Molecular characterisation of novel single stranded DNA viruses recovered from animal faeces

A thesis submitted in partial fulfilment of the requirements for the

Degree of Master of Science in Microbiology

University of Canterbury

By Alyssa Sikorski

University of Canterbury, New Zealand

2013

Table of Contents

Chapter 4: Two novel ssDNA viruses recovered from nesting material

Chapter 5: Conclusion

List of Tables

Chapter 1: Introduction

Chapter 2: Isolation of a novel ssDNA virus from porcine faeces: proof of concept

Chapter 3: Molecular characterisation of fourteen novel gemycircularvirus and gemycircularvirus-like isolates recovered from animal faecal samples in New Zealand

Chapter 4: Two novel ssDNA viruses recovered from nesting material

Chapter 5: Conclusion

List of Figures

Chapter 1: Introduction

Chapter 2: Isolation of a novel ssDNA virus from porcine faeces: proof of concept

Chapter 3: Molecular characterisation of fourteen novel gemycircularvirus and gemycircularvirus-like isolates recovered from animal faecal samples in New Zealand

Chapter 4: Two novel ssDNA viruses recovered from nesting material

Acknowledgements

It is with much gratitude that I would like to acknowledge my supervisor, Dr. Arvind Varsani, for his invaluable support and guidance over the past year, and for bringing me into his lab. Working in the virology lab group at the University of Canterbury has been a really enjoyable experience. Thank you Anisha Dayaram, Simona Kraberger, Daisy Stainton, Laurel Julian, Dorien Coray, and Ryan Catchpole for helping me find my way in the lab and for all the good laughs.

This thesis was possible because of friends and colleagues in the Biology Department at University of Canterbury who responded to my strange request and contributed tubes filled with animal faecal matter from their homes and from their travels around New Zealand. The process of sample collection was shortened significantly because of the help I received. Thank you to Melanie Massaro, Laura Young, Carina Davis, Aaron Stevens, Maggie Tisch, Matt Walters, Anisha Dayaram, Simona Kraberger, Daisy Stainton, Laurel Julian, Dorien Coray, and Ryan Catchpole.

I would also like to acknowledge everyone who has contributed to the publications that have been included in this thesis, including the virology lab group, Gerardo R. Argüello-Astorga, John Kearvell, and Simon Elkington, and my co-supervisor, Renwick C.J. Dobson.

Abstract

Recent metagenomic studies have shown that there is a higher diversity of ssDNA viruses in the environment than previously thought. While some viral families are well characterised, novel ssDNA isolates discovered with sequence-independent molecular techniques are often too divergent to fit within the currently established viral taxonomy. Several factors have contributed to the gap in knowledge, including: the (previously) high cost of sequencing, the disproportionate amount of research that occurs after a threat is identified, and the use of sequence-based molecular techniques to isolate viral sequences. Recent studies have begun to explore viral diversity in the environment, however, most of these studies have occurred outside New Zealand. Several benefits would come from uncovering the true ssDNA viral diversity and global distribution including improving the resolution of the current taxonomic structure for identifying unknown isolates and inferring possible virus-host relationships, and providing baseline data for the development of disease prevention and monitoring strategies. Studies specific to the New Zealand environment are essential. With its geographical isolation and Gondwana ancestry, New Zealand will possess a unique viral sequence space. Studies on local viral diversity and the spread of ssDNA viruses are going to be most relevant if they are conducted within the established ecosystems in New Zealand.

In this dissertation, a novel protocol was developed for exploring viral diversity in the New Zealand environment using basic molecular techniques and animal faecal samples. Design considerations included: identifying highly novel small circular viral sequences with DNA genomes without the use of specific primers, inflicting as little environmental impact as possible, and keeping the cost low. The faecal sampling approach does not require animal handling and therefore incorporates the use of viral reservoirs while remaining non-invasive. The molecular techniques in this protocol used non-specific rolling circle amplification (RCA) followed by restriction enzyme (RE) digests, cloning, and sequencing of the cloned genomes via sanger sequencing. This inexpensive exploratory method provided preliminary sequence information from which primers were designed for recovery of full viral genomes.

The success of this protocol was demonstrated by the recovery and molecular characterisation of a novel ssDNA virus isolate from a pig faecal sample, which was tentatively named porcine stool-associated circular virus (PoSCV). This protocol was then applied to sample viruses in the faecal matter from variety of domesticated, wild, and farmed animals in New

Zealand. The faecal samples were collected from the North and South Island of New Zealand as well as South East Island of the Chatham Islands (Rangatira). Several putative gemycircularviral isolates (novel viruses with similarities to geminiviruses and the recently discovered ssDNA virus infecting *Sclerotinia sclerotiorum*) were identified in the sequencing results based on BLASTx similarities to viral sequences available in public databases (GenBank). The full genomes of these isolates were recovered and characterised. Identification was based on genome organization, phylogenetic analysis of the replication associated protein (Rep), and full genome nucleotide pairwise identities. Fourteen novel ssDNA virus sequences relating to gemycircularviruses were discovered, of which ten were representative of new species (FaSCV-1, 2, 3, 4, 5, 6, 7, 8, 9, and 10) and three were identified as strains of the same species (FasGCV-1). Two additional isolates were discovered to be distantly related to these viruses: Ostrich faecal associated ssDNA virus (OfaV) and Rabbit faecal associated ssDNA virus (RfaV). Additionally, this protocol was used to recover novel ssDNA viruses from the nesting material of a dead Yellow-crowned Parakeet chick found in the Poulter Valley in the South Island of New Zealand. The nesting material likely contained faecal matter and thus represented another approach strategy for exploring ssDNA viruses in the environment. Two novel ssDNA isolates were discovered and molecularly characterised: *Cyanoramphus* nest-associated circular X virus (CynNCXV), and *Cyanoramphus* nest-associated circular K virus CynNCKV.

Chapter 1

Introduction

1.1 Single-stranded DNA virus classification

The International Committee on the Taxonomy of Viruses is a group of experts responsible for formal taxonomic classification of viruses after preliminary analysis, discussion, and debate by the public. Currently the ICTV recognizes seven ssDNA families: *Anelloviridae*, *Circoviridae*, *Geminiviridae*, *Inoviridae*, *Microviridae*, *Nanoviridae*, and *Parvoviridae* (King *et al.*, 2011) but the true diversity of single-stranded (ss) DNA viruses is unknown (Rosario & Breitbart, 2011). Introduction of metagenomic sequencing technologies, occurring in the last few years, is rapidly transforming the known diversity of ssDNA viruses. New genera have been proposed, such as Cyclovirus and Gemycircularvirus. The most recent ICTV report, the 9th report released in 2011, contains several unassigned viruses that may represent unnamed viral families and genera (King *et al.*, 2011). Classification will continue to change with the discovery of novel ssDNA viruses.

Virus classification is dependent on virion properties such as morphology, genome organisation, and mechanism of replication (Fenner & Maurin, 1976; King *et al.*, 2011). Most ssDNA viruses are morphologically similar with icosahedral shaped virions ranging in size from 12-30 nm (King *et al.*, 2011). The exceptions are the *Geminiviridae* family which has twinned icosahedral particles (King *et al.*, 2011) and the *Inoviridae* family which has rod-shaped or filamentous shaped particles. *Inoviridae* particle size ranges from 15 nm diameter to 200-400 nm length rod shaped and 7 nm diameter to 700-3500 nm length for filamentous shaped particles (King *et al.*, 2011).

The significance of the host has biased the range and extent of knowledge of known ssDNA viruses; the most thoroughly studied ssDNA viruses are those that cause disease in humans, and agricultural crops and livestock of significant economic importance. Of the known ssDNA viruses there are two families that infect prokaryotes and five that infect eukaryotes. Eukaryotic ssDNA viruses include those that infect plants and animals. *Geminiviridae* and *Nanoviridae* are known to infect plants while *Circoviridae*, *Anelloviridae* and *Parvoviridae* are known to infect animals, though only *Parvoviridae* has been found to infect invertebrate animals (King *et al.*, 2011). Prokaryote infecting ssDNA viruses are in the families *Inoviridae* and *Microviridae* (King *et al.*, 2011). Prokaryote infecting ssDNA viruses include those that infect bacteria, and in recent years there have been ssDNA viruses found infecting archaea (Mochizuki *et al.*, 2012; Pietilä *et al.*, 2009).

Single-stranded DNA viruses have high mutation rates. Specifically, nucleotide substitution rates for geminiviruses were estimated at 1.60 x 10^{-3} substitutions per site per year for DNA-A components and 1.33 x 10^{-4} substitutions per site per year for DNA-B components (Duffy & Holmes, 2009) while the nanovirus *Faba bean necrotic yellows virus* (FBNYV) is estimated to have a similar substitution rate of 1.78 x 10^{-3} substitutions per site per year (Grigoras *et al.*, 2010). The high substitution rates may be due to ssDNA molecules specifically being a target for mutagenic processes (Van Der Walt *et al.*, 2008). Recombination rates are also high amongst ssDNA viruses. Recombination of viruses is evident across all families of ssDNA. Particularly, a comparison of recombination breakpoints and recombination occurrence among a broad range of ssDNA viruses revealed some conserved patterns showing that recombination was lowest on open reading frames (ORFs) corresponding to structural proteins and highest in intergenetic regions (Lefeuvre *et al.*, 2009).

1.2 Single-stranded DNA viral genomes

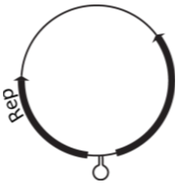
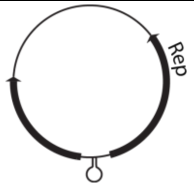
Eukaryotic ssDNA virus genomes are relatively small, with most single component genomes ranging from 2-6 kb and multiple component genome components ranging from 1-3 kb, while prokaryote infecting ssDNA genomes are slightly larger, ranging from 4-12 kb (King *et al.*, 2011). A small genome is possible because most ssDNA viruses rely on host polymerases and only one protein, the replication associated protein, for genome replication. Nanoviruses and some geminiviruses have multi-component genomes, comprised of six or eight genome components and one or two genome components respectively. These components are packaged into separate virions but they are all needed within an infected cell to fully function.

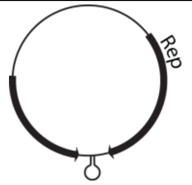
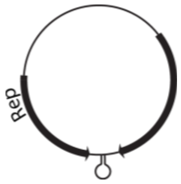
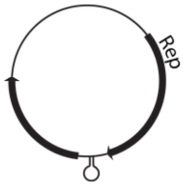
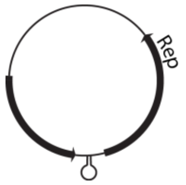
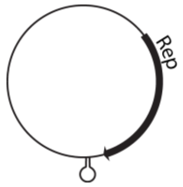
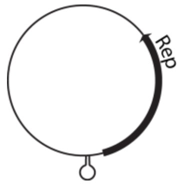
Individually packaged virions have been known to swap between two isolates. These reassortment events have been observed in *Nanoviridae* and *Geminiviridae*.

The genomes for most ssDNA viruses are circular or composed of circular components including *Anelloviridae*, *Circoviridae*, *Inoviridae*, *Microviridae*, *Nanoviridae*, and *Geminiviridae* (King *et al.*, 2011). The genomes of circular ssDNA have diverse organizations but contain conserved motifs within the replication associated protein (Rep) that indicate similarities in how they may replicate. The circular ssDNA viruses with similar conserved motifs are thought to replicate using the rolling circle replication (RCR) mechanism. *Parvoviridae* is the only known ssDNA virus with a linear genome. Linear ssDNA viruses in the *Parvoviridae* family have been shown to replicate using a similar mechanism known as the rolling hairpin replication (RHR) mechanism (Cotmore & Tattersall, 1996; Tattersall & Ward, 1976).

Single-stranded DNA viruses have been categorized into eight types based on open reading frame number and orientation in comparison to structural and sequence motifs such as the stem-loop element and the nonanucleotide motif (see Table 1.1) (Rosario *et al.*, 2009a). To organize by type, the ssDNA virus genomes are orientated with the assumption that the nonanucleotide motif is located on the positive-sense strand (Rosario *et al.*, 2009a).

Table 1.1 Types of circular ssDNA viruses oriented with the nonanucleotide motif in the positive sense strand. Adapted from Rosario *et al.* (2009a).

Genome Type	Example Illustration	Nonanucleotide motif and Rep in same strand	Single or Multiple Major ORFs	Ambisense or Unisense	Direction of ORFs compared to stem-loop
Type I		Yes	Multiple	Ambisense	ORFs read away from stem-loop
Type II		No	Multiple	Ambisense	ORFs read away from stem-loop

Type III		Yes	Multiple	Ambisense	ORFs read towards stem-loop
Type IV		No	Multiple	Ambisense	ORFs read towards stem-loop
Type V		Yes	Multiple	Unisense	ORFs read clockwise
Type VI		No	Multiple	Unisense	ORFs read counter-clockwise
Type VII		Yes	Single	N/A	ORF reads clockwise
Type VIII		No	Single	N/A	ORF reads counter clockwise

1.3 Rolling circle replication and the Rep protein

The RCR mechanism was first described by Gilbert and Dressler (1968) and has become well understood as a mechanism for replication of circular ssDNA such as that of bacteriophage, plasmids, and viruses. RCR mechanisms have been suggested for the genomes of circular ssDNA viruses such as geminiviruses and circoviruses, and the first confirmation of RCR as a mechanism for viral replication was in *Abutilon mosaic vir*us described in Jeske *et al.* (2001).

RCR for ssDNA viruses is thought to occur in the following manner: First, single stranded DNA is converted into a double-stranded intermediate. These are packed into mini-chromosomes (Abouzid *et al.*, 1988; Pilartz & Jeske, 1992). Once the Rep is expressed through host machinery, it cleaves DNA at the origin of replication. This creates an open circle replicative form that serves as the template for strand synthesis. Strand synthesis occurs in a continuous cyclical manner until is it displaced by the Rep protein and relegated into circular ssDNA replicates. Like RCR, RHR mechanisms also use a double stranded replication intermediate. Palindromic repeats on each end of the linear ssDNA genome self anneal to act as primers for host polymerase where new strand synthesis occurs in one direction until the other palindromic region is synthesized and then the direction switches (Cotmore & Tattersall, 1996).

Essential for RCR is the Rep. The Rep is highly conserved among geminiviruses, nanoviruses, circoviruses, and many uncharacterised small ssDNA viruses, and phylogenetic analysis of the Rep has shown an evolutionary connection between them (Martin *et al.*, 2011). Additionally, genome replication of *Porcine circovirus* (PCV) inside bacteria without use of PCV replication initiator protein suggests that circoviruses and other similar circular ssDNA viruses may have evolved from prokaryotic replicons (Cheung, 2006).

The origin of replication (*ori*) is a conserved RCR element marked by two types of DNA patterns and a putative stem-loop. The first DNA pattern is a conserve nonanucleotide motif at the top of the putative stem-loop structure. The nonanucleotide motif is usually a 9 bp sequence (NANTATTAC) with the cleavage site between the 7th and 8th base pair. Iterative sequences surround the putative hairpin structure. These are recognition and binding sites for the Rep protein.

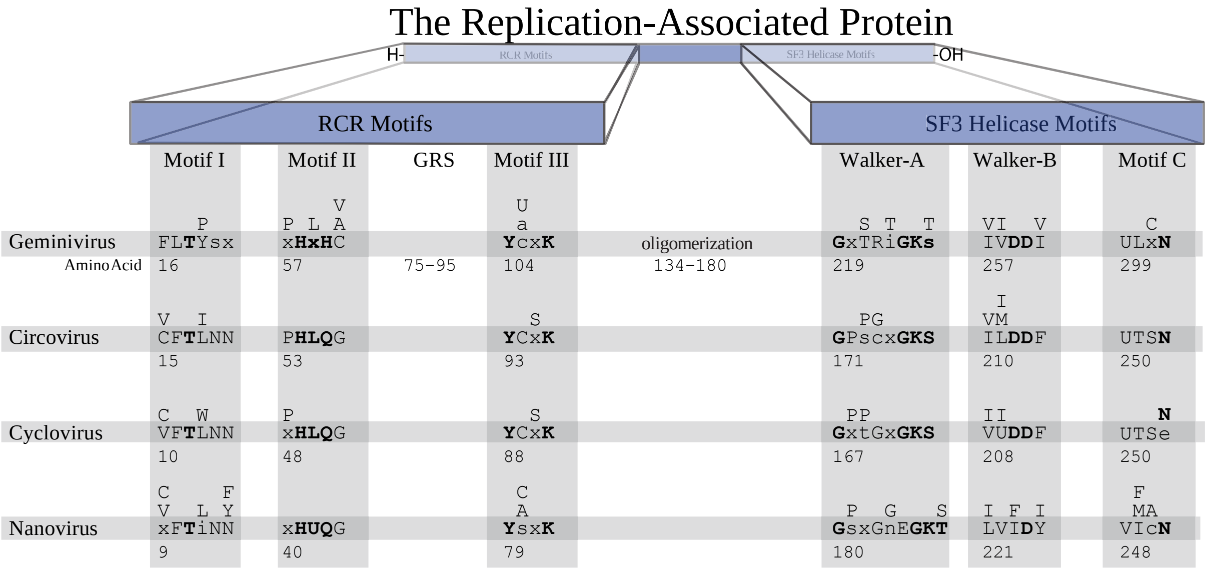


Figure 1.1 The sequences of the conserved motifs in the replication-associated proteins of geminivirus, circovirus, cyclovirus, and nanovirus in the order in which they are in the replication-associated protein and marked by amino acid positions, as has been adapted from Rosario *et al.* (2012b). The RCR motifs and the SF3 Helicase Motifs are described with a "U" representing any bulky amino-acid (I,L,V,M,F,Y, or W) and "x" representing any amino-acid (Rosario *et al.*, 2012b). Conserved residues are in bold. Lowercase letters indicate lower frequency of that amino acid at that position while uppercase indicates higher frequency(Rosario *et al.*, 2012b). 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).

1.3.1 *Helicase domain*

Helicases denature or unwind DNA by destabilization of hydrogen bonds between base pairs of double-stranded (ds) oligonucleotides. On the ssDNA viral Rep protein, helicase activity is used to unwind the dsDNA intermediate to ssDNA for nascent strand synthesis in a reaction dependent on energy acquired through nucleoside 5'-triphosphates (NTP) binding and hydrolysis (Gorbalenya *et al.*, 1990; Ilyina & Koonin, 1992). Several small DNA and RNA viral Rep helicases with similarities to ssDNA viral Reps have been classed in the SF3 helicase superfamily based on sequence identity (Gorbalenya *et al.*, 1990; Walker *et al.*, 1982). This superfamily is characterised by 3 main conserved amino acid sequence motifs which occur over a small <120nt stretch of the protein (Gorbalenya *et al.*, 1990). These conserved motifs are Walker-A [GxxxxGK(S/T)], Walker-B [hhxh(D/E)(D/E)] (Walker *et al.*, 1982), and motif C [h(T/S/x)(T/S/x)N] (Gorbalenya *et al.*, 1990) where "x" denotes any amino acid residue and "h" denotes any hydrophobic amino acid residue. A fourth amino acid motif has also been shown to be associated with SF3 helicases, named motif B', and is located between two of the previously identified motifs and may be involved in the coupling of NTPase activity and DNA binding (Koonin, 1993; Yoon-Robarts *et al.*, 2004).

Most of the predictions for ssDNA viral Rep helicase come from experimental data acquired from the solved structures of other SF3 helicases. Simian virus 40, a model for replication in this group, has been studied extensively through site directed mutagenesis (Fanning & Knippers, 1992; Pipas *et al.*, 1983; Schneider & Fanning, 1988; Shen *et al.*, 2005) and was the first structure to be solved (James *et al.*, 2003). Since, the helicase proteins of Human papillomavirus (Abbate *et al.*, 2004) and Adeno-associated virus type 2 (James *et al.*, 2003) have been solved. The helicase activity of a ssDNA viral Rep helicase domain was first confirmed in a geminivirus, *Tomato yellow leaf curl Sardinia virus* (Clérot & Bernardi, 2006). To date, most of the studies have been on geminiviruses, nanoviruses, and circoviruses. In the viral Rep SF3 helicases, the putative helicase domain is associated with an origin binding domain. Geminivirus helicases move in the 3' to 5' direction and are dependent on oligamerization state, and have similar function to other helicases.

All helicases share an NTP binding mechanism. NTPase activity is achieved through structural motifs on the helicase domain that form a "P-loop". The helicase domain contains is a NTP synthesis complex. The main NTPase motifs include Walker-A, Walker-B, and motif C. The first of the helicase motifs, Walker-A and Walker-B, were initially identified in enzymes requiring ATP (Walker *et al.*, 1982). Walker-A motif forms part of the P-loop that is used for nucleotide triphosphate recognition (Walker *et al.*, 1982). Motif C is located on the C-terminal end of the Rep protein. It is characterised as a series of hydrophobic residues followed by an Aspartic acid residue. A fourth motif, motif B, may be involved in coupling of ATP hydrolysis and DNA binding (Koonin, 1993; Yoon-Robarts *et al.*, 2004).

1.3.2 *N-terminal domain*

The N-terminal domain of the Rep protein contains the conserved RCR motifs and the specificity binding determinants (SPDs). There are three RCR motifs, RCR motif I, II, and III, that are conserved in the Rep proteins of ssDNA viruses, phage, and plasmids (Ilyina &

Koonin, 1992). The SPDs mediate binding between the Rep and the double stranded intermediate. When Rep bind to the iterative DNA sequences that surround the ori this is the first step in replication initiation. The SPDs, found in two regions of the N-terminal of the Rep protein, come together to form a beta sheet that is probably responsible for binding affinity (Londoño *et al.*, 2010). Nearby is RCR Motif I, with the conserved amino acid sequence Fu(t/u)(l/y)(t/p). The function of Motif I is still uncertain (Ilyina & Koonin, 1992). Motif II has the amino acid sequence (p/u)HuH where "u" denotes a large hydrophobic residue. The histadine residues on either side of the hydrophobic residue bind metal ions (Ilyina & Koonin, 1992; Koonin & Ilyina, 1993). Motif II is thought to coordinate the metal ions that are important for cleaving DNA during the nicking step of RCR (Ilyina & Koonin, 1992) and has been exhibited in several ssDNA viruses. Motif III is the site where DNA is cleaved and has the amino acid sequence YxxK, where "x" denotes any amino acid residue. The active tyrosine located on this domain is responsible for nicking the DNA, and then forms a covalent bond to the newly formed 5' end of the nicked strand. The N-terminal domain of the Rep protein catalyzes ligation of the new strands after replication (Laufs *et al.*, 1995).

1.4 Exploring ssDNA viruses in the environment

A primary reason for exploring the ssDNA sequence space is that it is grossly underestimated. Non-specific amplification of viral DNA from environmental samples followed by metagenomic sequencing techniques has revealed a large number of highly divergent ssDNA viruses, highlighting how inadequately viral diversity is represented by known ssDNA viruses in GenBank. Illuminating true viral diversity and spread has become possible with new technology but the methods are still being refined. Inexpensive metagenomic sequencing has only recently become available as have non-specific amplification methods for viral discovery.

Another significant objective is viral disease outbreak surveillance and prevention. Environmental sampling enables exploration of the entire viral sequence space, exposing a complex array of viruses interacting with the environment that goes beyond what has been discovered to be pathogenic to agricultural crops, humans, or livestock. Through environmental sampling, it is possible to identify viral species that may cause infection in the future. Also, increased knowledge of viral genome features through identification of sequence and structural motifs, the increased knowledge of the genetic diversity available for horizontal gene transfer, and increased knowledge of the host ranges of will all contribute to responding to or preventing future disease outbreak. Additionally, knowledge of a virome in both healthy and diseased conditions is a base on which to develop screening programs.

Environmental sampling is a simple approach that covers a wide spectrum for the purpose of capturing viral diversity. It is especially efficient if the sample medium has effectively concentrated viruses from the ecosystem, for it then provides a comprehensive snapshot of viral diversity without having to do high density sampling. One way to effectively concentrate viruses is through the use of specially designed machinery. Viruses can be concentrated mechanically by filtering large amounts of air or water. This method will have an environmental impact because of the large amounts of air and water that must run through the machinery. Natural viral accumulation, bioaccumulation, can also occur inside living organisms. The benefit of sampling living organisms for bioaccumulation of viruses will be dependent on the length of life of that organism and the spatial coverage that it has in the environment. Bioaccumulation can be exploited in living organisms through sampling of tissue, serum, or faecal matter. Sampling serum and tissue are both highly invasive to the environment. Animal tissue and serum sampling requires specific protocols and must account for animal welfare and environmental impact. Additionally, insect and plant tissue sampling requires removal of large quantities of tissue from the environment because higher density sampling is necessary to account for a low spatial coverage or short life span. Animal faecal sampling is the least invasive because only a small amount of faecal matter is needed and it does not require animal handling. The exception is sampling human faecal matter, which is highly invasive and requires strict protocols.

1.4.1 *Single-stranded DNA viruses in environmental samples*

1.4.1.1 Soil

The first attempt at exploring the diversity of ssDNA in the environment was carried out on soil from a rice paddy in Daejeon, Korea (Kim *et al.*, 2008). The technique utilized multiple displacement amplification (MDA) via Phi29 polymerase with random hexamers which had been previously successful for amplification of unknown ssDNA viruses. By not adding DNase, the ssDNA in the sample was preferentially amplified. The number of significant BLAST hits retrieved from the sequencing data was 50% with DNAse in the protocol and 90% without (Kim *et al.*, 2008). After amplification, the DNA was sheared using a HydroGene machine into DNA fragments that were then 'shotgun' cloned and sequenced (Kim *et al.*, 2008). From 1g of soil, 2.77 x 10^8 ± 0.47 x 10^8 viruses were estimated using epifluorescence microscopy (EFM) methods (Kim *et al.*, 2008). Assembled contigs with repeated sequences at the beginning and end were thought to be circular ssDNA virus genomes and BLAST analysis showed some had similarities to circoviruses and some had no significant similarities (Kim *et al.*, 2008). Back-to-back primers were designed and used in PCR to amplify and sequence the full genomes. This resulted in 18 of the 19 putative circular sequences confirmed and one sequence unsuccessful possibly because of chimeras (Kim *et al.*, 2008).

Another study investigated the ocean sediments from around the northwest Pacific (Yoshida *et al.*, 2013). Roche 454 pyrosequencing was used with two different amplification techniques: multiple displacement amplification (MDA) via Phi29 polymerase with random hexamers for ssDNA and linker-amplified shotgun library (LASL) for dsDNA (Yoshida *et al.*, 2013). Over 70% of the sequencing data had no matches in blast or had hits with E-values $>10^{-3}$ (Yoshida *et al.*, 2013). Of the putative eukaryotic ssDNA viruses identified based on blast hits with E-values $<10^{-3}$, many had hits to geminiviruses, circoviruses, nanoviruses, and microviruses, and novel unclassified viruses (Yoshida *et al.*, 2013). Yoshida *et al.* (2013) used MetaVir (http://metavir-meb.univ-bpclermont.fr/) to identify motifs in the sequences for sorting and aligning sequence reads before phylogenetic analysis. First, the long sequence reads were analyzed in separate phylogenetic analyses based on genetic markers in the major capsid protein and Rep protein domains, and secondly a phylogenetic analysis including more sequence reads of short length was automatically generated by MetaVir (Yoshida *et al.*, 2013). All phylogenetic analysis showed the highly divergent nature of the sequence reads obtained from the deep sea sediment, and that they likely represent novel ssDNA viruses (Yoshida *et al.*, 2013).

1.4.1.2 Water

Rosario *et al.* (2009a) searched for ssDNA viruses in metagenomic sequencing datasets from: coastal water of British Columbia and open ocean water from the Sargasso sea (Angly *et al.*, 2006), post wastewater treatment water (Rosario *et al.*, 2009b), and estuarine water from the Chesapeake Bay (Rosario *et al.*, 2009a). The datasets of assembled contigs contained sequences with similarities to circoviruses, nanoviruses, and geminiviruses (Rosario *et al.*, 2009a). Additionally, the potential full circular genomes that were assembled from contigs in the circovirus-like database were verified by inverse PCR (Rosario *et al.*, 2009a). It is important to note that not all sequences were verified, possibly due to the formation of chimeras (Rosario *et al.*, 2009a) or hybrids of different parent sequences that are falsely assembled into contigs (Kunin *et al.*, 2008). Ten novel ssDNA genomes were characterised with diverse genome architectures, such as differences in orientation of the ORFs and the number of ORFs pertaining to the Rep (Rosario *et al.*, 2009a).

Viral diversity has been observed in fresh water environments such as an Antarctic Lake (López-Bueno *et al.*, 2009), two fresh water lakes, Lakes Bourget and Pavin, France (Roux *et al.*, 2012a), and four perennial ponds in central Sahara, Mauritania (Fancello *et al.*, 2012). The greatest portion of the ssDNA viruses had hits to bacteria-infecting viruses; *Microviridae* family of ssDNA bacteriophages were high in samples from perennial ponds (Fancello *et al.*, 2012) and Antarctic Lake (López-Bueno *et al.*, 2009) and in fresh water lakes in France (Roux *et al.*, 2012a) specifically from which 81 novel *Microviridae* genomes were assembled from contigs (Roux *et al.*, 2012b). Also observed was a large portion of assembled contigs that related to circoviruses, specifically in the Antarctic lake and the French lakes the circovirus-like particles had low amino acid identity to known sequences (López-Bueno *et al.*, 2009; Roux *et al.*, 2012a).

1.4.1.3 Air

Viral spatial and temporal diversity was observed in the air above three different regions that varied in geographical location as well as the primary use of the land (Whon *et al.*, 2012). Data was collected over a period of 6 months using a connector-linked direct precipitation airsampler and a series of filters to select particles under 1µm (Whon *et al.*, 2012). Using Phi29 polymerase, they selectively amplified ssDNA and dsDNA viral portions for Roche 454 pyrosequencing sequencing (Whon *et al.*, 2012). The ssDNA fraction (from most abundant to least abundant) was: gemivirus-related, circovirus-related, microphage-related, and nanovirus-related (Whon *et al.*, 2012). Of the geminivirus-related ssDNA viruses, most were related to *Sclerotinia sclerotiorum hypovirulence-associated DNA virus* 1 (SsHADV-1) (Whon *et al.*, 2012; Yu *et al.*, 2010). Recently it has been proposed that these geminivirus-related ssDNA viruses might be part of a new genus of viruses that likely infect fungi, named gemycircularvirus (Rosario *et al.*, 2012a). It is also notable that the eukaryotic ssDNA virus portion in all of the air virome samples was higher than the ssDNA microphage portion which contrasts with studies on other types of environmental samples (Angly *et al.*, 2006; Kim *et al.*, 2008; Rosario *et al.*, 2009b; Roux *et al.*, 2012a) suggesting the uniqueness of the air virome (Whon *et al.*, 2012).

1.4.1.4 Invertebrates

Invertebrates accumulate diverse viruses from their niche environment, with their power for accumulation dependant on their feeding sources and mobility range. Using dragonflies, Rosario *et al.* (2011) isolated ssDNA viruses with a technique incorporating amplification of ssDNA by Phi29 polymerase with random hexamers and fragmented the viral DNA before cloning and sequencing. Twenty-one novel ssDNA cyclovirus genomes sharing >95% amino acid identify were isolated from 12 dragonfly specimens from the Kingdom of Tonga (Rosario *et al.*, 2011). In a follow-up study by Rosario *et al.* (2012a), 17 novel ssDNA viruses were discovered from the tissue of dragonflies collected from a variety of locations including Tonga, Puerto Rico, Florida and Florida Keys, USA, Bulgaria, Austria, Finland, Hungary, Germany, and Finland. Most of the novel ssDNA viruses discovered were determined to be cycloviruses (Rosario *et al.*, 2012a). Three of the novel isolates had Rep sequence similarity to circoviruses but the genome organizations were different from both cycloviruses and circoviruses (Rosario *et al.*, 2012a). Three other novel isolates had Rep similarity to geminiviruses, but with genome organizations and nonanucleotide motifs like that of the mycovirus SsHADV-1 (Rosario *et al.*, 2012a; Yu *et al.*, 2010). These were determined to be part of a new genus to be named

Gemycircularvirus (Rosario *et al.*, 2012a). Another putative cyclovirus was isolated from a Florida woods cockroach from Florida, USA, indicating that circoviruses have a wider range of hosts than previously thought (Padilla-Rodriguez *et al.*, 2013).

Top-end insect predators are at the apex of viral accumulation since the insects that they prey on have also been accumulating viruses from diverse environments and hosts. For environmental sampling this approach has potential to save time and lab resources. For example, two novel plant-infecting virus molecules (a mastrevirus and an alphasatellite) were discovered in dragonflies in Puerto Rico, and yet they had not appeared in the 60 plant tissue samples tested from the same region (Rosario *et al.*, 2013). The discoveries were particularly remarkable since the alphasatellite isolate likely represents a new lineage of alphasatellites in the New World, while the highly divergent mastrevirus isolate was one of only two mastreviruses found in the New World (Rosario *et al.*, 2013). The main disadvantage to invertebrate sampling is that it does not reveal the source of the infection. For example, in Rosario *et al.* (2013) no begomovirus samples were found in the region around the site of the alphasatellite discovery, though alphasatellites are usually associated with begomoviruses because they are incomplete on their own and require another helper virus. The novel mastrevirus that was found coinfecting the dragonfly sample could potentially be the helper virus, though this has never been documented before (Rosario *et al.*, 2013).

Broad viral diversity has been observed in whiteflies (Ng *et al.*, 2011a) and mosquitoes (Ng *et al.*, 2011b). The ssDNA portion was preferentially amplified before Roche 454 pyrosequencing (Ng *et al.*, 2011a; Ng *et al.*, 2011b). In whitefly samples, the majority of sequences had high nucleotide identity to known geminiviruses (Ng *et al.*, 2011a). In mosquito samples, the assembled sequences observed included those infecting animals, plants, and phage, and the majority of the sequences likely represent novel genomes (Ng *et al.*, 2011b). Additionally, some full genomes were verified using primer design and PCR (Ng *et al.*, 2011b). Two of the novel circular ssDNA viruses had unique genome organizations (Ng *et al.*, 2011b), one of which has been shown to group with other putative gemycircularviruses through phylogenetic analysis (Rosario *et al.*, 2012a).

In both Dayaram *et al.* (2013) and Dunlap *et al.* (2013), viruses have been isolated from the tissue of aquatic invertebrates. A novel ssDNA virus with sequence similarity to bacterial Rep-like sequences was isolated from a mollusc of the Avon-Heathcote Estuary in Christchurch, New Zealand using amplification by Phi29 polymerase with random hexamers, restriction enzyme digest of the amplified DNA, and selective cloning and sequencing (Dayaram *et al.*, 2013). Two novel circovirus-like viruses were isolated from copepods, a type of mesozooplankton, from Tampa Bay, Florida, using amplification by Phi29 polymerase with random hexamers and fragmenting viral DNA before cloning and sequencing (Dunlap *et al.*, 2013).

1.4.1.5 Sewage

Sampling from sewage systems essentially samples from a large number of sources, including the fact that sewage provides a fertile environment for a variety of viral hosts (Cantalupo *et al.*, 2011). Metagenomic analysis of DNA and RNA viruses was carried out on sewage samples from urban cities: Barcelona, Spain; Pittsburgh, Pennsylvania, United States; and Addis Ababa, Ethiopia (Cantalupo *et al.*, 2011); and in another similar study: Maiduguri, Nigeria; San Francisco, California, United States; Bangkok, Thailand; and Kathmandu, Nepal (Ng *et al.*, 2012). In both studies the metagenomic technique incorporated Roche 454 pyrosequencing after random PCR amplification of DNA and cDNA libraries. In Cantalupo *et al.* (2011) of 897,647 high-quality reads, the largest proportion 596,146 of the acquired sequence reads were determined to be novel, while the second largest proportion of reads (37,917) were bacteriophage, and the smallest proportion (8,491) were eukaryotic viruses. In both studies, a diversity of viruses were observed including those from ssDNA families, dsDNA families, (+) ssRNA families, and dsRNA families as well as novel genomes representing new families and new genera. Specifically, in Ng *et al.* (2012) novel circular ssDNA genomes with some similarities to members of *Geminiviridae* were identified, Nimivirus, Niminivirus, and Baminivirus.

In Blinkova *et al.* (2009), nested PCR was the main method employed for sewage samples from several coastal sewage facilities in the United States: Louisiana, Maryland, California, Washington, Oregon, Maine, North Carolina, New Jersey, Alabama, Wisconsin, and two sites in Florida. Nested PCR was based on sequences of newly characterised viruses of Cardiovirus, Cosavirus, Bocavirus, and Circovirus genera and resulted in a greater genetic diversity within each and a broader geographic distribution than previously known (Blinkova *et al.*, 2009).

1.4.1.6 Faeces

Faecal sampling uses the viral reservoir concept that has been demonstrated using invertebrates, though animal viral reservoirs also have the benefit of a longer life, therefore longer collection period, and a higher position on the food chain and therefore a broader scope. Additionally, the scope for viral discovery goes beyond sampling the animal viral reservoir and includes the viral communities present in faecal matter because of the opportunistic organisms that inhabit it. This includes bacteria, protozoa, and fungi as well as insect larvae and other invertebrates depending on how soon the faecal matter was collected (Delwart & Li, 2012). Faecal sampling is also ideal for a low-impact approach to investigating the viral diversity of an environment because it is non-invasive. Other viral discovery approaches that have used smaller viral reservoirs, such as sampling insects or molluscs, have required removal of the entire organism from the environment, while animal sampling does not. In contrast to sampling animal tissue or serum, faecal sampling requires no animal handling and samples can be retrieved after the animal has departed. Faecal sampling has been used for the discovery of diverse ssDNA viruses. Specifically, novel ssDNA viruses have been isolated from chimpanzee faeces (Blinkova *et al.*, 2010; Li *et al.*, 2010a), bovine faeces (Kim *et al.*, 2012), rodent faeces (Phan *et al.*, 2011), bat faeces (Ge *et al.*, 2011; Li *et al.*, 2010), badger and pine marten faeces (van den Brand *et al.*, 2012) and pig faeces (Sachsenröder *et al.*, 2012; Shan *et al.*, 2011).

In a leading study by Blinkova *et al.* (2010), random PCR amplification and metagenomic sequencing was used on viral DNA extracted from chimpanzee stool samples from Cameroon, Central African Republic, the Democratic Republic of the Congo, the Republic of the Congo, Tanzania, Uganda, and Rwanda. Three novel ssDNA viruses with BLASTx similarities to the Reps of circoviruses and nanoviruses were discovered and were named chimpanzee stool-associated circular viruses (ChiSCVs) (Blinkova *et al.*, 2010). They exhibited a unique ORF orientation with two major ORFs reading ambidirectionally towards the stem-loop element (Blinkova *et al.*, 2010). Nested PCR was also used further to investigate the presence of

ChiSCV-like viruses in the samples (Blinkova *et al.*, 2010). Seven ChiSCV genomes were determined by inverse PCR and characterised (Blinkova *et al.*, 2010). All ChiSCV genomes exhibited RCR motifs associated with other circular ssDNA viruses, but phylogenetic analysis of the Reps of ChiSCVs and representatives of geminiviruses, nanoviruses, and circoviruses grouped them into an independent clade (Blinkova *et al.*, 2010).

Li *et al.* (2010) also isolated novel ssDNA viruses from faecal samples of chimpanzees and humans, but through degenerate PCR based on sequences of circoviruses and CyCV1-PK5006, a unique circular ssDNA viral isolate that was previously discovered in a human faecal sample from Pakistan (Victoria *et al.*, 2009). The novel ssDNA viruses had ORF orientation similar to that of other circoviruses (ambidirectional ORFs, reading towards the putative stem-loop element) but they grouped into a distinct clade (Li *et al.*, 2010). The authors proposed they belonged to a new genus, Cyclovirus, to be included in the *Circoviridae* family (Li *et al.*, 2010). In the study cycloviruses were also found in human plasma and animal tissue (Victoria *et al.*, 2009), highlighting the value of initial discovery through faecal sampling.

Viral nucleic acid in pig faecal samples from a high density farm in North Carolina was amplified using random PCR, pyrosequenced, and sequence reads were assembled into contigs (Shan *et al.*, 2011). Thirteen percent of sequence reads had no BLASTx similarity, and 1% of the sequence reads had BLASTx matches to viruses in the ssDNA families *Circoviridae* or *Parvoviridae*, including those of high similarity and low similarity (Shan *et al.*, 2011). Four novel ssDNA viruses identified from the contigs as having similarities to circovirus Reps were pulled out with inverse PCR and characterised (Shan *et al.*, 2011). These novel genomes, named porcine circovirus-like viruses (Po-circo-like viruses) contained RCR motifs and N-terminal and P-loop domains associated with other ssDNA viruses (Shan *et al.*, 2011). They exhibited genome organizations different to that of previously discovered circovirus-like viruses (ChiSCVs) and cycloviruses (Blinkova *et al.*, 2010; Li *et al.*, 2010) with more than two ORFs arranged either unidirectionally or ambidirectionally transcribed away from the putative stem-loop element and phylogenetic analysis placed then in an independent clade (Shan *et al.*, 2011).

A study on rodent faecal samples from around California and Virginia used similar methods (Phan *et al.*, 2011). Forty-five per cent of sequence reads had no BLASTx hits, though a large proportion of the sequence matches to DNA viruses were to circoviruses (Phan *et al.*, 2011). Several novel ssDNA viruses were pulled out with inverse PCR, characterised, and described with several different types of genome organizations (Phan *et al.*, 2011). Phylogenetic analysis of the Reps with representative species of ssDNA families and other novel circovirus-like viruses revealed low relation to known viruses and the formation of several new clades, further attesting to the poor representation of ssDNA diversity available from the presently known ssDNA viral sequences (Phan *et al.*, 2011).

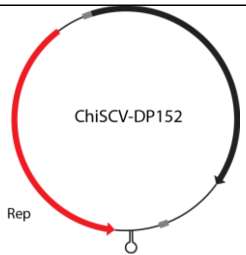
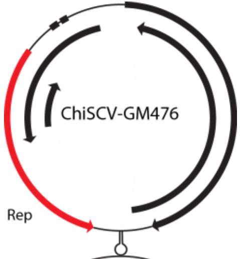
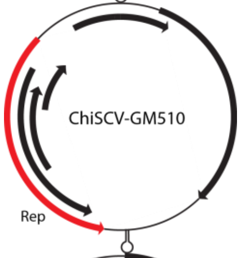
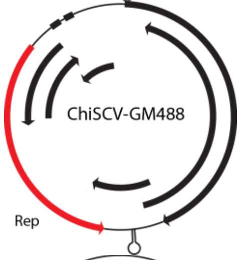
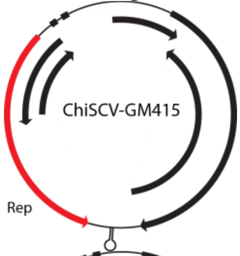
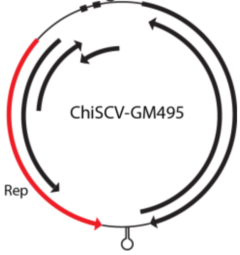
Viral nucleic acid in bat faecal samples from China was amplified using random PCR, sequenced by Solexa sequencing sequencing, and assembled into contigs (Ge *et al.*, 2012). Primers based on similarities to known circoviruses or circovirus-like viruses were also used on the amplified nucleic acid, and full consensus sequences were assembled using the PCR sequences and the previously assembled contigs (Ge *et al.*, 2011). Five full genomes were characterised from the data and four of these had similar genome organizations to other cycloviruses and grouped with cycloviruses during phylogenetic analysis of the Rep sequences of the novel viruses with representative circoviruses and cycloviruses, while the other novel genome was highly divergent and grouped into a monophyletic clade (Ge *et al.*, 2011). An additional 17 partial sequences representing putative novel ssDNA circovirus-like viruses were discovered through degenerate PCR based on the five novel sequences and other circovirus sequences, and phylogenetic analysis of partial Reps with other representative species from cycloviruses and circoviruses including the novel genomes from this study revealed five distinct groups within the circoviruses clade (Ge *et al.*, 2011).

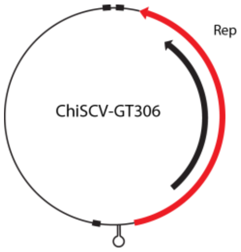
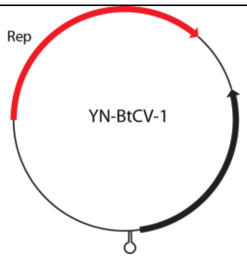
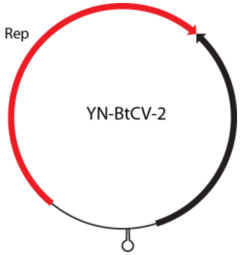
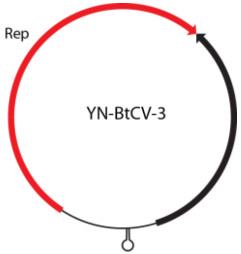
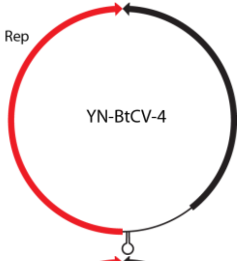
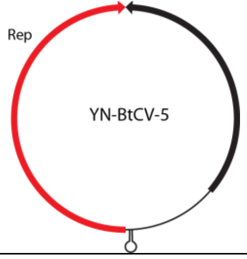
Bovine faecal samples from Korea were used with a PAN-PCR technique for amplification (Kim *et al.*, 2012). Fragments were cloned and sequenced, and contigs were assembled from the sequencing data (Kim *et al.*, 2012). Putative novel circular ssDNA virus, bovine stool associated circular DNA virus (BoSCV), was identified in the contigs, pulled out with inverse PCR, and characterised (Kim *et al.*, 2012). The genome organisation was similar to that of ChiSCVs with two major ORFs reading towards the putative stem-loop element (Kim *et al.*, 2012). Phylogenetic analysis of the Rep with representative species from *Circoviridae*, *Nanoviridae*, and *Geminiviridae* showed BoSCV clustering with ChiSCVs. A set of primers were designed based on BoSCV and degenerate PCR revealed similar sequences in other bovine and pig faecal samples (Kim *et al.*, 2012).

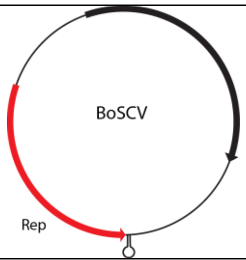
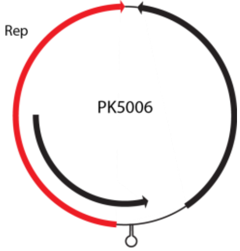
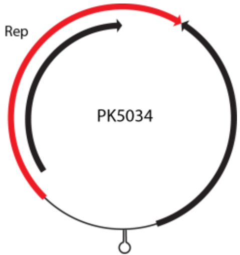
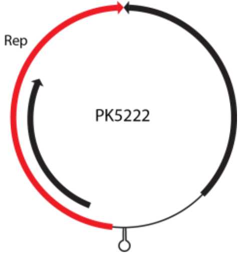
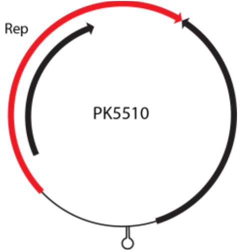
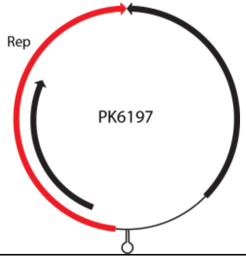
Both random PCR and Phi29 polymerase with random hexamers were used for amplification of viral nucleic acid from European badger and pine marten faecal samples before Roche 454 pyrosequencing and the contigs were assembled from sequencing reads (van den Brand *et al.*, 2012). Two putative novel ssDNA viruses from the contigs reads were pulled out with inverse PCR and characterised, both pertaining to the European (*Meles meles*) badger sample (van den Brand *et al.*, 2012). One of the novel isolates had similarity to other circoviruses and was named *Meles meles* circovirus-like virus (MmCVLV), and the other had similarity to geminiviruses and the fungal mycovirus SsHADV-1 (van den Brand *et al.*, 2012; Yu *et al.*, 2010) and was named *Meles meles* faecal virus (MmFV). The author proposed a new family for MmFV and SsHADV-1 (Yu *et al.*, 2010), *Briviviridae* (van den Brand *et al.*, 2012), but later analysis by Rosario *et al.* (2012a) after discoveries of similar viral isolates in dragonflies, cassava leaves (Dayaram *et al.*, 2012), and mosquitoes (Ng *et al.*, 2011b), suggested it belong to a new genus, Gemycircularvirus, within the *Geminiviridae* family.

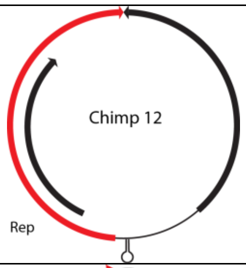
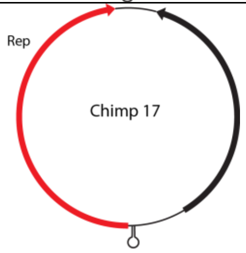
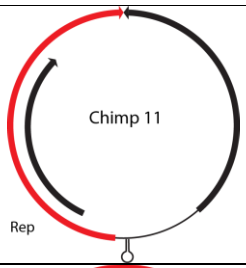
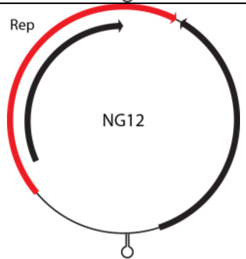
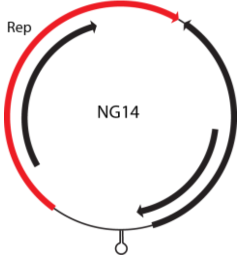
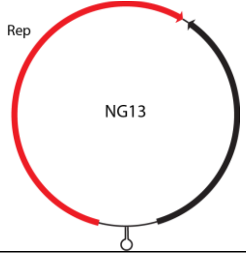
Full genome amplification was used on viral nucleic acid from pig faecal samples collected at an experimental animal facility in Germany (Sachsenröder *et al.*, 2012). The viral families/groups that were most abundant were *Microviridae* at 45.9%, ChiSCV-like viruses at 12%, and then *Picornaviridae* at 10.8% (Sachsenröder *et al.*, 2012). A novel ssDNA virus was pulled out with inverse PCR, characterised, and named pig stool-associated circular ssDNA virus (PigSCV) (Sachsenröder *et al.*, 2012). Sequence similarity was closest to ChiSCVs, though genome organization was dissimilar; Genome organization of PigSCV is unidirectional with the stem-loop element located on the same strand (Sachsenröder *et al.*, 2012). Phylogenetic analysis of the Rep with other ssDNA representative species showed PigSCV is related to the ChiSCVs but is on an distinct branch (Sachsenröder *et al.*, 2012). PCR was then used to positively detect PigSCV in other pig faecal samples from the same facility (Sachsenröder *et al.*, 2012).

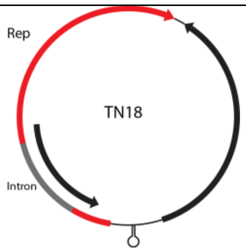
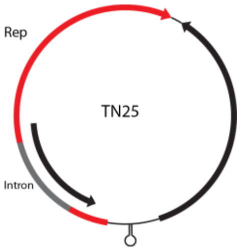
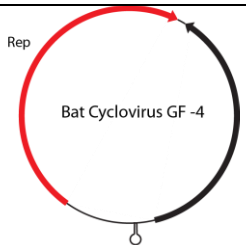
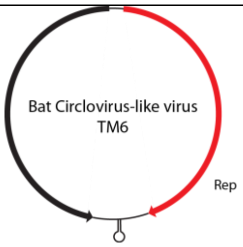
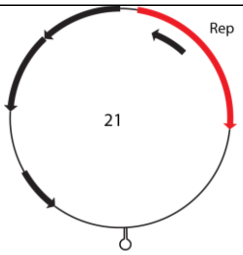
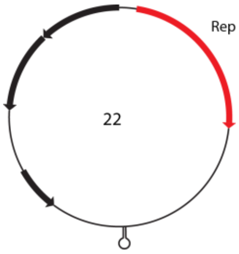
Table 1.2 Description of the ssDNA virus genomes isolated from faecal matter sources.

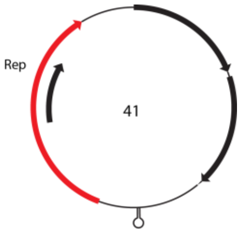
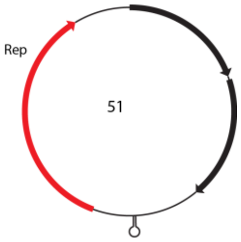
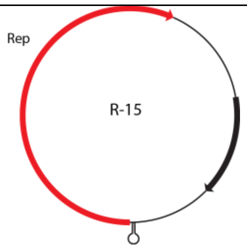
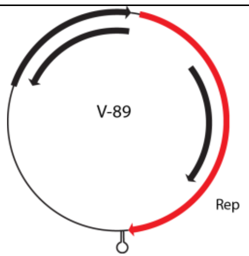
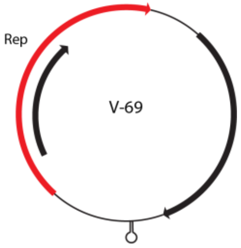
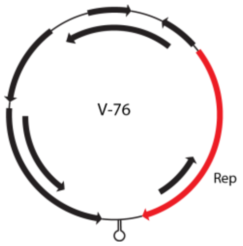
Source	Reference	GenBank Accession Number	Species Name	Genome	Genome Type	Genome Size
Chimpanzee faeces from Tanzania	(Blinkova *et al.*, 2010)	GQ351272	ChiSCV-DP152			2609 nt
		GQ351274	ChiSCV-GM476		IV	2637 nt
		GQ351275	ChiSCV-GM510		III	2589 nt
		GQ351276	ChiSCV-GM488			2638 nt
		GQ351277	ChiSCV-GM415			2639 nt
		GQ351273	ChiSCV-GM495			2640 nt

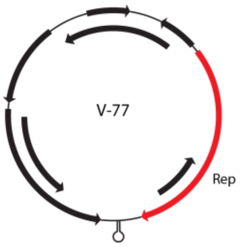
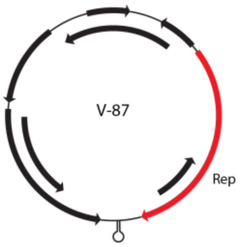
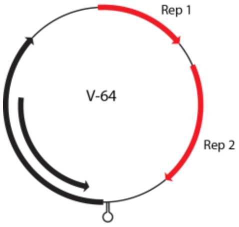
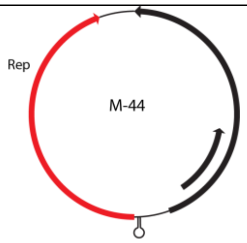
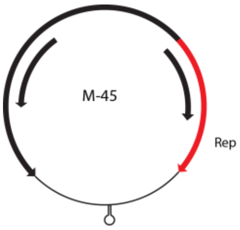
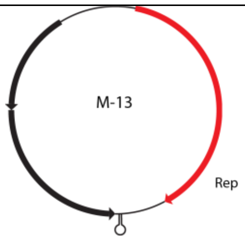
		GQ351278	ChiSCV-GT306		VIII	1198 nt
Bat faeces from China	(Ge *et al.*, 2011)	JF938078	YN-BtCV-1			2177 nt
		JF938079	YN-BtCV-2			1771 nt
		JF938080	YN-BtCV-3			1743 nt
		JF938081	YN-BtCV-4			1741 nt
		JF938082	YN-BtCV-5			1818 nt

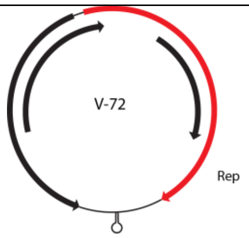
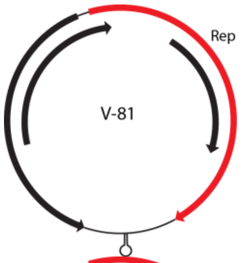
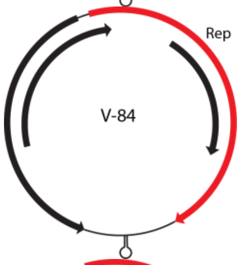
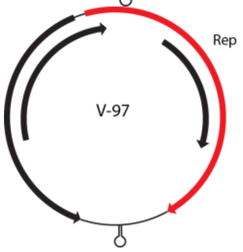
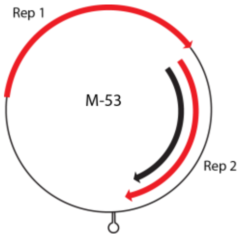
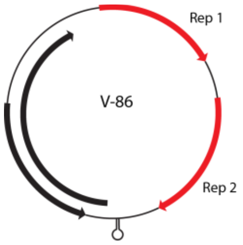
Bovine faeces from Korea	(Kim *et al.*, 2012)	JN634851	BoSCV CP11-49-3			2600 nt
Human faeces from Pakistan	(Li *et al.*, 2010)	GQ404844	CyCV-PK5006		II	1723 nt
		GQ404845	CyCV-PK5034		II	1780 nt
		GQ404846	CyCV-PK5222		II	1740 nt
		GQ404847	CyCV-PK5510		II	1759 nt
		GQ404848	CyCV-PK6197		II	1741 nt

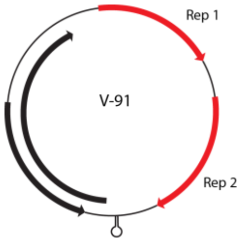
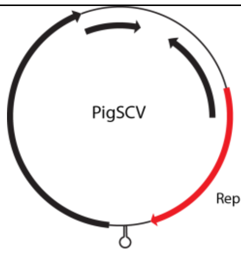
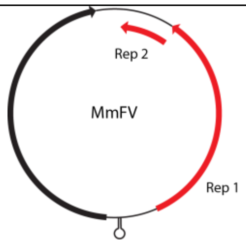
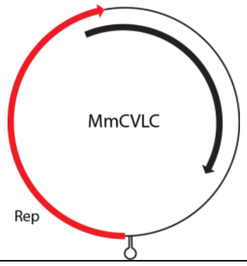
Chimpanzee faeces from Tanzania	(Li *et al.*, 2010)	GQ404850	CyCV-Chimp12		1747 nt
Chimpanzee faeces from Rwanda	(Li *et al.*, 2010)	GQ404851	CsaCV-chimp17		1935 nt
Chimpanzee faeces from Tanzania	(Li *et al.*, 2010)	GQ404849	CyCV-Chimp11		1750 nt
Human faeces from Nigeria	(Li *et al.*, 2010)	GQ404854	CyCV-NG12		1794 nt
		GQ404855	CyCV-NG14		1795 nt
		GQ404856	CyCV-NG13		1699 nt

Human faeces from Tunisia	(Li *et al.*, 2010)	GQ404858	CyCV-TN18			1867 nt
		GQ404857	CyCV-TN25			1867 nt
Bat faeces from California	(Li *et al.*, 2010)	HM228874	CyCV-GF4			1844 nt
Bat faeces from Texas	(Li *et al.*, 2010)	HM228876	bat circovirus-like virus TM6		IV	1696 nt
Pig faeces from North Carolina	(Shan *et al.*, 2011)	JF713716	po-circo-like virus 21			3912 nt
		JF713717	po-circo-like virus 22			3926 nt

		JF713718	po-circo-like virus 41		V	2904 nt
		JF713719	po-circo-like virus 51			2833 nt
Woodrat faeces from California	(Phan *et al.*, 2011)	JF755401	RodSCV-R-15		VII	1758 nt
Meadow vole faeces from Virginia	(Phan *et al.*, 2011)	JF755402	RodSCV-V-89			2069 nt
		JF755403	RodSCV-V-69		VI	2220 nt
		JF755404	RodSCV-V-76			3781 nt

		JF755405	RodSCV-V-77			3780 nt
		JF755406	RodSCV-V-87		I	3787 nt
		JF755407	RodSCV-V-64			2986 nt
House mouse faeces from Virginia	(Phan *et al.*, 2011)	JF755408	RodSCV-M-44			2294 nt
		JF755409	RodSCV-M-45		IV	2506nt
Pinyon mouse faeces from California	(Phan *et al.*, 2011)	JF755410	RodSCV-M-13		IV	2193nt

Meadow vole faeces from Virginia	(Phan *et al.*, 2011)	JF755411	RodSCV-V-72			2070 nt
		JF755412	RodSCV-V-81			2070 nt
		JF755413	RodSCV-V-84			2070 nt
		JF755414	RodSCV-V-97			2070 nt
		JF755415	RodSCV-M-53		VII	1124 nt
		JF755416	RodSCV-V-86		VI	2984 nt

		JF755417	RodSCV-V-91		VI	2984 nt
Pig faeces from Germany	(Sachsenröder *et al.*, 2012)	JQ023166	PigSCV			2459 nt
European badger faeces from The Netherlands	(van den Brand *et al.*, 2012)	JN704610	MmFV, *Meles meles* faecal virus			2,199 nt
		JQ085285	MmCVLV, *Meles meles* circovirus-like virus		III	2,218 nt

1.4.2 *Characterisation of novel ssDNA viruses*

Novel ssDNA viruses can be isolated from a sample using sequence-specific or sequence-independent techniques. Sequence-specific techniques (such as panviral microarrays and PCR with specific primers or degenerate primers) are limited in their ability to discover highly divergent novel ssDNA viruses. This is because microarrays rely on specific hybridization and PCR necessitates prior sequence knowledge for primer or degenerate primer set design. However, after novel ssDNA virus discovery through other techniques, sequence-specific PCR, in the form of inverse PCR, is essential for verifying the full genome sequence.

Novel viruses with sequence similarities to known viruses can be identified through the use of degenerate primers. In degenerate primer PCR, a set of degenerate primers are designed based on a virus sequence, conserved motif, or group of related sequences. Degenerate primers designed around conserved ssDNA virus motifs have led to the discovery of novel ssDNA viruses in dragonflies (Rosario *et al.*, 2012a) and human and animal faecal matter, tissue, serum (Li *et al.*, 2010). The main disadvantage of degenerate primer PCR is that it will not detect highly divergent genomes and each degenerate primer set will be specific to a certain group of viruses. In both studies, the degenerate primer PCR technique was used in conjunction with other metagenomic studies such as RCA with Phi29 DNA polymerase, restriction enzyme digest, cloning, and Sanger sequencing (Rosario *et al.*, 2012a), and for the other, random PCR and Roche 454 pyrosequencing (Li *et al.*, 2010).

Random PCR is a sequence-independent amplification technique that uses primers with a random sequence. The technique is used to amplify all nucleic acid present in a sample for the purpose of creating libraries for metagenomic sequencing. To amplify both DNA and RNA, the RNA must first be converted to cDNA in a separate experiment. Random PCR has been used to prepare libraries for total DNA and RNA nucleic acid analysis using Roche 454 pyrosequencing in such environmental samples as human faeces (Li *et al.*, 2010; Victoria *et al.*, 2009) sewage (Cantalupo *et al.*, 2011; Ng *et al.*, 2012) pig faeces (Shan *et al.*, 2011) rodent faeces (Phan *et al.*, 2011) European badger and pine martin faeces (van den Brand *et al.*, 2012) chimpanzee faeces (Li *et al.*, 2010), farm animal tissue (Li *et al.*, 2010), Sea lion tissue (Shan *et al.*, 2011), bat faecal matter, serum, and tissue (Donaldson *et al.*, 2010; Ge *et al.*, 2011; Li *et al.*, 2010) and using Solexa/Illumina pyrosequencing in bat faecal matter (Ge *et al.*, 2012). The two main types of next generation sequencing are Roche 454 pyrosequencing and Illumina/Solexa pyrosequencing. Roche 454 is the most successful method for viral metagenomics, possibly because of the longer sequence reads (Ge *et al.*, 2012). A metagenomic study on bats using Solexa pyrosequencing had high amounts of non-virus sequence data that may have been due to short sequence reads (Ge *et al.*, 2012); Residual RNA/DNA fragments after incomplete RNase and DNase breakdown during viral purification are not easily distinguished when all sequence reads are shorter. Random PCR has also been used to amplify nucleic acid prior to cloning and Sanger sequencing (Blinkova *et al.*, 2009; Blinkova *et al.*, 2010).

Sequence-independent rolling circle amplification (RCA) has been significantly valuable in the discovery of highly divergent novel ssDNA viruses. RCA is accomplished through the use of bacteriophage Phi29 DNA polymerase from *Bacillus subtilis* (Blanco *et al.*, 1989). It has been used in lab experiments for amplifying circular DNA templates. Bacteriophage Phi29 DNA polymerase is a high fidelity enzyme that has proofreading capabilities and high strand displacement that enables it to displace annealing strands before the 3' end proceeds (Esteban *et al.,* 1993). One bias to this technique is that it preferentially amplifies circular ssDNA templates (Edwards & Rohwer, 2005; Kim *et al.*, 2008; Kim *et al.*, 2011), but in the discovery of novel ssDNA viruses this can be an asset as most known ssDNA viruses are circular. For virus detection, RCA with bacteriophage Phi29 DNA polymerase was more comprehensive than PCR with specific primers; DNA-B components were detected in infected sample tissue using RCA with Phi29 DNA polymerase followed by digest with restriction enzymes when they were previously not detected with PCR, possibly because of low concentration (Inoue-Nagata *et al.*, 2004). Bacteriophage Phi29 DNA polymerase is able to overcome the issue of low copy number due to an exponential amplification rate of up to 10^7 (Nelson *et al.*, 2002). It does this through the use of random hexamers that bind to the template viral DNA at multiple locations and create several replication forks (Nelson *et al.*, 2002). The random hexamers also circumvent the need for prior knowledge of the sequence. This advantage was demonstrated in the amplification and isolation of a novel papillomavirus from infected bovine tissue (Rector *et al.*, 2004). Since these discoveries, RCA with bacteriophage Phi29 DNA followed by restriction enzyme digests, subcloning and sequencing has been used to amplify viral DNA for virus discovery (Dayaram *et al.*, 2012; Piasecki *et al.*, 2012; Rosario *et al.*, 2011; Shepherd *et al.*, 2008; Varsani *et al.*, 2011). This has been valuable for discovery of a wide range of novel ssDNA viruses from environmental samples such as dragonflies (Rosario *et al.*, 2012a) and cassava (Dayaram *et al.*, 2012). The main disadvantage to using this method of amplification is that is biased to circular ssDNA and there is a possibility for the formation of chimeras (Lasken & Stockwell, 2007).

RCA with bacteriophage Phi29 DNA polymerase can also be used to enrich the proportion of ssDNA in a sample before metagenomic sequencing (Haible *et al.*, 2006; Kim *et al.*, 2008 2006). This is ideal for the discovery of novel ssDNA because most ssDNA viruses have circular DNA templates which will be preferentially amplified above the other nucleic acid present. This method is simpler: it does not require the generation of random PCR primers. RCA with bacteriophage Phi29 DNA polymerase to enrich ssDNA before Roche 454 pyrosequencing or Solexa sequencing has been used for environmental samples such as fresh water (Fancello *et al.*, 2012; Roux *et al.*, 2012a), invertebrates (Ng *et al.*, 2011a; Ng *et al.*, 2011b), human faeces (Kim *et al.*, 2011; Minot *et al.*, 2011), sea turtle tissue (Ng *et al.*, 2009a), and rice paddy soil (Kim *et al.*, 2008).

The ssDNA enrichment step can also be followed by fragmentation, cloning and Sanger sequencing, and has been applied for screening viruses in: dragonflies (Rosario *et al.*, 2011), sea lions (Ng *et al.*, 2009b), and sea turtles (Ng *et al.*, 2009a). Fragmentation of RCA amplified viral DNA is useful in cases where knowledge about the types of sequences is not available such as relative size or restriction site(s). Like with NGS sequencing methods, contigs must be assembled from the sequencing data.

Viruses have been discovered in water (Rosario *et al.*, 2009a; Rosario *et al.*, 2009b; Roux *et al.*, 2012a), air (Whon *et al.*, 2012), and soil (Kim *et al.*, 2008; Yoshida *et al.*, 2013), substantiating that they are, in fact, everywhere. At present, they are not well understood; the current ssDNA virus taxonomy is unable to accommodate many novel viruses identified through non-specific amplification techniques. Methods that incorporate the latest metagenomic techniques with environmental sampling from different ecosystems and global regions will make it possible to eventually uncover the viral sequence space. To do this, a variety of samples from different regions of the world will need to be explored. Optimal methods should include an efficient sampling technique, such as the use of viral reservoirs, with amplification and sequencing techniques that are suitable for the type of viral nucleic acid being studied.

Chapter 2

Isolation of a novel ssDNA virus from porcine faeces: proof of concept

2.1 Abstract

Faecal matter has been used to investigate the presence of single-stranded (ssDNA) viruses in environments within Africa, North America, Europe, and Eastern Asia using next generation sequencing platforms. However, little is known about the diversity of viruses in the New Zealand environment. As a proof-of-concept study showing the viability of the concept of identifying single stranded DNA viruses in faecal matter within the New Zealand ecosystem, we investigate a novel approach using basic molecular techniques coupled with sanger sequencing. With these methods we recovered and characterised a novel ssDNA virus from porcine faecal matter. The novel ssDNA viral isolate was tentatively named porcine stool-associated circular virus (PoSCV). PoSCV has similarities at a genome level to other faecal viral isolates from pig, cow, and chimpanzees. It has two major open reading frames (ORFs) that are in a bidirectional orientation and two intergenic regions, each containing a putative stem-loop element. The putative capsid protein has <30% similarity to putative capsid proteins of Pig stool-associated single-stranded DNA virus (PigSCV) and Chimpanzee stool associated circular ssDNA virus (ChiSCV). The putative replication associated protein (Rep) has 50% similarity to the Rep of ChiSCV and ~30% similarity to bovine stool associated circular DNA virus (BoSCV). Our method which used simple molecular techniques is feasible for recovery and characterisation of novel ssDNA viruses from other faecal sources across New Zealand, and will contribute to understanding the ssDNA diversity in the New Zealand environment.

This chapter has been published:

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

2.2 Introduction

Viral metagenomic studies have shown that animal faeces contains a high diversity of viruses (Li *et al.*, 2011; Phan *et al.*, 2011) and thus can potentially be used to explore viral diversity in the environment. The first rounds of investigations on viral diversity in animal faecal samples have been within Eastern Asia, Africa, North America, and Europe, and only a few studies have recovered and characterised novel ssDNA viruses. Specifically, novel ssDNA viruses have been isolated in chimpanzee faeces from Tanzania (Blinkova *et al.*, 2010; Li *et al.*, 2010a) and Rwanda (Li *et al.*, 2010a), bovine faeces from Korea (Kim *et al.*, 2012), rodent faeces from the United States (Phan *et al.*, 2011), bat faeces from China (Ge *et al.*, 2011) and the United States (Li *et al.*, 2010b), badger faeces from the Netherlands (van den Brand *et al.*, 2012), and pig faeces from the United States (Shan *et al.*, 2011) and Germany (Sachsenröder *et al.*, 2012). Pig faeces has been a successful medium for exploration of novel ssDNA viruses. Sachsenröder *et al.* (2012) investigated the total viral nucleic acid in pig faecal samples and determined that 36% of all assembled contigs represented putative ssDNA viruses, and the low identity of the top tBLASTx scores show the divergent nature. In particular, tBLASTx analysis of the ssDNA contigs showed that 9% had matches to microviruses with identity of 54% based on highest E-value scores, 3.3% had matches to circoviruses with identity of 45%, and 0.83% had matches to ChiSCV-like viruses with identity of 68% (Sachsenröder *et al.*, 2012). All these methods have used a next generation sequencing platform for their investigations and have not necessarily had the genomes from the assemblies of the small contigs from these runs verified. Further, assemblies of next generation sequencing methods result in a consensus viral genome. This means that viruses with 5% diversity can easily be averaged out to yield a consensus genome, however, this is not a true viral genome that exists within that population but an in silico generated virus representing a population.

New Zealand has a unique biota due to its Gondwanan ancestry and geographic isolation for the past 80 million years (Cooper & Millener, 1993) and likely possesses a unique viral sequence space. No studies that have recovered and characterised novel ssDNA viruses from animal faecal matter have been conducted either in New Zealand or anywhere in Australasia. The ssDNA virus exploration in Australasia (including the Pacific region) has mainly focused on known pathogens such as *Banana bunchy top virus* (BBTV) (Stainton *et al.*, 2012), *Beak and feather disease virus* (BFDV) (Julian *et al.*, 2012; Massaro *et al.*, 2012; Ortiz-Catedral *et al.*, 2010; Ortiz-Catedral *et al.*, 2009; Varsani *et al.*, 2011), *Porcine circovirus* (PCV) (Garkavenko *et al.*, 2005), various novel ssDNA viruses isolated from dragonflies (Rosario et al., 2011, 2012), gastropod associated circular ssDNA virus (GaCV) (Dayaram et al., 2013) and known geminiviruses (Briddon *et al.*, 2010; Geering *et al.*, 2012; Hadfield *et al.*, 2011; Hadfield *et al.*, 2012; Kraberger *et al.*, 2012).

In order to better understand the distribution and diversity of ssDNA viruses in New Zealand, a protocol for viral particle purification from faecal samples using High Pure Viral Nucleic Acid Kit (Roche, USA) and DNA amplification using rolling circle amplification (RCA) via bacteriophage Phi29 DNA polymerase (Illustra^{TM} TempliPhi Kit, GE Healthcare) was tested. Described here is the protocol applied to a porcine faecal sample taken from the Cass Basin in New Zealand for a proof-of-concept. Through this method a novel ssDNA virus was recovered, tentatively named porcine stool-associated circular virus (PoSCV) and characterised from the sample.

2.3 Methods and materials

2.3.1 *Sample collection and viral particle purification*

A porcine faecal sample was collected from a domestic pig from a farm in the mid-Waimakariri intermontane river basin in the central South Island of New Zealand. The sample was stored at 4° C until processing. Approximately five grams of the faecal sample was added to 5 ml of SM buffer (0.1M NaCl, 50 mM Tris/HCI - pH 7.4, 10 mM $MgSO_4$) and vortexed to homogenize the sample. The homogenate was centrifuged at low speed centrifugation at 10,000 rpm for 10 minutes in an Eppendorf 5424R bench top centrifuge. The supernatant was recovered with a syringe and filtered using a 0.45 µm and then a 0.2 µm syringe filter (Sartorius Stedim Biotech, Germany). 200 µl of the filtrate was used for viral nucleic acid purification using the High Pure Viral Nucleic Acid Kit (Roche, USA) according to the manufacturer's isolation protocol for serum, plasma, or whole blood.

2.3.2 *Recovery and cloning of viral DNA using rolling circle amplification (RCA)*

The purified viral DNA was enriched for circular DNA using rolling circle amplification (RCA) using the Illustra^{TM} TempliPhi Kit (GE Healthcare) as previously described (Dayaram *et al.*, 2012; Piasecki *et al.*, 2012; Rosario *et al.*, 2011; Shepherd *et al.*, 2008; Varsani *et al.*, 2011). Briefly, 1 µl of viral nucleic acid was added to 5 µl of the manufacturer's provided denature buffer (10 mM Tris-HCl, pH 8.2, 0.5 mMEDTA) and heated in the PCR protocol: 94°C for 2 min, 4°C for 3 min. Following this, 5 µl of the reaction buffer (containing the Phi29 polymerase) was added and the reaction was incubated at 30°C for 16 hr. The RCA amplified

DNA was used in five separate restriction enzyme digest reactions (*Bam*HI, *Kpn*I, *Xmn*I, *Eco*RI, and *Eco*RV; see Table 2.1). In brief, 1 µl of restriction enzyme was used per 20 µl reaction with 1.5 µl RCA amplified DNA and 2 µl of restriction enzyme buffer. The digested DNA was run on a 0.7% agarose gel stained with SYBR® Safe DNA stain (Life Technologies, USA). Resulting fragments between 1.5 kb and 6 kb were excised from the gel, cleaned with Megaquick-spin™ Total Fragment DNA Purification Kit (iNtRON Biotechnology, Inc.) and cloned into corresponding plasmid vectors (see Table 2.1). Plasmids were transformed into *E.* coli strain DH5α using the heat shock method and screened using blue/white selection.

Table 2.1 Restriction Enzyme Digest Guide

Restriction Enzyme	Vector	Primers
***Bam*HI**	pUC19 (Fermentas, USA) digested with *Bam*HI	M13 F / R
***Kpn*I**	pSK digested with *Kpn*I	M13 F / R
***Xmn*I**	pUC19 (Fermentas, USA) digested with *Sma*I	M13 F / R
***Eco*RI**	pUC19 digested with *Eco*RI	M13 F / R
***Eco*RV**	pUC19 (Fermentas, USA) digested with *Sma*I	M13 F / R

Selected colonies were screened by PCR using M13 F /R primers and the resulting amplicons were resolved on a 0.7% agarose gel stained with SYBR® Safe DNA stain (Life Technologies, USA). Positive colonies (with inserts) were used to inoculate a 5ml Luria broth culture and grown overnight at 37°C in a shaking incubator. The plasmids from these overnight cultures were recovered using DNA-spin™ Plasmid DNA Purification Kit (iNtRON Biotechnology, Inc.) and sequenced by primer walking at Macrogen Inc. (Korea) using M13 forward and reverse primers (see Table 2.1).

2.3.3 *Recovery and characterisation of complete genome of novel ssDNA virus*

Based on the BLASTx (Altschul *et al.*, 1990) analysis of the resulting sequence reads, back-to-back primers were designed to be used in inverse PCR with the RCA amplified DNA as a template. The primer pairs as5-F 5'-CGA GCA TTT CCT GAA CCT TGT TAC-3' and as5-R 5'-GCG GAT TCT TCA GAC AGT TCA G-3' were designed based on similarities in the replication associated protein (Rep) gene region. Kapa HiFi HotStart DNA polymerase (Kapa Biosystems, USA) was used with the primers and the RCA template with the following PCR protocol: initial denaturation at 94°C for 2 min, followed by 25 cycles of 94°C for 15 s, 50°C for 15 s, 72°C for 1 min, and finishing with a 2 min extension at 72 °C to recover the full genome of the novel circular DNA virus. The full genomes were cloned into pJET 1.2 vectors (Fermentas, USA) and then transformed into DH5α *E.coli*. The pJET 1.2 vector is a suicide vector; only colonies with an insert should be able to grow on LB agar plates. The colonies were screened by PCR for the insert using pJET F/R primers, prior to inoculation of overnight cultures and recovery of the plasmid DNA. The resulting plasmid was purified using DNA-spin^{TM} Plasmid DNA Purification Kit (iNtRON Biotechnology, Inc.) and sequenced at Macrogen Inc. (Korea) by primer walking.

2.3.4 *Sequence Analysis*

The full genome was assembled from sequence reads using DNAMAN (version 5.2.9; Lynnon Biosoft). The Rep from the full genome sequence was aligned with Reps from all novel ssDNA genome isolates and representative species of circovirus, cyclovirus, geminivirus, and nanovirus (available in GenBank) using MUSCLE (Edgar, 2004) and manually edited in MEGA5 (Tamura *et al.*, 2011). A Maximum Likelihood (ML) phylogenetic tree was generated using the LG model of substitution and PHYML version 3 (Guindon *et al.*, 2010), with approximate likelihood-ratio test (aLRT) branch support (Anisimova & Gascuel, 2006). Branches with less than 60% aLRT support were collapsed with Mesquite (Version 2.75).

2.4 Results and discussion

2.4.1 *Application of method*

The viral particle purification protocol employed for this study was similar to that used on dragonflies to recover ssDNA viruses (Rosario *et al.*, 2011), with a few steps added specifically for viral nucleic acid extraction from faecal matter such as low speed centrifugation and filtration (Blinkova *et al.*, 2010; Li *et al.*, 2010a; Victoria *et al.*, 2009). A porcine faecal sample from the Cass Basin, New Zealand was prepared for viral DNA extraction according to the protocol. Our protocol assumes that there are intact virions shed in faecal matter and that these can be resuspended in a simple buffer by vortexing to indirectly homogenise the sample. Low speed centrifugation separates the larger and denser particles from the mixture, including host cells and single cell organisms, leaving intact virions in solution. Thereafter, particles larger than 0.2 µm in the supernatant are removed by sequential filtration through 0.45 µm and 0.2 µm syringe filters. This removes most bacterial cells and all eukaryotic host cells, but allows viral particles to pass through with the knowledge that most ssDNA virus particles are in the size range of 12-30 nm (King *et al.*, 2011). The main reason for recovering the virus rather working with total nucleic acid is to reduce the ‘background DNA noise’ and to eliminate bacterial plasmids, which, like circular ssDNA viruses, will be preferentially amplified by RCA.

Viral DNA was enriched with the Illustra^{TM} TempliPhi Kit (GE Healthcare) in 10 µl reactions as outlined in Shepherd *et al.* (2008) and successfully employed in other studies of ssDNA viruses (Dayaram *et al.*, 2012; Piasecki *et al.*, 2012; Rosario *et al.*, 2011; Varsani *et al.*, 2011). RCA using bacteriophage Phi29 DNA polymerase with random hexamers is a non-specific amplification technique that does not require prior knowledge about the sequence (Nelson *et al.*, 2002). It can also amplify circular DNA that is in low concentrations (Nelson *et al.*, 2002), thereby increasing the chances of discovering nucleic acid from novel ssDNA viruses.

Bacteriophage Phi29 DNA polymerase with random hexamers yields linear concatemer of nucleic acid, which can then be digested with appropriate restriction enzymes to yield unit length genomes. In this study five different restriction enzymes in separate digestion reactions were used to recover any ssDNA viruses in the range of 1.5-6 kb. The goal was to isolate full viral genomes from the concatemerized DNA. If the restriction site is present at a single location in the genome then that restriction enzyme digest of the concatemerized DNA would yield a full length genome in a linear state. Knowing that ssDNA viruses range from approximately 1.5-6 kb, except those with multiple-component genomes, a fragment of DNA comprised of approximately 2 kb from a *Bam*HI digest was excised from the gel, cloned, and sequenced. A preliminary BLASTx analysis showed that the 2 kb fragment *Bam*HI digest represented a putative ssDNA virus isolate. Cloned products from the restriction with other enzyme sites yielded no significant matches using BLASTx and hence were presumed to be fragments of DNA from either the host or other unknown sources and non-viral in nature.

2.4.2 *Full genome analysis of the novel ssDNA virus isolate*

It was highly possible that the fragment sequenced from the *Bam*HI digest did not represent a complete genome even though it was approximately 2 kb (the size of many ssDNA virus genomes) and exhibited putative Reps and Cps. If two identical restriction sites are located in close proximity to each other it is likely that the secondary smaller fragments may not be observed on a restriction digest gel. Such was the case here; the full genome of PoSCV obtained through inverse PCR was found to be 2589 nt, which was 308 nt larger than the original fragment sequenced from the *Bam*HI digest. It turned out that PoSCV has two *Bam*HI sites located 308 nt apart. This demonstrates why inverse PCR is always necessary to verify or identify the full genome sequence prior to characterisation.

The full genome sequence was analyzed using BLASTn and had highest similarity to Chimpanzee stool associated circular ssDNA virus (ChiSCV) isolate GT306 (GQ351278; 78% identity to over 10% coverage at an E value of 5 x 10^{-10}). Two major ORFs and 5 smaller ORFs less than 200 nt were identified using DNAMAN. The ORFs were analyzed with BLASTp to identify putative protein homologies. The largest ORF was 348 residues and was determined to be the putative capsid protein because it had high hits to ssDNA capsid proteins. The highest BLASTp similarity was to the ChiSCV putative capsid protein (#ADB24817; 34% identity over 85% coverage with an E-value of 4 x 10^{-47}), and some similarity to the putative capsid protein of Pig stool-associated single-stranded DNA virus (PigSCV; #YP_006331068; 26% identity over 97% coverage with an E-value of 2 x 10^{-24}).

The second largest ORF was 243 residues and was determined to be the putative Rep, as the highest BLASTp similarity was to the Rep of ChiSCV (#ADB24799; 51% identity over 88% coverage with and E-value of 4 x 10^{-9}). Some slight similarity was also detected to Bovine associated circular DNA virus spliced Rep (BoSCV; AEW47007; 87% identity over 33% coverage with and E-value of 2 x 10^{-19}), and some alphasatellites of nanoviruses and geminiviruses (30-35% identity over ~50% coverage with and E-value of 4 x 10^{-9} to 3 x 10^{-5}).

The novel ssDNA viral isolate was tentatively named Porcine stool-associated circular virus (PoSCV) and deposited in GenBank under the accession number JX274036. As illustrated in Figure 2.1, the putative Rep and capsid protein (Cp) are bidirectionally transcribed and are separated by two intergenic regions (IR). The long intergenic region (LIR) is 515 nt and the short intergenic region (SIR) is 292 nt. Two putative stem-loop elements, one in each of the IRs have been identified by way of secondary structure analysis of the nucleotide sequences. The nucleotide sequences of these intergeneitc regions are in Figure 2.1.

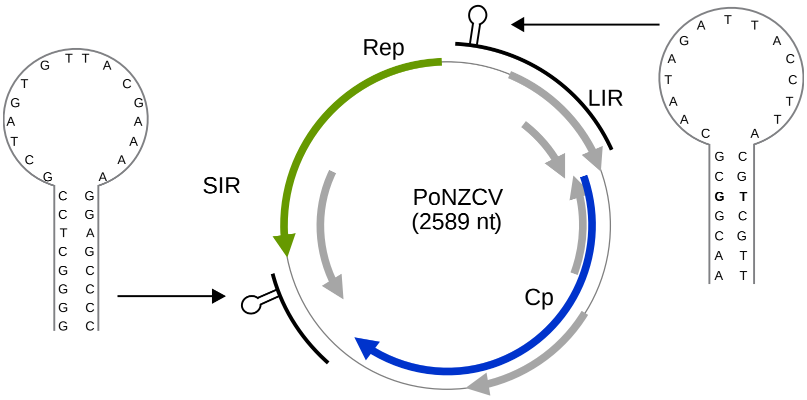


Figure 2.1 The genome of PoSCV. The two major ORFs are the putative coat protein (blue) and the putative Rep protein (green). ORFs less than 200 nt are in grey. Also depicted are the two putative stem-loop elements, one in each IR.

Table 2.2 Rep Motifs of other ssDNA viral isolates from faecal matter sources

					Rep motifs			
Viral isolate	**GenBank**	**Source**	**Motif 1**	**Motif 2**	**Motif 3**	**Walker A**	**Walker B**	**Reference**
PoSCV	JX274036	Pig	MTIPR	HWQIRI	YETKEGQY	ETGNVGKS	TIDTV	This Study
ChiSCV-DP152	GQ351272	Chimpanzee	MTTME	HWQVRC	YETKEGKY	QDGNMGKS	VIDIP	Blinkova et al., 2010
ChiSCV-GM495	GQ351273	Chimpanzee	MTTMQ	HWQVRC	YETKEGKY	EQGNMGKS	VIDIP	Blinkova et al., 2010
ChiSCV-GM476	GQ351274	Chimpanzee	MTTMQ	HWQVRC	YETKEGKY	EQGNMGKS	VIDIP	Blinkova et al., 2010
ChiSCV-GM510	GQ351275	Chimpanzee	MTTMQ	HWQVRC	YETKEGKY	EGGNMGKS	VIDIP	Blinkova et al., 2010
ChiSCV-GM488	GQ351276	Chimpanzee	MTTMQ	HWQVRC	YETKEGKY	EQGNMGKS	VIDIP	Blinkova et al., 2010
ChiSCV-GM415	GQ351277	Chimpanzee	MTTMQ	HWQVRC	YETKEGKY	EQGNMGKS	VIDIP	Blinkova et al., 2010
ChiSCV-GT306	GQ351278	Chimpanzee	-	-	-	QDGNMGKS	VIDIP	Blinkova et al., 2010
PigSCV	JQ023166	Pig	TSIPE	HYQCCI	YCRKTDNY	TIGGTGKS	WIDLP	Sachsenroder, 2012
CyCV-PK5006	GQ404844	Human	LGMTK	HYQFRG	YVYKDRNF	EKGNSGKT	IIDTP	Li et al., 2010a
CyCV-PK5034	GQ404845	Human	FTWND	HLQGFC	YCSKSGIF	GPPGTGKS	IIDDF	Li et al., 2010a
CyCV-PK5222	GQ404846	Human	FTWNN	HLQGFC	YCKKAGHW	GPPGSGKS	IIDDF	Li et al., 2010a
CyCV-PK5510	GQ404847	Human	FTWNN	HLQGFC	YCSKTGIF	GPPGTGKS	IIDDF	Li et al., 2010a
CyCV-PK6197	GQ404848	Human	FTWNN	HLQGFC	YCSKSGEF	GPPGSGKS	IIDDF	Li et al., 2010a
CyCV-Chimp11	GQ404849	Chimpanzee	FTWNN	HLQGFC	YCSKTGIF	GPPGTGKS	IIDDF	Li et al., 2010a
CyCV-Chimp12	GQ404850	Chimpanzee	FTWNN	HLQGYV	YCRKSGIF	GPPGSGKS	IIDDF	Li et al., 2010a
CsaCV-chimp17	GQ404851	Chimpanzee	FTWNN	HLQGYV	YCRKSGIF	GPPGSGKS	IIDDF	Li et al., 2010a
CyCV-NG12	GQ404854	Human	FTLNN	HLQGYL	YCSKEGDV	GPSGVGKS	VIDDF	Li et al., 2010a
CyCV-NG14	GQ404855	Human	FTLNN	HLQGFC	YCSKAGNF	GPTGSGKS	IIDDF	Li et al., 2010a
CyCV-NG13	GQ404856	Human	FTWNN	HLQGFC	YCSKTGNF	GPPGSGKS	IIDDF	Li et al., 2010a
CyCV-TN25	GQ404857	Human	FTLNN	HLQGFF	YCSKSGNI	GPPGCGKS	IIDDF	Li et al., 2010a
CyCV-TN18	GQ404858	Human	WTLNN	HLQGFC	YCSKSGEV	GPTGSGKS	IIDDF	Li et al., 2010a
CyCV-GF4	HM228874	Bat	WTLNN	HLQGFS	YCSKSGEV	GPTGSGKS	IIDDF	Li et al., 2010b
Batcirco-like	HM228875	Bat	FTWNN	HLQGFC	YCSKAGDF	GEPGTGKS	IIDDF	Li et al., 2010b
po-circo-like21	JF713716	Pig	FTLNN	HIQGFI	YCKKSGTF	GESGAGKT	VMDDF	Phan et al., 2011
po-circo-like22	JF713717	Pig	FTINN	HIQGYL	YCTKEDQV	GPPGKGKS	VIDDW	Phan et al., 2011
po-circo-like41	JF713718	Pig	FTINN	HIQGYL	YCTKEDQV	GPPGKGKS	IIDDW	Phan et al., 2011
po-circo-like51	JF713719	Pig	FTINN	HIQGYF	YCSKEGNV	GPPGSGKS	VMDDY	Phan et al., 2011
RodSCV-R-15	JF755401	Rodent	FTINN	HIQGYM	YCSKEDQI	GPAGSGKT	LIDDF	Phan et al., 2011
RodSCV-V-89	JF755402	Rodent	LTYAQ	HLHAWV	YVTKDGHY	GRAGCGKS	ILDDL	Phan et al., 2011
RodSCV-V-69	JF755403	Rodent	WVGTK	HVQFVV	YCSKSETR	GPSGSGKS	ILDDF	Phan et al., 2011
RodSCV-V-76	JF755404	Rodent	GTLNA	HLQMYV	YCMKEDTR	GPSGTGKS	IFDDF	Phan et al., 2011
RodSCV-V-77	JF755405	Rodent	LTYPQ	HLHAWV	YVMKDGDT	GRPGCGKS	VMNDV	Phan et al., 2011
RodSCV-V-87	JF755406	Rodent	LTYPQ	HLHAWV	YVMKDGDT	GRPGCGKS	VMNDV	Phan et al., 2011
RodSCV-M-44	JF755408	Rodent	LTYPQ	HLHAWV	YVMKDGDT	GRPGCGKS	VMNDV	Phan et al., 2011
RodSCV-M-45	JF755409	Rodent	FTIPD	HWQLIC	YVWKDDTC	GSTGMGKS	VIDEF	Phan et al., 2011

RodSCV-M-13	JF755410	Rodent	FTSYV	HYQGYA	YCKKDDCR	GDTGVGKT	-	Phan et al., 2011
RodSCV-V-72	JF755411	Rodent	WVFTR	HIQGYA	YAMKEDTR	GPSGSGKS	-	Phan et al., 2011
RodSCV-V-81	JF755412	Rodent	WLGTK	HVQFVV	YCTKEDTR	GPSGTGKS	ILDDF	Phan et al., 2011
RodSCV-V-84	JF755413	Rodent	WLGTK	HVQFVV	YCTKEDTR	GPSGTGKS	ILDDF	Phan et al., 2011
RodSCV-V-97	JF755414	Rodent	WLGTK	HVQFVV	YCTKEDTR	GPSGTGKS	ILDDF	Phan et al., 2011
RodSCV-M-53	JF755415	Rodent	WLGTK	HVQFVV	YCTKEDTR	GPSGTGKS	ILDDF	Phan et al., 2011
YN-BtCV-1	JF938078	Bat	FTINN	HLQGYC	YCKKEGDF	GSTGTGKS	IIEDF	Ge et al., 2011
YN-BtCV-2	JF938079	Bat	LLTIP	HWQLVV	YVWKEDTS	GRTGAGKS	VIDEF	Ge et al., 2011
YN-BtCV-3	JF938080	Bat	FTLNN	HLQGFC	YCSKAGNF	GPPGSGKS	IIDDF	Ge et al., 2011
YN-BtCV-4	JF938081	Bat	FTWNN	HLQGYA	YCSKAGEI	GLPGTGKS	IIDDF	Ge et al., 2011
YN-BtCV-5	JF938082	Bat	FTWNN	HLQGFC	YCKKSGDF	GPPGSGKS	IVDDF	Ge et al., 2011
BoSCVCP11-49-3	JN634851	Bovine	FTWNN	HLQGFC	YCSKSGDF	GPPGTGKS	VIDDF	Kim et al., 2011
MmFV	JN704610	Badger	LTYAQ	HLHAFV	YAIKDGDV	GETRLGKT	-	Van den Brand et al., 2011
RodSCV-V-86	JF755416	Rodent	FLTIN	HMHGII	YLEKTSGT	GTTGSGKS	ILDDL	Phan et al., 2011
RodSCV-V-91	JF755417	Rodent	FLTIN	HMHGII	YLEKTSGT	GTTGSGKS	ILDDL	Phan et al., 2011

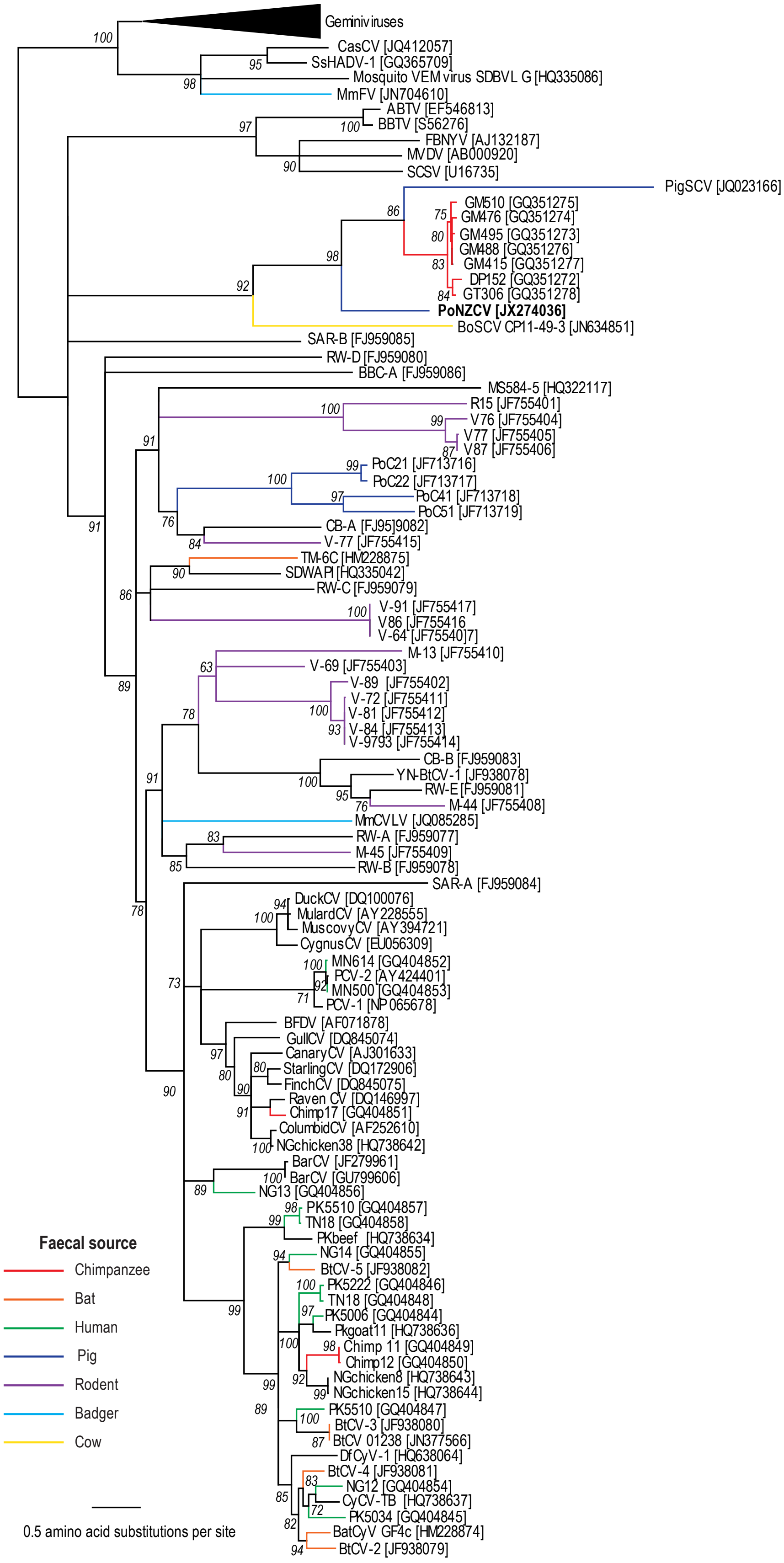


Figure 2.2 Maximum Likelihood Phylogenetic Tree of the Reps of PoSCV with Reps of other novel ssDNA isolates and representative species from ssDNA families and genera. The phylogenetic tree was constructed in PHYML (version 3) and the LG model, and editted manually. Branches with less than 60% aLRT support were collapsed with Mesquite (version 2.75). The colored branches represent isolates from faecal sources which are identified in the legend.

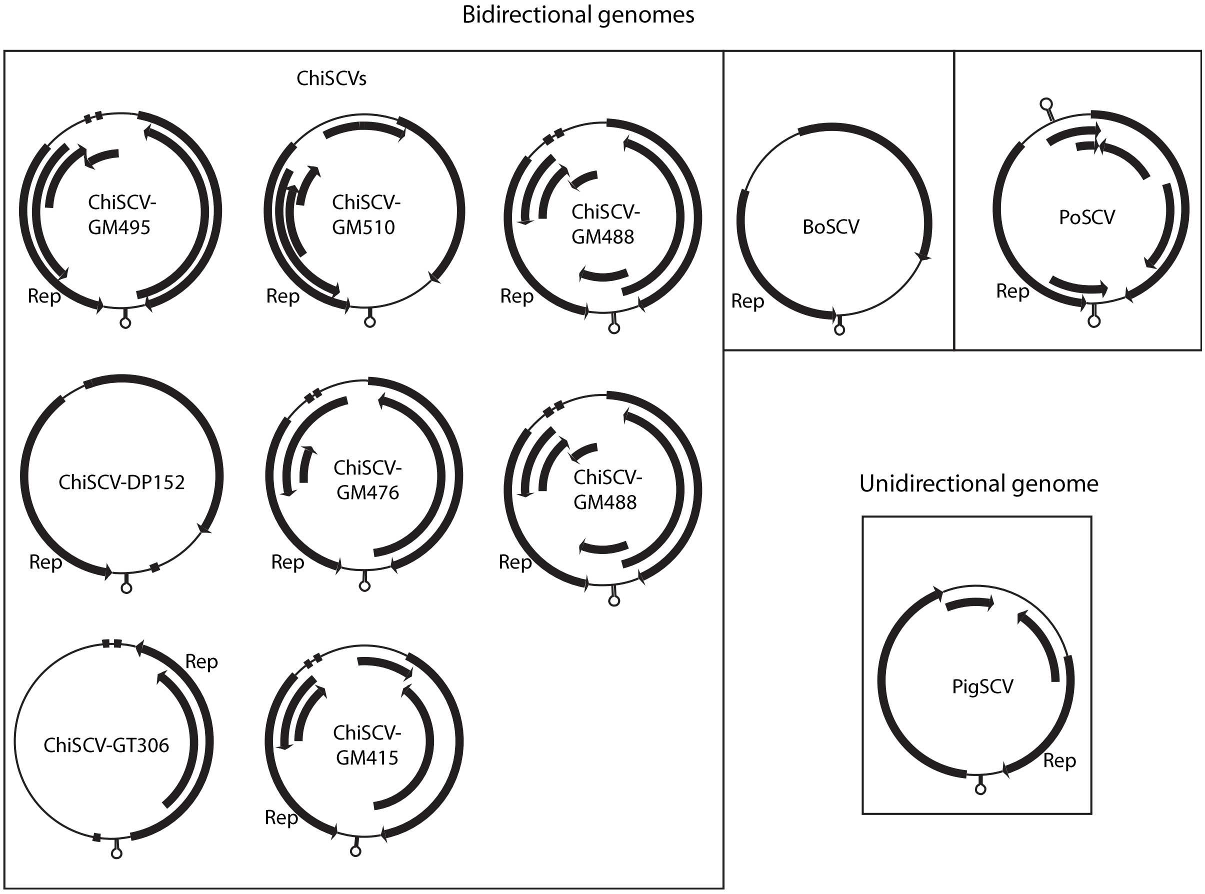


Figure 2.3 The genome organizations of all the isolates that form a unique clade with PoSCV during phylogenetic analysis of the Reps: ChiSCVs, BoSCV, and PigSCV. PigSCV is the only isolate with a unidirectional genome.

The putative Rep amino acid sequence contains both the rolling circle replication (RCR) and SF3 Helicase motifs that are reviewed in Rosario *et al.* (2012b). Motifs from other ssDNA viral isolates are listed in Table 2.2. All RCR motifs were identified in PoSCV: Motif I (MTIPR), Motif II (HWQIRI), and Motif III (YETKEGQY). For the SF3 Helicase motifs, walker A

(TIDTV) and walker B (GNVGKSW) were identified, but not Motif C. The closest homologue with a solved solution structure is the Rep of *Faba bean necrotic yellows virus* (FBNYV; RCSB Protein Data Bank; ID 2HWT), with which PoSCV only shares ~15% amino acid identity.

2.4.3 *Phylogenetic analysis of the Rep of PoSCV*

Phylogenetic analysis of the Rep of PoSCV (see Figure 2.3) with those of other novel ssDNA isolates and representative species from ssDNA families and genera revealed that PoSCV clusters with ChiSCVs (Blinkova *et al.*, 2010), BoSCV (Kim *et al.*, 2012), PigSCV (Sachsenröder *et al.*, 2012). However, PoSCV is in a unique clade that is not shared with any other known ssDNA isolates.

2.4.4 *Relationship of PoSCV to ChiSCVs, BoSCV, and PigSCV*

Though they cluster together in phylogenetic analysis of the Reps, PoSCV, ChiSCVs, and BoSCV share moderate Rep amino acid identities. The amino acid sequence of the PoSCV Rep shared highest amino acid pairwise identity (49.8-51.5%) with Reps of ChiSCV isolates (MEGA5; calculated with pairwise deletion of gaps). Pairwise identity of PoSCV was 33.5% with BoSCV and 28.2% with PigSCV. Also, PigSCV has a unidirectional genome organization while all the other members of this clade have bidirectional genome organizations.

Comparative analysis of the SIR nucleotide sequences revealed that BoSCV and ChiSCV have conserved putative stem-loop elements located after the stop codon of the Rep gene, while LIRs of PoSCV, BoSCV, and ChiSCV did not have conserved stem-loop elements. Also, the SIRs of PoSCV and ChiSCV share a similar organization of repeated sequences around the stem-loop element. Thus the origin of replication (ori) is likely the stem-loop element that is

Intergenic regions

PoSCV-LIR

```
 <- Rep start                 <    stem loop element     >
CATAAACGCCGTATTGCGAAAGCAATTTATAACGGCGCAATAGATTACCTTACGTCGTTTACGACTATAACAAAAAGG
AGTTATGTTGGTTATGGCATACTACAGAAGGCGCAGAAGCTACGGAAGAAGGACTTACGGTCGCAGGTATCATAGAAG
GAGGCGCTACTGATGGTTGCTCCGGTTATCGTCGTTGCCGGTGCGATGGCCGGTTCTTCCATTCTCAACTACGTTCTC
AATCAGAAGACTACCCGCGCTCAGATCAGACAGTCCGATTATGTCTCGGATTATTCAAGGCGTTACAATGCGGAAAAT
AGAAGATATTGGGCCGATTACTACAAGAACACGGGTTTTCGTCCTAGGTATCCTATGCGCTCAGGTGCTGAGTACAAT
CTTAGTGCTCTGTATGGTGCTCGTACCTCTAGGGCCAATGCTATTGCCTCTCAGTATCGTGCTGGTGCTGGTGTGGGT
GTATCTGGCGCTTATGGACTTAATGCGATCTATAGAAAAAGGTGATAAAAATG
                                                 -> CP start
```

PoSCV-SIR

```
 * CP stop
TGATGAAGGTAATCATGAAAATCATGATCAAGGTCCGTGATTGGTTGGATGCACGGATCGAATATCACAAAGAACCGT
AAACATCGGGGTAATTCTCCGGTTAAACCTCTTCCTTTTTGTTGCTAATTTACCGGCAAGTAACCTTTGTCTTCTGTG
GGGTAAGGGCCCCCGAATAGGCGGAGCCTCCCCAGTGGAACGTTCTCAGCGTAGGGGTCTTTCCTCCACCCCCCTATA
CCCCTAAAGGGGTATAGAAGGGGCTCCGCTAGTGTTACGAAAAGGAGCCCCCTCGGCCTCCTCA
                   <       stem loop element     >          * Rep stop
```

Rep gene upstream sequences (homologous segment)

```
PoSCV
GCAACCATCAGTAGCGCCTCCTTCTATAGTACCTGCGACCGTAAGTCCTTCTTCCGTAGCTTCTGCGCCTTCTGTAGT
ATGCCATAACCAACATAACTCCTTTTTGTTATAGTCGTAAACGACGTAAGGTAATCTATTGCGCCGTTATAAATTGCT
TTCGCAATACGGCGTTTATG
                 -> Rep start
ChiSCV-GM476
AGCGCCTTCCATACGACCTTCTTCCATACGACCTGCGCCTATATCCATAAGCCATTTTCATTACCTCTGCAACTAGTT
AAGCTAGTAGCGTTACAAGCAAGCGCACCGCTATTACATGAGCATTCGCATCAAAATCCACAAGTGCTATAAAGGGC
TGTCCACGACGGCCTCTCGAATATG
                      -> Rep start
```

Figure 2.4 Above: The intergenic regions of PoSCV. Below: The homologous nucleotide sequence segments located upstream of the Reps of PoSCV and ChiSCV-GM476.

located in the SIR (see Figure 2.1). The nonanucleotide motifs located at the putative ori of PoSCV is “TAGTGTTAC”, which is different than the conserved nonanucleotide motifs of nanoviruses, circoviruses, and cycloviruses which have a “NANTATTAC” nucleotide sequence (where “N” denotes any residue). BoSCV and ChiSCVs also have divergent nonanucleotide motifs. Based on similarities in the Rep and capsid protein sequences, stem-loop elements, and genome organizations, PoSCV, BoSCV, and ChiSCV are probably each representative of new genera in a new family.

2.5 Concluding remarks

The protocol used here is quick and inexpensive when compared to methods that employ next generation sequencing. Sanger sequencing is much less expensive than next generation sequencing and it does not produce large databases of sequence data that must be sorted. Furthermore, both sequencing methods are still limited by the sequence verification step, usually inverse PCR with internal primers and Sanger sequencing, which is critical before genome characterisation because of the potential for chimeras associated with sequence assemblies into contigs. Chimeras are most likely the reason for failure to verify some of the putative novel ssDNA viruses from next generation sequencing data in previous studies that investigated novel ssDNA viruses in environmental samples (Kim *et al.*, 2008; Rosario *et al.*, 2009a).

Investigating animal faecal matter for novel ssDNA viruses may forecast emerging pathogens in the environment, which often originate inside animal reservoirs (Woolhouse & Gaunt, 2007). Novel virus sequences identified with this protocol can be used for screening and surveillance to protect proactively against disease outbreak. New Zealand pigs have already been found to carry pathogenic circular ssDNA viruses such as *Porcine circovirus* type 2 (PCV2), which is of particular concern worldwide because of the associated development of a fatal condition called postweaning multisystemic wasting syndrome (PMWS) (Garkavenko *et al.*, 2005). Monitoring for novel ssDNA viruses with non-sequence-specific techniques is optimal because ssDNA viruses have such high rates of mutation. Additionally, with faecal sampling it is not only the animal’s own virome that is explored, but also the viruses that are infecting elements of the animal’s environment such as its food or water source, and the viruses infecting the organisms that have inhabited the faecal matter. Faecal matter is an ideal way to explore novel viruses in ecosystems because it eliminates the need for animal handling, thereby reducing the impact on the animal as well as the environment. As a non-invasive approach it also expands the range of animals that can be investigated to include those that are protected, such as some of New Zealand’s threatened endemic bird species. The next aim will be to further investigate ssDNA viral diversity using this method on a variety of animals from different ecosystems around New Zealand.

Chapter 3

Molecular characterisation of fourteen novel gemycircularvirus and gemycircularvirus-like isolates recovered from animal fecal samples in New Zealand

3.1 Abstract

The presence of single-stranded DNA (ssDNA) virus in the fecal matter of a variety of New Zealand animals was investigated, from which several sequences relating to the novel circular ssDNA virus group, Gemycircularvirus, were recovered. The full viral genomes were analyzed in an effort to better characterise a facet of ssDNA virus diversity in New Zealand. The methods for the discovery of novel ssDNA virus sequences were based on non-specific amplification by rolling circle amplification (RCA), followed by restriction enzyme (RE) digests or specific primers based on previously obtained next generation sequences. Fourteen novel ssDNA virus sequences relating to gemycircularviruses were discovered, eight with the use of specific primers, and six by RE digest followed by inverse PCR. Ten species of gemycircularvirus were identified based on phylogenetic analysis of the Replication associated proteins (Reps), genome architecture, and full nucleotide pairwise identities and were named Fecal associated gemycircularvirus-1, 2, 3, 4, 5, 6, 7, 8, 9, and 10. Three of the isolates with >79% genome-wide nucleotide pairwise identity were determined to be different strains of the same species, FasGCV-1. Two sequences appear to be divergent from gemycircularviruses based on the phylogenetic analysis of the Rep and only share 56-60% full nucleotide pairwise identity with gemycircularviruses. These isolates are tentatively named Ostrich fecal-associated ssDNA virus (OfaV) and Rabbit fecal-associated ssDNA virus (RfaV). The gemycircularvirus isolates discovered here come from a range of animal sources, which indicates a greater distribution of these viruses in the environment than previously thought.

3.2 Introduction

Animal fecal matter is a convenient source for the recovery and characterisation of novel ssDNA viruses from ecosystems for the purpose of exploring viral diversity and sequence space. Specifically this has been demonstrated through the recovery of diverse novel single-stranded DNA (ssDNA) viruses in the fecal matter of chimpanzees (Blinkova *et al.*, 2010; Li *et al.*, 2010a), bovine (Kim *et al.*, 2012), rodents (Phan *et al.*, 2011), bats (Ge *et al.*, 2011; Li *et al.*, 2010b), badgers (van den Brand *et al.*, 2012), and pigs (Sachsenröder *et al.*, 2012; Shan *et al.*, 2011). The protocol that was used for the recovery and isolation of ssDNA viruses from pig fecal matter described previously (see Chapter 2) has been further applied to investigate ssDNA viruses in fecal samples from a variety of animals across New Zealand. This included domesticated and wild animals from different niches within the New Zealand environment such as farmed cows, pigs, and chickens, wild pigs, geese, and hares, birds from the Chatham Islands, and seals. To contribute to the understanding of ssDNA viral sequence space, we endeavoured to find and characterise a group of related viruses among the novel sequence isolates which thus may share evolutionary and mechanistic relationships. Of the all novel ssDNA sequences obtained through fecal sampling, several had BLASTx (Altschul *et al.*, 1990) similarities to newly discovered gemycircularviruses. Gemycircularvirus is a newly proposed genus awaiting approval from the International Committee of Taxonomy of Viruses (ICTV) (Rosario *et al.*, 2012a). The proposed inclusion of this genus is based on seven closely related, recently discovered ssDNA viruses that share sequence similarity to geminiviruses but have unique genome organizations and cluster into a distinct clade during phylogenetic analysis of the highly conserved replication-associated protein (Rep) sequences (Rosario et al., 2012). The most highly studied gemycircularvirus, *Sclerotinia sclerotiorum* hypovirulence-associated DNA virus 1 (SsHADV-1), is a fungal-infecting virus, and the high similarities shared among gemycircularviruses in the Rep motifs and genome organizations suggest that they probably are a group of fungal-infecting viruses (Yu *et al.*, 2010).

SsHADV-1 was the first fungal-infecting circular ssDNA virus discovered (Yu *et al.*, 2010). Culturing experiments verified that SsHADV-1 in fact has a fungal host and specifically is able to infect the hyphae of its host extracellularly (Yu *et al.*, 2010, Yu *et al.*, 2013). Phylogenetic analysis of SsHADV-1 with known ssDNA viruses at that point in time showed that it clustered most closely with ssDNA viruses of the *Geminiviridae* family. SsHADV-1 also has geminivirus Rep sequence motif (GRS) that is characteristic of geminiviruses (Nash *et al.*, 2011; Palmer & Rybicki, 1998; Yu *et al.*, 2010). In geminiviruses, the GRS is crucial for replication initiation; mutants with altered GRS regions were unable to complete the cleavage reaction in replication initiation and were not able to infect their hosts (Nash *et al.*, 2011). However, whole virion particles observed by transmission electron microscopy showed SsHADV-1 has singular isometric virion particles while geminiviruses have twinned quasi-icosahedral virion particles (Briddon 2003).

The Rep amino acid sequence of SsHADV-1 has the rolling circle replication (RCR) motifs I, II, III that are important for rolling circle replication of circular ssDNA viruses and plasmids (Ilyina & Koonin, 1992; Koonin & Ilyina, 1993). As highlighted in Chapter 1, the RCR amino acid motifs are significant for the functionality of the Rep protein during replication. RCR motif I may be important for binding of dsDNA during replication and RCR motif II is involved in divalent metal coordination for the cleavage of DNA (Hafner *et al.*, 1997; Ilyina & Koonin, 1992; Koonin & Ilyina, 1993; Laufs *et al.*, 1995; Steinfeldt *et al.*, 2006). Motif III has a conserved tyrosine and is the site of cleavage as the tyrosine forms a covalent bond with the 5' end of the cleaved strand (Laufs *et al.*, 1995). SsHADV-1 also has the SF3 helicase motifs, conserved amino acid motifs with functional significance in the helicase domain of the Rep protein of ssDNA viruses (Gorbalenya *et al.*, 1990; Walker *et al.*, 1982). The helicase domain unwinds the dsDNA intermediate during replication with energy obtained through NTP binding and hydrolysis (Gorbalenya *et al.*, 1990; Ilyina & Koonin, 1992)

Also in the proposed gemycircularvirus group is a recently characterised ssDNA virus isolate from cassava leaves from Ghana: Cassava associated circular DNA virus (CasCV; Dayaram *et al.*, 2012). In silico modelling of the Rep protein of CasCV with a geminivirus Rep showed that the locations of RCR motifs, the GRS, and helicase motifs were conserved, and are probably functionally similar (Dayaram *et al.*, 2012). Phylogenetic analysis of the Reps showed that an isolate from European badger feces from the Netherlands, Meles meles fecal virus (MmFV; van den Brand *et al.*, 2012), and an isolate from mosquitoes of California, USA, mosquito VEM SDBVL-G, (Ng *et al.*, 2011) also clustered with CasCV and SsHADV-1 into a unique clade (Dayaram *et al.*, 2012). Additional gemycircularviruses were isolated from dragonflies from the Kingdom of Tonga, and USA (Florida), Dragonfly associated circular viruses-1, 2, and 3 (DfasCV-1, DfasCV-2 and DfasCV-3). All three isolates had spliced Reps with the RCR motifs, GRS, and helicase motifs (Rosario *et al.*, 2012a). All of the gemycircularvirus isolates discovered to date exhibit a bidirectional type II genome organization similar to cycloviruses and geminiviruses (Rosario *et al.*, 2012b). Type II genomes have the putative origin of replication, a stem-loop element with a conserved nonanucleotide motif, located on the opposite strand from the Rep (see Table 1.1) (Rosario *et al.*, 2012b).

Similar sequences to the Rep of SsHADV-1 are found as integrons in the genomes of fungi (Lui et al., 2011). Several of these integrons have high similarity to the Rep of SsHADV-1 and other ssDNA viruses and hence may have originated from gemycircularviruses that once integrated into the host genome (Lui et al., 2011). Also interestingly, phylogenetic studies revealed that geminiviruses and gemycircularviruses may have evolved from an ancestor that predates the separation of plants and fungi (Liu *et al.*, 2011).

Fourteen closely related novel ssDNA virus isolates recovered from fecal samples are characterised and described here. Genome organization and phylogenetic analysis of the Rep sequences indicates that twelve of these are putative gemycircularviruses, with three belonging to the same species. Previously there were seven proposed genomes in the gemycircularvirus genus. This more than doubles the known isolates in the gemycircularviruses group. By looking at a group of related novel isolates we are able to effectively expand a specific area of viral diversity and make comparisons between the viral genome sequences within that group. This study demonstrates how exploration of ssDNA viruses in fecal matter and focusing on a specific area of viral diversity contributes significantly to illuminating the ssDNA viral sequence space.

3.3 Materials and methods

3.3.1 *Sample collection and virus particle purification*

Fecal samples were collected from a variety of animals across New Zealand. Samples were stored at 4° C until processing, or -20° C for samples collected before 2012. To extract viral DNA, the High Pure Viral Nucleic Acid Kit (Roche, USA) was used according to the manufacturer’s instructions. Sample preparation before viral nucleic acid extraction was according to the methods described in Chapter 2. In short, five grams of fecal matter was added with a disposable spatula to 5 ml of SM buffer (0.1M NaCl, 50 mM Tris/HCI - pH 7.4, 10 mM $MgSO_4$), homogenized, and centrifuged (10,000 rpm for 10 minutes). The supernatant was filtered through 0.45 µm and 0.2 µm syringe filters (Sartorius Stedim Biotech, Germany) and 200 µl of sample filtrate was used for viral isolation.

3.3.2 *Recovery and cloning of viral DNA using rolling circle amplification (RCA)*

As described previously, 1 µl of viral nucleic acid was used to enrich circular ssDNA using the Illustra^{TM} TempliPhi Kit (GE Healthcare). Five restriction enzymes (*Bam*HI, *Kpn*I, *Xmn*I, *Eco*RI, and *Eco*RV) were used to restrict the RCA enriched product from each sample. Restriction fragments were resolved on a 0.7% agarose gel and fragments in the size range of 1.5 - 6 kb were excised from the agarose gel, cleaned with the Megaquick-spin^{TM} Total Fragment DNA Purification Kit (iNtRON Biotechnology, Korea) and cloned into plasmid vectors cut with the appropriate restriction enzyme. The plasmids were transformed into *E*. coli strain DH5α and grown overnight in Luria broth culture at 37°C. Plasmids were recovered with the DNA-spin^{TM} Plasmid DNA Purification Kit (iNtRON Biotechnology, Korea) and sequenced by primer walking at Macrogen Inc. (Korea).

3.3.3 *Recovery and characterisation of complete genome of novel ssDNA virus*

Back-to-back primers were designed based on the BLASTx (Altschul *et al.*, 1990) analysis of sequence reads. PCR was undertaken using RCA enriched DNA as a template and the Kapa HiFi HotStart DNA polymerase (Kapa Biosystems, USA). Full genomes were ligated into pJET 1.2 blunt vector (Fermentas, USA) and transformed in to *E*. coli strain DH5α. The recovered plasmids were sequenced at Macrogen Inc. (Korea) by primer walking.

3.3.4 *Inverse PCR and primer PCR*

Back-to-back primers were designed for inverse PCR based on sequences obtained through the RE digest method. These primer sets were used on the RCA product to pull out full genomes. Secondly, these inverse PCR primer sets were used to screen all the fecal samples. Thirdly, the fecal samples were screened using back-to-back primer sets designed based on assembled contigs from a next generation sequencing run (another study). These back-to-back primers were designed inside the conserved Reps of putative viral genome sequences. PCR with the Kapa HiFi HotStart DNA polymerase (Kapa Biosystems, USA) was as follows: denaturation at 94°C for 2 min, 25 cycles of 94°C for 15 s, 55°C for 15 s, 72°C for 1 min, and a final 2 min extension at 72 °C. PCR products were then run on a 0.7% gel and bands of the required size were excised, cleaned and sequenced as above.

Table 3.1 All of the primer sets were used to screen all the fecal samples. The primers that begin with "fec_s_" were taken from another study that used next generation sequencing. Primers were designed in the conserved Rep of the putative viral genome sequences. The other primers were initially designed for inverse PCR to pull out full genomes based on the sequence information received through the restriction enzyme digests.

Primer name	Primer sequence
fec_s_22	F: CGTCTAACACCTCTTCGACCA R: ATTCGCCGAATCCAACCCC
fec_s_1	F: TGTCTTCGTGGAGCATTATCGT R: TGGTCGCTGGAGGGCTATCT
fec_s_34	F: TCCGCCAGTTCCGTAAAGTAAAGT R: AAAATCCACCAAGAAGCTTCTGA
as35	F:TCAAGCACGCGGTTGATGTTT R:CTATTGTCTTAAGGACGGGGA
P14Xmn1	F: CTTACGGCTCATTCCTCTTCG R: GCCCTCCTCAATGTCACCT
P22Xmn1	F: AAAGTTGGTTGAATGGGGCAG R: TTCCCCTTGTACTTGTCTGTGA
P21Xmn1	F: ATAGGATAATGTACTCTGATCATCG R: GAACCTGAGTCAACCTTCTTC
as24Xmn1	F:CTTACCGATCTTGCCCTCCTC R:TAATCTTCTGCTCTGCGCGG
P24Kpn1	F: GTCCCTCAGCTGATCCCTTG R: CATGGTCTCACGGGATTATTC

3.3.5 *Sequence Analysis*

Full genomes were assembled using DNAMAN (version 5.2.9; Lynnon Biosoft). Sequences with BLASTx (Altschul *et al.*, 1990) hits to gemycircularviruses were selected. Putative open reading frames were identified using DNAMAN (version 5.2.9; Lynnon Biosoft). The Reps and the spliced Reps were found through manual identification of the RCR motifs, the GRS, and the helicase motifs. Pairwise similarity score matrices of the full nucleotide sequences and amino acid sequences of the Rep and Coat protein (Cp) were generated with SDT (Version 1.0) (Muhire *et al.*, 2013).

Reps of the novel isolates were aligned with the Reps of other gemycircularviruses, integrons, and representatives of the *Geminiviridae* family using MUSCLE (Edgar, 2004). The alignment was edited manually for construction of a maximum likelihood phylogenetic tree using PHYML (version 3) (Guindon *et al.*, 2010) using the LG model. Approximate likelihood-ratio test (aLRT) branch support (Anisimova & Gascuel, 2006) was used, and branches with less than 60% aLRT support were collapsed (Mesquite, version 2.75).

Table 3.2 Novel ssDNA isolates recovered from fecal matter of animals across New Zealand. If the sequence was discovered through RE digest, the enzyme is specified.

Name of Isolate	Method of isolation	Source	Date of Sample	Location
Ostrich fecal associated ssDNA virus (OfaV)	PCR with fec_s_22 primers	Ostrich	2011	Christchurch, Canterbury, South Island
Rabbit fecal associated ssDNA virus (RfaV)	Digest with *Xmn*1 and inverse PCR	Rabbit	2009	Chilton Valley, Cass Basin, South Island
FasSCV-10	Digest with *Xmn*1 and inverse PCR	Starling	2009	Rangatira, Chatham Islands
FasGCV-9	Digest with *Xmn*1 and inverse PCR	Blackbird	2011	Rangatira, Chatham Islands
FasGCV-8	Digest with *Xmn*1 and inverse PCR	Black Robin	2011	Rangatira, Chatham Islands
GFasGCV-7	Digest with *Xmn*1 and inverse PCR	Duck	2012	Christchurch, Canterbury, South Island
FasGCV-6	PCR with fec_s_22 primers	CI Warbler	2011	Rangatira, Chatham Islands
FasGCV-5	Digest with *Kpn*1 and inverse PCR	CI Warbler	2011	Rangatira, Chatham Islands
FasGCV-4	PCR with fec_s_1 primers	Seal	2012	Half Moon Bay, Kaikoura, South Island
FasGCV-3	PCR with fec_s_1 primers	CI Warbler	2011	Rangatira, Chatham Islands
FasGCV-2	PCR with fec_s_34 primers	Pig	2011	Cass Basin, South Island
FasGCV-1 [NZ_as41_Sheep_2009]	PCR with fec_s_1 primers	Sheep	2009	Sugarloaf, Cass Basin, South Island
FasGCV-1 [NZ_P21_Black_robin_2011]	PCR with fec_s_1 primers	Black Robin	2011	Rangatira, Chatham Islands
FasGCV-1 [NZ_P9_Black_bird_2009]	PCR with fec_s_1 primers	Blackbird	2009	Rangatira, Chatham Islands

3.4 Results and discussion

Fourteen novel ssDNA viruses that group with gemycircularviruses were recovered from fecal matter using the methods described here. Based on full nucleotide pairwise identities, genome architecture, conserved motifs, phylogenetic analysis, and pairwise comparisons of the Reps and Cps, 10 novel species of gemycircularviruses were identified. The novel gemycircularviruses have been tentatively named Fecal-associated gemycircularvirus

(FasGCV)-1, 3, 4, 5, 6, 7, 8, 9, and 10. Three of the fourteen novel isolates potentially belong to the same species, FasGCV-1 [NZ_as41_Sheep_2009], FasGCV-1 [NZ_P21_Black_robin_2011], and FasGCV-1 [NZ_P9_Black_bird_2009] as they share >79% genome wide pairwise identity. Two other novel isolates have similarities to gemycircularviruses, and may be part of the same family that is yet to be identified. These two isolates do not cluster with the other gemycircularvirus isolates and hence they have been tentatively named Ostrich fecal-associated ssDNA virus (OfaV) and Rabbit fecal-associated ssDNA virus (RfaV).

Six novel isolates were discovered with the random digest technique followed by inverse PCR to pull out full genomes. Eight were discovered using specific primers which had been designed based on next generation sequencing data. Notably, the sequences acquired with the use of the next generation contig based primers are closely related; the three strains of FasGCV-1 and FasGCV-2, 3, 4, and 6 all have close relationships within the gemycircularvirus clade. A greater diversity within the group was accessed by the restriction digests and inverse PCR technique; FasGCV-5, 7, 8, 9, and 10, and RfaV all cluster more closely with other isolates based on the phylogenetic analysis of the Rep.

3.4.1 *Full genome analysis of the gemycircularviruses*

The FasGCV isolates and previously identified gemycircularvirus isolates SsHADV-1, CasCV, DfasCV-1, DfasCV-2, DfasCV-3, MmFV, and Mosquito VEM virus SDBVL-G share 57-78% genome-wide nucleotide pairwise identity (Figure 3.1). FasGCV-1 [NZ-as41-Sheep-2009], FasGCV-1 [NZ-P21-Black robin-2011], and FasGCV-1 [NZ-P9-Black bird-2009] share 79-97% genome-wide nucleotide pairwise identity (Figure 3.1) and therefore were determined to be the same species. Like other gemycircularviruses, the FasGCV isolates have genome sizes ranging from approximately 2.1 - 2.2kb and have a bidirectional genome organization with a Rep and Cp being transcribed away from the putative ori (Figure 3.2). The putative origin of replication (ori), a stem loop element with a conserved nonanucleotide motif, are located on the strand containing the Cp gene. These features give the genomes a type II circular ssDNA genome structure (Rosario *et al.*, 2012b). FasGCV-1, 2, 3, 4, 5, 8, 9, and 10 have one intergenic region containing the putative ori while the Cp and Rep genes overlap at the 3’ ends. FasGCV-6 and FasGCV-7 do not have overlapping ORFs and therefore they have two intergenic regions (Figure 3.2). In contrast to the rest of the isolates, FasGCV-6 has the largest intergenic region between the 3’ ends of the Rep and Cp (Figure 3.2).

The FasGCV isolates and the previously identified gemycircularvirus isolates all have a spliced Rep gene with the exception of SsHADV-1. The introns in the FasGCV isolates are 114-179 nucleotides (nt) in length. The FasGCV Rep genes (without inrons) range from 922-1027 nt in length and the Cp genes range from 782-929 nt in length.

All of the gemycircularviruses discovered to date have a unique nonanucleotide motif, the most common being “TAATATTAT” which is exhibited by 12 of the 19 species of gemycircularviruses (Table 3.3). Variations include the “AAATAACTT” in MmFV, “TAATACTAT” in FasGCV-3, and “TAATGTTATG” in FasGCV-4, 8, and 9 (Table 3.3).

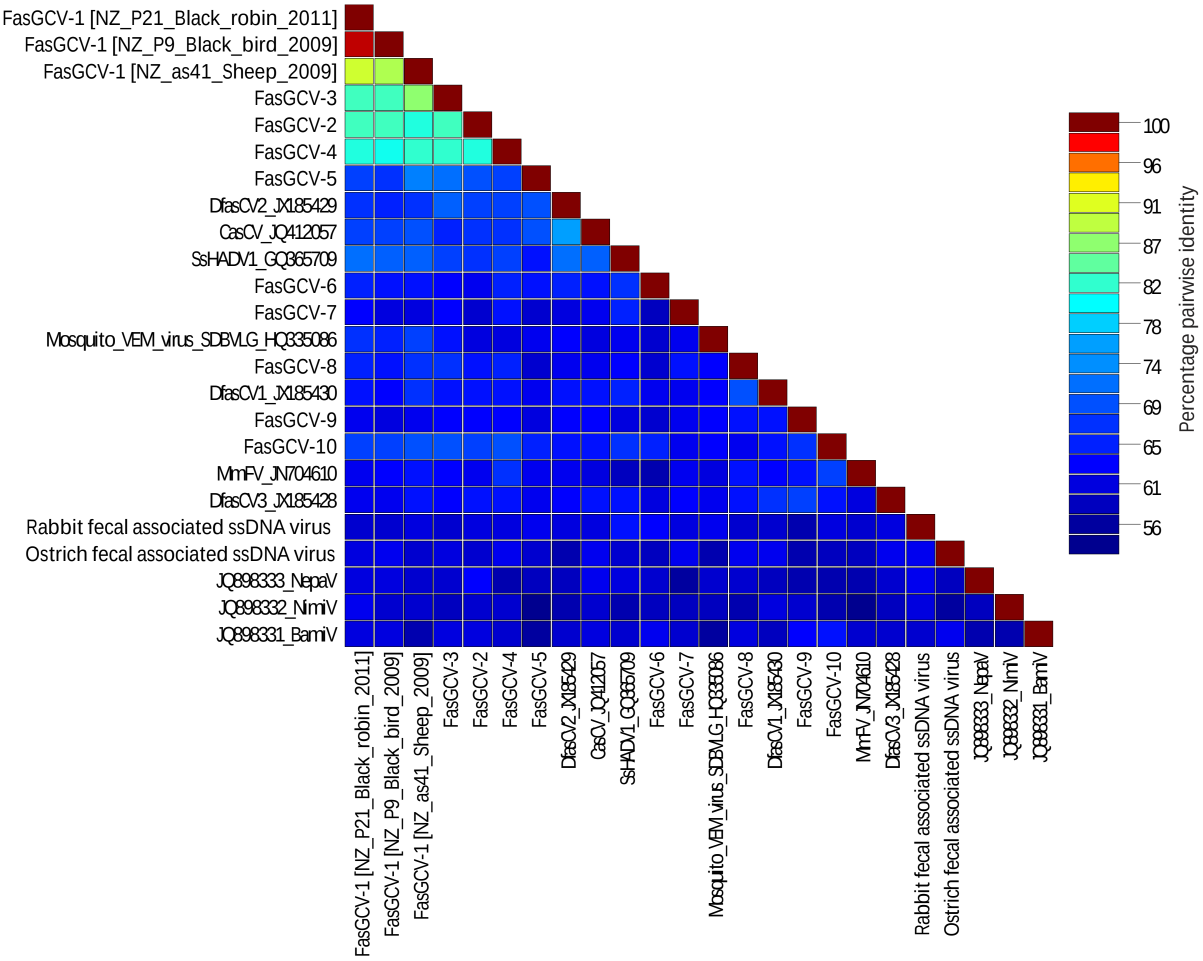


Figure 3.1 Nucleotide pairwise comparisons of the full sequences of gemycircularviruses and related isolates.

3.4.1 *Full genome analysis of two novel ssDNA viruses*

The two novel ssDNA isolates, OfaV (2122 nt) and Rabbit fecal associated ssDNA virus, RfaV (2162 nt), also have type II genome organizations (Figure 3.2) but they only share 56-60% full genome pairwise identity with gemycircularviruses (Figure 3.1). OfaV and RfaV have two intergenic regions (Figure 3.2). RfaV has the larger intergenic region between the 3' ends of the Rep and Cp gene and OfaV has the larger intergenic region between the 5' ends of the Rep and Cp gene (Figure 3.2). The genes of OfaV and RfaV are 1092 nt and 1081 nt (Reps), and 932 nt and 939 nt (Cps) respectively. Like gemycircularviruses, RfaV has a spliced Rep gene that contains a 146 nt intron and has the conserved "TAATATTAT" nonanucleotide motif at the putative ori (Table 3.3). Conversely, OfaV has a unique "TAACATTGA" nonanucleotide motif (Table 3.3) and does not have a spliced Rep.

3.4.2 *Analysis of the Rep and Cp*

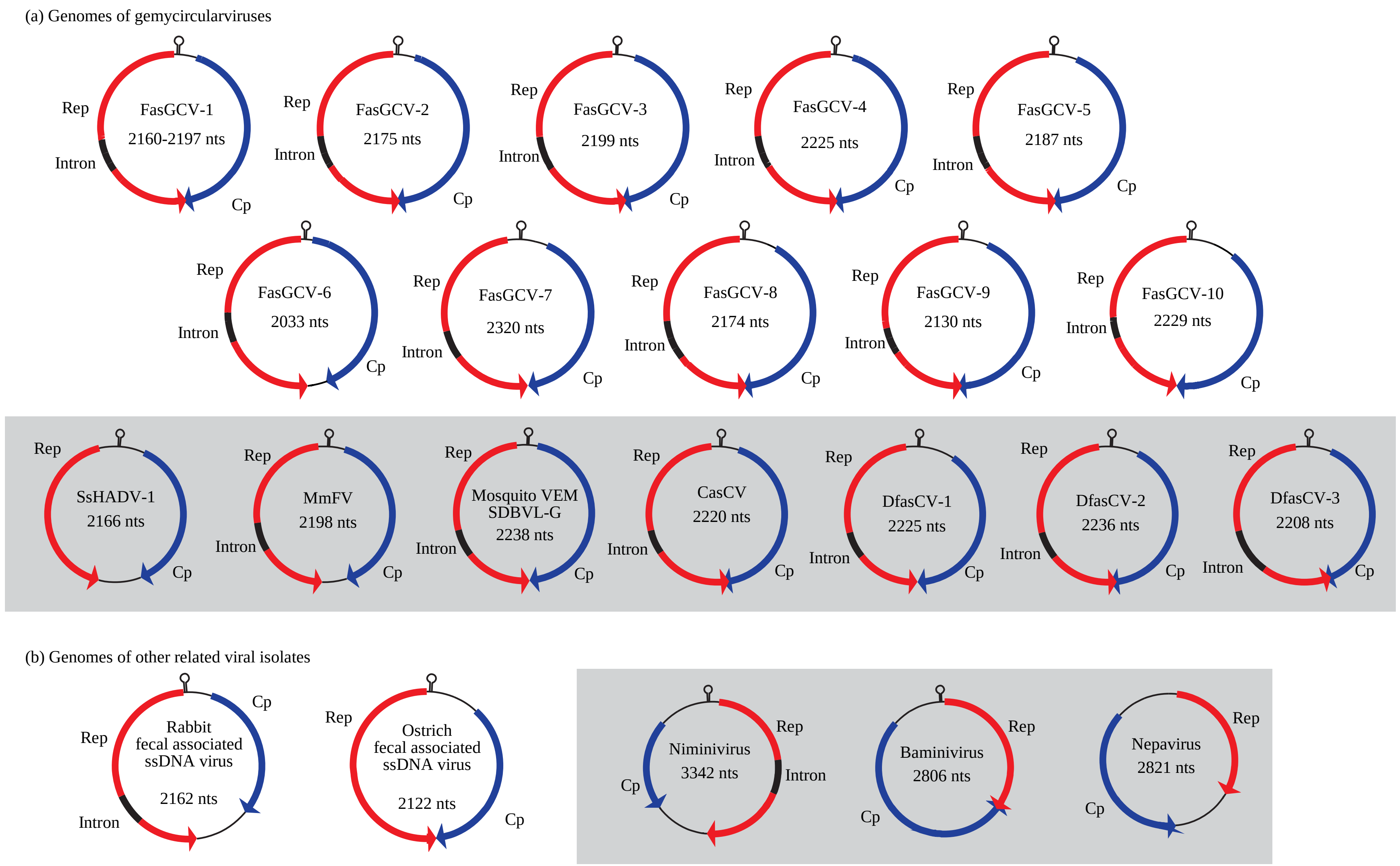
The gemycircularvirus isolates share from 36-90% pairwise identity in the amino acid Rep sequence (Figure 3.3). Additionally, there is significant conservation in the unique RCR, GRS, and helicase motifs. The unique RCR motifs I, II, and III have similarity to the geminivirus motifs which are "FLT(Y/P)sx", "(P/x)H(L/x)H(V/A/C)", and "(Y(U/a/c)xK" respectively. For all gemycircularvirus isolates discovered thus far, the RCR motif I is "L(L/V/I)T(Y/F)(A/P)" (Table 3.3). Most gemycircularvirus isolates have "H(L/x)HxF" conserved in RCR motif II (Table 3.3). This contains the leucine residue often located between the metal-binding histadine residues in geminiviruses, with an additional unique conserved phenylalanine. The exception is FasGCV-6 which only has one conserved histadine residue. Motif III for the gemycircularvirus isolates is "Y(A/T)(I/T/C)KD", still containing the essential tyrosine (Table 3.3).

The novel gemycircularvirus isolates also shared the NTP-binding domain previously found conserved across gemycircularviruses (Dayaram *et al.*, 2012). Within the NTP-binding domain the SF3 helicase motifs Walker A and Walker B could be identified (Table 3.3). The helicase motifs Walker-A and Walker-B in the novel isolates share similarity to the helicase motifs in geminiviruses which are Walker A "Gx(S/T)R(T/i)GK(s/T)" and Walker B "(V/I)(I/V)DD(V/I)", with the exception of FasGCV-1[NZ-P9-Black bird-2009] which has a highly divergent Walker A motif (Table 3.3).

OfaV and RfaV only share 28-39% pairwise identity in the Rep with gemycircularviruses (Figure 3.4). OfaV Rep has RCR motif I, II, and II, a GRS, and a Walker-B consistent with gemycircularviruses (Table 3.3). However, Walker-A is highly divergent with a "RRRRTRQT" amino acid sequence (Table 3.3). The RCR motifs II and III of Rabbit fecal-associated ssDNA are not that similar to those of the gemycircularviruses (Table 3.3).

Table 3.3 The conserved motifs of all the novel isolates (in bold) and previously identified gemycircularviruses. The nonanucleotide motifs of the novel isolates are similar to previously identified gemycircularviruses. The RCR conserved motifs I, II, and III for ssDNA viruses are usually Fu(t/u)(l/y)(t/p), (p/u)HuH, and YxxK respectively, and more specifically geminiviruses are FLT(Y/P)sx, (P/x)H(L/x)H(V/A/C), and (Y(U/a/c)xK respectively. The unique Rep motifs in gemycircularviruses and the novel isolates in this study show high conservation. The helicase motifs Walker-A and Walker-B in the novel isolates share similarity to the helicase motifs in geminiviruses which are Gx(S/T)R(T/i)GK(s/T) and (V/I)(I/V)DD(V/I). The GRS motif has some similarity to geminiviruses with several conserved "HPNI" sequences but with a unique C terminal end.

Isolate Name	Nonanucleotide Motif	Motif I	Motif II	GRS	Motif III	Walker-A	Walker- B
FasGCV-1 [NZ_as41_Sheep_2009]	TAATATTAT	LLTYA	HLHAFVD	DVFDVGGRHPNLVPSY	YAIKD	GDTRLGKT	VFDDM
FasGCV-1 [NZ_P21_Black_robin_2011]	TAATATTAT	LLTYA	HLHAFVD	DVFDVGGRHPNLVPSY	YAIKD	GDTRLGKT	VFDDM
FasGCV-1 [NZ_P9_Black_bird_2009]	TAATATTAT	LLTYA	HLHAFVD	DVFDVGGRHPNLVPSY	YAIKD	AFPAWLDV	AIDDM
FasGCV-2	TAATATTAT	LLTYA	HLHAFCD	DVFDVGGRHPNVMPSF	YATKD	GDTRLGKT	VFDDM
FasGCV-3	TAATACTAT	LLTYA	HLHAFCD	DVFDVGGRHPNLVPSY	YAIKD	GDTRLGKT	VFDDM
FasGCV-4	TAATGTTAT	LLTYA	HLHAFCD	DVFDVGGFHPNIEASR	YAIKD	GDTRLGKT	VFDDM
FasGCV-5	TAATATTAT	LVTYP	HLHVFCD	DIFDVGGFHPNIERSK	YACKD	GDALTGKT	VIDDI
FasGCV-6	TAATATTAT	LLTYA	HLHCFAD	RIFDVDGRHPNVVPSR	YAIKD	GPSLTGKT	VLDDI
FasGCV-7	TAATATTAT	LLTYP	TSHCFLD	RIFDIQGHHPNIERVG	YTIKD	GETRLGKT	IFDDL
FasGCV-8	TAATGTTAT	LLTFP	HLHAFFM	RVFDVDGRHPNVVRGY	YAIKD	GPTRLGKT	IFDDM
FasGCV-9	TAATGTTAT	LLTYA	HLHAFVD	RVFDVQGHHPNVEPSR	YAIKD	GPTRTGKT	VFDDF
FasSCV-10	TAATATTAT	CPHYL	HLHAFCD	RRFDVDGYHPNVQPFG	YAIKD	GESRLGKT	VFDDM
OfaV	TAACATTGA	FLTYS	HFHAFIL	RIFDFDGLHPNIESVR	YTKKD	RRRRTRQT	VFDDI
RfaV	TAATATTAT	FLTYS	HYHVLVA	RIFDVGGCHPNFKSVR	YCLKD	GRSRLGKT	VMDDI
SsHADV-1	TAATATTAT	LLTYA	HLHCFAE	DVFDVDGHHPNITKSR	YAIKD	GPSQTGKT	VFDDI
CasCV	TAATATTAT	LITYA	HLHCFID	DIFDVDGRHPNIEPSW	YAIKD	GDSRSGKT	IFDDI
DfasCV-1	TAATATTAT	LLTYP	HLHAFFM	RVFDVDGHHPNIVRGY	YATKD	GDTRLGKT	VFDDM
DfasCV-2	TAATATTAT	LVTYP	HLHCFAD	DIFDVDGCHPNIQPST	YAIKD	GESRTGKT	IFDDI
DfasCV-3	TAATATTAT	LLTYA	HYHAFFM	RIFDIDGYHPNILSGR	YATKD	GPSRTGKT	VFDDI
MmFV	AAATAACTT	LLTYA	HLHAFVD	RRFDVEGFHPNIAPCG	YAIKD	GETRLGKT	VLDDM
Mosquito_VEM_virus_SDBVL_G	TAATATTAT	LLTYA	HFHAFLD	RFWDIAGRHPNIARVG	YAIKD	GPSRTGKT	VFDDI


(a) Genomes of gemycircularviruses

(b) Genomes of other related viral isolates

Figure 3.2 (a) The genomes of the novel gemycircularviruses with all of the previously characterised gemycircularviruses (shaded) all have type II genome organizations with putative ori located on the strand opposite the Rep, and all except SsHADV-1 have spliced Reps. (b) The genomes of more divergent isolates that have similarities to gemycircularviruses, Rabbit fecal associated ssDNA virus (RfaV) and Ostrich fecal associated ssDNA virus (OfaV), and previously identified isolates Niminivirus, Baminivirus, and Nepavirus (shaded).

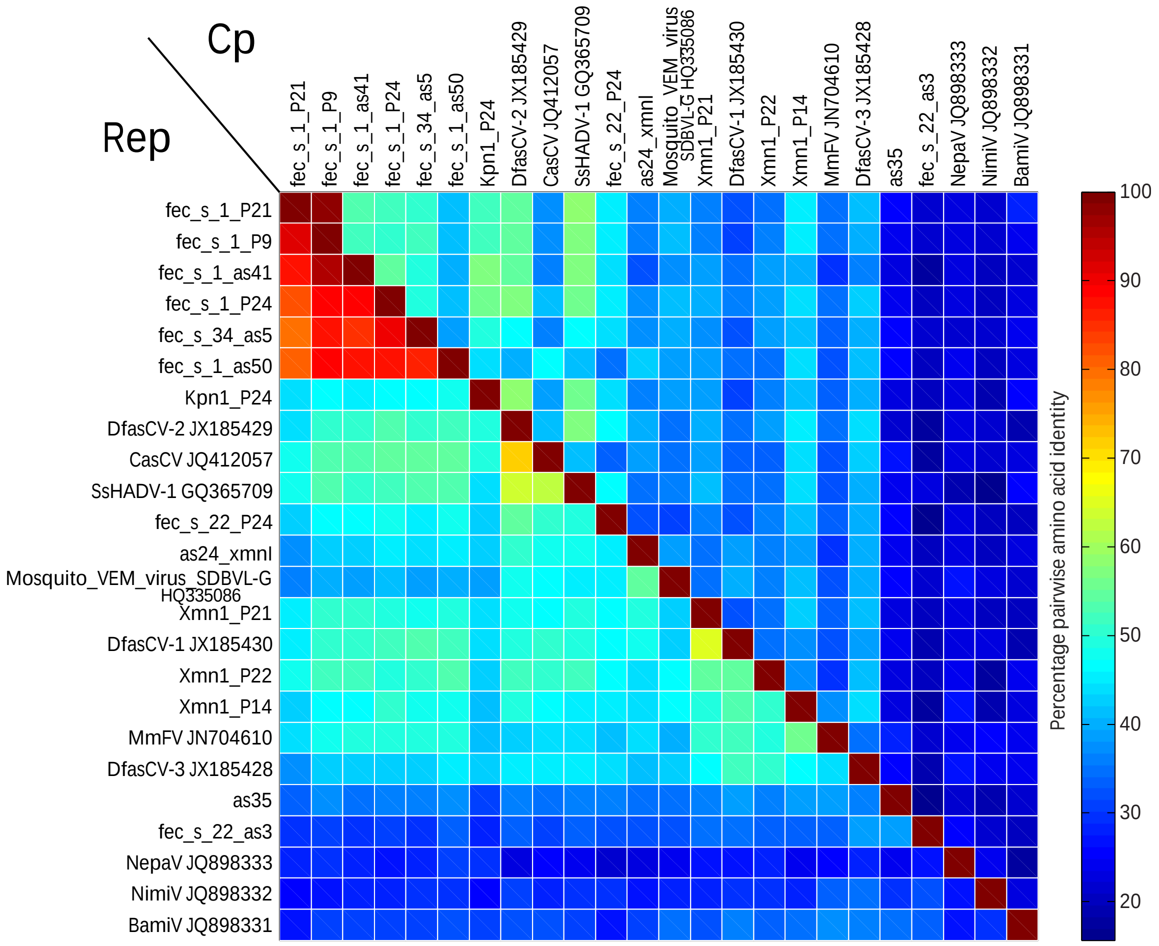


Figure 3.3 Pairwise comparisons of the novel isolates in this study and related gemycircularviruses and geminiviruses. The pairwise comparison of the Cp is inverted above the pairwise comparison of the Rep.

Like geminiviruses, the Reps of OfaV, RfaV, and the FasGCVs contain a conserved GRS (Nash *et al.*, 2011) (Table 3.3). The GRS motifs are consistent with the findings of Rosario *et al.* (2012a) in that the gemycircularvirus Reps lack the conserved "QxAK" amino acid sequence that is conserved on the C terminus end of geminivirus GRS motifs.

FasGCV-1 [NZ-P21-Black robin-2011] and FasGCV-1 [NZ-P9-Black bird-2009] share >98% pairwise identity in the Cp while sharing only ~53% pairwise identity with strain FasGCV-1 [NZ-as41-Sheep-2009], an isolate from sheep fecal matter. This is interesting as FasGCV-1 [NZ-P21-Black robin-2011] and FasGCV-1 [NZ-P9-Black bird-2009] are both isolated from avian sources from the Chatham Island and FasGCV-1 [NZ-as41-Sheep-2009] was isolated from sheep fecal matter on the South Island of New Zealand.

3.4.3 *Phylogenetic analysis*

The novel FasGCV isolates all cluster together in a clade with the previously identified gemycircularviruses in the phylogenetic analysis of the Reps (Figure 3.4). FaSGCV-1, -2, -3, and -4 all cluster in a clade within the gemycircularvirus group and share greater than 80% amino acid pairwise identity in the Rep sequences. Additionally, FasGCV-6 clusters with isolates SsHADV-1, DfasCV-2, and CasCV. Clustering close by but not within the gemycircularvirus clade are the integrons from truffle genomes (*Tuber melanosporum*) (Liu *et al.*, 2011). Slightly more distantly related are the integrons from other fungal genomes including *Nectria haematococca*, *Aspergillus fumigates*, and *Magnaporthe orzae* in another

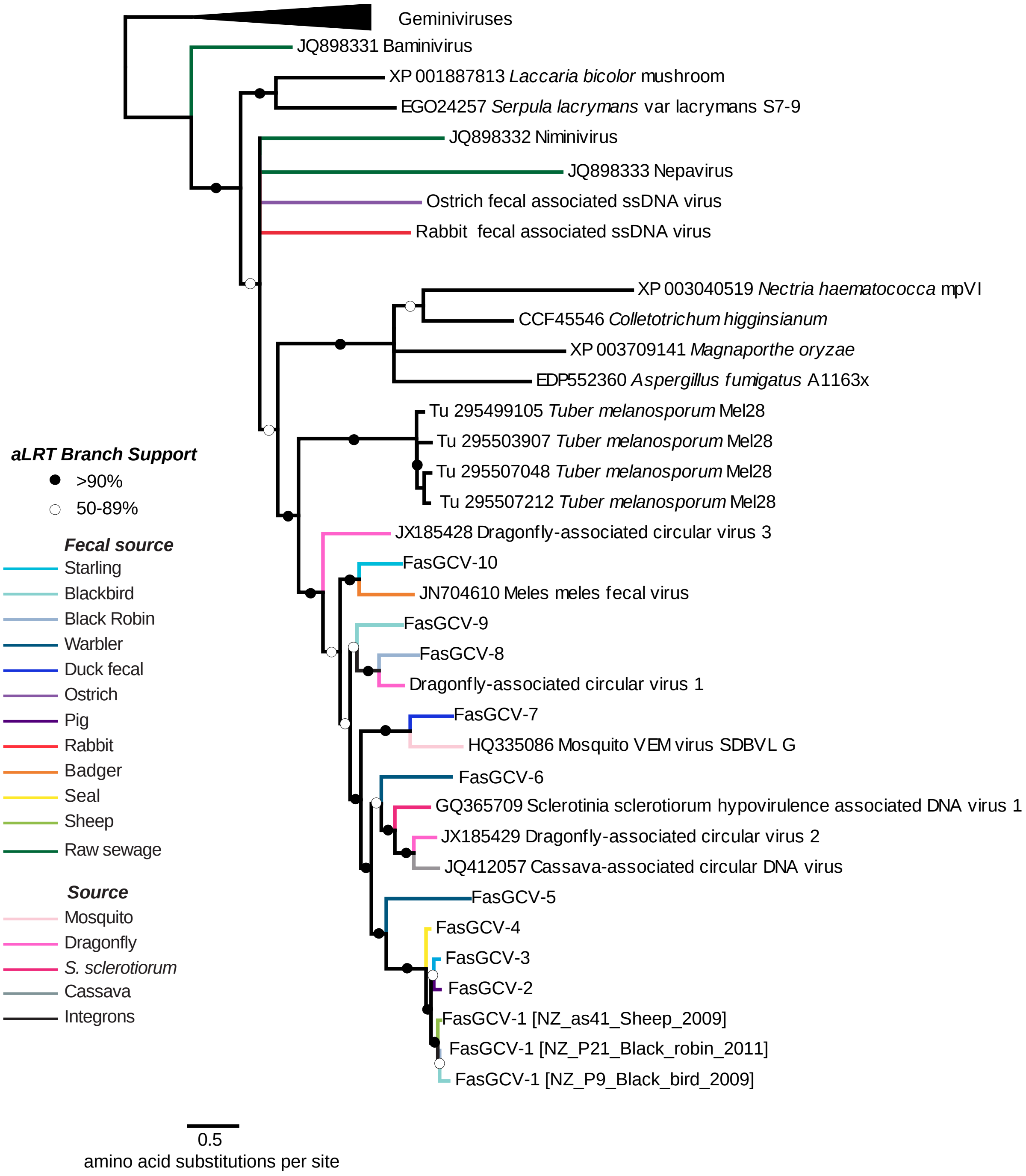

Geminiviruses
JQ898331 Baminivirus
XP 001887813 *Laccaria bicolor* mushroom
EGO24257 *Serpula lacrymans* var lacrymans S7-9
JQ898332 Niminivirus
JQ898333 Nepavirus
Ostrich fecal associated ssDNA virus
Rabbit fecal associated ssDNA virus
XP 003040519 *Nectria haematococca* mpVI
CCF45546 *Colletotrichum higginsianum*
XP 003709141 *Magnaporthe oryzae*
EDP552360 *Aspergillus fumigatus* A1163x
Tu 295499105 *Tuber melanosporum* Mel28
Tu 295503907 *Tuber melanosporum* Mel28
Tu 295507048 *Tuber melanosporum* Mel28
Tu 295507212 *Tuber melanosporum* Mel28
JX185428 Dragonfly-associated circular virus 3
FasGCV-10
JN704610 Meles meles fecal virus
FasGCV-9
FasGCV-8
Dragonfly-associated circular virus 1
FasGCV-7
HQ335086 Mosquito VEM virus SDBVL G
FasGCV-6
GQ365709 Sclerotinia sclerotiorum hypovirulence associated DNA virus 1
JX185429 Dragonfly-associated circular virus 2
JQ412057 Cassava-associated circular DNA virus
FasGCV-5
FasGCV-4
FasGCV-3
FasGCV-2
FasGCV-1 [NZ_as41_Sheep_2009]
FasGCV-1 [NZ_P21_Black_robin_2011]
FasGCV-1 [NZ_P9_Black_bird_2009]
aLRT Branch Support
● >90%
○ 50-89%
Fecal source
Starling
Blackbird
Black Robin
Warbler
Duck fecal
Ostrich
Pig
Rabbit
Badger
Seal
Sheep
Raw sewage
Source
Mosquito
Dragonfly
S. sclerotiorum
Cassava
Integrons
0.5
amino acid substitutions per site

Figure 3.4 Phylogenetic analysis of the Rep sequences of previously identified gemycircularviruses, the FasGCV isolates, Ostrich fecal associated ssDNA virus isolate, Rabbit fecal associated ssDNA virus isolate, and related sequences such as integrons from fungal genomes, Niminivirus, Baminivirus. The tree is rooted with geminiviruses.

clade (Liu *et al.*, 2011). The integrons of fungal origin lay basal to gemycircularviruses further supporting the idea that gemycircularviruses may infect fungi or have ancestors that interacted closely with fungi. Also included in the phylogenetic tree are isolates that have similarities to gemycircularviruses, fecal isolates OfaV and RfaV, and sewage isolates Niminivirus, Baminivirus, and Nepavirus (Ng *et al.*, 2012). These novel isolates each form separate branches showing their divergent nature.

3.5 Concluding remarks

Through the investigation of ssDNA viruses in fecal matter we have significantly expanded the known diversity of ssDNA viruses. A significant contribution was made to the gemycircularvirus genus through the addition of twelve novel isolates and two distantly related isolates. Six of the isolates were discovered through restriction enzyme digests while eight were discovered through screening with specific primers based on next generation sequence contig data. This demonstrates that the non-specific digest technique followed by inverse PCR is at least as viable as the more expensive next generation sequencing method. The fourteen isolates were recovered from a variety of fecal sources and from animals of different locations and environments suggesting that gemycircularviruses are prevalent in the New Zealand environment. Additionally, gemycircularviruses may exist within a variety of environmental niches given that the sources for gemycircularvirus recovery to date have included bird fecal matter, sheep fecal matter, seal fecal matter, and whole dragonflies and mosquitoes.

It is likely that the hosts of gemycircularviruses are various fungi because of the high similarity to SsHADV-1(Yu *et al.*, 2010) and the phylogenetic relationship to integrons from fungal genomes (Lui *et al.*, 2011). Furthermore, it is well known that fungi colonise fecal matter and the majority of gemycircularviruses isolates now come from fecal matter. However, culturing experiments and infectivity studies would need to be conducted to verify the hosts of these viruses. Additional studies could look at the viral diversity around the highly divergent isolates in this study. For example OfaV and RfaV that fall into an area of the phylogenetic tree that is not well characterised could be used to design degenerate primers, which could then be used to screen environmental samples.

Chapter 4

Two novel ssDNA viruses recovered from nesting material

4.1 Abstract

A desiccated dead Yellow-crowned Parakeet chick was found inside a nest in the Poulter Valley in the South Island of New Zealand. The nest and parakeet were investigated for ssDNA viruses using the protocol described in Chapter 2. This protocol was originally designed for exploration of ssDNA in faecal samples and it utilizes sequence-independent molecular techniques coupled with inexpensive sanger sequencing. The protocol was chosen because it was previously successful for virus recovery from faecal matter and bird nests are naturally going to contain some faecal matter. Two novel ssDNA isolates were discovered, while known ssDNA viruses specific to Psittacine birds, such as BFDV, were not detected. The two novel ssDNA viruses are tentatively named CynNCXV (2308 nt) and CynNCKV (2087 nt). Both isolates are circular ssDNA viruses with genome architectures similar to circoviruses with open reading frames (ORFs) in a bidirectional orientation. The larger ORFs exhibited BLAST similarity to replication associated proteins (Reps) and had the rolling circle replication motifs I, II, III, as well as two of the helicase motifs, walker A and walker B. The putative CynNCKV Rep has 30% similarity to Picobiliphyte nano-like virus (Picobiliphyte M5584-5; 66-88% coverage; e-value of 5 x 10^{-33}) and the putative CynNCXV Rep has 33% similarity to Rodent stool associated virus (RodSCV M-45; 92-94% coverage; e-value of 5 x 10^{-31}). Putative stem-loop elements with conserved nonanucleotide motifs were also identified. Interestingly, there was no full genome BLASTn (Altschul et al., 1990) similarity for either isolate to known ssDNA viruses in GenBank. A maximum likelihood phylogenetic tree of the Reps shows the divergent nature of both ssDNA virus isolates.

This chapter is in press:

Sikorski, A., Kearvell, J., Elkington, S., Dayaram, A., Argüello-Astorga, G. R., Varsani, A. (2013) Novel ssDNA viruses discovered in Yellow-crowned Parakeet (Cyanoramphus auriceps) nesting material. Archives of Virology DOI: 10.1007/s00705-013-1642-6 (In press)

4.2 Introduction

Approximately 59% of New Zealand endemic bird species were either threatened or extinct as of 2008 (Miskelly *et al.*, 2008). It has been shown that population bottlenecks associated with the threatened status reduces immunocompetence (Hale & Briskie, 2007), and therefore leaves them at greater risk for disease outbreak. There are several circoviruses known to infect birds, and *Beak and feather disease* virus (BFDV) is a particular threat to Psittacine birds. A recent and comprehensive investigation of BFDV in New Zealand conducted by Massaro *et al.* (2012) recovered and characterised 31 full BFDV genomes from New Zealand birds using specific primers designed on a conserved region of the BFDV Rep. This was the first detection of BFDV on the South Island, and it identified eight Yellow-crowned Parakeets infected in Eglignton Valley (Massaro *et al.*, 2012). Only one other study has characterised full virus genomes of BFDV in New Zealand (Ortiz-Catedral *et al.*, 2010), and thus there is a need for baseline data to understand regional BFDV diversity and spread (Massaro *et al.*, 2012). For example, worldwide analysis of full BFDV genomes revealed 14 strains of BFDV, inter-strain as well as intra-strain recombination events, and that some of the assumed BFDV isolates may actually represent new species (Varsani *et al.*, 2011). Diverse strains of BFDV may be present in New Zealand with low similarity to known strains, as well as completely novel species of circular ssDNA viruses.

A dead and desiccated yellow-crowned parakeet (*Cyanoramphus auriceps*) chick was found inside a nest in the Poulter Valley on the South Island of New Zealand. The cause of death was unknown. In this chapter we screen the dead parakeet and its nesting material for any circular ssDNA viruses with the protocol that was outlined in Chapter 2. This protocol was chosen because it was previously successful for virus recovery from faecal matter, and bird nests are naturally going to contain some faecal matter from both the mother and the chicks. Also, by screening for any circular ssDNA viruses we are encompassing BFDV and other avian circoviruses as well as highly divergent species that have not yet been discovered. In this application, the protocol is an effective tool for sequence-independent exploration of ssDNA viruses in the dead parakeet and its nest. Two novel isolates were recovered and characterised that have highly novel sequences: they would not have been discovered with sequence-specific techniques.

4.3 Methods and materials

4.3.1 *Sample collection and viral particle purification*

Using the protocol designed for the recovery of novel ssDNA viruses from faecal matter, two novel ssDNA viruses were recovered from nesting material and scrapings of a dead yellow-crowned parakeet (*Cyanoramphus auriceps*). The nest was contained in two layers of snap-lock bags and held at 6°C. In a lamina flow hood, to prevent contamination of the sample as well as the lab, pieces of the nest and the desiccated chick were cut with a scalpel blade and approximately 5 grams was placed in a conical tube. Approximately 5 ml of SM buffer (0.1 M NaCl, 50 mM Tris/HCl [pH 7.4], and 10 mM MgSO4) was added to the conical tube and vortexed until homogenized. The sample was treated as previously described in Chapter 2. Briefly, the homogenized mixture was centrifuged for 10 minutes at 10,000 RPM, filtered with 0.45 µm and 0.22 µm syringe filters (Sartorius Stedim Biotech, Germany), and used for viral nucleic acid extraction with the High Pure Viral Nucleic Acid Kit (Roche, USA). Viral DNA was amplified by rolling circle amplification (RCA) via the Illustra^{TM} TempliPhi Kit (GE Healthcare).

4.3.2 *Recovery and cloning of viral DNA*

A *Kpn*I digest and an *Xmn*I digest of the concatemerized viral DNA each yielded a clear band representing approximately 2 kb linear fragments of DNA on a 0.7% agarose gel. These fragments were excised from the gel, cleaned, and cloned into plasmids as described previously in Chapter 2. The *Kpn*I fragment was cloned into *Kpn*I cut pUC19 (Fermentas, USA) and the *Xmn*I fragment was cloned into pJET2.1 (Fermentas, USA).

BLASTx (Altschul *et al.*, 1990) analysis of the cloned sequences revealed that each were representative of different novel ssDNA viruses. Back-to-back primers were designed based on the clones sequences in order to pull out full genomes with inverse PCR: P-as1*Xmn*I F 5'-GAG TAC GAC TGG TTT GCA GAT TAC G-3' and P-as1*Xmn*I R 5'-TGG GTC TGT ATC CCA CGT TAT CTC-3' for the *Xmn*I isolate, and P-as1*kpn*I-F 5'-TGC ATC CTC TTC GAC GAC TGG-3', P-as1*kpn*I 5'-TTT CTG TGC CGC ATA GCC GTT C-3' for the *Kpn*I isolate. Fragments were cloned into pJET 1.2 suicide vector (Fermentas,USA) and transformed into DH5α *E.coli*. The inserts were screened by PCR using pJET F/R primers and the colonies that were positive for the insert were used for overnight cultures. The plasmid was recovered with the DNA-spin^{TM} Plasmid DNA Purification Kit (iNtRON Biotechnology, Inc.) and sequenced at Macrogen Inc. (Korea) by primer walking.

4.3.3 *Sequence analysis*

Full genomes were assembled from sequence reads with DNAMAN (version 5.2.9; Lynnon Biosoft). The Reps of the two novel isolates discovered here were aligned with Reps of other novel ssDNA genome isolates and representative species of circovirus, cyclovirus, geminivirus, and nanovirus from GenBank (as of 24th of Aug 2012). Reps were first aligned with T-coffee (Notredame *et al.*, 2000) and edited manually. A maximum likelihood phylogenetic tree was constructed from the alignment using PHYML (version 3) (Guindon *et al.*, 2010), the LG model, and approximate likelihood-ratio test (aLRT) branch support (Anisimova & Gascuel, 2006). Branches were collapsed that had less than 60% aLRT support (Mesquite, version 2.75).

4.4 Results and discussion

4.4.1 *Sequence analysis*

The two novel isolates will be referred to as *Cyanoramphus* nest-associated circular X virus (CynNCXV) and *Cyanoramphus* nest-associated circular K virus CynNCKV. The verified full genome sequences of CynNCXV and CynNCKV are 2308 nucleotides (nt) and 2087 nt respectively. Interestingly, full genome BLASTn (Altschul *et al.*, 1990) searches for both isolates revealed no similarity to any sequences available in GenBank. The two novel isolates were determined to have similar architecture with the two major open reading frames (ORFs) identified with DNAMAN positioned in a bidirectional orientation. However, CynNCXV has two intergenic regions (IRs), a long intergenic region (LIR) of 362 nt and a short intergenic region (SIR) of 158 nt, while CynNCKV only has one IR that is 369 nt in length.

BLASTx (Altschul *et al.*, 1990) of the ORFs revealed that the largest ORFs of both CynNCXV

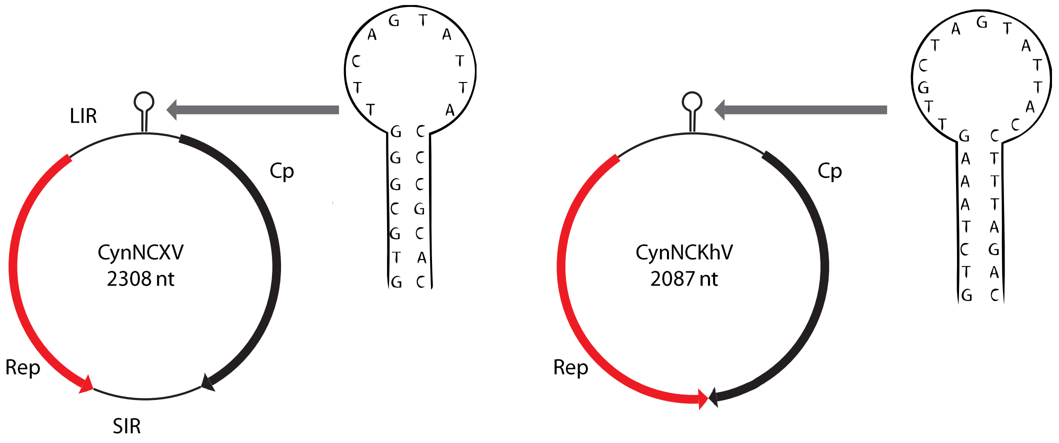


Figure 4.1 The genomes of CynNCXV and CynNCKV. The putative stem-loops are enlarged. The major ORFs are the Rep protein (red), and putative Coat proteins (Cp; black).

Table 4.1 The table of RCR motifs and SF3 helicase motifs. Motif C was not identified.

Isolate name	RCR motifs			SF3 helicase motifs	
	Motif 1	Motif 2	Motif 3	Walker A	Walker B
CynNCXV	FTVNN	HLQGFI	YCTKSCG	AEGNTGKT	VIFDF
CynNCKV	FTSYD	HFQGYA	YCRKEST	GVTGIGKT	ILFDD

(102 nt) and CynNCKV (900 nt) encode the putative Replication-associated proteins (Reps). The putative Rep of CynNCXV has significant similarity to the Reps of nanoviruses and ~30% to Picobiliphyte nano-like virus (e-value 5 x 10^{-33}; 66-88% coverage). The Rep of CynNCKV has significant similarity to Reps of Gardia intestinalis integrons and 33% to a Rodent stool associated ssDNA virus (AEM05797) (e-value 5 x 10^{-31}; 92-94% coverage). RCR Motifs I, II, III and superfamily 3 (SF3) helicase motifs walker A and B as reviewed by Rosario *et al.* (2012b) have been identified in the Rep amino acid sequences (see Table 4.4).

The second largest ORFs of CynNCXV (768 nt) and CynNCKV (861 nt) had no BLASTx (Altschul *et al.*, 1990) hits but are regarded as the putative capsid proteins. Stem-loop elements were identified in intergenetic regions just upstream from the putative coat protein genes. Conserved nonanucleotide sequence motifs were identified on the top of the stem-loop elements as “CAGTATTAC” for CynNCXV and “TAGTATTAC” for CynNCKV (see Figure 4.1).

4.4.2 *Phylogenetic analysis*

The Rep of CynNCXV clusters with the Reps of Picobiliphyte nano-like virus (Yoon *et al.*, 2011) and Dragonfly orbiculatusviruses (Rosario *et al.*, 2012a). The Rep of CynNCKV clusters with the Reps of Rodent stool associated virus M-45 (Phan *et al.*, 2011) and reclaimed water associated ssDNA viruses RW-B and RW-A (Rosario *et al.*, 2009b).

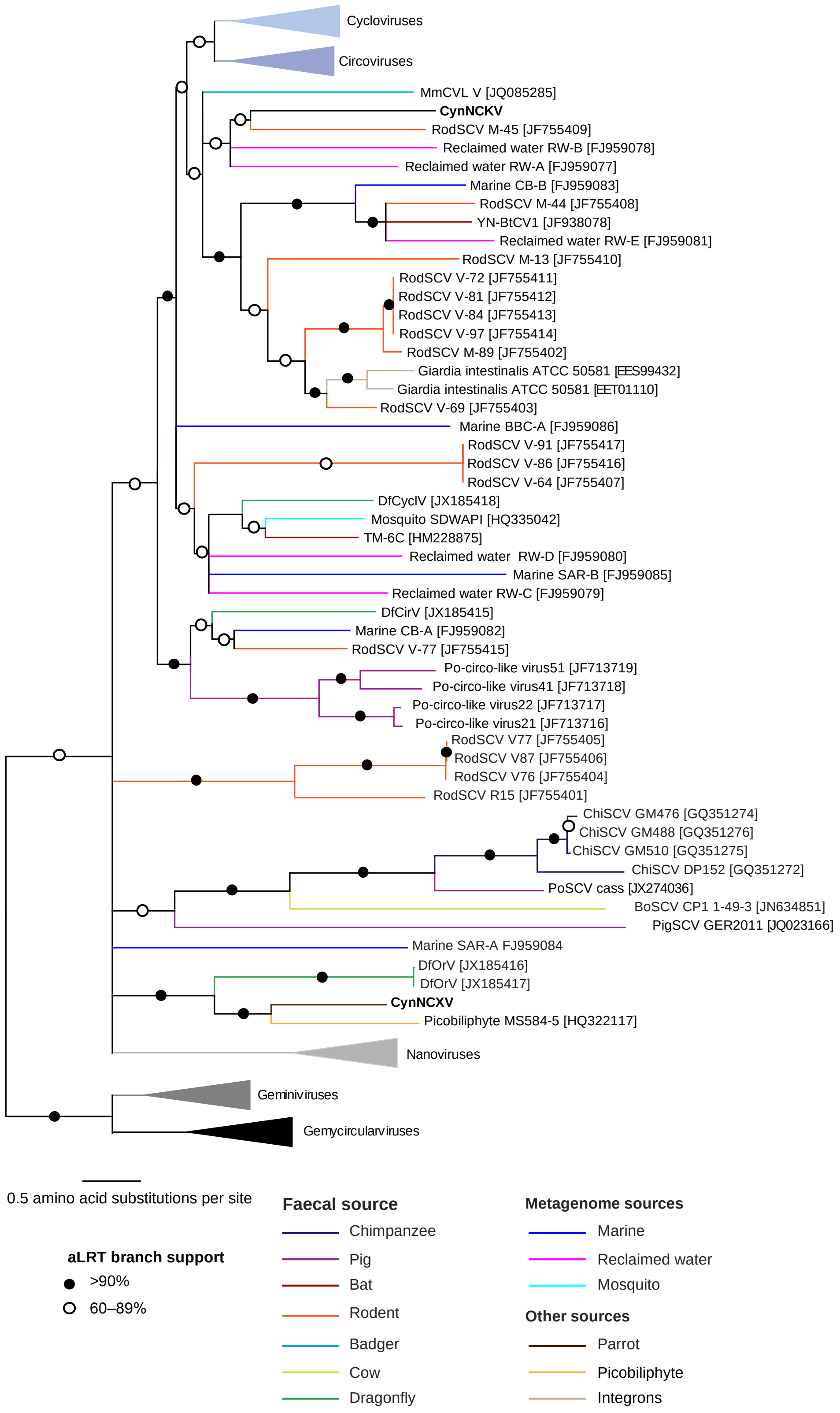


Figure 4.2 This maximum likelihood phylogenetic tree contains the Reps of the two novel isolates discovered here, CynNCXV and CynNCKV, as well as Reps of other novel ssDNA genome isolates, and Reps of representative species of circovirus, cyclovirus, geminivirus, and nanovirus from GenBank (24th of Aug 2012). This maximum likelihood phylogenetic tree was generated with the LG model and approximate likelihood-ratio test (aLRT) branch support (Anisimova & Gascuel, 2006) using PHYML (version 3). Branches with less than 60% aLRT support (Mesquite, version 2.75) were collapsed. Reps were first aligned with T-coffee (Notredame *et al.*, 2000) and edited manually.

4.5 Concluding remarks

The previously described protocol (see Chapter 2) was successfully applied for the recovery of novel ssDNA viruses from a Yellow-crowned Parakeet chick and nesting material. Previously the protocol was used for inexpensive and sequence-independent recovery of novel ssDNA viruses from faecal samples in order to investigate the viruses of an ecosystem, while it is used here to recover viruses from a dead chick and its nesting material without the use of specific primers. A significant challenge to protecting against emerging viral pathogens and associated disease outbreaks is the unknown viral sequence space. However, the molecular techniques used here overcome the need for specific knowledge about viral sequences.

The hosts of the viral isolates characterised here are unknown. Infectivity studies would need to be conducted in order to identify the hosts as well as severity of infection; however this is not possible due to conservation of these birds. As it stands, the virus could have been involved in the death of the parakeet, a secondary infection present because of the organisms involved in the decaying of the animal, or a virus infecting the nesting material. However, this protocol can continue to be used, possibly by the Department of Conservation, as an inexpensive diagnostic tool to investigate a wider range of ssDNA virus infections in a shorter period of time. Additionally, potential emerging pathogens identified through this approach can afterwards be screened for in specific PCR. This proactive approach to viral disease prevention will allow time for the development of mitigation plans before a full blown outbreak occurs, and thus assist in the conservation of endemic birds that are particularly vulnerable because of low population numbers.

GenBank Accession #s: JX908740, JX908739

Chapter 5

Conclusion

This dissertation set out to investigate environmental ssDNA viral diversity, and has identified several novel ssDNA viruses within the New Zealand environment. Viral diversity and sequence space is not only important to improve the resolution of the viral taxonomy, but it also provides a basis for investigating the significance of future novel viral isolates, and creates sequence data for monitoring changes in viral communities. The bulk of viral discovery research in the past has focused on characterising specific groups of viruses with direct and proven significance to human health and economically important crops or livestock and as a consequence the diversity and spread illustrated by known viral sequences has created a taxonomic framework that is unable to accommodate novel viruses being discovered in the environment. Broad environmental viral diversity studies that seek to identify novel viruses without prioritizing based on limited existing knowledge of their implications has recently become feasible due to advances in technologies of inexpensive sequencing and non-specific amplification, however, these studies are being carried out in regions of the world outside New Zealand and do not account for the unique New Zealand viral sequence space.

The aim of this dissertation was to optimize an affordable protocol for novel ssDNA virus discovery in low resource settings, such as New Zealand, and collect some baseline ssDNA virome data in New Zealand. Specifically, this study sought to:

1. Establish a method and approach for the recovery of novel ssDNA viruses from the environment in New Zealand that has a low impact on the environment while also being inexpensive
2. Characterise novel isolates in New Zealand and explore how they fit into the current world-wide viral diversity framework

5.1 Main findings

The viability of a method for recovering novel ssDNA viruses from the environment is demonstrated in Chapter 2. Porcine stool-associated circular virus (PoSCV) was recovered from a pig faecal sample using the methods described in Chapter 2 and has less than 30% similarity to isolates in the putative capsid protein, only 50% similarity to known isolates in the replication associated protein, and no significant overall nucleotide similarity. The success of non-specific amplification of ssDNA viruses in faecal samples by way of rolling circle amplification (RCA) and its value as an inexpensive tool for viral discovery is highlighted by the recovery of PoSCV. This chapter also establishes faecal sampling as a viable non-invasive sampling technique, as neither the animal nor the environment needed to be disturbed for sample collection. Faecal samples from New Zealand animals hold ssDNA viruses that have not yet been identified. Animals are bioaccumulators of viruses from the environment.

In Chapter 3 a group of related novel isolates are characterised in order to substantially expand the knowledge of a facet of ssDNA virology. The assessment of a group of related novel isolates makes it possible to focus on an aspect of viral diversity and assess viral genome sequences as they relate to each other. Fourteen highly novel isolates are characterised and incorporated into the framework of current ssDNA diversity using phylogenetic analysis of the Reps, genome architecture, and full genome nucleotide pairwise identities. The findings provided in this chapter more than double the number of species previously included in this clade and provide substantial support for the inclusion of gemycircularvirus as a new genus. Ten species of a novel genus, gemycircularvirus, are identified and given the names Faecal associated gemycircularvirus (FasCV) -1, 2, 3, 4, 5, 6, 7, 8, 9, and 10. Through analysis of all of the gemycircularviral sequences it is further substantiated that gemycircularvirus isolates contain a unique version of the Geminivirus Rep Sequence (GRS) motif that is usually characteristic of plant-infecting geminiviruses. While the gemycircularviruses are demonstrated to be separate from geminiviruses through phylogenetic analysis of the Rep sequences, it is also significant to note that most ssDNA viruses outside geminiviruses have not been found to contain a GRS motif. One of the members of the gemycircularvirus clade, SsHADV-1, has undergone infectivity studies showing that it is a fungal-infecting virus, and its host is the plant pathogen *Sclerotinia sclerotiorum* (Yu *et al.*, 2010). Since animal faeces contains many secondary microbial inhabitants, such as fungi, the gemycircularviruses genus may in fact be comprised of fungal-infecting viruses. Additional findings in this chapter are the two isolates that are determined to fall outside the gemycircularvirus clade based on their genome wide identities (56-60% full nucleotide pairwise identity with gemycircularviruses) and the phylogenetic relationships of their replication associated proteins (Rep).

Chapter 4 demonstrated a potential practical application of the novel ssDNA virus discovery protocol that was outlined in Chapter 2. In this chapter the protocol is used on 5 grams of mixed material from the Yellow-crowned Parakeet nest and not on 5 grams of pure faecal matter.

Nevertheless, two novel isolates *Cyanoramphus* nest-associated circular X virus (CynNCXV) and *Cyanoramphus* nest-associated circular K virus CynNCKV were successfully recovered. This alternative method overcomes the limitation of needing to find pure faecal matter when the aim is to explore a very specific ecosystem. The novel viruses isolated from the nest had little sequence identity to know viruses in GenBank and so they would not have been discovered using regular PCR techniques. This is significant because a Yellow-crowned Parakeet chick was found dead inside the nest and the cause of death was unknown.

Table 5.1 Faecal samples used in this thesis were collected from animals in New Zealand. The faecal samples were collected from the South Island, Rangatira of the Chatham Islands, and the North Island.

Sample name	Animal	Sample collector	Year of sampling	Locality	Region	Viral isolate(s)
as1	Yellow-crowned parakeet	John Kearvell and Simon Elkington	2012	Poulter Valley, Cass Basin	South Island	*Cyanoramphus* nest-associated circular X virus (CynNCXV) and *Cyanoramphus* nest-associated circular K virus (CynNCKV)
as2	Orange-fronted parakeet	John Kearvell and Simon Elkington	2012	Hawdon Valley, Cass Basin	South Island	
as3	Ostrich	Darren Smalley	2011	Canterbury	South Island	Ostrich faecal associated ssDNA virus (OfaV)
as4	Muscovy duck	Darren Smalley	2011	Canterbury	South Island	
as5	domestic pig	Virology lab group	2011	Cass Basin	South Island	Porcine stool-associated circular virus (PoSCV) and Faecal-associated gemycircularvirus 2 (FasGCV-2)
as6	Eclectus parrot	Frank	2011	Canterbury	South Island	
as7	Duck	Virology lab group	2012	Canterbury	South Island	
as8	Geese	Virology lab group	2012	Canterbury	South Island	
as9	Chicken	Virology lab group	2011	Canterbury	South Island	
as10	Ostrich	Darren Smalley	2011	Canterbury	South Island	
as11	Chicken	Dorien Coray	2011	Christchurch	South Island	
as12	Chicken	Virology lab group	2011	Canterbury	South Island	
as13	Chicken	Virology lab group	2011	Canterbury	South Island	
as14	Chicken	Virology lab group	2010	Riccarton, Christchurch	South Island	
as15	Cat	Virology lab group	2011	Canterbury	South Island	

as16	Sheep	Virology lab group	2009	Canterbury	South Island	
as17	Llama	Ryan Catchpole	2009	West Melton	South Island	
as18	Chicken	Ryan Catchpole	2009	West Melton	South Island	
as19	Hare	Laura Young	2010	Mt. Misery, Cass Basin	South Island	
as20	Wild pig	Laura Young	N/A	Cass Basin	South Island	
as21	Hare	Laura Young	2009	Sugarloaf, Cass Basin	South Island	
as22	Chamois	Laura Young	2009	Sugarloaf, Cass Basin	South Island	
as23	Cow	Matt Roche	2012	Horarota, Canterbury	South Island	
as24	Duck	Virology lab group	2012	Ilam, Christchurch	South Island	FasGCV-7
as25	Wild pig	Aaron Stevens	2012	Jollybrook River, Canterbury	South Island	
as26	Deer	Aaron Stevens	2012	Jollybrook River, Canterbury	South Island	
as27	Chicken	Ryan Catchpole	2012	West Melton	South Island	
as28	Duck	Ryan Catchpole	2012	Riccarton, Christchurch	South Island	
as29	Llama	Ryan Catchpole	2012	West Melton	South Island	
as30	Horse	Ryan Catchpole	2012	West Melton	South Island	
as32	Horse	Ryan Catchpole	2012	West Melton	South Island	
as35	Rabbit	Laura Young	2009	Chilton Valley, Cass Basin	South Island	Rabbit faecal associated ssDNA virus (RfaV)
as36	Hare	Laura Young	2009	Chilton Valley, Cass Basin	South Island	
as37	Possum	Laura Young	2009	Middle Bush	South Island	
as38	Possum	Laura Young	2009	Sugarloaf, Cass Basin	South Island	
as39	Deer	Laura Young	2009	Chilton Valley, Cass Basin	South Island	
as40	Hedgehog	Laura Young	2009	Cass Basin	South Island	
as41	Sheep	Laura Young	2009	Sugarloaf, Cass	South Island	FasGCV-1 strain NZ_as41_Sheep_2009
as42	Kaka	Auckland zoo	2011	Auckland Zoo, Auckland	North Island	
as43	Antipode parakeet	Auckland zoo	2011	Auckland Zoo, Auckland	North Island	
as47	Sheep	Jenny Ladley	2012	Wakefield	South Island	
as48	Cow	Jenny Ladley	2012	Wakefield	South Island	
as49	Domestic pig	Jenny Ladley	2012	Wakefield	South Island	
as50	Seal	Virology lab group	2012	Half Moon Bay, Kaikoura	South Island	FasGCV-4
as51	Sheep	Virology lab group	2012	Half Moon Bay, Kaikoura	South Island	
as52	Cow	Virology lab group	2012	Kaikoura	South Island	

as53	Wild dog	Emily Tighe	2012	Orana Park, Christchurch	South Island	
as54	Wild dog	Emily Tighe	2012	Orana Park, Christchurch	South Island	
as55	Wild dog	Emily Tighe	2012	Orana Park, Christchurch	South Island	
as56	Hare	Laura Young	2011	Cass Basin	South Island	
as57	Hare	Laura Young	2011	Cass Basin	South Island	
as58	Bird	Laura Young	2011	Cass Basin	South Island	
as59	Possum	Laura Young	2011	Cass Basin	South Island	
as60	Bird	Laura Young	2011	Cass Basin	South Island	
as61	Possum	Laura Young	2011	Cass Basin	South Island	
as62	Pig	Laura Young	2011	Cass Basin	South Island	
as63	Possum	Laura Young	2011	Cass Basin	South Island	
as64	Pig	Laura Young	2011	Cass Basin	South Island	
P1	Black Robin	Melanie Massaro	2008	Rangatira/South East Island	Chatham Islands	
P2	Blackbird	Melanie Massaro	2008	Rangatira/South East Island	Chatham Islands	
P3	Tomtit	Melanie Massaro	2008	Rangatira/South East Island	Chatham Islands	
P4	Warbler	Melanie Massaro	2008	Rangatira/South East Island	Chatham Islands	
P5	Dunnock	Melanie Massaro	2008	Rangatira/South East Island	Chatham Islands	
P6	Silvereye	Melanie Massaro	2008	Rangatira/South East Island	Chatham Islands	
P7	Starling	Melanie Massaro	2008	Rangatira/South East Island	Chatham Islands	
P8	Black Robin	Melanie Massaro	2009	Rangatira/South East Island	Chatham Islands	
P9	Blackbird	Melanie Massaro	2009	Rangatira/South East Island	Chatham Islands	FasGCV-1 strain NZ_P9_Black_bird_2009
P10	Tomtit	Melanie Massaro	2009	Rangatira/South East Island	Chatham Islands	
P11	Warbler	Melanie Massaro	2009	Rangatira/South East Island	Chatham Islands	
P12	Dunnock	Melanie Massaro	2009	Rangatira/South East Island	Chatham Islands	
P13	Silvereye	Melanie Massaro	2009	Rangatira/South East Island	Chatham Islands	
P14	Starling	Melanie Massaro	2009	Rangatira/South East Island	Chatham Islands	FasGCV-10
P15	Black Robin	Melanie Massaro	2010	Rangatira/South East Island	Chatham Islands	
P17	Tomtit	Melanie Massaro	2010	Rangatira/South East Island	Chatham Islands	
P18	Warbler	Melanie Massaro	2010	Rangatira/South East Island	Chatham Islands	
P19	Dunnock	Melanie Massaro	2010	Rangatira/South East Island	Chatham Islands	
P20	Silvereye	Melanie Massaro	2010	Rangatira/South East Island	Chatham Islands	
P21	Black Robin	Melanie Massaro	2011	Rangatira/South East Island	Chatham Islands	FasGCV-8 and

						FasGCV-1 strain NZ_P21_Black_robin _2011
P22	Blackbird	Melanie Massaro	2011	Rangatira/South East Island	Chatham Islands	FasGCV-9
P23	Tomtit	Melanie Massaro	2011	Rangatira/South East Island	Chatham Islands	
P24	Warbler	Melanie Massaro	2011	Rangatira/South East Island	Chatham Islands	FasGCV-5, FasGCV-6, and FasGCV-3
P25	Dunnock	Melanie Massaro	2011	Rangatira/South East Island	Chatham Islands	
P26	Silvereye	Melanie Massaro	2011	Rangatira/South East Island	Chatham Islands	

5.2 Limitations of the study

This study has evaluated an inexpensive way to recover novel ssDNA viruses from environmental samples without relying on prior sequence knowledge, and thus DNA amplification using rolling circle amplification (RCA) via bacteriophage Phi29 DNA polymerase (Illustra^{TM} TempliPhi Kit, GE Healthcare) was the amplification technique selected. While this technique is more affordable than next generation sequencing and does not require specific primers, it does have a bias towards circular ssDNA templates. This limitation was understood before the study began, and since most ssDNA viruses have circular genomes it was accepted as a means to enrich the ssDNA in the sample and increase the likelihood of detection.

Another limitation was encountered when highly novel isolates had very low hits to known viruses in GenBank. When very little information is available on similar sequences it becomes too difficult to tentatively identify genes or to develop useful or insightful phylogenetic analysis. In such cases the logical response is to add the sequences to GenBank so that they are accessible to anyone for further research. In this manner, the sequences are still contributing to developing a greater understanding of ssDNA viral diversity because they will be available for future GenBank searches and may facilitate the research of another lab group.

5.1 Future research

The study in Chapter 4 has recovered and characterised novel ssDNA viruses from samples taken from a vulnerable species of parrot within New Zealand, the Yellow-crowned parakeet. Monitoring programs could be developed to further investigation the viral communities on and around these animals, in healthy and diseased conditions. As a cost-effective and non-invasive protocol has already been developed here, conservation organizations like the Department of Conservation could implement it easily with minimal personnel training being required for sample collection and sample analysis is relatively cheap. In this manner faecal samples can be collected from any of the protected birds that are being investigating in the field as part of the Natural Heritage Management System. The importance of these birds to New Zealand biodiversity highlights this as a particularly significant area to focus future research.

The aim of improving the resolution of the viral taxonomy is not likely to be achieved any time soon, and so there is plenty of opportunity for continued investigations of ssDNA viral diversity. Even so, illuminating the viral sequence space is only the beginning. Further explorations into the ecology of these viruses, including interactions with other viruses as well as with the environment, and changes in viral communities will be areas of research that will be possible once the viral sequence space is explored.

References

Abbate, E. A., Berger, J. M. & Botchan, M. R. (2004). The X-ray structure of the papillomavirus helicase in complex with its molecular matchmaker E2. *Genes & development* **18**, 1981-1996.

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.

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., Liu, H., Mahaffy, J. M., Mueller, J. E., Nulton, J., Olson, R., Parsons, R., Rayhawk, S., Suttle, C. A. & Rohwer, F. (2006). The marine viromes of four oceanic regions. *PLoS biology* **4**, e368-2131.

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

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**, 3507-3513.

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

Briddon, R. W., Martin, D. P., Owor, B. E., Donaldson, L., Markham, P. G., Greber, R. S. & Varsani, A. (2010). A novel species of mastrevirus (family Geminiviridae) isolated from Digitaria didactyla grass from Australia. *Archives of Virology* **155**, 1529-1534.

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

Chen, L., Rojas, M., Kon, T., Gamby, K., Xoconostle-Cazares, B. & Gilbertson, R. (2009). A severe symptom phenotype in tomato in Mali is caused by a reassortant between a novel recombinant begomovirus (Tomato yellow leaf curl Mali virus) and a betasatellite. *Molecular plant pathology* **10**, 415-430.

Cheung, A. K. (2004). Palindrome regeneration by template strand-switching mechanism at the origin of DNA replication of porcine circovirus via the rolling-circle melting-pot replication model. *Journal of Virology* **78**, 9016-9029.

Cheung, A. K. (2006). Rolling-circle replication of an animal circovirus genome in a theta-replicating bacterial plasmid in Escherichia coli. *Journal of Virology* **80**, 8686-8694.

Clérot, D. & Bernardi, F. (2006). DNA helicase activity is associated with the replication initiator protein rep of tomato yellow leaf curl geminivirus. *Journal of Virology* **80**, 11322-11330.

Cooper, R. A. & Millener, P. R. (1993). The New Zealand biota: Historical background and new research. *Trends in Ecology & Evolution* **8**, 429-433.

Cotmore, S. F. & Tattersall, P. (1996). Parvovirus DNA Replication. *Cold Spring Harbor Monograph Archive* **31**, 799-813.

Dayaram, A., Goldstien, S., Zawar-Reza, P., Gomez, C., Harding, J. S. & Varsani, A. (2013). Novel ssDNA 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., Jaschke, A., Hadfield, J., Baschiera, M., Dobson, R. C., Offei, S. K., Shepherd, D. N., Martin, D. P. & Varsani, A. (2012). Molecular characterisation of a novel cassava associated circular ssDNA virus. *Virus Res* **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.

Donaldson, E. F., Haskew, A. N., Gates, J. E., Huynh, J., Moore, C. J. & Frieman, M. B. (2010). Metagenomic analysis of the viromes of three North American bat species: viral diversity among different bat species that share a common habitat. *Journal of Virology* **84**, 13004-13018.

Duffy, S. & Holmes, E. C. (2009). Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. *Journal of General Virology* **90**, 1539-1547.

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.

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

Esteban, J., Salas, M. & Blanco, L. (1993). Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. *Journal of Biological Chemistry* **268**, 2719-2726.

Fancello, L., Trape, S., Robert, C., Boyer, M., Popgeorgiev, N., Raoult, D. & Desnues, C. (2012). Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. *The ISME Journal* **7**, 359-369.

Fanning, E. & Knippers, R. (1992). Structure and function of simian virus 40 large tumor antigen. *Annual review of biochemistry* **61**, 55-85.

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

Garkavenko, O., Elliott, R. B. & Croxson, M. C. (2005). Identification of pig circovirus type 2 in New Zealand pigs. *Transplantation proceedings* **37**, 506-509.

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.

Ge, X., Li, Y., Yang, X., Zhang, H., Zhou, P., Zhang, Y. & Shi, Z. (2012). Metagenomic analysis of viruses from bat fecal samples reveals many novel viruses in insectivorous bats in China. *Journal of Virology* **86**, 4620-4630.

Geering, A. W., Thomas, J., Holton, T., Hadfield, J. & Varsani, A. (2012). Paspalum striate mosaic virus: an Australian mastrevirus from Paspalum dilatatum. *Archives of Virology* **157**, 193-197.

Gilbert, W. & Dressler, D. (1968). DNA replication: the rolling circle model. *Cold Spring Harbor Symposia on Quantitative Biology* **33**, 473-484.

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.

Grigoras, I., Timchenko, T., Grande-Pérez, A., Katul, L., Vetten, H.-J. & Gronenborn, B. (2010). High variability and rapid evolution of a nanovirus. *Journal of Virology* **84**, 9105-9117.

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.

Hadfield, J., Martin, D. P., Stainton, D., Kraberger, S., Owor, B. E., Shepherd, D. N., Lakay, F., Markham, P. G., Greber, R. S. & Briddon, R. W. (2011). Bromus catharticus striate mosaic virus: a new mastrevirus infecting Bromus catharticus from Australia. *Archives of Virology* **156**, 335-341.

Hadfield, J., Thomas, J. E., Schwinghamer, M. W., Kraberger, S., Stainton, D., Dayaram, A., Parry, J. N., Pande, D., Martin, D. P. & Varsani, A. (2012). Molecular characterisation of dicot-infecting mastreviruses from Australia. *Virus research* **166**, 13-22.

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.

Hale, K. A. & Briskie, J. V. (2007). Decreased immunocompetence in a severely bottlenecked population of an endemic New Zealand bird. *Animal Conservation* **10**, 2-10.

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**, 3279-3285.

Inoue-Nagata, A. K., Albuquerque, L. C., Rocha, W. B. & Nagata, T. (2004). A simple method for cloning the complete begomovirus genome using the bacteriophage φ29 DNA polymerase. *Journal of virological methods* **116**, 209-211.

James, J. A., Escalante, C. R., Yoon-Robarts, M., Edwards, T. A., Linden, R. M. & Aggarwal, A. K. (2003). Crystal structure of the SF3 helicase from adeno-associated virus type 2. *Structure* **11**, 1025-1035.

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.

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

Khan, S. A. (1997). Rolling-circle replication of bacterial plasmids. *Microbiology and molecular biology reviews* **61**, 442-455.

Kim, H. K., Park, S. J., Nguyen, V. G., 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.

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.

Kim, M. S., Park, E. J., Roh, S. W. & Bae, J. W. (2011). Diversity and abundance of single-stranded DNA viruses in human feces. *Applied and environmental microbiology* **77**, 8062-8070.

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

Koonin, E. V. & Ilyina, T. V. (1993). Computer-assisted dissection of rolling circle DNA replication. *Biosystems* **30**, 241-268.

Koonin, E. V. (1993). A common set of conserved motifs in a vast variety of putative nucleic acid-dependent ATPases including MCM proteins involved in the initiation of eukaryotic DNA replication. *Nucleic acids research* **21**, 2541-2547.

Kraberger, S., Thomas, J. E., Geering, A. D., Dayaram, A., Stainton, D., Hadfield, J., Walters, M., Parmenter, K. S., van Brunschot, S. & Collings, D. A. (2012). Australian monocot-infecting mastrevirus diversity rivals that in Africa. *Virus research*.

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K. & Hugenholtz, P. (2008). A bioinformatician's guide to metagenomics. *Microbiology and molecular biology reviews* **72**, 557-578.

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

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* **92**, 3879.

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.

Li, L., Kapoor, A., Slikas, B., Bamidele, O. S., Wang, C., Shaukat, S., Masroor, M. A., Wilson, M. L., Ndjango, J. B., Peeters, M., Gross-Camp, N. D., Muller, M. N., Hahn, B. H., Wolfe, N. D., Triki, H., Bartkus, J., Zaidi, S. Z. & Delwart, E. (2010a). Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces. *Journal of Virology* **84**, 1674-1682.

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

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

Liu, H., Fu, Y., Li, B., Yu, X., Xie, J., Cheng, J., Ghabrial, S. A., Li, G., Yi, X. & Jiang, D. (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.

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.

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., Taylor, G., Greene, T., van de Wetering, J., van de Wetering, M., Pryde, M., Dilks, P., Heber, S., Steeves, T. E., Walters, M., Shaw, S., Ortiz-Catedral, L., Potter, J., Farrant, M., Brunton, D. H., Hauber, M., Jackson, B., Bell, P., Moorhouse, R., McInnes, K., Varsani, A., Julian, L., Galbraith, J. A., Kurenbach, B., Kearvell, J., Kemp, J., van Hal, J. & Elkington, S. (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.

Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S. A., Wu, G. D., Lewis, J. D. & Bushman, F. D. (2011). The human gut virome: Inter-individual variation and dynamic response to diet. *Genome research* **21**, 1616-1625.

Miskelly, C. M., Dowding, J. E., Elliott, G. P., Hitchmough, R. A., Powlesland, R. G., Robertson, H. A., Sagar, P. M., Scofield, R. P. & Taylor, G. A. (2008). Conservation status of New Zealand birds, 2008. *Notornis* **55**, 117-135.

Mochizuki, T., Krupovic, M., Pehau-Arnaudet, G., Sako, Y., Forterre, P. & Prangishvili, D. (2012). Archaeal virus with exceptional virion architecture and the largest single-stranded DNA genome. *Proceedings of the National Academy of Sciences* **109**, 13386-13391.

Muhire, B., Martin, D. P., Brown, J. K., Navas-Castillo, J., Moriones, E., Zerbini, F. M., Rivera-Bustamante, R., Malathi, V., Briddon, R. W. & Varsani, A. (2013). A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). *Archives of Virology*, 1-14.

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. *Science Signaling* **85**, 1182.

Nelson, J. R., Cai, Y. C., Giesler, T. L., Farchaus, J. W., Sundaram, S. T., Ortiz-Rivera, M., Hosta, L. P., Hewitt, P. L., Mamone, J. A. & Palaniappan, C. (2002). TempliPhi, phi29 DNA polymerase based rolling circle amplification of templates for DNA sequencing. *BioTechniques* **32**, S44-S47.

Ng, T. 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., Manire, C., Borrowman, K., Langer, T., Ehrhart, L. & Breitbart, M. (2009a). 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., 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., Suedmeyer, W. K., Wheeler, E., Gulland, F. & Breitbart, M. (2009b). Novel anellovirus discovered from a mortality event of captive California sea lions. *Journal of General Virology* **90**, 1256-1261.

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

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.

Ortiz-Catedral, L., Kurenbach, B., Massaro, M., McInnes, K., Brunton, D. H., Hauber, M. E., Martin, D. P. & Varsani, A. (2010). A new isolate of beak and feather disease virus from endemic wild red-fronted parakeets (Cyanoramphus novaezelandiae) in New Zealand. *Archives of virology* **155**, 613-620.

Ortiz-Catedral, L., McInnes, K., Hauber, M. E. & Brunton, D. H. (2009). First report of beak and feather disease virus (BFDV) in wild Red-fronted Parakeets (Cyanoramphus novaezelandiae) in New Zealand. *Emu* **109**, 244-247.

Palmer, K. E. & Rybicki, E. P. (1998). The molecular biology of mastreviruses. *Advances in virus research* **50**, 183-234.

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.

Piasecki, T., Kurenbach, B., Chrzastek, K., Bednarek, K., Kraberger, S., Martin, D. P. & Varsani, A. (2012). Molecular characterisation of an avihepadnavirus isolated from Psittacula krameri (ring-necked parrot). *Archives of Virology* **157**, 585-590.

Pietilä, M. K., Roine, E., Paulin, L., Kalkkinen, N. & Bamford, D. H. (2009). An ssDNA virus infecting archaea: a new lineage of viruses with a membrane envelope. *Molecular microbiology* **72**, 307-319.

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

Pipas, J. M., Peden, K. & Nathans, D. (1983). Mutational analysis of simian virus 40 T antigen: isolation and characterization of mutants with deletions in the T-antigen gene. *Molecular and cellular biology* **3**, 203-213.

Pita, J., Fondong, V., Sangare, A., Otim-Nape, G., Ogwal, S. & Fauquet, C. (2001). Recombination, pseudorecombination and synergism of geminiviruses are determinant keys to the epidemic of severe cassava mosaic disease in Uganda. *Journal of General Virology* **82**, 655-665.

Rector, A., Tachezy, R. & Van Ranst, M. (2004). A sequence-independent strategy for detection and cloning of circular DNA virus genomes by using multiply primed rolling-circle amplification. *Journal of Virology* **78**, 4993-4998.

Rokyta, D., Burch, C., Caudle, S. & Wichman, H. (2006). Horizontal gene transfer and the evolution of microvirid coliphage genomes. *Journal of bacteriology* **188**, 1134-1142.

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

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

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. (2012b). A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Archives of Virology* **157**, 1851-1871.

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

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.

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., Colombet, J., Sime-Ngando, T. & Debroas, D. (2012a). Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. *PloS one* **7**, e33641.

Roux, S., Krupovic, M., Poulet, A., Debroas, D. & Enault, F. (2012b). Evolution and Diversity of the Microviridae Viral Family through a Collection of 81 New Complete Genomes Assembled from Virome Reads. *PloS one* **7**, e40418.

Sachsenröder, J., Twardziok, S., Hammerl, J. A., Janczyk, P., Wrede, P., Hertwig, S. & Johne, R. (2012). Simultaneous Identification of DNA and RNA Viruses Present in Pig Faeces Using Process-Controlled Deep Sequencing. *PloS one* **7**, e34631.

Saunders, K., Lucy, A. & Stanley, J. (1991). DNA forms of the geminivirus African cassava mosaic virus consistent with a rolling circle mechanism of replication. *Nucleic acids research* **19**, 2325-2330.

Schneider, J. & Fanning, E. (1988). Mutations in the phosphorylation sites of simian virus 40 (SV40) T antigen alter its origin DNA-binding specificity for sites I or II and affect SV40 DNA replication activity. *Journal of Virology* **62**, 1598-1605.

Shackelton, L. A., Hoelzer, K., Parrish, C. R. & Holmes, E. C. (2007). Comparative analysis reveals frequent recombination in the parvoviruses. *Journal of General Virology* **88**, 3294-3301.

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

Shen, J., Gai, D., Patrick, A., Greenleaf, W. B. & Chen, X. S. (2005). The roles of the residues on the channel β-hairpin and loop structures of simian virus 40 hexameric helicase. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 11248-11253.

Shepherd, D. N., Martin, D. P., Lefeuvre, P., Monjane, A. L., Owor, B. E., Rybicki, E. P. & Varsani, A. (2008). A protocol for the rapid isolation of full geminivirus genomes from dried plant tissue. *Journal of Virological Methods* **149**, 97-102.

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

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.

Stenger, D. C., Revington, G. N., Stevenson, M. C. & Bisaro, D. M. (1991). Replicational release of geminivirus genomes from tandemly repeated copies: evidence for rolling-circle replication of a plant viral DNA. *Proceedings of the National Academy of Sciences* **88**, 8029-8033.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**, 2731-2739.

Tattersall, P. & Ward, D. C. (1976). Rolling hairpin model for replication of parvovirus and linear chromosomal DNA. *Nature* **263**, 106.

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

Van Der Walt, E., Martin, D. P., Varsani, A., Polston, J. E. & Rybicki, E. P. (2008). Experimental observations of rapid Maize streak virus evolution reveal a strand-specific nucleotide substitution bias. *Virology journal* **5**, 104.

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.

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. *The 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**, 945-951.

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.

Woolhouse, M. & Gaunt, E. (2007). Ecological Origins of Novel Human Pathogens. *Critical Reviews in Microbiology* **33**, 231-242.

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 (New York, NY* **332**, 714-717.

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.

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. & Huang, J. (2010). A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. *Proceedings of the National Academy of Sciences* **107**, 8387-8392.

Yu, X., Li, B., Fu, Y., Xie, J., Cheng, J., Ghabrial, S. A., Li, G., Yi, X. & Jiang, D. (2013). Extracellular transmission of a DNA mycovirus and its use as a natural fungicide. *Proceedings of the National Academy of Sciences* **110**, 1452-1457.