. A. (1976b). Ultrastructural observations on epizootic neoplasia and lytic virus infection in bivalve mollusks. *Progress in experimental tumor research* **20**, 283.

Farley, C. A., Foster, W. S., Banfield, W. G. & Kasnic, G. (1972). Oyster Herpes-Type Virus. *Science* **178**, 759-&.

Gibbs, M. J., Smeianov, V. V., Steele, J. L., Upcroft, P. & Efimov, B. A. (2006). Two families of rep-like genes that probably originated by interspecies recombination are represented in viral, plasmid, bacterial, and parasitic protozoan genomes. *Molecular biology and evolution* **23**, 1097-1100.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology* **59**, 307-321.

Gunkel, G. & Streit, B. (1980). Mechanisms of bioaccumulation of a herbicide (atrazine, s-triazine) in a freshwater mollusc *Ancylus fluviatilis* (müll) and a fish *Coregonus fera* (jurine). *Water Research* **14**, 1573-1584.

Hewson, I., Eaglesham, J. B., Höök, T. O., LaBarre, B. A., Sepúlveda, M. S., Thompson, P. D., Watkins, J. M. & Rudstam, L. G. (2013). Investigation of viruses in *Diporeia* spp. from the Laurentian Great Lakes and Owasco Lake as potential stressors of declining populations. *Journal of Great Lakes Research* **39**, 499-506.

Heyraud-Nitschke, F., Schumacher, S., Laufs, J., Schaefer, S., Schell, J. & Gronenborn, B. (1995). Determination of the origin cleavage and joining domain of geminivirus Rep proteins. *Nucleic Acids Research* **23**, 910.

Hickman, A. B. & Dyda, F. (2005). Binding and unwinding: SF3 viral helicases. *Curr Opin Struct Biol* **15**, 77-85.

Hine, P., Wesney, B. & Hay, B. (1992). Herpesvirus associated with mortalities among hatchery-reared larval Pacific oysters, Crassostrea gigas. *Dis Aquat Org* **12**, 135-142.

Kim, K.-H., Chang, H.-W., Nam, Y.-D., Roh, S. W., Kim, M.-S., Sung, Y., Jeon, C. O., Oh, H.-M. & Bae, J.-W. (2008). Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Applied and environmental microbiology* **74**, 5975-5985.

Krupovic, M. (2013). Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. *Curr Opin Virol* **3**, 578-586.

Labonté, J. M. & Suttle, C. A. (2013). Previously unknown and highly divergent ssDNA viruses populate the oceans. *The ISME journal* **7**, 2169-2177.

Le Deuff, R. M., Nicolas, J.-L., Renault, T. & Cochennec, N. (1994). Experimental transmission of a herpes-like virus to axenic larvae of Pacific oyster, Crassostrea gigas. *Bulletin of the European Association of fish Pathologists* **14**, 69-72.

LeDeuff, R. M., Renault, T. & Gerard, A. (1996). Effects of temperature on herpes-like virus detection among hatchery-reared larval Pacific oyster Crassostrea gigas. *Diseases of Aquatic Organisms* **24**, 149-157.

Liu, H., Fu, Y., Li, B., Yu, X., Xie, J., Cheng, J., Ghabrial, S. A., Li, G., Yi, X. & other authors (2011). Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. *BMC Evolutionary Biology* **11**, 276.

López-Bueno, A., Tamames, J., Velázquez, D., Moya, A., Quesada, A. & Alcamí, A. (2009). High diversity of the viral community from an Antarctic lake. *Science* **326**, 858-861.

Maddison, W. & Maddison, D. (2011). Mesquite 2.75: a modular system for evolutionary analysis.

McDaniel, L. D., Rosario, K., Breitbart, M. & Paul, J. H. (2014). Comparative metagenomics: Natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. *Environmental microbiology* **16**, 570-585.

Meyers, T. R., Burton, T., Evans, W. & Starkey, N. (2009). Detection of viruses and virus-like particles in four species of wild and farmed bivalve molluscs in Alaska, USA, from 1987 to 2009. *Diseases of aquatic organisms* **88**, 1-12.

Morley, N. (2010). Interactive effects of infectious diseases and pollution in aquatic molluscs. *Aquatic toxicology* **96**, 27-36.

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

Ng, T. F. F., Alavandi, S., Varsani, A., Burghart, S. & Breitbart, M. (2013). Metagenomic identification of a nodavirus and a circular ssDNA virus in semi-purified viral nucleic acids from the hepatopancreas of healthy *Farfantepenaeus duorarum* shrimp. *Dis Aquat Org* **105**, 237-242.

Ng, T. F. F., Willner, D. L., Lim, Y. W., Schmieder, R., Chau, B., Nilsson, C., Anthony, S., Ruan, Y., Rohwer, F. & other authors (2011). Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. *PloS one* **6**, e20579.

Nicolas, J., Comps, M. & Cochennec, N. (1992). Herpes-like virus infecting Pacific oyster larvae, C. gigas. *Bull Eurr Assoc Fish Pathol* **12**, 11-13.

Price, M. N., Dehal, P. S. & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one* **5**, e9490.

Renault, T. & Novoa, B. (2004). Viruses infecting bivalve molluscs. *Aquatic Living Resources* **17**, 397-409.

Rosario, K. & Breitbart, M. (2011). Exploring the viral world through metagenomics. *Current Opinion in Virology* **1**, 289-297.

Rosario, K., Duffy, S. & Breitbart, M. (2009a). Diverse circovirus-like genome architectures revealed by environmental metagenomics. *Journal of General Virology* **90**, 2418-2424.

Rosario, K., Duffy, S. & Breitbart, M. (2012a). A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Archives of virology* **157**, 1851-1871.

Rosario, K., Nilsson, C., Lim, Y. W., Ruan, Y. & Breitbart, M. (2009b). Metagenomic analysis of viruses in reclaimed water. *Environmental microbiology* **11**, 2806-2820.

Rosario, K., Dayaram, A., Marinov, M., Ware, J., Kraberger, S., Stainton, D., Breitbart, M. & Varsani, A. (2012b). Diverse circular single-stranded DNA viruses discovered in dragonflies (Odonata: Epiprocta). *Journal of General Virology* **93**, 2668-2681.

Rosario, K., Dayaram, A., Marinov, M., Ware, J., Kraberger, S., Stainton, D., Breitbart, M. & Varsani, A. (2012c). Diverse circular ssDNA viruses discovered in dragonflies (Odonata: Epiprocta). *Journal of General Virology* **93**, 2668-2681.

Rosario, K., Marinov, M., Stainton, D., Kraberger, S., Wiltshire, E. J., Collings, D. A., Walters, M., Martin, D. P., Breitbart, M. & other authors (2011). Dragonfly cyclovirus, a novel single-stranded DNA virus discovered in dragonflies (Odonata: Anisoptera). *Journal of General Virology* **92**, 1302-1308.

Sikorski, A., Argüello-Astorga, G. R., Dayaram, A., Dobson, R. C. J. & Varsani, A. (2013a). Discovery of a novel circular single-stranded DNA virus from porcine faeces. *Archives of virology* **158**, 283-289.

Sikorski, A., Massaro, M., Kraberger, S., Young, L. M., Smalley, D., Martin, D. P. & Varsani, A. (2013b). Novel myco-like DNA viruses discovered in the faecal matter of various animals. *Virus Research* **177**, 209-216.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. & Birol, İ. (2009). ABySS: a parallel assembler for short read sequence data. *Genome research* **19**, 1117-1123.

Steinfeldt, T., Finsterbusch, T. & Mankertz, A. (2006). Demonstration of nicking/joining activity at the origin of DNA replication associated with the rep and rep′ proteins of porcine circovirus type 1. *Journal of virology* **80**, 6225-6234.

Suttle, C. A. (2007). Marine viruses—major players in the global ecosystem. *Nature Reviews Microbiology* **5**, 801-812.

van den Brand, J. M. A., van Leeuwen, M., Schapendonk, C. M., Simon, J. H., Haagmans, B. L., Osterhaus, A. D. M. E. & Smits, S. L. (2011). Metagenomic analysis of the viral flora of pine marten and European badger feces. *Journal of Virology* **86**, 2360-2365.

van der Oost, R., Heida, H. & Opperhuizen, A. (1988). Polychlorinated biphenyl congeners in sediments, plankton, molluscs, crustaceans, and eel in a freshwater lake: Implications of using reference chemicals and indicator organisms in bioaccumulation studies. *Archives of Environmental Contamination and Toxicology* **17**, 721-729.

Walsh, K., Dunstan, R., Murdoch, R., Conroy, B., Roberts, T. & Lake, P. (1994). Bioaccumulation of pollutants and changes in population parameters in the gastropod mollusc *Austrocochlea constricta*. *Archives of Environmental Contamination and Toxicology* **26**, 367-373.

Wright, E. A., Heckel, T., Groenendijk, J., Davies, J. W. & Boulton, M. I. (1997). Splicing features in *maize streak virus* virion and complementary sense gene expression. *The Plant Journal* **12**, 1285-1297.

Yoon, H. S., Price, D. C., Stepanauskas, R., Rajah, V. D., Sieracki, M. E., Wilson, W. H., Yang, E. C., Duffy, S. & Bhattacharya, D. (2011). Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* **332**, 714-717.

Yoshida, M., Takaki, Y., Eitoku, M., Nunoura, T. & Takai, K. (2013). Metagenomic Analysis of Viral Communities in (Hado) Pelagic Sediments. *PloS one* **8**, e57271.

Zawar-Reza, P., Argüello-Astorga, G. R., Kraberger, S., Julian, L., Stainton, D., Broady, P. A. & Varsani, A. (2014). Diverse small circular single-stranded DNA viruses identified in a freshwater pond on the McMurdo Ice Shelf (Antarctica). *Infection, Genetics and Evolution* **26**, 132-138.

Chapter 7

High viral diversity of circular replication-associated protein encoding single-stranded DNA viruses in the Lake Sarah ecosystem, Cass Basin, New Zealand

Contents

7.1 Abstract

Over the past decade there has been an increase in environmental metagenomic studies focusing on viruses in various ecosystems. Some studies have concentrated on organisms in both aquatic and terrestrial habitats that serve as useful ‘sampling tools’ for viral surveillance. These studies have led to the discovery of a large number of circular replication associated protein (Rep) - encoding single-stranded (CRESS) DNA viruses. In this study we identify 139 novel CRESS DNA viruses whose complete genomes were recovered from benthic sediment, water, *Echyridella menzeisi*, *Procordulia grayi*, *Xanthocnemis zealandica, Musculium novazelandiae,* oligochaeta, *Chironomus zealandicus, Potamopyrgus antipodarum* and *Physella acuta* present in Lake Sarah situated in the Cass Basin, New Zealand. Analysis of the CRESS DNA viral genomes recovered show they are all highly diverse. Nonetheless, some of these viruses together with ones previously identified in other studies are beginning to form groups of closely related sequences. We note that molluscs and benthic sediment are great ‘sampling tools’ for surveillance and assessing diversity of CRESS DNA viruses in some ecosystems.

7.2 Introduction

Chapter 7 looks at using metagenomic methods to explore CRESS DNA viral diversity within the Lake Sarah ecosystem, whilst also establishing relationships between sample types and CRESS DNA viral distribution. Viruses are the most abundant entities on Earth, and play a significant role in regulating organisms at all trophic levels in environments (Breitbart & Rohwer, 2005; Koonin *et al.*, 2006; Suttle, 2007). However, understanding viral dynamics and the 'flow' of viruses in ecosystems amongst different organisms has been difficult to examine until recently. The introduction of viral metagenomic methods has transformed how viruses, both novel and known, are discovered and characterised. These techniques have enabled a wealth of novel circular replication associated protein (Rep) encoding ssDNA (CRESS) viruses to be discovered from a range of different environments including sea water (Angly *et al.*, 2006; Labonté & Suttle, 2013), deep-sea vents (Yoshida *et al.*, 2013), rice paddy field soil (Kim *et al.*, 2008), Antarctic lakes (López-Bueno *et al.*, 2009), aquifers (Smith *et al.*, 2013), reclaimed water (Rosario *et al.*, 2009b) and fresh water lakes (Roux *et al.*, 2012). These metagenomic methods have led to a better understanding of viral communities within environmental samples.

In addition, studies have also examined the diversity of CRESS DNA viruses in various ecosystems by looking at different insects (Ng *et al.*, 2013; Ng *et al.*, 2011a; Ng *et al.*, 2011b; Padilla-Rodriguez *et al.*, 2013; Rosario *et al.*, 2012b; Rosario *et al.*, 2011) (Chapter 2, 3 and 4). Vector enabled metagenomics (VEM) have been employed to explore the role insects play in the transmission of geminiviruses, whilst also shedding light on CRESS DNA viral diversity (Ng *et al.*, 2011a; Ng *et al.*, 2011b; Rosario *et al.*, 2014; Rosario *et al.*, 2013). A recent study demonstrated that plant pathogens such as geminiviruses and their related satellites can be detected in insect predators such as dragonflies (Rosario *et al.*, 2013), whilst further studies on dragonflies and their larvae have identified various novel CRESS DNA viruses (Dayaram *et al.*, 2014a; Dayaram *et al.*, 2013c; Rosario *et al.*, 2012b; Rosario *et al.*, 2011).

In addition, animal-infecting viruses such as the *Starling circovirus*, have been detected in natural bioconcentrators such as molluscs (Dayaram *et al.*, 2013a). Diverse CRESS DNA viruses have been shown to be associated with different types of molluscs as well as crustaceans and prophages (Dayaram *et al.*, 2013b; Dunlap *et al.*, 2013; Hewson *et al.*, 2013; McDaniel *et al.*, 2014; Ng *et al.*, 2013; Yoon *et al.*, 2011) (Chapter 5). These studies have demonstrated that insects and other invertebrates can potentially be used to monitor plant pathogens and may also act as viral reservoirs for ecosystems potentially due to mechanisms that enable them to accumulate.

The use of viral metagenomics is not limited to the discovery of novel viruses, it can be useful for understanding viral dynamics in different ecosystems. Examining virus - host interactions in ecosystems as a whole is a difficult concept to understand without first having the baseline data CRESS DNA viruses circulating in an ecosystem, so it is not possible to identify these viruses in other ecosystems. There are only a handful of studies to date that have used metagenomic approaches to create viral networks looking at evolutionary relationships and host - virus interaction (Labonté & Suttle, 2013; Rozenblatt-Rosen *et al.*, 2012).

Most viral metagenomic studies have only examined viral diversity within a specific niche within an ecosystem. This study looks at the CRESS DNA viral diversity across different sample types (water, benthic sediment, oligochaet worms, midge larvae, dragonfly and damselfly larvae, cockles, gastropods and muscles) across the Lake Sarah ecosystem gain some insight into the ‘flow of viruses in the ecosystem by attempting to create viral networks and access what sample types may potentially act as ideal ‘sampling tools’ for the surveillance of CRESS DNA viruses.

7.3 Materials and methods

7.3.1 Sample collection

Samples were collected from Lake Sarah, near Cass in the South Island of New Zealand on the 11th of December 2012. Samples collected from the lake bottom were *Echyridella menzeisi* (mussels) (n=3), dragonfly larvae *Procordulia grayi* (n=17) and damselfly larvae *Xanthocnemis zealandica* (n=22). *Musculium novazelandiae* (cockles) (n=46), worms Oligochaeta (n=33) as well as midge larvae *Chironomus zealandicus* (n=27) and two types of gastropods *Potamopyrgus antipodarum* (n=26) and *Physella acuta* (n=33), 40ml of benthic sediment and 40ml of water were collected from the lake. Samples were refrigerated upon collection. All samples were washed in sterile distilled water prior to processing. In addition, we collected water and benthic sediment samples.

7.3.2 Viral DNA extraction and enrichment of circular DNA

Samples were grouped according to species and homogenised in SM buffer [0.1 M NaCl, 50 m M Tris/HCl (pH 7.4), 10 mM MgSO4] at a ratio of 10 ml buffer per 5 g of tissue and centrifuged (10,000 x g for 10 min) to pellet tissue debris. The supernatant was then filtered first through a 0.45 µm, followed by a 0.2 µm syringe filter (Sartorius Stedim Biotech, Germany). Circular DNA present in the viral DNA extract was then preferentially amplified using rolling circle amplification (RCA) using TempliPhi 2000 (GE Healthcare, USA) as previously described (Dayaram *et al.*, 2014a; Dayaram *et al.*, 2013b; Dayaram *et al.*, 2012; Rosario *et al.*, 2012b; Rosario *et al.*, 2011; Sikorski *et al.*, 2013a).

7.3.3 Sequencing viral DNA, de novo assembly and identification of viral sequences

The enriched select RCA products were then combined: *P. grayi* and *X. zealandica* larvae samples, the midge larvae *C. zealandicus* with the Oligochaeta samples. In total, DNA from seven samples were prepared and sequenced on an Illumina HiSeq 2000 (Illumina, USA) platform at the Beijing Genomics Institute (Hong Kong). The resulting paired end reads were then *de novo* assembled using ABySS V1.3.5 (Simpson *et al.*, 2009) with kmer =64. The *de novo* assembled contigs >500nts with similarities to CRESS DNA viral proteins were then identified by BLASTx analysis using KoriBLAST v4.1 (Korilog SARL, Bioinformatics Solutions, France) and this process was repeated for all seven datasets.

7.3.4 Recovery of CRESS DNA viral genomes

Based on the BLASTx analysis, full genomes were then verified by PCR using back-to-back primers listed in Table 7.1 and Kapa HiFi HotStart polymerase (Kapa Biosystems, USA). The fragments were gel purified and cloned into pJET1.2 (Thermo Fisher, USA) plasmid vector and Sanger sequenced at Macrogen Inc. (South Korea) by primer walking. Viral genomes were then assembled from the Sanger sequencing reads using DNA Baser Sequence Assembler (version 4.16; Heracle Biosoft S.R.L., Romania). An initial analysis of the full genomes was then carried out using BLASTx and tBLASTx (Altschul et al., 1990) followed by annotation of the genomes including identifying the major open reading frames (ORFs), conserved motifs and putative stem-loop structures.

7.3.5 Analysis of the replication-associated protein

A data set was created using all Reps from all CRESS DNA viruses available on GenBank (downloaded 29th October 2014). These viral reps, including the Reps from the viruses described in this study, were aligned using MAFFT (Katoh & Standley, 2013) and MUSCLE (Edgar, 2004) with manual editing. The alignment generated was used to deduce a maximum likelihood phylogenetic tree using a JTT + CAT model with approximate likelihood ratio test (aLRT) branch support (Anisimova & Gascuel, 2006) using FastTree version 2.1.7 (Price *et al.*, 2010). Branches with less than 80% aLRT support were collapsed using Mesquite (version 2.75) (Maddison & Maddison, 2011).

This method was again used to create maximum likelihood phylogenetic trees for both the Rep and CP of one sequence with significant similarities to a group of viruses we have tentatively named chipoviruses (chimpanzee and porcine viruses).

Table 7.1: Details of primer sequences used to recover viral genomes in this study.

Sequence	Forward primer	Reverse primer
LSaCV-1	CCCAAACGAGAACGATGGCTAATAACCAG	TACTCGTTCCAGTTGCCCGCGGAT
LSaCV-2	CGCCATATTCGACGATTTTGAGGACTGG	TACTCCACGTCGTCCCTTATAAGGTCCA
LSaCV-3	ATTCTGCGGACGACATGTTCTACAATGATTGCG	CAACCGAGGTCAACCCATCTACTGGTC
LSaCV-4	TCTCCCGGTAGGTTGGAAACTGACAG	GCAACCTTCGTTTCCTCATCCGTCGG
LSaCV-5	ATTGTGTTCCAAGAAGAGCGCGGCGC	GTACTTAAAACTCCCTGCATCAGCTCCATC
LSaCV-6	ATTCCGTGCCACTGCTAAGAACTTTGCC	TTCTTCTTAGGGCTCATCACTCCGAGATCA
LSaCV-7	CGAAGCCGAGCGTTCATCTTTACCTGG	ACTGGGAGGCATTGAGAAGACATTGATTTTCAG
LSaCV-8	GAAAAATACCCCCCGAAATTCCCCGAAAATG	GCGGAATTTCACGGCAGGGGGAAG
LSaCV-9	GGTGATTCAGGAACTCCTCATCTCCAAGG	GACCTCTCGGCCGATGATCCAGTATTTG
LSaCV-10	CAACGACTATTATGAAAATGGGGAGCTACCTG	TTGTCTTTCTTACAATAAGTTGCAGCTTGT
LSaCV-11	AATTGGACTCTGGCTGACTATGACAAAGTATTGG	ATTTATAGTGAACGTCCAGCGCTTCGCTG
LSaCV-12	CGATGTCAGATTTATTGGGTACGCCCA CG	GTGATGGATTGAATGTATGTTTCAATGGCGTCG
LSaCV-13	GGGATTTGTTCAAGACTATTGCGTCGCC	ATTGACTCTCAAAAGCCGTGAAACCCCATC
LSaCV-14	CATCTGCAAGGATTCCTCCACTTCAAAAATCCC	AGGGGTACCGGTCGTCGGCG
LSaCV-15	TGGGGCTGGCAAACCGAAAAGTGCC	CTCAGCAATCTCTGGTGGCATGGTCTC
LSaCV-16	TTTGAGACCATGCCACCAGAGATTGCTG	AAGATGCCATTGTCCCTCGTATGCCG
LSaCV-17	AATTTGGAACCTGGGAACCAACCTTTGCAC	CTTCCGGGGGAGTAACTTGAGTATCTTC
LSaCV-18	CGCAGTATACGTTATTCAATGAGATGCCTGC	ACTCAAATGCGGTGAACGCCCAACG
LSaCV-19	CACCTCATTTGCAAGGTTATATTGAACTCCATCAC	TGCCTTCTGCGCCCACCTCAAACC
LSaCV-20	GATGCGCTGCACAAGCAGAGCTCTG	TAGAATAGGGTACTGGGTCTCATAAGCGGTAG
LSaCV-21	GGGCTAACTTCAAGCCACTGACGAGTG	TCGGGAAGAGTTTCTTCACGAAGGAGAGAC
LSaCV-22	CAGACGAGGACAACATAGCGCAGGC	CTGGGATGGGGTTATTTATCGTTACACTCCAAC
LSaCV-23	GAGGATGATTGTAATCGGCTCCGAACCC	CGTATAGTTATTCAAAGTGAAACATACGTTGCC
LSaCV-24	CCGGCTGGAAGCTTACCGGTCAG	GGATCTTGGTCCACGCATTAACTTCCTCC
LSaCV-25	CGTGAAAACGAAGCATATTGCAGTAAGGAAGGC	AAAATTGCCACGCATTATTTCAATATGCGCTGG
LSaCV-26	GAGGCTTCCATTAGATGGGGATGCGTG	GTAAAGAGCCTCTTTCGTAAGGGTTGTCTGC
LSaCV-27	CTTCCTCAGTCATGGGATCCTGATTCTTATGAC	AAGCACATCCAAATCGTCAGGATTATTCAAAGC
LSaCV-28	GCAAATCCTGATTTGCAGTTATCGTCTCGTAC	AACATCATCACCACCAATGATACCACCTCC
LSaCV-29	TTGGGGAACCGCGTGACTCTGAAAAG	AAGTCCAAGGACCATCGGCTCTGGT
LSaCV-30	GCGAGTACCCTTTGAAGACCGCTG	TCAAGCTGTCCTCGGATATATTGAACTCCG
LSaCV-31	CCTCCAGCTTGCTGGAGCAACTTATCTTA	TCTTCCCAGTCGAGGAGACCTTCGTA
LSaCV-32	ACGGCGCCCAAGACGATGATGTACA	GAACACGATGGCGCCCTGTATGTG
LSaCV-33	AACCCTTTGGAGTACTCCCAGACCTTTAC	GTAGAACGTGTCCATGTTGAGCCAACG
LSaCV-34	ATCTCGATACTATGTCGTCGGAGTTGAGG	AGATCTGGTAATCTCCCGATAATGGCTGAG
LSaCV-35	CATGTTCCGTCGTGATCTTTCCACAGA	ATGTCGTACATTTCGACCTGTACAGGTGC
LSaCV-36	GGGAACTCCTCATCTTCAAGGCTATATCAG	GTCTCGCCAACTTCTCTTCCAATGACTG
LSaCV-37	CCTTGATTGGTTACCTGGTGTACATTTGGAG	ATACGACCAGCACGACACTGGGAG
LSaCV-38	AGTCGCACGTTTTGAATCGCACATGCG	TCGCTGTCATCATAGTTGTTGAGTGTCATACAC
LSaCV-39	CAGCTCCGTTGGCAAAGAGCTAGAGAAT	CTCTGCTCGACCTTTGTTATCATTAGTAACTGG
LSaCV-40	GCACATGGTGTGGAACTCTGAACAACTACT	GACTACTCTGTGGACCTCCCCCA
LSaCV-41	GAAGTCTGTGCTCAACTCGAACTTGGAACT	ATTTGGAGAGGGGATGGTACAGAGCCA
LSaCV-42	CATCCTATTGGTACTGCTGAACCTCCAAGT	CAGTAACCAGTCACAGTCATCACCCTCA
LSaCV-43	AACGCATGGGTGCTCTGCATAAGCTC	TGGGCTTAGCGAAGTTGACATATCCTTGGA
LSaCV-44	AGAAGTGCTGCAGCAATGGAGTATGTTTGG	TGTGGGCTCAATGTGAGCTTCATCTCCA
LSaCV-45	AAGCGTGGCAACCTGCACCTACTACAT	GTAAGGAAAGCAGCCAGCTCTGCG
LSaCV-46	ACGGTGACTACACAGTAGGCCAGG	ACTGTGGTATGACATAGGGCTCGTCAGA
LSaCV-47	CGAAGGGGTCTGACCAAGAGAACGA	CCGGCTCATTATGAGAAGGCTCACCA
LSaCV-48	GACGAACCGTCCTGGACAGACACA	GTGCTCAGGGTTGTTGTCTGTGAAGATC
LSaCV-49	ATGCCATTGGGAGCTCACCAGAAGTGA	GTGGGGAGCCACTTCTTAAGCTTTGC
LSaCV-50	GGGAACCCAACAACAAGCAATAGACTATTGC	TTTCTGGGTTCCCAATGAATTTTTGGAAAGTCCG
LSaCV-51	GCTTTTGACGCACCATTTGCATCAGGGG	GGCGCCTTTATACTTGTGGTGACCCAG
LSaCV-52	GTCGCACATGGTGTGGAACTCTGAACAA	TACTCTGTGGACCTCCCCCAGTG
DflaCV-3	GACTGGCACTGGGAAATCTAGAACT	GAACCCCAGAACACATAGCAGG
DflaCV-5	CTAACCCACAAATGAATAATGTTGCC	GGTAATGGATGATGTGATAGATGG
DflaCV-6	AATGGTGCACGGATAGGATTG	TTTGTCTTACGGATCCAACG
DflaCV-8	GAGAGCAGAGGGACATAGTACGG	CGGGTAGGGATAGTACGGGGTA
AHEaCV-8	ACTAGTAAAACAGAGGCGACGAAGG	CGGTTCAAGATAGTTAGGGGGC
LSaCM-1	GGATATTCGTCCTGAACAACCCGCG	AGTACTTGGTCTTCCCTGCGTAGTCAG
LSaCM-2	GGAAGAACCACTCTTCTCTTATCTAATCATAGGC	ATGAGTGCTCTGTAGTCTTCTTCAGTCCAATTAC
LSaCM-3	GGTGAGGGATTCACAGTCCCAGACC	TTCAAGCGTATATGATGTGAACATAAATCTACG
LSaCM-4	CTATCCAAGCTTGGAGCTATGCTACCAAGG	GAGTACCATCGGCAGGCCGGAG
LSaCM-5	TACAAGGCGTAATCTGCTTCAACGAAAAAATTCGGG	GATGCGGAGTTCCGGATTCTCCGAC
LSaCM-6	CGTCTCCAGAGGCTGGTGCTTTACC	AGTGGTTTCTCGATTGATACGCTAGGGATGG
LSaCM-7	TACACTGCTTGGAAAGTACCAACTATGGGTGG	GCACACATTTCTAAATTGTTGAGGTCTTTTAGCG
LSaCM-8	CTAGACAACTCGGCGTGCGTTGTATTGC	ATCAGCTGCCGCAGCCGATTACGTG
LSaCM-9	CGCTCGTTAAAGTGTGCATATCAAGTGTTCG	AAGCGAGTCTCTCTCTTCTTCCGTGTAATTG
LSaCM-10	TCACTCTTTCGCGAAGCTATAGCAATCAGGC	GGCCTGAGACGACCTTAAATCGTCTCT
LSaCM-11	CAGTGGAAGAACAAGAAACGTAAGCTTGGCG	CCATTGTTGTTTCTGAGCTGCTGCTCTCATTTG

The Rep and CP of this virus were assembled into datasets with those from pig-stool-associated circular ssDNA virus PisaCV), bovine-stool-associated circular virus (BoSVC), porcine-stool-associated circular virus (PoSCV), turkey-stool-associated circular virus (TuSCV), chimpanzee-stool-associated circular virus (ChiSV) and Odonata-associated circular virus 5 and 21 (OdasCV-5, -21) aligned using MUSCLE (Edgar, 2004).

Further BLASTp analysis was carried out to compare the recovered putative Reps from the viral genomes to the NCBI non-redundant protein database. This was repeated for the putative coat proteins (CP). All pairwise identities were calculated using SDT v1 (Muhire *et al.*, 2014).

7.3.6 Viral distribution analysis based on functional feeding groups of samples

Relationships between the different samples based on their virus assemblages were analysed between all sample types, but also grouped into potentially important ecological groupings. One grouping, functional feeding groups (FFG), was based on known feeding mechanisms of the different organisms. These FFG groupings were browsers (Oligochaeta, *C. zealandicus*, *P. acuta* and *P. antipodarum*), filters (*E. menzeisi* and *M. novazelandiae*) and predators (*P. grayi* and *X. zealandica*), along with the potential environmental virus sources, water and benthic sediment. The second ecological grouping divided potential prey species into those that had morphological defenses (e.g., hard shells) or undefended (e.g., soft bodies) which may influence predator feeding and potential virus transfer. These grouping were water, benthic sediment, defended (*E. menzeisi*, *M. novazelandiae*, *P. acuta* and *P. antipodarum*), undefended (Oligochaeta and *C. zealandicus*) and predators (*P. grayi* and *X. zealandica*).

Statistical analysis was performed using the Jaccard index for presence/absence data was performed using the vegan package in R v.3.0.2 program (Oksanen, 2013). The data was first transformed into a binary matrix based on the presence or absence of a given virus in a FFP and sample type. The Jaccard's distance was calculated for the presence-absence of viruses within samples and a hierarchical cluster analysis was then performed, this was based on the dissimilarity index (1=samples share no viruses, 0=samples share all viruses). Dendrograms were then created for the analysis using unweighted pair-group method with arithmetic average linkage method to establish how similar the different samples were based on their virus assemblages.

7.4 Results and Discussion

7.4.1 Characterisation of viral genomes

One hundred and thirty nine circular viral sequences were recovered from seven different samples types (*M. novazelandiae* n=27, Oligochaeta and *C. zealandicus* n=16, *E. menziesi* n=26, water n=8, *P. antipodarum* and *P. acuta* n=29, *P. grayi* and *X. zealandica* n=11 and benthic sediment n=22) sampled in Lake Sarah, Cass Basin, New Zealand (43.0491 S, 171.7767 E). These represent 52 novel CRESS DNA viruses which we have named Lake Sarah-associated circular DNA virus 1 through to 52 (LSaCV-1 to 52) as well as six previously described CRESS DNA viruses including dragonfly larvae-associated circular DNA viruses -3, -5, -6, -8 and -10 (DflaCV-3,-5,-6,-8 and -10) and Avon-Heathcote estuary-associated circular DNA virus 8 (AHEaCV-8). The CRESS DNA viral genomes ranged in size from ~1550 to ~6400 nt (Figure 7.1 A and B). All the genomes had at least two major open reading frames (ORFs) that had varying organisations. Further BLASTp analysis of the major ORFs indicated that some encoded for a putative Rep or putative CP, with some genomes having additional major ORFs with no credible BLASTp hits. Both LSaCV-2 and LSaCV-30 had putative spliced Rep with a 155 nt and 35 nt intron respectively; introns in the Reps are common in members of the mastrevirus genus (Wright *et al.*, 1997) other spliced Reps having been recently identified in other CRESS DNA viruses (Dayaram *et al.*, 2014b; Dayaram *et al.*, 2015; Dayaram *et al.*, 2012; Kraberger *et al.*, 2014; Ng *et al.*, 2011b; Ng *et al.*, 2014; Rosario *et al.*, 2012b; Sikorski *et al.*, 2013b). All genomes have at least one intergenic regions (IR), with 48 genomes having both short intergenic regions (SIR) and long intergenic regions (LIR) and 10 having only LIR. The intergenic regions also contain the origin of replication (*ori*) where replication is initiated, this region usually contains a conserved nonanucleotide motif that is common across most CRESS DNA viruses (Table 7.2).

7.4.2 Circular DNA molecules

In addition to the 139 CRESS DNA genomes, 31 circular DNA molecules were recovered from the seven Lake Sarah samples, representing 11 unique molecules (Figure 7.1 C).

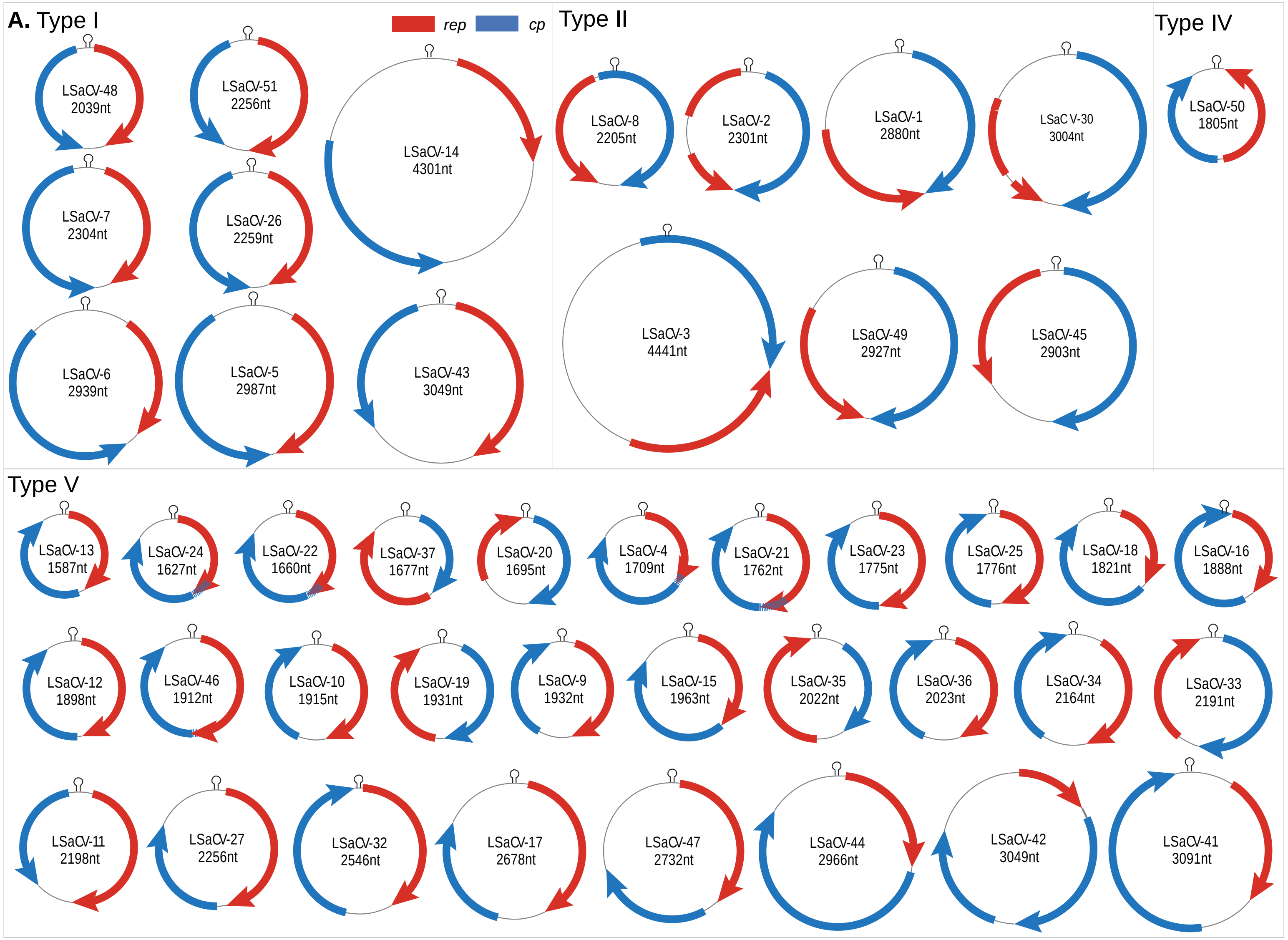


Figure 7.1: (A) Types of genome organisations of the novel CRESS DNA viruses recovered (figure continued)

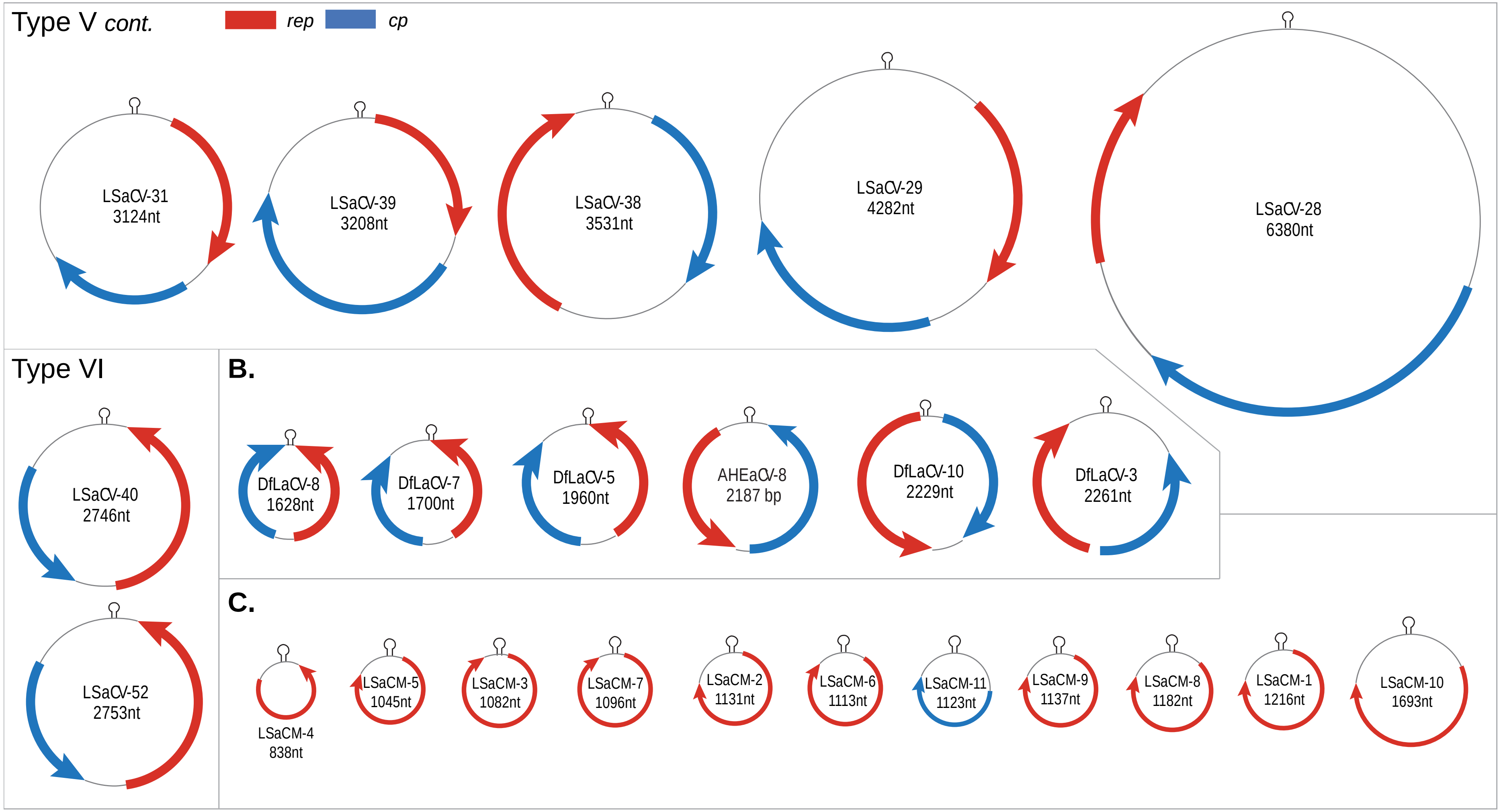


Figure 7.1: (B) CRESS DNA virus genome organisations similar to previously identified ones from Lake Sarah, Lake Hawdon, Lake Donne and Avon Heathcote estuary (Chapters 3 and 6) (C) Genome organisation of small circular molecules

Table 7.2: Viral isolates recovered and conserved motifs identified in the Rep and the putative nonanucleotide sequences.

Sequence	RCR			SF3 Helicase motifs			Nonanucleotide
	Motif I	Motif II	Motif III	Walker-A	Walker-B	Motif C	
LSaCV-1	LLTIN	HWQLLVVF	YVWKEDT	GRTGTGKS	VVLDEFR	ITSN	TAATTATTT
LSaCV-2	FFLTY	HVHVCFTL	YVKKDGN	GASRTGKT	AIFDDFE	ICSN	TAATATTAA
LSaCV-3	FHLTY	HTHALFKF	YHEKAPV	GSSGMGKT	IVFDDMS	ITCE	TAATATTCT
LSaCV-4	ITINN	HYQGMLTT	YVHKSDT	--------	VAVNPMW	ITCE	TAGTATTAC
LSaCV-5	FTLNN	HLQGYAQR	YCTKDAD	GPTGLGKS	IVIDDYR	ITRD	TAGTATTAC
LSaCV-6	LTFPQ	HLHALISF	YIKKDGK	SPPNAGKT	IIVDDST	ITRD	TAATATTAA
LSaCV-7	FTWNN	HLQGYIRF	YCHKEGD	GLSGVGKT	ILFDDMD	VTSQ	TATTATTAC
LSaCV-8	LTFPQ	HLHIYLTF	YITKTDK	SPPNCGKT	VVIDEFK	ILSN	TAATATTAA
LSaCV-9	FTLNN	HLQGYASL	YCSKGGN	GEPGVGKS	VIIDDFG	ITSN	TAGTATTAC
LSaCV-10	FTLNN	HLQGYIYF	YCKKDNN	GPTGTGKS	VLIDDMR	ITSC	TAGTATTAC
LSaCV-11	FTINN	HLQGFVHL	YCSKEDT	ATGGTGKT	VVIFDFT	CFSN	TAATACTGT
LSaCV-12	FTCNN	HLQGVVVL	YITPHAK	GPTGSGKS	VYIEDLG	VTSN	TATTATTAC
LSaCV-13	FTAFE	HYQGYIRT	YCKKSAT	--------	IYDIKDQ	LFTN	TAGTATTAC
LSaCV-14	FTLNN	HLQGFLHF	YCSKEDP	GPPGTGKT	VIIDDFY	ITSN	TATTATTAC
LSaCV-15	FTAYE	HYQGYLML	YCRKKDT	GQPDLPDR	VTKLVGE	IVQN	TAGTATTAC
LSaCV-16	FLNTR	HLQCYLYT	YCSKEKI	GATGTGKS	VVIDDMR	ITSC	CAGTATTAC
LSaCV-17	-----	HIQGFIQF	YCMKEDT	GPTNTGKT	VLFDEFD	ITAT	TAGTATTAC
LSaCV-18	FTAYE	HYQGWLRT	YCNKTET	AIEEYGKD	LSPEELT	ILSD	TAGTATTAC
LSaCV-19	FTLNN	HLQGYIEL	YCRKGGN	SIGGVGKS	IIVIDVP	IFAN	TTAATTATC
LSaCV-20	FTAYE	HRQGYVRT	YCQKSAT	LSHRPGTK	VCLDRQT	AWCN	TAGTATTAC
LSaCV-21	MVINN	HIQAWLKL	YAQKLDK	RYKLHGDT	VFVSDVF	PTTN	TAGTATTAA
LSaCV-22	VTINN	HYQLMVKT	YVHKDDT	GIETPEER	LVLLDQC	STSE	TAGTATTAC
LSaCV-23	FTLNN	HLQGYIEF	YCEFADY	GSPGSGKT	LLLDDYD	ITTN	TAGTATTAC
LSaCV-24	ITINN	HFQGMLQT	YVSKEET	VEEFGRRR	ATVDELV	ITRH	TAGTATTAC
LSaCV-25	LTNFN	TLSMFCAS	YCSKEGK	GPAGTGKT	ILFDDVE	FTSN	TAGTATTAC
LSaCV-26	FLTYP	HLHAVFAF	YCEKEDP	GATRLGKT	AVIDDLS	YLTN	TAATATTAA
LSaCV-27	FTLNN	HLQGYLYF	YCTKEEG	GPTGTGKT	VIIVDEF	FTSN	TAATACTTA
LSaCV-28	YTFNN	HLQGFIKF	YCKKDQD	ELGNSGKS	VVFMDCA	YLSN	TATATTAAC
LSaCV-29	VTKNR	HYQGYLEL	YATKEDT	GPSGTGKS	YIIDDFK	FTSD	TTTTAATAT
LSaCV-30	LTIPH	HWQLLVAF	YVWKEDT	GRSGSGKS	VVLDEFR	GTLI	TAGTATTAC
LSaCV-31	FTVNN	HLQGYIYF	YCSKEEG	GPSGTGKS	CIIDEVD	LCSN	CACTATATT
LSaCV-32	FTLNN	HLQGFVVF	YCQKEGF	GPAGVGKS	VLIEDID	ITSQ	TAGTATTAC
LSaCV-33	FTCYK	HIQGMAYN	YCTKAKS	GKSGKGKT	IIIDDLD	VTTQ	TTATATTAC
LSaCV-34	FTLNN	HLQGYASF	YCTKDGD	GPPGVGKS	CILDDFG	VTSN	TAGTATTAC
LSaCV-35	-----	-------	------	--------	VGANEDL	VTSV	AACCTTACG
LSaCV-36	FTLNN	HLQGYIRF	YCSKDGD	GDPGVGKS	VIIDDFG	VTSN	TAGTATTAC
LSaCV-37	FRQTK	HFQGALQC	YVMKEET	--------	VQTDIQT	LVSI	TAGTATTAC
LSaCV-38	MTLNN	HLQCFFSL	YCKKEEN	GPTGTGKS	VLIEDMD	VTSN	TAGTATTAC
LSaCV-39	FTWNN	HLQGFCSF	YCSKAGE	GPTGTGKS	VYLEDID	VTSN	TATTATTAC
LSaCV-40	GTLNN	HLQMACSF	YCMKEGD	GPSGTGKT	VLFDDFR	ITSC	CAGTATTAC
LSaCV-41	FWLCT	HYQLVAAF	YVGKEET	GSTGTGKS	VVVDEFR	ITSN	TAGTATTAC
LSaCV-42	FTIHS	HLQGALQL	YCSKSET	GPTGTGKT	LVIDDFD	ITSN	CAGTATTAC
LSaCV-43	FTVNF	HLQGYVNF	YCTKEEG	GPTGTGKS	LVFDEFR	FSTN	TGATAACCT
LSaCV-44	ILTIP	HWQMVVFL	YVWKQET	GPTGTGKS	VVIDEFR	ITSN	CAGTATTAC
LSaCV-45	FLTYP	HLHACIQY	YCRKDGN	GESGCGKT	IIFDDVD	FTCN	TAATATTAC
LSaCV-46	ITDYE	HWQLYCYT	YGQKEET	ECGGSGKS	YVIFDLA	IMSN	TAGTATTAC
LSaCV-47	FTVHR	HVQAYVQF	YCTKEAS	GPTNTQKT	VLWQDFD	LTSN	TAGTATTAC
LSaCV-48	FTDNN	HLQGYVEF	YASKEAT	GPSGCGKT	VVLDEFY	ITSN	GCCTAATAT
LSaCV-49	FIGTC	HWQFVVQC	YVWKEET	GPTGTGKS	VIIDEFR	ICSN	GTATGTCGC
LSaCV-50	YTANN	HIQGYIEF	YCMKGGD	GLTGSGKT	VIIDDFD	ITCE	GCACGATTA
LSaCV-51	LTYSR	HYHILVKF	YISDPSK	GPSRIGKT	LIMDDFS	WLCN	AACAAGCCA
LSaCV-52	FTINN	HLQGACIL	YCTKEDT	GPTGVHKT	AIFDDFR	ITTP	TAGTATTAC
DflaCV-3	FTLNN	HLQGFVYF	YCSKDGD	GAAGTGKS	VVMDDMD	VTSQ	AACTATTAC
DflaCV-5	FTAFV	HWQGYVEF	YCKKDMK	--------	IIVRHAK	ITPL	TAGTATTAC
DflaCV-6	FTLFH	HLQGYLYY	YCKKEGD	GSTGCGKT	IILDDFH	ITCE	TAGTATTAC
DflaCV-8	ITLFG	HSHMYLRF	YCWSEGD	GPTGTGKT	LANAPSP	YCSM	TAGTATTAC
AHEaCV-8	FRYNA	HYQGRMSL	YVIKEDT	PIGDLGKS	YIVDMPR	VFTN	CCTACTTAC
LSaCM-1	-----	HLQGYIEM	YAKKDNT	GKTGSGKT	ILIDEFN	ICSN	TCACCTGCA
LSaCM-2	FTKNN	HLQGYAEF	YCMKDGD	GPAGSGKT	VIINDVG	ITSQ	CATCTTTAT
LSaCM-3	FTSYT	HLQGYIEF	YCMKLDT	GRPGTGKT	VIIDDFS	ITSN	TAGTATTAC
LSaCM-4	-----	--------	-------	--------	-------	----	CTCGTTTAC
LSaCM-5	FTQNN	QLQGVICF	YCKKDGD	GEAGTGKS	VLIEDFD	VTSN	CATTATTAC
LSaCM-6	FTDNE	HLQGVIVF	YCKKEGD	GATGAGKS	VIFDDFR	ITAP	CAGTATTAC
LSaCM-7	FTDNE	HLQGVIVF	YCKKEGD	GATGAGKS	VIFDDFR	ITAP	TAGTATTAC
LSaCM-8	FWLLT	HWQVIVAF	YVWKDDT	GSTGTGKS	VVMDEFR	ITSN	TAGTATTAC
LSaCM-9	YTLNN	HLQGYVQF	YCKKDGD	GPTGTGKS	VIIEEAD	VISN	TAGTATTAC
LSaCM-10	FTWDT	HYQGTLTL	YVTKDEG	EKGNSGKS	VVIFTNL	IFTN	TAATATTAC

These ranged in size from 838 nt to 1693 nt and we have tentatively named these Lake Sarah-associated circular DNA molecules (LSaCM) -1 through to -11. The sequences of LSaCM have at least one major ORF. BLASTp analysis of the ORFs show that LSaCM-1 through to LSaCM-10 encode for putative Reps (Table 7.3) whereas LSaCM-11, sampled in benthic sediment, Oligochaeta, *C. zealandicus, E. menziesi, M. novazelandiae, P. antipodarum* and *P. acuta,* encodes for a putative CP sharing ~ 29% pairwise identity to that of SI00529 (JX904443) (Table 7.4). A putative *ori* was identified for each LSaCM, with all containing a replication stem-loop structure A putative nonanucleotide motif was identified in each LSaCM which varied among molecules (Table 7.2)..Given the size of these molecules and the fact they only contain one major ORF, it is possible they represent defective genomes (Casado *et al.*, 2004; Hadfield *et al.*, 2012; Jeske *et al.*, 2001; van der Walt *et al.*, 2009) that have large deletions. Alternatively they could represent small genome components of multipartite CRESS DNA viral genomes, such as those found in the family *Nanoviridae* or satellite molecules associates with CRESS DNA viruses similar to the association observed with alpha satellites and begomoviruses (Mubin *et al.*, 2010; Romay *et al.*, 2010).

7.4.3 Analysis of the viral Reps and CPs of LSaCVs and LSaCMs

Phylogenetic analysis of the Reps of both LSaCV and LSaCM illustrate the significant diversity of these sequences (Figure 7.2) indicating that they do not group with the current genera of *Geminiviridae*, *Nanoviridae* and *Circoviridae*. The Reps of LSaCV-1 and LSaCV-30 share amino acid identity 73.6% (recovered from gastropods, benthic sediment, cockles and worms). The Rep of LSaCV-25 shares the highest amino acid identity, 89.1% with that of AHEaCV -17 (KM874343). Reps of LSaCV-9, -34 and -36 share between 56 and 65% pairwise identity with that of sewage-associated circular DNA virus (SaCV) -17 (KM821752), whereas the Rep of LSaCV-39 shares 67.6% identity with that of McMurdo Ice Shelf pond-associated circular DNA virus (MpaCDV) -3 (KJ547648).

From the initial phylogenetic analysis it was evident that there were two well supported clades of CRESS DNA viruses in the maximum likelihood phylogenetic tree that fall outside the current ssDNA viral families. It is apparent that LSaCV-37, -13, 18, -15, -20, -21, -22, -4, and -24 form a clade with Odonata-associated circular DNA virus (OdasCV) -1 (KM598393) and Sewage-associated circular DNA virus (SaCV) -15 (Figure 7.2 and 7.3).

Table 7.3: Top BLASTp hits to Lake Sarah viral Reps recovered

Viral sequence	Accession #	BLASTp Hit	Identity	E-value	Accession #
LSaCV-1	KP153390	DflaCV-2	60%	6x10^{-114}	YP_009001739
LSaCV-2	KP153394	SaCV-4	42%	8x10^{-41}	AIF34812
LSaCV-3	KP153395	MpaCDV-2	49%	3x10^{-23}	YP_009047130
LSaCV-4	KP153397	YN-BtCV China	36%	9x10^{-11}	AEL28804
LSaCV-5	KP153402	GOM02962	39%	1x10^{-57}	AGA18347
LSaCV-6	KP153403	BCTV	41%	3x10^{-16}	ACM44491
LSaCV-7	KP153404	*Diporeia sp.* CV	54%	3x10^{-95}	AGG39811
LSaCV-8	KP153405	SaCV-14	35%	5x10^{-48}	AIF34806
LSaCV-9	KP153407	SI03705	39%	1X10^{-48}	AGA18440
LSaCV-10	KP153408	SOG03994	49%	5X10^{-75}	AGA18263
LSaCV-11	KP153409	CynNCXV	43%	6X10^{-82}	YP_009021888
LSaCV-12	KP153410	SI00197	36%	6X10^{-43}	AGA18388
LSaCV-13	KP153412	GOM00012	32%	7X10^{-6}	AGA18289
LSaCV-14	KP153414	EeCV	52%	3X10^{-84}	YP_009000900
LSaCV-15	KP153417	ToLCNDV alphasatellite	42%	1X10^{-9}	AFD61610
LSaCV-16	KP153420	SOG00164	52%	1X10^{-80}	AGA18246
LSaCV-17	KP153424	SI00850	32%	2X10^{-26}	AGA18409
LSaCV-18	KP153427	Cuban alphasatellite 1	36%	6X10^{-9}	YP_008169853
LSaCV-19	KP153429	SI00793	47%	2X10^{-62}	AGA18407
LSaCV-20	KP153434	*F.duorarum* circovirus	36%	3X10^{-7}	AGS47835
LSaCV-21	KP153437	BFDV	40%	7x10^{-7}	AEL30250
LSaCV-22	KP153441	Mosquito VEM SDRBAJ	39%	3x10^{-8}	AEF58777
LSaCV-23	KP153442	*Diporeia sp.* CV	84%	0	AGG39815
LSaCV-24	KP153443	YN-BtCV-1	32%	3x10^{-7}	AEL87784
LSaCV-25	KP153445	MpaCDV-7	55%	9x10^{-76}	YP_009047142
LSaCV-26	KP153446	TYDV-A	34%	2x10^{-46}	AFD63074
LSaCV-27	KP153523	BatCV	44%	7x10^{-58}	YP_007974237
LSaCV-28	KP153451	MpaCDV-5	41%	5x10^{-65}	YP_009047137
LSaCV-29	KP153454	PKbeef23	32%	8x10^{-19}	ADU76993
LSaCV-30	KP153456	18-LDMD	28%	5x10^{-19}	AGS36233
LSaCV-31	KP153459	BMLRV alphasatellite 1	40%	0.13	YP_009021872
LSaCV-32	KP153464	GOM03161	41%	9x10^{-58}	AGA18366
LSaCV-33	KP153469	SI03701	33%	1x10^{-30}	AGA18438
LSaCV-34	KP153471	SI03705	41%	1x10^{-49}	AGA18440
LSaCV-35	KP153472	SAR-B	26%	0.14	YP_003084139
LSaCV-36	KP153475	SI03705	39%	1x10^{-49}	AGA18440
LSaCV-37	KP153476	hs1	32%	2x10^{-4}	YP_009022029
LSaCV-38	KP153484	19-LDMD	41%	5X10^{-50}	AGS36235
LSaCV-39	KP153485	MpaCDV-3	67%	7X10^{-140}	YP_009047132
LSaCV-40	KP153488	CyCV-NG13	36%	1X10^{-39}	ADD62475
LSaCV-41	KP153494	DflaCV-2	64%	1X10^{-116}	YP_009001739
LSaCV-42	KP153496	RW-B	33%	5x10^{-40}	YP_003084285
LSaCV-43	KP153499	BatCV	34%	5x10^{-30}	YP_007974237
LSaCV-44	KP153500	YN-BtCV-1	60%	2x10^{-109}	AEL87784
LSaCV-45	KP153501	Nepavirus	33%	3x10^{-39}	YP_009021041
LSaCV-46	KP153502	SaCV-10	32%	3x10^{-25}	AIF34798
LSaCV-47	KP153504	MmCV	32%	5x10^{-26}	AEW49399
LSaCV-48	KP153505	CanCV	41%	8x10^{-54}	AGI42840
LSaCV-49	KP153507	SaCV-6	49%	1x10^{-77}	AIF34816
LSaCV-50	KP153511	DflaCV-6	38%	2x10^{-52}	YP_009001747
LSaCV-51	KP153522	*S. lacrymans* S7.9	33%	2x10^{-42}	XP_007321836
LSaCV-52	KP153490	SaCV-11	44%	5X10^{-75}	AIF34799
LSaCM-1	KP153359	11-LDMD	34%	4x10^{-24}	AGS36210
LSaCM-2	KP153360	DflaCV-3	44%	5x10^{-69}	AHH31467
LSaCM-3	KP153361	21-LDMD	39%	4x10^{-45}	AGS36240
LSaCM-4	KP153363	GOM03228	29%	4x10^{-6}	AGA18372
LSaCM-5	KP153364	SOG00182	60%	3x10^{-109}	AGA18247
LSaCM-6	KP153368	DflaCV-3	40%	7x10^{-67}	AHH31467
LSaCM-7	KP153370	SI03931	38%	3x10^{-48}	AGA18448
LSaCM-8	KP153371	DflaCV-2	63%	7x10^{-110}	YP_009001739
LSaCM-9	KP153377	19-LDMD	46%	3x10^{-68}	AGS36235
LSaCM-10	KP153379	*C. closterium*	34%	3x10^{-27}	YP_009029088
DflaCV-3	KP153447	DflaCV-3	99%	0	AHH31467
DflaCV-5	KP153524	DflaCV-5	100%	0	YP_009001745
DflaCV-6	KP153527	DflaCV-6	99%	0	YP_009001747
DflaCV-8	KP153526	DflaCV-8	97%	0	YP_009001751
DflaCV-10	KP153516	DflaCV-10	98%	0	AHH31482
AHEaCV-8	KP153528	Marine virus	37%	3x10^{-48}	GAC77869

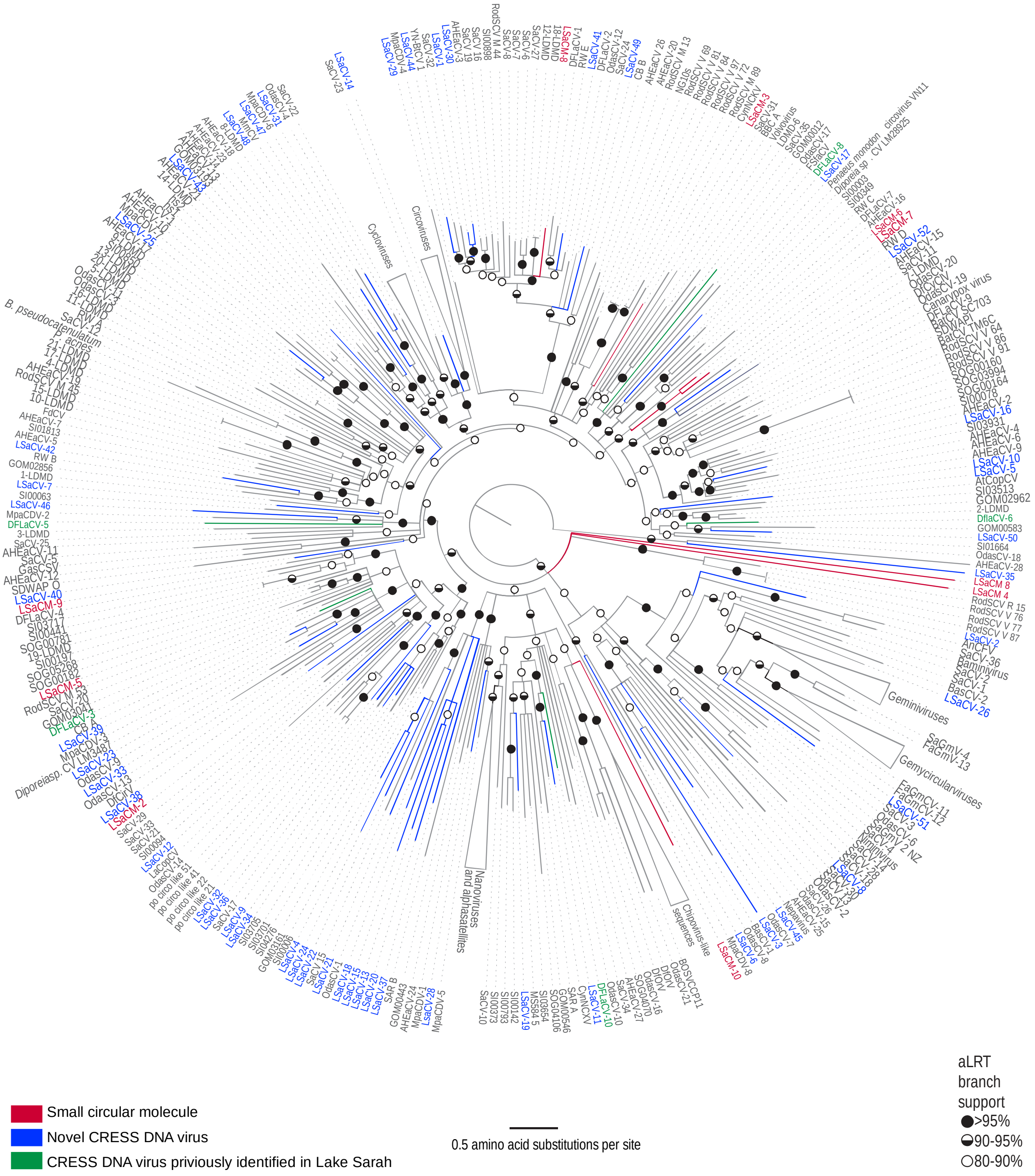


Figure 7.2: Maximum likelihood phylogenetic tree of all CRESS DNA viral Rep sequences with aLRT branch support and mid-point rooted. See Table 7.5 for GenBank accession numbers associated with the acronyms used in the figure. Sections of the tree are enlarged in Figure 7.3 and 7.4.

Table 7.4: Summary top five amino acid pairwise identities of the Rep proteins and the top five amino acid pairwise identities of the CP proteins computed in SDT. Blank space indicates no credible result.

Query	Accession	Rep pairwise identity	% identity	CP pairwise identity	% identity
LSaCV-1	KP153390	LSaCV-30 [KP153456]	73.55	SaCV-19 [KM821754]	27.8
LSaCV-1	KP153390	SaCV-32 [KM821767]	64.02	AHEaCV-3 [KM874295]	27.6
LSaCV-1	KP153390	AHEaCV-3 [KM874295]	62.74	SaCV-33 [KM821768]	24.4
LSaCV-1	KP153390	SaCV-8 [KJ547632]	62.21		
LSaCV-1	KP153390	DflaCV-2 [KF738874]	61.74		
LSaCV-2	KP153394	SaGmV-2 [KJ547642]	43.05		
LSaCV-2	KP153394	SaCV-4 [KJ547628]	42.18		
LSaCV-2	KP153394	OdasCV-6 [KM598389]	41.11		
LSaCV-2	KP153394	DfaCV-1 [JX185430]	40.89		
LSaCV-2	KP153394	SaCV-2 [KJ547626]	40.86		
LSaCV-3	KP153395	AHEaCV-25 [KM874355]	33.86		
LSaCV-3	KP153395	OdasCV-7 [KM598390]	32.85		
LSaCV-3	KP153395	MpaCDV-2 [KJ547647]	31.45		
LSaCV-3	KP153395	PanSV-A [L39638]	29.83		
LSaCV-3	KP153395	DflaCV-7 [KF738881]	29.76		
LSaCV-4	KP153397	SaCV-15 [KM821750]	54.45	CfCV [JQ011377]	29.9
LSaCV-4	KP153397	AHEaCV-18 [KM874346]	34.78	SaCM-6 [KM877828]	26.2
LSaCV-4	KP153397	GuCV [JQ685854]	34.39		
LSaCV-4	KP153397	21-LDMD [KF133828]	33.51		
LSaCV-4	KP153397	1-LDMD [KF133807]	33.51		
LSaCV-5	KP153402	AtCopCV [JQ837277]	41.5		
LSaCV-5	KP153402	SI03931 [JX904581]	40.61		
LSaCV-5	KP153402	GOM02962 [JX904333]	39.53		
LSaCV-5	KP153402	OdasCV-20 [KM598406]	38.35		
LSaCV-5	KP153402	CanCV [AFK82575]	38.19		
LSaCV-6	KP153403	SaCV-18 [KM874346]	33.55		
LSaCV-6	KP153403	LSaCV-8 [KP153405]	31.23		
LSaCV-6	KP153403	OdasCV-8 [KM598391]	30.58		
LSaCV-6	KP153403	Niminivirus [JQ898332]	30.26		
LSaCV-6	KP153403	BeYDV [AM849096]	30.25		
LSaCV-7	KP153404	Diporeia sp. CV LM3487 [KC248416]	54.55	AHEaCV-25 [KM874355]	25
LSaCV-7	KP153404	MpaCDV-3 [KJ547648]	45.62	SaCV-24 [KM821759]	24.7
LSaCV-7	KP153404	SaCV-21 [KM821756]	44.36		
LSaCV-7	KP153404	19-LDMD [KF133826]	43.96		
LSaCV-7	KP153404	DflaCV-3 [KF738875]	42.55		
LSaCV-8	KP153405	SaCV-18 [KM874346]	40.61	DflaCV-3 [KF738875]	26.9
LSaCV-8	KP153405	SaCV-14 [KJ547625]	36.42	SaCV-34 [KM821769]	25
LSaCV-8	KP153405	SaCV-28 [KM874362]	34.88		
LSaCV-8	KP153405	SaCV-26 [KM874359]	33.94		
LSaCV-8	KP153405	CpRV [GU256532]	31.35		
LSaCV-9	KP153407	SaCV-17 [KM821752]	56.47		
LSaCV-9	KP153407	CanineCV [KC241982]	43.4		
LSaCV-9	KP153407	CanCV [AFK82575]	42.91		
LSaCV-9	KP153407	CanineCV [JQ821392]	42.91		
LSaCV-9	KP153407	SI03705 [JX904561]	42.59		
LSaCV-10	KP153408	AHEaCV-9 [KM874315]	56.18	OdasCV-16 [KM598411]	2.09E-02
LSaCV-10	KP153408	AHEaCV-6 [KM874304]	55.13		
LSaCV-10	KP153408	LSaCV-16 [KP153420]	55.06		
LSaCV-10	KP153408	OdasCV-20 [KM598406]	51.15		
LSaCV-10	KP153408	SI03931 [JX904581]	49.81		
LSaCV-11	KP153409	SOG04070 [JX904144]	41.64	CynNCXV [JX908739]	31.5
LSaCV-11	KP153409	MS584-5 [HQ322117]	41.63	GKaV [AB698917]	27
LSaCV-11	KP153409	LSaCV-19 [KP153429]	41.32		
LSaCV-11	KP153409	SI03654 [JX904548]	40.74		
LSaCV-11	KP153409	CynNCXV [JX908739]	40.12		
LSaCV-12	KP153410	SI00197 [JX904420]	37.17		
LSaCV-12	KP153410	RodSCV-M-53 [JF755415]	36.5		
LSaCV-12	KP153410	OdasCV-14 [KM598397]	36.29		
LSaCV-12	KP153410	LaCopCV [JF912805]	35.89		
LSaCV-12	KP153410	SaCV-20 [KM821755]	35.77		
LSaCV-13	KP153412	LSaCV-20 [KP153434]	35.63		
LSaCV-13	KP153412	LSaCV-18 [KP153427]	35.45		
LSaCV-13	KP153412	LSaCV-15 [KP153417]	34.78		
LSaCV-13	KP153412	LSaCV-24 [KP153443]	30.11		
LSaCV-13	KP153412	OdasCV-1 [KM598393]	30.05		
LSaCV-14	KP153414	NGchicken8 [HQ738643]	49.63		
LSaCV-14	KP153414	NG13 [GQ404856]	49.31		
LSaCV-14	KP153414	NGchicken15 [HQ738644]	49.26		
LSaCV-14	KP153414	PK5510 [GQ404847]	49.08		
LSaCV-14	KP153414	DfCyV-2 [JX185423]	48.9		
LSaCV-15	KP153417	LSaCV-18 [KP153427]	42.08		
LSaCV-15	KP153417	SaCV-15 [KM821750]	32.34		
LSaCV-15	KP153417	LSaCV-20 [KP153434]	31.75		
LSaCV-15	KP153417	PaLCVA-India [AFA26472]	30.63		
LSaCV-15	KP153417	AYVSGA [NP_579867]	30.56		
LSaCV-16	KP153420	AHEaCV-9 [KM874315]	52.99		
LSaCV-16	KP153420	AHEaCV-6 [KM874304]	51.14		
LSaCV-16	KP153420	SI03931 [JX904581]	50.38		
LSaCV-16	KP153420	OdasCV-20 [KM598406]	49.06		
LSaCV-16	KP153420	SOG00164 [JX904076]	48.13		
LSaCV-17	KP153424	GOM02856 [JX904312]	34.42	SaCV-16 [KM821751]	34
LSaCV-17	KP153424	PKbeef23 [HQ738634]	33.58	SaCV-18 [KM821753]	32.6
LSaCV-17	KP153424	PKgoat21 [HQ738635]	33.58	SaCV-14 [KJ547625]	32.3
LSaCV-17	KP153424	SI00850 [JX904473]	33.58	SaCV-11 [KJ547622]	31.7
LSaCV-17	KP153424	12-LDMD [KF133819]	33.47	SaCV-29 [KM821764]	30.5

Query	Accession	Rep pairwise identity	% identity	CP pairwise identity	% identity
LSaCV-18	KP153427	LSaCV-20 [KP153434]	35.39	SaCV-20 [KM821755]	23.5
LSaCV-18	KP153427	SI03701 [JX904559]	32.8	OdasCV-18 [KM598402]	26
LSaCV-18	KP153427	LSaCV-37 [KP153476]	32.07		
LSaCV-18	KP153427	BBTV-SAT [ACJ36782]	31.07		
LSaCV-18	KP153427	ABTV [EF546807]	30.94		
LSaCV-19	KP153429	SI00793 [JX904469]	47.35		
LSaCV-19	KP153429	SI00142 [JX904416]	44.06		
LSaCV-19	KP153429	SI03654 [JX904548]	42.42		
LSaCV-19	KP153429	MS584-5 [HQ322117]	40.23		
LSaCV-19	KP153429	DflaCV-10 [KF738884]	39.34		
LSaCV-20	KP153434	BBTV-SAT [AAA61875]	32.43	OdasCV-1 [KM598393]	25.8
LSaCV-20	KP153434	BBTV-SAT [AAG44003]	31.89	SaCM-6 [KM877828]	21
LSaCV-20	KP153434	SI01813 [JX904523]	31.72		
LSaCV-20	KP153434	FdCV [KC441518]	31.69		
LSaCV-20	KP153434	SaCV-29 [KM821764]	31.22		
LSaCV-21	KP153437	LSaCV-4 [KP153397]	29.57		
LSaCV-21	KP153437	LSaCV-24 [KP153443]	27.19		
LSaCV-21	KP153437	AHEaCV-18 [KM874346]	26.56		
LSaCV-21	KP153437	AHEaCV-21 [KM874350]	26.38		
LSaCV-21	KP153437	SaCV-15 [KM821750]	26.34		
LSaCV-22	KP153441	LSaCV-4 [KP153397]	43.78	SaCM-10	28.6
LSaCV-22	KP153441	SaCV-15 [KM821750]	38.5	Penaeus monodon VN11	27.7
LSaCV-22	KP153441	LSaCV-24 [KP153443]	32.55	AHEaCV-21 [KM874350]	27.5
LSaCV-22	KP153441	SaCV-22 [KM821757]	30.69	BaCV [GU799606]	26.7
LSaCV-22	KP153441	OdasCV-4 [KM598408]	30.24	SaCM-7 [KM877829]	24.7
LSaCV-23	KP153442	SaCV-12 [KJ547623]	39.92		
LSaCV-23	KP153442	RhFeCV [JQ814849]	39.1		
LSaCV-23	KP153442	SI00850 [JX904473]	38.91		
LSaCV-23	KP153442	1-LDMD [KF133807]	38.85		
LSaCV-23	KP153442	11-LDMD [KF133818]	38.2		
LSaCV-24	KP153443	SaCV-15 [KM821750]	51.22	OdasCV-1 [KM598393]	30.2
LSaCV-24	KP153443	LSaCV-4 [KP153397]	50		
LSaCV-24	KP153443	*P.monodon* VN11 [KF481961]	34.48		
LSaCV-24	KP153443	SaCV-25 [KM821760]	34.11		
LSaCV-24	KP153443	SaCV-17 [KM821752]	32.23		
LSaCV-25	KP153445	AHEaCV-17 [KM874343]	89.09	AHEaCV-17 [KM874343]	93.4
LSaCV-25	KP153445	AHEaCV-10 [KM874320]	47.64	AHEaCV-18 [KM874346]	93.4
LSaCV-25	KP153445	MpaCDV-7 [KJ547652]	45.82	SaCV-16 [KM821751]	29.3
LSaCV-25	KP153445	CanCV [AFK82575]	37.11	MpaCDV-7 [KJ547652]	29.1
LSaCV-25	KP153445	13-LDMD [KF133820]	35.94	OdasCV-10 [KM598412]	27.8
LSaCV-26	KP153446	CpCDV [AM933135]	37.29	Niminivirus	30.1
LSaCV-26	KP153446	DoYMV [AM157413]	36.18	OdasCV-15 [KM598398]	29.8
LSaCV-26	KP153446	ChCDV [AM850136]	36.12		
LSaCV-26	KP153446	DfaCV-1 [JX185430]	35.84		
LSaCV-26	KP153446	BeYDV [AM849096]	35.45		
LSaCV-27	KP153523	AHEaCV-14 [KM874332]	44.37	SaCV-18 [KM821753]	37.1
LSaCV-27	KP153523	BatCV XOR7 [KC339249]	42.28	SaCV-14 [KJ547625]	34.9
LSaCV-27	KP153523	AHEaCV-13 [KM874329]	42.21	SaCV-29 [KM821764]	33.9
LSaCV-27	KP153523	AHEaCV-18 [KM874346]	41.61	SaCV-11 [KJ547622]	32.7
LSaCV-27	KP153523	RodSCV-M-45 [JF755409]	40.7	AHEaCV-5 [KM874301]	32.6
LSaCV-28	KP153452	MpaCDV-5 [KJ547650]	44.95	ClorDNAV [AB844272]	28.9
LSaCV-28	KP153452	MpaCDV-1 [KJ547646]	40.07	ClorDNAV01 [AB553581]	27.5
LSaCV-28	KP153452	AHEaCV-24 [KM874354]	39.18	CtenDNAV [AB597949]	24.2
LSaCV-28	KP153452	DflaCV-10 [KF738884]	38.64		
LSaCV-28	KP153452	DflaCV-10a [KF738885]	38.31		
LSaCV-29	KP153454	AHEaCV-5 [KM874301]	33.68		
LSaCV-29	KP153454	GOM02856 [JX904312]	33.45		
LSaCV-29	KP153454	NG10 [ADF80742]	33.33		
LSaCV-29	KP153454	CyCV-TB [HQ738637]	30.92		
LSaCV-29	KP153454	YN-BtCV5 [JF938082]	30.92		
LSaCV-30	KP153456	AHEaCV-3 [KM874295]	61.18	SaCV-32 [KM821767]	31.5
LSaCV-30	KP153456	SaCV-32 [KM821767]	60.92	AHEaCV-3 [KM874295]	23.2
LSaCV-30	KP153456	SaCV-19 [KM821754]	59	Nepavirus [JQ898333]	22.1
LSaCV-30	KP153456	12-LDMD [KF133819]	58.08		
LSaCV-30	KP153456	SI00898 [JX904478]	57.81		
LSaCV-31	KP153459	MpaCDV-6 [KJ547651]	37.17	SaCV-11 [KJ547622]	28.7
LSaCV-31	KP153459	OdasCV-18 [KM598401]	37.13	SaCV-28 [KM821763]	28.6
LSaCV-31	KP153459	OdasCV-18 [KM598402]	37.13	SaCV-18 [KM821753]	26.8
LSaCV-31	KP153459	SI03931 [JX904581]	36.78	FaCM-3 [KJ547619]	26.4
LSaCV-31	KP153459	OdasCV-18 [KM598403]	36.03	AllMDV [KC202818]	23.7
LSaCV-32	KP153464	SaCV-21 [KM821756]	43.7		
LSaCV-32	KP153464	DflaCV-3 [KF738875]	41.73		
LSaCV-32	KP153464	DlaCV-3a [KF738876]	41.73		
LSaCV-32	KP153464	RodSCV-M-53 [JF755415]	41.7		
LSaCV-32	KP153464	GOM03041 [JX904344]	41.11		
LSaCV-33	KP153469	SI03701 [JX904559]	35.46		
LSaCV-33	KP153469	SaCV-29 [KM821764]	34.03		
LSaCV-33	KP153469	GOM03161 [JX904368]	33.97		
LSaCV-33	KP153469	OdasCV-9 [KM598392]	33.46		
LSaCV-33	KP153469	MpaCDV-3 [KJ547648]	33.33		
LSaCV-34	KP153470	SaCV-17 [KM821752]	59.78	SaCV-17 [KM821752]	26.7
LSaCV-34	KP153470	LSaCV-36 [KP153474]	56.16		
LSaCV-34	KP153470	LSaCV-9 [KP153407]	56.16		
LSaCV-34	KP153470	RaCV [DQ146997]	43.35		
LSaCV-34	KP153470	SaCV-21 [KM821756]	41.92		
LSaCV-35	KP153472	NG10 [ADF80742]	26.56		
LSaCV-35	KP153472	RodSCV-M-45 [JF755409]	25.87		
LSaCV-35	KP153472	PisaCV FUJ1 [JX305998]	25.23		
LSaCV-35	KP153472	CaCV [AJ301633]	24.6		
LSaCV-35	KP153472	DflaCV-1 [KF738873]	24.32		

Query	Accession	Rep pairwise identity	% identity	CP pairwise identity	% identity
LSaCV-36	KP153474	SaCV-17 [KM821752]	65.83	SaCV-17 [KM821752]	24.5
LSaCV-36	KP153474	LSaCV-9 [KP153407]	59.42		
LSaCV-36	KP153474	RaCV [DQ146997]	41.89		
LSaCV-36	KP153474	GOM03161 [JX904368]	41.22		
LSaCV-36	KP153474	SgCV [JQ011377]	41.2		
LSaCV-37	KP153476	DfCyV-4 [KC512917]	33.7		
LSaCV-37	KP153476	hs1 [JX559621]	32.24		
LSaCV-37	KP153476	hs2 [JX559622]	31.69		
LSaCV-37	KP153476	DfCyV-8 [KC512920]	30.21		
LSaCV-37	KP153476	RodSCV-V-69 [JF755403]	29.53		
LSaCV-38	KP153483	RodSCV-M-53 [JF755415]	43.87	SaCV-13 [KJ547624]	38.8
LSaCV-38	KP153483	GOM03041 [JX904344]	42.44		
LSaCV-38	KP153483	SaCV-20 [KM821755]	41.79		
LSaCV-38	KP153483	SaCV-21 [KM821756]	40.52		
LSaCV-38	KP153483	19-LDMD [KF133826]	40.49		
LSaCV-39	KP153486	MpaCDV-3 [KJ547648]	67.63	MpaCDV-3 [KJ547648]	32.1
LSaCV-39	KP153486	Diporeia sp. CV-LM3487 [KC248416]	46.72	SaCV-1 [KJ547620]	26.6
LSaCV-39	KP153486	GOM03041 [JX904344]	44.81	Nepavirus [JQ898333]	24.7
LSaCV-39	KP153486	SI00197 [JX904420]	43.53	SaCV-32 [KM821767]	22.5
LSaCV-39	KP153486	CB-A [FJ959082]	43.28	SaCV-33 [KM821768]	21.8
LSaCV-40	KP153488	OdasCV-20 [KM598406]	44.03	SaCV-29 [KM821764]	32.3
LSaCV-40	KP153488	AHEaCV-2 [KM874290]	40.71	SaCV-11 [KJ547622]	32
LSaCV-40	KP153488	LSaCV-50 [KP153511]	40.22	SaCV-18 [KM821753]	30
LSaCV-40	KP153488	AHEaCV-6 [KM874304]	39.92	SaCV-14 [KJ547625]	29.4
LSaCV-40	KP153488	SOG03994 [JX904139]	38.72	AHEaCV-5 [KM874301]	28.7
LSaCV-41	KP153494	DflaCV-2 [KF738874]	64.53	SaCV-27 [KM821762]	25.1
LSaCV-41	KP153494	18-LDMD [KF133825]	61.45	SaCV-6 [KJ547630]	22.9
LSaCV-41	KP153494	12-LDMD [KF133819]	60.63		
LSaCV-41	KP153494	SaCV-7 [KJ547631]	59.09		
LSaCV-41	KP153494	LSaCM-8 [KP153371]	59		
LSaCV-42	KP153496	SaCV-12 [KJ547623]	36.8	SaCV-25 [KM821760]	29.5
LSaCV-42	KP153496	RW-B [FJ959078]	36.04	SaCV-26 [KM821761]	29.3
LSaCV-42	KP153496	OdasCV-20 [KM598406]	35.85	OdasCV-11 [KM598394]	29
LSaCV-42	KP153496	15-LDMD [KF133822]	35.59	SaCV-29 [KM821764]	28.5
LSaCV-42	KP153496	SI00850 [JX904473]	35.5	FaCM-3 [KJ547619]	28.5
LSaCV-43	KP153499	SaCV-12 [KJ547623]	38.76	SaCV-28 [KM821763]	31.2
LSaCV-43	KP153499	SaCV-22 [KM821757]	36.56	FaCM-3 [KJ547619]	29.7
LSaCV-43	KP153499	GOM03193 [JX904377]	34.98	SaCV-29 [KM821764]	26.9
LSaCV-43	KP153499	SI04276 [JX904605]	34.66	SaCV-18 [KM821753]	24.4
LSaCV-43	KP153499	AHEaCV-14 [KM874332]	34.48	AHEaCV-5 [KM874301]	24.4
LSaCV-44	KP153500	SaCV-16 [KM821751]	59.92	PoBoV1 [HM053693]	35.2
LSaCV-44	KP153500	YN-BtCV-1 [JF938078]	59.32	PoBoV3 JF429834]	31.4
LSaCV-44	KP153500	AHEaCV-3 [KM874295]	58.94	PoBoV4 [JF429835]	31.4
LSaCV-44	KP153500	SaCV-19 [KM821754]	57.79	MPV-4 [FJ440683]	31
LSaCV-44	KP153500	SaCV-8 [KJ547632]	56.49	GBoV1 [HM145750]	27.8
LSaCV-45	KP153501	OdasCV-7 [KM598390]	41.73	SaCV-1 [KJ547620]	27.8
LSaCV-45	KP153501	OdasCV-15 [KM598398]	39.16	SaCV-33 [KM821768]	26.5
LSaCV-45	KP153501	AHEaCV-25 [KM874355]	38.37	MpaCDV-3 [KJ547648]	26.1
LSaCV-45	KP153501	Nepavirus [JQ898333]	34.44	SaCV-19 [KM821754]	26
LSaCV-45	KP153501	MpaCDV-2 [KJ547647]	33.71	OdasCV-7 [KM598390]	24.5
LSaCV-46	KP153502	MpaCDV-5 [KJ547650]	35.27	SaCV-4 [KJ547628]	28.4
LSaCV-46	KP153502	AtCopCV [JQ837277]	34.33		
LSaCV-46	KP153502	SaCV-10 [KJ547621]	32.86		
LSaCV-46	KP153502	SOG04070 [JX904144]	32.23		
LSaCV-46	KP153502	AHEaCV-20 [KM874348]	31.93		
LSaCV-47	KP153504	SI00850 [JX904473]	34.73	SaCV-11 [KJ547622]	34.8
LSaCV-47	KP153505	AtCopCV [JQ837277]	33.79	SaCV-18 [KM821753]	29.6
LSaCV-47	KP153506	MmCV [JQ085285]	33.33	SaCV-28 [KM821763]	28
LSaCV-47	KP153507	5-LDMD [KF133812]	33.21	SaCV-29 [KM821764]	27.4
LSaCV-47	KP153504	20-LDMD [KF133827]	33.21	FaCM-3 [KJ547619]	25.8
LSaCV-48	KP153505	CanCV [AFK82575]	42.07	Niminivirus [JQ898332]	28.3
LSaCV-48	KP153505	CanineCV [JQ821392]	42.07	AHEaCV-14 [KM874332]	26.4
LSaCV-48	KP153505	CanineCV [KC241982]	41.7	DflaCV-4 [KF738877]	25
LSaCV-48	KP153505	AHEaCV-18 [KM874346]	41.61	OdasCV-15 [KM598398]	24.1
LSaCV-48	KP153505	SaCV-12 [KJ547623]	41.11	BasCV-1 [KJ938716]	23.7
LSaCV-49	KP153507	OdasCV-12 [KM598395]	52.99	SaCV-1 [KJ547620]	27.3
LSaCV-49	KP153507	RW-E [FJ959081]	50.95	MpaCDV-3 [KJ547648]	26.9
LSaCV-49	KP153507	SaCV-6 [KJ547630]	50.93	Nepavirus [JQ898333]	26.2
LSaCV-49	KP153507	SaCV-27 [KM821762]	48.86	SaCV-32 [KM821767]	25.4
LSaCV-49	KP153507	SaCV-32 [KM821767]	48.7	SaCV-19 [KM821754]	22.6
LSaCV-50	KP153511	batCV-SC703 [JN857329]	41.89	OdasCV-16 [KM598411]	28.6
LSaCV-50	KP153511	NG12 [GQ404854]	40.46	SaCV-20 [KM821755]	27
LSaCV-50	KP153511	OdasCV-18 [KM598401]	40.45		
LSaCV-50	KP153511	PK5510 [GQ404847]	40.31		
LSaCV-50	KP153511	OdasCV-18 [KM598402]	39.7		
LSaCV-51	KP153522	SaGmV-10a [KJ547644]	36.51	DfCyclV [JX185418]	32.9
LSaCV-51	KP153522	OdasCV-6 [KM598389]	36.04	SaCV-4 [KJ547628]	32.1
LSaCV-51	KP153522	SaCV-2 [KJ547626]	35.66	OdasCV-2 [KM598399]	28.6
LSaCV-51	KP153522	CpCDV [AM933135]	35.06	AHEaCV-27 [KM874360]	27.7
LSaCV-51	KP153522	FaGmCV-10 [KF371632]	34.94	OdasCV-10 [KM598412]	24.5
LSaCV-52	KP153490	AHEaCV-15 [KM874336]	45	SaCV-34 [KM821769]	27
LSaCV-52	KP153490	SaCV-11 [KJ547622]	43.92	DfCyclV [JX185418]	25.7
LSaCV-52	KP153490	AtCopCV [JQ837277]	42.3		
LSaCV-52	KP153490	SI00850 [JX904473]	41.42		
LSaCV-52	KP153490	RodSCV-M-53 [JF755415]	40.61		
AHEaCV-8	KP153528	AHEaCV-8 [KM874310]	100	AHEaCV-8 [KM874310]	99.7
AHEaCV-8	KP153528	AHEaCV-29 [KM874367]	63.43	GM510 [GQ351275]	25.9
AHEaCV-8	KP153528	AHEaCV-22 [KM874351]	39.41	AHEaCV22 [KM874351]	25
AHEaCV-8	KP153528	ChiSV-GT306 [GQ351278]	36.6	AHEaCV29 [KM874366]	24
AHEaCV-8	KP153528	TuSCV [KF880727]	33.91		
DflaCV-10	KP153516	DflaCV-10 [KF738884]	95.62	DflaCV-10 [KF738885]	62.9
DflaCV-10	KP153516	DflaCV-10a [KF738885]	92.33		
DflaCV-10	KP153516	CynNCXV [JX908739]	43.12		
DflaCV-10	KP153516	SI03654 [JX904548]	39.68		
DflaCV-10	KP153516	LSaCV-11 [KP153409]	39.44		

Query	Accession	Rep pairwise identity	% identity	CP pairwise identity	% identity
DflaCV-3	KP153448	DflaCV-3 [KF738875]	99.3	DflaCV-3[KF738875]	99
DflaCV-3	KP153448	DflaCV-3a [KF738876]	98.6	SaCV-18 [KM821753]	32.8
DflaCV-3	KP153448	Diporeia sp. CV LM3487 [KC248416]	50.74	SaCV-30 [KM821765]	29.4
DflaCV-3	KP153448	SaCV-21 [KM821756]	47.21	SaCV-14 [KJ547625]	29.3
DflaCV-3	KP153448	19-LDMD [KF133826]	44.6	SaCV-11 [KJ547622]	28.7
DflaCV-5	KP153524	DflaCV-5 [KF738878]	100	DflaCV-5 [KF738878]	100
DflaCV-5	KP153524	DflaCV-5a [KF738879]	100	DflaCV-5a [KF738879]	
DflaCV-5	KP153524	RodSCV-V-69 [JF755403]	34.25	AHEaCV-23 [KM874353]	35
DflaCV-5	KP153524	RodSCV-M-45 [JF755409]	33.51	OdasCV-19 [KM598404]	28.9
DflaCV-5	KP153524	AHEaCV-6 [KM874304]	32.13	SaCV-15 [KM821750]	26.4
DflaCV-6	KP153527	DflaCV-6 [KF738880]	99.64	DflaCV-6 [KF738880]	93.4
DflaCV-6	KP153527	AHEaCV-19 [KM874347]	39.71	SaCM-7 [KM877829]	29.7
DflaCV-6	KP153527	DfCyV-3 [JX185424]	38.86	SaCM-6 [KM877828]	26.3
DflaCV-6	KP153527	StCV [KC846095]	38.35	AHEaCV-21 [KM874350]	25
DflaCV-6	KP153527	FWCasCyV [JX569794]	38.1		
DflaCV-8	KP153526	DflaCV-8 [KF738882]	98.8	DflaCV-8 [KF738882]	100
DflaCV-8	KP153526	SaCV-25 [KM821760]	35.22	DfCyclV [JX185418]	27.4
DflaCV-8	KP153526	6-LDMD [KF133813]	34.86		
DflaCV-8	KP153526	FdCV [KC441518]	34.75		
DflaCV-8	KP153526	AHEaCV-11 [KM874323]	33.47		
LSaCM-1	KP153359	DfCyV-5 [JX185427]	35.51		
LSaCM-1	KP153359	DFCyV-5 [JX185426]	35.34		
LSaCM-1	KP153359	Chimp11 [GQ404849]	35.21		
LSaCM-1	KP153359	Chimp12 [GQ404850]	35.21		
LSaCM-1	KP153359	PK5006 [GQ404844]	34.56		
LSaCM-2	KP153360	DFLaCV-3 [KF738875]	43.7		
LSaCM-2	KP153360	DFLaCV-3a [KF738876]	43.33		
LSaCM-2	KP153360	SaCV-33 [KM821768]	43.14		
LSaCM-2	KP153360	SOG05268 [JX904185]	43.08		
LSaCM-2	KP153360	CB-A [FJ959082]	42.86		
LSaCM-3	KP153361	21-LDMD [KF133828]	39.33		
LSaCM-3	KP153361	15-LDMD [KF133822]	37.14		
LSaCM-3	KP153361	SI04276 [JX904605]	36.1		
LSaCM-3	KP153361	11-LDMD [KF133818]	35.19		
LSaCM-3	KP153361	YN-BtCV3 [JF938080]	34.44		
LSaCM-4	KP153363	BarCV [JF279961]	29.59		
LSaCM-4	KP153363	SI03513 [JX904541]	29.33		
LSaCM-4	KP153363	BarCV [GU799606]	28.4		
LSaCM-4	KP153363	PeYDV [EU921828]	27.85		
LSaCM-4	KP153363	SaCV-8 [KJ547632]	26.97		
LSaCM-5	KP153364	SOG00182 [JX904077]	61.54		
LSaCM-5	KP153364	SOG05268 [JX904185]	59.77		
LSaCM-5	KP153364	GOM03041 [JX904344]	48.87		
LSaCM-5	KP153364	SI00197 [JX904420]	47.76		
LSaCM-5	KP153364	RodSCV-M-53 [JF755415]	47.74		
LSaCM-6	KP153368	AHEaCV-9 [KM874315]	42.25		
LSaCM-6	KP153368	SI03931 [JX904581]	39.15		
LSaCM-6	KP153368	PK5034 [GQ404845]	38.17		
LSaCM-6	KP153368	DfCyV-3 [JX185424]	37.98		
LSaCM-6	KP153368	SOG03994 [JX904139]	37.93		
LSaCM-7	KP153370	RodSCV-M-53 [JF755415]	43.54		
LSaCM-7	KP153370	LSaCV-38 [KP153483]	43.27		
LSaCM-7	KP153370	SaCV-21 [KM821756]	43.18		
LSaCM-7	KP153370	SI00197 [JX904420]	42.07		
LSaCM-7	KP153370	LSaCV-32 [KP153464]	41.92		
LSaCM-8	KP153371	DFLaCV-2 [KF738874]	62.26		
LSaCM-8	KP153371	SaCV-7 [KJ547631]	59.25		
LSaCM-8	KP153371	SaCV-6 [KJ547630]	58.74		
LSaCM-8	KP153371	SaCV-27 [KM821762]	57.36		
LSaCM-8	KP153371	18-LDMD [KF133825]	56.27		
LSaCM-9	KP153377	RodSCV-M-53 [JF755415]	47.76		
LSaCM-9	KP153377	19-LDMD [KF133826]	45.65		
LSaCM-9	KP153377	SI03717 [JX904562]	45.32		
LSaCM-9	KP153377	GOM03041 [JX904344]	44.4		
LSaCM-9	KP153377	DFLaCV-3 [KF738875]	43.48		
LSaCM-10	KP153379	DfOrV [JX185416]	31.97%		
LSaCM-10	KP153379	AHEaCV-24 [KM874354]	30.76%		
LSaCM-10	KP153379	ChiSV-GT306 [GQ351278]	30.57%		
LSaCM-10	KP153379	AHEaCV-22 [KM874351]	30.26%		
LSaCM-10	KP153379	MpaCDV-8 [KJ547653]	29.75%		
LSaCM-11	KP153384			SI00529 [JX904443]	29
LSaCM-11	KP153384			SI04136 [JX904591]	28
LSaCM-11	KP153384			SI03746 [JX904565]	28
LSaCM-11	KP153384			4-LDMD [KF133811]	27
LSaCM-11	KP153384			Mosquito VEM SDRBAJ [HQ335087]	26

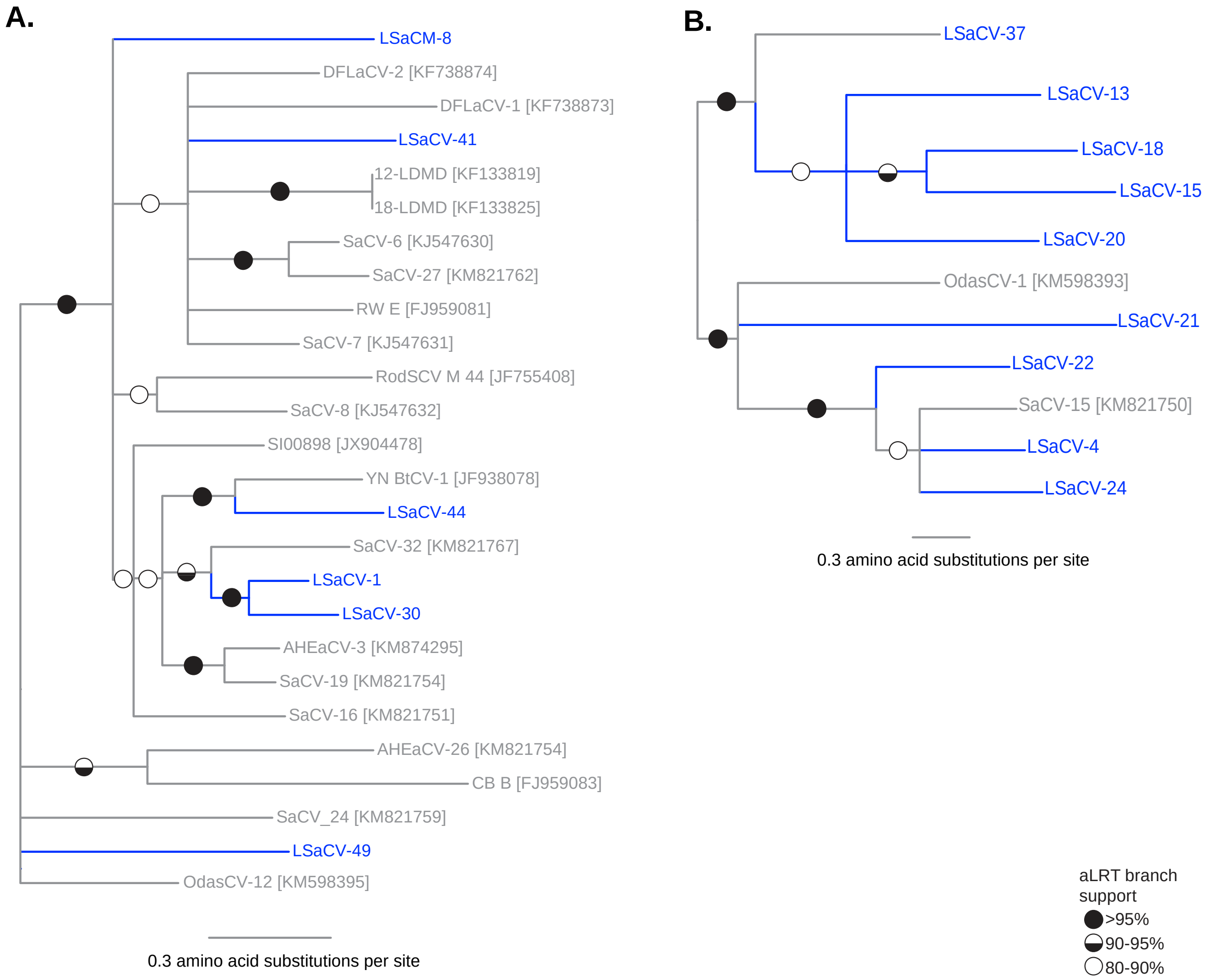


Figure 7.3: Enlarged sections of Figure 7.2. Maximum likelihood phylogenetic tree with aLRT branch support of the closely related CRESS DNA viral Rep sequences that do not fall within current recognised CRESS DNA viral families.

LSaCM-8 and LSaCV-41, -44, -1, -30 and -49 also form a clade with CRESS DNA viruses recovered from sewage, recycled water, dragonfly larvae, rodent stools, marine water, dragonflies and bat guano (Dayaram *et al.*, 2014a; Dayaram *et al.*, 2014b; Dayaram *et al.*, 2015; Kraberger *et al.*, 2014; McDaniel *et al.*, 2014; Phan *et al.*, 2011; Rosario *et al.*, 2009a) (Figure 7.3). With the growing number of CRESS DNA viruses discovered it is likely that these clades will become resolved so they are able to become formally classified.

Conservation in CRESS DNA viral genomes is usually present in the conserved domains of the Rep. These domains include the rolling circle replication (RCR) motifs and superfamily three (SF3) helicase motifs which play different roles in initiating replication of CRESS DNA viruses (Londoño *et al.*, 2010). The level of conservation observed in the Rep is due to the role it plays in mediating the binding of the Rep with the dsDNA intermediate via the specificity binding determinants (SPDs) (Londoño *et al.*, 2010). The precise function of the motifs is still unknown, however, RCR motif I is thought to recognise iteron sequences when co-ordinating the binding of the Rep. RCR motif II and III covalently bind a conserved tyrosine residue to the DNA that initiate DNA cleavage of the dsDNA intermediate when the Rep binds (Steinfeldt *et al.,* 2006).

The SF3 helicase motifs are located in the helicase domain. All motifs exhibit helicase activity during RCR forming a "P-loop" structure that acts as a deoxyribonucleotide triphosphate (dNTP) binding domain (Gorbalenya *et al.*, 1990; Walker *et al.*, 1982). In most cases these motifs were identified in all viral Reps, however, the Reps LSaCV-35 and LSaCM-4 were missing all RCR motifs and some of the SF3 helicase motifs (Table 7.2). Reps from LsaCV-4, -13,-37 and DflaCV-5 were also missing the SF3 helicase Walker-A motif, whereas LsaCV-17 and LSaCM-1 were missing RCR motif I (Table 7.2).

The LSaCM that encode putative Reps also contained rolling circle replication (RCR) motifs I, II and II as well as the SF3 helicase motifs and varied in the degree of conservation, with LSaCM-4 not containing any of the motifs (Table 7.2). Overall, there is some level of conservation observed in the Reps isolated in this study in the RCR and SF3 helicase motifs.

Of the 139 viral genomes recovered representing 52 unique viral species. Of all the viral genomes recovered in this study, the putative CPs in 18 genomes did not share significant pairwise identity with those from previously characterised CRESS DNA viruses (Table 7.4), illustrating that we can only speculate that these ORFs encode CPs. The CP of LsaCV-25 shared ~93.4% pairwise amino acid identity with AHEaCV-17 (KM874343). CPs of LSaCV-27 and -38 shared 37.1% and 38.8% pairwise identity respectively with those of SaCV-18 (KM821753) and SaCV-13 (KJ547624). The only small circular molecule to have a major ORF that hit to a CP was LSaCM-11, this shared ~28.29% pairwise identity with viral isolates recovered from the marine metagenome from the Sannich Inlet, British Columbia (Labonté & Suttle, 2013). The low pairwise identity observed in CPs from the LSaCV and LSaCM viral sequences suggest there are very low levels of amino acid conservation in this protein, which has been evident throughout most novel CRESS DNA viral CP analysis.

7.4.4 Analysis of AHEaCV-8 with other Chipovirus-like sequences

AHEaCV-8, previously isolated from samples (*A. stutchburyi*, *P. subtriangulata*, *A. crenata* and benthic sediment) collected from the Avon-Heathcote estuary (Chapter 6) was also recovered from the mussel species *E. menzeisi*. These viral isolates are most closely related to the previously described chipovirus sequences. These include CRESS DNA viruses recovered from chimpanzee, bovine, turkey and porcine stools, as well as dragonflies *(Libellula quadrimaculata* and *Erythrodiplax fusca),* benthic sediment and various mollusc species. The genomes of these viruses vary between ~1,800 nt to ~2,500 nts, and have two major ORFs that have varying genome organisation. Both major ORFs are relatively conserved across these viral proteins and encode either a Rep or CP.

AHEaCV-8 has a unidirectionally transcribed genome (2,187 nt). The isolate recovered in this study named AHEaCV-8-LSMU-2014, shares 100% Rep and ~99.7% CP pairwise identity with the previously described AHeaCV-8 isolates (Table 7.4), and overall between ~98.4-98.6% genome wide nucleotide identity. The Rep of AHEaCV-8-LSMU-2014 shares between ~25-63% amino acid identity with other chipovirus sequences whilst the CP shares ~18-25% amino acid identity with other chipovirus CP sequences. Further phylogenetic analysis of the Rep and CP (Figure 7.4) indicates that the Rep of AHEaCV-8 groups with the other chipovirus sequences whilst the CP falls basal to the chipoviruses CPs, clustering with AHEaCV-29 and AHEaCV-22. Given the difference in genome size and architecture, low amino acid identities to other chipoviruses sequences in both major proteins, and that only the Rep of AHEaCV-8 groups with the chipoviruses; it suggests that this virus is only distantly related to chipovirus. This is the second report of AHEaCV-8 being isolated from molluscs, suggesting that the molluscs may be potential hosts for these viruses.

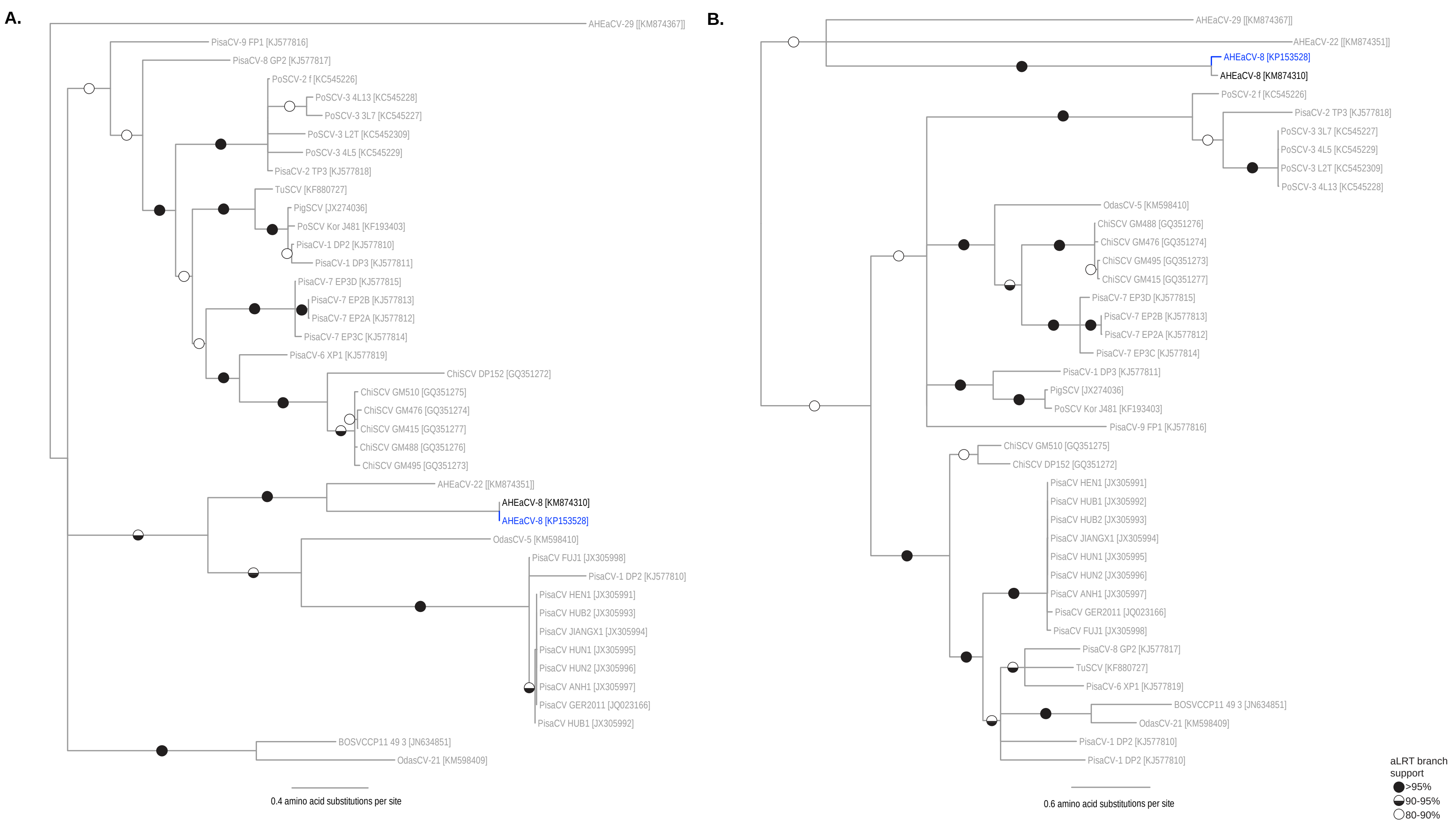


Figure 7.4: Maximum likelihood phylogenetic trees with aLRT branch support and mid-point rooted (a) Enlarged sections of Figure 7.2. showing the Reps of chipoviruses (b) Phylogenetic tree of the CPs of chipoviruses.

However, given the concentrating nature of molluscs, it is possible that these viruses may be present in the ecosystem and as thus are concentrated in the molluscs as a result.

7.4.5 DflaCV isolates

DflaCV-3, -5, -6, -8 and -10 were recovered from different sample types (water, benthic sediment, oligochaeta, *C. zealandicus, E. menziesi, P. antipodarum, P. acuta, X. zealandica, P. grayi* and *M. novazelandiae)* from Lake Sarah. These viral genomes vary in size ~1,600 to 2,200 nt with all having bidirectionally organised ORFs which encode for a putative Rep and CP (Figure 7.1). At a nucleotide level the genomes shared 95-100% identity with the previously described isolates (Dayaram et al., 2014a). Phylogenetic analysis of the Reps of the viral isolates show that they fall with the previously described DflaCVs, but are distantly related to most other CRESS DNA viruses (Figure 7.2). In addition, pairwise identities of both major proteins highlight that they share 95-100% Rep and 93-100% CP amino acid pairwise identity (Table 7.4). Further full genome nucleotide pairwise identity showed that the previously described DflaCV-5 (isolate NZPG2LG) and DflaCV-5 (isolate NZPG7LS) shared 100% and 99.8% pairwise identity with the DfLaCV-5 isolated recovered in this study. DflaCV-6 (isolate NZPG9LD) and DflaCV-8 (isolate NZPG5LS) shared 99.9% and 99.6% pairwise identity with DflaCV-6 genomes recovered in this study. Whilst the previously described DflaCV-3 (isolate NZPG1LG) shared between 97.8-98.8% pairwise identity with the four DflaCV-3 isolates recovered in this study. The initial DflaCV10 (isolate NZXZ1LH) from Lake Hawdon shared between 83.7-87.3% pairwise identity and DflaCV-10 (isolate NZXZ2LS) from Lake Sarah shared between 85.3-88.1% identity with the six isolates from this study.

Given that the previous study isolated DflaCV-5, -8 and -10 from Lake Sarah from *X. zealandica* and *P.grayi* (Dayaram *et al.*, 2014a), it is not surprising that these viral genomes have been recovered from the same location in water, benthic sediment, *E. menzeisi,* Oligochaeta and *C. zealandicus* samples. DflaCV-3 was originally isolated from *P. grayi* samples taken from Lakes Grasmere and Lake Hawdon which are 650 m and 7.2 km respectively from Lake Sarah (Figure 7.5 and 7.6). However, DflaCV-6 was initially recovered from *P. grayi* sampled at Lake Donne, over 171 km from Lake Sarah.

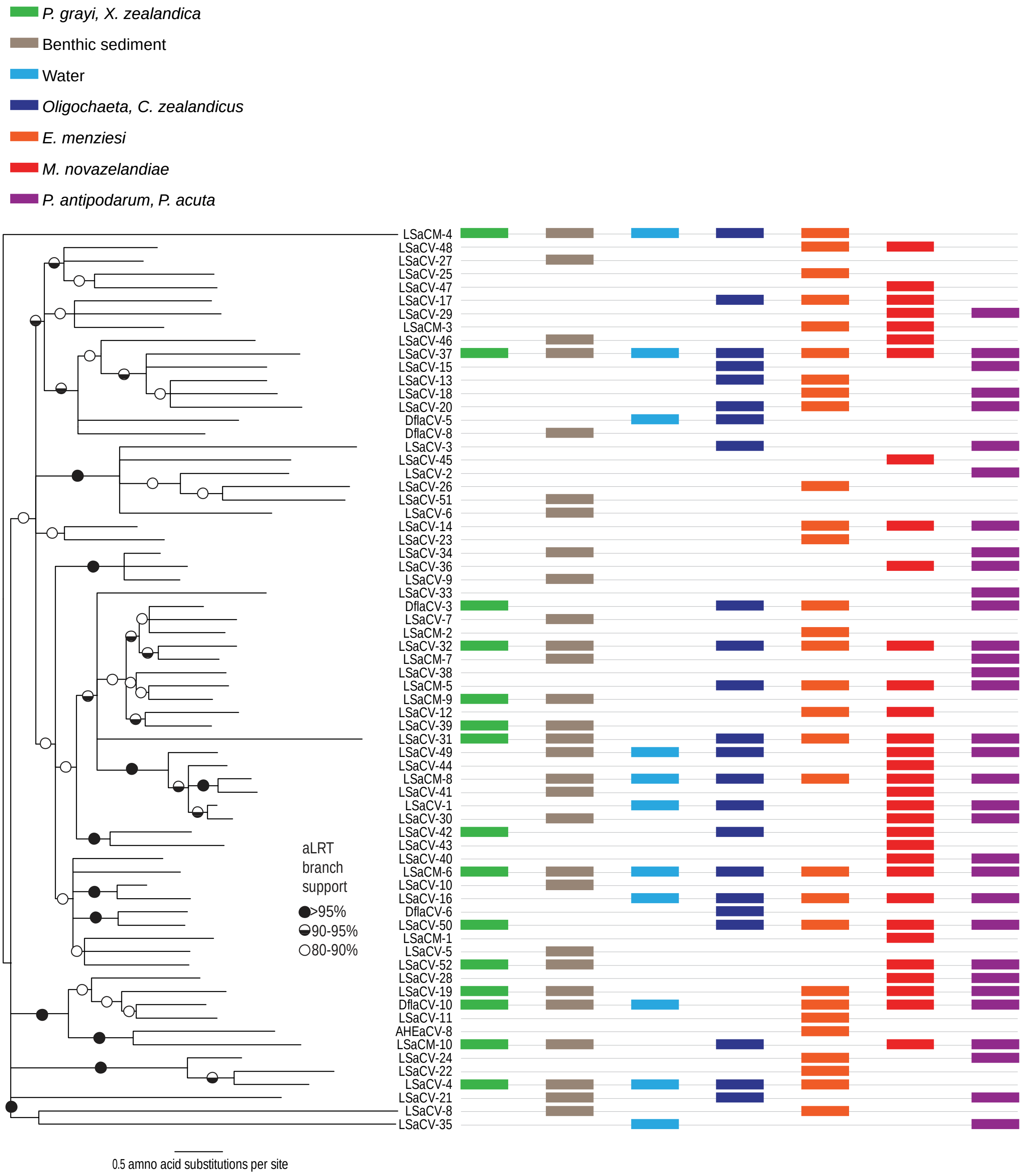


Figure 7.5: Maximum likelihood phylogenetic tree mid-point rooted using aLRT branch support of all Rep sequences recovered from Lake Sarah and the isolation source of each isolate.

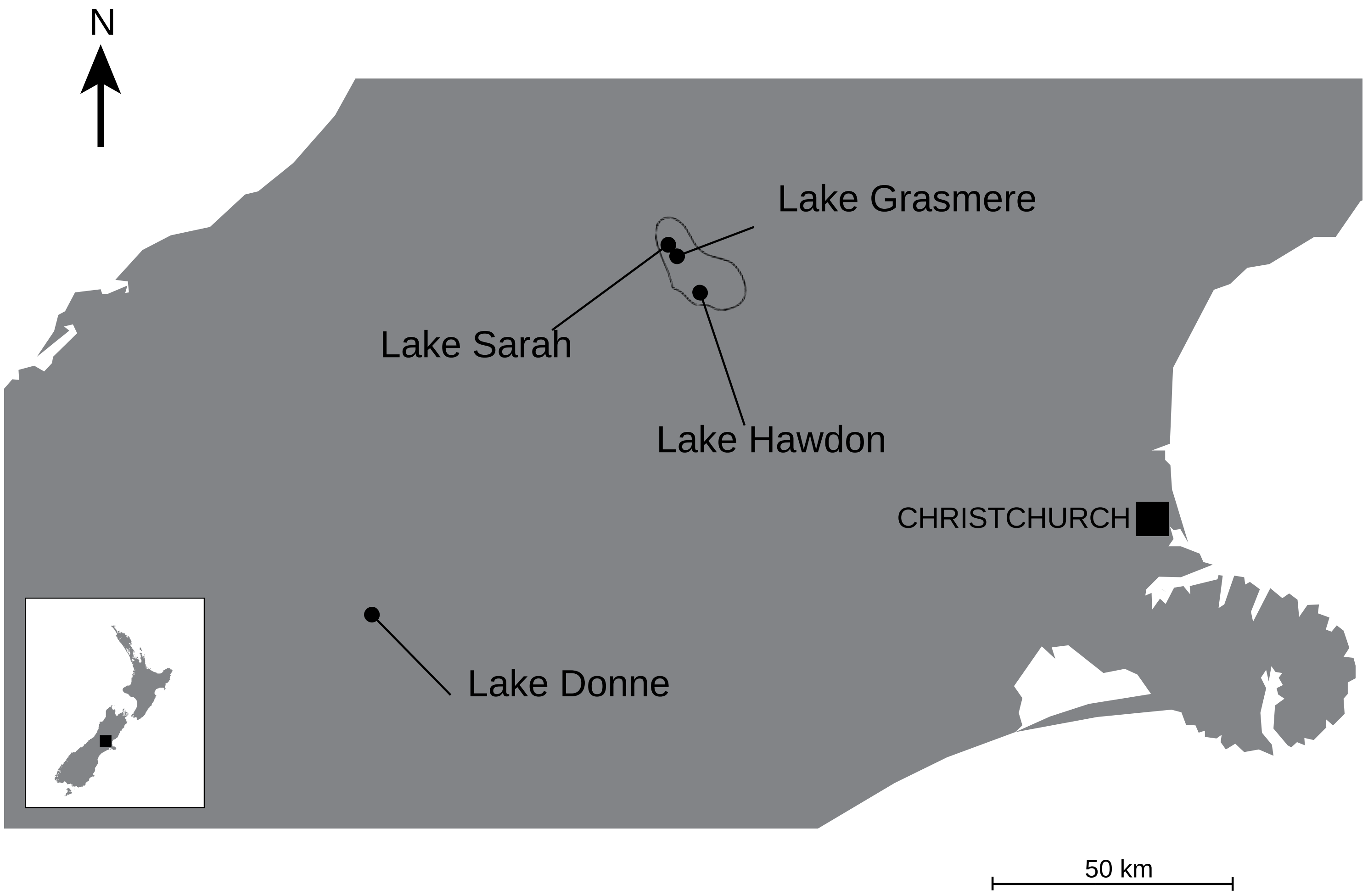


Figure 7.6: Map of the partial region of the South Island of New Zealand showing the four sampling locations of all dragonfly larvae-associated viruses (DflaCVs). The dotted circle indicates the Cass Basin area.

Given the close proximity of Lake Hawdon and Lake Grasmere in relation to Lake Sarah it is not surprising that these viruses are also present in the larvae as there ecosystem are very close, and further suggests that DflaCVs are potentially ubiquitous in freshwater environments. In addition, as dragonflies are mobile and can travel up to 1,000 km (Anderson, 2009) and lay their larvae in different lakes in the vicinity, it can be assumed that there may be a 'flow' of these viruses between lakes in the Cass Basin. This also may account for why DflaCV-3 is also present in Lake Donne.

7.4.6 Relationship between sample types based on viral distribution

The viral distribution analysis gives an indication of how different organisms and samples are related to each other based on the presence or absence of the CRESS DNA viruses isolated in this study. Both FFG and defended-undefended cluster dendrograms showed a relationship between the molluscs , Oligochaeta and *C. zealandicus.* Overall, limited interactions between sample types based on the presence and absence of CRESS DNA viruses was observed (Figure 7.7). As molluscs are filter feeders they are closely associated with water and sediment. However, no association was found between mollusc species (defended), water or sediment (Figure 7.7). This was also the case for Odonata larvae (predators) which would likely be associated with soft bodied undefended prey such as Oligochaeta and *C. zealandicus.* Mollusc species share closer relationships based on viral presence which is what would be expected as they are all filter feeders with similar feeding behaviours, diets and morphology. The Odonata (predators) shared the closest relationship with the benthic sediment based on the CRESS DNA viruses isolated from them. As Odonata larvae live in the benthic sediment of freshwater environments it could be expected that their CRESS DNA viruses isolated would be similar.

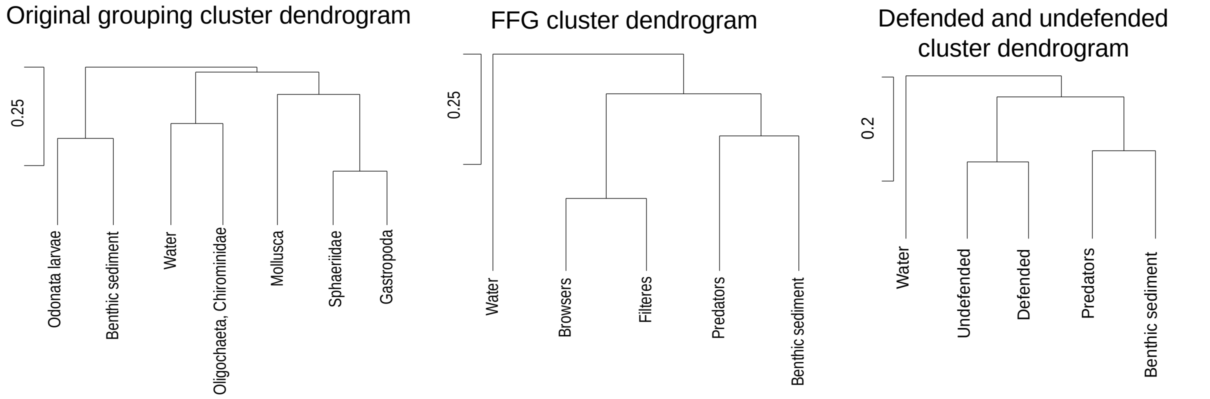


Figure 7.7: (A) Original cluster denogram Odonata larvae (*P. grayi* and *X. zealandica*), benthic sediment, water, Oligochaeta and *C. zealandicus*, Mollusca (*E. menzeisi*), Sphaeriidae (*M. novazelandiae*) and Gastropoda (*P. acuta* and *P. antipodarum*). FFG cluster dendrogram, water, browsers (Oligochaeta, *C. zealandicus P. acuta* and *P. antipodarum*), filters (*E. menzeisi* and *M. novazelandiae,*), predators (*P. grayi* and *X. zealandica*) and benthic sediment. Hierarchical cluster dendrogram water, undefended (Oligochaeta and *C. zealandicus*), defended (*E. menzeisi*, *M. novazelandiae*, *P. acuta* and *P. antipodarum*), predators *P. grayi* and *X. zealandica*) and benthic sediment.

7.5 Concluding remarks

In this study 139 new viral sequences were recovered, representing 52 novel CRESS DNA viruses and 31 circular DNA molecules representing 11 unique molecules. These were isolated from benthic sediment, water, oligochaetes, four species of molluscs, midge larvae and Odonata larvae sampled in Lake Sarah in the Cass Basin, New Zealand. The isolation of small circular molecules has been reported in previous metagenomic studies (Kraberger *et al.*, 2014; Rosario *et al.*, 2012a; Seguin *et al.*, 2014). The function of most of these small circular molecules is still unknown. However, some have been associated with geminiviruses and are presumed to be alphasatellites as they usually have one ORF which encodes for a Rep (Idris *et al.*, 2014). As some alphasatellites have been shown to provide viruses with a selective advantage (Nawaz-ul-Rehman *et al.*, 2010) it is possible that some of these molecules being discovered that encode putative Reps are not just defective genomes but may be satellites associated with other CRESS DNA viruses.

An analysis of the putative ‘flow’ of LSaCVs and LSaCMs between different samples (*P. grayi, X. zealandica,* benthic sediment, water, Oligochaeta, *C. zealandicus, E. menziesi, M. novazelandiae, P. antipodarum* and *P. acuta*) show that the viral sequences were present in multiple samples (Figure 7.5). Combining phylogenetic analysis of the Rep sequences with the isolation source provides a visual representation of links between phylogenetic relatedness and isolation source. Most viruses recovered were from the four species of molluscs (59%) followed by the benthic sediment (15.8%) and the worms and midge larvae (11.5%). Whilst the least number were recovered from water (5.8%) and the dragonfly and damselfly larvae (7.9%). This is similar to the results from the Chapter 6, that found most CRESS DNA viruses were isolated from molluscs (84.8%) rather than benthic sediment (15.2%), further illustrating that molluscs represent a great surveillance sampling tool for the monitoring of CRESS DNA viruses in ecosystems. It is interesting to note that most viruses that were recovered from the dragonfly and damselfly larvae were also isolated from the benthic sediment. As the larvae live within the benthic sediment (Corbet & Brooks, 2008), it is not surprising that similar CRESS DNA viruses are observed across both sample types.

Sclerotinia sclerotiorum hypovirulence-associated DNA virus 1 (SsHADV-1) was originally isolated in China (Yu *et al.*, 2010). This virus has now been isolated from benthic river sediments from New Zealand (Kraberger *et al.*, 2013) and two species of dragonflies sampled in Oklahoma, United States of America (Chapter 4). This is an example of how the continuing identification of CRESS DNA viruses expands the overall baseline data. This in turn enables these viruses to be identified easily in other environments.

Examples of how this baseline data is becoming useful is the isolation of AHEaCV-8 and the five DflaCVs recovered from across the seven sample types from Lake Sarah. AHEaCV-8 was originally isolated from samples taken from the Avon-Heathcote estuary, including three different species of molluscs and benthic sediment (Chapter 6) and has now been identified in *E. menziesi* from Lake Sarah. The additional discovery of five DflaCVs isolates from different sample types and separate lakes again illustrates how known CRESS DNA viruses can be used as probes to identify these viruses in other organisms and ecosystems. DflaCV-3, -6 and -10 were identified in samples from Lake Sarah, where they were previously not identified (Chapter 3), suggests there is movement of these viruses between lakes or that DflaCVs are potentially prevalent in freshwater ecosystems. As DflaCVs were not detected in adult dragonfly samples from the same ecosystem this suggests that there may be another factor responsible for the flow of these viruses between ecosystems. Overall, the CRESS DNA viruses being discovered are not limited by specific ecosystems and there may be movement of these viruses between environments.

Finally, cluster analysis of organisms based on feeding behaviour, and morphology did not reflect the CRESS DNA viral distribution across sample types. This analysis demonstrates that understanding viral interactions is complex and requires a better understanding of virus-host interactions which will help to further understand how these viruses are transmitted and the flow of them in ecosystems. Understanding the viral distribution and diversity in ecosystems is vital baseline data that is needed to further shed light on potential hosts, host-virus evolution and to access the movement of CRESS DNA viruses between ecosystems.

GenBank accession numbers: KP153359 – KP153528

Table 7.5: List of full CRESS DNA viral sequences and their accession number in Fig 7.1., Fig 7.2. and Fig 7.3.

Acronym	Genbank #	Acronym	Genbank #	Acronym	Genbank #	Acronym	Genbank #	Acronym	Genbank #	Acronym	Genbank #
10-LDMD	KF133817	Baminivirus	KJ938716	FaGmCV-11	KF371631	LSaCV-16-LSWO-2013	KP153422	MCMValpha	HM163578	RodSCV V 81	JF755412
11-LDMD	KF133818	Baminivirus	JQ898331	FaGmCV-12	KF371630	LSaCV-16-LSGA-2013	KP153423	MiCVDL-13	KJ020099	RodSCV V 84	JF755413
12-LDMD	KF133819	BarCV	GU799606	FaGmCV-1a	KF371643	LSaCV-17-LSMU-2013	KP153424	MiSV	D00800	RodSCV V 86	JF755416
13-LDMD	KF133820	BarCV	JF279961	FaGmCV-1b	KF371642	LSaCV-17-LSCO-2013	KP153425	MmCV	JQ085285	RodSCV V 87	JF755406
14-LDMD	KF133821	BasCV-1	KJ938716	FaGmCV-1c	KF371641	LSaCV-17-LSWO-2013	KP153426	MmFV	JN704610	RodSCV V 91	JF755417
15-LDMD	KF133822	BasCV-2	KM510191	FaGmCV-2	KF371640	LSaCV-18-LSMU-2013	KP153427	Mosquito VEM SDBVL-G	HQ335086	RodSCV V 97	JF755414
16-LDMD	KF133823	BatCV SC703	JN857329	FaGmCV-3	KF371639	LSaCV-18-LSGA-2013	KP153428	MpaCDV-1	KJ547646	RW-A	FJ959077
17-LDMD	KF133824	BatCV TM6C	HM228875	FaGmCV-4	KF371638	LSaCV-19-LSMU-2013	KP153429	MpaCDV-2	KJ547647	RW-B	FJ959078
18-LDMD	KF133825	batCyVGF4c	HM228874	FaGmCV-5	KF371637	LSaCV-19-LSSO-2013	KP153430	MpaCDV-3	KJ547648	RW-C	FJ959079
19LDMD	KF133826	BBC-A	FJ959086	FaGmCV-6	KF371636	LSaCV-19-LSCO-2013	KP153431	MpaCDV-4	KJ547649	RW-D	FJ959080
1-LDMD	KF133807	BBTV	EU531473	FaGmCV-7	KF371635	LSaCV-19-LSGA-2013	KP153432	MpaCDV-5	KJ547650	RW-E	FJ959081
20-LDMD	KF133827	BBTV	EU531473	FaGmCV-8	KF371634	LSaCV-19-LSLA-2013	KP153433	MpaCDV-6	KJ547651	SaCV-1	KJ547620
21-LDMD	KF133828	BBTV-SAT	L32166	FaGmCV-9	KF371633	LSaCV-20-LSMU-2013	KP153434	MpaCDV-7	KJ547652	SaCV-10	KJ547621
2-LDMD	KF133808	BCSMV	HQ113104	FaGmV-13	KJ938717	LSaCV-20-LSWO-2013	KP153435	MpaCDV-8	KJ547653	SaCV-11	KJ547622
3-LDMD	KF133810	BCTIV	EU273818	FBNSV	GQ150778	LSaCV-20-LSGA-2013	KP153436	MS584-5	HQ322117	SaCV-12	KJ547623
4-LDMD	KF133811	BCTIV	EU273818	FBNYV	AJ132187	LSaCV-21-LSMU-2013	KP153437	MSRV	JQ624880	SaCV-13	KJ547624
5-LDMD	KF133812	BCTIV	JX082259	FbSLCV-2	JX094281	LSaCV-21-LSWO-2013	KP153438	MSV	AF329881	SaCV-14	KJ547625
6-LDMD	KF133813	BCTV	M24597	FdCV	KC441518	LSaCV-21-LSSO-2013	KP153439	MVDV	AB000920	SaCV-15	KM821750
7-LDMD	KF133814	BDV	AM922261	FiCV	DQ845075	LSaCV-21-LSGA-2013	KP153440	MYMCalpha	FN675297	SaCV-16	KM821751
8-LDMD	KF133815	BeYDV	AM849096	FSfaCV	KF246569	LSaCV-22-LSMU-2013	KP153441	Nepavirus	JQ898333	SaCV-17	KM821752
9-LDMD	KF133816	BFDV	AF071878	FWCasCyV	JX569794	LSaCV-23-LSMU-2013	KP153442	NG10	GQ404895	SaCV-18	KM821753
ABTV	EF546807	BGYMV	D00201	GasCSV	KC172652	LSaCV-24-LSMU-2013	KP153443	NG12	GQ404854	SaCV-19	KM821754
ACMV	GQ204107	BMCTV	AY134867	GCFaV	JQ901105	LSaCV-24-LSGA-2013	KP153444	NG13	GQ404856	SaCV-2	KJ547626
ACMV	J02057	BOSVCCP11493	JN634851	GoCV	DQ192280	LSaCV-25-LSMU-2013	KP153445	NG14	GQ404855	SaCV-2	KJ547626
AHEaCV-1	KM874291	BSCTV	X97203	GOM00012	JX904192	LSaCV-26-LSMU-2013	KP153446	NGchicken15	HQ738644	SaCV-20	KM821755
AHEaCV-1	KM874290	BtCV	JN377566	GOM00443	JX904231	DflaCV-3-LSLA-2013	KP153447	NGchicken8	HQ738643	SaCV-21	KM821756
AHEaCV-10	KM874322	CaCV	AJ301633	GOM00546	JX904245	DflaCV-3-LSGA-2013	KP153448	Niminivirus	KJ938716	SaCV-22	KM821757
AHEaCV-10	KM874321	Canarypoxvirus	NP955176	GOM00583	JX904250	DflaCV-3-LSMU-2013	KP153449	OdasCV-1	KM598393	SaCV-23	KM821758
AHEaCV-10	KM874320	CanCV	JQ821392	GOM02856	JX904312	DflaCV-3-LSWO-2013	KP153450	OdasCV-10	KM598412	SaCV-24	KM821759
AHEaCV-10	KM874319	CB-A	FJ959082	GOM02962	JX904333	LSaCV-28-LSGA-2013	KP153451	OdasCV-11	KM598394	SaCV-25	KM821760
AHEaCV-11	KM874327	CB-B	FJ959083	GOM03041	JX904344	LSaCV-28-LSCO-2013	KP153452	OdasCV-12	KM598395	SaCV-26	KM821761
AHEaCV-11	KM874326	CCDaV	JQ920490	GOM03161	JX904368	LSaCV-28-LSLA-2013	KP153453	OdasCV-13	KM598396	SaCV-27	KM821762
AHEaCV-11	KM874325	CFDV	M29963	GOM03193	JX904377	LSaCV-29-LSGA-2013	KP153454	OdasCV-14	KM598397	SaCV-28	KM821763
AHEaCV-11	KM874324	CGMV	AF029217	GuCV	DQ845074	LSaCV-29-LSCO-2013	KP153455	OdasCV-16	KM598411	SaCV-29	KM821764
AHEaCV-11	KM874323	ChCDV	AM850136	GuCV	JQ685854	LSaCV-30-LSSO-2013	KP153456	OdasCV-17	KM598400	SaCV-3	KJ547627
AHEaCV-12	KM874328	Chimp11	GQ404849	HJasCV	KF413620	LSaCV-30-LSGA-2013	KP153457	OdasCV-18	KM598401	SaCV-3	KJ547627
AHEaCV-13	KM874331	Chimp12	GQ404850	HrCTV	U49907	LSaCV-30-LSCO-2013	KP153458	OdasCV-19	KM598404	SaCV-30	KM821765
AHEaCV-13	KM874330	chimp17	GQ404851	hs1	JX559621	LSaCV-31-LSGA-2013	KP153459	OdasCV-2	KM598399	SaCV-31	KM821766
AHEaCV-13	KM874329	ChiSCV DP152	GQ351272	hs2	JX559622	LSaCV-31-LSSO-2013	KP153460	OdasCV-20	KM598406	SaCV-32	KM821767
AHEaCV-14	KM874335	ChiSCV GM415	GQ351277	HuCyV-5841A	KF726986	LSaCV-31-LSCO-2013	KP153461	OdasCV-21	KM598409	SaCV-33	KM821768
AHEaCV-14	KM874334	ChiSCV GM476	GQ351274	HuCyV-7046A	KF726987	LSaCV-31-LSWO-2013	KP153462	OdasCV-3	KM598407	SaCV-34	KM821769
AHEaCV-14	KM874333	ChiSCV GM488	GQ351276	HuCyV-7078A	KF726984	LSaCV-31-LSLA-2013	KP153463	OdasCV-4	KM598408	SaCV-35	KM821770
AHEaCV-14	KM874332	ChiSCV GM495	GQ351273	HuCyV-7081A	KF726985	LSaCV-32-LSGA-2013	KP153464	OdasCV-5	KM598410	SaCV-36	KM821748
AHEaCV-15	KM874339	ChiSCV GM510	GQ351275	LaCopCV	JF912805	LSaCV-32-LSSO-2013	KP153465	OdasCV-7	KM598390	SaCV-37	KM821749
AHEaCV-15	KM874338	ChiSCV GT306	GQ351278	LSaCM-1-LSCO-2013	KP153359	LSaCV-32-LSCO-2013	KP153466	OdasCV-8	KM598391	SaCV-4	KJ547628
AHEaCV-15	KM874337	CLCRV	AM501481	LSaCM-2-LSMU-2013	KP153360	LSaCV-32-LSMU-2013	KP153467	OdasCV-9	KM598392	SaCV-4	KJ547628
AHEaCV-15	KM874336	ClGMV	DQ641692	LSaCM-3-LSMU-2013	KP153361	LSaCV-32-LSWO-2013	KP153468	ODV	AM296025	SaCV-5	KJ547629
AHEaCV-16	KM874342	CoCV	KF738868	LSaCM-3-LSCO-2013	KP153362	LSaCV-33-LSGA-2013	KP153469	PanSV-A	L39638	SaCV-6	KJ547630
AHEaCV-16	KM874341	CoGMV	EU636712	LSaCM-4-LSSO-2013	KP153363	LSaCV-34-LSSO-2013	KP153471	PCV1	AF012107	SaCV-7	KJ547631
AHEaCV-16	KM874340	CpCDV	AM933135	LSaCM-5-LSWO-2013	KP153364	LSaCV-35-LSGA-2013	KP153472	PCV1-2	FJ655418	SaCV-8	KJ547632
AHEaCV-17	KM874345	CpCV	GU256530	LSaCM-5-LSCO-2013	KP153365	LSaCV-35-LSWA-2013	KP153473	PCV2	AY424401	SaCV-9	KJ547633
AHEaCV-17	KM874344	CpRV	GU256532	LSaCM-5-LSMU-2013	KP153366	LSaCV-36-LSGA-2013	KP153474	PeCTV	EF501977	SaGmV-10a	KJ547644
AHEaCV-17	KM874343	CpYV	JN989439	LSaCM-5-LSGA-2013	KP153367	LSaCV-36-LSCO-2013	KP153475	PeYDV	EU921828	SaGmV-10b	KJ547645

Acronym	Genbank #	Acronym	Genbank #	Acronym	Genbank #	Acronym	Genbank #	Acronym	Genbank #	Acronym	Genbank #
AHEaCV-18	KM874346	CSMV	M20021	LSaCM-6-LSSO-2013	KP153368	LSaCV-37-LSGA-2013	KP153476	PigSCV	JX274036	SaGmV-12	KJ547642
AHEaCV-19	KM874347	CyCV	EU056309	LSaCM-6-LSGA-2013	KP153369	LSaCV-37-LSMU-2013	KP153477	PisaCV ANH1	JX305997	SaGmV-2	KJ547642
AHEaCV-2	KM874294	CyCV-TB	HQ738637	LSaCM-7-LSSO-2013	KP153370	LSaCV-37-LSLA-2013	KP153478	PisaCV FUJ1	JX305998	SaGmV-3	KJ547643
AHEaCV-2	KM874293	CynNCKV	JX908740	LSaCM-8-LSGA-2013	KP153371	LSaCV-37-LSSO-2013	KP153479	PisaCV GER2011	JQ023166	SaGmV-4	KJ547634
AHEaCV-2	KM874292	CynNCXV	JX908739	LSaCM-8-LSCO-2013	KP153372	LSaCV-37-LSWA-2013	KP153480	PisaCV HEN1	JX305991	SaGmV-5	KJ547635
AHEaCV-20	KM874349	CyVN-cs1	KF031471	LSaCM-8-LSWO-2013	KP153373	LSaCV-37-LSWO-2013	KP153481	PisaCV HUB1	JX305992	SaGmV-6	KJ547636
AHEaCV-20	KM874348	CyVN-hcf	KF031466	LSaCM-8-LSSO-2013	KP153374	LSaCV-37-LSCO-2013	KP153482	PisaCV HUB2	JX305993	SAR-A	FJ959084
AHEaCV-21	KM874350	CyVN-hcf1	KF031465	LSaCM-8-LSWA-2013	KP153375	LSaCV-38-LSGA-2013	KP153483	PisaCV HUN1	JX305995	SAR-B	FJ959085
AHEaCV-22	KM874352	CyVN-hcf3	KF031467	LSaCM-8-LSMU-2013	KP153376	LSaCV-38-LSCO-2013	KP153484	PisaCV HUN2	JX305996	SCSV	AAA68022
AHEaCV-22	KM874351	CyVN-hcf4	KF031468	LSaCM-9-LSLA-2013	KP153377	LSaCV-39-LSGA-2013	KP153485	PisaCV JIANGX1	JX305994		
AHEaCV-23	KM874353	CyVN-hcf5	KF031469	LSaCM-9-LSSO-2013	KP153378	LSaCV-39-LSSO-2013	KP153486	PK5006	GQ404844		
AHEaCV-24	KM874354	CyVN-ps1	KF031470	LSaCM-10-LSLA-2013	KP153379	LSaCV-39-LSLA-2013	KP153487	PK5034	GQ404845		
AHEaCV-25	KM874358	DfaCV-1	JX185430	LSaCM-10-LSCO-2013	KP153380	LSaCV-40-LSCO-103520-13	KP153488	PK52222	GQ404846		
AHEaCV-25	KM874357	DfasMV	JX458740	LSaCM-10-LSWO-2013	KP153381	LSaCV-40-LSGA-102269-13	KP153489	PK5510	GQ404847		
AHEaCV-25	KM874356	DfCirV	JX185415	LSaCM-10-LSGA-2013	KP153382	LSaCV-52-LSCO-2013	KP153490	PK6197	GQ404848		
AHEaCV-25	KM874355	DfCyClV	JX185418	LSaCM-10-LSGA-102070-13	KP153383	LSaCV-52-LSSO-2013	KP153491	PKbeef23	HQ738634		
AHEaCV-26	KM874359	DfCyV-1	HQ638049	LSaCM-11-LSWA-2013	KP153384	LSaCV-52-LSGA-2013	KP153492	PKgoat11	HQ738636		
AHEaCV-27	KM874361	DfCyV-2	JX185423	LSaCM-11-LSWO-2013	KP153385	LSaCV-52-LSLA-2013	KP153493	PKgoat21	HQ738635		
AHEaCV-27	KM874360	DfCyV-3	JX185424	LSaCM-11-LSGA-2013	KP153386	LSaCV-41-LSCO-2013	KP153494	PNYDV	GU553134		
AHEaCV-28	KM874365	DfCyV-3	JX185428	LSaCM-11-LSMU-2013	KP153387	LSaCV-41-LSSO-2013	KP153495	pocircolike21	JF713716		
AHEaCV-28	KM874364	DfCyV-4	JX185425	LSaCM-11-LSSO-2013	KP153388	LSaCV-42-LSCO-2013	KP153496	pocircolike22	JF713717		
AHEaCV-28	KM874363	DfCyV-4	KC512917	LSaCM-11-LSCO-2013	KP153389	LSaCV-42-LSWO-2013	KP153497	pocircolike41	JF713718		
AHEaCV-28	KM874362	DFCyV-5	JX185426	LSaCV-1-LSCO-2013	KP153390	LSaCV-42-LSLA-2013	KP153498	pocircolike51	JF713719		
AHEaCV-29	KM874368	DFCyV-5	JX185427	LSaCV-1-LSGA-2013	KP153391	LSaCV-43-LSCO-2013	KP153499	PoSCV 2	KC545226		
AHEaCV-29	KM874367	DfCyV-6	KC512918	LSaCV-1-LSWO-2013	KP153392	LSaCV-44-LSCO-2013	KP153500	PoSCV 33L7	KC545227		
AHEaCV-29	KM874366	DfCyV-7	KC512919	LSaCV-1_LSWA-2013	KP153393	LSaCV-45-LSCO-2013	KP153501	PoSCV 34L13	KC545228		
AHEaCV-3	KM874297	DfCyV-8	KC512920	LSaCV-2-LSGA-2013	KP153394	LSaCV-46-LSSO-2013	KP153502	PoSCV 34L5	KC545229		
AHEaCV-3	KM874296	DflaCV-1	KF738873	LSaCV-3-LSWO-2013	KP153395	LSaCV-46-LSCO-2013	KP153503	PoSCV 3L2T	KC5452309		
AHEaCV-3	KM874295	DflaCV-10	KF738884	LSaCV-3-LSGA-2013	KP153396	LSaCV-47-LSCO-2013	KP153504	PoSCV-1 DP2	KJ577810		
AHEaCV-4	KM874300	DflaCV-10a	KF738885	LSaCV-4-LSWO-2013	KP153397	LSaCV-48-LSCO-2013	KP153505	PoSCV-1 DP3	KJ577811		
AHEaCV-4	KM874299	DflaCV-2	KF738874	LSaCV-4-LSSO-2013	KP153398	LSaCV-48-LSMU-2013	KP153506	PoSCV-2 TP3	KJ577818		
AHEaCV-4	KM874298	DflaCV-3	KF738875	LSaCV-4-LSWA-2013	KP153399	LSaCV-49-LSGA-2013	KP153507	PoSCV-6 XP1	KJ577819		
AHEaCV-5	KM874303	DflaCV-3a	KF738876	LSaCV-4-LSMU-2013	KP153400	LSaCV-49-LSSO-2013	KP153508	PoSCV-7 EP2-A	KJ577812		
AHEaCV-5	KM874302	DflaCV-4	KF738877	LSaCV-4-LSLA-2013	KP153401	LSaCV-49-LSCO-2013	KP153509	PoSCV-7 EP2-B	KJ577813		
AHEaCV-5	KM874301	DflaCV-5	KF738878	LSaCV-5-LSSO-2013	KP153402	LSaCV-49-LSWA-2013	KP153510	PoSCV-7 EP3-C	KJ577814		
AHEaCV-6	KM874308	DflaCV-5a	KF738879	LSaCV-6-LSSO-2013	KP153403	LSaCV-50-LSCO-2013	KP153511	PoSCV-7 EP3-D	KJ577815		
AHEaCV-6	KM874306	DflaCV-6	KF738880	LSaCV-7-LSSO-2013	KP153404	LSaCV-50-LSGA-2013	KP153512	PoSCV-8 GP2	KJ577817		
AHEaCV-6	KM874305	DflaCV-7	KF738881	LSaCV-8-LSCO-2013	KP153405	LSaCV-50-LSMU-2013	KP153513	PoSCV-9 FP1	KJ577816		
AHEaCV-6	KM874304	DflaCV-8	KF738882	LSaCV-8-LSSO-2013	KP153406	LSaCV-50-LSWO-2013	KP153514	PoSCV-Kor J481	KF193403		
AHEaCV-6	KM874307	DflaCV-9	KF738883	LSaCV-9-LSSO-2013	KP153407	LSaCV-50-LSLA-2013	KP153515	Propionibacterium	EFS79573		
AHEaCV-7	KM874309	DfOrV	JX185416	LSaCV-10-LSSO-2013	KP153408	DflaCV-10-LSCO-2013	KP153516	PSMV	JF905486		
AHEaCV-8	KM874314	DfOrV	JX185417	LSaCV-11-LSMU-2013	KP153409	DflaCV-10-LSWA-2013	KP153517	RaCV	DQ146997		
AHEaCV-8	KM874313	Diporeia sp CV LM28925	KC248425	LSaCV-12-LSCO-2013	KP153410	DflaCV-10-LSSO-2013	KP153518	RhFeCV	JQ814849		
AHEaCV-8	KM874311	Diporeia sp CV LM3487	KC248416	LSaCV-12-LSMU-2013	KP153411	DflaCV-10-LSMU-2013	KP153519	RodSCV M 13	JF755410		
AHEaCV-8	KM874310	DoYMV	AM157413	LSaCV-13-LSMU-2013	KP153412	DflaCV-10-LSLA-2013	KP153520	RodSCV M 44	JF755408		
AHEaCV-8	KM874312	DSV	M23022	LSaCV-13-LSWO-2013	KP153413	DflaCV-10-LSGA-2013	KP153521	RodSCV M 45	JF755409		
AHEaCV-9	KM874318	DuCV	AY228555	LSaCV-14-LSMU-2013	KP153414	LSaCV-51-LSMU-2013	KP153522	RodSCV M 53	JF755415		
AHEaCV-9	KM874317	DuCV	DQ100076	LSaCV-14-LSCO-2013	KP153415	LSaCV-27-LSSO-2013	KP153523	RodSCV M 89	JF755402		
AHEaCV-9	KM874316	EcmlV	HF921477	LSaCV-14-LSGA-2013	KP153416	DflaCV-5-LSWA-2013	KP153524	RodSCV R 15	JF755401		
AHEaCV-9	KM874315	ECSV	FJ66563	LSaCV-15-LSWO-2013	KP153417	DflaCV-5-LSWO-2013	KP153525	RodSCV V 64	JF755407		
AnCFV	KJ938716	ECSV	FJ66563	LSaCV-15-LSGA-2013	KP153418	DflaCV-8-LSSO-2013	KP153526	RodSCV V 69	JF755403		
AnCFV	KJ938716	EMSV	JF508490	LSaCV-15-LSMU-2013	KP153419	DflaCV-6-LSMU-2013	KP153527	RodSCV V 72	JF755411		
AtCopCV	JQ837277	ESV	EU244915	LSaCV-16-LSMU-122865-13	KP153420	AHEaCV-8-LSMU-2013	KP153528	RodSCV V 76	JF755404		
AYVSalpha	AJ416153	FaGmCV-10	KF371632	LSaCV-16-LSWA-2013	KP153421	MaMPRV	AY044133	RodSCV V 77	JF755405		

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.

Anderson, R. C. (2009). Do dragonflies migrate across the western Indian Ocean? *Journal of tropical ecology* **25**, 347-358.

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., Chan, A. M., Haynes, M., Kelley, S. & other authors (2006). The marine viromes of four oceanic regions. *PLoS biology* **4**, e368.

Anisimova, M. & Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic biology* **55**, 539-552.

Breitbart, M. & Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? *Trends in microbiology* **13**, 278-284.

Casado, C., Javier, O. G., Padron, E., Bean, S., McKenna, R., Agbandje-McKenna, M. & Boulton, M. (2004). Isolation and characterization of subgenomic DNAs encapsidated in" single" T= 1 isometric particles of Maize streak virus. *Virology* **323**, 164-171.

Corbet, P. & Brooks, S. (2008). *Dragonflies*, vol. 106. Harpercollins Pub Ltd.

Dayaram, A., Galatowitsch, M., Harding, J. S., Argüello-Astorga, G. R. & Varsani, A. (2014a). Novel circular DNA viruses identifiedin *Procordulia grayi* and *Xanthocnemis zealandica* larvae using metagenomic approaches. *Infection, Genetics and Evolution*, 134-141.

Dayaram, A., Goldstien, S., Zawar-Reza, P., Gomez, C., Harding, J. S. & Varsani, A. (2013a). Identification of starling circovirus in an estuarine mollusc (*Amphibola crenata*) in New Zealand using metagenomic approaches. *Genome announcements* **1**, e00278-00213.

Dayaram, A., Goldstien, S., Zawar-Reza, P., Gomez, C., Harding, J. S. & Varsani, A. (2013b). Novel single stranded DNA virus recovered from estuarine Mollusc (Amphibola crenata) whose replication associated protein (Rep) shares similarities with Rep-like sequences of bacterial origin. *Journal of General Virology* **94**, 1104-1110.

Dayaram, A., Potter, K., Pailes, R., Moline, A. B., Marinov, M., Rosenstein, D. D. & Varsani, A. (2014b). Identification of diverse circular Rep-encoding DNA viruses in dragonflies and damselflies of Arizona and Oklahoma. . *In Press*.

Dayaram, A., Goldstien, S., Argüello-Astorga, G. R., Zawar-Reza, P., Gomez, C., Harding, J. S. & Varsani, A. (2015). Diverse circular Rep encoding ssDNA viruses circulating amongst molluscs at the Avon-Heathcote estuary in Christchurch, New Zealand. *Infection, Genetics and Evolution*, 1-18.

Dayaram, A., Opong, A., Jäschke, A., Hadfield, J., Baschiera, M., Dobson, R. C. J., Offei, S. K., Shepherd, D. N., Martin, D. P. & other authors (2012). Molecular characterisation of a novel cassava associated circular ssDNA virus. *Virus Research* **166**, 130-135.

Dayaram, A., Potter, K. A., Moline, A. B., Rosenstein, D. D., Marinov, M., Thomas, J. E., Beitbart, M., Rosario, K., Argüello-Astorga, G. R. & other authors (2013c). High global diversity of cycloviruses amongst dragonflies. *Journal of General Virology* **94**, 1827-1840.

Dunlap, D. S., Ng, T. F. F., Rosario, K., Barbosa, J. G., Greco, A. M., Breitbart, M. & Hewson, I. (2013). Molecular and microscopic evidence of viruses in marine copepods. *Proceedings of the National Academy of Sciences* **110**, 1375-1380.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797.

Gorbalenya, A. E., Koonin, E. V. & Wolf, Y. I. (1990). A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses. *FEBS letters* **262**, 145-148.

Hadfield, J., Thomas, J., Schwinghamer, M., Kraberger, S., Stainton, D., Dayaram, A., Parry, J., Pande, D., Martin, D. & other authors (2012). Molecular characterisation of dicot-infecting mastreviruses from Australia. *Virus Research* **166**, 13-22.

Hewson, I., Eaglesham, J. B., Höök, T. O., LaBarre, B. A., Sepúlveda, M. S., Thompson, P. D., Watkins, J. M. & Rudstam, L. G. (2013). Investigation of viruses in *Diporeia* spp. from the Laurentian Great Lakes and Owasco Lake as potential stressors of declining populations. *Journal of Great Lakes Research* **39**, 499-506.

Idris, A., Al-Saleh, M., Piatek, M. J., Al-Shahwan, I., Ali, S. & Brown, J. K. (2014). Viral Metagenomics: Analysis of Begomoviruses by Illumina High-Throughput Sequencing. *Viruses* **6**, 1219-1236.

Jeske, H., Lütgemeier, M. & Preiß, W. (2001). DNA forms indicate rolling circle and recombination-dependent replication of Abutilon mosaic virus. *The EMBO journal* **20**, 6158-6167.

Katoh, K. & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772-780.

Kim, K.-H., Chang, H.-W., Nam, Y.-D., Roh, S. W., Kim, M.-S., Sung, Y., Jeon, C. O., Oh, H.-M. & Bae, J.-W. (2008). Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Applied and environmental microbiology* **74**, 5975-5985.

Koonin, E. V., Senkevich, T. G. & Dolja, V. V. (2006). The ancient Virus World and evolution of cells. *Biol Direct* **1**, 29.

Kraberger, S., Stainton, D., Dayaram, A., Zawar-Reza, P., Gomez, C., Harding, J. S. & Varsani, A. (2013). Discovery of Sclerotinia sclerotiorum hypovirulence-associated virus-1 in urban river sediments of Heathcote and Styx rivers in Christchurch city, New Zealand. *Genome announcements* **1**, e00559-00513.

Kraberger, S., Argüello-Astorga, G. R., Greenfield, G. L., Galilee, C., Law, D., Martin, D. P. & Varsani, A. (2014). Characterisation of a diverse range of Rep-encoding ssDNA viruses recovered from a sewage treatment oxidation pond. *Infection, Genetics and Evolution* **In Review**.

Labonté, J. M. & Suttle, C. A. (2013). Previously unknown and highly divergent ssDNA viruses populate the oceans. *The ISME journal* **7**, 2169-2177.

Londoño, A., Riego-Ruiz, L. & Argüello-Astorga, G. R. (2010). DNA-binding specificity determinants of replication proteins encoded by eukaryotic ssDNA viruses are adjacent to widely separated RCR conserved motifs. *Archives of virology* **155**, 1033-1046.

López-Bueno, A., Tamames, J., Velázquez, D., Moya, A., Quesada, A. & Alcamí, A. (2009). High diversity of the viral community from an Antarctic lake. *Science* **326**, 858-861.

Maddison, W. & Maddison, D. (2011). Mesquite 2.75: a modular system for evolutionary analysis.

McDaniel, L. D., Rosario, K., Breitbart, M. & Paul, J. H. (2014). Comparative metagenomics: Natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. *Environmental microbiology* **16**, 570-585.

Mubin, M., Shahid, M., Tahir, M., Briddon, R. & Mansoor, S. (2010). Characterization of begomovirus components from a weed suggests that begomoviruses may associate with multiple distinct DNA satellites. *Virus genes* **40**, 452-457.

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

Nawaz-ul-Rehman, M. S., Nahid, N., Mansoor, S., Briddon, R. W. & Fauquet, C. M. (2010). Post-transcriptional gene silencing suppressor activity of two non-pathogenic alphasatellites associated with a begomovirus. *Virology* **405**, 300-308.

Ng, T. F. F., Alavandi, S., Varsani, A., Burghart, S. & Breitbart, M. (2013). Metagenomic identification of a nodavirus and a circular ssDNA virus in semi-purified viral nucleic acids from the hepatopancreas of healthy *Farfantepenaeus duorarum* shrimp. *Dis Aquat Org* **105**, 237-242.

Ng, T. F. F., Duffy, S., Polston, J. E., Bixby, E., Vallad, G. E. & Breitbart, M. (2011a). Exploring the Diversity of Plant DNA Viruses and Their Satellites Using Vector-Enabled Metagenomics on Whiteflies. *PloS one* **6**, e19050.

Ng, T. F. F., Willner, D. L., Lim, Y. W., Schmieder, R., Chau, B., Nilsson, C., Anthony, S., Ruan, Y., Rohwer, F. & other authors (2011b). Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. *PloS one* **6**, e20579.

Ng, T. F. F., Chen, L.-F., Zhou, Y., Shapiro, B., Stiller, M., Heintzman, P. D., Varsani, A., Kondov, N. O., Wong, W. & other authors (2014). Preservation of viral genomes in 700-y-old caribou feces from a subarctic ice patch. *Proceedings of the National Academy of Sciences*, 201410429.

Padilla-Rodriguez, M., Rosario, K. & Breitbart, M. (2013). Novel cyclovirus discovered in the Florida woods cockroach *Eurycotis floridana* (Walker). *Archives of virology* **158**, 1389-1392.

Phan, T. G., Kapusinszky, B., Wang, C., Rose, R. K., Lipton, H. L. & Delwart, E. L. (2011). The Fecal Viral Flora of Wild Rodents. *PLoS pathogens* **7**, e1002218.

Price, M. N., Dehal, P. S. & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one* **5**, e9490.

Romay, G., Chirinos, D., Geraud-Pouey, F. & Desbiez, C. (2010). Association of an atypical alphasatellite with a bipartite New World begomovirus. *Archives of virology* **155**, 1843-1847.

Rosario, K., Duffy, S. & Breitbart, M. (2009a). Diverse circovirus-like genome architectures revealed by environmental metagenomics. *Journal of General Virology* **90**, 2418-2424.

Rosario, K., Duffy, S. & Breitbart, M. (2012a). A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Archives of virology* **157**, 1851-1871.

Rosario, K., Nilsson, C., Lim, Y. W., Ruan, Y. & Breitbart, M. (2009b). Metagenomic analysis of viruses in reclaimed water. *Environmental microbiology* **11**, 2806-2820.

Rosario, K., Capobianco, H., Ng, T. F. F., Breitbart, M. & Polston, J. E. (2014). RNA Viral Metagenome of Whiteflies Leads to the Discovery and Characterization of a Whitefly-Transmitted Carlavirus in North America. *PloS one* **9**, e86748.

Rosario, K., Padilla-Rodriguez, M., Kraberger, S., Stainton, D., Martin, D. P., Breitbart, M. & Varsani, A. (2013). Discovery of a novel mastrevirus and alphasatellite-like circular DNA in dragonflies (Epiprocta) from Puerto Rico. *Virus Research*, 231-237.

Rosario, K., Dayaram, A., Marinov, M., Ware, J., Kraberger, S., Stainton, D., Breitbart, M. & Varsani, A. (2012b). Diverse circular ssDNA viruses discovered in dragonflies (Odonata: Epiprocta). *Journal of General Virology* **93**, 2668-2681.

Rosario, K., Marinov, M., Stainton, D., Kraberger, S., Wiltshire, E. J., Collings, D. A., Walters, M., Martin, D. P., Breitbart, M. & other authors (2011). Dragonfly cyclovirus, a novel single-stranded DNA virus discovered in dragonflies (Odonata: Anisoptera). *Journal of General Virology* **92**, 1302-1308.

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., Colombet, J., Sime-Ngando, T. & Debroas, D. (2012). Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PloS one* **7**, e33641.

Rozenblatt-Rosen, O., Deo, R. C., Padi, M., Adelmant, G., Calderwood, M. A., Rolland, T., Grace, M., Dricot, A., Askenazi, M. & other authors (2012). Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins. *Nature* **487**, 491-495.

Seguin, J., Rajeswaran, R., Malpica-Lopez, N., Martin, R. R., Kasschau, K., Dolja, V. V., Otten, P., Farinelli, L. & Pooggin, M. M. (2014). De novo reconstruction of consensus master genomes of plant RNA and DNA viruses from siRNAs. *PloS one* **9**, e88513.

Sikorski, A., Kearvell, J., Elkington, S., Dayaram, A., Argüello-Astorga, G. R. & Varsani, A. (2013a). Novel ssDNA viruses discovered in yellow-crowned parakeet (*Cyanoramphus auriceps*) nesting material. *Archives of virology*, 1603-1607.

Sikorski, A., Massaro, M., Kraberger, S., Young, L. M., Smalley, D., Martin, D. P. & Varsani, A. (2013b). Novel myco-like DNA viruses discovered in the faecal matter of various animals. *Virus Research* **177**, 209-216.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. & Birol, İ. (2009). ABySS: a parallel assembler for short read sequence data. *Genome research* **19**, 1117-1123.

Smith, R. J., Jeffries, T. C., Roudnew, B., Seymour, J. R., Fitch, A. J., Simons, K. L., Speck, P. G., Newton, K., Brown, M. H. & other authors (2013). Confined aquifers as viral reservoirs. *Environmental microbiology reports* **5**, 725-730.

Steinfeldt, T., Finsterbusch, T. & Mankertz, A. (2006). Demonstration of nicking/joining activity at the origin of DNA replication associated with the rep and rep′proteins of porcine circovirus type 1. *Journal of virology* **80**, 6225-6234.

Suttle, C. A. (2007). Marine viruses—major players in the global ecosystem. *Nature Reviews Microbiology* **5**, 801-812.

van der Walt, E., Rybicki, E. P., Varsani, A., Polston, J., Billharz, R., Donaldson, L., Monjane, A. L. & Martin, D. P. (2009). Rapid host adaptation by extensive recombination. *Journal of general virology* **90**, 734-746.

Walker, J. E., Saraste, M., Runswick, M. J. & Gay, N. J. (1982). 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. *The EMBO journal* **1**, 945.

Wright, E. A., Heckel, T., Groenendijk, J., Davies, J. W. & Boulton, M. I. (1997). Splicing features in *maize streak virus* virion and complementary sense gene expression. *The Plant Journal* **12**, 1285-1297.

Yoon, H. S., Price, D. C., Stepanauskas, R., Rajah, V. D., Sieracki, M. E., Wilson, W. H., Yang, E. C., Duffy, S. & Bhattacharya, D. (2011). Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* **332**, 714-717.

Yoshida, M., Takaki, Y., Eitoku, M., Nunoura, T. & Takai, K. (2013). Metagenomic Analysis of Viral Communities in (Hado) Pelagic Sediments. *PloS one* **8**, e57271.

Yu, X., Li, B., Fu, Y., Jiang, D., Ghabrial, S. A., Li, G., Peng, Y., Xie, J., Cheng, J. & other authors (2010). A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. *Proceedings of the National Academy of Sciences* **107**, 8387.

Chapter 8

Discussion and future research

Contents

8.1 Overview

Over the five years there has been a significant increase in the use of next generation sequencing (NGS) technologies due to the decreasing costs and establishment of core NGS sequencing facilities and private service providers. These sequencing methods (reviewed in Chapter 1) have dramatically changed the field of virology, with viruses that were previously unable to be cultured under traditional laboratory setting are now easily identified without prior knowledge. These sequencing technologies can be used to identify the virome of an ecosystem. This in turn had led to a vast increase in the number of circular replication-associated protein (Rep) encoding single-stranded (CRESS) DNA viruses being discovered from a range of different samples sourced globally. It is clear from many of these different studies that there is a huge diversity of CRESS DNA viruses (Angly *et al.*, 2006; Breitbart *et al.*, 2003; Breitbart *et al.*, 2002; Delwart, 2007; Fancello *et al.*, 2013; Kim *et al.*, 2008; Ng *et al.*, 2012; Rosario & Breitbart, 2011; Rosario *et al.*, 2009; Roux *et al.*, 2012; Whon *et al.*, 2012; Yoshida *et al.*, 2013). It is also apparent that specific sample types such as human faecal matter (Li *et al.*, 2010), invertebrates (Ng *et al.*, 2013; Padilla-Rodriguez *et al.*, 2013; Rosario *et al.*, 2012b), animals tissue (Li *et al.*, 2013) and animal faeces (Sikorski *et al.*, 2013; van den Brand *et al.*, 2011) harbour diverse CRESS DNA viruses; with many of these samples types having the potential to be used as surveillance tools for the monitoring of CRESS DNA viruses in different environments. Overall, viral metagenomics techniques have been instrumental towards understanding CRESS DNA viral ecology and have enabled rapid advances in this field with the numerous amounts of baseline data being produced that will benefit many disciplines.

The aim of this PhD research was to examine CRESS DNA viral diversity in both terrestrial and aquatic ecosystems. The main objective of my research was to generate baseline data of the CRESS DNA viruses in top end insect predators such as dragonflies and damselflies as well as natural concentrators such as different species of molluscs. Understanding the diversity of viruses present in these organisms will enable us to develop new molecular tools for surveillance in order to monitor these viruses in ecosystems. Finally, accessing CRESS DNA viral diversity throughout different samples in the same ecosystem will give a better idea of the diversity these organisms harbour, whilst providing insight into viral dynamics such as possible distribution and flow of these viruses in an environment.

8.2 Major findings

Table 8.1: Summary of all CRESS DNA viral genomes recovered from this research. Includes: chapter, viral isolate, GenBank accession number and isolation source.

Chapter 2	
Viral isolates	Isolation source
DfCyV-4 [KC512917]	*Erythemis simplicicollis*
DfCyV-4 [KC512916]	*Aeshna multicolour*
DfCyV-6 [KC512918]	*Aeshna multicolour*
DfCyV-7 [KC512919]	*Xanthocnemis zealandica*
DfCyV-8 [KC512920]	*Orthetrum Sabina*

Chapter 3	
Viral isolates	Isolation source
DflaCV-1 [KF738873]	*Procordulia grayi*
DflaCV-2 [KF738874]	*Procordulia grayi*
DflaCV-3 [KF738875]	*Procordulia grayi*
DflaCV-3 [KF738876]	*Procordulia grayi*
DflaCV-4 [KF738877]	*Procordulia grayi*
DflaCV-5 [KF738878]	*Procordulia grayi*
DflaCV-5 [KF738879]	*Procordulia grayi*
DflaCV-6 [KF738880]	*Procordulia grayi*
DflaCV-7 [KF738881]	*Procordulia grayi*
DflaCV-8 [KF738882]	*Procordulia grayi*
DflaCV-9 [KF738883]	*Procordulia grayi*
DflaCV-10 [KF738884]	*Xanthocnemis zealandica*
DflaCV-10 [KF738885]	*Xanthocnemis zealandica*

Chapter 4	
Viral isolates	Isolation source
SsHADV-1 [KM598383]	*Erythemis simplicicollis*
SsHADV-1 [KM598382]	*Ischnura ramburii*
SsHADV-1 [KM598384]	*Pantala hymenaea*
DfasCV-4 [KM598385]	*Ischnura posita*
DfasCV-4 [KM598386]	*Pantala hymenaea*
DfasCV-5 [KM598388]	*Libellula saturata*
DfasCV-5 [KM598387]	*Rhionaeschna multicolor*
OdasCV-1 [KM598393]	*Ischnura ramburii*
OdasCV-2 [KM598399]	*Ischnura posita*
OdasCV-3 [KM598407]	*Ischnura ramburii*
OdasCV-4 [KM598408]	*Ischnura posita*
OdasCV-5 [KM598410]	*Libellula quadrimaculata*
OdasCV-6 [KM598410]	*Libellula quadrimaculata*
OdasCV-7 [KM598410]	*Libellula quadrimaculata*
OdasCV-8 [KM598391]	*Libellula quadrimaculata*
OdasCV-9 [KM598392]	*Erythemis simplicicollis*
OdasCV-10 [KM598412]	*Libellula quadrimaculata*
OdasCV-11 [KM598394]	*Erythemis simplicicollis*
OdasCV-12 [KM598395]	*Libellula quadrimaculata*
OdasCV-13 [KM598396]	*Libellula quadrimaculata*
OdasCV-14 [KM598397]	*Erythrodiplax fusca*
OdasCV-15 [KM598398]	*Libellula quadrimaculata*
OdasCV-16 [KM598411]	*Libellula quadrimaculata*
OdasCV-17 [KM598400]	*Libellula quadrimaculata*
OdasCV-18 [KM598401]	*Libellula quadrimaculata*
OdasCV-18 [KM598402]	*Libellula quadrimaculata*
OdasCV-18 [KM598403]	*Libellula saturata*
OdasCV-19 [KM598404]	*Libellula quadrimaculata*
OdasCV-19 [KM598405]	*Pachydiplax longipennis*
OdasCV-20 [KM598406]	*Erythemis simplicicollis*
OdasCV-21 [KM598409]	*Erythrodiplax fusca*

Chapter 5	
Viral isolates	Isolation source
GaCSV [KC172652]	*Amphibola crenata*

Chapter 6	
Viral isolates	Isolation source
AHEaCV-1 [KM874290]	*Austrolvenus stutchburyi*
AHEaCV-1 [KM874291]	*Austrolvenus stutchburyi*
AHEaCV-2 [KM874292]	*Austrolvenus stutchburyi*
AHEaCV-2 [KM874293]	*Austrolvenus stutchburyi*
AHEaCV-2 [KM874294]	*Austrolvenus stutchburyi*
AHEaCV-3 [KM874295]	*Austrolvenus stutchburyi*
AHEaCV-3 [KM874296]	*Austrolvenus stutchburyi*
AHEaCV-3 [KM874297]	*Amphibola crenata*
AHEaCV-4 [KM874298]	*Austrolvenus stutchburyi*
AHEaCV-4 [KM874299]	*Austrolvenus stutchburyi*
AHEaCV-4 [KM874300]	*Austrolvenus stutchburyi*
AHEaCV-5 [KM874301]	*Austrolvenus stutchburyi*
AHEaCV-5 [KM874302]	*Austrolvenus stutchburyi*

Chapter 7	
Viral isolates	Isolation source
LSaCV-1 [KP153390]	*Musculium novazelandiae*
LSaCV-1 [KP153391]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-1 [KP153392]	*Chironomus zealandicus*
LSaCV-1 [KP153392]	Water
LSaCV-2 [KP153394]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-3 [KP153395]	*Chironomus zealandicus*
LSaCV-3 [KP153396]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-4 [KP153397]	*Chironomus zealandicus*
LSaCV-4 [KP153398]	Sediment
LSaCV-4 [KP153399]	Water
LSaCV-4 [KP153400]	*Echyridella menziesi*
LSaCV-4 [KP153401]	*Procordulia grayi, Xanthocnemis zealandica*
LSaCV-5 [KP153402]	Sediment
LSaCV-6 [KP153403]	Sediment
LSaCV-7 [KP153404]	Sediment
LSaCV-8 [KP153405]	*Musculium novazelandiae*
LSaCV-8 [KP153406]	Sediment
LSaCV-9 [KP153407]	Sediment
LSaCV-10 [KP153408]	Sediment
LSaCV-11 [KP153409]	*Echyridella menziesi*
LSaCV-12 [KP153410]	*Musculium novazelandiae*
LSaCV-12 [KP153411]	*Echyridella menziesi*
LSaCV-13 [KP153412]	*Echyridella menziesi*
LSaCV-13 [KP153413]	*Chironomus zealandicus*
LSaCV-14 [KP153414]	*Echyridella menziesi*
LSaCV-14 [KP153415]	*Musculium novazelandiae*
LSaCV-14 [KP153416]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-15 [KP153417]	*Chironomus zealandicus*
LSaCV-15 [KP153418]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-15 [KP153419]	*Echyridella menziesi*
LSaCV-16 [KP153420]	*Echyridella menziesi*
LSaCV-16 [KP153421]	*Echyridella menziesi*
LSaCV-16 [KP153422]	Water
LSaCV-16 [KP153423]	*Chironomus zealandicus*
LSaCV-17 [KP153424]	*Echyridella menziesi*
LSaCV-17 [KP153425]	*Musculium novazelandiae*
LSaCV-17 [KP153426]	*Chironomus zealandicus*
LSaCV-18 [KP153427]	*Echyridella menziesi*
LSaCV-18 [KP153428]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-19 [KP153429]	*Echyridella menziesi*
LSaCV-19 [KP153430]	Sediment
LSaCV-19 [KP153431]	*Musculium novazelandiae*
LSaCV-19 [KP153432]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-19 [KP153433]	*Procordulia grayi, Xanthocnemis zealandica*
LSaCV-20 [KP153434]	*Echyridella menziesi*
LSaCV-20 [KP153435]	*Chironomus zealandicus*
LSaCV-20 [KP153436]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-21 [KP153437]	*Echyridella menziesi*
LSaCV-21 [KP153438]	*Chironomus zealandicus*
LSaCV-21 [KP153439]	Sediment
LSaCV-21 [KP153440]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-22 [KP153441]	*Echyridella menziesi*
LSaCV-23 [KP153442]	*Echyridella menziesi*
LSaCV-24 [KP153443]	*Echyridella menziesi*
LSaCV-24 [KP153444]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-25 [KP153445]	*Echyridella menziesi*
LSaCV-26 [KP153446]	*Echyridella menziesi*
DflaCV-3 [KP153447]	*Procordulia grayi, Xanthocnemis zealandica*
DflaCV-3 [KP153448]	*Potamopyrgus antipodarum, Physella acuta*
DflaCV-3 [KP153449]	*Echyridella menziesi*
DflaCV-3 [KP153450]	*Chironomus zealandicus*
LSaCV-28 [KP153451]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-28 [KP153452]	*Musculium novazelandiae*
LSaCV-28 [KP153453]	*Procordulia grayi, Xanthocnemis zealandica*
LSaCV-29 [KP153454]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-29 [KP153455]	*Musculium novazelandiae*
LSaCV-30 [KP153456]	Sediment
LSaCV-30 [KP153457]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-30 [KP153458]	*Musculium novazelandiae*
LSaCV-31 [KP153459]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-31 [KP153460]	Sediment

AHEaCV-5 [KM874303]	*Amphibola crenata*
AHEaCV-6 [KM874304]	*Austrolvenus stutchburyi*
AHEaCV-6 [KM874305]	*Austrolvenus stutchburyi*
AHEaCV-6 [KM874306]	*Amphibola crenata*
AHEaCV-6 [KM874307]	Benthic sediment
AHEaCV-6 [KM874308]	*Paphies subtriangulata*
AHEaCV-7 [KM874309]	*Austrolvenus stutchburyi*
AHEaCV-8 [KM874310]	*Austrolvenus stutchburyi*
AHEaCV-8 [KM874311]	*Austrolvenus stutchburyi*
AHEaCV-8 [KM874312]	*Paphies subtriangulata*
AHEaCV-8 [KM874313]	*Amphibola crenata*
AHEaCV-8 [KM874314]	Benthic sediment
AHEaCV-9 [KM874315]	*Austrolvenus stutchburyi*
AHEaCV-9 [KM874316]	*Austrolvenus stutchburyi*
AHEaCV-9 [KM874317]	*Amphibola crenata*
AHEaCV-9 [KM874318]	Benthic sediment
AHEaCV-10 [KM874319]	*Austrolvenus stutchburyi*
AHEaCV-10 [KM874320]	*Austrolvenus stutchburyi*
AHEaCV-10 [KM874321]	Benthic sediment
AHEaCV-10 [KM874322]	*Amphibola crenata*
AHEaCV-11 [KM874323]	*Austrolvenus stutchburyi*
AHEaCV-11 [KM874324]	*Austrolvenus stutchburyi*
AHEaCV-11 [KM874325]	*Austrolvenus stutchburyi*
AHEaCV-11 [KM874326]	Benthic sediment
AHEaCV-11 [KM874327]	*Paphies subtriangulata*
AHEaCV-12 [KM874328]	*Austrolvenus stutchburyi*
AHEaCV-13 [KM874329]	*Austrolvenus stutchburyi*
AHEaCV-13 [KM874330]	*Amphibola crenata*
AHEaCV-13 [KM874331]	Benthic sediment
AHEaCV-14 [KM874332]	*Austrolvenus stutchburyi*
AHEaCV-14 [KM874333]	*Amphibola crenata*
AHEaCV-14 [KM874334]	Benthic sediment
AHEaCV-14 [KM874335]	*Paphies subtriangulata*
AHEaCV-15 [KM874336]	*Austrolvenus stutchburyi*
AHEaCV-15 [KM874337]	*Austrolvenus stutchburyi*
AHEaCV-15 [KM874338]	Benthic sediment
AHEaCV-15 [KM874339]	*Paphies subtriangulata*
AHEaCV-16 [KM874340]	*Austrolvenus stutchburyi*
AHEaCV-16 [KM874341]	*Austrolvenus stutchburyi*
AHEaCV-16 [KM874342]	Benthic sediment
AHEaCV-17 [KM874343]	*Austrolvenus stutchburyi*
AHEaCV-17 [KM874344]	*Amphibola crenata*
AHEaCV-17 [KM874345]	*Paphies subtriangulata*
AHEaCV-18 [KM874346]	*Amphibola crenata*
AHEaCV-19 [KM874347]	*Amphibola crenata*
AHEaCV-20 [KM874348]	*Amphibola crenata*
AHEaCV-20 [KM874349]	*Paphies subtriangulata*
AHEaCV-21 [KM874350]	*Paphies subtriangulata*
AHEaCV-22 [KM874351]	Benthic sediment
AHEaCV-22 [KM874352]	*Paphies subtriangulata*
AHEaCV-23 [KM874353]	*Paphies subtriangulata*
AHEaCV-24 [KM874354]	*Paphies subtriangulata*
AHEaCV-25 [KM874355]	*Austrolvenus stutchburyi*
AHEaCV-25 [KM874356]	*Amphibola crenata*
AHEaCV-25 [KM874357]	Benthic sediment
AHEaCV-25 [KM874358]	*Paphies subtriangulata*
AHEaCV-26 [KM874359]	*Paphies subtriangulata*
AHEaCV-27 [KM874360]	*Austrolvenus stutchburyi*
AHEaCV-27 [KM874361]	*Paphies subtriangulata*
AHEaCV-28 [KM874362]	*Amphibola crenata*
AHEaCV-28 [KM874363]	Benthic sediment
AHEaCV-28 [KM874364]	*Paphies subtriangulata*
AHEaCV-28 [KM874365]	*Austrolvenus stutchburyi*
AHEaCV-29 [KM874366]	*Austrolvenus stutchburyi*
AHEaCV-29 [KM874367]	*Austrolvenus stutchburyi*
AHEaCV-29 [KM874368]	*Paphies subtriangulata*
LSaCV-31 [KP153461]	*Musculium novazelandiae*
LSaCV-31 [KP153462]	*Chironomus zealandicus*
LSaCV-31 [KP153463]	*Procordulia grayi, Xanthocnemis zealandica*
LSaCV-32 [KP153464]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-32 [KP153465]	Sediment
LSaCV-32 [KP153466]	*Musculium novazelandiae*
LSaCV-32 [KP153467]	*Echyridella menziesi*
LSaCV-32 [KP153468]	*Chironomus zealandicus*
LSaCV-33 [KP153469]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-34 [KP153471]	Sediment
LSaCV-35 [KP153472]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-35 [KP153473]	Water
LSaCV-36 [KP153474]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-36 [KP153475]	*Musculium novazelandiae*
LSaCV-37 [KP153476]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-37 [KP153477]	*Echyridella menziesi*
LSaCV-37 [KP153478]	*Procordulia grayi, Xanthocnemis zealandica*
LSaCV-37 [KP153479]	Sediment
LSaCV-37 [KP153480]	Water
LSaCV-37 [KP153481]	*Chironomus zealandicus*
LSaCV-37 [KP153482]	*Musculium novazelandiae*
LSaCV-38 [KP153483]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-38 [KP153484]	*Musculium novazelandiae*
LSaCV-39 [KP153485]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-39 [KP153486]	Sediment
LSaCV-39 [KP153487]	*Procordulia grayi, Xanthocnemis zealandica*
LSaCV-40 [KP153488]	*Musculium novazelandiae*
LSaCV-40 [KP153489]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-52 [KP153490]	*Musculium novazelandiae*
LSaCV-52 [KP153491]	Sediment
LSaCV-52 [KP153492]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-52 [KP153493]	*Procordulia grayi, Xanthocnemis zealandica*
LSaCV-41 [KP153494]	*Musculium novazelandiae*
LSaCV-41 [KP153495]	Sediment
LSaCV-42 [KP153496]	*Musculium novazelandiae*
LSaCV-42 [KP153497]	*Chironomus zealandicus*
LSaCV-42 [KP153498]	*Procordulia grayi, Xanthocnemis zealandica*
LSaCV-43 [KP153499]	*Musculium novazelandiae*
LSaCV-44 [KP153500]	*Musculium novazelandiae*
LSaCV-45 [KP153501]	*Musculium novazelandiae*
LSaCV-46 [KP153502]	Sediment
LSaCV-46 [KP153503]	*Musculium novazelandiae*
LSaCV-47 [KP153504]	*Musculium novazelandiae*
LSaCV-48 [KP153505]	*Musculium novazelandiae*
LSaCV-48 [KP153506]	*Echyridella menziesi*
LSaCV-49 [KP153507]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-49 [KP153508]	Sediment
LSaCV-49 [KP153509]	*Musculium novazelandiae*
LSaCV-49 [KP153510]	Water
LSaCV-50 [KP153511]	*Musculium novazelandiae*
LSaCV-50 [KP153512]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-50 [KP153513]	*Echyridella menziesi*
LSaCV-50 [KP153514]	*Chironomus zealandicus*
LSaCV-50 [KP153515]	*Procordulia grayi, Xanthocnemis zealandica*
DflaCV-10 [KP153516]	*Musculium novazelandiae*
DflaCV-10 [KP153517]	Water
DflaCV-10 [KP153518]	Sediment
DflaCV-10 [KP153519]	*Echyridella menziesi*
DflaCV-10 [KP153520]	*Procordulia grayi, Xanthocnemis zealandica*
DflaCV-10 [KP153521]	*Potamopyrgus antipodarum, Physella acuta*
LSaCV-51 [KP153522]	*Echyridella menziesi*
LSaCV-27 [KP153523]	Sediment
DflaCV-5 [KP153524]	Water
DflaCV-5 [KP153525]	*Chironomus zealandicus*
DflaCV-8 [KP153526]	*Sediment*
DflaCV-6 [KP153527]	*Echyridella menziesi*
AHEaCV-8 [KP153528]	*Echyridella menziesi*

8.2.1 General overview

As part of this PhD research 268 novel CRESS DNA viruses have been indentified in various sample types (adult Odonata, Odonata larvae, seven molluscs species, oligochaeta, water and benthic sediment) originating from three different countries (New Zealand, Australia and the United States of America) (Table 8.1). This has dramatically increased the number and diversity of CRESS DNA genomes currently available in the NCBI GenBank database.

Chapters 2, 3, 4 and 7 explore the CRESS DNA viral diversity in insect predators dragonflies and damselflies (Odonata), and further compare the difference in CRESS DNA viruses isolated from the larvae (aquatic) and the adults (terrestrial) in their given ecosystems. Of the eleven different Odonata species sampled in the United States of America, Australia, and New Zealand 36 CRESS DNA viral genomes were recovered (Chapters 2 and 4). The diversity of the recovered viral genomes from Odonata in Chapters 2 and four were highly diverse. Dragonfly cyclovirus genomes were recovered from *Orthetrum sabina, Xanthocnemis zealandica* and *Aeshna multicolour* samples described in Chapter 2. Twenty four CRESS DNA viruses were isolated from three species of dragonfly and damselfly larvae sampled in various lakes in the South Island of New Zealand, the first report of CRESS DNA viruses in dragonfly larvae. The CRESS DNA viruses recovered from the adult dragonfly and damselflies differed to those isolated from the larvae, indicating that the CRESS DNA viruses circulating in terrestrial and aquatic ecosystems are diverse. Overall, these chapters begin to demonstrate that previously identified DfLaCVs, SsHADV-1 and DfCyVs as well as novel CRESS DNA viruses can be identified easily using the NGS and can be further verified using PCR with sequence specific primers.

In Chapters 5, 6 and 7 CRESS DNA viral diversity in molluscs, which are known to be natural concentrators, are described. One hundred and forty nine CRESS DNA virus genomes were identified and recovered from seven species of molluscs and sediment samples, both collected from two sample sites in New Zealand. The first isolated CRESS DNA viruses in molluscs GasCV (Chapter 5) was subsequently identified in other molluscs sampled from the same location. Molluscs in general have proved to be an extremely useful sampling tool for viral surveillance. Chapters 6 and 7 start to report CRESS DNA viral diversity and distribution in ecosystems. In both chapters the work reveals the high diversity of CRESS DNA viruses in ecosystems. CRESS DNA viruses were isolated from other sample types such as benthic sediment (n=34; Chapter 6 and 7), Oligochaeta and midge larvae (n=16; Chapter 7) and water (n= 8; Chapter 7). This helped to show the viral distribution of CRESS DNA viruses across different sample types in the same ecosystem.

8.2.2 Viral metagenomics

Viral metagenomics and shotgun sequencing in combination with rolling circle amplification (RCA) and multiple displacement amplification (MDA) have proven to be important tools for determining CRESS DNA viral diversity in various organisms and ecosystems without any prior knowledge (Kraberger *et al.*, 2014; Rosario *et al.*, 2009, 2012a; Sikorski *et al.*, 2013). These techniques have led to the discovery of diverse CRESS DNA viruses, most of which do not fall within the currently recognised viral taxonomy, with the majority of sequences having no similarities to sequences currently available on the NCBI GenBank database.

This research, as part of this thesis, shows the development from traditional restriction enzyme digestion of RCA products to NGS pipelines to identify novel genomes in different samples types. In Chapters 2 and 3 CRESS DNA viral genomes were recovered using restriction enzyme digests, which rely on the virus having a specific nucleotide cleavage site for the enzyme to cleave the circular DNA. Only eleven out of the 268 viral genomes recovered in this study were isolated using these methods. In Chapters 3, 5, 6, and 7, NGS methods are used to recover CRESS DNA viruses from both environmental and invertebrate samples. The results from these chapters demonstrate the effectiveness of combining RCA with NGS to identify potential CRESS DNA viruses and then verifying the full genomes with sequence specific primers and PCR.

Metagenomics to examine CRESS DNA viral genomes in invertebrates was initially performed to investigate whether Odonata were feeding on smaller insects vectoring plant viruses. This concept of vector enabled metagenomics (VEM) has been applied in multiple studies to investigate virus diversity in different insect populations that could potentially be vectoring viruses (Ng *et al.*, 2011a; Rosario *et al.*, 2013). Chapters 2 and 4 further test the VEM approach by applying it to dragonflies and damselflies which are top insect predators in their ecosystems. The VEM approach was successfully used identify novel CRESS DNA viruses including SsHADV-1, a mycovirus shown to infect the fungus *Sclerotinia sclerotiorum*. VEM-based studies by Rosario *et al.* (2013) resulted in the identification of the first mastrevirus in the new world, recovered from dragonfly samples (*Erythrodiplax fusca*) from Puerto Rico.

Metagenomic approaches were used to investigate CRESS DNA viral diversity in natural concentrators such as molluscs and in environmental samples such as water and benthic sediment (Chapters 5, 6 and 7). As molluscs are natural concentrators of their environment and are often used as bio-indicators of contamination in the local aquatic environment (Baršienė *et al.*, 2002; van der Oost *et al.*, 1988), it is plausible that they may indirectly concentrate CRESS DNA viruses using the same mechanism. Over 149 CRESS DNA viral genomes were recovered from molluscs. The increase of known CRESS DNA viruses as a result of metagenomic sequencing will help answer questions about the viral biome of freshwater lakes and estuaries, and how they differ globally. Some of these the early studies on viral biome comparisons have suggested that viral types differ between locations (Angly *et al.*, 2006).

8.2.3 Subsequent CRESS DNA virus isolation in different ecosystems

Although metagenomics was used to identify potential CRESS DNA viruses in samples, all viral genomes were subsequently verified using specifically designed back-to-back primers (Chapters 2, 3, 4, 5, 6 and 7). The use of metagenomics in this study has shown that there are limitations of the Illumina sequencing data in the assembly process. Error rates increase with longer read lengths which in turn effect contig assembly (Zhou *et al.*, 2010). Comparing the *de novo* assembled contigs against genomes recovered using PCR with back-to-back primers, that are subsequently cloned and Sanger sequenced; enabled verification of the *de novo* assembly data and provided for an option to archive the unit length of genomic DNA. These specific back-to-back primers can be used as probes to detect and recover these viral genomes in other sample types. Examples of this in this study are the recovery of DfLaCV 3, 5, 6, 8 and 10 (Chapter 3 and 7) that were first detected in dragonfly and damselfly larvae samples and subsequently discovered in a variety of sample types from Lake Sarah (Chapter 7). This was also the case with Avon-Heathcote estuary-associated circular DNA virus (AHEaCV) -8. This viruses was first identified in mollusc samples from the Avon-Heathcote estuary and further isolated from a mollusc sample from Lake Sarah. These examples show further applications of these primers as probes and demonstrates the CRESS DNA viruses being isolated are not necessarily limited to specific ecosystems.

8.2.4 Evolutionary relationships of novel CRESS DNA viruses and circular molecules

The CRESS DNA viruses recovered in this study are extremely diverse with many viruses being distantly related to many of the known classified viruses.

The Rep is the protein mainly used to infer phylogenetic relationships between CRESS DNA viruses as it is the most conserved (Martin *et al.*, 2011). All CRESS DNA viruses recovered as part of this thesis work examined the conserved rolling circle replication (RCR) and superfamily 3 (SF3) helicase motifs to identify possible relationships of novel CRESS DNA viruses with other known viral families. For the most part, almost all the viral genomes as well as some of the circular molecules, showed some levels of conservation in both RCR and SF3 helicase motifs (Chapters 2, 3, 4, 5, 6 and 7) as well as variations of the nonanucleotide motif. RCR motif two and three as well as SF3 helicase Walker-A motifs were consistently the most highly conserved across most of the novel CRESS DNA viruses identified.

Iteron sequences that are found in the intergenic region were identified in many of the novel CRESS DNA viruses recovered. These sequences varied between genomes (Chapter 2 and 3) and are known to play a role as *cis*-acting elements as they interact with DNA binding specificity determinants (SPDs) during viral replication (Londoño *et al.*, 2010). In Chapter 2 the relationship between SPDs and iteron sequences were determined and shows a relationship between a specific core iteron sequence and a specific SPD region. However, SPD sequence regions identified differed across cycloviruses indicating that different genomes have different binding affinities to different iteron sequences.

Phylogenetic analysis of the Rep protein encoded by the viral genomes recovered illustrates that some of the viruses (DfCyV-4, 6, 7 and 8) recovered fall within the recently proposed ssDNA viral family *Cycloviridae* (Chapter 2). As recombination is known to be a mechanism of evolution employed by circular ssDNA viruses such as geminiviruses, circoviruses and nanoviruses (Julian *et al.*, 2012; Lefeuvre *et al.*, 2009; Martin *et al.*, 2011; Massaro *et al.*, 2012; Varsani *et al.*, 2011) further recombination analysis of all cycloviruses isolated to date revealed evidence of both intra- and inter-species recombination (Chapter 2). The results from this analysis indicate that recombination may play a wider role in the evolution of many other CRESS DNA viruses and with the discovery of more CRESS DNA viruses forming new clades, it is likely that further recombination events between these viruses will be identified.

Phylogenetic analysis carried out in Chapter 6 indicates a closely related group of CRESS DNA viruses referred to as “chipoviruses”. These viruses have been recovered from faecal matter, insect material and mollusc tissue, however, the hosts of these viruses are yet to be identified. Unlike many other CRESS DNA viruses being identified, this novel group of viruses show high amino acid conservation in both the Rep and coat protein (CP). Twelve viral isolates (Chapter 6 and 7) were identified that fall into this proposed group. The conservation in both the Rep and CP of these viruses indicate that some CRESS DNA viruses may have mechanisms conserving both major proteins.

The identification of GasCV (Chapter 5) indicates some association between bacteria and CRESS DNA viruses. This virus was subsequently isolated again from different mollusc samples taken from the Avon-Heathcote estuary which suggests that these viruses might be more common than previously thought (Chapter 6). Further to this, several CRESS DNA viral sequences have been identified in the genomes of other organisms including plants, animals, fungi and protists (Liu *et al.*, 2011), with many of the transferred genes being conserved and functional within the host genomes. The close association with the Rep of GasCV with Rep-like sequences found in various bacteria suggests that CRESS DNA viruses can integrate into host genomes and one can speculate on their role in horizontal gene transfer. Examples of this are geminivirus rep-like elements identified in the tobacco genome (Bejarano *et al.*, 1996; Lefeuvre *et al.*, 2011) and rep-like genes have also been found in protozoans *Entamoeba histolytica* and *Giardia intestinalis* (Gibbs *et al.*, 2006; Liu *et al.*, 2011).

8.2.5 CRESS DNA viral distribution and flow of viruses

The use of metagenomic sequencing to identify viral networks in ecosystems is an entirely novel concept. Chapter 6 and 7 provide some insights into how metagenomic sequencing methods can be used on different samples to help understand the 'flow' of CRESS DNA in an ecosystem. This can ultimately be used to infer viral networks, however, clearly the current bottleneck in this endeavour is due to poor knowledge of viruses in ecosystems globally. The networking concept has been applied to other areas of virology such as a study that examined the interaction between variations in tumour viruses and host target proteins that enable cancer genes to be identified (Rozenblatt-Rosen *et al.*, 2012). Other studies have examined the diversity between CRESS DNA viruses in temperate and subtropical seawater using cluster networks (Labonté & Suttle, 2013). The idea of creating networks between similar viruses using different parameters will hopefully shed light on virus-host interactions. However, Chapters 6 and 7 have demonstrated that for this to be achieved successfully a large amount of baseline data of CRESS DNA viruses circulating in ecosystems initially needs to be established and catalogued. In Chapter 7 an investigation on how different sample types are interacting at an ecological level is undertaken, although no significant interactions between viral distribution and samples types were established. Further to this, any study aiming to look at the flow of viruses between different organisms in an environment needs to have a thorough understanding of virus-host interaction as well as an understanding at an ecological level of how different organisms are interacting, as all these factors will influence the flow of viruses in an ecosystem. Creating viral networks will not only be useful for indentifying potential hosts of CRESS DNA viruses, and could possibly provide insights into virus-host co-evolution and the movement of these viruses.

8.2.6 Taxonomy of emerging CRESS DNA viruses

The ICTV has already established taxonomic guidelines for the classification of viruses, which for CRESS DNA currently includes five different families (Chapter 1). Novel CRESS DNA viruses have been discovered from a broad range of samples including animals, plants, faecal matter, fungi, bacteria and various environmental samples. The diversity of the CRESS DNA viruses being discovered (Chapters 2-7) through metagenomics and NGS sequencing from a variety of samples highlights that many of these viruses are yet to be assigned to a family. However, the resent proposal of the two new genera Gemycircularvirus and Cyclovirus, indicate the emerging need for further CRESS DNA viral taxonomy. Many of the viruses in this studied have not been formerly classified and have only been tentatively named as new groups of viruses are emerge.

Of the 268 viral genomes recovered in this research, 13 genomes recovered from dragonfly samples have been classified within the proposed cyclovirus genus. This was achieved by computing pairwise genome-wide identities of the full genomes of CRESS DNA viruses currently assigned to this genus. A threshold of 75% genome-wide identity was established for the species demarcation criteria for cycloviruses, the is in place for circoviruses (Biagini, 2012). The genome-wide nucleotide identities were calculated in SDT v1.2 (Muhire *et al.*, 2014).

Seven viral genomes isolated from dragonfly samples from the United States of America (USA) (Chapter 4) showed sequence similarity to other viruses that have been tentatively classified within the gemycircularvirus genus. Experimental evidence has shown that gemycircularviruses may infect fungi as SsHADV-1 was shown to infect the fungus *Sclerotinia sclerotiorum*. Nonetheless, it is not yet known if other gemycircularviruses infect fungi. Gemycircularviruses have two major ORFs one of which encode a Reps with the conserved GRS motifs indicating similarity to Reps from the *Geminiviridae* family (Nash *et al.*, 2011). Gemycircularviruses may also have putative spliced Reps that requires the splicing of the intron for function, or in the case of SsHADV-1, have a fully functioning Rep. Gemycircularviruses are also recognised by their conserved nonanucleotide motif, in addition to RCR and SF3 helicase motifs present in the Rep. CRESS DNA viruses that fall within this genus have been classified by comparing genome-wide pairwise identity, with a demarcation threshold of >55% currently used to assign viruses to this genus (Dayaram *et al.*, 2012; Kraberger *et al.*, 2013; Kraberger *et al.*, 2014; Ng *et al.*, 2011b; Sikorski *et al.*, 2013; van den Brand *et al.*, 2011) and a >78% genome-wide identity threshold for species classification (Kraberger *et al.*, 2014; Sikorski *et al.*, 2013).

The demarcation criteria in conjunction with the genome-wide pairwise identities calculated using SDT v1.2 or similar tools could potentially be adopted for further studies that identify potential cyclovirus and gemycircularviruses in metagenomic studies. New viral genera and families might also be established using similar taxonomy techniques, using both genome organisation, motifs and genome-wide pairwise identities to classify closely related groups of CRESS DNA viruses.

8.3 Future research

Undertaking this PhD research has unravelled many aspects that could be further investigated. The use of NGS has enabled large amounts of data to be generated, however, it is also vital that these large amounts of data are dealt with appropriately and exploited to their full potential.

- Further, using the specific primers developed throughout this research to identify these CRESS DNA viruses in other ecosystems.
- Use primers to identify previously reported viruses in Lake Sarah in other surrounding lakes in the Cass Basin to determine the distribution of these viruses.
- Determine the hosts of the CRESS DNA viruses identified by carrying out infectivity assays. This will help to understand how these viruses might be moving in ecosystems.
- Once hosts have been identified, transmission studies of these viruses will further show how these viruses are moving between hosts.
- Use data from transmission and as well as metagenomic data to create viral networks viruses in given ecosystems.

References

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., Chan, A. M., Haynes, M., Kelley, S. & other authors (2006). The marine viromes of four oceanic regions. *PLoS biology* **4**, e368.

Baršienė, J., Bučinskienė, R. & Jokšas, K. (2002). Cytogenetic damage and heavy metal bioaccumulation in molluscs inhabiting different sites of the Neris River. *Ekologija* **2**, 52-57.

Bejarano, E. R., Khashoggi, A., Witty, M. & Lichtenstein, C. (1996). Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. *Proceedings of the National Academy of Sciences* **93**, 759-764.

Biagini, P., M. Bendinelli, S. Hino, L. Kakkola, A. Mankertz, C. Niel, H. Okamoto, S. Raidal, C. G. Teo, and D. Todd (2012). *Family - Circoviridae, Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*. San Diego: Elsevier.

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J. M., Nulton, J., Salamon, P. & Rohwer, F. (2003). Metagenomic analyses of an uncultured viral community from human feces. *Journal of bacteriology* **185**, 6220-6223.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., Azam, F. & Rohwer, F. (2002). Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences* **99**, 14250-14255.

Dayaram, A., Opong, A., Jäschke, A., Hadfield, J., Baschiera, M., Dobson, R. C. J., Offei, S. K., Shepherd, D. N., Martin, D. P. & other authors (2012). Molecular characterisation of a novel cassava associated circular ssDNA virus. *Virus Research* **166**, 130-135.

Delwart, E. L. (2007). Viral metagenomics. *Reviews in medical virology* **17**, 115-131.

Fancello, L., Trape, S., Robert, C., Boyer, M., Popgeorgiev, N., Raoult, D. & Desnues, C. (2013). Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. *The ISME journal* **7**, 359-369.

Gibbs, M. J., Smeianov, V. V., Steele, J. L., Upcroft, P. & Efimov, B. A. (2006). Two families of rep-like genes that probably originated by interspecies recombination are represented in viral, plasmid, bacterial, and parasitic protozoan genomes. *Molecular biology and evolution* **23**, 1097-1100.

Julian, L., Lorenzo, A., Chenuet, J.-P., Bonzon, M., Marchal, C., Vignon, L., Collings, D. A., Walters, M., Jackson, B. & other authors (2012). Evidence of multiple introductions of beak and feather disease virus into the Pacific islands of Nouvelle-Caledonie (New Caledonia). *Journal of General Virology* **93**, 2466-2472.

Kim, K.-H., Chang, H.-W., Nam, Y.-D., Roh, S. W., Kim, M.-S., Sung, Y., Jeon, C. O., Oh, H.-M. & Bae, J.-W. (2008). Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Applied and environmental microbiology* **74**, 5975-5985.

Kraberger, S., Stainton, D., Dayaram, A., Zawar-Reza, P., Gomez, C., Harding, J. S. & Varsani, A. (2013). Discovery of Sclerotinia sclerotiorum hypovirulence-associated virus-1 in urban river sediments of Heathcote and Styx rivers in Christchurch city, New Zealand. *Genome announcements* **1**, e00559-00513.

Kraberger, S., Argüello-Astorga, G. R., Greenfield, G. L., Galilee, C., Law, D., Martin, D. P. & Varsani, A. (2014). Characterisation of a diverse range of Rep-encoding ssDNA viruses recovered from a sewage treatment oxidation pond. *Infection, Genetics and Evolution* **In Review**.

Labonté, J. M. & Suttle, C. A. (2013). Previously unknown and highly divergent ssDNA viruses populate the oceans. *The ISME journal* **7**, 2169-2177.

Lefeuvre, P., Lett, J.-M., Varsani, A. & Martin, D. (2009). Widely conserved recombination patterns among single-stranded DNA viruses. *Journal of virology* **83**, 2697-2707.

Lefeuvre, P., Harkins, G. W., Lett, J.-M., Briddon, R. W., Chase, M. W., Moury, B. & Martin, D. P. (2011). Evolutionary time-scale of the begomoviruses: evidence from integrated sequences in the Nicotiana genome. *PloS one* **6**, e19193.

Li, L., Kapoor, A., Slikas, B., Bamidele, O. S., Wang, C., Shaukat, S., Masroor, M. A., Wilson, M. L., Ndjango, J. B. N. & other authors (2010). Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces. *Journal of Virology* **84**, 1674.

Li, L., McGraw, S., Zhu, K., Leutenegger, C. M., Marks, S. L., Kubiski, S., Gaffney, P., Cruz Jr, F. N. D., Wang, C. & other authors (2013). Circovirus in tissues of dogs with vasculitis and hemorrhage. *Emerging infectious diseases* **19**, 534.

Liu, H., Fu, Y., Li, B., Yu, X., Xie, J., Cheng, J., Ghabrial, S. A., Li, G., Yi, X. & other authors (2011). Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. *BMC Evolutionary Biology* **11**, 276.

Londoño, A., Riego-Ruiz, L. & Argüello-Astorga, G. R. (2010). DNA-binding specificity determinants of replication proteins encoded by eukaryotic ssDNA viruses are adjacent to widely separated RCR conserved motifs. *Archives of virology* **155**, 1033-1046.

Martin, D. P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P. & Varsani, A. (2011). Recombination in eukaryotic single stranded DNA viruses. *Viruses* **3**, 1699-1738.

Massaro, M., Ortiz-Catedral, L., Julian, L., Galbraith, J. A., Kurenbach, B., Kearvell, J., Kemp, J., van Hal, J., Elkington, S. & other authors (2012). Molecular characterisation of beak and feather disease virus (BFDV) in New Zealand and its implications for managing an infectious disease. *Archives of virology* **157**, 1651-1663.

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

Nash, T. E., Dallas, M. B., Reyes, M. I., Buhrman, G. K., Ascencio-Ibanez, J. & Hanley-Bowdoin, L. (2011). Functional Analysis of a Novel Motif Conserved across Geminivirus Rep Proteins. *Journal of Virology* **85**, 1182.

Ng, T. F. F., Alavandi, S., Varsani, A., Burghart, S. & Breitbart, M. (2013). Metagenomic identification of a nodavirus and a circular ssDNA virus in semi-purified viral nucleic acids from the hepatopancreas of healthy *Farfantepenaeus duorarum* shrimp. *Dis Aquat Org* **105**, 237-242.

Ng, T. F. F., Duffy, S., Polston, J. E., Bixby, E., Vallad, G. E. & Breitbart, M. (2011a). Exploring the Diversity of Plant DNA Viruses and Their Satellites Using Vector-Enabled Metagenomics on Whiteflies. *PloS one* **6**, e19050.

Ng, T. F. F., Marine, R., Wang, C., Simmonds, P., Kapusinszky, B., Bodhidatta, L., Oderinde, B. S., Wommack, K. E. & Delwart, E. (2012). High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. *Journal of virology* **86**, 12161-12175.

Ng, T. F. F., Willner, D. L., Lim, Y. W., Schmieder, R., Chau, B., Nilsson, C., Anthony, S., Ruan, Y., Rohwer, F. & other authors (2011b). Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. *PloS one* **6**, e20579.

Padilla-Rodriguez, M., Rosario, K. & Breitbart, M. (2013). Novel cyclovirus discovered in the Florida woods cockroach *Eurycotis floridana* (Walker). *Archives of virology* **158**, 1389-1392.

Rosario, K. & Breitbart, M. (2011). Exploring the viral world through metagenomics. *Current Opinion in Virology* **1**, 289-297.

Rosario, K., Duffy, S. & Breitbart, M. (2009). Diverse circovirus-like genome architectures revealed by environmental metagenomics. *Journal of General Virology* **90**, 2418-2424.

Rosario, K., Duffy, S. & Breitbart, M. (2012a). A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Archives of virology* **157**, 1851-1871.

Rosario, K., Padilla-Rodriguez, M., Kraberger, S., Stainton, D., Martin, D. P., Breitbart, M. & Varsani, A. (2013). Discovery of a novel mastrevirus and alphasatellite-like circular DNA in dragonflies (Epiprocta) from Puerto Rico. *Virus Research*, 231-237.

Rosario, K., Dayaram, A., Marinov, M., Ware, J., Kraberger, S., Stainton, D., Breitbart, M. & Varsani, A. (2012b). Diverse circular ssDNA viruses discovered in dragonflies (Odonata: Epiprocta). *Journal of General Virology* **93**, 2668-2681.

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., Colombet, J., Sime-Ngando, T. & Debroas, D. (2012). Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PloS one* **7**, e33641.

Rozenblatt-Rosen, O., Deo, R. C., Padi, M., Adelmant, G., Calderwood, M. A., Rolland, T., Grace, M., Dricot, A., Askenazi, M. & other authors (2012). Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins. *Nature* **487**, 491-495.

Sikorski, A., Massaro, M., Kraberger, S., Young, L. M., Smalley, D., Martin, D. P. & Varsani, A. (2013). Novel myco-like DNA viruses discovered in the faecal matter of various animals. *Virus Research* **177**, 209-216.

van den Brand, J. M. A., van Leeuwen, M., Schapendonk, C. M., Simon, J. H., Haagmans, B. L., Osterhaus, A. D. M. E. & Smits, S. L. (2011). Metagenomic analysis of the viral flora of pine marten and European badger feces. *Journal of Virology* **86**, 2360-2365.

van der Oost, R., Heida, H. & Opperhuizen, A. (1988). Polychlorinated biphenyl congeners in sediments, plankton, molluscs, crustaceans, and eel in a freshwater lake: Implications of using reference chemicals and indicator organisms in bioaccumulation studies. *Archives of Environmental Contamination and Toxicology* **17**, 721-729.

Varsani, A., Regnard, G. L., Bragg, R., Hitzeroth, I. I. & Rybicki, E. P. (2011). Global genetic diversity and geographical and host-species distribution of beak and feather disease virus isolates. *Journal of General Virology* **92**, 752-767.

Whon, T. W., Kim, M.-S., Roh, S. W., Shin, N.-R., Lee, H.-W. & Bae, J.-W. (2012). Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere. *Journal of virology* **86**, 8221-8231.

Yoshida, M., Takaki, Y., Eitoku, M., Nunoura, T. & Takai, K. (2013). Metagenomic Analysis of Viral Communities in (Hado) Pelagic Sediments. *PloS one* **8**, e57271.

Zhou, X., Ren, L., Li, Y., Zhang, M., Yu, Y. & Yu, J. (2010). The next-generation sequencing technology: a technology review and future perspective. *Science China Life Sciences* **53**, 44-57.