PHYLOGENOMICS, BIOGEOGRAPHY, AND ECOLOGY OF CIRCULAR REP-ENCODING SSDNA VIRUSES AFFILIATED WITH MICROCRUSTACEANS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Kalia Sakaye Irinaga Bistolas

August 2019

PHYLOGENOMICS, BIOGEOGRAPHY, AND ECOLOGY OF CIRCULAR REP-ENCODING SSDNA VIRUSES AFFILIATED WITH MICROCRUSTACEANS

Kalia Sakaye Irinaga Bistolas, Ph.D.

Cornell University 2019

Small crustaceans populate nearly every aquatic ecosystem, often representing the most diverse and abundant component of metazoan communities. These taxa comprise a rich lexicon of unexplored viral diversity. Circular replication initiator protein-encoding ssDNA (CRESS-DNA) viruses are cosmopolitan members of microcrustacean viral consortia. Despite their prevalence among invertebrates, the paradigms that govern CRESS-DNA virus phylogenomics, biogeography, and ecology in arthropods are largely unknown. This dissertation couples viromic sequencing, viral surveillance, and metazoan transcriptomics to examine novel ssDNA virus genotypes, infer the influence of metazoan phylogenetics on viral distribution, and assess the relationship between these viruses and the biology of putative invertebrate hosts. Viromes from microcrustacean populations from disparate aquatic habitats were sequenced to compare CRESS-DNA virus genomes on a multi-ecosystem scale, enabling identification of 215 putatively novel and microdiverse genotypes. Identification of endogenized viral elements established a paleovirological record of historical arthropod-virus interactions. Pairwise sequence comparison indicated that novel CRESS-DNA viruses shared similarity with genotypes recovered from similar microcrustaceans. Several genotypes – specifically those associated with benthic amphipods of genus *Diporeia* – were further explored to resolve the role of microcrustacean speciation on CRESS-DNA virus biogeography and determine the potential influence of viral load on microcrustacean ecology. Among three identified viral genotypes associated with *Diporeia*

spp., one (LM29173) was both prevalent and recurrent among amphipod populations in the Laurentian Great Lakes. Occurrence of this genotype coincided with amphipod haplotype demographics, indicating the potential for host specificity. Transcriptomes from both amphipod haplotype were compared, with load of LM29173 corresponding with significant over- or under-expression of >2,000 *de novo* assembled amphipod transcripts, though no change in microcrustacean nutritional quality was detected. Collectively, these studies demonstrate the diversity of CRESS-DNA viruses among microcrustaceans and address fundamental questions about their ecogenomics in non-model arthropod systems.

BIOGRAPHICAL SKETCH

In 2014, Kalia Bistolas joined the Department of Microbiology at Cornell University, where she studies the ecogenomics of ssDNA viruses associated with aquatic invertebrates under the direction of Dr. Ian Hewson. Her doctoral research couples molecular methods with bioinformatic approaches to foster discovery of new viruses and explore their ecology and biogeography. Kalia holds a bachelors degree in Biology from Occidental College (Los Angeles, CA) where she graduated *summa cum laude* studying symbioses between microbes and neotropical insects. As an undergraduate, she was also involved in characterizing the effects of elevated pCO_2 on the biochemistry and physiology of marine microalgae (Western Washington University, Anacortes, WA). Her experiences prior to graduate school helped spark her interest in environmental microbiology and viral ecology, which has been the focus of her research pursuits ever since. Her graduate research career has sent Kalia across the globe, deep into the lakes in Cornell's backyard, and into a maze of nested virtual machines. She intends to use her foundation in both fieldwork and bioinformatics to continue investigating how viruses interact with non-model hosts in aquatic ecosystems.

To the generations of women who came before me, upon whose sacrifices I stand to scale this summit

"In the universe suddenly restored to its silence, the myriad wondering little voices of the earth rise up...each atom of that stone, each mineral flake of that night-filled mountain, in itself forms a world."

– A. Camus

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my Ph.D. advisor, Dr. Ian Hewson, for the fundamental role he has played in my doctoral studies. Ian has provided me with the guidance, mentorship, and expertise that I needed, while allowing me the independence to explore new research interests. He has been an invaluable source of advice, encouragement, and support. His patience with me cannot be overstated and I look forward to future collaboration.

This dissertation was guided by the many intellectual contributions of my committee members, Professors Lars Rudstam and Gary Blissard. Their instruction and counsel have been vital to my professional development and shaped the direction of my research. I am thankful to Professors Esther Angert, Tory Hendry, and Dan Buckley, among others in the Department of Microbiology, for their continued input and instruction in research and teaching.

I would like to thank my colleague, Elliot Jackson, for his support and advice through the years - his research efforts have been essential to my graduate education and his camaraderie has been essential to preserving my sanity. I am also greatly indebted to Francine Arroyo, Ryan Besemer, Chris Derito, Alex Vompe, Jason Button, Erin Eggleston, Mitchell Johnson, Mike Petassi, Jordan Rede, Julie Brown, Sue Pierre, Armanda Rocco, Melanie Smee, and fellow participants of the R/V Blue Heron UNOLS Chief Scientist training cruise (2017), among many others for both their academic assistance and moral support. Thanks also to Lillian Henry, Patti Brenchley, Tina Daddona, Shirley Cramer, Michelle Cole, Joe Peters, and the staff at CIRTL for their administrative support, and the Cornell University Genomics Facility (BRC) and Stable Isotope Laboratory (COIL) for sequencing and elemental analysis services and Jim Watkins for his expertise and collection help. Thank you to my wonderful Ithaca family, especially the Leskys and Weinbergers, for their gentle encouragement and endless humor. I am grateful for Professors S. Goffredi, G. North, and B. Braker for their unfounded, yet unwavering belief that I can grow into something resembling a scientist.

Finally, my graduate studies would not have been possible without family – by blood and by choice. From the first moment that I experimented with motor oil and sand in the garage to the drive down the coast in search of “sea bugs,” they have been there, keeping me going through a satellite in space.

This research was supported by NSF – 135696 and DGE-1144153 with additional support from the U.S. Environmental Protection Agency, Cooperative Agreement GL 00E01184-0 to Cornell.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACT	β-Actin
ANIb	Average Nucleotide Identity by BLAST
CAI	Codon Adaptation Index
Cap	CRESS-DNA virus structural capsid-encoding gene
cDNA	Complementary DNA
CRESS-DNA	Circular Replication protein Encoding Single Stranded DNA (virus)
Ct	Cycle Threshold
DEG	Differentially Expressed Gene
EF1A	Elongation Factor-1α
ENC, or N_c	Effective Number of Codons
EVE	Endogenized Viral Element
FDR	False Discovery Rate
HUH	Histidine-hydrophobe-Histidine endonuclease motif
ICTV	International Committee on Taxonomy of Viruses
KOG	euKaryotic Orthologous Group
MCV	Microcrustacean CRESS-DNA Virus-like contig
MNV	Multiple Nucleotide Variant
NMHC	Non-Muscular myosin Heavy Chain
No-RT	No Reverse Transcription control
NTC	No Template Control
ORF	Open Reading Frame
PCV	Porcine Circovirus
PMWS	Postweaning Multisystemic Wasting Syndrome
qPCR	Quantitative Polymerase Chain Reaction
RCR	Rolling Circle Replication
Rep	CRESS-DNA virus replication initiator gene
RPKM	Reads Per Kilobase of transpcript per Million mapped reads
RSCU	Relative Synonymous Codon Usage
RT-qPCR	Reverse Transcriptase Quantitative Polymerase Chain Reaction
SE	Standard Error
SF3	Superfamily 3 helicase domain
SNV	Single Nucleotide Variant
SRA	Short Read Archive
UBQ	Ubiquitin-conjugating enzyme E2
VLP	Virus-Like Particle
%GC	Percent guanine-cytosine content
%GC3	Percent guanine-cytosine content at the 3rd codon position
%IDrep	Pairwise percent identity of the *rep* ORF

CHAPTER 1

INTRODUCTION

The collective sum of aquatic viruses contain more than several hundred thousand gigabases of unique genetic information, much of which remains uncharted (Middelboe & Brussaard, 2017; Steward et al, 2000; Suttle, 2007). The quantity of viruses within this "virosphere" outstrips the number of stars in the known universe by ~10 millionfold, comprising both a vast genetic reservoir and a profoundly influential force on the ecology and evolution of their cellular hosts (Mann, 2005; Thingstad, 2000). In 1892, Dmitri Ivanovsky first described viruses as "filterable infectious agents", entities that were unseen, inert but infectious, neither alive nor dead (Flint et al, 2015; Iwanowski et al, 1892). In a correspondingly elusive fashion, the often catastrophic effects of these obligate parasites are frequently observed long before the causative agent is identified and characterized. For example, the causative agents of smallpox and polio epidemics (variola virus and poliovirus, respectively) were both discovered or isolated in the 18th and 19th centuries, despite origins in Mesopotamia and early Egypt (1000-1500BCE; Geddes et al, 2006; Nathanson & Kew et al, 2010). Therefore, throughout history, understanding or mitigating the consequences of infection among humans or agroeconomically-relevant species has taken precedence over disentangling the intrinsic ecology of viral populations (Flint et al, 2015). However, once an approach was devised to directly enumerate viruses in aquatic ecosystems, it became apparent that viruses were not only numerically dominant in the natural world, but likely have a much greater impact on the ecology of cellular organisms than culture-based plaque assays implied (Bergh et al, 1989). In recent years, renewed efforts to define the ecology of viruses associated with non-model organisms and increased accessibility to the tools required to examine them has led to a more comprehensive understanding of the central role these viruses play in global nutrient cycling and their contribution to ecosystem structure and function.

1.1 Ecological contribution of viruses in the hydrosphere | The quantity of carbon (C) contained within aquatic viruses is equivalent to over 9.5 billion people (assuming 0.2 femtograms C

per free virus-like particle and 21kg C per adult human; West et al, 2009; Jover et al, 2014). Carbon, nitrogen (N), phosphorus (P), sulfur (S), and other trace elements sequestered as viral biomass are labile components of the hydrosphere, with degradation, propagation, and turnover occurring on the order of hours to weeks in surface water. Yet beyond their static elemental stoichiometry, viruses also likely infect all metazoans and a significant quantity of the $>10^{29}$ metabolically active microbial cells in aquatic ecosystems, amending their role in biogeochemical cycling. In total, viruses are likely responsible for 8-43% of microbial mortality through lytic infection, imposing top-down control on cellular communities (Fuhrman, 1999). Therefore, more important than the standing stock of elements captured within viral particles is arguably the composition and quantity of cellular debris produced through viral lysis. Viral lysate, operationally defined by size as dissolved or particulate organic matter (DOM or POM, respectively) is often sequestered (POM) or recycled via the "viral shunt," (both DOM and POM) facilitating further productivity. This viral shunt refers to the wholesale lysis of microbial cells by bacteriophage, effectively redirecting bioavailable resources towards re-assimilation by microbial cells and bypassing utilization by upper trophic level consumers. Models suggest that approximately 27% of bacterial respiration is stimulated by this shunt (Fuhrman, 1999). N and P-rich products of cell lysis are readily incorporated into new nucleic acids and organically-complexed iron is accrued by nearby siderophores. Concurrently, viral lysis of abundant microbial populations curtails dominant taxa in cyclical boom-and-bust infection dynamics, redistributing resources and maintaining microbial diversity (referred to as the "kill-the-winner" hypothesis). Via this mechanism, the viral shunt provides bottom-up control of microbial community productivity, recycling 6-26% of photosynthetically fixed organic C for uptake by heterotrophic bacteria, ultimately resulting in a net effect of oxidizing organic matter and replenishing inorganic nutrients for utilization by other taxa (Fuhrman, 1999).

Viruses of metazoans are responsible for a significant component of nonconsumptive (i.e. non-predatory) mortality, driving POM sequestration, triggering zoonotics, and reorganizing community trophic structure in relation to viral host susceptibility (de Lorgeril, 2018; Munn 2006).

By some estimates, viruses turn over as much as $1.5x10^{14}$ kg carbon $year^{-1}$ stored in macro- and microorganisms, equivalent to 30x the standing quantity of C in the ocean (Fuhrman, 1999; Jover et al, 2014; Thingstad, 2000). The mechanisms employed by viruses to execute this magnitude of nutrient turnover underpin the complex, continuously evolving interactions between metazoans, microbes, and their abiotic environment (Breitbart, 2012; Lima-Mendez et al, 2015; Rohwer & Vega Thurber, 2009; Suttle, 2007). We have reached an era where we may begin to parse both the scope of viral diversity and chart the operation of individual viral particles in fluid and tissue matrices.

1.2 Value of viral discovery **|** The advent of high-throughput sequencing (HTS) technologies has been instrumental in describing spatiotemporal patterns of viral populations, subverting limitations imposed by the lack of universally conserved "marker genes" (Brum & Sullivan, 2015; Brum, 2016; Roux et al, 2015). Early studies of environmental viral populations largely relied on *a priori* information to target a subset of viruses with known molecular signatures (Edwards & Rohwer, 2005). For example, algal-virus-like DNA polymerase gene homology allowed identification of potentially novel members of the Phycodnaviridae (Chen et al, 1996), while the ubiquity of structural gene g20 among cyanophage facilitated discovery of new populations in distinct ecosystems (Short & Suttle, 2004). While critical to investigating the ecology of somewhat established viral groups and still widely and effectively used today, these methods served as frontrunners for a generation of technologies dedicated to the untargeted capture of whole viral communities ("viromes") on scales ranging from specific organismal tissues to complete biomes. Co-development of tools to purify virus-like particles (VLPs), aid sequence independent and near-quantitative nucleic acid amplification, and facilitate computational assembly/annotation of novel viral genes and genomes has stimulated the emergence of viral discovery efforts on par with early pursuits to sequence the human microbiome (Roux et al, 2016). These studies suggest that the majority of sequence diversity on this planet perhaps exists below the cellular size fraction (Gregory et al, 2019).

However, our understanding of the permutations of viral genomic diversity and their

ecological significance is restricted by our inability to identify new viral sequences and determine the nature of their association with host taxa. These functionally and taxonomically unannotated sequences, or "viral dark matter," currently comprise 63–93% of viral sequence space (Brum & Sullivan, 2015; Hurwitz & Sullivan, 2013). Although this ratio is rapidly diminishing through consortia-based sequencing initiatives (Angly et al, 2006; Brum et al, 2013; Hurwitz & Sullivan, 2013; Karsenti et al, 2011; Needham et al, 2013; Paez-Espino et al, 2019, Parsons et al, 2013), these unannotated sequences currently impede sensitive and accurate characterization of viral community structure and evolution (Roux et al, 2015). Despite the known shortcomings of data exclusion (i.e. the omission of draft sequences failing to meet annotation or metadata requirements for sequence dissemination), only a miniscule fraction of viral sequences detected and hypothetically annotated in viromes are represented in accessible data repositories. This is predominantly due to a lack of sufficient validation and obstructs comparative analyses between ecosystems (Roux et al, 2019). Recently, successful sequence annotation using innovative computational techniques (such as proteomic capsid subunit clustering, oligomer signature network inference, profile hidden Markov models queries, among others) has begun to provide corroborative validation, allowing a more holistic perspective of viral diversity. Predictably, this has released a deluge of viral discoveries beyond known taxonomic families (Gregory et al, 2019; Paez-Espino et al, 2019). In alignment with these advances, the International Committee on the Taxonomy of Viruses (ICTV) has admitted viral sequences discovered via viromic sequencing into the viral taxonomic hierarchy that demonstrate the necessary qualities of bona fide viruses but fail to be cultured or isolated (Simmonds et al, 2017). Comparative viromics and genomic surveillance within this taxonomic framework is broadly applicable, lending insight into viral ecology across a breadth of scales, including: (1) community composition (e.g. macroecology, co-infection dynamics, β-diversity, biogeography, etc), (2) population ecology (e.g. microdiversity, quasispecies delineation, variant selection, etc), and (3) gene flow (e.g. transmission/horizontal gene transfer, protein clustering, etc). Together, discovery of novel genotypes begins to illuminate the intimate affiliation between viruses and their hosts and infer the

mechanisms that precipitate ongoing strategies to evade, resist, utilize, or otherwise adapt to coexisting populations (Middelboe & Brussarard, 2017; Rohwer, 2009).

1.3 Viruses with circular ssDNA genomes are ubiquitous **|** Widespread viromic sequencing has revealed the unprecedented diversity of viruses with ssDNA genomes. Among these, those encoding a characteristic homologous replication initiation protein (*rep*) within a single, circular DNA molecule arranged in either a sense or ambisense orientation represents the most common ssDNA viral supergroup. These "circular replication initiator protein-encoding single stranded DNA viruses," abbreviated "CRESS-DNA viruses," are highly abundant, infecting all domains of cellular life, from bacteria (e.g. bacteriophage; *Microviridae, Inoviridae*) and archaea (*Pleolipoviridae*, possibly *Smacoviridae*; Díez-Villaseñor & Rodriguez-Valera, 2019) to eukaryotes (*Bacilladnaviridae, Circoviridae, Geminiviridae, Genomoviridae, Nanoviridae*, possibly *Smacoviridae;* Rosario et al, 2012; Varsani & Krupovic, 2018; Zhao et al, 2019). Indeed, this supergroup comprises the majority of eukaryotic ssDNA viral families delineated by ICTV (six of seven as of June, 2019, with anelloviruses encoding a nonhomologous but functionally similar *rep*). One viromic study identified over a hundred previously undescribed, genetically unique groups of ssDNA viruses in a single sequencing effort, capturing novel viral groups from marine ecosystems in tropics to the subtropics (Labonté & Suttle, 2013). Another study (Ng et al, 2015) identified novel CRESS-DNA viruses in a cohort of humans with acute gastroenteritis, finding indication of these ssDNA viruses in 28-67% of samples, whereas Wang et al (2018) found genotypes in cattle herds, with proposed implications for livestock yield. CRESS-DNA viruses are also globally pervasive elements of natural reservoirs, and have been found in agricultural soils (Kim et al, 2008), benthic sediments (Yoshida et al, 2013), sewage effluent (Kraberger et al, 2015), and water columns (López-Bueno et al, 2009; Rosario et al, 2009; Zawar-Reza et al, 2014) from ecosystems across a range of latitudes. These genomes are found in publicly available viromes and metagenomes from around the word, in nearly every niche that has been investigated (for partial representation, see Table 1.1).

Vertebrate		
Mammalia		
Carnivora	**Mustelidae**	*van den Brand et al, 2012*
	(pine marten, mink, ferret badger)	*Zaccaria et al, 2016*
		Lian et al, 2014
		Smits et al, 2013
	Canidae	*Li et al, 2013*
	(fox, dog)	*Zaccaria et al, 2016*
	Hominidae (chimpanzee)	*Blinkova et al., 2010*
		Li et al, 2010
	Otariidae (fur seal)	*Sikorski et al, 2013*
Rodentia	**Muroidea**	*Hansen et al, 2015*
	(rat, mouse)	*Sachsenroder et al, 2014*
		Phan et al, 2011
Artiodactyla	**Bovidae**	*Li et al, 2010*
	(cow, goat)	*Kim et al, 2012*
	Camelidae (camel)	*Woo et al, 2014*
Chiroptera	**Pteropodidae**	*Ge et al, 2011,*
		Li et al, 2010,
	Rhinolophidae	*Lima et al, 2015*
	Hipposideridae	
	Vespertilionidae	
	Miniopteridae	
	Molossidae (bat)	
Reptilia & Amphibia		
Testudines	**Chelonioidea** (sea turtle)	*Ng et al, 2009*
Anura	**Hylidae** (treefrog)	*Tarján et al, 2014*
	Bufonidae (toad)	*Tarján et al, 2014*
Squamata	**Helodermatidae** (gila monster)	*Somayaji et al, 2018*
	Pythonidae (python)	*Altan et al, 2019*
Actinopterygii		
Gobiiformes	**Gobiidae** (goby)	*Tarján et al, 2014*
Cypriniformes	**Cyprinidae** (bream, barbel)	*Lorincz et al, 2011*
Anguilliformes	**Anguillidae** (European eel)	Doszpoly et al, 2014

Invertebrate		
Cnidaria		
Scleractinia	**Merulinidae** (stony coral)	*Soffer et al, 2014*
Alcyonacea	**Plexauridae**	*Rosario et al, 2015*
	Primnoidae (soft coral)	
Porifera		
Haplosclerida	**Lubomirskiidae** (sponge)	*Butina et al, 2019*
Echinodermata		
Forcipulatida	**Asteriidae** (echinoderm)	*Fahsbender et al, 2015*
Temnopleuroida	**Toxopneustidae** (urchin)	*Rosario et al, 2015*
Mollusca		
Basommatophora	**Amphibolidae** (mollusc)	*Dayaram et al, 2013b*
Venerida	**Veneridae**	*Dayaram et al, 2013c*
	Mesodesmatidae	*Dayaram et al, 2015,2016*
Mytilida	**Mytilidae** (mussel)	*Rosario et al, 2015*
Neotaenioglossa	**Littorinidae** (snail)	*Rosario et al, 2015*
Ctenophora		
Lobata	**Bolinopsidae**	*Breitbart et al, 2015*
Beroida	**Beroidae** (ctenophore)	
Arthropoda		
Amphipoda	**Pontoporeiidae**	*Hewson et al, 2013b*
		Bistolas et al, 2017
	Gammaridae	*Rosario et al, 2015*
Decapoda	**Palaemonidae**	*Rosario et al, 2015*
	Sicyoniidae	
	Portunidae	
	Diogenidae	
	Penaeidae (shrimp/crab)	*Pham et al, 2014*
		Ng et al, 2013
Cladocera	**Daphniidae**	*Hewson et al, 2013a*
Calanoida	**Acartiidae**	*Dunlap et al, 2013*
	Pontellidae (copepod)	*Eaglesham & Hewson, 2013*
Coleoptera	**Gyrinidae**	*Rosario et al, 2018*
	Curculionidae (beetle)	
Hemiptera	**Aleyrodidae** (whitefly)	*Nakasu et al, 2017*
Diptera	**Culicidae**	*Garigliany et al, 2015*
	(mosquito)	*Ng et al, 2011*
	Simuliidae	*Kraberger et al, 2019a*

Aves		
Anseriformes	**Anatidae** (duck, swan, goose)	*Halami et al, 2008*
		Stenzel et al, 2015
		Matczuk et al, 2015
		Zhang et al, 2013
		Hattermann et al, 2003
		Cha et al, 2014
	Laridae (gull)	*Todd et al, 2007*
Charadriiformes	**Phasianidae** (turkey)	*Reutger et al, 2014*
Galliformes	**Columbidae** (columbid)	*Mankertz et al, 2000*
Columbiformes	**Corvidae** (raven)	*Stewart et al, 2006*
Passeriformes	**Fringillidae** (finch)	*Rinder et al, 2015*
	Sturnidae (starling)	*Johne et al, 2006*
	Paridae (chickadee)	*Hanna et al, 2015*
	Fanniidae	*Rosario et al, 2018*
	Calliphoridae (fly)	
Hymenoptera	**Apidae** (honeybee)	*Kraberger et al, 2019b*
	Solenopsis	*Rosario et al, 2018*
	Formicidae (ant)	
Odonata	**Aeshnidae**	*Dayaram et al, 2013*
	Coenagrionidae	*Rosario et al, 2012*
	Corduliidae	*Rosario et al, 2011*
	Libellulidae (dragonfly, damselfly)	
Orthoptera	**Gryllidae** (cricket)	*Pham et al, 2013*
		Rosario et al, 2018
	Acrididae (grasshopper)	
Blattodea	**Blattidae** (cockroach)	*Padilla-Rodriguez et al, 2013*
Isoptera	**Termitidae** (termite)	*Kerr et al, 2018*
Polydesmida	**Paradoxosom-atidae** (millipede)	*Rosario et al, 2018*
Parasitiformes	**Varroidae** (mite)	*Kraberger et al, 2018*
Araneae	**Dysderidae**	*Rosario et al, 2018*
	Segestriidae	
	Araneidae	
	Theridiidae	
	Agelenidae	
	Cybaeidae	
	Pimoidae	
	Tetragnathidae	

Table 1.1 | *A brief selection of vertebrate and invertebrate taxa in which novel CRESS-DNA viruses have been identified via high throughput sequencing*. The discovery of new genotypes massively expands the possible host range of these ssDNA viruses and contributes to sequence repositories, aiding in recognition of more genotypes. Underlined = class (vertebrate) or phylum (invertebrate); first column = family; bold = order; parentheses = example common names.

The majority of information about the epidemiology, ecology, structural attributes, and pathogenicity of metazoan CRESS-DNA viruses is derived from vertebrate-associated genotypes. These well-characterized viruses include common or destructive pathogens of livestock or companion animals, which currently serve as the only available archetypes for molecular (and particularly, culture-based) characterization of environmental/invertebrate-associated CRESS-DNA viruses. For example, porcine circovirus 2 (PCV2), the etiological agent responsible for post-weaning multisystemic wasting syndrome (PMWS) in livestock herds, accrues net losses between 3 and 110USD per pig within affected herds, earning a ranking among the top three economically destructive pathogens within the swine industry (Gillespie et al, 2009; Alarcon et al, 2013). The economic relevance of this pathogen and structural similarities to a nonpathogenic serotype, PCV1, has shaped its significance as a model for vertebrate ssDNA viruses, and the only available reference for environmental CRESS-DNA viruses. Study of PCV2 has lent insight into viral capsid structure (Khayat et al, 2011), intracellular and infection dynamics (Cao et al, 2015), pathogenicity (Fenaux et al, 2003), and epidemiology (Rose et al, 2012) in vertebrate CRESS-DNA viruses. Secondly, Beak and Feather Disease Virus (BFDV, *Circoviridae*) is responsible for persistent immunosuppression in Psittacine avian hosts (Old and New World parrots), causing feather loss and tissue necrosis (Ritchie et al, 1989), offering potential phylogenomic context for the evolutionary relationship between CRESS-DNA viral families (Eastwood et al, 2014; Niagro et al, 1998). Finally, Canine Circovirus (CaCV, *Circoviridae*) represents an emergent enteric pathogen associated with vasculitis and hemorrhage, which may (or may not) play a significant role in co-infection dynamics in domestic dogs (Li et al, 2013).

Unclassified putative CRESS-DNA viruses are also common among non-model undomesticated vertebrates, including birds, bats, small carnivores, bony fish, and amphibians (Delwart & Li, 2012; Todd et al, 2011; see Table 1.1). However, their affiliation with these metazoans does not provide proof of infection. CRESS-DNA viruses have been detected in all major lineages of terrestrial arthropods (Rosario et al, 2018) and a range of aquatic invertebrates (Table

1.1), though the nature of the affiliation with these potential hosts has been entirely uninvestigated for the majority of genotypes. Therefore, despite their ubiquity, the impacts of CRESS-DNA viruses on the biology or ecology of non-model hosts and their ultimate contribution to metazoan non-consumptive (i.e. non-predatory) mortality remain largely unknown. These discoveries aid in resolving the possible metazoan hosts enabling the propagation and evolution of ssDNA viral populations, exploring host-virus interactions and impact on metazoan physiology, and advance our knowledge of both viral diversity and impact on ecosystem structure.

1.4 Extraordinary ecoevolutionary characteristics of CRESS-DNA viruses **|** CRESS-DNA viruses are elegant in their simplicity. Their genomes are among the smallest known to infect metazoans, ranging from 1-6kb (King et al, 2011). At the most rudimentary level, these viruses must encode two open reading frames ("ORFs," or colloquially among viruses, potential genes), responsible for replication (*rep*) and capsid formation (*cap*). These two ORFs represent the minimum requirement for viral existence: without the ability to propagate genomic information via nucleic acid replication (*rep*) and to produce a protein shell to package this information to commute between prospective hosts (*cap*), these genomes would fail the quintessential definition of a virus. While many CRESS-DNA viruses also encode auxiliary genes related to receptor recognition, immune system evasion, manipulation of host metabolism, pathogenesis, etc (though usually fewer than ten), many retain their inherent minimalism (Rosario et al, 2015). This simplicity may be futher facilitated by a uniquely minimally pathogenic relationship with the host (i.e. where propagation of new virions relies primarily on differential resource allocation, rather than immediate cell lysis), where fewer pathogenicity-associated ORFs may be indicative of a more commensal interaction. Evolutionary jettisoning of genomic "baggage" and maintenance of small genome sizes may accelerate replication rates and permit smaller (perhaps more efficient) capsid volume. This irreducibility renders CRESS-DNA viruses attractive to the study of ssDNA viral evolution, including virome community cohesion and genotype distribution.

Variant-driven evolution – Despite their diminutive genomes and number of essential genes

(or perhaps as a corollary to these factors), CRESS-DNA viruses are a trove of miscellaneous genera, likely generated through reiterative, rapid acquisition of mutations and recombination events. CRESS-DNA viruses evolve at rates comparable to RNA viruses, predominantly due to the lability of ssDNA relative to dsDNA (CRESS-DNA viruses exhibit 10^{-3} to 10^{-4} substitutions site^{-1} year^{-1} whereas most RNA viruses exhibit 10^{-2} to 10^{-5} substitutions site^{-1} year^{-1}; Duffy et al, 2008; Duffy & Holmes, 2007, 2009; Jenkins et al, 2002; Sanjuán et al, 2010). In addition to streamlining replication, small genome size may also decrease mutational load, avoiding disadvantages associated with mutation-driven fitness costs and potentially resulting in greater genomic "resilience" in a variety of environments. Characterizing the spectra of variants across ssDNA viral genomes and the ecological consequence of genotype variant densities within hosts or ecosystems is a first step in determining the role of error-prone rapid mutation and competition between variants in ssDNA virus evolution (Duffy et al, 2008; Martin et al, 2011). The majority of extant CRESS-DNA viruses putatively exist in microdiverse populations, enabling more permutations for success (i.e. viral transmission, infection, propagation) in a variety of environmental and host-specific conditions. Differential fitness of genotypes carrying variants may influence viral epidemiology by facilitating cross-species transmission or accelerating emergence. Therefore, this trait is likely responsible for the vast and growing emergence of new genera, including potential agricultural pathogens (Duffy & Holmes, 2007, 2009).

Variability in genome architecture – As a result of rapid evolution and recombination, novel CRESS-DNA viral genomes with unannotated ORFs sharing little or no similarity (or average nucleotide identity) to known genes, or that do not retain the ORF orientation of known CRESS-DNA viruses have been computationally recovered from viromes, complicating taxonomic categorization. Viruses rely entirely on hosts to provide basic mechanisms for genomic replication, yet high-throughput sequencing of well-characterized circoviruses imply that additional factors beyond host DNA polymerase fidelity, such as genomic architecture, genome size/composition, replication mechanism, host-driven epigenetic alterations, or selection to avoid sequence-specific

(e.g. RNAi/miRNA) defense mechanisms also likely influence viral evolution (Franzo et al, 2018). Therefore, these viruses are instead characterized by genomic architecture and presence of conserved motifs, as these motifs may retain indistinguishable function, despite codon degeneracy and saturated rates of synonymous variation (Aiewsakun & Katzourakis 2016). CRESS-DNA viruses may comprise monopartite or multipartite genomes, with the orientation of the *rep* ORF relative to structural ORFs and the putative origin of replication (i.e. ambisense or unisense, with the nonanucleotide motif denoting the positive sense strand) segregating novel genomes into eight tractable groups. This nomenclature provides a functional reference to first identify and consolidate unclassified - often computationally identified and environmental – viral genotypes (Rosario et al, 2015).

Replication initiation – The presence of the *rep* ORF is a defining factor for CRESS-DNA viruses, retaining both functional and structural homology. As such, *rep* serves as a hallmark of this diverse group and often assists in bioinformatic identification of new genotypes. Rep is essential for the initiation of rolling circle replication, nicking and freeing a 3' hydroxyl group within a canonical nonanucleotide (9-mer) motif in a covalently closed stem loop structure, priming rolling circle replication (motif: NANTATT*AC, where *indicates *nic* site upstream of recognition sequence and N indicates ambiguous nucleotides; Rosario et al, 2015; Figure 1.1). *rep* ORFs often contain multiple internal start codons immediately downstream from the initial start codon, allowing for leaky scanning or readthrough depending on codon position, and include two functional regions: one containing an HUH domain (where HUH refers to histidine-hydrophobe-histidine followed by two additional tyrosines) involved in replication initiation and termination via ssDNA breaking and joining, and one superfamily 3 helicase domain (SF3) involved in replication elongation (Chandler et al, 2013; Rosario et al, 2015; Figure 1.1). SF3, including Walker A, B, and Motif C terminates in a catalytic arginine finger (AAA+ ATPase), potentially utilized in elongation activity, but not universally present (Gorbalenya et al, 1990).

Interestingly, rolling circle motifs I-III within the 5' HUH domain is homologous to HUH

domains present in all three domains of cellular life, potentially lending insight into its evolutionary origins (Chandler et al, 2013). Within these organisms, the HUH domain performs site-specific cleavage and ligation in ssDNA templates, with functions ranging from intron homing, repetitive extragenic palindromic sequence processing, conjugative plasmid transfer, DNA transposons targeting, etc (Campos-Olivas, 2002; Chandler et al, 2013). Despite the variety of functions HUH-containing Rep-like enzymes execute, the close association between HUH domains and SF3 appears unique to ssDNA viral genomes and the general conservation of this configuration among these viruses demonstrate their utility as identifiable features of CRESS-DNA viruses (Chandler et al, 2013; Laufs et al, 1995). The endonuclease activity of the HUH domain in *rep* has also been proposed as a key element explaining the proclivity of ssDNA viruses for recombination. The modularity of *rep*, coupled with widespread evidence of co-occurring but recombined genomes, suggest that the adaptive value of recombination among circular ssDNA viral genomes outweighs the high frequency of dead-end deleterious outcomes (Lefeuvre et al, 2009; Martin et al, 2011), also facilitating host switching to access more favorable infectious niches. While it remains unconfirmed experimentally, it has also been proposed that the origin of CRESS-DNA viruses involves capsid acquisition by a HUH-containing mobile genetic element (Krupovic et al, 2009).

Capsid – Several studies have proposed a narrative in which *rep*-encoding ssDNA (e.g. plasmids, transposons, etc) have acquired an icosahedral capsid from ssRNA viruses (Gibbs & Weiller, 1999; Kazlauskas et al, 2017; Krupovic et al, 2009; Krupovic & Bamford, 2008), though the observed similarities between selfish genetic elements and viruses may alternatively arise if these elements are remnants of ancestral viruses that have lost capsids. Recombination between viral *cap* ORFs of different families also appears feasible, as circovirus, bacilladnavirus, and nodavirus capsids may also originate from ssRNA viruses, blurring the lines between previously defined virus groups (Rybiki 1994). Further convoluting this ancestry, some CRESS-DNA viruses appear to contain capsids highly similar to extant tombusviruses (+ssRNA, Baltimore group IV). Members of this group have been designated the "*Cruciviridae,*" and have been identified in ecosystems ranging

from geothermally active lakes to peatlands (Diemer & Stedman, 2012, Steel et al, 2016; Dayaram et al 2016, Quaiser et al, 2016; Bistolas et al, 2017a, Hewson et al, 2013a, Kuprovic et al, 2015).

CRESS-DNA capsid proteins largely lack sequence conservation (on average <60% sequence similarity; Rosario et al, 2015), instead epitomizing one of the most divergent and often pleiotropic, components of CRESS-DNA viral genomes. Therefore, annotation is primarily reliant on the presence of an arginine (or basic amino acid)-rich N-terminus predicted to interact with packaged DNA (Crowther et al, 2003; Heath et al, 2006) and the identification of intrinsically disordered regions (those enabling conformational flexibility, conferring greater multi-functionality to residues; Rosario et al, 2015b; Niagro et al, 1998). Mammalian and avian circovirus *cap* ORFs have been predicted to evolve under both positive (Brandao et al, 2010; Kundua et al, 2012; Liao et al, 2015) and purifying selection (Firth et al, 2009; Hughes & Piontkivska, 2008), often with conflicting conclusions drawn for the same viral genome. Furthermore, while single mutations in the capsid protein of mammalian circoviruses leads to distinct antigentic properties (Allemandou et al, 2011), little is known about the role of these changes in the emergence of viral phenotypes.

Despite this variability, the capsids of well-characterized (vertebrate or plant) CRESS-DNA viruses are predominantly icosahedral (T=1), ~28-32 kDa (Payne 2016), 15nm-32nm single or twin icosahedra (King et al, 2011), and highly resilient to environmental assaults. For example, physicochemical properties of the model metazoan pathogen, PCV2, enable viral infectivity despite treatment with acid (pH 3 per Mcnulty et al, 1984), heat (56°C and 70°C for 15min), alcohol, or chloroform (per Feldman & Wang 1961; Allan et al, 1994). This structural stability may underlie the endurance of these viruses in natural reservoirs and residence time in vectors, potentially explaining their widespread nature.

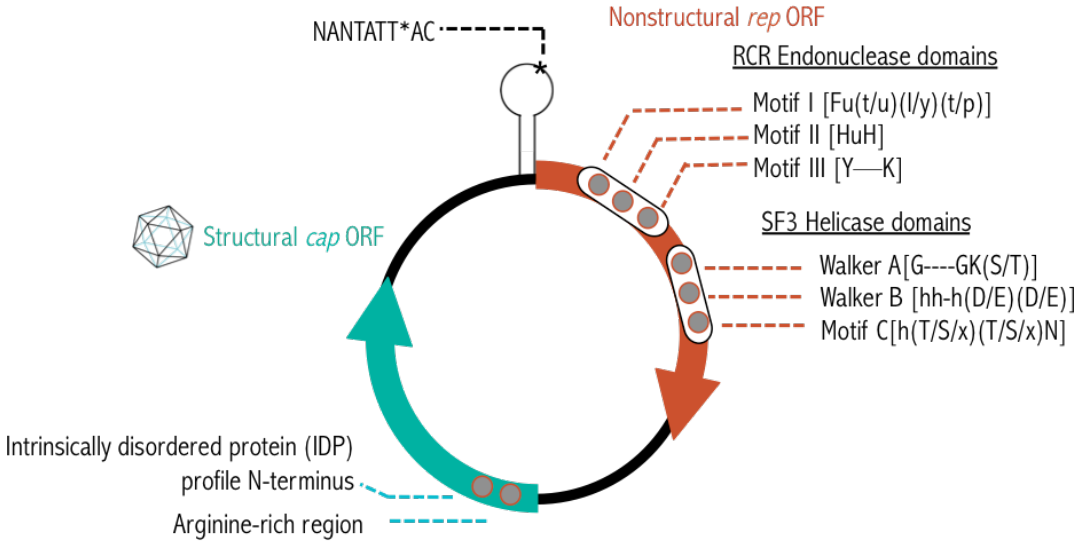

Figure 1.1 | *General schematic of archetype CRESS-DNA virus genome*, including two characteristic open reading frames (*rep, cap*) and domains associated with rolling circle replication and capsid.

1.5 CRESS-DNA virus-host interactions | *Model genotypes* - Most of what we know about the interface between CRESS-DNA viruses and metazoan hosts is derived from isolated, cultured, vertebrate-associated genotypes. Resolved crystal structures reveal that the capsid of porcine circovirus 2 (PCV2) is comprised of canonical β-barrel "jelly roll" folds with accompanying large side chains (Cao et al, 2015; Khayat et al, 2011), a common tertiary structure among icosahedral capsid subunits, including both ssDNA and ssRNA viruses (Khayat & Johnson 2011). Heparan sulfate and chondroitin sulfate B glycosaminoglycans (GAG) have both been proposed as targeted receptors for PCV2, with the capsid sulfate ions predicted to facilitate clatherin-mediated endocytosis and entry into monocytes (Cao et al, 2015). Once PCV2 is free of the endosome, this capsid protein recruits host retrograde motor protein, dynein, for intracellular transport via microtubules towards the nucleus, with capsids as direct ligands of this transporter. The mechanism of dynein recruitment is still not well understood, though it is speculated random collisions between the capsid N-terminus and motor proteins may facilitate the interaction. Among invertebrate and environmental CRESS-DNA viruses, *cap* ORFs are rarely conserved across families and ecosystems. Therefore, although the function of PCV-*cap* has been essentially resolved, this may contribute only

broad generalizations to our knowledge of CRESS-DNA virus infection dynamics. Because these structural proteins play a primary and potentially pleiotropic role in host-virus interactions, it is difficult to predict the relationship between *cap* sequence composition and implications for host-virus interactions without a clearer picture of phylogenetic context and intracellular interactions.

Invertebrate genotypes - CRESS-DNA viruses have been identified within viromes of invertebrates from nearly all biomes presently explored, ranging from arachnids to gastropods (Table 1.1). For decades, it has been well established that insects are capable of vectoring geminiviruses, and circular ssDNA viral genomes recurrently circulate within insect populations (Dietzgen et al, 2016; Mansoor et al, 2003). However, the application of high throughput sequencing triggered a wave of new viral discoveries of unclassified, bioinformatically identified invertebrate-associated genotypes – those affiliated with invertebrates, but without confirmation that they infect invertebrates. Therefore, "CRESS-DNA virus" serves as a catchall designation for incredibly diverse genotypes to avoid contaminating nascent taxonomic classification schemes. While novel CRESS-DNA viruses have been described consistently throughout history, the first putative invertebrate genotypes formally designated "CRESS-DNA viruses" were identified in 2012 among the Odonata (dragonflies; Rosario et al, 2012). It is now common to identify CRESS-DNA viruses among insect viromes, and phytophagous and hematophagous insects have been hypothesized as key vectors for potential pathogens of agricultural crops (e.g. *geminiviruses*, *nanoviruses*; Dietzgen et al, 2016) or vertebrates (e.g. *circoviruses*; Kraberger et al, 2019; Wang et al, 2018). To our knowledge, the pathogenicity of no CRESS-DNA virus has been confirmed in any invertebrate host. While a cnidarian-associated CRESS-DNA virus was proposed as a candidate pathogen associated with multifactorial bleaching and white-plague-like tissue loss in *Montastraea annularis* (Soffer et al, 2014), and another was implicated in lesion development and tissue loss in an echinoderm (Fahsbender et al, 2015), it is clear that these are merely correlative observations, and tremendous caution must be exercised as no classical histopathological or diagnostic molecular methods have been sufficiently applied.

Crustacean genotypes - In 2014, CRESS-DNA virus-like sequences were identified as integrated components of crustacean genomes (Metegnier et al, 2015; Thézé et al, 2014). Concurrently, CRESS-DNA viral consortia were observed in association with marine calanoid copepods (*Acartia tonsa, Labidocera aestiva*; Dunlap et al, 2013), freshwater brachiopods (*Daphnia mendotae* and *Daphnia retrocurva*; Hewson et al, 2013a), freshwater amphipods (*Diporeia spp.* Hewson et al, 2013b), and marine decapods (*Penaeus monodon* and *Farfantepenaeus duorarum*; Pham et al, 2014 and Ng et al, 2013, respectively), with prevalence and copy number often corresponding with population density or disease states. While tempting to associate the presence of these ssDNA viruses with crustacean fitness, these studies were purely correlative with spatiotemporal fluxes in crustacean population density and require additional investigation (Table 1.2). Despite the correlative nature of these studies, they collectively triggered an interest in the potential ecological impact of these ssDNA viruses on crustacean ecology.

Genotype	**Putative host**	**Correlation with population flux**	**Examples of Confounding factors**
LM29173 (Hewson et al, 2013b)	*Diporeia spp.*	Detection correlates with dramatic population decline in the mid-1990s,	Viral distribution may more closely chart haplotype distribution
LaCopCV, AtCopCV (Dunlap et al, 2013)	*Acartia tonsa, Labidocera aestiva*	Viral load observed during periods of population turnover (spring/fall)	Detection at these periods may be a result of modification in copepod ontogeny, seasonality, growth rate, etc
DMClaHV (Hewson et al, 2013a)	*Daphnia*	Viral prevalence greatest immediately prior to population decline	Potentially attributable to accelerated transmission of a relatively benign virus within high density populations, rather than indicative of the causative factor in a population decline driven by resource limitation, or associated with an epibiont
PmCV-1 (Pham et al, 2014)	*Penaeus monodon*	Viral genotype detected among "diseased" shrimp in aquaculture	Genotype may also be present in asymptomatic shrimp; were not examined

Table 1.2 | *Description of CRESS-DNA virus discoveries among crustaceans* – genotype detection often correlates with changes in population dynamics of crustaceans, but these fluxes in population density may also be attributable to simultaneous confounding factors.

1.6 Significance of microcrustacean hosts | It is generally accepted that the Pancrustacea are present in every biome, dominating in both biomass and total species richness. In particular, aquatic ecosystems are structured by complex, continuously evolving interactions between these metazoans, microbes, and their abiotic environment (Lima-Mendez et al, 2015), modulated by trophic-level exchanges ranging from viral parasitism to conspecific competition to mutualistic symbiosis. There are over 30,000 extant species of crustaceans, with microcrustaceans (e.g. mesozooplankton, meiobenthic fauna, etc) often representing the prevailing mesograzers within aquatic ecosystems, responsible for both top-down and bottom-up control of primary productivity (Dole-Olivier et al, 2000). For example, crustacean grazing may reduce epiphyte load, whereas sloppy feeding and excretion products may mediate availability of N, P, and other essential nutrients and elevate photosynthetic capacity in otherwise oligotrophic habitats. Conversely, zooplankton are essential links between DOM/POM and ichthyofauna, exerting grazing pressure on autotrophs and governing the spatial organization and trophic architecture within these communities (Clark et al, 2001; González et al, 2019; Neill, 1975). Indeed, by some estimates, a single species of copepod (*Calanus finmarchicus*) may export more than 3 million tons of carbon in the North Atlantic annually, indicating the importance of these organisms to atmospheric CO_2 drawdown, macrofauna migration, microbial nutrient uptake in the photic zone, etc (Jónasdóttir et al, 2015). Therefore, through both their consumption of biomass, and the act of being consumed, microcrustaceans often define energy and carbon flow through aquatic food webs.

As with all species, disease is hypothetically pervasive among these mesograzers, and likely detracts from net feeding behavior, skewing the quantity of biomass consumed by higher trophic levels relative to that sequestered (e.g. as POM/marine snow; Figure 1.2). Microcrustacean mortality is poorly constrained, with nonconsumptive mortality attributable to parasitism and disease often missing from productivity calculations (Hirst & Kiørboe, 2002). Due to their cosmopolitan nature, rapid evolution, and unusual genomic attributes, CRESS-DNA viral discoveries, integrated with investigation of their distribution, evolution, host specificity, and host impact are therefore

indispensible to understanding ecologically relevant host-virus interactions and nutrient flux in aquatic ecosystems.

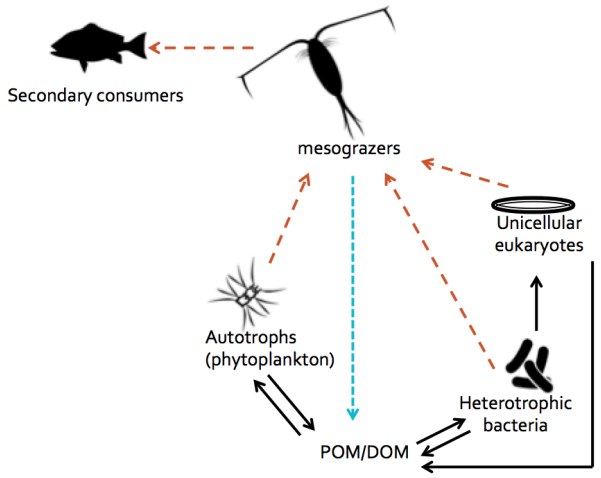

Figure 1.2 | *Simplified schematic of carbon flow to crustacean mesograzers and secondary consumers in aquatic ecosystems highlighting the role of metazoan viral infection on the bioavaliability of resources for microbial and metazoan* - Red (dashed) lines indicate reduced carbon flow to mesograzers as a product of viral infection (nonconsumptive mortality, reduced herbivory, etc), resulting in reduced biomass available to upper trophic level organisms. Blue line indicates greater recycling of carbon from mesograzers to POM/DOM pool through viral lysis and production of cell debris, likely resulting in utilization by microbial taxa.

1.7 Dissertation aims | The rich viral consortia associated with microcrustacean holobionts likely play a significant role in the physiology of the metazoan, and as a result, their function within the broader ecosystem. Comparative analyses of ambient viruses associated with these species illustrate distinct patterns of viruses on organismal, population, species and community levels. By exploring the diversity and dynamics of virus-microcrustacean associations, we may begin to capture the wider picture of CRESS-DNA virus phylogenomics and biogeography. Assuming the evolutionary potential of ssDNA viruses and widespread abundance of aquatic microcrustaceans, we may ostensibly extrapolate an immense global diversity and incidence of CRESS-DNA viral taxa. While momentous efforts are underway to evaluate host specificity, habitat distribution, and genomic diversity of ssDNA bacteriophage on a global scale (Gregory et al, 2019; Hurwitz & Sullivan, 2013), very few have yet to describe distributions of metazoan-associated viruses. Hence, this dissertation leverages viromes derived from a range of aquatic microcrustaceans from temperate to subtropical

marine and freshwaters to identify novel CRESS-DNA viral genotypes and explore the patterns that drive viral distribution and diversity across local and global scales. This dissertation represents a compilation of manuscripts, published or pending, intended to address three cardinal questions:

(1) What parameters modulate viral diversity and genomic composition?

(2) How are CRESS-DNA viruses biogeographically distributed?

(3) What discernable impact, if any, do CRESS-DNA viruses have on microcrustacean ecology?

To address these questions, 31 haplotype-specific viromes and 12 metatranscriptomes from 23 unique species of microcrustacean (amphipods, copepods, brachiopods, and isopods) were collected via plankton tows, sediment grabs, and manual sampling of intertidal nereocystis forests, detrital sponges, rocky intertidal sediments, or kelp wracks from the Sargasso and Salish seas, Laurentian Great Lakes, Eastern Australia, among other sites. These viromes provided the foundation for viral discovery on a fragmented, but global scale, allowing investigation of CRESS-DNA virus phylogenomics, biogeography, and population microdiversity. Viromes from this variety of species addressed the potential endemism and distribution of CRESS-DNA viruses, further explored within haplotype-specific viromes affiliated with the amphipod, *Diporeia spp*. from the Great Lakes, indicating the potential for microcrustaceans to serve as reservoirs for durable virions. Finally, gene expression and tissue stoichiometry in *Diporeia spp*. was examined to delineate the impact of CRESS-DNA virus infection on organismal scales. Collectively, this effort - though limited by taxonomic scope and scale - aims to investigate the ecoevolutionary characteristics of microcrustacean-associated CRESS-DNA viruses in an effort to advance our understanding of their role in aquatic ecosystem structure and function.

1.8 References

Aiewsakun P, Katzourakis A. 2016. Time-dependent rate phenomenon in viruses. *J Virol*. 90(16):7184-95. doi:10.1128/JVI.00593-16.

Alarcon P, Rushton J, Wieland B. 2013. Cost of post-weaning multi-systemic wasting syndrome and porcine circovirus type-2 subclinical infection in England - an economic disease model. *Prev Vet Med*. 110(2):88-102. doi:10.1016/j.prevetmed.2013.02.010.

Allan GM, Ellis JA. 2000. Porcine circoviruses: a review. *J Vet Diagn Invest*. 12(1):3-14. doi:10.1177/104063870001200102.

Allan GM, Phenix KV, Todd D, McNulty MS. 1994. Some biological and physico-chemical properties of porcine circovirus. *Zentralbl Vet Med B*. 41(1):17-26. doi:10.1111/j.1439-0450.1994.tb00201.x.

Allemandou A, Grasland G, Hernandez-Nignol AC, Kéranflec'h A, Cariolet R, Jestin A. 2011. Modification of PCV-2 virulence by substitution of the genogroup motif of the capsid protein. *Veterinary Research*. 42:54 doi:10.1186/1297-9716-42-54.

Altan E, Kubiski SV, Burchell J, Bicknese E, Deng X, Delwart E. 2019. The first reptilian circovirus identified infects gut and liver tissues of black-headed pythons. *Vet Res*. 16;50(1):35. doi:10.1186/s13567-019-0653-z.

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F. 2006. The marine viromes of four oceanic regions. *PLoS Biol*. 4(11):e368. doi:10.1371/journal.pbio.0040368.

Bistolas KSI, Besemer RM, Rudstam LG, Hewson I. 2017. Distribution and inferred evolutionary characteristics of a chimeric ssDNA virus associated with intertidal marine isopods. *Viruses*. 9(12): 361. doi:10.3390/v9120361.

Bistolas KSI, Jackson EW, Watkins Jm, Rudstam LG, Hewson I. 2017a. Distribution of circular single-stranded DNA viruses associated with benthic amphipods of genus *Diporeia* in the Laurentian Great Lakes. *Fresh Biol*. 62(7):1220-1231. doi:10.1111/fwb.12938.

Bergh O, Børsheim KY, Bratbak G, Heldal M. 1989. High abundance of viruses found in aquatic environments. *Nature*. 10;340(6233):467-8. doi:10.1038/340467a0.

Blinkova O, Victoria J, Li Y, Keele BF, Sanz C, Ndjango JB, Peeters M, Travis D, Lonsdorf EV, Wilson ML, Pusey AE, Hahn BH, Delwart EL. 2010. Novel circular DNA viruses in stool samples of wild-living chimpanzees. *J Gen Virol*. 91(1):74-86. doi:10.1099/vir.0.015446-0.

Brandao PE, de Souza SP, de Castro AMMG, Richtzenhain LJ. 2010. The Cap gene of porcine circovirus type 2 (PCV2) evolves by positive selection *in vitro. Braz. J. Vet. Res. Anim. Sci., São*

Paulo. 47(3): 209-212.

Breitbart M, Benner BE, Jernigan PE, Rosario K, Birsa LM, Harbeitner RC, Fulford S, Graham C, Walters A, Goldsmith DB,.Berger SA, Nejstgaard JC. 2015 Discovery, prevalence, and persistence of novel circular single-stranded DNA viruses in the ctenophores *Mnemiopsis leidyi* and *Beroe ovata*. *Front Microbiol*. 6:1427. doi:10.3389/fmicb.2015.01427.

Breitbart M. 2012. Marine viruses: Truth or dare. *Annu Rev Mar Sci.* 4:425–448. doi:10.1146/annurev-marine-120709-142805.

Brum JR, Ignacio-Espinoza JC, Kim EH, Trubl G, Jones RM, Roux S, VerBerkmoes NC, Rich VI, Sullivan MB. 2016. Illuminating structural proteins in viral "dark matter" with metaproteomics. *Proc Natl Acad Sci USA*. 113(9):2436-41. doi: 10.1073/pnas.1525139113.

Brum JR, Schenck RO, Sullivan MB. 2013. Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *ISME J*. 7(9):1738-51. doi:10.1038/ismej.2013.67.

Brum JR, Sullivan MB. 2015. Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat Rev Microbiol*. 13:147–159. doi:10.1038/nrmicro3404.

Butina TV, Bukin Y, Khanaev IV, Kravtsova L, Maikova OO, Tupikin A, Kabilov M, Belikov SI. 2019. Metagenomic analysis of viral communities in diseased Baikal sponge *Lubomirskia baikalensis. Limnol Oceanogr.* 1: 155-162. doi:10.31951/2658-3518-2019-A-1-155.

Campos-Olivas R, Louis JM, Clerot D, Gronenborn B, Gronenborn AM. 2002. The structure of a replication initiator unites diverse aspects of nucleic acid metabolism. *Proc Natl Acad Sci.* 99:10310–10315. doi: 10.1073/pnas.152342699.

Cao J, Lin C, Wang H, Wang L, Zhou N, Jin Y, Liao M, Zhou J. 2015. Circovirus transport proceeds via direct interaction of the cytoplasmic dynein IC1 subunit with the viral capsid protein. *J Virol*. 89(5):2777-91. doi:10.1128/JVI.03117-14.

Cha SY, Song ET, Kang M, Wei B, Seo HS, Roh JH, Yoon RH, Moon OK, Jang HK. 2014. Prevalence of duck circovirus infection of subclinical pekin ducks in South Korea. *J Vet Med Sci.* 76(4):597-9. doi:10.1292/jvms.13-0447

Chandler M, de la Cruz F, Dyda F, Hickman AB, Moncalian G, Ton-Hoang B. 2013. Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. *Nat Rev Microbiol*. 11(8):525-38. doi: 10.1038/nrmicro3067.

Chen F, Suttle CA, Short SM. 1996. Genetic diversity in marine algal virus communities as revealed by sequence analysis of DNA polymerase genes. *Appl Environ Microbiol*. 62(8): 2869–2874.

Clark DR, Aazem KV, Hays GC. 2001. Zooplankton abundance and community structure over a 4000 km transect in the northeast Atlantic. *J Plankton Res*. 23(4): 365-372. doi:10.1093/plankt/23.4.365 .

Crowther RA, Berriman JA, Curran WL, Allan GM, Todd D. 2003. Comparison of the structures of three circoviruses: chicken anemia virus, porcine circovirus type 2, and beak and feather disease virus. *J Virol*. 77(24):13036-41. doi:10.1128/jvi.77.24.13036-13041.2003.

Dayaram A, Potter KA, Moline AB, Rosenstein DD, Marinov M, Thomas JE, Breitbart M, Rosario K, Argüello-Astorga GR, Varsani A. 2013a. High global diversity of cycloviruses amongst dragonflies. *J Gen Virol*. 94(8):1827-40. doi:10.1099/vir.0.052654-0.

Dayaram A, Goldstien S, Zawar-Reza P, Gomez C, Harding JS, Varsani A. 2013a. Identification of starling circovirus in an estuarine mollusc (*Amphibola crenata*) in New Zealand using metagenomic approaches. *Genome Announc*.1(3): e00278-13. doi:10.1128/genomeA.00278-13.

Dayaram A, Goldstien S, Zawar-Reza P, Gomez C, Harding JS, Varsani A. 2013b. Novel ssDNA virus recovered from estuarine Mollusc (*Amphibola crenata*) whose replication associated protein (Rep) shares similarities with Rep-like sequences of bacterial origin. *J Gen Virol*.94(5):1104-10. doi:10.1099/vir.0.050088-0.

Dayaram A, Goldstien S, Argüello-Astorga GR, Zawar-Reza P, Gomez C, Harding JS, Varsani A. 2015. Diverse small circular DNA viruses circulating amongst estuarine molluscs. *Infect Genet Evol*. 31:284-95. doi: 10.1016/j.meegid.2015.02.010.

Dayaram A, Galatowitsch ML, Argüello-Astorga GR, van Bysterveldt K, Kraberger S, Stainton D, Harding JS, Roumagnac P, Martin DP, Lefeuvre P, Varsani A. 2016. Diverse circular replication-associated protein encoding viruses circulating in invertebrates within a lake ecosystem. *Infect Genet Evol*. 39:304-16. doi: 10.1016/j.meegid.2016.02.011 21.

De Lorgeril J, Lucasson A, Petton B, Toulza E, Montagnani C, Clerissi C, Vidal-Dupiol J, Chaparro C, Galinier R, Escoubas JM, Haffner P, Dégremont L, Charrière GM1, Lafont M1, Delort A, Vergnes A, Chiarello M, Faury N, Rubio T, Leroy MA, Pérignon A, Régler D, Morga B, Alunno-Bruscia M, Boudry P, Le Roux F, Destoumieux-Garzón D, Gueguen Y, Mitta G. 2018. Immune-suppression by OsHV-1 viral infection causes fatal bacteraemia in Pacific oysters. *Nat Commun.* 11;9(1):4215. doi:10.1038/s41467-018-06659-3.

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 Res*.164(1-2):114-21.

doi:10.1016/j.virusres.2011.11.021.

Diemer GS, Stedman KM. 2012. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. *Biol. Direct*, 7:13. doi:10.1186/1745-6150-7-13.

Dietzgen RG, Mann KS, Johnson KN. 2016. Plant virus–insect vector interactions: current and potential future research directions. *Viruses*. 8(11): 303. doi:10.3390/v8110303.

Díez-Villaseñor C, Rodriguez-Valera F. 2019. CRISPR analysis suggests that small circular single-stranded DNA smacoviruses infect Archaea instead of humans. *Nat Commun*. 10(1):294. doi:10.1038/s41467-018-08167-w.

Dole-Olivier MJ, Galassi DMP, Marmonier P, Creuzé Des Châtelliers M. 2000. The biology and ecology of lotic microcrustaceans. *Freshwater Biology*. 44(1):63-91. doi:10.1046/j.1365-2427.2000.00590.x 47.

Doszpoly A, Tarján ZL, Glávits R, Müller T, Benkő M. 2014. Full genome sequence of a novel circo-like virus detected in an adult European eel *Anguilla anguilla* showing signs of cauliflower disease. *Dis Aquat Organ*. 109(2):107-15. doi:10.3354/dao02730.

Duffy S, Holmes EC. 2007. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus Tomato Yellow Leaf Curl Virus. *J Virol*. 82(2): 957-965. doi:10.1128/JVI.01929-07.

Duffy S, Holmes EC. 2009. Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. *J Gen Virol*. 90:1539-1547, doi: 10.1099/vir.0.009266-0 80.

Duffy S, Shackelton LA , Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet*. 9:267-276 27. doi:10.1038/nrg2323.

Dunlap DS, Ng TFF, Rosario K, Barbosa JG, Greco AM, Breitbart M, Hewson I. 2013. Molecular and microscopic evidence of viruses in marine copepods. *Proc Nat Acad Sci USA*. 110:1375-1380 . doi:10.1073/pnas.1216595110.

Eaglesham JB, Hewson I. 2013. Widespread detection of circular replication initiator protein (*rep*)-encoding ssDNA viral genomes in estuarine, coastal and open ocean net plankton. *Mar Ecol Prog Ser.* 494:65-72 15. doi:10.3354/meps10575.

Eastwood JR, Berg ML, Ribot RFH, Raidal SR, Buchanan KL, Walder KR, Bennett ATD. 2014. Phylogenetic analysis of beak and feather disease virus across a host ring-species complex. *Proc Nat Acad Sci*. 14153–14158, doi:10.1073/pnas.1403255111.

Edwards RA, Rohwer F. 2005. Viral metagenomics. *Nature Rev Microbiol*. 3, 504–510. doi:10.1038/nrmicro1163.

Fahsbender E, Hewson I, Rosario K, Tuttle AK,Varsani A, Breitbart M. 2015. Discovery of a novel circular DNA virus in the Forbes sea star, *Asterias forbesi*. *Arch Virol*. 160(9):2349–2351 22. doi:10.1007/s00705-015-2503-2.

Feldman HA, Wang SS. 1961. Sensitivity of various viruses to chloroform. *Proc Soc Exp Biol Med*. 106:736-8. doi:10.3181/00379727-106-26459.

Fenaux M, Opriessnig T, Halbur PG, Meng XJ. 2003. Immunogenicity and pathogenicity of chimeric infectious DNA clones of pathogenic porcine circovirus type 2 (PCV2) and nonpathogenic PCV1 in weanling pigs. *J Virol*. 77(20):11232-43. doi:10.1128/jvi.77.20.11232-11243.2003.

Firth C, Charleston MA, Duffy S, Shapiro B, Holmes EC. 2009. Insights into the evolutionary history of an emerging livestock pathogen: Porcine Circovirus 2. *J Virol*. 83(24):12813-12821. doi:10.1128/JVI.01719-09.

Flint J, Racaniello VR, Rall GR, Skalka AM. 2015. Principles of Virology, 4th ed. *American Society for Microbiology*. Washington DC, USA.

Franzo G, Segales J, Tucciarone CM, Cecchinato M, Drigo M. 2018. The analysis of genome composition and codon bias reveals distinctive patterns between avian and mammalian circoviruses which suggest a potential recombinant origin for Porcine circovirus 3. *PLoS One*. 13(6):e0199950. doi:10.1371/journal.pone.0199950.

Fuhrman JA. 1999. Marine viruses and their biogeochemical and ecological effects. *Nature*. 399:541–548. doi:10.1038/21119.

Garigliany MM, Börstler J, Jöst H, Badusche M, Desmecht D, Schmidt-Chanasit J, Cadar D. 2015. Characterization of a novel circo-like virus in *Aedes vexans* mosquitoes from Germany: evidence for a new genus within the family *Circoviridae*. *J Gen Virol*. 96(4):915-20. doi:10.1099/vir.0.000036.

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. *J Gen Virol*. 92(11):2646-2653. doi:10.1099/vir.0.034108-0.

Geddes AM. The history of smallpox. *Clin Dermatol*. 24(3):152-157. doi:10.1016/j.clindermatol.2005.11.009.

Gibbs MJ, Weiller GF. 1999. Evidence that a plant virus switched hosts to infect a vertebrate and then recombined with a vertebrate-infecting virus. *Proc Natl Acad Sci* 96:8022-8027. doi:10.1073/pnas.96.14.8022.

Gillespie J, Opriessnig T, Meng XJ, Pelzer K, Buechner-Maxwell V. 2009. Porcine circovirus type 2 and porcine circovirus-associated disease. *J Vet Intern*. 23(6):1151-63. doi:10.1111/j.1939-1676.2009.0389.x.

González CE, Rubén E, Antonio B, Wolfgang S. 2019. Zooplankton taxonomic and trophic community structure across biogeochemical regions in the Eastern South Pacific. *Front Mar Sci*. doi:10.3389/fmars.2018.00498.

Gorbalenya AE, Koonin EV, Wolf Y. 1990. A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses. *FEBS Lett*. 262(1):145-8. doi:10.1016/0014-5793(90)80175-I.

Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A, Ardyna M, Arkhipova K, Carmichael M, Cruaud C, Dimier C, Domínguez-Huerta G, Ferland J, Kandels S, Liu Y, Marec C1 Pesant S, Picheral M, Pisarev S, Poulain J, Tremblay JÉ, Vik D; Tara Oceans Coordinators, Babin M, Bowler C, Culley AI, de Vargas C, Dutilh BE, Iudicone D, Karp-Boss L, Roux S, Sunagawa S, Wincker P, Sullivan MB. 2019. Marine DNA viral macro- and microdiversity from pole to pole. *Cell*. 177(5):1109-1123.e14. doi:10.1016/j.cell.2019.03.040.

Halami MY, Nieper H, Müller H, Johne R. 2008. Detection of a novel circovirus in mute swans (*Cygnus olor*) by using nested broad-spectrum PCR. *Virus Res*. 132(1-2):208-12. doi:10.1016/j.virusres.2007.11.001.

Hanna ZR, Runckel C, Fuchs J, DeRisi JL, Mindell DP, Van Hemert C, Handel CM, Dumbacher JP. 2015. Isolation of a complete circular virus genome sequence from an Alaskan black-capped chickadee (*Poecile atricapillus*) gastrointestinal tract sample. *Genome Announc*. 3(5):e01081-15. doi:10.1128/genomeA.01081-15.

Hansen TA, Fridholm H, Frøslev TG, Kjartansdóttir KR, Willerslev E, Nielsen LP, Hansen AJ. 2015. New type of papillomavirus and novel circular single stranded DNA virus discovered in urban *Rattus norvegicus* using circular DNA enrichment and metagenomics. *PLoS One.* 10(11):e0141952. doi:10.1371/journal.pone.0141952.

Hattermann K, Schmitt C, Soike D, Mankertz A 2003. Cloning and sequencing of duck circovirus (DuCV). *Arch Virol*. 148(12):2471-80. doi:10.1007/s00705-003-0181-y.

Heath L, Williamson AL, Rybicki EP. 2006. The capsid protein of beak and feather disease virus binds to the viral DNA and is responsible for transporting the replication-associated protein into the nucleus. *J Virol*. 80(14):7219-25. doi:10.1128/JVI.02559-05.

Hewson I, Ng G, Li W, LaBarre BA, Aguirre I, Barbosa JG, Breitbart M, Greco AW, Kearns CM, Looi A, Schaffner LR, Thompson PD, Hairston NG. 2013a. Metagenomic identification, seasonal dynamics, and potential transmission mechanisms of a *Daphnia*−associated single−stranded DNA virus in two temperate lakes. *Limnol Oceanogr*. 58: 1605– 1620. doi:10.4319/lo.2013.58.5.1605 17.

Hewson I, Eaglesham JB, Höök TO, LaBarre BA, Sepúlveda MS, Thompson PD, Watkins JM, Rudstam LG. 2013b. Investigation of viruses in *Diporeia* spp. from the Laurentian Great Lakes and Owasco Lake as potential stressors of declining populations. *J Great Lakes Res*. 39:499–506. doi:10.1016/j.jglr.2013.06.006 16.

Hirst AG, Kiørboe T. 2002. Mortality of marine planktonic copepods: global rates and patterns. *MEPS*. 230:195-209. doi:10.3354/meps230195.

Hughes AL, Piontkivska H. 2008. Nucleotide sequence polymorphism in circoviruses. *Infect Genet Evol.* 8(2):130-8. doi: 10.1016/j.meegid.2007.11.001.

Hurwitz BL, Sullivan MB. 2013. The Pacific Ocean virome (POV): A marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS ONE*. 8(2):e57355. doi:10.1371/journal.pone.0057355.

Iwanowski, D. 1892. Über die Mosaikkrankheit der Tabakspflanze. Bulletin Scientifique publié par l'Académie Impériale des Sciences de Saint-Pétersbourg / Nouvelle Serie III. 35: 67–70. Translated into English in Johnson, J., Ed. (1942) Phytopathological classics (St. Paul, Minnesota: American Phytopathological Society) No. 7, pp. 27–-30.

Jenkins GM, Rambaut A, Pybus OG, Holmes EC. 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J Mol Evol*. 54(2):156-65. doi:10.1007/s00239-001-0064-3.

Johne R, Fernández-de-Luco D, Höfle U, Müller H. 2006. Genome of a novel circovirus of starlings, amplified by multiply primed rolling-circle amplification. *J Gen Virol*. 87(5):1189-95. doi:10.1099/vir.0.81561-0.

Jónasdóttir SH, Visser AW, Richardson K, Heath MR. 2015. Seasonal copepod lipid pump promotes carbon sequestration in the deep North Atlantic. *Proc Natl Acad Sci U S A*. 112(39):12122-6. doi:10.1073/pnas.1512110112.

Jover LF, Effler TC, Buchan A, Wilhelm SW, Weitz JS. 2014. The elemental composition of virus particles: implications for marine biogeochemical cycles. *Nat Rev Microbiol*. 12(7):519-28. doi:10.1038/nrmicro3289.

Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, Raes J, Sullivan M, Arendt D, Benzoni F, Claverie JM, Follows M, Gorsky G, Hingamp P, Iudicone D, Jaillon O, Kandels-Lewis S, Krzic U, Not F, Ogata H, Pesant S, Reynaud EG, Sardet C, Sieracki ME, Speich S, Velayoudon D, Weissenbach J, Wincker P; Tara Oceans Consortium. 2011. A holistic approach to marine eco-systems biology. *PLoS Biol*. 9(10):e1001177. doi:10.1371/journal.pbio.1001177.

Kazlauskas D, Dayaram A, Kraberger S, Goldstien S, Varsani A, Krupovic M. 2017. Evolutionary history of ssDNA bacilladnaviruses features horizontal acquisition of the capsid gene from ssRNA nodaviruses. *Virology*. 504:114-121. doi:10.1016/j.virol.2017.02.001

Kerr M, Rosario K,Baker CCM, Breitbart M. 2018. Discovery of four novel circular single-stranded DNA viruses in fungus-farming termites. *Genome Announc*. 6(17): e00318-18. doi:10.1128/genomeA.00318-18

Khayat R, Brunn N, Speir JA, Hardham JM, Ankenbauer RG, Schneemann A, Johnson JE. 2011. The 2.3-angstrom structure of porcine circovirus 2. *J Virol*. 85(15):7856-62. doi:10.1128/JVI.00737-11

Khayat R, Johnson JE. 2011. Pass the jelly rolls. *Structure*. 19(7):904-6. doi:10.1016/j.str.2011.06.004.

Kim HK, Park SJ, Nguyen VG, Song DS, Moon HJ, Kang BK, Park BK. 2012. Identification of a novel single-stranded, circular DNA virus from bovine stool. *J Gen Virol*. 93:635–639. doi:10.1099/vir.0.037838-0.

Kim KH, Chang HW, Nam YD, Roh SW, Kim MS, Sung Y, Jeon CO, Oh HM, Bae JW. 2008. Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Appl Environ Microbiol*. 74(19): 5975-5985 39. doi:10.1128/AEM.01275-08.

King AMQ, Adams MJ, Carstens EB, Lefkowitz, EJ. 2011. Virus Taxonomy: Ninth Report of the International Committee on the Taxonomy of Viruses. *Elsevier Inc*. ISBN 978-0-12-384684-6

Kraberger S, Argüello-Astorga GR, Greenfield LG, Galilee C, Law D, Martin DP, Varsani A. 2015. Characterization of a diverse range of circular replication-associated protein encoding DNA viruses recovered from a sewage treatment oxidation pond. *Infect Genet Evol*. 31:73-86. doi:10.1016/j.meegid.2015.01.001.

Kraberger S, Schmidlin K, Fontenele RS, Walters M, Varsani A. 2019a. Unravelling the single-stranded DNA virome of the New Zealand blackfly. *Viruses*. 11(6):E532. doi:10.3390/v11060532.

Kraberger S, Cook CN, Schmidlin K, Fontenele RS, Bautista J, Smith B, Varsani A. 2019b. Diverse single-stranded DNA viruses associated with honey bees (Apis mellifera). Infect Genet Evol. 71:179-188. doi:10.1016/j.meegid.2019.03.024.

Kraberger S, Visnovsky GA, van Toor RF, Male MF, Waits K, Fontenele RS, Varsani A. 2018. Genome sequences of two single-stranded DNA viruses identified in *Varroa destructor*. *Genome Announc*. 6(9): e00107-18. doi:10.1128/genomeA.00107-18.

Krupovic M, Bamford DH. 2008. Virus evolution: how far does the double β-barrel viral lineage extend? *Nat Rev Micro*. 6:941–948. doi:10.1038/nrmicro2033.

Krupovic M, Ravantti JJ, Bamford DH. 2009. Geminiviruses: a tale of a plasmid becoming a virus. *BMC Evol Biol*. 9:112. doi:10.1186/1471-2148-9-112.

Krupovic M, Zhi N, Li J, Hu G, Koonin EV, Wong S, Shevchenko S, Zhao K, Young NS. 2015. Multiple layers of chimerism in a single-stranded DNA virus discovered by deep sequencing. *Genome Biol Evol*. 7(4):993-1001. doi:10.1093/gbe/evv034.

Kundua S, Faulkes CG, Greenwood AG , Jones CG, Kaiser P, Lyne OD, Black SA, Chowrimootoo A, Groombridge A. 2012. Tracking viral evolution during a disease outbreak: the rapid and complete selective sweep of a circovirus in the endangered echo parakeet. *J Virol*. 86(9):5221-5229 95. doi:10.1128/JVI.06504-11.

Labonté JM, Suttle CA. 2013. Previously unknown and highly divergent ssDNA viruses populate the oceans. *ISME J*. 7: 2169–2177. doi:10.1038/ismej.2013.110.

Laufs J, Jupin I, David C, Schumacher S, Heyraud-Nitschke F, Gronenborn B. 1995. Geminivirus replication: genetic and biochemical characterization of Rep protein function, a review. *Biochimie*. 77(10):765-73. doi:10.1016/0300-9084(96)88194-6.

Lefeuvre P, Lett JM, Varsani A, Martin DP. 2009. Widely conserved recombination patterns among single-stranded DNA viruses. *J Virol*. 83(6):2697-707. doi:10.1128/JVI.02152-08.

Li L, Kapoor A, Slikas B, Bamidele OS, Wang C, Shaukat S, Masroor MA, Wilson ML, Ndjango JB, Peeters M, Gross-Camp ND, Muller MN, Hahn BH, Wolfe ND, Triki H, Bartkus J, Zaidi SZ, Delwart E. 2010. Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces. *J Virol*. 84(4):1674-82. doi:10.1128/JVI.02109-09.

Li L, McGraw S, Zhu K, Leutenegger CM, Marks SL, Kubiski S, Gaffney P, Dela Cruz FN Jr, Wang C, Delwart E, Pesavento PA. 2013. Circovirus in tissues of dogs with vasculitis and hemorrhage. *Emerg Infect Dis.* 19(4): 534–541. doi:10.3201/eid1904.121390.

Li L, Victoria JG, Wang C, Jones M, Fellers GM, Kunz TH, Delwart E. 2010. Bat guano virome: predominance of dietary viruses from insects and plants plus novel mammalian viruses. *J Virol*. 84(14):6955-65. doi:10.1128/JVI.00501-10.

Li L, Shan T, Soji OB, Alam MM, Kunz TH, Zaidi SZ, Delwart E. 2011. Possible cross-species transmission of circoviruses and cycloviruses among farm animals *J Gen Virol.* 92 (4), 768-772. doi:10.1128/JVI.02109-09.

Lian H, Liu Y, Li N, Wang Y, Zhang S, Hu R. 2014. Novel circovirus from mink, China. *Emerg Infect Dis.* 20(9):1548-50. doi:10.3201/eid2009.140015.

Liao PC, Wang KK, Tsai SS, Liu HJ, Huang BH, Chuang KP. 2015. Recurrent positive selection and heterogeneous codon usage bias events leading to coexistence of divergent pigeon circoviruses. *J Gen Virol*. 96(8):2262-73. doi:10.1099/vir.0.000163.

Lima FES, Cibulski SP, Dall Bello AG, Mayer FQ, Witt AA, Roehe PM, d'Azevedo PA. 2015. A novel chiropteran circovirus genome recovered from a brazilian insectivorous bat species. *Genome Announc*. 3(6):e01393-15. doi:10.1128/genomeA.01393-15.32.

Lima-Mendez G, Faust K, Henry N, Decelle J, Colin S, Carcillo F, Chaffron S, Ignacio-Espinosa JC, Roux S, Vincent F, Bittner L, Darzi Y, Wang J, Audic S, Berline L, Bontempi G, Cabello AM, Coppola L, Cornejo-Castillo FM, d'Ovidio F, De Meester L, Ferrera I, Garet-Delmas MJ, Guidi L, Lara E, Pesant S, Royo-Llonch M, Salazar G, Sánchez P, Sebastian M, Souffreau C, Dimier C, Picheral M, Searson S, Kandels-Lewis S; Tara Oceans coordinators, Gorsky G, Not F, Ogata H, Speich S, Stemmann L, Weissenbach J, Wincker P, Acinas SG, Sunagawa S, Bork P, Sullivan MB, Karsenti E, Bowler C, de Vargas C, Raes J. 2015. Determinants of community structure in the global plankton interactome. *Science*. 348: 6237. doi:10.1126/science.1262073.

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(5954):858-61. 43. doi:10.1126/science.1179287.

Lorincz M, Cságola A, Farkas SL, Székely C, Tuboly T. 2011. First detection and analysis of a fish circovirus. *J Gen Virol*. 92(8):1817-21. doi:10.1099/vir.0.031344-0.

Mankertz A, Hattermann K, Ehlers B, Soike D. 2000. Cloning and sequencing of columbid circovirus (coCV), a new circovirus from pigeons. *Arch Virol*.145(12):2469-79. doi:10.1007/s007050070002.

Mann, N.H. 2005. The third age of phage. *Plos Biol.* 3(5):e182. doi:10.1371/journal.pbio.0030182.

Mansoor S, Briddon RW, Zafar Y, Stanley J. 2003. Geminivirus disease complexes: an emerging threat. *Trends Plant Sci*. 8(3):128-34. doi:10.1016/S1360-1385(03)00007-4.

Martin DP, Biagini P, Lefeuvre P, Golden M, Roumagnac P, Varsani A. 2011. Recombination in eukaryotic single stranded DNA viruses. *Viruses*. 3(9):1699-738. doi:10.3390/v3091699.

Matczuk AK, Krawiec M, Wieliczko A. 2015. A new duck circovirus sequence, detected in velvet scoter (*Melanitta fusca*) supports great diversity among this species of virus. *Virol J.* 12:121. doi:10.1186/s12985-015-0352-y.

McNulty MS, Allan GM, Connor TJ, McFerran JB, McCracken RM. 1984. An entero-like virus associated with the runting syndrome in broiler chickens. *Avian Pathol*. 13(3):429-39. doi:10.1080/03079458408418545.

Metegnier G, Becking T, Chebbi MA, Giraud I, Moumen B, Schaack S, Cordaux R, Gilbert C. 2015. Comparative paleovirological analysis of crustaceans identifies multiple widespread viral groups. *Mob DNA*. 6:16. doi:10.1186/s13100-015-0047-3.

Middelboe, M., Brussaard, C. 2017. Marine viruses: key players in marine ecosystems. *Viruses*, 9(10): 302. doi:10.3390/v9100302.

Munn CB. 2006. Viruses as pathogens of marine organisms—from bacteria to whales. *J Mar Biol Assoc UK*. 86(3): 453-467. doi:10.1017/S002531540601335X.

Nakasu EYT, Melo FL, Michereff-Filho M, Nagata T, Ribeiro BM, Ribeiro SG, Lacorte C, Inoue-Nagata AK. 2017. Discovery of two small circular ssDNA viruses associated with the whitefly *Bemisia tabaci*. *Arch Virol*. 162(9):2835-2838. doi:10.1007/s00705-017-3425-y.

Nathanson N, Kew OM. 2010. From emergence to eradication: The epidemiology of poliomyelitis deconstructed. *Am J Epidemiol*. 172(11): 1213–1229. doi:10.1093/aje/kwq320.

Needham DM, Chow CE, Cram JA, Sachdeva R, Parada A, Fuhrman JA. 2013. Short-term observations of marine bacterial and viral communities: patterns, connections and resilience. *ISME J*. 7(7):1274-85. doi:10.1038/ismej.2013.19.

Neill W. 1975. Experimental studies of microcrustacean competition, community composition and efficiency of resource utilization. *Ecology*. 56(4):809–826. doi:10.2307/1936293.

Ng TFF, Alavandi S, Varsani A, Burghart S, Breitbart M. 2013. Metagenomic identification of a nodavirus and a circular ssDNA virus in semi-purified viral nucleic acids from the hepatopancreas of healthy *Farfantepenaeus duorarum* shrimp. *Dis Aquat Org.* 105:237-242 18. doi: 10.3354/dao02628.

Ng TFF, Manire C, Borrowman K, Langer T, Ehrhart L, Breitbart M. 2009. Discovery of a novel single-stranded DNA virus from a sea turtle fibropapilloma by using viral metagenomics. *J Virol.* (6):2500-9. doi: 10.1128/JVI.01946-08.

Ng TFF, Willner DL, Lim YW, Schmieder R, Chau B, Nilsson C, Anthony S, Ruan Y, Rohwer F, Breitbart M. 2011. Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. *PLoS One.* 6(6): e20579. doi: 10.1371/journal.pone.0020579.

Ng TFF, Zhang W, Sachsenröder J, Kondov NO, da Costa AC, Vega E, Holtz LR, Wu G, Wang D, Stine CO, Antonio M, Mulvaney US, Muench MO, Deng X, Ambert-Balay K, Pothier P, Vinjé J, Delwart E. 2015. A diverse group of small circular ssDNA viral genomes in human and non-human primate stools. *Virus Evol*. 1(1):vev017. doi:10.1093/ve/vev017.

Niagro FD, Forsthoefel AN, Lawther RP, Kamalanathan L, Ritchie BW, Latimer KS, Lukert PD. 1998. Beak and feather disease virus and porcine circovirus genomes: intermediates between the geminiviruses and plant circoviruses. *Arch Virol*.143:1723–1744. doi:10.1007/s007050050412.

Padilla-Rodriguez M, Rosario K, Breitbart M. 2013. Novel cyclovirus discovered in the Florida woods cockroach *Eurycotis floridana* (Walker). *Arch Virol*. 158(6):1389-92. doi:10.1007/s00705-013-1606-x.

Paez-Espino D, Roux S, Chen IA, Palaniappan K, Ratner A, Chu K, Huntemann M, Reddy TBK, Pons JC, Llabrés M, Eloe-Fadrosh EA, Ivanova NN, Kyrpides NC. 2019. IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Res.* 47(D1):D678-D686. doi:10.1093/nar/gky1127.

Parsons RJ, Breitbart M, Lomas MW, Carlson CA. 2011. Ocean time-series reveals recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso Sea. *ISME J.* 6(2):273-84. doi:10.1038/ismej.2011.101.

Payne, SL. 2016. Chapter 30: Other small DNA viruses. Book: *Viruses* doi:10.1016/B978-0-12-803109-4.00030-1.

Pham HT, Bergoin M, Tijssen P. 2013. *Acheta domesticus Volvovirus*, a novel single-stranded circular DNA virus of the house cricket. *Genome Announc*. 1(2):e0007913. doi:10.1128/genomeA.00079-13.

Pham HT, Yu Q, Boisvert M, Van HT, Bergoin M, Tijssen P. 2014. A circo-like virus isolated from *Penaeus monodon* shrimps. *Genome Announc*. 2(1): e01172-13. doi:10.1128/genomeA.01172-13.19.

Phan TGAK, Wang BA, Rose CA, Lipton RKA, Delwart HLA, Eric E. 2011. The fecal viral flora of wild rodents. *PLoS. Pathog*. 7 e1002218. doi:10.1371/journal.ppat.1002218.

Quaiser A, Krupovic M,, Dufresne A, Francez AJ, Roux S. 2016. Diversity and comparative genomics of chimeric viruses in Sphagnum-dominated peatlands. *Virus Evol*. 2(2):vew025. doi:10.1093/ve/vew025.

Reuter G, Boros A, Delwart E, Pankovics P. 2014. Novel circular single-stranded DNA virus from turkey faeces. *Arch Virol.*159:2161–2164. doi:10.1007/s00705-014-2025-3

Rinder M, Schmitz A, Peschel A, Korbel R. 2015. Complete genome sequence of a novel circovirus from zebra finch. *Genome Announc*. 3(3): e00560-15. doi:10.1128/genomeA.00560-15.

Ritchie BW, Niagro FD, Lukert PD, Steffens WL 3rd, Latimer KS. 1989. Characterization of a new virus from cockatoos with psittacine beak and feather disease. *Virology*. 171(1):83-8. doi:10.1016/0042-6822(89)90513-8.

Rohwer F, Vega Thurber R. 2009. Viruses manipulate the marine environment. *Nature*. 459: 207–21. doi:10.1038/nature08060.

Rosario K, Dayaram A, Marinov M, Ware J, Kraberger S, Stainton D, Breitbart M, Varsani A. 2012. Diverse circular ssDNA viruses discovered in dragonflies (Odonata: *Epiprocta*). *J Gen Virol.* 93(12):2668-81. doi: 10.1099/vir.0.045948-0.

Rosario K, Duffy S, Breitbart M. 2012. A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Arch Virol*. 157(10):1851-71. doi:10.1007/s00705-012-1391-y.

Rosario K, Marinov M, Stainton D, Kraberger S, Wiltshire EJ, Collings DA, Walters M, Martin DP, Breitbart M, Varsani A. 2011. Dragonfly cyclovirus, a novel single-stranded DNA virus discovered in dragonflies (Odonata: *Anisoptera*). *J Gen Virol.* 92(6):1302-8. doi:10.1099/vir.0.030338-0.

Rosario K, Mettel KA, Benner BE, Johnson R, Scott C, Yusseff-Vanegas SZ, Baker CCM, Cassill DL, Storer C, Varsani A, Breitbart M. 2018. Virus discovery in all three major lineages of terrestrial arthropods highlights the diversity of single-stranded DNA viruses associated with invertebrates. *PeerJ*. 11;6:e5761. doi: 10.7717/peerj.5761.

Rosario K, Nilsson C, Lim YW, Ruan Y, Breitbart M. 2009. Metagenomic analysis of viruses in reclaimed water. *Environ Microbiol.* 11(11):2806-20. doi: 10.1111/j.1462-2920.2009.01964.x. 41.

Rosario, K, Schenck RO, Harbeitner RC, Lawler SN, Breitbart M. 2015. Novel circular single-stranded DNA viruses identified in marine invertebrates reveal high sequence diversity and consistent predicted intrinsic disorder patterns within putative structural proteins. *Front Microbiol*. 6:696. doi:10.3389/fmicb.2015.00696 23.

Rose N, Opriessnig T, Grasland B, Jestin A. 2012. Epidemiology and transmission of porcine circovirus type 2 (PCV2). *Virus Res.* 164(1-2):78-89. doi: 10.1016/j.virusres.2011.12.002.

Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, Kuhn JH, Lavigne R, Brister JR5, Varsani A, Amid C, Aziz RK, Bordenstein SR, Bork P, Breitbart M, Cochrane GR, Daly RA, Desnues C, Duhaime MB, Emerson JB, Enault F, Fuhrman JA22, Hingamp P, Hugenholtz P, Hurwitz BL, Ivanova NN, Labonté JM, Lee KB, Malmstrom RR, Martinez-Garcia M, Mizrachi IK, Ogata H, Páez-Espino D, Petit MA, Putonti C, Rattei T, Reyes A, Rodriguez-Valera F, Rosario K, Schriml L, Schulz F, Steward GF, Sullivan MB, Sunagawa S, Suttle CA, Temperton B, Tringe SG, Thurber RV, Webster NS, Whiteson KL, Wilhelm SW, Wommack KE, Woyke T, Wrighton KC, Yilmaz P, Yoshida T, Young MJ, Yutin Z, Allen LZ, Kyrpides NC, Eloe-Fadrosh EA. 2019. Minimum information about an uncultivated virus genome (MIUViG). *Nat Biotechnol*. 27(1):29-37.doi:10.1038/nbt.4306.

Roux S, Hallam SJ, Woyke T, Sullivan MB. 2015. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife* 2015;4:e08490. doi:10.7554/eLife.08490.

Roux S, Solonenko NE, Dang VT, Poulos BT, Schwenck SM, Goldsmith DB, Coleman ML, Breitbart M, Sullivan MB. 2016. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ*. 4:e2777. doi:10.7717/peerj.2777.

Rybicki EP. 1994. A phylogenetic and evolutionary justification for three genera of *Geminiviridae Arch Virol*. 139:49-77. doi:10.1007/BF01309454.

Sachsenroder J, Braun A, Machnowska P, Ng TF, Deng X, Guenther S, Bernstein S, Ulrich RG, Delwart E, Johne R. 2014. Metagenomic identification of novel enteric viruses in urban wild rats and genome characterization of a group A rotavirus. *J Gen Virol.* 95:2734–2747. doi:10.1099/vir.0.070029-0.

Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. 2010. Viral mutation rates. *J Virol.* 84(19):9733-9748. doi:10.1128/JVI.00694-10.

Short CM, Suttle CA. 2004. Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. *Appl Environ Microbiol*. 71(1):480–486. doi:10.1128/AEM.71.1.480-486.2005.

Sikorski A, Dayaram A, Varsani A. 2013. Identification of a novel circular DNA virus in New Zealand fur seal (*Arctocephalus forsteri*) fecal matter. *Genome Announc*. 1(4): e00558-13. doi:10.1128/genomeA.00558-13.

Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B, Hull R, King AM, Koonin EV, Krupovic M, Kuhn JH, Lefkowitz EJ, Nibert ML, Orton R, Roossinck MJ, Sabanadzovic S, Sullivan MB, Suttle CA, Tesh RB, van der Vlugt RA, Varsani A, Zerbini FM. 2017. Consensus statement: Virus taxonomy in the age of metagenomics. *Nat Rev Microbiol*. 15(3):161-168. doi:10.1038/nrmicro.2016.177.

Smits SL, Raj VS, Oduber MD, Schapendonk CM, Bodewes R, Provacia L, Stittelaar KJ, Osterhaus AD, Haagmans BL. 2013. Metagenomic analysis of the ferret fecal viral flora. *PLoS ONE*. 8:e71595. doi:10.1371/journal.pone.0071595.

Soffer N, Brandt ME, Correa AM, Smith TB, Thurber RV. 2014. Potential role of viruses in white plague coral disease. *ISME J*. 8(2):271-83. doi:10.1038/ismej.2013.137.

Somayaji V, DeNardo D, Wilson Sayres MA, Blake M, Waits K, Fontenele RS, Kraberger S, Varsani A. 2018. Genome sequence of a single-stranded DNA virus identified in gila monster feces. *Microbiol Resour Announc*. 7(7): e00925-18. doi:10.1128/MRA.00925-18

Steel O, Kraberger S, Sikorski A, Young LM, Catchpole RJ, Stevens AJ, Ladley JJ, Coray DS, Stainton D, Dayaram A, Julian L, van Bysterveldt K, Varsani A. 2016. Circular replication-associated protein encoding DNA viruses identified in the faecal matter of various animals in New Zealand. *Infect Genet Evol*. 43:151-64. doi: 10.1016/j.meegid.2016.05.008.

Stenzel T, Farkas K, Varsani A. 2015. Genome sequence of a diverse goose circovirus recovered from greylag goose. *Genome Announc*. 3(4): e00767-15. doi: 10.1128/genomeA.00767-15.

Steward GF, Montiel JL, Azam F. 2000. Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. *Limnol Oceanogr* 45(8):1697–1706. doi:10.4319/lo.2000.45.8.1697.

Stewart ME, Perry R, Raidal SR. 2006. Identification of a novel circovirus in Australian ravens (*Corvus coronoides*) with feather disease. *Avian Pathol*. 35(2):86-92. doi:10.1080/03079450600597345.

Suttle, C.A. 2007. Marine viruses — major players in the global ecosystem. *Nat Rev Microbiol*, 5: 801– 812. doi:10.1038/nrmicro1750.

Tarján ZL, Pénzes JJ, Tóth RP, Benkő M. 2014. First detection of circovirus-like sequences in amphibians and novel putative circoviruses in fishes. *Acta Vet Hung*. 62(1):134-44. doi:10.1556/AVet.2013.061.

Thézé J, Leclercq S, Moumen B, Cordaux R, Gilbert C. 2014. Remarkable diversity of endogenous viruses in a crustacean genome. *Genome Biol Evol*. 6(8):2129-40. doi:10.1093/gbe/evu163.

Thingstad, T.F. 2000. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol Oceanogr*. 45 (6): 1320–1328. doi:10.4319/lo.2000.45.6.1320.

Todd D, McNulty MS, Adair BM, Allan GM. 2011. Animal circoviruses. *Adv Virus Re*s. 57:1-70. doi:10.1016/S0065-3527(01)57000-1.

Todd D, Scott AN, Fringuelli E, Shivraprasad HL, Gavier-Widen D, Smyth JA. 2007. Molecular characterization of novel circoviruses from finch and gull. *Avian Pathol*. 36(1):75-81. doi:10.1080/03079450601113654.

van den Brand JM, van Leeuwen M, Schapendonk CM, Simon JH, Haagmans BL, Osterhaus AD, Smits SL. 2012. Metagenomic analysis of the viral flora of pine marten and European badger feces. *J Virol*. 86(4):2360-5. doi:10.1128/JVI.06373-11.

Varsani A, Krupovic M. 2018. Smacoviridae: a new family of animal-associated single-stranded DNA viruses. *Arch Virol*. 163(7):2005-2015. doi:10.1007/s00705-018-3820-z.

Wang B, Sun LD, Liu HH, Wang ZD, Zhao YK, Wang W, Liu Q. 2018. Molecular detection of novel circoviruses in ticks in northeastern China. *Ticks Tick Borne Dis*. 9:836–839. doi:10.1016/j.ttbdis.2018.03.017.

Wang H, Li S, Asif M, Yang S, Wang X, Shen Q Shan T, Deng X, Li J, Hua X, Cui L, Delwart E, Zhang W. 2018. Plasma virome of cattle from forest region revealed diverse small circular ssDNA viral genomes. *Virol J*. 15:11. doi:10.1186/s12985-018-0923-9.

West TO, Marland G, Singh N, Bhaduri BL, Roddy AB. 2009. The human carbon budget: an estimate of the spatial distribution of metabolic carbon consumption and release in the United States. *Biogeochemsitry*. 94(1): 29-41. doi:10.1007/s10533-009-9306-z.

Woo PC, Lau SK, Teng JL, Tsang AK, Joseph M, Wong EY, Tang Y, Sivakumar S, Bai R, Wernery R, Wernery U, Yuen KY. 2014. Metagenomic analysis of viromes of dromedary camel fecal samples reveals large number and high diversity of circoviruses and picobirnaviruses. *Virology*. 471–473:117–125. doi:10.1016/j.virol.2014.09.020.

Yoshida M, Takaki Y, Eitoku M, Nunoura T, Takai K. 2013. Metagenomic analysis of viral communities in (hado)pelagic sediments. *PLoS One*. 8(2):e57271. doi:10.1371/journal.pone.0057271.

Zaccaria G, Malatesta D, Scipioni G, Felice ED, Campolo M, Casacciaa C, Savinia G, Sabatinoa DD, Lorussoa A. 2016. Circovirus in domestic and wild carnivores: an important opportunistic agent? *Virology*. 490: 69–74 34. doi:10.1016/j.virol.2016.01.007.

Zawar-Reza P, Arguello-Astorga GR, Kraberger S, Julian L, Stainton D, Broady PA, Varsani A. 2014. Diverse small circular single-stranded DNA viruses identified in a freshwater pond on the McMurdo Ice Shelf (Antarctica). *Infect Genet Evol*. 26:132–138. doi:10.1016/j.meegid.2014.05.018.

Zhang Z, Jia R, Wang M, Lu Y, Zhu D, Chen S, Yin Z, Wang Y, Chen X, Cheng A. 2013. Complete genome sequence of the novel duck circovirus strain GH01 from Southwestern China. *Genome Announc*. 1(1):e00166-12. doi:10.1128/genomeA.00166-12.

Zhao L, Rosario K, Breitbart M, Duffy S. 2019. Eukaryotic circular *rep*-encoding single-stranded DNA (CRESS-DNA) viruses: ubiquitous viruses with small genomes and a diverse host range. *Advances in Virus Research*, Ch.3 Vol. 103:71-133. doi:10.1016/bs.aivir.2018.10.001.

CHAPTER 2

MICROCRUSTACEANS HARBOR COSMOPOLITAN AND MICRODIVERSE CIRCULAR SSDNA VIRUSES

2.1 Abstract **|** Circular replication initiator-protein encoding single stranded DNA (CRESS-DNA) viruses are pervasive components of crustacean viral consortia. Often encoding only two ORFs, these are among the smallest viruses known to affiliate with invertebrates. This study assembled 215 microcrustacean-associated CRESS-DNA virus-like contigs (MCVs) from 23 unique host species (31 viromes), illustrating their frequent detection in aquatic ecosystems. MCVs associated with microcrustaceans of similar metataxonomic group or which occupied overlapping biogeographical ranges shared greater ORF nucleotide identity and codon usage patterns relative to those recovered from unrelated metataxonomic groups or distant collection sites. This may be indicative of evolutionary convergence or shared origin, aiding in genotype classification and denoting that both host and biogeography likely contribute to genotype similarity. Shared MCVs harbored among disparate microcrustacean taxa may limit net MCV diversity. However, read recruitment patterns provide evidence of host specificity, suggesting that microcrustaceans may act as secondary hosts for CRESS-DNA viral genotypes. Accumulation of virome-specific variants may further indicate a high degree of temporal and taxonomic specificity. Novel MCVs did not accrue nonsynonymous variants uniformly, with less conservation in intergenic regions and greater frequencies of A/T-rich variants among malacostracans, mirroring codon usage paradigms. Endogenized viral elements provided paleovirological confirmation that viruses similar to these novel MCVs once infected microcrustacean hosts. Therefore, MCVs identified in this study likely exemplify a highly diverse group of arthropod viruses with specific host ranges, weak gene homology, and rapid evolutionary rate that may be optimally employed to better understand the distribution and ecology of ssDNA viruses in aquatic arthropods.

2.2 Introduction | Viruses revise, exchange, and transmit genetic information throughout all biomes. These viruses are not only the most abundant biological entities on the planet (Suttle, 2005), but represent an immense reservoir of genetic diversity, much of which governs cellular community composition and evolve on observable scales (Duffy et al, 2008; Rohwer & Thurber, 2009; Roux et al, 2015; Wommack et al, 2000). By definition, these obligate parasites exist to replicate and advance genotype fitness by adapting to host physiology and refining transmission strategies. As such, interpretation of viral ecology is contingent on host and environmental context. Therefore, investigating viral consortia provides strategic opportunities to better understand patterns of selection, particularly in understudied non-model metazoan systems, and observe how ecosystems shape the evolution of some of the smallest biological entities.

In aquatic ecosystems, viruses maintain food web dynamics through host lysis, metabolic manipulation, nutrient cycling, and horizontal gene transfer (Rohwer 2009). In recent years, principle initiatives for viral discovery have been driven by untargeted whole-community sequencing approaches (viromics; Breitbart et al, 2002; Paez-Espino et al, 2017; Roux et al, 2015, Simmonds et al, 2017). Within this framework, viruses with circular replication initiator protein-encoding single stranded DNA ("CRESS-DNA") genomes, represent a disproportionate diversity of novel and unclassified DNA viruses relative to those with linear or dsDNA genomes, largely due to artificial enrichment in virome construction and sequencing (i.e. via pre-filtration and the use of Φ29 polymerase pre-amplification). These non-enveloped, icosahedral viruses typically contain small genomes (<6kb) and broadly include bacteriophage (*Microviridae*), plant pathogens (*Geminiviridae, Nanoviridae*), and metazoan-associated viruses (*Anelloviridae, Circoviridae, Parvoviridae*), among others (Eaglesham et al 2013; Labonté & Suttle 2013; Rosario et al, 2012; Rosario et al, 2018).

At minimum, CRESS-DNA viral genomes encode a nonstructural replication initiator protein (*rep*) and a structural capsid protein (*cap*), proliferating via a conserved rolling circle replication (RCR) mechanism (Rosario et al, 2012). These viruses exhibit rapid rates of evolution

approaching those of RNA viruses, an impressive degree of genomic diversity (<80% genome-wide identity), and flexible genome architecture (Duffy et al, 2008; Rosario et al, 2012). Therefore, CRESS-DNA viruses identified in viromic datasets are identified via conservation of *rep* and ORF orientation relative to a putative origin of replication (nonanucleotide motif; Rosario et al, 2012). Viral surveillance demonstrates that these viruses are detected in an assortment of environments, ranging from soils and sediments (Kim et al, 2008; Yoshida et al, 2013) to water columns and wastewater effluents (Kraberger et al, 2015; Labonté et al, 2013; Rosario et al, 2009; Roux et al, 2013) from the Antarctic to the tropics (López-Bueno, 2009; Tamaki et al, 2012; Zawar-Reza et al, 2014). Likewise, CRESS-DNA viruses appear to be pervasive and diverse members of non-model aquatic invertebrate viromes capable of widespread distribution (Eaglesham et al, 2013; Rosario et al, 2015; Rosario et al, 2018; Shulman & Davidson, 2017). For example, CRESS-DNA viruses have been identified in ctenophores (Breitbart et al, 2015), cnidarians (Soffer et al, 2014), gastropods (Dayaram et al, 2016; Rosario et al, 2015), echinoderms (Fahsbender et al, 2015; Jackson et al, 2016), and crustaceans (Bistolas et al, 2017; Dunlap et al, 2013; Hewson et al, 2013a,b), among others.

This study explores CRESS-DNA viruses associated with abundant populations of aquatic microcrustaceans – arthropod mesograzers typically less than 10 millimeters in length - from disparate ecosystems. Copepods, amphipods, isopods, and brachiopods exhibit a range of ecological strategies, serving as critical links within aquatic food webs by maintaining the export of carbon and other macronutrients, attenuating primary productivity, and serving as food for upper trophic level consumers. Therefore, microcrustaceans often encompass the greatest phylogenetic and functional diversity of metazoans within aquatic ecosystems and play numerous roles in nutrient cycling through growth, respiration, grazing, excretion, and mortality (Dole-Oliver et al, 2000; Gismervik, 1997; Neill, 1975). The degree of microcrustacean mortality, metabolic change, and impact on population dynamics attributable to their viruses remains unknown.

High-throughput sequencing suggests that ssDNA viruses comprise a significant component

of microcrustacean viral consortia and recent discoveries are a critical step in understanding ecologically relevant host-virus interactions and expanding our knowledge of diversity within these populations. Nevertheless, studies focused solely on viral discovery miss the opportunity to address fundamental, hypothesis-driven questions about viral ecology and evolution (Holmes, 2007). High rates of evolutionary change (Duffy et al, 2008), environmental persistence (Allen et al, 1994), and apparent ubiquity in association with invertebrates (Rosario et al, 2015, 2018) render microcrustacean-associated CRESS-DNA viruses, or "MCVs", ideal candidates to explore the patterns and drivers of host-associated viral diversity.

This study aims to leverage MCVs as a tractable clade for community-wide viromic sequencing to infer ssDNA virus biogeography, ecology, and evolutionary dynamics from viral phylogenomics. Through inter- and intra-ecosystem sequencing of representative microcrustacean viral consortia, this study aims to (1) classify novel CRESS-DNA viral constituents of the metazoan holobiont based on genomic characteristics and define the microcrustacean and ecosystem-specific parameters that may contribute to the selection for these genomic features, (2) determine if computational tools may lend insight into the conundrum of widely detected, yet microcrustacean-specific CRESS-DNA viral genotypes, (3) evaluate the microdiversity of novel MCV genotypes to better understand viral protein conservation, putative host phylogeny, and biogeography, and finally, (4) assess crustacean genomes for endogenized viral elements to construct a paleovirological record of CRESS-DNA virus infection in microcrustaceans. While purely computational (and therefore hypothetical), these estimates may provide a foundation to further explore the widespread distribution and rapid evolution of ssDNA viruses in aquatic ecosystems to assess the contribution of CRESS-DNA viruses to microcrustacean mortality.

2.3 Methods **|** *Specimen collection* – DNA viral assemblages associated with nine copepod, nine amphipod, four isopod, and one brachiopod species (31 haplotype-specific libraries) from abundant microcrustacean populations in disparate aquatic habitats were sequenced to compare associated

DNA viral phylogeography on a multi-ecosystem scale (Table 2.1). Due to variation in sample site and species, collection was modified accordingly: (1) benthic specimens were collected via Ponar benthic samplers, sieved to remove sediment, and rinsed *in situ*, (2) zooplankton were collected via net tows and rinsed *in situ*, (3) littoral species were individually collected and rinsed in phosphate-buffered saline (PBS). All samples were then individually selected via forceps, flash frozen, and transported in liquid nitrogen to be preserved at -80°C.

Virome preparation and sequencing – Efforts to understand the scope and ecological significance of viral diversity in invertebrates has been aided tremendously by the emergence of accessible high throughput sequencing technology. To generate viromes, specimens were homogenized (5-10min; 2.0mm BashingBead™ Lysis Tubes, Zymo Research, Irvine, CA, USA) and processed per Ng et al (2010), with modifications detailed in Vega Thurber (2009) and Bistolas et al (2017a). Homogenates were filtered to reduce cellular contamination (0.2µm PES syringe filtration, VWR International, Radnor, PA, USA), precipitated via polyethylene glycol to concentrate biomass (10% PEG-8000 by weight), resuspended and enzymatically digested to diminish non-encapsidated nucleic acids (2.5 U DNase I, 0.25 U RNase and 1 U Benzonase, Sigma-Aldrich, St. Louis, MO, USA). Nucleic acids were then extracted with a kit that utilizes the chaotropic agent, guanidinium thiocyanate, to disrupt capsids prior to solid phase extraction on a silica spin column (ZR viral extraction kit, Zymo Research, Irvine, CA, USA). Resulting DNA template underwent isothermal rolling circle amplification per manufacturer instructions (multiple displacement amplification, or "MDA"; Genomiphi Whole Genome Amplification Kit, GE Healthcare, Little Chalfont, UK), with successful reactions confirmed via PicoGreen quantification (ThermoFisher, Waltham, MA, USA) and gel electrophoresis. Products were fragmented (400–600bp) and ligated to adapter oligomers (Nextera XT DNA Library Preparation Kit; Illumina, San Diego, CA, USA) prior to 2x250bp paired-end Illumina MiSeq sequencing (Cornell University Core Laboratories Center, Ithaca, NY, USA). As our objective was to non-quantitatively identify novel CRESS-DNA viruses, viromes were deliberately enriched for small, circular, ssDNA templates at multiple stages of preparation and

sequencing (e.g. 0.02μm filtration, pre-amplification virion concentration, use of enzymes TURBO™ DNase and ϕ29 polymerase, nucleic acid extraction kit with a single elution, etc; Brinkman et al, 2018).

Viral annotation – Viromic sequencing generated >80 million reads after trimming and quality control (CLC Genomics Workbench v.8.5.1, Qiagen, Hilden, Germany, FastQC v.0.11.8), with sequencing depth reaching saturation in most libraries (Table 2.1; SI Table 2.1; SI Figure 2.1). Reads were assembled *de novo* within sequence libraries to generate >$15.3x10^4$ contiguous sequences ("contigs"; average N50=1772.9±93.3SE) using CLC Genomics Workbench and extended via SPAdes (v.3.6.2, Bankevich et al, 2012). Contigs were compared to the NCBI non-redundant and viral databases (nr and gbvrl; BLASTx v. 2.7.1, Diamond v.0.9.17, e-value<$1x10^{-5}$; Altschul et al, 1990), and those annotated as CRESS-DNA virus-like contigs were further validated for characteristic genomic features, completeness, and novelty (detailed in SI Table 2.2).

Assessment of phylogenetic context – Trimmed MCV *rep* ORFs, associated best BLASTx hits (113 unique genomes), and 462 CRESS-DNA viruses retrieved from public databases (including circoviruses, anelloviruses, geminiviruses, genomoviruses, invertebrate-associated CRESS-DNA viruses, and other unclassified environmental CRESS-DNA viruses, retrieved from NCBI, 2018) were utilized for phylogenetic comparison. Single read archive (NCBI-SRA) databases were further collated for read recruitment and phylogenetic comparison. Reciprocal comparisons between *rep* ORF sequences via tBLASTx (e-value<$1x10^{-5}$) were visualized via Cytoscape v.3.6.1. ORF orientation was detected and visualized in CLC Genomics Workbench (v.8.5.1). Predicted MCV *rep* ORFs were aligned in Muscle (v.3.8 Edgar et al, 2004) on both nucleotide and predicted amino acid levels. However, maximum likelihood phylogenies implemented in PhyML (v.3.0, Lefort et al, 2017) yielded uninformative clades with low branch support, prompting non-alignment based clustering methods to estimate phylogenetic similarity. Therefore, feature frequency profiling (FFP) was implemented per Sims et al (2009), including assessment for optimal kmer length (n=10), with resulting profiles visualized via multidimensional scaling.

Metataxonomic Group	Species	Collection Site	Latitude	Longitude	Salinity [1]	General ecosystem description	Collection Method [2]	Collection mo/year	Reads after trimming (X10 ^{6})	N50	Mean contig length (nt)	Total contigs	% Reads assembled
Amphipod	*Ampelisca spp.*	Moreton Bay Research Station, AUS	-27.495	153.401	M	Intertidal, Sand flats	NS	12/15	3.59	2044	1611	2956	95.0
	Diporeia spp.	Lake Cayuga, USA	42.707	-76.722	F	Freshwater Lake	PS	8/14	3.06	1879	1502	1671	94.7
		Lake Superior, USA	48.297	-88.966	F		PS	8/14	1.53	3234	1914	1830	93.0
		Lake Michigan, USA	42.383	-87.000	F		PS	8/14	1.51	1587	1378	2264	91.1
		Lake Huron, USA	45.278	-82.452	F		PS	8/14	2.24	2638	1749	2200	95.5
	Echinogammarus spp.	Lake Erie, USA	41.892	-83.197	F		PS	8/14	2.73	2515	1745	1056	96.4
	Gammaridean amphipoda	Heron Island Research Station, AUS	-23.442	151.913	M	Intertidal, Reef	NS	12/15	2	2094	1608	5264	88.9
			-23.442	151.913	M		NS	12/15	3.35	1413	1284	25031	71.8
		Moreton Bay Research Station, AUS	-27.495	153.401	M	Intertidal, Sand flats	NS	12/15	2.63	2210	1625	5273	90.3
		Salish Sea, USA	48.427	-122.902	M	Sound	II	9/15	4.42	1460	1305	6157	94.2
	Hyperia spp.	Sargasso Sea	27.667	-64.767	M	Oligotrophic pelagic marine	PT	10/14	3.42	1592	1379	71	97.4
	Sunamphitoe spp.	Port Townsend, USA	48.142	122.782	M	Kelp beds, nearshore	II	9/15	3.94	1841	1502	5912	93.3
		USC Wrigley Marine Science Center, USA	33.445	-118.484	M		II	1/15	1.73	2580	1783	977	97.3
	Themisto spp.	Salish Sea, USA	48.427	-122.902	M	Sound	PT	1/16	3.88	1328	1221	6682	92.4
Copepod	*Acartia spp.*	Sargasso Sea	21.667	-65.667	M	Oligotrophic Pelagic Marine	PT	10/14	0.61	2191	1634	308	94.3
		Sargasso Sea	21.667	-65.667	M		PT	10/14	4.74	1843	1474	161	94.7
	Candacia spp.	Sargasso Sea	29.667	-64.767	M		PT	10/14	3.34	1736	1424	969	96.8
	Clausocalanus spp.	Sargasso Sea	29.667	-64.767	M		PT	10/14	3.04	1110	1071	7788	71.9
	Diacyclops thomasi	Lake Michigan, USA	42.383	-87.000	F	Freshwater lake	PT	10/14	3.04	1874	1467	1882	95.0
	Diaptomus minutus	Oneida Lake, USA	43.201	-76.075	F		PT	4/15	0.41	1452	1307	2050	83.7
	Limnocalanus macrurus	Lake Michigan, USA	42.383	-87.000	F		PT	8/14	3.02	1261	1143	7609	69.5
	Macrosetella gracilis	Sargasso Sea	24.667	-65.217	M	Oligotrophic pelagic marine	PT	10/14	3	1411	1234	5788	78.1
	Miracia efferata	Sargasso Sea	29.667	-64.767	M		PT	10/14	5.49	1376	1265	635	95.1
		Sargasso Sea	29.667	-64.767	M		PT	10/14	4.17	1662	1366	1631	96.0
	Skistodiaptomus oregonensis	Lake Owasco, USA	42.707	-76.722	F	Freshwater lake	PT	8/14	2.98	1172	1104	8445	66.1
Brachiopod	*Daphnia mendotae*	Oneida Lake, USA	43.020	-75.949	F		PT	4/15	0.64	1631	1400	3449	83.7
			43.020	-75.949	F		PT	5/15	0.296	1345	1269	957	79.5
Isopod	*Gnorimosphaeroma oregonensis*	Port Townsend,USA	48.142	122.782	M	Rocky intertidal	II	9/15	3.93	1448	1289	20911	83.3
	Idotea (Pentidotea) resecata	USC Wrigley Marine Science Center, USA	33.445	118.484	M		II	1/15	2.83	1236	1179	6051	70.1
	Idotea (Pentidotea) wosnesenskii	Port Townsend, USA	48.142	122.782	M		II	9/15	3.78	1246	1202	12112	89.1
	Sphaeroma spp.	Heron Island Research Station, AUS	-23.442	151.913	M	Intertidal, Reef	NS	12/15	2.88	2550	1843	5164	93.4

Table 2.1 | *Taxonomic and biogeographic metadata associated with 31 viromes sequenced in this study*. [1]F=Freshwater, M=Marine; [2]PT: plankton tow, PS: ponar sampler, NS: nearshore seine/algae shakes, II: individually selected from intertidal vegetation with forceps.

Genomic content and codon usage patterns – Total %GC content, %GC3 content (guanine and cytosine presence at third codon position), and α-/β-diversity of oligomer diversity (i.e. the diversity of 10-mers within each genotype) were compared among putatively novel MCV and *rep* ORFs. These parameters were computed via webserver CAIcal (genomes.urv.es/CAIcal/), Feature Frequency Profiling (FFP; Sims et al, 2009), and RStudio v.1.1.456. Furthermore, biases in codon usage arise when viruses utilize non-random codons to encode for the same amino acid, possibly to better facilitate replication by host polymerases or maintain genomic stability. Codon usage patterns of complete MCV *rep* ORFs were evaluated using CodonW v.1.4.4 (http://codonw.sourceforge.net/). Codon usage heterogeneity (i.e. "bias") was assessed via normalized effective number of codons (ENC, N=20-61), where greater number of observed codons denotes random codon usage and therefore reduced bias. Likewise, relative synonymous codon usage (RSCU) was applied to quantify this bias, where over- or under-utilization of AT- or GC-terminating codons represented directional bias (RSCU > 1.6 or RSCU < 0.6, respectively). These metrics were targeted to compare parameters conceivably governing *rep* ORF selection and to establish co-variation between sequence composition and host specific or biogeographic patterns.

Sequence conservation and genotype distribution – Percent identity ($\%ID_{rep}$) over the length of *rep* ORFs was assessed using pairwise alignment-based comparison via the sequence demarcation tool (SDT v.1.2, Muhire et al, 2014), where *rep* ORFs with pairwise identity >95% delineated single sequence clusters. Average genome nucleotide identity via BLAST-based pairwise comparison (ANIb) among putatively novel MCVs was calculated in Gegenees (v.3.1) with 25nt step size, 50nt fragment size, and 15% threshold cutoff (Ågren et al, 2012). Genotype distribution via detection in read recruitment over contig length and average coverage of contigs within libraries was visualized via histograms in Rstudio (90% similarity, 90% read length; CLC Genomics Workbench v.8.5.1; RStudio v.1.1.456). Comparison of read recruitment within sequence libraries relative to between libraries (“cross-recruitment”) was performed via CLC workbench (90% similarity, 80% read length). Prevalence, frequency, and probability distribution of single nucleotide variants (SNVs),

multinucleotide variants (MNVs), rearrangements, insertions, and deletions were identified via a multinomial model for low frequency variant calling per site-dependent read coverage implemented in CLC workbench. Predicted nonsynonymous and synonymous consequences of variation were detected when ORFs were successfully translated and normalized to ORF length and coverage.

Host prediction and evidence of genomic integration – Endogenized viral elements (EVEs) were detected by comparing 23 publicly available and assembled crustacean genomes (2 amphipods, 5 brachiopods, 10 copepods, 4 decapods, and 2 isopods) from the NCBI Genomes database (SI table 2.3) against 215 putatively novel MCV genomes and 462 reference CRESS-DNA genomes via BLASTx (e-value $<1X10^{-5}$). Detected CRESS-DNA viral EVEs were distinguished between potentially novel and previously characterized sequences through comparison and manual alignment to all CRESS-DNA viruses archived in NCBI (Genbank). To determine identities of neighboring ORFs and detect possible transposable and repeat elements, sequences were trimmed *in silico* within 2kb of predicted EVE sites and annotated via RepBase (v. 23.04, Bao et al, 2015), Dfam (v.2.9, Hubley et al, 2016), and NCBI non-redundant (nr) databases. Internal stop codons and frameshifts were detected via translation and comparison to the closest comparative sequence.

2.4 Results & Discussion

2.4.i. Discovery of novel microcrustacean CRESS-DNA viruses (MCVs)

Viromic discovery - CRESS-DNA viruses were detected in 76% of microcrustacean species, and 22 of 31 viromes. Due to sequencing depth and biases specific to multiple displacement amplification ("MDA"; e.g. overamplification of circular, ssDNA templates via ϕ29 polymerase rolling circle replication; see Appendix III), detection of these ssDNA viruses was predictable (Dean et al, 2001; Kim et al, 2008), yet their prevalence may also indicate that these genotypes circulate widely among aquatic mesograzer populations collected from a breadth of ecosystems. The characteristically cosmopolitan nature and diversity of CRESS-DNA viruses offers an opportunity to explore the parameters that constrain ssDNA viral composition and dispersal in aquatic ecosystems.

MCVs were predominantly detected via sequence similarity to nonstructural *rep* ORFs. Among 215 MCVs identified, 200 contained complete, translatable *reps* with clearly defined start/stop codons, with 173 of these sharing <95% nucleotide identity across *rep* (utilized as a functional threshold for viral quasispecies cluster or "species" demarcation). 27 sequences exhibited >95% *rep* ORF pairwise identity, but averaged 86.0% identity over the full length of the contig. These sequences were clustered and considered components of the same MCV. Furthermore, for the purposes of this investigation, contigs comprising <80% sequence identity were considered distinct from known viruses in public databases.

These putatively novel, *de novo* assembled MCVs were small, ranging in length from 580bp-5748bp (average ~2kb), and shared sequence similarity to a range of pre-existing CRESS-DNA viruses loosely affiliated with a variety of ecosystems and putative hosts including aquatic sediments, pelagic marine viruses, insects, and vertebrates, among others (SI Table 2.2; SI File 2.1; Shulman et al, 2017). Non-microcrustacean affiliated CRESS-DNA viral genomes typically span 2-6kb in length (average 2.25kb among best-BLASTx hit), indicating that contigs in microcrustacean viromes likely represent partial or truncated genomes. Due to limitations in assembly algorithms that conservatively aim to minimize assembly of dissimilar reads, it remains unclear which contigs represent complete CRESS-DNA viral genomes. However, 17 CRESS-DNA virus-like contigs circularized, potentially denoting complete MCV genomes.

Novel CRESS-DNA viruses share significant similarity to other invertebrate CRESS-DNA viruses – Microcrustacean associated MCVs predominantly shared sequence similarity to other uncultured invertebrate viruses or virioplankton from aquatic ecosystems. CRESS-DNA viruses newly identified via viromic sequencing often do not share genome-wide sequence similarity or hallmark characteristics specific to traditionally demarcated ssDNA families, complicating taxonomic classification and host pairing (Simmonds et al, 2017). Rapid accumulation of sequence variants and potential for lateral gene transfer or genomic reorganization impede clear separation between diverse ssDNA virus clades. Therefore, novel genotypes associated with non-model hosts -

including those identified in this study - may fall along a taxonomic spectrum where early common ancestors are unknown or recurring, reflected in the difficulty to compare novel CRESS-DNA viruses via linear multiple sequence alignment. However, network-based visualization techniques (Figure 2.1) demonstrate that MCV *rep* ORFs (solid nodes) share sequence similarity to other previously identified invertebrate-associated CRESS-DNA viruses (open nodes, with similarity represented by network edges; tBLASTx, e-value $<10^{-5}$ against the NCBI-nr database and reciprocal sequence database).

26% of these MCV *rep* ORFs shared similarity with another novel *rep* ORF from a different host virome ("reciprocal hit," Figure 2.1, bold edges), most of which (96.4%) shared both putative invertebrate macro-taxonomic group and general abiotic habitat characteristics, despite exhibiting >80% dissimilarity in rep ORF identity. It should be noted that tBLASTx comparisons were made against reciprocal (and therefore, small) databases of locally identified MCVs, possibly overestimating significant hits (i.e. e-value $<10^{-5}$). Novel MCVs shared *rep* sequence similarity to MCVs derived from crustaceans of the same order 80.0% of the time, implying a relationship between microcrustacean taxonomy and viral genotype occurrence. For example, MCVs associated with calanoid copepods in the Sargasso Sea shared sequence similarity (reciprocal blast hit) to calanoid copepods in the Laurentian Great Lakes, despite over 2,500km of geographic separation and a vast separation in abiotic conditions (lacustrine v. marine), potentially indicating convergent evolutionary patterns. Likewise, different malacostraca with overlapping and nearly identical intertidal kelp forest habitats in the Salish Sea, but lacking phylogenetic similarity (differ in taxonomic order), shared similar MCVs (quantified via percent identity below), which may suggest habitat-wide dispersion of viral particles, viral vectoring, or neutral/selective accumulation of variants towards sympatric speciation from a single genotype. Collectively, these similarities may infer host and habitat driven ssDNA viral evolution, including convergent evolution or divergence from a similar ancestral genotype.

Sequences from short-read archives constructed from pelagic marine and freshwater (Great

Lakes) viromes were also recruited to novel CRESS-DNA virus-like genotypes in this study, providing evidence of the widespread nature of these ssDNA viruses and the potential for CRESS-DNA viruses to associate with or infect a range of non-model organisms in the water column (single read archive libraries associated with bioprojects PRJEB4419, Sunagawa et al, 2015, and PRJNA262540, Kim et al, 2015, respectively).

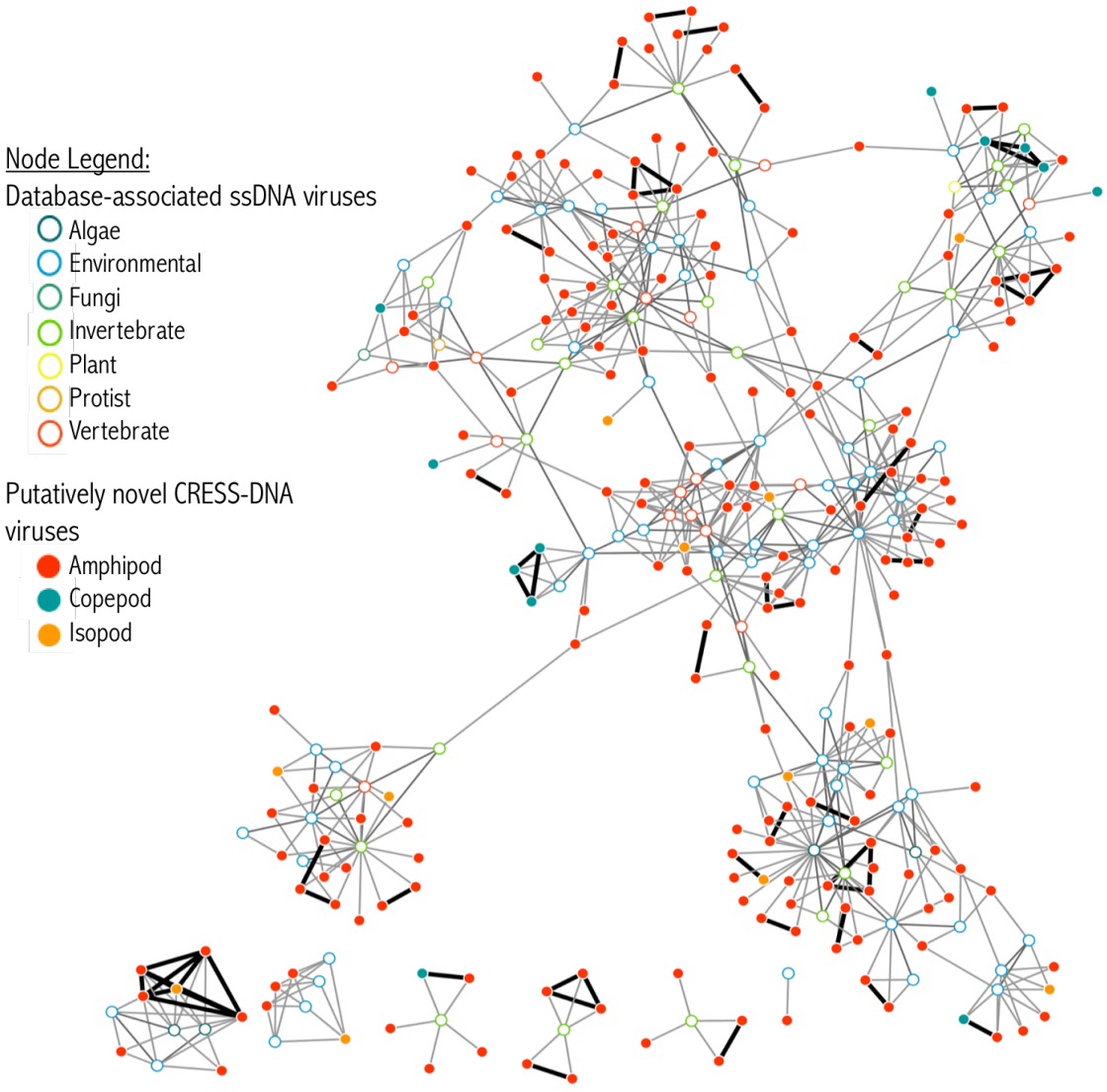

Figure 2.1 | *Network visualization* - illustrating connectivity between putatively novel microcrustacean CRESS-DNA *rep* ORFs (solid nodes) and known CRESS-DNA viruses (open nodes), where edges indicate sequence similarity by reciprocal BLAST hit (tBLASTx, e-value<10^{-5}). 25.6% of novel CRESS-DNA viruses share a reciprocal BLAST hit with a rep ORF from a different virome (bold edges), suggesting similarity between MCV genotypes.

2.4.ii. Genomic composition of novel CRESS-DNA viruses: host phylogeny and biogeographic distance likely drive CRESS-DNA virus genomic dissimilarity

Putative host taxonomy contextualizes viral similarity better than viral taxonomy – Detecting distant sequence homology beyond *rep* ORFs and resolving the evolutionary relationships between MCVs within the context of preexisting ssDNA virus phylogenies posed a particular challenge. The majority of invertebrate CRESS-DNA viral sequences in public databases (i.e. NCBI-Genbank, IMG/VR) share only 40 – 60% sequence similarity (Rosario et al, 2012, 2015). However, in addition to alignment-free sequence similarity-based networks, oligomer (i.e. k-mer) frequency profiles have been demonstrated to aid construction of phylogenetic clusters when traditional multiple-sequence alignment methods are not well suited (Sievers et al, 2018; Sims & Kim, 2011). At the time of analysis (Nov, 2018), novel CRESS-DNA viral contigs were compared to 298 publicly available putative CRESS-DNA viral genomes with complete *rep* ORFs (excluding 358 plant-associated geminiviruses, serving as an outgroup) with complete *rep* ORFs, including 123 affiliated with invertebrates (n=74 of which are aquatic; Figure 2.2).

Initial analyses of metagenome-derived viral sequences illustrate that phylogenies constructed using k-mer binning cluster viral genomes in host-specific groups (Kapoor et al, 2010), rather than viral taxonomy. As predicted, oligomer signature similarity measures of β-diversity (Jensen Shannon divergence) via feature frequency profiling reflected distinct clades of CRESS-DNA viruses that largely parallel the gross evolutionary history of their hosts (ANOSIM $p>0.05$). These MCV 10-mer signatures were most similar to other invertebrate associated circoviruses, suggesting that clustering by host taxonomy supersedes clustering by viral classification. This may further facilitate host-virus pairing and indicates that paradigms of viral taxonomy should prioritize host over genotype characterization. However, within these clusters, MCVs were not accurately parsed by geographic location, or other habitat characteristics via *rep* ORF k-mer signature. Due to their resilience to harsh environmental conditions (Allan et al, 1994) and ubiquity in aquatic environments (Labonté & Suttle, 2013; López-Bueno et al, 2009; Rosario et al, 2009, 2015; Zawar-

Reza et al, 2014), novel CRESS-DNA virus sequences are routinely identified in environmental metaviromes and have been detected in water column samples (e.g. Lake Michigan, Tara Oceans, etc; Bistolas et al, 2017a; Gregory et al, 2019). Therefore, this analysis may serve as a proof-of-concept for predictive host matching among these environmentally derived viral contigs: by applying k-mer frequencies to interrogate partial and complete CRESS-DNA viral genomes in non-host associated samples, it may be possible to predict relationships between CRESS-DNA viruses and possible invertebrate host phyla.

Bias in codon usage patterns and low %GC content is characteristic of MCVs – On average, MCV *rep* ORFs from this study comprised significantly lower oligomer richness (α-diversity) after normalization, relative to CRESS-DNA virus *reps* affiliated with other metazoan viromes ($p<1x10^{-15}$,Welch's ANOVA). This may indicate selective or neutral trends towards small k-mer repertoires among MCVs. Codon degeneracy provides an opportunity to evaluate this sequence simplicity in essential (presumably conserved) ORFs; biases towards small codon usage repertoires among MCVs may drive low α-diversity in oligomer profiles and sequence clustering in multiple dimensional scaling (Figure 2.2).

Effective number of codons (N_c) was applied as a bulk measure of unique codons used within a genome normalized to amino acid diversity/length, revealing that there were a significantly smaller number of codons employed among MCV *rep* ORFs ($p<0.0001$, one-way analysis of means, Figure 2.3). Codon heterogeneity often represents selection-driven bias to utilize specific codons to encode synonymous amino acids (Shackelton et al, 2006). This was reflected by a measure of relative synonymous codon usage (RSCU), indicating that specific synonymous codons are over-/under-represented in microcrustacean associated sequences in comparison to other viral groups or CRESS-DNA viral *rep* ORFs associated with vertebrate hosts ($p<0.0001$, Figure 2.3). Interestingly, anelloviruses (often associated with vertebrate and mosquito viromes; Shi et al, 2015; Shulman & Davidson, 2017) exhibited similar codon usage bias. These groups also formed tight clusters in multiple dimension scaling (MDS) or shared ecological/biogeographic characteristics with MCVs,

perhaps suggesting less divergence between groups after normalization due to reduced generation time or denoting similarity among adaptive pressures exerted upon these viruses. As hypothesized, if CRESS-DNA virus burst size is large, putative hosts prevalent, and environmental stability robust, it follows that similarities in selective pressure (rather than low frequency of infection and rate of viral turnover) may be responsible for commonalities in codon usage bias.

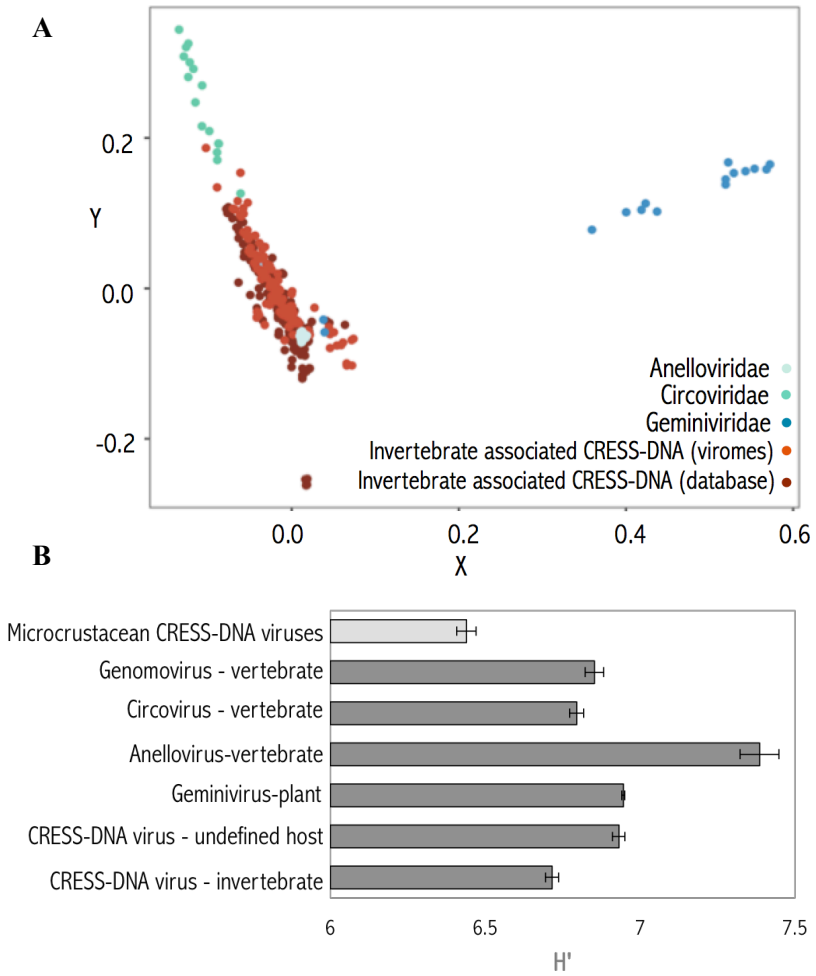

Figure 2.2 | *Non-alignment based classification of novel CRESS-DNA* – **(A)** MDS of feature frequency profiles (FFP: alignment-free sequence composition comparison utilizing 10nt k-mer) of MCV *rep* ORFs and ORFs from publicly derived genomes, illustrating that novel MCV sequence composition clusters with other invertebrate associated circoviruses. (B) oligomer α-diversity of microcrustacean *rep* (Shannon diversity index of FFP, average±SE) relative to *rep* from publicly derived genotypes, including CRESS-DNA virus-like viroplankton and terrestrial invertebrate-associated genomes.

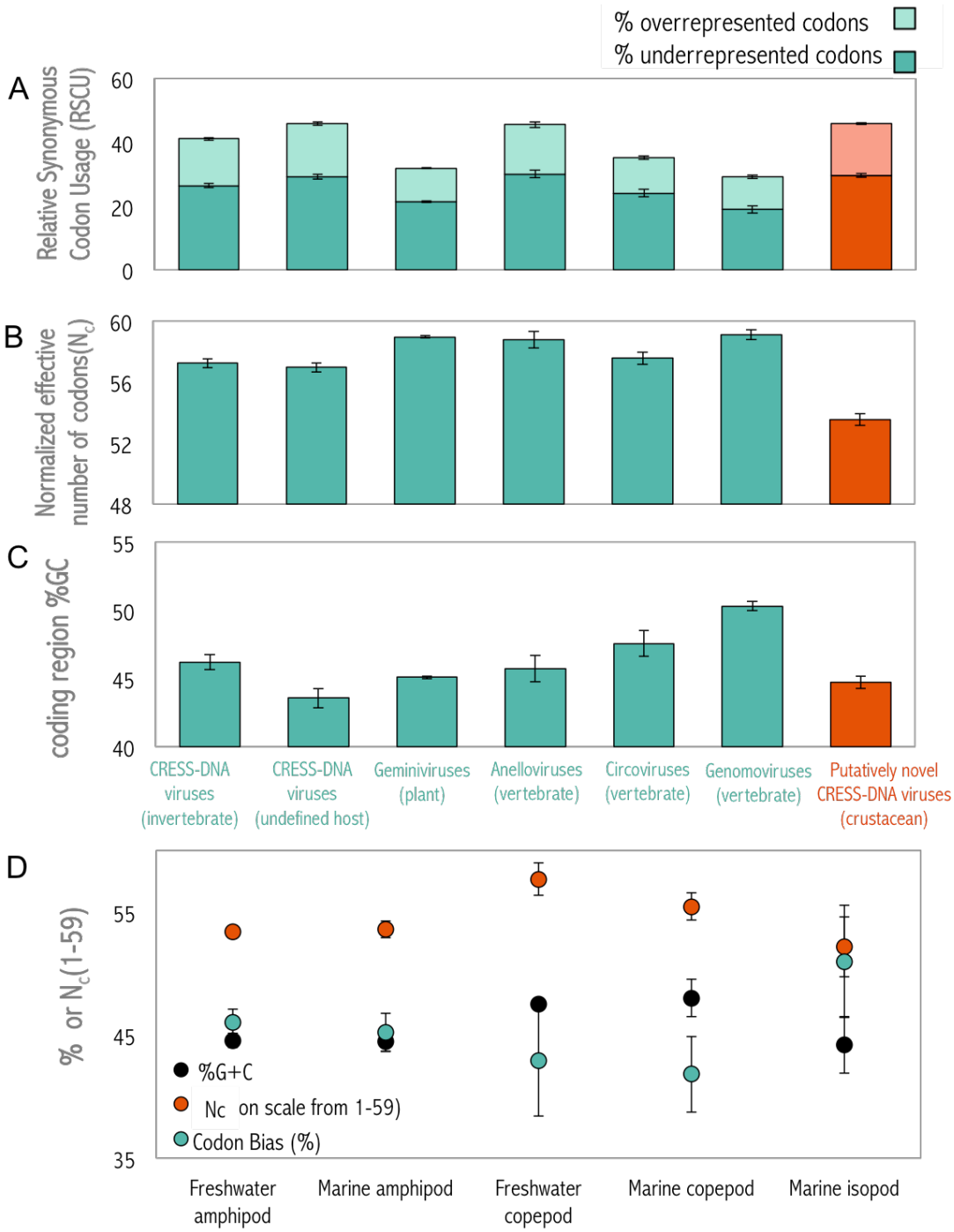

Figure 2.3 | *Codon usage patterns of novel CRESS-DNA viruses relative to previously classified CRESS-DNA viruses* – (A) Relative synonymous codon usage (RSCU) illustrating over and under-representation of utilized codons per *rep* ORF, (B) total effective number of codons (N_c) utilize to encode ORF, and (C) %GC bias in MCVs relative to CRESS-DNA viral genotypes in public databases (average±SE). (D) Measures of codon repertoire (%GC, N_c, or RSCU) parsed by putative crustacean host and habitat.

Codon usage biases in MCVs may be driven guanine/cytosine content (%GC), specifically in the frequency of codons terminating in these nucleotides due to less stringent binding of cognate tRNAs (%GC3). Net %GC content in MCV *rep* ORFs was significantly lower (%GC and %GC3; $p<.00001$, t-test) than viral groups associated with vertebrate or viroplankton hosts. Furthermore, %GC content of MCVs was most similar to that of viroplankton relative to terrestrial invertebrate genomes ($p>0.05$; Tukey). It was conceivable that characteristics specific to aquatic invertebrate hosts or their habitat may bias CRESS-DNA *reps* towards the use of synonymous codons terminating in adenine/thymine, preserving amino acid sequences while predisposing the ORF towards low %GC.

While the majority of MCV *rep* ORFs were shorter than ORFs from viroplankton or vertebrate viruses (average±SE; 235.5±6.4nt vs. 349.7±2.9nt), codon usage biases remained when normalized to length. These patterns in ORF codon usage biases and sequence similarity were most discernable among the malacostraca-associated CRESS-DNA viral contigs. Isopods and amphipods (both freshwater and marine) exhibited significantly lower N_c (average 53.4±0.4SE, $p<0.05$, t-test) and ORF-wide %GC (average 44.5±0.5SE, $p<0.05$, t-test) among MCVs. Congruently, marine and freshwater copepod-affiliated viruses exhibited lower codon usage biases (RSCU) relative to malacostraca-associated CRESS-DNA viruses (42.1%). While the strategy of codon usage bias categorization falls short in concretely or concisely differentiating between phylogenetic viral groups, it may suggest that adaptive and non-adaptive forces specific to orders of hosts may influence the composition of essential coding regions.

Endpoint comparison of microcrustacean-associated CRESS-DNA viruses (MCVs) may lend insight into selective pressures – Measures of %GC content, oligomer frequency, and codon usage biases, suggest that co-evolution with contemporary (or former) hosts and their ecosystems may bias MCVs towards specific genetic signatures. Therefore, comparing MCVs with both similar and contrasting ecosystems/hosts may guide broad inferences of viral ecology based on endpoint genomic composition. While network based analyses exhibited connectivity between MCVs and

viroplankton/vertebrate viruses, one-to-one pairwise identity over MCV *rep* ($\%ID_{rep}$) and average nucleotide identity (ANIb) may provide a quantitative measure within and between microcrustacean libraries. Most microcrustacean libraries contained multiple unique MCVs (72.73%), facilitating this comparison and indicating that there are conceivably more than one discrete genotypic population of virus affiliated with a given species or organism (e.g. coinfection, prior infection, or associated with diet/epibionts/other non-crustacean hosts). Indeed, on average, libraries contained 2.87 (±1.49SE) distinct MCV per thousand quality-controlled reads mapped to virus-like contigs), with the greatest normalized number of MCVs from amphipod associated libraries (average 17.27 MCVs $library^{-1}$) and libraries from freshwater ecosystems (16.13 MCVs $library^{-1}$). MCVs assembled within these same libraries (implicitly from the same organism, host species, collection site, ecosystem, and virome preparation conditions) exhibited the greatest pairwise identity over the length of the *rep* ORF (average 60.18%, $p>0.05$, Mann-Whitney U) and whole-contig similarity (average 36.08%ANIb, $p<0.001$ Mann-Whitney U) relative to MCVs assembled from different libraries (59.96% identity and 35.27% ANIb, respectively). This may reflect inherent biases in sequencing methods, or alternatively indicate that MCVs sharing host and biogeographic parameters share similar genomic characteristics.

To assess the correlation between host/biogeography variables and viral genomic composition, sequence resemblance was compared both within *rep* ORFs and across full contigs normalized to length. Predictably, there was considerably greater pairwise percent identity between conserved *rep* ORFs (average 59.99±0.0003%SE) in comparison to whole contigs (average nucleotide identity by tBLASTx; 35.36±0.0253%SE ANIb over contig; $p>0.05$; t-test). This provides further evidence of both (1) conservation of the CRESS-DNA viral replication mechanism and (2) hyper-variability of non-*rep*-encoding intergenic or structural ORF regions as a product of rapid accumulation of single nucleotide variants (SNVs) in circular ssDNA viruses (Firth et al, 2009; Rosario et al, 2012).

Affiliation with specific microcrustaceans may significantly influence MCV phylogenomics.

Putative microcrustacean hosts were clustered into macrotaxonomic group by order (e.g. Amphipoda, Isopoda) to serve as a broad proxy for phylogenetic context. There were significantly greater $\%ID_{rep}$ and contig-wide ANIb similarities within metataxonomic groups, exceeding that of species-specific comparisons ($p<0.001$, Kruskal-Wallis rank sum), supporting the possibility of host-driven viral selection. This trend continued with other macrotaxonomic groups, with amphipod-associated genotypes exhibiting greatest similarities to other amphipod-associated genotypes ($p<0.05$, Dunn), and isopod-associated genotypes exhibiting greatest similarities to other isopod-associated genotypes ($p<0.001$, Dunn). Greatest differences in genomic composition were observed between MCVs associated with malacostracans and copepods (which serve as a phylogenetic outgroup to the malacostraca), regardless of specific species, individuals, or libraries, ($p<0.001$, Wilcoxon rank sum with continuity correction), mirroring codon usage patterns. However, even when copepods are removed from statistical analysis, pairwise viral similarity remains higher when contigs share macrotaxonomic group or species.

Although possibly syllogistic, this observation may provide evidence that phylogenetic context (e.g. microcrustacean identity) influences both codon usage patterns and overall genomic composition of associated MCV. The $\%ID_{rep}$ and ANIb from MCVs affiliated with the same microcrustacean species, but with different sequencing libraries were marginally higher than those from different species ($p>0.05$, Wilcoxon rank sum). This is likely a reflection of low sampling depth of the same species across multiple sites (i.e. single/few libraries species^{-1}) and low taxonomic resolution between species in the Vericrustacea. Therefore, while the caveat of run biases or cross-contamination may convolute conclusions drawn from pairwise MCV comparison, observations of MCV similarity despite different sequencing runs support these inferences.

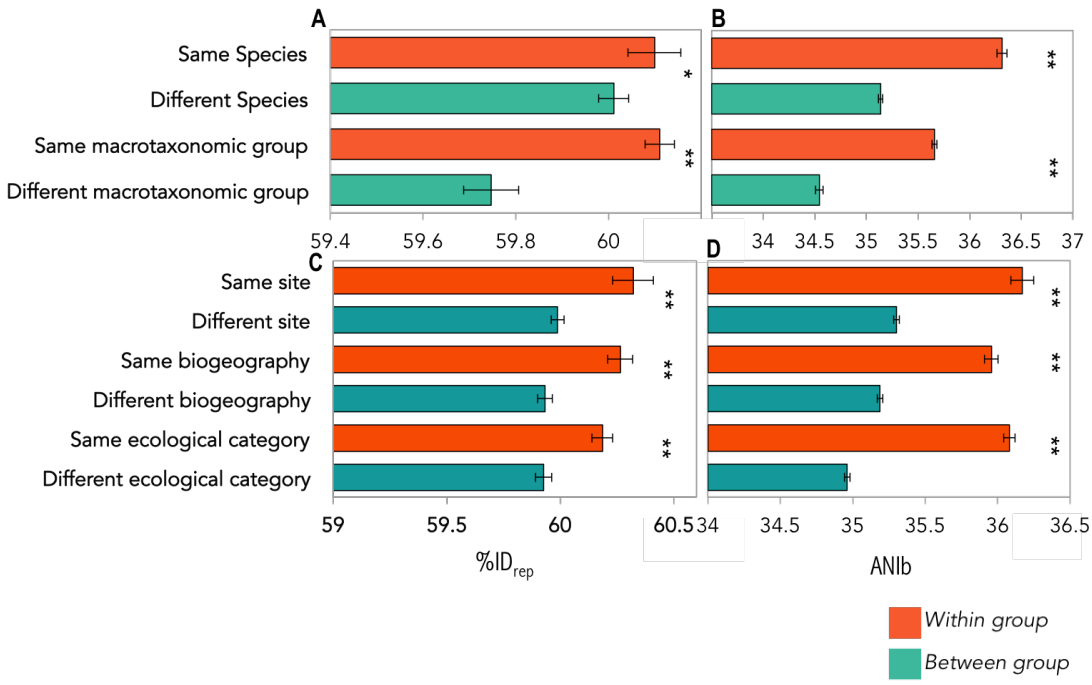

Figure 2.4 | *Within and between-group comparison of MCV genotypes with similar or dissimilar host attributes* - Top: comparison of *rep* ORF (%ID_{rep}; A) and contig similarity (ANIb; B) with sequences recovered from similar microcrustacean species and macrotaxonomic group relative to those from dissimilar species or metataxonomic group. Bottom: comparison of *rep* ORF (%ID_{rep}; C) and contig similarity (ANIb; D) with genotypes collected from the same sites (see Table 2.1), those collected from the same ecosystems (biogeography), and those collected from ecosystems sharing ecological characteristics (see SI Figure 2.2) relative to genotypes collected from dissimilar ecosystems. Average±SE*; multi-way ANOVA, * p<0.05, **p<0.01.* Despite remarkable geographic distances between collection sites, viral contigs derived from copepods exhibited the greatest average pairwise percent %IDrep (63.30%±1.48SE) and whole-contig ANIb (43.43%±2.57SE) when compared to MCVs derived from other copepods.

Affiliation with specific biogeographic conditions may significantly influence MCV phylogenomics. There was also a robust correlation between ecosystem-specific variables and MCV similarity. MCVs sharing comparable geographic or habitat-specific characteristics (ranging from salinity to symbiosis with *nereocystis* spp.) exhibited significantly greater %ID_{rep} or ANIb ($p<0.0001$ each; Wilcoxon rank sum and Welch's ANOVA), though this finding may be influenced by microcrustacean niche distribution or range (SI Figure 2.2; SI Figure 2.3). Despite substantial geographic distance, MCVs identified within similar ecosystems shared significantly greater pairwise genomic similarities (e.g. intertidal vs. pelagic $p<0.0001$, Welch's ANOVA; marine vs.

freshwater $p<0.05$, Wilcoxon rank sum, etc; Figure 2.4). Interestingly, environmental variables had a more apparent impact on average pairwise whole-contig ANIb than *rep* ORF identity, denoting that biogeographic or abiotic factors may be congruent with environment-specific selection in less conserved genomic regions, such as structural ORFs or intergenic sequences.

No single metadata factor adequately predicts MCV genomic composition. Regardless of taxonomic- or biogeographic-specific factors, $\%ID_{rep}$ did not consistently predict full-contig ANIb ($R^2<0.08$) suggesting that *rep* may be subject to different selection paradigms relative to surrounding genetic neighborhoods containing structural or other hypothetical ORFs. Greater pairwise identity of *rep* provides additional evidence of conservation (or convergent evolution) of *rep* across ecosystems and species. While there is a significant correlation between similarities in host/ecosystem variables and genomic content ($\%ID_{rep}$ and ANIb), suggesting non-neutral selection/evolution, no single biogeographic- or host-specific factor adequately predicts genotype similarity ($R^2<0.1$ log, exponential, or linear regressions; SI Figure 2.3). While putative host identity (metataxonomic group) may have a more substantial correlation with MCV genomic content ($p<0.001$, GLMM), interaction effects between virome characteristics suggest that multiple factors, such as host identity, ecosystem context, stochastic events, and other unknown variables likely govern viral composition.

2.4.iii. Diversity and biogeographic distribution of microcrustacean-associated CRESS-DNA viruses (MCVs): widespread, yet endemic genotypes

Not all MCVs are unique across microcrustacean viromes – Sequence comparisons demonstrate the diversity of MCVs that may exist within ecosystems, species, and individuals, suggesting a vast global diversity of CRESS-DNA viruses. Yet evidence of *rep* conservation (or convergent evolution) and whole-contig similarity (ANIb) from analogous hosts or similar ecosystems suggest that viral sequence space is potentially shared between metazoan hosts, or within communities/ecosystems, ultimately constraining biocomplexity. Several contigs sharing >95% *rep* ORF pairwise identity were collapsed into single genome bins, but did not share the same sequence

library (n=37), also conceivably indicating widespread distribution of similar genotypes. Of these "non-unique" MCVs from divergent libraries, 78.4% shared putative host macrotaxonomic group, 29.7% species, 21.6% haplotype, and 64.8% biogeographic traits and habitat characteristics. Furthermore, 91.1% of these MCVs sharing >95%ID_{rep} were shared among malacostracan libraries, again illustrating similarities between malacostracan-associated CRESS-DNA viral genotypes relative to copepod-associated viral genotypes. In this sense, the majority of these contigs likely share *rep* ORF similarities due to shared host/ecosystem variables, regardless of whole contig dissimilarities (average ANIb 91.8%±1.3SE).

Read cross-recruitment and potential for widespread MCV distribution may decrease extrapolated predictions of CRESS-DNA viral diversity – Remaining contigs that share *rep* similarity but do not share host/environmental characteristics may be indicative of regular exchange between discrete aquatic systems and widespread viral distribution. This hypothesis is corroborated by detection of similar invertebrate CRESS-DNA viruses across large geographic ranges. For example, Eaglesham and Hewson (2013) noted the widespread distribution of targeted CRESS-DNA viruses among net plankton from estuaries, coastal waters, and the open ocean. Other studies have described similar viruses among co-occurring ctenophores and copepods (Breitbart et al, 2015), and among multiple trophic level organisms within a temperate lake (Dayaram et al, 2016). Similar studies in terrestrial insects and vertebrate communities support observations of a polyphyletic phylogeny and identification of similar viral genotypes among a variety of sample sites. Therefore total CRESS-DNA viral diversity identified via viromics may be governed by co-occurrence or suprainfection and associated genetic exchange in addition to *rep* conservation or convergent evolution mediated by similarity in host/ecosystem.

Detection of similar MCVs from libraries sharing geographic proximity or niche overlap may further implicate gene flow as a driving influence on CRESS-DNA viral diversity. Therefore, these MCVs may exhibit distance decay-like patterns of compositional similarity (Nekola & White, 2004). To assess the potential distribution and sequence conservation of CRESS-DNA viral

genotypes within microcrustacean species sharing similar habitat/niche or phylogenetic similarity, quality controlled reads from independent libraries were cross-recruited to MCV *rep* ORFs. Overall, cross-library read recruitment confirms that CRESS-DNA viruses are widespread among aquatic arthropods. MCVs cross-recruited a total of >2.6 million reads after quality control (average $618nt^{-1}$ coverage, though likely overrepresented due amplification biases). Rapid evolutionary rates and the accumulation of variants per sequence among ssDNA viruses may impede read recruitment, reflected in heterogeneity of read mapping along consensus sequences. Despite these variables, due to commonalities in sequence similarity, contigs often recruit reads from other libraries, with 47.0% of contigs recruiting reads beyond the library of origin.

Possible evidence of host specificity, despite widespread distribution – While recent discoveries and library cross-recruitment may indicate global presence of CRESS-DNA viral genotypes, we propose that the relationship between distribution and microcrustacean infection remains unclear. This inference is further substantiated by highly skewed read recruitment, with the majority of CRESS-DNA virome reads associating with a single "dominant" MCV. Therefore, while a diversity of MCVs may be identifiable in a microcrustacean host, it is possible that the majority may represent transient, remnant, or slow/non-replicative members. Read recruitment may serve as an initial proxy for viral co-occurrence, with several obvious caveats, including the randomized overamplification of circular ssDNA templates. Nevertheless, ranking the standardized abundance of MCV-recruiting reads by order of magnitude allows a computational foothold to evaluate the potential for replication (infection). On average, the MCV genotype with the greatest coverage recruited nearly half (47.0%) of all MCV-like reads within that library and exhibited $1.16x\text{-}9.0x10^5$ (average $7.99\ x10^4$) greater coverage than the genotype with the lowest coverage (SI Figure 2.4). This intra-library heterogeneity in read recruitment may be symptomatic of amplification biases or suggest that only a few MCVs are specific to that organism, while others target different (including non-microcrustacean) hosts. While microcrustacean viromes harbor a range of genotypes, it is possible that they target affiliated non-arthropod unicellular hosts (e.g. "hitchhiking" epibionts,

protist parasites, etc) or represent past infections (residual infections).

Such specificity has been demonstrated by quantifying recurrent viral genotypes associated with lacustrine amphipods by qPCR (Bistolas et al, 2017a). For example, while one viral genotype (LM29173) was detected among all populations of amphipod via viromics, it was only detected in meaningful copy numbers in certain populations. Genotypes differentiated between amphipods on a haplotype-specific level, despite a plethora of potential horizontal transmission routes between lake ecosystems. Thus, these genotypes exhibited both habitat- and host-specific endemism, perhaps alluding to rapid selection that was not captured by single timepoint sampling or non-specific priming and sequencing. Similarly, a crucivirus identified within these libraries exhibited a narrow temporal and geographic range of detection among the isopod *Idotea wosnesenskii* (one sampling season, one site; Bistolas et al, 2017b). These investigations also mirror narrow host ranges observed among emergent vertebrate-associated CRESS-DNA viruses. These observations support the hypothesis that aquatic invertebrate CRESS-DNA viruses may be endemic to specific geographic ranges, with active infection limited by parameters including host species/haplotype range. Therefore, members of viral assemblages may not intimately interact with host cell receptors or metabolism and the widespread distribution of these genotypes are likely mediated by membership within an ecosystem and the density or diversity of coexisting metazoans.

Studies indicate that vertebrate-associated CRESS-DNA viruses are capable of retaining their genetic integrity in environmental reservoirs, including within hosts and coexisting genera and as free virus-like particles (VLPs; Allan et al, 1994). Both porcine circovirus 2 (PCV2) and human Torque Teno viruses are resistant to temperatures and antiseptics that would destroy cellular organisms, enabling their durability in biotic and abiotic reservoirs (Allan et al, 1994; Bendinelli et al, 2001). Therefore, it follows that CRESS-DNA viruses in aquatic ecosystems are likely capable of the widespread distribution observed among MCVs. However, while CRESS-DNA viruses are prevalent and may be detected among a broad range of metazoans and environmental samples, it remains possible that viral replication is limited to a small percentage of genotypes and putative

hosts. Consequently, the structure of CRESS-DNA viral consortia is likely influenced by more variables than host phylogeography. Evidence of endemism (Bistolas et al, 2017a; Jackson et al, 2015), transience (Bistolas et al, 2017b), seasonal recurrence (Hewson et al, 2013; Whon et al, 2012), and non-microcrustacean infections (Bistolas et al, 2017b; Dayaram et al, 2016) perhaps support this hypothesis. Based on this host specificity, we may ostensibly extrapolate an immense global diversity of CRESS-DNA viral taxa. However, the widespread detection of similar CRESS-DNA viral genotypes indicates a finite scope of viral diversity, constrained to the limits of CRESS-DNA viral adaptive potential.

2.4.iv. Microcrustacean ssDNA viruses exhibit population-wide microdiversity

Read recruitment indicates that novel MCVs, like many RNA viruses, likely evolve in conjunction with a milieu of non-consensus sequences ("quasispecies") that alter the fitness landscape through competition between genotype variants. 158 MCVs contained 3,409 unique sites of variation above threshold detection (1.01×10^{-2} sites nt^{-1}), calculated through high-quality short reads recruited to a consensus contig. Therefore, the putatively novel MCVs identified in this study likely represent the highest frequency consensus sequence among reads harboring potential variants. This degree of variation may indicate that MCVs share adaptive properties with well-characterized vertebrate CRESS-DNA viruses, which exhibit rapid mutation and recombination rates (Duffy et al, 2008; Lefeuvre et al, 2009). For example, deep sequencing of vertebrate circoviruses indicate that CRESS-DNA virus nucleotide substitution rates (10^{-3} to 10^{-4} substitutions $site^{-1}$ $year^{-1}$) are comparable to those of RNA viruses (10^{-2} to 10^{-5} substitutions $site^{-1}$ $year^{-1}$; Jenkins et al, 2002; Sanjuán et al, 2010), likely due to ssDNA lability. Despite implications that these CRESS-DNA viruses evolve rapidly, most analyses focus on surveillance of consensus sequences in relation to host taxa, but may overlook the presence of complex co-occurring microdiverse populations of variant sequences which may play significant roles in virus fitness, evolution, or pathogenicity.

Single nucleotide variants exist in high density among CRESS-DNA populations – Single

nucleotide variants (SNVs) were the most commonly identified deviation from consensus sequences, with an overall prevalence of 75.84%, representing a continuous spectrum of sequence conservation (probability >96.3%; Figure 2.5a). Isopod associated viral genotypes contained the greatest quantity of SNVs ($8.7x10^{-7}$ SNVs reads^{-1} reference length^{-1}) relative to other metataxonomic groups, and MCVs derived from lacustrine viromes contained significantly greater prevalence of SNVs relative to MCVs from other ecosystems. However, the average frequency of SNVs (percent sequences within population with detected variant) was greatest in amphipods (average 11.58%) and marine ecosystems (13.93%), potentially indicating low frequency SNVs are likely not maintained among original isopod-associated or lacustrine viral populations (SI Figure 2.5). Overall, SNVs were detected at low frequencies (11.32%±0.36SE per site), but above thresholds for random SNV incorporation due to sequencing error within quality controlled reads.

In addition to harboring the greatest overall frequency of SNVs ($p<0.001$; Welch's ANOVA), amphipod-associated CRESS-DNA viruses also exhibited the greatest cumulative raw quantity of variants (n=3188). Beyond SNVs, other mechanisms, such recombination, reassortment, duplications, frameshifts, and gene transfers are also likely responsible for the recently observed richness of CRESS-DNA virus clades. Multiple nucleotide variants (MNVs, average 2.39nt, range 2-9nt), where more than one nucleotide is exchanged but produce no frameshift were commonly observed (average frequency 14.27%±1.07SE), followed by deletions (8.19%±1.08SE), insertions (8.09%±1.12SE), and replacements (7.06%±2.45SE). Interestingly, while 6.6x less prevalent relative to SNVs, MNVs exhibited greater frequencies per variant site, potentially indicating that these MNVs may be more likely to persist and comprise a greater proportion of the viral "population" (SI Figure 2.5b). Alternatively, this greater frequency may indicate that these sites may be more mutable relative to the surrounding sequences without fitness cost to transmission, infection, or propagation.

Putative host phylogeny and biogeography may play a role in the accumulation of sequence variants – Phylogenetically distant microcrustacean species with niche overlap shared more similar MCV variant frequencies relative to phylogenetically similar species that were biogeographically

distant, potentially indicating that collection site variables may play a role in driving variant accumulation. Likewise, the average frequency of variants also differs significantly between marine and freshwater genotypes ($p<0.05$, Mann–Whitney U),

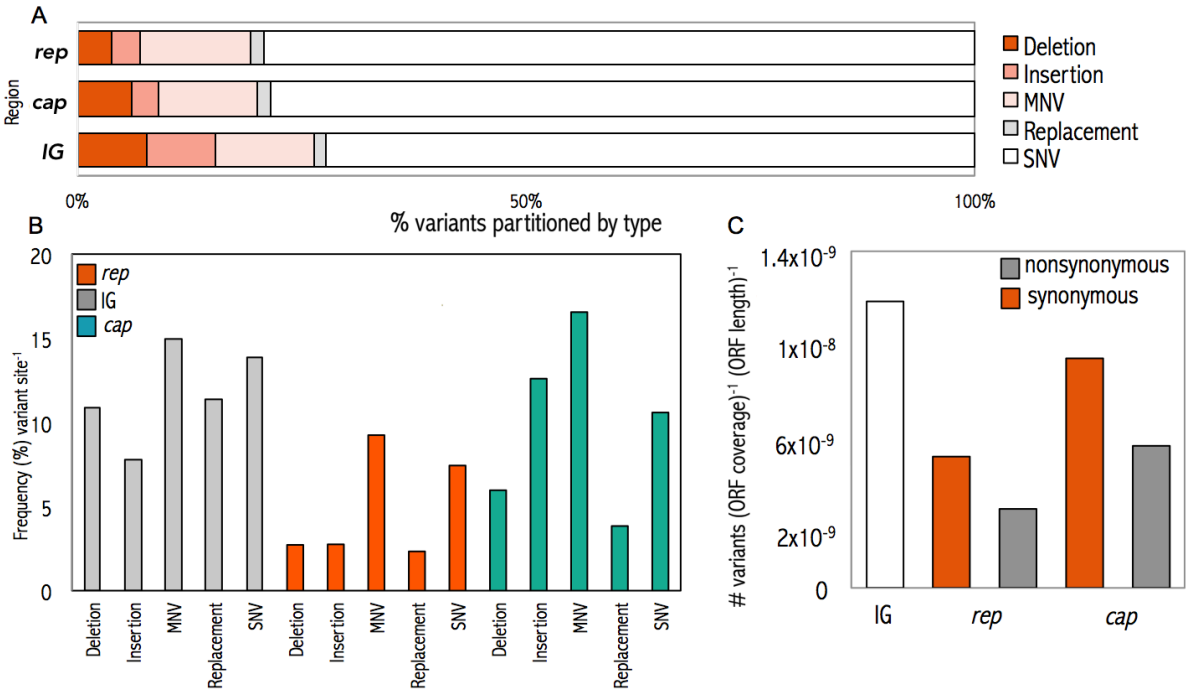

Figure 2.5 | *MCVs exhibit nucleotide variability within viromes, suggesting genotypes exist as microdiverse populations* - (A) Total proportion of variants (prevalence), partitioned by type within coding regions cap and rep, and noncoding intergenic region (IG), illustrating the majority of variants are single nucleotide variants. SNV: single nucleotide variant, MNV: multiple nucleotide variant. (B) Frequency of variant per region (rep, cap, or intergenic region) and type, indicating percentage of sequences carrying the variant within sequence "population." (C) Ratio of variants potentially eliciting synonymous or nonsynonymous changes at the amino acid level within coding regions.

collection site ($p<0.001$, Welch's ANOVA), and general site classification (intertidal/pelagic/benthic lacustrine; $p<0.001$, Welch's ANOVA). One genotype, CRESS-DNA contig LM29173, was identified among the same amphipod species haplotype within the Laurentian Great Lakes (distribution originally reported in Bistolas et al, 2017a; SI Figure 2.6). Interestingly, the total composition, richness, and frequency of sequence variants varied significantly between separate, but adjacent host populations within each lake. Furthermore, relative to the consensus sequence of LM29173 derived from Lake Michigan amphipods, variants from Lake Huron amphipods exhibited significantly lower frequencies (average 49.92%) than Lake Erie amphipods (average 92.58%; $p<0.001$ Mann–Whitney U). Of note, Lakes Michigan and Huron are connected via the Straits of Mackinac and often regarded as the same hydrological system, whereas Lake Erie is insulated via directional flow of the St. Clair and Detroit Rivers. The high incidence of variant/selective differences within this genotype between lakes may indicate the absence of overwhelming gene flow and sequential allopatric divergence due to a degree of isolation between lakes over time (inter-population variability), potentially contributing to observed CRESS-DNA viral diversity. This observation suggests that a collection of variables, including abiotic conditions, host physiology, genetics, and density likely determine the cumulative composition of variants (genotype richness, evenness, genetic similarity, etc) within microcrustacean populations.

Density and frequency of genomic variation differs in predicted protein-encoding sequences (ORFs) – Variant distribution was not uniform. Greater densities of variable sites were typically observed in non-protein coding regions, or intergenic sequences (>150%; Figure 2.5b). In alignment with this observation, 73.34% of SNVs, 72.03% of MNVs, and 83.63% of deletions, 80.60% of insertions, and 73.39% of replacements (normalized to read depth/length) were identified in regions outside of the *rep* ORF, indicative of low levels of sequence conservation. The average frequency of deletions (10.89%±1.68SE; probability 99.12%) within the intergenic region outnumbered the frequency of insertions (7.76%±1.15SE; probability 99.99%), which may lend insight into the minimal genome size of many arthropod CRESS-DNA viruses, as less-essential sequences may be

lost over time to facilitate genome replication.

In microcrustacean viruses, the putatively immunogenic *cap* ORF, when detected, was less conserved relative to *rep* ORFs, and the prevalence of non-synonymous variants exceeded that of synonymous variants (Ka/Ks=1.51). However, there was a greater frequency of synonymous variants, suggesting that variation in the *cap* ORF may not be well maintained within a population (5.20% reads carrying variant at non-synonymous site versus 10.28% of reads carrying variant at synonymous site). As a putatively structural protein, non-synonymous variation in *cap* may have explicit fitness consequences for receptor recognition and host infection. While single mutations in the capsid protein of mammalian CRESS-DNA viruses leads to viral particles with distinct antigenic properties (Allemandou et al, 2011), little is known about the role of these changes in the emergence of viral genotypes in arthropods. Conversely, *rep* ORFs exhibited significantly greater overall conservation compared to *cap*, yet only a marginally lower Ka/Ks ratio (1.42). As *rep* is an essential element in viral replication, we hypothesize that non-synonymous mutations may be immediately deleterious to the virus (i.e. prevent replication). Conversely, while variation in *cap* may exclude host recognition and replication, some mutations may not completely alter infection dynamics and allow propagation of some variant genotypes, regardless of reduced sequence conservation.

Potential functional consequences of nucleotide substitutions on translated sequences – In both ORFs, the frequency of synonymous variants exceeded the frequency of non-synonymous variants by >1.5fold (prevalence 62.68%), indicating that synonymous variants were likely selectively tolerated relative to non-synonymous variants (frequency: 13.20%±0.68SE vs. 7.22%±0.41SE, $p<0.001$, Kruskall-Wallis; Figure 2.5c). This observation further supports the proposed definition of steady-state mutation-selection balance of population genetics, indicating that, despite a degree of stochasticity in the introduction of variants, selection likely drives the establishment of a high-frequency consensus sequence. In both coding regions, the total number microcrustacean population-wide variants are likely enormously underestimated due to extinction of non-viable (i.e. not propagating/successfully infectious) genotypes and the frequency of remaining

variants may reflect a range of viral success. This differential “fitness” of genotypic variants within a viral population may influence epidemiology by facilitating cross-species transmission, accelerating emergence, or driving selection.

Among malacostracan-associated viral contigs, many synonymous substitutions represented maintained shifts in guanine(G)- or cytosine(C)-terminating codons to adenine (A)- or tyrosine(T)-terminating codons, indicating a continued progression towards greater %AT content, as observed by malacostracan viral codon usage patterns. Within all microcrustacean libraries, the overall ratio of variant sites with %GC gain, loss, and constancy remained stable (approximately 1:1:1, with marginally more sites with %GC gain in isopods by 0.022% and marginal loss in copepods by 0.075%; $p>0.5$, Kruskal-Wallis rank sum). However, a switch from AT-rich variants to GC-rich variants ("%GC gain") occurred at lower frequency in malacostracan-associated viral contigs relative to copepod-associated contigs (ratio of %GC loss: %GC gain 1.07:1 in amphipods, 1.59:1 in isopod viral contigs, and 0.60:1 in copepod, respectively). This may indicate that malacostracan viruses tend to lose variants that alter A or T into G or C at a greater frequency than the reverse.

This observation appears to be consistent across salinity (freshwater/marine) and ecosystem types (intertidal, lake, pelagic, etc), and may be largely driven by maintenance (greater observed frequency) of AT-rich replacements or MNVs among malacostracan viruses and higher frequency of copepod SNVs switching AT to GC, particularly in intergenic regions (1.24:0.88:1 average ratio of GC gain: loss: constancy among malacostracan sequence variants in intergenic region versus 1.58:3.39:1 ratio of GC gain: loss: constancy in copepod contig intergenic regions).

Variants signified switching between same nitrogenous base types (47.74%), rather than a transformation from pyrimidine to purine (25.95%) or vice versa (27.78%). This was particularly evident among synonymous variants, where same base-type variants comprised 62.52% of variant sites. However, purine to pyrimidine base shifts were more frequent (retained or proliferated) across both coding regions (1.30x more frequent in *rep*; 1.08x more frequent in *cap* in comparison to noncoding regions) and among both nonsynonymous and synonymous (1.23x more frequent in in

synonymous and 1.26x more frequent in nonsynonymous sites) relative to pyrimidine to purine shifts. This trend is not observed in intergenic regions, where pyrimidine to purine and purine to pyrimidine shifts appear to share equivalent frequencies in site occurrence. Therefore, coding regions may be biased towards directional maintenance resulting in base-type accumulation in ssDNA molecules.

As ambisense genomes, preamplification, and high-throughput sequencing obscures strandedness. Therefore, purine to pyrimidine variants do not indicate a clear directional bias (pyrimidine to purine may be equally likely if the opposite strand is packaged), but may specify that coding regions undergo different selective pressures relative to intergenic regions resulting in disparities in pyrimidine/purine composition. Variation in percent purine content in coding regions was significantly lower (13.97%) relative to variation in percent purine content in noncoding regions (bimodal distribution, 22.15% variation; $p<0.001$; F-test, SI Figure 2.7), which may indicate directional selection. In dsDNA, these biases should be negligible due to Chargaff's rule. However, CRESS-DNA viruses characteristically package ssDNA, and bias in purine to pyrimidine composition may impact inter-host strand stability (Arenz et al, 2007; Svinarchuk 1995), intra-host immune evasion and infection dynamics (Cristillo et al, 2001), or indicate polynucleotide tracts (Wurtzer et al, 2006).

2.4.v. Endogenized viral elements provide paleovirological evidence of historical CRESS-DNA infection in microcrustaceans

Putatively novel MCV *rep*s were utilized as "bait" to identify 321 endogenized viral elements (EVE) in 23 crustacean genomes (19 unique species) derived from public databases (NCBI-Genome database, accessed 08/2018, SI Table 2.3). EVEs - virus-like sequences that integrate into eukaryotic genomes - may lend insight into previous host tropism or provide a snapshot of ancestral genotypes related to fast evolving extant viruses (Metegnier et al, 2015). Conservation of CRESS-DNA virus *rep* provides a basis to query metazoan genomes for similar sequences, which signify

integration events into a host germ line and continued vertical inheritance. To accomplish this integration, a virus encoding a *rep* ORF similar to an extant virus must actively infect a non-somatic cell of the putative host, confirming the functional host for that CRESS-DNA viral genotype.

Of detected EVEs, 131 (40.81%) were previously characterized. However, to our knowledge, the remaining 190 *rep*-like sequences were previously undetected and represent possible infection and integration events. These findings indicate that discovery of extant viruses may be relevant for the identification of EVEs, and therefore hold paleovirological significance. Furthermore, EVEs may provide a portrait of previous genotypes, as integrated sequences are subject to host-specific, rather than virus-specific evolutionary paradigms (Gilbert et al, 2014; Gilbert & Feschotte, 2010), and can serve as synapomorphies or resolve the history of viral infections (Belyi et al, 2010; Katzourakis & Gifford, 2010;Liu et al, 2011; Thézé et al, 2015). While potential computational misassemblies between extant viral genomes and host genomes may overestimate the quantity of unique EVEs, the use of MCVs associated with microcrustaceans of a different species as BLAST query "bait" may also underestimate this number.

On average, 61.99% of EVEs contained a frameshift and 30.22% contained a premature internal stop codon, suggesting they are degenerate, non-coding sequences. Remaining *rep*-like ORFs lacked nonsense mutations, potentially indicating exaption or recent endogenization. However, the mechanism of endogenization of non-retroviral ssDNA viruses in eukaryotic genomes remains unclear (Krupovic & Forterre, 2015). 32.4% of *rep*-like crustacean EVEs were within the vicinity of transposable elements (TEs), and 41.41% were within the vicinity of putative retrotransposable/retroviral elements (both within 2kb of *rep*-like ORFs), which may lend insight into site-specific integration. However, class II DNA transposable elements, namely helitrons, were also identified within 2kb of 54.67% of putative EVEs. Helitrons represent a group of TEs that encode a *rep*-like protein domain responsible for rolling circle replication via a homologous mechanism to CRESS-DNA viral replication (Kapitonov & Jurka, 2007). While these TEs encode a rolling circle catalytic domain, they typically utilize a superfamily I helicase, unlike the CRESS-

DNA virus *rep*-encoded superfamily 3 helicase (S3H), and did not co-annotate with putatively viral EVEs. Like previous observations (Dennis et al, 2018), the adjoining sequences flanking newly identified putative endogenized *rep* ORFs were not comparable, suggesting that these integration events did not occur consistently at specific sites.

As proposed as a mechanism for helitron acquisition, this may be achieved via host endonuclease capture and integration of DNA - including that from infecting viral genotypes - at locations of eukaryotic dsDNA damage to facilitate repair, which results in non-site specific endogenization of *rep*-like sequences (filler DNA model Kapitonov & Jurka, 2007). Alternatively, as CRESS-DNA viral *rep* ORFs encode prerequisite endonuclease, NTPase, and DNA binding domains required for integration, *rep* may autonomously facilitate its own integration and excision and rely on host polymerase for passive proliferation, mimicking the activity of other RCR mobile genetic elements (notably those also encoding HUH domains Chandler et al, 2013; Kapitonov & Jurka, 2007). Collectively, these events provide additional evidence of active infection of a crustacean cell by ancestral CRESS-DNA viral genotypes and may allude to the origin of CRESS-DNA viruses or selfish helitron-like transposable elements.

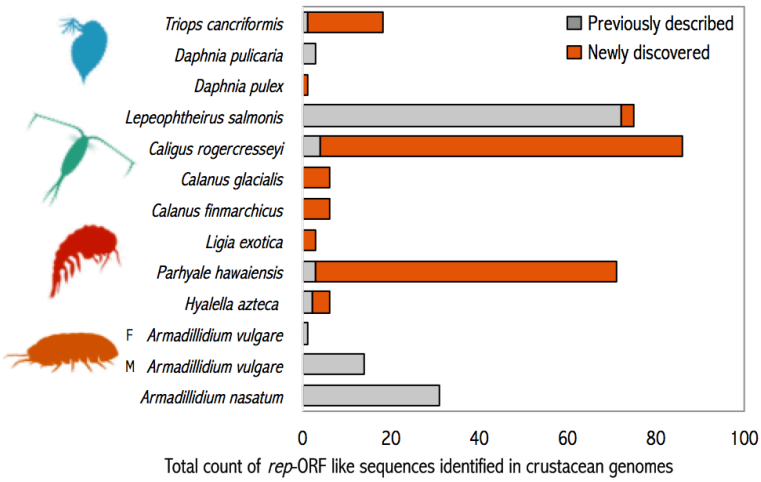

Figure 2.6 | *Putatively novel CRESS-DNA viruses may aid in the discovery of endogenized viral elements in crustaceans genomes* – Count of endogenized viral elements (EVEs) sharing sequence similarity by BLAST (e-value < 10^{-5}) to MCV (*rep*) within publicly available crustacean whole genome assemblies.

2.5 Conclusions **|** High-throughput sequencing confirms that ssDNA viruses are widespread and constitute a significant component of viral richness within the microcrustacean holobiont. CRESS-DNA viruses, in particular, appear to exhibit a massive range of viral diversity, as reflected by both alignment-based and alignment-free methods. Permutations of genome architecture, codon usage patterns, and variant accumulation far exceed the ability to taxonomically characterize novel genotypes within this clade.

This study leverages viromics to explore the utility of pairing host- and biogeography-specific characteristics with viral genotype composition as a computational endpoint strategy to determine which factors may influence the distribution and evolutionary characteristics of CRESS-DNA viruses. Limited codon repertoires illustrate that malacostracan-associated viruses contain low intragenomic α-diversity among MCV coding sequences, likely driven by limited %GC. Yet variant detection suggests that the emergence of novel genotypes is not rare and contributes to the overall diversity of CRESS-DNA viruses. Directional biases in variant accumulation and overall genetic heterogeneity between replication-associated ORFs (*rep)* indicate that subpopulations of these genotypes may also be microdiverse, with different selective pressures acting on *rep,* relative to structural proteins or intergenic regions. Additional exploration of ORF entropy may define the characteristics that constrain local and global spatial distribution and intra-host evolution.

Overall, this exploration proposes a scenario in which CRESS-DNA viruses may be specific to microcrustacean populations, with biogeography defined by host/ecosystem parameters, yet designates these metazoans as possible reservoirs for non-host specific genotypes. By capitalizing on evidence of integration events within microcrustacean genomes (endogenous viral elements), we may differentiate between viruses previously infecting microcrustaceans and spurious identification of unaffiliated viruses. This paleovirological record indicates that CRESS-DNA viruses, as a whole, infect microcrustaceans. However, microcrustaceans are often indiscriminate feeders, potentially capable of concentrating organic material, including virus-like particles (VLPs) and provide

substrates for chitin-bound epibionts. Therefore, while viromics lends insight into a range of viral genotypes and their widespread distribution, greater investigation of biogeographical distribution and host are necessary to determine the impact of these genotypes on specific organisms and populations.

Mesograzer-affiliated viruses continuously arbitrate ecosystem interactions, with impacts on nutrient availability and community composition. Therefore, differentiating between minimally pathogenic viruses and those that significantly alter behavior, mortality, and macronutrient cycling is critical to understanding both biotic and abiotic conditions in aquatic ecosystems. With genomic minimalism, rapid evolution, and ubiquity, microcrustacean-associated CRESS-DNA viruses may be well suited as models for metazoan ssDNA viral ecology.

2.6 *References*

Ågren J, Sundström A, Håfström T, Segerman B. 2012. Gegenees: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. *PLoS One*. doi:10.1371/journal.pone.0039107.

Allan GM, Phenix KV, Todd D, McNulty MS. 1994. Some biological and physico-chemical properties of porcine circovirus. *Zentralbl Veterinary med B*. 41(1):17-26. doi:10.1111/j.1439-0450.1994.tb00201.x.

Allemandou A, Grasland G, Hernandez-Nignol AC, Kéranflec'h A, Cariolet R, Jestin A. 2011. Modification of PCV-2 virulence by substitution of the genogroup motif of the capsid protein. *Vet Res*. 42:54. doi: 10.1186/1297-9716-42-54.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215:403–410. doi:10.1016/S0022-2836(05)80360-2.

Arenz C, Zeitz O. 2007. DNA Made of Purines Only. *Chem Biol*. 14(5):467-469. doi:10.1016/j.chembiol.2007.05.001.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 19(5): 455–477. doi:10.1089/cmb.2012.0021.

Bao W, Kojima, KK, Kohany O. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 6:11. doi:10.1186/s13100-015-0041-9.

Belyi VA, Levine AJ, Skalka AM. 2010. Sequences from ancestral single-stranded DNA viruses in vertebrate genomes: the parvoviridae and circoviridae are more than 40 to 50 million years old. *J Virol.* 84(23):12458–12462. doi:10.1128/JVI.01789-10.

Bendinelli M, Pistello M, Maggi F, Fornai C, Freer G, Vatteroni ML. Molecular properties, biology, and clinical implications of TT virus, a recently identified widespread infectious agent of humans. *Clin Microbiol Rev.* 14 (1) 98-113. doi:10.1128/CMR.14.1.98-113.2001.

Bistolas KSI, Besemer RM, Rudstam LG, Hewson I. 2017b. Distribution and inferred evolutionary characteristics of a chimeric ssDNA virus associated with intertidal marine isopods. *Viruses*. 9(12):361. doi:10.33909120361.

Bistolas KSI, Jackson EW, Watkins Jm, Rudstam LG, Hewson I. 2017a. Distribution of circular single-stranded DNA viruses associated with benthic amphipods of genus *Diporeia* in the Laurentian Great Lakes. *Fresh Biol*. 62(7):1220-1231. doi:10.1111/fwb.12938.

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F. 2002. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA*. 99(22):14250-5. doi:10.1073/pnas.202488399.

Breitbart M, Benner BE, Jernigan PE, et al. 2015 Discovery, prevalence, and persistence of novel circular single-stranded DNA viruses in the ctenophores *Mnemiopsis leidyi* and *Beroe ovata*. *Front Microbiol*. 6:1427. doi:10.3389/fmicb.2015.01427.

Brinkman NE, Villegas EN, Garland JL, Keeley SP. 2018. Reducing inherent biases introduced during DNA viral metagenome analyses of municipal wastewater. *PLoS ONE*. 13(4):e0195350. doi: 10.1371/journal.pone.0195350.

Chandler M, de la Cruz F, Dyda F, Hickman AB, Moncalian G, Ton-Hoang B. 2013. Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. *Nat Rev Microbiol*. 11(8):525-38. doi: 10.1038/nrmicro3067.

Cristillo AD, Mortimer JR, Barrette IH, Lillicrap TP, Forsdyke DR. 2001. Double-stranded RNA as a not-self alarm signal: to evade, most viruses purine-load their RNAs, but some (HTLV-1, Epstein-Barr) pyrimidine-load. *J Theor Biol.* 208(4):475-91. doi:10.1006/jtbi.2000.2233.

Dayaram A, Galatowitsch ML, Argüello-Astorga GR, van Bysterveldt K, Kraberger S, Stainton D, Harding JS, Roumagnac P, Martin DP, Lefeuvre P, Varsani A. 2016. Diverse circular replication-

associated protein encoding viruses circulating in invertebrates within a lake ecosystem. *Infect Genet Evol*. 39:304-16. doi:10.1016/j.meegid.2016.02.011.

Dean FB, Nelson JR, Giesler TL, Lasken RS. 2001. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res*. 11(6): 1095–1099. doi:10.1101/gr.180501.

Dennis TPW, de Souza WM, Marsile-Medun S, Singer JB, Wilson SJ, Gifford RJ. 2018. The evolution, distribution and diversity of endogenous circoviral elements in vertebrate genomes. *Virus Res*. 262:15-23. doi:10 .1016/j.virusres.2018.03.014.

Dole-Olivier MJ, Galassi DMP, Marmonier P, Creuzé Des Châtelliers M. 2000. The biology and ecology of lotic microcrustaceans. *Freshwater Biol*. doi:10.1046/j.1365-2427.2000.00590.x 47.

Duffy S, Shackelton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Gen.* 9:267-276. doi:10.1038/nrg2323

Dunlap DS, Ng TFF, Rosario K, Barbosa JG, Greco AM, Breitbart M, Hewson I. 2013. Molecular and microscopic evidence of viruses in marine copepods. *Proc Nat Acad Sci USA*. 110:1375-1380. doi:10.1073/pnas.1216595110.

Eaglesham JB, Hewson I. 2013. Widespread detection of circular replication initiator protein (*rep*)-encoding ssDNA viral genomes in estuarine, coastal and open ocean net plankton. *Mar Ecol Prog Ser.* 494:65-72. doi:10.3354/meps10575.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792– 1797. doi:10.1093/nar/gkh340.

Fahsbender E, Hewson I, Rosario K, Tuttle AK,Varsani A, Breitbart M. 2015. Discovery of a novel circular DNA virus in the Forbes sea star, *Asterias forbesi. Arch Virol*. 160(9):2349–2351. doi:10.1007/s00705-015-2503-2

Firth C, Charleston MA, Duffy S, Shapiro B, Holmes EC. 2009. Insights into the evolutionary history of an emerging livestock pathogen: porcine circovirus 2. *J Virol*. 83(24):12813–12821. doi:10.1128/JVI.01719-09.

Gilbert C, Feschotte C. 2010. Genomic fossils calibrate the long-term evolution of hepadnaviruses. *PLoS Biol*. 2010;8:e1000495. doi:10.1371/journal.pbio.100049.

Gilbert C, Meik JM, Dashevsky D, Card DC, Castoe TA, Schaack S. 2014. Endogenous

hepadnaviruses, bornaviruses and circoviruses in snakes. *Proc Biol Sci.* 281(1791):20141122. doi:10.1098/rspb.2014.1122.

Gismervik I. 1997. Stoichiometry of some marine planktonic crustaceans. *J Plankton Res*. 19(2): 279–285. doi:10.1093/plankt/19.2.279.

Hewson I, Chow C, Fuhrman JA. 2010. Ecological role of viruses in aquatic ecosystems. *eLS*. doi:10.1002/9780470015902.a0022546.

Hewson I, Eaglesham JB, Höök TO, LaBarre BA, Sepúlveda MS, Thompson PD, Watkins JM, Rudstam LG. 2013b. Investigation of viruses in *Diporeia* spp. from the Laurentian Great Lakes and Owasco Lake as potential stressors of declining populations. *J Great Lakes Res.* 39:499–506.10.1016/j.jglr.2013.06.006.

Hewson I, Ng G, Li W, LaBarre BA, Aguirre I, Barbosa JG, Breitbart M, Greco AW, Kearns CM, Looi A, Schaffner LR, Thompson PD, Hairston NG. 2013a. Metagenomic identification, seasonal dynamics, and potential transmission mechanisms of a Daphnia−associated single−stranded DNA virus in two temperate lakes. *Limnol Oceanogr*. 58:1605– 1620. doi:10.4319/lo.2013.58.5.1605.

Holmes E. 2007. Viral evolution in the genomic age. *PLoS*. doi:10.1371/journal.pbio.0050278.

Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AF, Wheeler TJ. 2016. The Dfam database of repetitive DNA families. *Nucleic Acids Res*. 44(1):D81-9. doi:10.1093/nar/gkv1272.

Jackson B, Varsani A, Holyoake C, Jakob-HoffIan R, Robertson I, McInnes K, Empson R, Gray R, Nakagawa, Warren K. 2015. Emerging infectious disease or evidence of endemicity? A multi-season study of beak and feather disease virus in wild red-crowned parakeets (*Cyanoramphus novaezelandiae*). *Arch Virol*. 160: 2283. doi:10.1007/s00705-015-2510-3.

Jackson EW, Bistolas KSI, Button JB, Hewson I. 2016. Novel circular single-stranded DNA viruses among an asteroid, echinoid and holothurian (phylum: echinodermata). *PLoS One*. 11(11): e0166093. doi:10.1371/journal.pone.0166093.

Jenkins GM, Rambaut A, Pybus OG, Holmes EC. 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J Mol Evol.* 54(2):156-65. doi:10.1007/s00239-001-0064-3.

Kapitonov VV, Jurka J. 2007. Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends Genet*. 23(10):521-9. doi: 10.1016/j.tig.2007.08.004 .

Kapoor A, Simmonds P, Lipkin WI, Zaidi S, Delwart E. 2010. Use of nucleotide composition analysis to infer hosts for three novel picorna-like viruses. *J Virol*. 84(19):10322-8. doi:10.1128/JVI.00601-10.

Katzourakis A, Gifford RJ. 2010. Endogenous viral elements in animal genomes. *PLoS Genet*. 6(11): e1001191. doi:10.1371/journal.pgen.1001191.

Kim KH, Chang HW, Nam YD, Roh SW, Kim MS, Sung Y, Jeon CO, Oh HM, Bae JW. 2008. Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Appl Environ Microbiol*. 74(19): 5975-5985. doi:10.1128/AEM.01275-08.

Kim Y, Aw TG, Teal TK, Rose JB. 2015. Metagenomic investigation of viral communities in ballast water. *Environ Sci Technol*. 49(14):8396–8407. doi: 10.1021/acs.est.5b01633.

Kraberger S, Argüello-Astorga GR, Greenfield LG, Galilee C, Law D, Martin DP, Varsani A. 2015. Characterization of a diverse range of circular replication-associated protein encoding DNA viruses recovered from a sewage treatment oxidation pond. *Infect Genet Evol*. 31:73-86. doi:10.1016/j.meegid.2015.01.001.

Krupovic M, Forterre P. 2015. Single-stranded DNA viruses employ a variety of mechanisms for integration into host genomes. *Ann NY Acad Sci.* 1341:41-53. doi:10.1111/nyas.12675.

Labonté JM, Suttle CA. 2013. Previously unknown and highly divergent ssDNA viruses populate the oceans. *ISME J*. 7: 2169–2177. doi:10.1038/ismej.2013.110.

Lefeuvre P, Lett JM, Varsani A, Martin DP. 2009. Widely conserved recombination patterns among single-stranded DNA viruses. *J Virol*. 83(6):2697-707. doi:10.1128/JVI.02152-08.

Lefort V, Longueville JE, Gascuel O. 2017. SMS: Smart model selection in PhyML. *Mol Biol Evol*. 34(9):2422-2424. doi:10.1093/molbev/msx149.

Liu H, Fu Y, Li B, Yu X, Xie J, Cheng J, Ghabrial SA, Li G, Yi X, Jiang D. 2011. Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. *BMC Evol Biol*. 11:276. doi:10.1186/1471-2148-11-276.

López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, Alcamí A. 2009. High diversity of the viral community from an Antarctic lake. *Science*. 326(5954):858-61. doi:10.1126/science.1179287.

Metegnier G, Becking T, Chebbi MA, Giraud I, Moumen B, Schaack S, Cordaux R, Gilbert C. 2015.

Comparative paleovirological analysis of crustaceans identifies multiple widespread viral groups. *Mob DNA*. 6:16. doi:10.1186/s13100-015-0047-3.

Muhire BM, Varsani A, Martin DP. 2014. SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. *PLoS One*. 9(9):e108277. doi:10.1371/journal.pone.0108277.

Neill W. 1975. Experimental studies of microcrustacean competition, community composition and efficiency of resource utilization. *Ecology*. 56(4):809–826.

Nekola JC, White PS. 2004. The distance decay of similarity in biogeography and ecology. J Biogeogr. 26(4):867-878 doi:10.1046/j.1365-2699.1999.00305.x.

Ng TFF, Willner D, Nilsson C, Lim YW, Schmieder R, Chau B, Ruan Y, Rohwer F, Breitbart M. 2010. Vector-based metagenomics for animal virus surveillance. *Int J Infect Dis*. 14:e378. doi:https://doi.org/10.1016/j.ijid.2010.02.461.

Paez-Espino, D, Chen IMA, Palaniappan K, Ratner A, Chu K, Szeto E, Pillay M, Huang J, Markowitz VM, Nielsen T, Huntemann M, Reddy TBK, Pavlopoulos GA, Sullivan MB, Campbell BJ, Chen F, McMahon K, Hallam SJ, Denef V, Cavicchioli R, Caffrey SM, Streit WR, Webster J, Handley KM, Salekdeh GH, Tsesmetzis N, Setubal JC, Pope PB, Liu WT, Rivers AR, Ivanova NN, Kyrpides NC. 2017. IMG/VR: a database of cultured and uncultured DNA viruses and retroviruses. *Nucleic Acids Res*. 45(1): D457–D465. doi:10.1093/nar/gkw1030.

Rohwer F, Thurber RV. 2009. Viruses manipulate the marine environment. *Nature*. 459:207–212. doi:10.1038/nature08060. 8.

Rosario K, Duffy S, Breitbart M. 2012. A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Arch Virol*. 157(10):1851-71. doi:10.1007/s00705-012-1391-y.

Rosario K, Mettel KA, Benner BE, Johnson R, Scott C, Yusseff-Vanegas SZ, Baker CCM, Cassill DL, Storer C, Varsani A, Breitbart M. 2018. Virus discovery in all three major lineages of terrestrial arthropods highlights the diversity of single-stranded DNA viruses associated with invertebrates. *PeerJ*. e5761. doi:10.7717/peerj.5761.

Rosario K, Nilsson C, Lim YW, Ruan Y, Breitbart M.2009. Metagenomic analysis of viruses in reclaimed water. *Environ Microbiol*. 11(11):2806-20. doi:10.1111/j.1462-2920.2009.01964.x.

Rosario, K, Schenck RO, Harbeitner RC, Lawler SN, Breitbart M. 2015. Novel circular single-stranded DNA viruses identified in marine invertebrates reveal high sequence diversity and

consistent predicted intrinsic disorder patterns within putative structural proteins. *Front Microbiol*. 6:696. doi:10.3389/fmicb.2015.00696.

Roux S, Hallam SJ, Woyke T, Sullivan MB. 2015. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife*. 4:e08490. doi:10.7554/eLife.08490.

Roux S, Krupovic M, Debroas D, Forterre P, Enault F. 2013. Uncontaminated viromes reveal the abundance and diversity of metabolism genes in environmental viruses. *Open Biology*.

Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. 2010. Viral mutation rates. *J Virol.* 84(19):9733-9748. doi:10.1128/JVI.00694-10.

Shackelton LA, Parrish CR, Holmes EC. 2006. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. *J Mol Evol*. 62(5):551-63. doi:10.1007/s00239-005-0221-1.

Shi C, Liu Y, Hu X, Xiong J, Zhang B, Yuan Z. 2015. A metagenomic survey of viral abundance and diversity in mosquitoes from Hubei province. *PLoS One*. 10(6):e0129845. doi:10.1371/journal.pone.0129845.

Shulman LM & Davidson I. 2017. Viruses with circular single-stranded DNA genomes are everywhere! *Ann Rev Virol.* 4:159-180. doi:10.1146/annurev-virology-101416-041953.

Sievers A, Wenz F, Hausmann M, Hildenbrand G. 2018. Conservation of k-mer composition and correlation contribution between introns and intergenic regions of animalia genomes. *Genes*. 9(10): 482. doi:10.3390/genes9100482.

Simmonds P, Adams MJ, Benko M, Breitbart M, Brister R, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B, Hull R, King AMQ, Koonin EV, Krupovic M, Kuhn JH, Lefkowitz EJ, Nibert ML, Orton R, Roossinck MJ, Sabanadzovic S, Sullivan MB, Suttle CA, Tesh RB, van der Vlugt RA, Varsani A, Zerbini FM. 2017. Virus taxonomy in the age of metagenomics. *Nature Rev Microbiol.* 15(3):161-168. doi:10.1038/nrmicro.2016.177.

Sims GE, Jun SR, Wu GA, Kim SH. 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc Natl Acad Sci USA*. 106(8):2677-82. doi:10.1073/pnas.0813249106.

Sims GE, Kim SH. 2011. Whole-genome phylogeny of *Escherichia coli/Shigella* group by feature frequency profiles (FFPs). *Proc Natl Acad Sci USA.* 108:8329–34. doi:10.1073/pnas.1105168108.

Soffer N, Brandt ME, Correa AM, Smith TB, Thurber RV. 2014. Potential role of viruses in white plague coral disease. *ISME J*. 8(2):271-83. doi:10.1038/ismej.2013.137.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, d'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez G, Poulain J, Poulos BT, Royo-Llonch M, Sarmento H, Vieira-Silva S, Dimier C, Picheral M, Searson S, Kandels-Lewis S; Tara Oceans coordinators, Bowler C, de Vargas C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB, Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG, Bork P. 2015. Ocean plankton. Structure and function of the global ocean microbiome. *Science*. 348(6237):1261359. doi:10.1126/science.1261359.

Suttle CA.2005. Viruses in the sea. *Nature*. 437:356-361. doi:10.1038/nature04160.

Svinarchuk F, Monnot M, Merle A, Malvy C, Fermandjian S. 1995. The high stability of the triple helices formed between short purine oligonucleotides and SIV/HIV-2 vpx genes is determined by the targeted DNA structure. *Nucleic Acids Res*. 23(19):3831-6. doi:10.1093/nar/23.19.3831.

Tamaki H, Zhang R, Angly FE, Nakamura S, Hong PY, Yasunaga T, Kamagata Y, Liu WT. 2012. Metagenomic analysis of DNA viruses in a wastewater treatment plant in tropical climate. *Environ Microbiol*. 14(2):441-52. doi: 10.1111/j.1462-2920.2011.02630.x.

Thézé J, Leclercq S, Moumen B, Cordaux R, Gilbert C. Remarkable diversity of endogenous viruses in a crustacean genome. *Genome Biol Evol*. 6(8):2129-40. doi:10.1093/gbe/evu163.

Vega Thurber R, Haynes M, Breitbart M, Wegley L, Rohwer F. 2009. Laboratory procedures to generate viral metagenomes. *Nature Protoc*. 4: 470-83. doi:10.1038/nprot.2009.10.

Whon TW, Kim MS, Roh SW, Shin NR, Lee HW, Bae JW. 2012. Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere. *J Virol*. 86(15):8221-8231. doi:10.1128/JVI.00293-12.

Wommack KE, Colwell RR. 2000. Virioplankton: viruses in aquatic ecosystems. *Microbiol Mol Biol Rev*. 64(1):69-114. doi:10.1128/mmbr.64.1.69-114.2000.

Wurtzer S, Goubard A, Mammano F, Saragosti S, Lecossier D, Hance AJ, Clavel F. 2006. Functional central polypurine tract provides downstream protection of the human immunodeficiency virus type 1 genome from editing by APOBEC3G and APOBEC3B. *J Virol.* 80(7):3679–3683. doi:10.1128/JVI.80.7.3679-3683.2006.

Yoshida M, Takaki Y, Eitoku M, Nunoura T, Takai K. 2013. Metagenomic analysis of viral communities in (hado)pelagic sediments. *PLoS One.* 8(2):e57271. doi:10.1371/journal.pone.0057271.

Zawar-Reza P, Arguello-Astorga GR, Kraberger S, Julian L, Stainton D, Broady PA, Varsani A. 2014. Diverse small circular single-stranded DNA viruses identified in a freshwater pond on the McMurdo Ice Shelf (Antarctica). *Infect Genet Evol*. 26:132–138. doi:10.1016/j.meegid.2014.05.018.

CHAPTER 3

DISTRIBUTION OF CIRCULAR SINGLE-STRANDED DNA VIRUSES ASSOCIATED WITH BENTHIC AMPHIPODS OF GENUS *DIPOREIA* FROM THE LAURENTIAN GREAT LAKES[1]

3.1 Abstract **|** Benthic amphipods of genus *Diporeia* are key indicators of ecosystem health and were the most abundant invertebrate in the four deep Laurentian Great Lakes until the mid-1990s. The mechanism(s) responsible for declines in *Diporeia* abundance in Lakes Huron, Michigan and Ontario remain contentious. A previous study identified a circular replication initiator protein-encoding single-stranded (CRESS) DNA virus, LM29173, associated with *Diporeia* populations. This study surveyed the distribution of CRESS-DNA viral genotypes in amphipods from several populations and investigated the relationship between viral presence and nutritional quality of the amphipod. We illustrated that CRESS-DNA virus-like sequences are common among *Diporeia* viromes, and are detectable as free particles in virioplankton. Among three identified CRESS-DNA viral genotypes (LM29173, LM122 and LH481), LM29173 is a recurrent and prevalent constituent of Lake Michigan and Lake Huron *Diporeia*-associated viral consortia, with genotype load averaging 726–3,643 copies mg^{-1} of tissue. LM29173 abundance coincides with amphipod haplotype demographics, and was more prevalent among *Diporeia* from the southern clade (Lakes Michigan, Huron, New York Finger Lakes) than northern clade (Lake Superior), irrespective of the state of population decline. We expected that the association of CRESS-DNA viruses would impart costs to amphipod physiology. However, viral load of LM29173 did not correlate with amphipod nutritional quality (lipid content, C:N stoichiometry).

---

[1]Presented with minor amendment from the original published article:
Bistolas KSI, Jackson EW, Watkins JM, Rudstam LG, Hewson I. 2017. Distribution of circular single-stranded DNA viruses associated with benthic amphipods of genus *Diporeia* in the Laurentian Great Lakes. *Freshwater Biology* 62:1220-1231. DOI: 10.1111/fwb.12938

[1]Supplementary material may be accessed at: https://doi.org/10.1111/fwb.12938

3.2 Introduction **|** Since the early 1990s, benthic invertebrate communities in the Laurentian Great Lakes have undergone rapid and comprehensive restructuring with the introduction of invasive dreissenid bivalves and concurrent decline in historically dominant burrowing amphipods of genus *Diporeia* (Barbiero et al., 2011; Nalepa, Fanslow, & Lang, 2009). *Diporeia* utilize detrital material, contribute to the greater pool of dissolved organic matter and serve as key food resources for several fish species in the Great Lakes (Fitzgerald & Gardner, 1993; Gardner, Nalepa, Frez, Cichocki, & Landrum, 1985; Guiguer & Barton, 2002; Halfon, Schito, & Ulanowicz, 1996). While these organisms have been previously observed at densities exceeding 14,000/m^2 (60–80% of profundal benthic biomass, Cook & Johnson, 1974), annual monitoring programs depict progressive and severe decline in *Diporeia* populations (up to 90%) at shallow (<50m) and mid-depth (30–70m) contours in four of the five Great Lakes, excluding Lake Superior (Auer, Auer, Urban, & Auer, 2013; Barbiero et al., 2011; Birkett, Lozano, & Rudstam, 2015). These changes in benthic community composition have been linked to alterations in lake-wide biogeochemical cycling and may be responsible for variation in condition, growth and maturation of commercially and ecologically important fish species, including lake whitefish (*Coregonus clupeaformis*; Barbiero, Rockwell, Warren, & Tuchman, 2006; Pothoven, Nalepa, Schneeberger, & Brandt, 2001; Rennie, Sprules, & Johnson, 2009).

Despite their ecological significance, the causative factor(s) responsible for the decline in *Diporeia* remain contentious. While population decline largely coincides with the increase in exotic filter-feeding mussels (*Dreissena spp*.), a direct causative link implicating the invasive species in *Diporeia* decline has not been well established (Nalepa et al., 2006; Watkins, Rudstam, Mills, & Teece, 2012; Watkins et al., 2007). For example, decline in some Lake Ontario and Lake Michigan amphipod populations preceded dreissenid incursion (Nalepa et al., 2009; Vanderploeg, Liebig, Nalepa, Fahnenstiel, & Pothoven, 2010; Watkins et al., 2007), and there is evidence that mussels and amphipods are capable of successful coexistence (i.e. Finger Lakes benthic communities in central New York; Watkins et al., 2012). Hypotheses attributing changes in *Diporeia* abundance to niche

displacement and bivalve-imposed food limitation (either through the interception of detrital deposits or mitigation of calcite-dependent phytoplankton blooms) have been refuted both *in situ* and in mesocosm trials (Dermott, Bonnell, & Jarvis, 2005; Nalepa, Fanslow, & Foley, 2005; Nalepa et al., 2006; Ryan, Sepulveda, Nalepa, & Höök, 2012; Watkins et al., 2012). *Diporeia* population densities are also poorly associated with variation in lake productivity, benthic flux of organic carbon and fish predation (Watkins et al., 2012). Eukaryotic parasites (gregarines, helminthes, Ciliophora, Haplosporidia, Microsporidia, etc.) and bacterial groups, including members with 16s rRNA sequence similarity to known pathogens, have been identified in association with *Diporeia* hosts (Messick, Overstreet, Nalepa, & Tyler, 2004; Winters, Fitzgerald, Brenden, Nalepa, & Faisal, 2014; Winters, Marsh, Brenden, & Faisal, 2015). However, patterns of microbial pathogen detection do not suggest their contribution to variation in *Diporeia* densities among sites (Winters et al., 2014, 2015). Another potential driver of decline is viruses (Hewson et al., 2013a). Here, we focus on DNA viruses associated with *Diporeia*.

As exemplified in a 2013 study (Hewson et al., 2013a), viral metagenomics (metaviromics) is an effective platform for characterizing the composition and diversity of viral taxa associated with *Diporeia*. Hewson et al. (2013a) identified several circular ssDNA virus-like sequences uniquely associated with *Diporeia* from Lake Michigan, which were not prevalent in lakes with stable amphipod populations (Lake Superior, Owasco Lake). One putative viral genotype, LM29173, shares sequence similarity and genome architecture with Type V circular replication initiator protein-encoding single-stranded DNA (CRESS-DNA) viruses identified in aquatic environments (e.g. reclaimed water-associated circovirus-like genome RW-E; Rosario, Duffy, & Breitbart, 2009). Like other Type V CRESS-DNA genomes, LM29173 is unisense, contains a canonical nonanucleotide motif (NANTATTAC) origin of replication ('ori') and a *rep*-encoding open reading frame (ORF) on the positive-sense strand (Hewson et al., 2013a; Rosario, Duffy, & Breitbart, 2012). This genotype represents a prevalent constituent of 2008–2009 amphipod viromes, as it was detected in ~90% of individuals evaluated from Lake Michigan (Hewson et al., 2013a). This study continues investigation

of CRESS-DNA viruses in contemporary amphipod populations by sampling from Lake Michigan, Lake Superior and Owasco Lake (Hewson et al., 2013a). To expand on collection efforts in 2006–2011, we also targeted amphipods from declining Lake Huron populations and two additional populations (Cayuga and Seneca Lakes). Our 2014 collections increase sample sizes by a factor of 3–16 per original sample site to assess statistical relevance of viral distributions.

While putative CRESS-DNA viruses have been associated with aquatic crustaceans, including copepods (Dunlap et al., 2013), cladocerans (Hewson et al., 2013a) and decapods (Gudenkauf & Hewson, 2016; Rosario, Schenck, Harbeitner, Lawler, & Breitbart, 2015), and have been linked to fluxes in invertebrate population densities (Hewson et al., 2013a), the degree to which they alter invertebrate ecological fitness remains unclear. To date, it remains unknown if invertebrate-associated CRESS-DNA viruses precipitate disease or mortality in metazoans. This study aims to resolve the distribution of several amphipod-associated CRESS-DNA virus genotypes between distinct but adjacent *Diporeia* populations, which vary in biogeographic range, depth and co-occurrence with invasive dreissenids, among other factors. In addition, because *Diporeia* serve as energetic resources for upper trophic-level organisms (e.g. slimy sculpin *Cottus cognatus*, lake whitefish *Coregonus clupeiformis*, alewife *Alosa pseudoharengus*, etc.; Wells, 1980), we seek to determine if association with putative CRESS-DNA viruses impact lipid content and elemental composition of the amphipod, potentially reducing the nutritional quality of this prey species.

3.3 Methods **|** *Viral discovery* - Amphipods from Lakes Superior, Michigan, Huron and Erie were collected in August–September, 2014, at depths ranging from 8.5 to 167.6m at EPA-designated stations (U.S. Environmental Protection Agency, 2009). Finger Lakes (Seneca Lake, Owasco Lake, Cayuga Lake, New York, USA, Fig S3.1) populations were sampled at 25–80m, within 2km of recreational boat launches. Specimens were collected using a Ponar benthic sampler, sieved (500μm) to remove sediment, rinsed in virus-free water to isolate animals from environmental viruses and individually frozen at 80°C.

Metaviromic libraries were prepared from 10 animals from each lake. Specimens were homogenized with 2.0-mm Bead Beater tubes (10min; Zymo Research, Irvine, CA, USA) and viral particles were purified per Ng et al. (2010; Fig S3.2). Briefly, homogenates were syringe filtered (0.2µm PES; VWR International, Radnor, PA, USA), precipitated with 10% PEG-8000 by weight overnight at 4°C, centrifuged at 15,000xg to remove the supernatant and resuspended in 1ml 0.02µm filtered nuclease-free H_2O. Suspensions were treated with 200µl $CHCl_3$. Aqueous fractions were enzymatically digested with nucleases (2.5U DNase I, 0.25U RNase and 1U Benzonase) at 37°C for 3hr to eliminate non-encapsidated nucleic acids, then amended with EDTA to stop digestion. Aliquots of purified virus were extracted using the ZR viral extraction kit (Zymo Research) and isothermally amplified through Genomiphi Whole Genome Amplification (GE Healthcare, Little Chalfont, UK) per manufacturer protocols. This amplification method preferentially amplifies small, circular ssDNA genomes, and allows for enrichment of CRESS-DNA viral genomes. Amplification was confirmed by Pico Green incorporation and gel visualization. Samples were fragmented (400–600bp) and prepared by Truseq Nano kit prior to 2x250 bp sequencing via Illumina MiSeq (Cornell University Core Laboratories Center, Ithaca, NY, USA).

Bioinformatic identification of CRESS-DNA viruses - Sequences were trimmed for read quality, ambiguous nucleotides, read length and Illumina adapters/barcodes, then assembled *de novo* on CLC Genomics Workbench into contiguous sequences ("contigs"; ver.8.5.1; Qiagen, Hilden, Germany: mismatch cost 2, insertion cost 3, deletion cost 3, length fraction 0.80, similarity 0.80, minimum contig length 500nt, vote for non-specific matches). Viral discovery efforts have previously revealed virus-like sequences associated with high-throughput sequencing pipelines (Naccache et al., 2013). Therefore, it was essential to establish a negative control to distinguish which sequences were artifacts attributable to sequencing preparations. Contigs were compared to libraries generated via parallel high-throughput sequencing of a virus-free, nuclease-free water template (BLASTn; Altschul, Gish, Miller, Myers, & Lipman, 1990), and contigs with significant sequence similarity to control library reads (e-value $<1x10^{-5}$) were discarded from downstream

analyses.

Remaining contigs were compared against the NCBI non-redundant (nr) database (BLASTx, e-value <$1x10^{-5}$; Altschul et al., 1990). CRESS-DNA virus-like contigs were assessed for completeness (circularization), read coverage (50% similarity over 80% length) and presence of a canonical ori (NANTATTAC) within a covalently closed stem loop structure (Mfold Web Server; Zuker, 2003). ORFs within CRESS-DNA virus-like contigs were demarcated using GetORF and annotated via BLASTx (e-value <$1x10^{-5}$). Putative replication initiator (*rep*) ORFs were additionally assessed for the presence of rolling circle replication (RCR) and SF3 viral helicase motifs common among eukaryotic CRESS-DNA viruses (Rosario et al., 2012). Predicted amino acid sequences of translated ORFs were evaluated on the Kyte-Doolittle hydrophobicity scale to lend preliminary insight into peptide composition and structure (CLC Workbench ver.8.5.1; Qiagen). Metavirome data are available in GenBank under bioproject PRJNA344369.

Prevalence & load of viral genotypes - Nucleic acids from individual amphipods from the Great Lakes (n=199) and Finger Lakes (n=35) were extracted in randomized sets via the ZR96-Tissue & Insect DNA kit (Zymo Research). Virioplankton samples (280–492ml) were collected at depths ranging from 1.8 to 165.4m and pre-filtered (0.22µm). Viral sized particles were captured on a 0.02µm Anotop filter and eluted by incubating/vortexing filters with 800µl ZR Viral DNA buffer (Zymo Research), then backwashing viral particles into a collection tube. Viral DNA was extracted using ZR Viral-DNA Kit (Zymo Research).

Prevalence and load of viral genotype LM29173 were assessed using quantitative PCR (qPCR) using TaqMan reagents (TaqMan Universal Master Mix II, no UNG; Applied Biosystems, Foster City, CA, USA) and employing pre-established primers/probes designed within the LM29173-*rep* ORF (Hewson et al., 2013a; Fig. S3.3). Two additional genotypes were identified via viromic analyses (LH481, Lake Huron *Diporeia*, accession no. KX982251; LM122; Lake Michigan *Diporeia*, accession no. KX982252) and primers/probes were designed using Primer3 (Rozen & Skaletsky, 2000) to target associated *rep* ORFs. Duplexed reactions targeting LM122 and LH481

genotypes were optimized for reaction efficiency (standards run independently exhibited equivalent reaction efficiency and limits of quantification as when run in duplex), dynamic range (standards R^2>0.98), primer/probe concentration and annealing temperature to ensure non-biased quantification. Reactions utilized TaqMan Multiplex Master Mix (Applied Biosystems). Amplification was performed on a StepOnePlus™ Real-Time PCR system (Applied Biosystems) and thermocycling and reaction parameters are detailed in Fig S3.3. Duplicate eight-fold standard dilutions were included in each run of duplicate samples.

DNA-RNA dual extractions (ZR-Duet DNA-RNA Miniprep kit; Zymo Research) were utilized to assess viral transcription. Extracted RNA was aliquoted, and either reverse transcribed (RT) via Superscript III (Invitrogen, Carlsbad, CA, USA, per manufacturer instructions) or subjected to an identical reaction without reverse transcriptase as a no-RT control. Samples were considered positive when congruent no-RT controls were negative. Each sample was then evaluated via duplicate qPCR reactions with congruent, duplicate no-RT controls (Fig S3.3).

Cycle threshold (Ct) and quantity were determined via StepOnePlus (ver.2.3; Applied Biosystems) software and valid runs were defined by reaction efficiency (>94%) and standard regression linearity (R^2 >0.98). The lower limits of quantification for LM122 and LH481 were calculated per reaction and Ct values averaged 31.4 and 33.2 (equivalent to 20.7 and 29.7 genome copies), respectively. Quantities were corrected for extraction volume and standardized by animal wet weight. Samples were considered positive when technical replicates were both positive and Ct standard deviation was <0.5 (StepOnePlus ver.2.3; Applied Biosystems). Prevalence was defined as the number of amphipods within a site with positive viral detection. Load was defined as mean copy number of viral genotype among positive samples.

Amphipod haplotype demographics - Whole animals were extracted (Tissue-Insect Extraction Kit; Zymo Research) and mitochondrial cytochrome c oxidase subunit I (COI) sequences were amplified using primers LCO1490 and HCO2198 (Folmer, Black, Hoeh, Lutz, & Vrijenhoek, 1994). Thermocycling conditions included: initial denaturation at 94°C for 2min, 35 cycles of 30s at

94°C, 60s at 50°C and 60s at 72°C each, and final extension for 6min at 72°C. Products were gel-purified (Zymo Research), cloned (pGEM-T vector; Promega, Madison, WI, USA) and used to transform JM109 competent *Escherichia coli* (Invitrogen). Plasmid inserts were extracted via Zyppy™ Plasmid Miniprep Kit (Zymo Research) and Sanger sequenced (Cornell University Core Laboratories Center, Ithaca, NY, USA).

Sanger sequences and those curated from Genbank (McCalla, 2010; Pilgrim, Scharold, Darling, & Kelly, 2009; Usjak, 2010) were globally aligned in MUSCLE with default parameters (Edgar, 2004). The optimal model for tree construction (GTR + G + T) was determined by hLRT, BIC and AIC model selection consensus and used to generate a maximum likelihood tree with 1,000 bootstraps and topical configurations of internal branches defined via NJ (CLC workbench ver.8.5.1; Qiagen). Additional COI sequences were deposited into Genbank (accession no. KY042235–KY042242).

Assessment of amphipod nutritional quality – After desiccation at 60°C for 48hr, animals were pooled within sites, such that each sample was >2 mg (minimum biomass for downstream analysis). Lipids were extracted via a modification in Bligh and Dyer (1959). Briefly, dried samples were homogenized with micropestles, amended with 100µl chloroform:methanol (2:1), vortexed for 30s, settled for 30min, then centrifuged for 10min at 1,400xg. Supernatant was decanted and the process was repeated two more times. Samples were then re-dried at 60°C for 48hr to remove remaining solvent and weighed. Percent lipid was calculated as change in dry weight divided by initial dry weight. Remaining samples were combusted using a NC2500 elemental analyzer (Carlo Erba, Italy) to determine carbon and nitrogen composition at the Cornell University Stable Isotope Laboratory (COIL, Ithaca, NY, USA). Internal standards (BCBG, 0.17% N and 0.43% C; Methionine standard 0.10% N and 1.98% C, every 10 samples) confirmed accuracy and precision of the instrument.

***3.4 Results** | I. Discovery of CRESS-DNA Viruses Associated with Diporeia* - High-throughput

sequencing efforts generated 1.6–3.2 million reads per amphipod virome. On average, reads assembled into 1,804 contigs with a mean length of 1,658nt (Table 3.1). The majority of contigs (76–88%) could not be annotated or were associated with cellular organisms, including the amphipod genome, congruent with previous viromic analyses (Hewson et al., 2013a). An additional 1.8–9.6% of contigs were identified as artifacts generated by the next-generation sequencing pipeline, sharing high sequence similarity to a control virome prepared using 0.02μm-filtered nuclease-free H_2O as a template.

CRESS-DNA viruses represented a common metazoan-associated viral group shared between libraries, collectively assembling the greatest number of contigs (117 of 322 total contigs, Fig S3.4) and recruiting the greatest number of reads (i.e. recruiting >1.5 million reads, representing >149-fold more reads than all other viral groups combined excluding phage, Fig S3.4). The high relative proportion of small, circular, ssDNA virus contigs in amphipod metaviromes is largely a product of rolling circle amplification employing Φ29 polymerase (Kim et al., 2008; Roux et al., 2016). Preferential amplification of CRESS-DNA viruses biases assessments of viral community composition and bars comparison to other aquatic viromes (i.e. virioplankton). However, these biases are advantageous in the discovery and detection of novel CRESS-DNA virus sequences.

Among CRESS-DNA viral contigs, 29% shared sequence similarity to contigs assembled from viromes specific to other lakes. 39% of CRESS-DNA virus-like contigs contained a canonical ori (NANTATTAC). Two ORFs that shared sequence similarity to CRESS-DNA non-structural *rep* and structural *cap* sequences (BLASTx, e-value $<10^{-5}$) were identified in 17.1% of CRESS-DNA virus-like contigs. Two putative CRESS-DNA virus genotypes recovered from virome analysis represent novel (<70% sequence similarity to known genotypes) sequences and share sequence similarity with environmental CRESS-DNA sequences from aquatic environments (Fig 3.1; marine hadopelagic sediments, Yoshida, Takaki, Eitoku, Nunoura, & Takai, 2013; and freshwater ponds Zawar-Reza et al., 2014; BLASTx, e-value $<10^{-5}$, Fig 3.2), and were further investigated for features characteristic of CRESS-DNA viruses (Fig S3.2).

	Finger Lakes *Diporeia*	Lake Superior *Diporeia*	Lake Michigan *Diporeia*	Lake Huron *Diporeia*	Lake Erie *Echinogammarus*	Total
Total number of reads	3,247,374	1,627,416	1,664,514	2,350,106	2,889,726	11,779,136
Reads after trimming	3,055,195	1,530,578	1,513,034	2,244,766	2,730,661	11,074,234
% assembled into contigs	94.65	92.97	91.09	95.481	96.42	94.12 (Mean)
Mean contig length (nt)	1,502	1,914	1,378	1,749	1,745	1,657.6 (Mean)
Total contigs	1,671	1,830	2,264	2,200	1,056	9,021
Annotated contigs	875	1,682	2,034	2,132	804	7,527
Contigs ID with host genome	2	9	1	4	3	19
Contigs ID with laboratory contaminants	32	175	183	139	20	549
Contigs ID with cellular contaminants	659	1,230	1,589	1,847	679	6,004
Contigs associated with viruses	182	268	261	142	102	955
Reads recruited to contigs associated with metazoan viruses	102,341	16,611	225,867	13,182	247,207	605,208

Table 3.1 |*Summary of viral metagenome (metavirome) assembly and annotation*. Metaviromes were constructed from 10 amphipods per lake and subjected to 2 × 250 bp paired-end Illumina Miseq sequencing, resulting in 1.6–3.2 million reads per amphipod metavirome. Over 9,000 contiguous sequences (contigs) were assembled de novo in CLC workbench (v.8.5.1, de Bruijn) with an average length of 1,658 nt (metric for success of assembly). Contigs were annotated by BLASTx (e-value $<1 \times 10^{-5}$) to identify sequences relevant to viral consortia (n = 955). Note that libraries were pre-amplified via Φ29 polymerase and therefore biased towards circular ssDNA genomes.

LM122 (1955nt) was identified in Lake Michigan libraries (1179 average read depth per base) and represents a complete circular molecule encoding an ori, CAGTATTAC, within a closed 32nt stem loop (ΔG = 8.78). This putative Type I (ambisense) CRESS-DNA virus contains a non-structural *rep* ORF (262aa) with evidence of characteristic rolling circle replication motifs (residues characteristic of RCR motif 1 and RCR motif 3) and SF3 viral helicase domains (residues characteristic of a Walker-A motif and Motif C), in addition to a putative *cap* ORF (257aa) arranged in an ambisense orientation (Fig 3.2b). LH481 (2657nt) is a Type VI (unisense, with an antisense-oriented Rep relative to the ori) genome from Lake Huron *Diporeia* (229 average read depth per base) encoding two antisense ORFs: a 481aa *cap* ORF, and a 283aa *rep* ORF with residues consistent with a Walker-A motif, a Walker-B motif and a partial Motif C region. The ori (TATTATTAC) is enclosed in a 37nt stem loop (ΔG = 0.54, Figure 2c). Both novel genomes contain two ORFs: one ORF with significant sequence similarity to *rep* of other invertebrate or aquatic CRESS-DNA viruses, and a second (putative *cap*), which shared little similarity to other viral sequences. These findings corroborate the conclusion that *rep* ORFs are generally more conserved than *cap* ORFs. New putative *cap* ORFs lacked similarity to known *cap* ORFs, and did not contain a series of basic amino acids in the N-terminal region, as noted in other eukaryotic CRESS-DNA virus *cap* ORFs. Therefore, these ORFs are assumed to encode structural capsid proteins, but additional expression-based analyses are required to confirm their function. Genotypes were classified based on genome architecture per Rosario et al. (2012).

LM29173, a Type V CRESS-DNA viral genotype identified in a 2013 survey of *Diporeia* viromes from Lake Michigan (animals collected in 2008–2009, Hewson et al., 2013a), was present in 2014 Lake Michigan metaviromes and exhibited the greatest degree of read coverage (12,139 average depth per base) among the three targeted genotypes (Fig 3.2a). This sequence recruited 8.8% of CRESS-DNA virus affiliated reads, predominantly from Lake Michigan and Lake Huron *Diporeia* libraries.

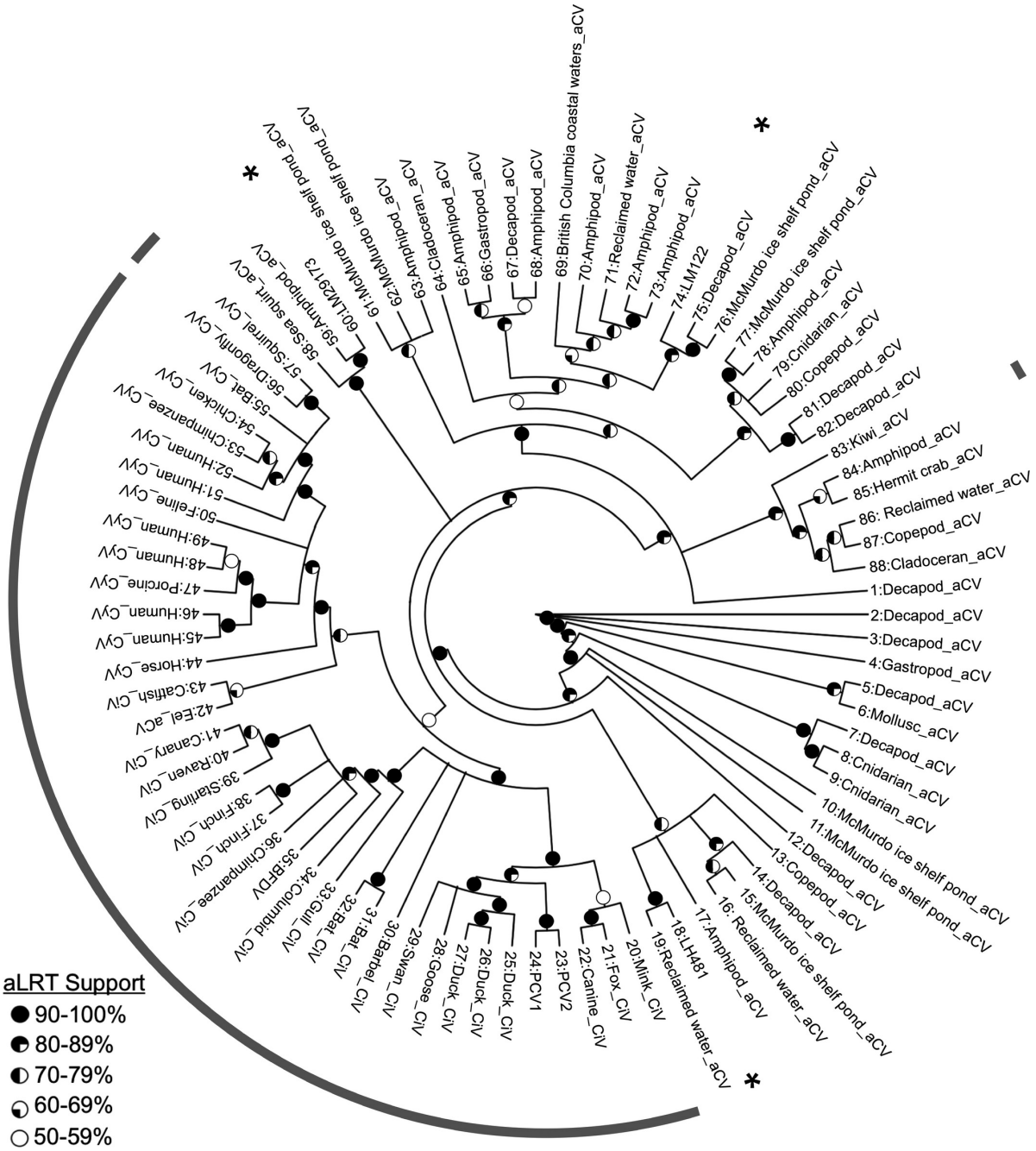

Figure 3.1 | *Phylogenetic context* - Maximum likelihood phylogeny of CRESS-DNA virus *rep* ORFs (LG + G, ~260AA) describing a subset of organisms/aquatic environments in which CRESS-DNA viruses have been identified (i.e. representative viral genotypes identified in environmental aquatic viromes or genotypes associated with avian, mammalian or invertebrate putative hosts). Terminal nodes indicate putative host, with specific information detailed in Figure S8. This phylogeny includes viral genomes of genus circovirus (CiV) and genus cyclovirus (CyV), in addition to unclassified CRESS-DNA viruses identified via metaviromics (aCV, associated circular viruses). Internal node symbols indicate approximate likelihood ratio test (aLRT) branch support (Anisimova & Gascuel, 2006). Grey bar indicates vertebrate hosts; asterisks indicate target CRESS-DNA genotypes relevant to this study.

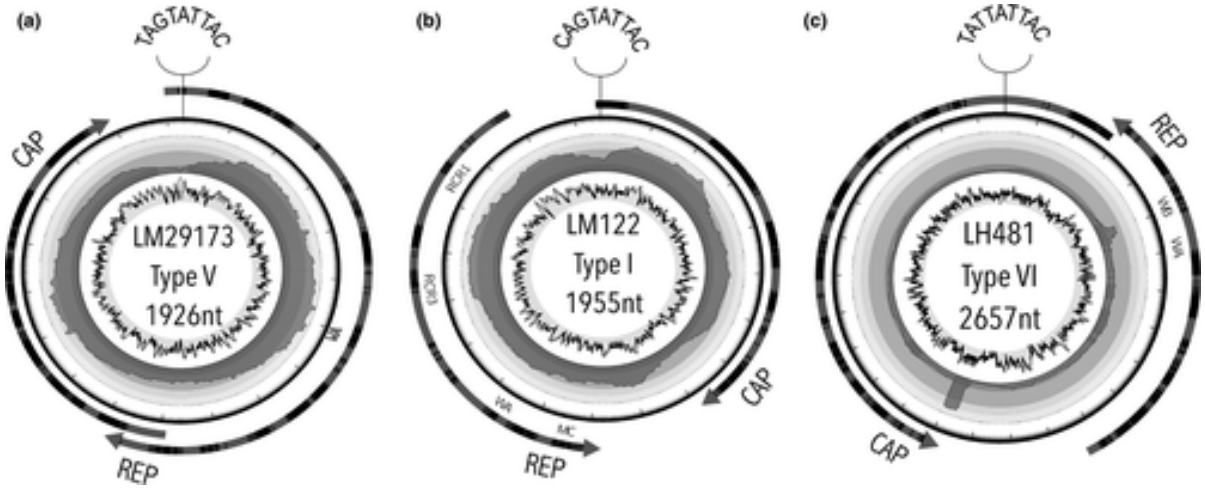

Figure 3.2 | *Characterization of target CRESS-DNA virus-like genotypes* – (a) Lake Michigan Diporeia-derived genotype LM29173, (b) Lake Michigan *Diporeia*-derived genotype LM122, and (c) Lake Huron Diporeia-derived genotype LH481. From inside to outside, tracks describe %GC content (grey indicates >%GC), read coverage and ORFs (RCR 1–3 = Rolling Circle Motif 1–3, WA = Walker A, MC = Motif C, per Rosario et al., 2012; shaded bars indicate predicted residue hydrophobicity per Kyte and Doolittle, 1982, where darker colour represents greater hydrophobicity). Stem loop structures describe putative origins of replication (ori, NANTATTAC). LM29173 genome derived from Hewson et al. (2013a).

II. CRESS-DNA virus genotype distribution – LM29173 prevalence and average load in Lake Michigan (97%, 3,643 copies mg^{-1}) significantly exceeded those in Lakes Huron (58%, 1,116 copies mg^{-1}, Games-Howell, $p < .01$, Ruxton & Beauchamp, 2008) and Superior (28%, 726 copies mg^{-1}, Games-Howell, $p < .001$, Ruxton & Beauchamp, 2008), corresponding to populations that have experienced the greatest proportional decline (prevalence: $\chi^2_{(2,n=199)} = 67.01$, $p < 2x10^{-25}$; load: Welch's ANOVA, $F_{2,70.8} = 7.86$, $p = .0008$; Fig 3.3). While LM122 and LH481 were detected in over 50% of sampled Great Lakes *Diporeia* (63%, average prevalence of LM122; 71%, average prevalence of LH481, Fig 3.3a), average load was negligible (97 and 165 copies mg^{-1}, respectively; Kruskal–Wallis Rank Sum with Dunn's Test per lake, $p < .001$; Fig 3.3b). Wet weight of individual amphipods typically ranged from 1-5 mg.

Patterns of LM29173 load in *Diporeia* are not well described by station (Welch's ANOVA, $F_{9,24.18} = 2.2059$, $p = .06$; Fig S3.6) within and among lakes. However, COI sequences indicate that *Diporeia* populations form two distinct sub-species clades, with Lake Superior comprising a northern clade, and Lakes Michigan, Huron and Ontario comprising a southern clade. Lake Superior amphipods exhibit reduced LM29173 load in comparison to amphipods from southern clade populations. This is corroborated by *Diporeia* from the Finger Lakes in Central New York (Cayuga, Owasco, and Seneca Lakes), which are members of the southern clade and exhibit comparable interactions with LM29173 (54% prevalence, 797 copies mg^{-1} tissue; Fig 3.4).

Relationship between LM29173 association and amphipod nutritional quality – Prevalence of LM29173 did not correlate with amphipod lipid content (Δ% dry weight of whole animal). However, average viral copy number correlated with an increase in amphipod lipid content (Fig 3.5a, b). Likewise, C:N stoichiometry of lipid-free animals increased coincident with LM29173 copy number (Fig 3.5c, d).

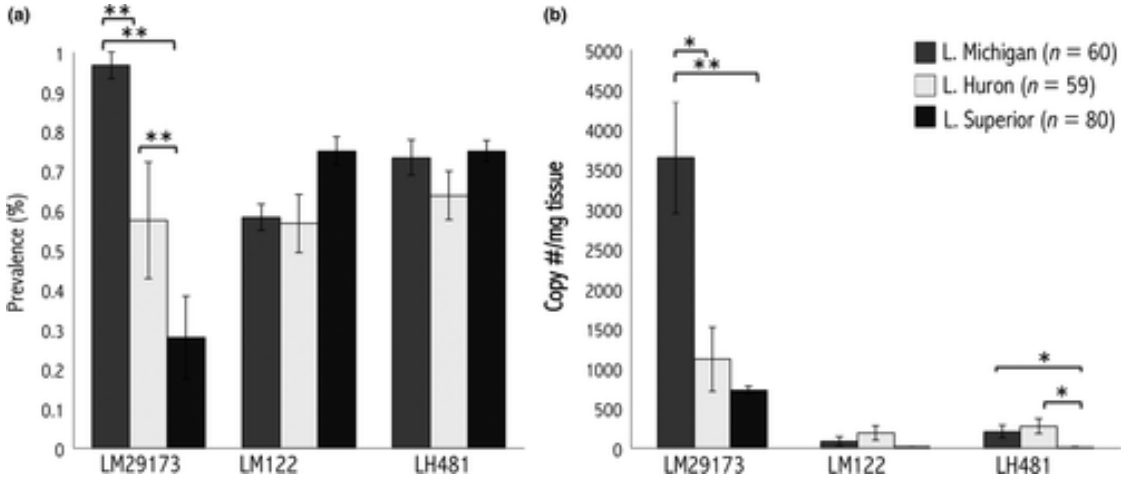

Figure 3.3 | *Quantification of prevalence and load of three target CRESS-DNA virus genotypes (qPCR)* – (a) Average genotype prevalence (±1 SE) defined as the percentage of organisms positive for viral genotype ($\chi 2 = 67.037$, Dunn's test post hoc) and (b) average genotype load (±1 SE) defined as mean copy number of viral genotype per unit wet weight within positive organisms (Welch's ANOVA, $p < .005$; Games–Howell post hoc). (a and b) $^{**}p \leq .001$, $^{*}p \leq .05$

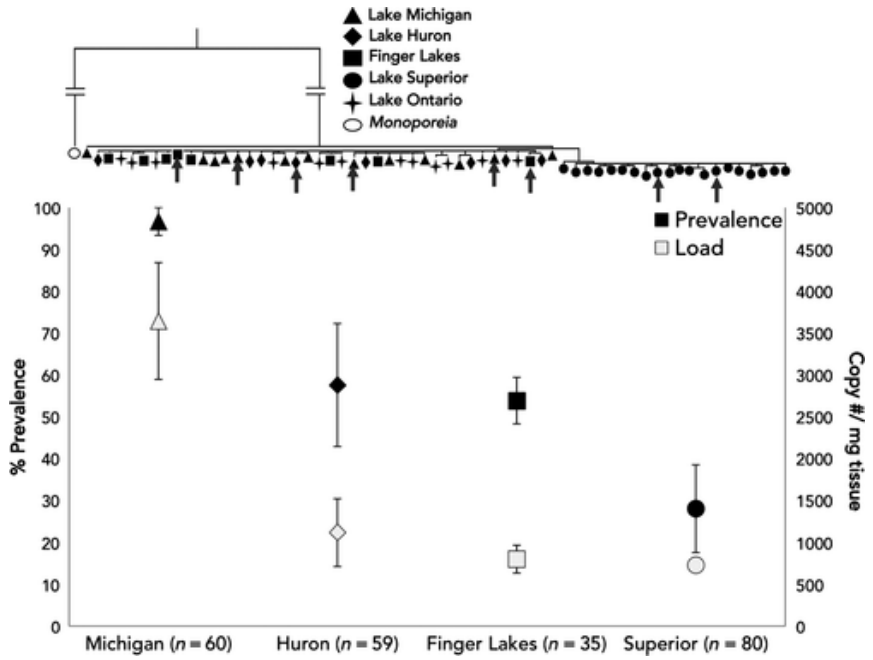

Figure 3.4 | *Haplotype specificity of CRESS-DNA viral distribution* - Average prevalence (±1 *SE*, filled shapes, e.g. ●) and load (per mg tissue, ±1 *SE*, unfilled shapes, e.g. ○) of viral genotype LM29173 in the Great Lakes (Michigan, Huron, Superior), and the Finger Lakes (Seneca, Owasco, Cayuga) aligned with GTR maximum parsimony tree of *Diporeia* cytochrome oxidase I (CO1, 538 nt alignment) sequences. Phylogeny node shape corresponds to amphipod population (open circle = *Monoporeia* out-group). Arrows correspond to new CO1 sequences contributed by this study. Out-group branch is collapsed

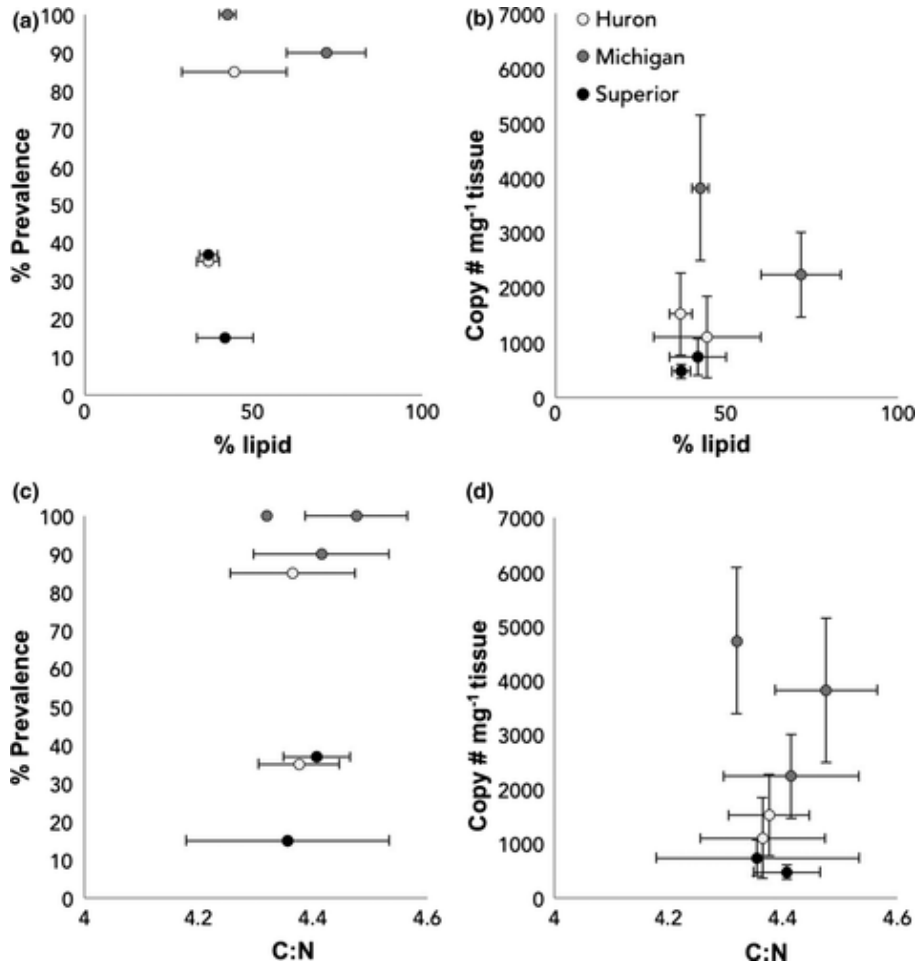

Figure 3.5 | *Amphipod nutritional quality relative to viral load* - Relationship between amphipod lipid content (% lipid = change in dry weight after lipid removal divided by initial dry weight) and LM29173 prevalence (a; Pearson's product–moment correlation, $t = 1.02$, $df = 8$, $p = .34$) and load mg^{-1} tissue wet weight (b; Copy number ±1 *SE*, Pearson's product–moment correlation, $t = 2.35$, $df = 5$, $p = .07$) and amphipod molar C:N and LM29173 prevalence (c; Pearson's product–moment correlation, $t = 0.490$, $df = 5$, $p = .64$) and load (d; Copy number ±1 *SE*, Pearson's product–moment correlation, $t = 0.0810$, $df = 5$, $p = .94$)

3.5 Discussion **|** As important detritivores and nutritional resources for higher trophic levels, benthic meiofauna are significant to ecosystem function and effective biomarkers of habitat condition. Despite their ecological significance, little is known about the viral consortia associated with invertebrates in the Laurentian Great Lakes. This study demonstrates that CRESS-DNA viral elements are common constituents of *Diporeia* viromes in the Laurentian Great Lakes. Metaviromic analyses identified additional CRESS-DNA virus genotypes, and determined that LM29173 is a prevalent and recurrent member of viral consortia in Lakes Michigan and Huron, but not Lake Superior. Distributions of LM29173 are likely governed by amphipod haplotype demographics. Although data imply that CRESS-DNA viruses are not associated with reduced nutritional quality of amphipods (lipid content and elemental stoichiometry), their potential impact on *Diporeia* behavior, physiology and mortality remains unknown.

Viral discovery – Diporeia from Great Lakes Michigan, Huron and Superior were selected for viral discovery efforts to identify potentially relevant viral sequences associated with declining and stable amphipod populations. We also examined viromes of Lake Erie *Echinogammarus* and New York Finger Lakes *Diporeia* to determine if viruses are specific to *Diporeia* in Great Lakes ecosystems. Invasive amphipods of genus *Echinogammarus* are genetically and ecologically divergent from *Diporeia*, and provided an outgroup to determine if targeted viral genotypes are genus specific. Finger Lakes *Diporeia* populations are not currently in decline, but share haplotype similarity to Lake Michigan and Huron *Diporeia* and therefore offered insight into the role of *Diporeia* genetics on virome composition.

Among *Diporeia* viromes, CRESS-DNA viruses were the most commonly observed metazoan-associated viral group (Fig S3.4). This is a reflection of biases inherent to Φ29 polymerase-dependent rolling circle amplification techniques, which enrich for small, circular DNA templates (Kim et al., 2008; Roux et al., 2016). The majority of reads in Lake Erie *Echinogammarus* viromes are also associated with CRESS-DNA viruses, indicating that the presence of this viral group may be widespread among amphipods in the Great Lakes. Three putative CRESS-DNA virus-

like contigs were identified within their libraries and further assessed for (1) completeness (contig and ORF length), (2) presence of an energetically favorable stem loop structure encoding a nonanucleotide motif (ori) and (3) ORF orientation, order and functional domains (Figure S3.2).

LM122 and LH481 were identified in Lake Michigan and Lake Huron viromes, respectively. Both were most similar to CRESS- DNA viruses identified in aquatic environments (Yoshida et al., 2013; Zawar-Reza et al., 2014; Fig 3.2). Like other CRESS-DNA viruses recovered from these habitats, LM122 and LH481 contain diverse genome architectures (Fig 3.2): LM122 encodes two ORFS in an ambisense organization, whereas LH481 encodes two ORFs in a unisense orientation. Viral genotype LM29173, a Type V CRESS-DNA genome identified in previous viromes was also detected in viromes with high read coverage (Fig 3.2). Of the 11 circular genomes closed in the 2013 analysis of *Diporeia* viral consortia, only LM3635 (270 total reads recruited) and LM29173 (133,697 total reads recruited) were detected. This may indicate the transience of most CRESS-DNA virus-like elements in *Diporeia*, excluding LM29173.

CRESS-DNA virus genotype distribution – The total number of reads recruited to target CRESS-DNA virus-like sequences in Lakes Michigan and Huron *Diporeia* was 2.6-fold higher than in Lake Superior or Finger Lakes amphipods (Fig S3.4), coinciding with populations experiencing the greatest degree of decline (Barbiero et al., 2011). However, this may be a reflection of inconsistent amplification biases that arose during metavirome sample preparation (Φ29 rolling circle amplification). Therefore, this heterogeneity was further investigated on an individual-organism basis via qPCR to determine the range and distribution of three putative CRESS-DNA genotypes.

Consistent with 2013 findings (Hewson et al., 2013a), LM29173 is a dominant and recurrent constituent of *Diporeia* viral consortia in Lake Michigan. Genotypes LM122 and LH481 were present in the majority of all *Diporeia* collected from Lakes Michigan, Huron and Superior (Fig 3.3). However, corresponding load was negligible compared to the average LM29173 load per lake (Fig 3.3b), indicating that these genotypes may have different tropisms or temporal distributions in

amphipods, or do not specifically infect *Diporeia*. Conversely, LM29173 was consistently detected in significantly greater copy number (Fig 3.3) in *Diporeia* spanning a wide geographic range. Although this does not preclude the possibility that LM29173 is dietary in origin or affiliated with a eukaryotic symbiont of *Diporeia*, it remains unlikely due to (1) the indiscriminate nature of amphipod detritivory, (2) the abundance of ssDNA genome copies within single organisms, (3) the consistency of viral detection among *Diporeia* of a range of ontogenies, population densities and locations, (4) the lack of known eukaryotic symbionts affiliated with *Diporeia* and (5) the presence of LM29173-specific RNA transcripts recovered from tissues. LM29173 transcripts were detected in six of eight LM29173-positive samples from Lake Michigan, indicating active replication of the viral genotype within tissues (Fig S3.5b). Microscopy (TEM) is required to provide conclusive evidence of CRESS-DNA virus presence within specific amphipod tissues.

Distributions of LM29173 largely coincided with the status of *Diporeia* populations in the Great Lakes: viral prevalence and load in Lake Michigan and Huron significantly exceeded that in Lake Superior (Fig 3.3). However, amphipod genetic divergence may play a role in governing patterns of viral distribution. Pilgrim et al. (2009) reports the presence of two genetically distinct clades of *Diporeia* populating the Great Lakes: a southern clade, including Lakes Michigan, Huron and Ontario, and a northern clade including Lake Superior. This study corroborates this conclusion and provides additional cytochrome c oxidase I (COI) sequences as evidence. These haplotype clusters hypothetically arose in the Pleistocene (circa 650,000 years ago) due to severe population bottlenecks and patterns of repopulation from isolated refugia after periodic North American glacial retreats (McCalla, 2010; Pilgrim et al., 2009; Usjak, 2010). Although not divergent enough to delineate species identities (Pilgrim et al., 2009), haplotype demographics may, in part, influence patterns of LM29173 prevalence and load (Fig 3.4). COI sequences indicate that *Diporeia* populations from the Finger Lakes (Cayuga, Owasco and Seneca Lakes, New York, USA) are members of a distinct southern clade. These populations exhibit prevalence and load of LM29173 comparable to other southern clade populations, but have not exhibited population declines (Dermott

et al., 2005; Watkins et al., 2012). Therefore, patterns of viral distribution correspond predominantly to amphipod haplotype demographics, rather than environmental variables (i.e. abiotic factors specific to sample site), indicating that bottlenecks produced by historic glaciation events generate sub-species variability that may govern patterns of viral association.

Genotype LM29173 was also consistently detected in virioplankton from Lakes Michigan, Huron and Superior, ranging from 0 to 135.5 m above the benthos (Fig S3.5a and S3.7). However, greatest copy numbers were detected in Lake Michigan (>25-fold higher than in Lakes Huron or Superior), potentially indicating the ability of this virus to persist as virioplankton, congruent with characteristics of cultured, agro-economically relevant CRESS-DNA viruses (Allan, Phenix, Todd, & McNulty, 1994; Raidal & Cross, 1994). Alternatively, LM29173 may infect a range of other organisms in the water column. Similarly, sequences from a short-read dataset constructed from Lake Michigan and Ontario ballast water viromes (Kim, Aw, Teal, & Rose, 2015) recruit to target CRESS-DNA virus genomes (0.3–1,9679 average read depth per base, 50% read length, 80% sequence similarity) denoting the possibility of viral nucleic acid transport for 2–8 days in a commercial holding tank. Despite the environmental stability and consequent potential of LM29173 to distribute to multiple lakes, this genotype is predominantly detected in the southernmost Great Lakes (Michigan, Huron).

Impact of LM29173 association on amphipod nutritional quality – Although previous studies illustrate a correlation between invertebrate and CRESS-DNA population dynamics (Hewson et al., 2013b), the mechanism and degree to which CRESS-DNA viruses alter invertebrate ecological fitness remain unclear. The distinct correlation between amphipod population status and CRESS-DNA virus dynamics indicates that association with LM29173 may be an indicator of organism stress, or that the virus imposes some cost upon amphipod physiology. Because *Diporeia* serve as conduits of energy to benthic icthyofauna, impacts of viruses on amphipod nutritional quality could impact higher trophic level organisms (Gardner et al., 1985). However, proxy measurements for amphipod nutritional content (lipid content, C:N stoichiometry, Fig 3.5) were not reduced with

increased LM29173 load. Therefore, LM29173 association either does not significantly impact amphipod physiology, or consequences of association are confounded by changes in amphipod metabolism.

For example, increases in lipid content or C:N stoichiometry in association with high viral load could be a function of lower conspecific density and, therefore, reduced competition for food in populations with high incidence of LM29173. In this scenario, low viral prevalence and dense populations of *Diporeia* may enhance the effects of conspecific competition and result in a high quantity of organisms with low net nutritional quality in Lake Superior. This relationship may indicate that prevalence of LM29173 is not predicated upon nutritional deficiency or conspecific competition. The correlation between high viral load and lipid content could also be a reflection of changes in amphipod feeding preference/behavior in the presence of the virus to counteract potential impacts of infection (Povey, Cotter, Simpson, & Wilson, 2013). In addition, differences in lake ecosystem physical and biochemical features (e.g. dissolved organic carbon, nutrient cycling, pH, temperature, etc.) may further complicate the relationship between amphipod nutritional quality and LM29173.

Although LM29173 is differentially abundant in declining *Diporeia* populations from Lake Michigan and Lake Huron compared to populations from Lake Superior, it is unlikely that this virus is specifically responsible for *Diporeia* decline. LM29173 is also prevalent in stable amphipod populations from the Finger Lakes, indicating that viral distribution is possibly a function of amphipod genetics. Study of CRESS-DNA virus pathogenicity is necessary to establish tropism and epidemiology. Understanding of the consequence(s) and mechanism(s) of CRESS-DNA virus association in invertebrates will lend insight into both viral and amphipod ecology. Studies discerning the costs of viral association on an individual's behavior or physiology may also explain the ability of some amphipod haplotype clusters to resist, tolerate or clear CRESS-DNA viruses.

3.6 References

Allan GM, Phenix KV, Todd D, McNulty MS. 1994. Some biological and physico-chemical properties of porcine circovirus. *Zentralbl Veterinary med B*. 41(1):17-26. doi:10.1111/j.1439-0450.1994.tb00201.x

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

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

Auer, M. T., Auer, N. A., Urban, N. R., & Auer, T. 2013. Distribution of the amphipod *Diporeia* in Lake Superior: The ring of fire. *Journal of Great Lakes Research*, 39, 33–46.

Barbiero, R. P., Rockwell, D. C., Warren, G. J., & Tuchman, M. L. 2006. Changes in spring phytoplankton communities and nutrient dynamics in the eastern basin of Lake Erie since the invasion of *Dreissena spp*. *Canadian Journal of Fisheries and Aquatic Sciences*, 63, 1549–1563.

Barbiero, R. P., Schmude, K., Lesht, B. M., Riseng, C. M., Warren, G. J., & Tuchman, M. L. 2011. Trends in *Diporeia* populations across the Laurentian Great Lakes, 1997–2009. *Journal of Great Lakes Research*, 37, 9–17.

Birkett, K. S., Lozano, S., & Rudstam, L. G. 2015. Long-term trends in Lake Ontario's benthic macroinvertebrate community from 1994– 2008. *Aquatic Ecosystem Health & Management*, 18, 76–85.

Bligh, E. G., & Dyer, W. J. 1959. A rapid method of total lipid extraction and purification. *Canadian Journal of Biochemistry and Physiology*, 37, 911–917.

Cook, D. G., & Johnson, M. G. 1974. Benthic macroinvertebrates of the St. Lawrence Great Lakes. *Journal of the Fisheries Research Board of Canada*, 31, 763–782.

Dermott, R., Bonnell, R., & Jarvis, P. 2005. Population status of the amphipod *Diporeia* in eastern North American lakes with or without *Dreissena*. *Verhandlungen Internationale Vereinigung für Theoretische und Angewandte Limnologie,* 29, 880–886.

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 of the United States of America, 110, 1375– 1380.

Edgar, R. C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32, 1792– 1797.

Fitzgerald, S. A., & Gardner, W. S. 1993. An algal carbon budget for pelagic-benthic coupling in Lake Michigan. *Limnology and Oceanography*, 38, 547–560.

Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, 3, 294–299.

Gardner, W. S., Nalepa, T. F., Frez, W. A., Cichocki, E. A., & Landrum, P. F. 1985. Seasonal patterns in lipid content of Lake Michigan macroinvertebrates. *Canadian Journal of Fisheries and Aquatic Sciences*, 42, 1827–1832.

Gudenkauf, B. M., & Hewson, I. 2016. Comparative metagenomics of viral assemblages inhabiting four phyla of marine invertebrates. *Frontiers in Marine Science*, 3, 23. doi:10.3389/fmars.2016.00023

Guiguer, K. R. R. A., & Barton, D. R. 2002. The trophic role of *Diporeia* Amphipoda in Colpoys Bay Georgian Bay benthic food web: A stable isotope approach. *Journal of Great Lakes Research*, 28, 228– 239.

Halfon, E., Schito, N., & Ulanowicz, R. E. 1996. Energy flow through the Lake Ontario food web: Conceptual model and an attempt at mass balance. *Ecological Modelling*, 86, 1–36.

Hewson, I., Eaglesham, J. B., Höök, T. O., LaBarre, B. A., Sepulveda, M. S., Thompson, P. D., Watkins, J.M., Rudstam, L. 2013a. Investigation of viruses in *Diporeia* spp. from the Laurentian Great Lakes and Owasco Lake as potential stressors of declining populations. *Journal of Great Lakes Research*, 39, 499–506.

Hewson, I., Ng, G., Li, W., LaBarre, B. A., Aguirre, I., Barbosa, J. B., … Hairston, N. G. Jr. 2013b. Metagenomic identification, seasonal dynamics, and potential transmission mechanisms of a *Daphnia*-associated single-stranded DNA virus in two temperate lakes. *Limnology and Oceanography*, 58, 1605–1620.

Kim, Y., Aw, T. G., Teal, T. K., & Rose, J. B. 2015. Metagenomic investigation of viral communities in ballast water. *Environmental Science & Technology*, 49, 8396–8407.

Kim, K. H., Chang, H. W., Nam, Y. D., Roh, S. W., Kim, M. S., Sung, Y., . . . Bae, J. W. 2008. Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Applied and Environmental Microbiology*, 74, 5975–5985.

McCalla, S. G. 2010. Patterns of genetic diversity in *Diporeia* in the Laurentian Great Lakes. Master's thesis, Order No. 1490680. Available from ProQuest Dissertations & Theses Global. 861341831. Retrieved from https://search.proquest.com/docview/861341831? accountid=10267

Messick, G. A., Overstreet, R. M., Nalepa, T. F., & Tyler, S. 2004. Preva- lence of parasites in amphipods *Diporeia* spp. from Lakes Michigan and Huron, USA. *Diseases of Aquatic Organisms*, 59, 159–170.

Naccache, S. N., Greninger, A. L., Lee, D., Coffey, L. L., Phan, T., Rein, Weston, A., ... Chiu, C. Y. 2013. The perils of pathogen discovery: Origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *Journal of Virology*, 87, 11966–11977.

Nalepa, T. F., Fanslow, D. L., & Foley, A. J. III 2005. Spatial patterns in population trends of the amphipod *Diporeia spp.* and *Dreissena* mussels in Lake Michigan. *Verhandlungen Internationale Vereinigung für Theoretische und Angewandte Limnologie*, 29, 426–431.

Nalepa, T. F., Fanslow, D. L., Foley, A. J. III, Lang, G. A., Eadie, B. J., & Quigley, M. A. 2006. Continued disappearance of the benthic amphipod *Diporeia* spp. in Lake Michigan: Is there evidence for food limitation? *Canadian Journal of Fisheries and Aquatic Sciences*, 63, 872–890.

Nalepa, T. F., Fanslow, D. L., & Lang, G. A. 2009. Transformation of the offshore benthic community in Lake Michigan: Recent shift from the native amphipod *Diporeia* spp. to the invasive mussel Dreissena *rostriformis bugensis*. *Freshwater Biology*, 54, 466–479.

Ng, T. F. F., Willner, D., Nilsson, C., Lim, Y. W., Schmieder, R., Chau, B., . . . Breitbart, M. 2010. Vector-based metagenomics for animal virus surveillance. *International Journal of Infectious Diseases*, 14, e378.

Pilgrim, E. M., Scharold, J., Darling, J. A., & Kelly, J. R. 2009. Genetic structure of the benthic amphipod *Diporeia* Amphipoda: *Pontoporeiidae* and its relationship to abundance in Lake Superior. *Canadian Journal of Fisheries and Aquatic Sciences*, 66, 1318–1327.

Pothoven, S. A., Nalepa, T. F., Schneeberger, P. J., & Brandt, S. B. 2001. Changes in diet and body condition of Lake Whitefish in southern Lake Michigan associated with changes in benthos. *North American Journal of Fisheries Management*, 21, 876–883.

Povey, S., Cotter, S. C., Simpson, S. J., & Wilson, K. 2013. Dynamics of macronutrient self-medication and illness-induced anorexia in virally infected insects. *Journal of Animal Ecology*, 83, 245–255.

Raidal, S. R., & Cross, G. M. 1994. The haemagglutination spectrum of psittacine beak and feather disease virus. *Avian Pathology*, 23, 621– 630.

Rennie, M. D., Sprules, W. G., & Johnson, T. B. 2009. Factors affect- ing the growth and condition of lake whitefish Coregonus clu- peaformis. *Canadian Journal of Fisheries and Aquatic Sciences*, 66, 2096–2108.

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

Rosario, K. S., Duffy, S., & Breitbart, M. 2012. A field guide to eukaryotic circular single-stranded DNA viruses: Insights gained from metagenomics. *Archives of Virology*, 157, 1851–1871.

Rosario, K., Schenck, R. O., Harbeitner, R. C., Lawler, S. N., & Breitbart, M. 2015. Novel circular single-stranded DNA viruses identified in marine invertebrates reveal high sequence diversity and consistent predicted intrinsic disorder patterns within putative structural proteins. *Frontiers in Microbiology,* 6, 696. doi:10.3389/ fmicb.2015.00696

Roux, S., Solonenko, N. E., Dang, V. T., Poulos, B. T., Schwenck, S. M., Goldsmith, D. B., …Sullivan, M. B. 2016. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ*, 4, e2777.

Rozen, S., & Skaletsky, H. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods in Molecular Biology, 132, 365–386.

Ruxton, G. D., & Beauchamp, G. 2008. Time for some a priori thinking about post hoc testing. *Behavioral Ecology*, 19, 690–693.

Ryan, D. J., Sepu lveda, M. S., Nalepa, T. F., & Ho€o€k, T. O. 2012. Spatial variation in RNA:DNA ratios of *Diporeia* spp. in the Great Lakes region. *Journal of Great Lakes Research*, 38, 187–195.

U.S. Environmental Protection Agency. 2009. National coastal condition assessment quality assurance project plan 2008–2012. United States EPA, Office of Water, Office of Wetlands, Oceans and Watersheds, Washington, DC. EPA/841-R-09-004.

Usjak, S. 2010. Comparative phylogeography of North American *Diporeia hoyi* and *Gammarus lacustris* order: Amphipoda. Master's thesis. Retrieved from UWSpace. http://hdl.handle.net/10012/4902

Vanderploeg, H. A., Liebig, J. R., Nalepa, T. F., Fahnenstiel, G. L., & Pothoven, S. A. 2010. Dreissena and the disappearance of the spring phytoplankton bloom in Lake Michigan. *Journal of Great Lakes Research*, 36, 50–59.

Watkins, J. M., Dermott, R., Lozano, S. J., Mills, E. L., Rudstam, L. G., & Scharold, J. V. 2007. Evidence for remote effects of dreissenid mussels on the amphipod *Diporeia*: Analysis of Lake Ontario benthic surveys, 1972–2003. *Journal of Great Lakes Research*, 33, 642.

Watkins, J. M., Rudstam, L. G., Mills, E. L., & Teece, M. A. 2012. Coexistence of the native benthic amphipod *Diporeia* spp. and exotic dreissenid mussels in the New York Finger Lakes. Journal of Great Lakes Research, 38, 226–235.

Wells, L. 1980. Food of alewives, yellow perch, spottail shiners, trout- perch, and slimy and fourhorn sculpins in southeastern Lake Michigan. Technical Paper no. 98, U.S. Fish and Wildlife Service.

Winters, A. D., Fitzgerald, S., Brenden, T. O., Nalepa, T. F., & Faisal, M. 2014. Spatio-temporal dynamics of parasites infecting *Diporeia* spp. Amphipoda, *Gammaridae* in southern Lake Michigan USA. *Journal of Invertebrate Pathology*, 121, 37–45.

Winters, A. D., Marsh, T. L., Brenden, T. O., & Faisal, M. 2015. Analysis of bacterial communities associated with the benthic amphipod *Diporeia* in the Laurentian Great Lakes Basin. *Canadian Journal of Microbiology*, 61, 72–81.

Yoshida, M. T., Takaki, Y., Eitoku, M., Nunoura, T., & Takai, K. 2013. Metagenomic analysis of viral communities in hadopelagic sediments. *PLoS* ONE, 8, e57271.

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

Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31, 3406–3434.

CHAPTER 4

GENE EXPRESSION OF BENTHIC AMPHIPODS (GENUS: *DIPOREIA)* IN RELATION TO A CIRCULAR SSDNA VIRUS ACROSS TWO LAURENTIAN GREAT LAKES[2]

4.1 Abstract **|** Circular *rep*-encoding ssDNA (CRESS-DNA) viruses are common constituents of invertebrate viral consortia. Despite their ubiquity and sequence diversity, the effects of CRESS-DNA viruses on invertebrate biology and ecology remain largely unknown. This study assessed the relationship between the transcriptional profile of benthic amphipods of genus *Diporeia* and the presence of the CRESS-DNA virus, LM29173, in the Laurentian Great Lakes to provide potential insight into the influence of these viruses on invertebrate gene expression. Twelve transcriptomes derived from *Diporeia* were compared, representing organisms from two amphipod haplotype clades (Great Lakes Michigan and Superior, defined by COI barcode sequencing) with varying viral loads (up to 3×10^6 genome copies organism^{-1}). Read recruitment to *de novo* assembled transcripts revealed 2,208 significantly over or underexpressed contigs in transcriptomes with above average LM29173 load. Of these contigs, 31.5% were assigned a putative function. The greatest proportion of annotated, differentially expressed transcripts were associated with functions including: (1) replication, recombination, and repair, (2) cell structure/biogenesis, and (3) post-translational modification, protein turnover, and chaperones. Contigs putatively associated with innate immunity displayed no consistent pattern of expression, though several transcripts were significantly overexpressed in amphipods with high viral load. Quantitation (RT-qPCR) of target transcripts, non-muscular myosin heavy chain, β-actin, and ubiquitin-conjugating enzyme E2, corroborated transcriptome analysis and indicated that Lake Michigan and Lake Superior amphipods with high LM29173 load exhibit lake-specific trends in gene expression. While this investigation provides the first comparative survey of the transcriptional profile of invertebrates of variable CRESS-DNA viral load, additional inquiry is required to define the scope of host-specific responses to potential infection.

[2]Presented with minor amendment from the original published article:
Bistolas KSI, Rudstam LG, Hewson I. 2017. Gene expression of benthic amphipods (genus: *Diporeia*) in relation to a circular ssDNA virus across two Laurentian Great Lakes. *PeerJ* 5:e3810 doi:10.7717/peerj.3810
[2]Supplementary material may be accessed at: https://doi.org/10.7717/peerj.3810

4.2 Introduction | Circular *rep*-encoding ssDNA (CRESS-DNA) virus genomes are small (~1.7–6kb), circular molecules which encode, at minimum, major open reading frames *rep* (replication initiator protein) and *cap* (structural capsid protein; Rosario, Duffy & Breitbart, 2012; Rosario et al., 2015). Eukaryotic CRESS-DNA viruses broadly encompass ssDNA viruses that infect plants (*Geminiviridae, Nanoviridae*), and metazoans (*Circoviridae, Anelloviridae*; Dunlap et al., 2013; Rosario et al., 2017; Rosario, Duffy & Breitbart, 2012), and include common and important pathogens of ecologically or commercially relevant vertebrates. For example, beak and feather disease virus (BFDV, *Circoviridae*) is responsible for persistent immunosuppression in avian hosts (Eastwood et al., 2014) and porcine circoviruses infect domestic swine, manifesting sub-clinically (PCV1) or eliciting postweaning multisystemic wasting syndrome (PMWS, PCV2; Allan & Ellis, 2000). The use of culture-independent (viromic) approaches has led to the discovery and characterization of an extraordinary diversity of novel ssDNA viruses in environmental reservoirs and non-model invertebrates (Labonté & Suttle, 2013; Rosario & Breitbart, 2011; Rosario et al., 2017; Rosario, Duffy & Breitbart, 2012; Roux et al., 2016). To date, the etiology, pathology, and association between ssDNA viruses and any invertebrate remain wholly unknown. This study utilized whole transcriptome sequencing to investigate the relationship between a CRESS-DNA virus and benthic amphipods of genus *Diporeia* from the Laurentian Great Lakes.

Circular *rep*-encoding ssDNA viruses have been identified in association with several major aquatic invertebrate phyla, including the Annelida, Arthropoda, Chaetognatha, Cnidaria, Ctenophora, Echinodermata, and Mollusca, among others (Breitbart et al., 2015; Dayaram et al., 2016; Dunlap et

al., 2013; Eaglesham & Hewson, 2013; Fahsbender et al., 2015; Kibenge & Godoy, 2016; Hewson et al., 2013a, 2013b; Jackson et al., 2016; Rosario et al., 2015; Soffer et al., 2013). These viruses appear to be biogeographically widespread, taxonomically diverse, and common constituents of crustacean viromes (Dunlap et al., 2013; Hewson et al., 2013a, 2013b; Labonté & Suttle, 2013; Rosario et al., 2015, 2017; Rosario, Duffy & Breitbart, 2012). However, little is known about the role of CRESS-DNA viruses in mediating crustacean ecology, physiology, and mortality. Because no immortal crustacean cell line currently exists, propagation of crustacean-associated CRESS-DNA viruses *in vitro* remains intractable, and the unknown nature of CRESS-DNA virus tropism and infection dynamics in these systems impedes targeted sequencing of virus-infected cells. Furthermore, many microcrustaceans cannot be reared or maintained effectively in aquaria without significant physiological stress and high incidence of mortality, hindering *in vivo* infection experiments. Therefore, we implemented a whole-organism comparative transcriptome sequencing (transcriptomics) approach in evaluating the relationship between the presence of CRESS-DNA viral genotype, LM29173, and benthic crustaceans (genus: *Diporeia*) in Great Lakes ecosystems.

Diporeia are historically abundant benthic meiofauna in the Laurentian Great Lakes (Auer et al., 2013; Barbiero et al., 2011; Birkett, Lozano & Rudstam, 2015; Guiguer & Barton, 2002). These amphipods influence lake-wide biogeochemistry and mediate relationships between spring diatom blooms and upper trophic level consumers through detritivory and sediment bioturbation (Gardner et al., 1985; Guiguer & Barton, 2002; Halfon, Schito & Ulanowicz, 1996; Wells, 1980). Localized and precipitous declines in several *Diporeia* populations have prompted exploration of their viral consortia (Bistolas et al., 2017; Hewson et al., 2013a). Viromic sequencing has documented a common and recurrent CRESS-DNA virus genotype, LM29173, frequently detected in impacted *Diporeia* populations in Lakes Michigan and Huron, but rare among specimens from stable Lake Superior populations (Bistolas et al., 2017; Hewson et al., 2013a). It is also prevalent among amphipods from the deep, glacial Finger Lakes of Central New York (Seneca, Cayuga, and Owasco

Lakes). Previous DNA barcoding of maternally inherited cytochrome c oxidase I (COI) sequences (Pilgrim et al., 2009; Bistolas et al., 2017) have revealed sub-species genetic variation between impacted and stable populations, with *Diporeia* from Lakes Michigan, Huron, Ontario, Erie, and the Finger Lakes comprising a southern lake haplotype clade, and amphipods from Lake Superior comprising a northern lake haplotype clade. While LM29173 is more abundant in *Diporeia* from declining southern populations than stable northern populations, no advances have been made to describe the impact of this CRESS-DNA virus on amphipod biology. This study offers preliminary insight into the relationship between LM29173 and gene expression in amphipods from both haplotype clades, and provides transcriptional targets for further investigation. Specific objectives of this study were to (1) investigate the association between LM29173 presence/load and the transcriptional profile of *Diporeia*, (2) determine if detected changes in gene expression are specific to distinct *Diporeia* haplotypes, and (3) explore the effect of LM29173 presence on amphipod transcription of innate immunity regulators/effectors.

4.3 Methods **|** *Sample collection and transcriptome preparation - Diporeia* were collected in August–September, 2014, via Ponar benthic sampler from Great Lakes Michigan and Superior at EPA-designated stations (Fig. 4.1, Table S4.1; United States Environmental Protection Agency, 2012). Organisms were sieved to remove sediment (500μm), rinsed, and immediately individually frozen at −80 °C.

Nucleic acids were extracted from individual amphipods via ZR-Duet™ DNA/RNA MiniPrep kit (Zymo Research, Irvine, CA, USA). Presence and genome load (copy number) of LM29173 was determined via qPCR per Hewson et al. (2013a) using SsoAdvanced™ Universal Probes Supermix (Bio-Rad Laboratories, Hercules, CA, USA), corrected for total extraction volume, and standardized by organism wet weight (mg). Two samples with the highest and two samples with the lowest copy numbers organsim^{-1} of LM29173 from each of three stations (Lake Michigan 27 and

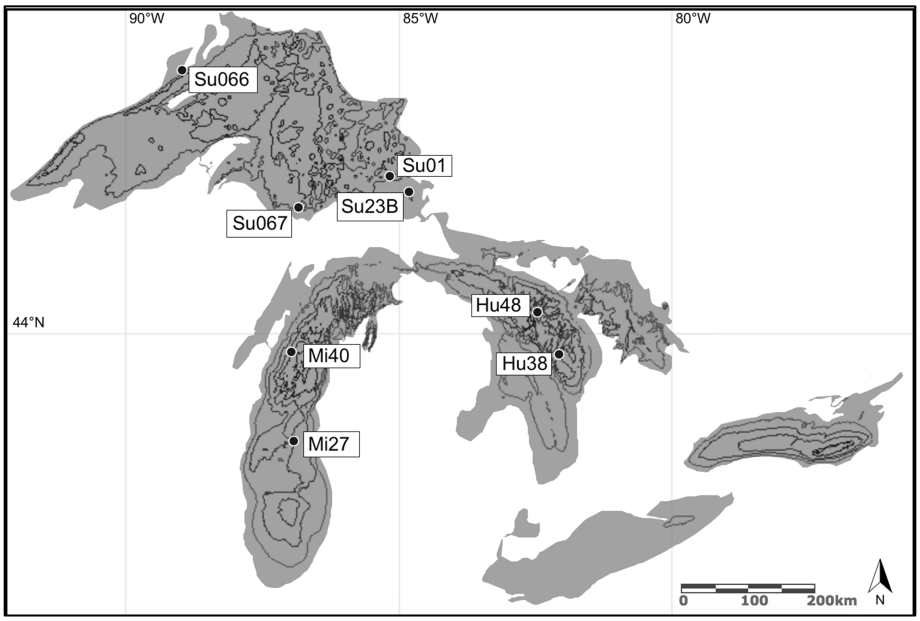

Figure 4.1 | *Amphipod collection sites in the Laurentian Great Lakes (August–September, 2014).* Collection locations are congruent with EPA-Great Lakes National Program Office (GLNPO) designated stations (United States Environmental Protection Agency, 2012). Specimens were collected on the R/V Lake Guardian via Ponar benthic sampler. Bathymetry data was provided by NOAA National Geophysical Data Center's Marine Geology & Geophysics Division (NGDC/MGG) and the NOAA Great Lakes Environmental Research Laboratory (GLERL). Map service published and hosted by Esri Canada© 2012 under Attribution-NonCommercial 2.5 Canada (CC BY-NC 2.5 CA) license https://creativecommons.org/licenses/by-nc/2.5/ca/. DOI:10.7717/peerj.3810/fig-1

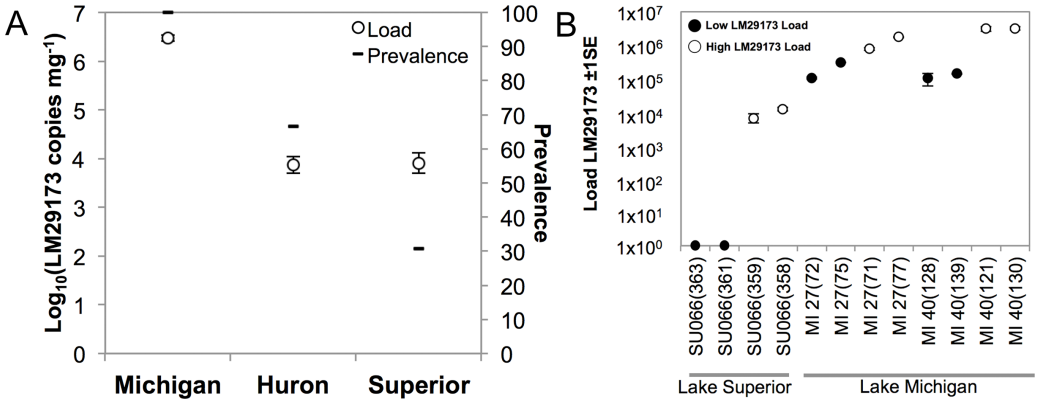

Figure 4.2 | *Quantitative detection of LM29173* – (A) Prevalence and average load (log_{10} transformed copy number mg^{-1} of tissue±1SE) of CRESS-DNA virus genotype LM29173 in amphipods from Great Lakes Michigan, Huron, and Superior. Viral load was significantly greater in Lake Michigan than Lakes Huron (Games–Howell post hoc t =7.30, p=3.1x10^{-9}) or Superior (Games–Howell post hoc t =7.30, p=3.0×10^{-9}); (B) Load of LM29173 (copy number $organism^{-1}$) in amphipods selected for transcriptome sequencing.

40, Lake Superior 066; Fig. 4.2, Fig. S4.1; United States Environmental Protection Agency, 2012) were selected for transcriptome preparation (n = 12 total transcriptomes, n = 4 per station). For selected samples, RNA fractions were further enzymatically digested with TurboDNAse (Thermo Fisher Scientific, Waltham, MA, USA) for 15min to reduce co-extracted DNA. Ribosomal RNA was depleted via mRNA-ONLY™ mRNA Isolation Kit (Epicentre, Madison, WI, USA), and remaining RNA was reverse transcribed and amplified via the TransPlex® Complete Whole Transcriptome Amplification Kit (WTA2; Sigma-Aldrich, Saint Louis, MO, USA) per manufacturer instructions. Resulting cDNA libraries were quantified via PicoGreen fluorescence and prepared for sequencing using a Nextera XT DNA Library Preparation Kit (Illumina, San Diego, CA, USA). Resulting libraries were subjected to 2×250 bp paired-end sequencing on an Illumina MiSeq at the Cornell University Core Laboratories Center (Ithaca, NY, USA). Libraries were deposited in Genbank (accession: PRJNA379017; SRR5341776–SRR5341788).

Transcriptome assembly and comparison of transcript expression - Reads were trimmed for quality (quality score < 0.05, modified-Mott trimming algorithm), ambiguous nucleotides (n = 0), length (50 nt ≤ length ≤ 251 nt), and Illumina adapters via CLC Genomics Workbench (v.8.5.1; Qiagen, Hilden, Germany). Reads mapped to SILVA rRNA databases (90% identity, 50% coverage via CLC Genomics Workbench; http://www.arb-silva.de/) were excluded from assembly. Remaining reads were then assembled *de novo* using Trinity on the Galaxy bioinformatics platform per default parameters (National Center for Genome Analysis Support, Indiana University Pervasive Technology Institute; Table S4.2). Resulting contigs were further clustered via CD-HIT-EST to reduce isoform redundancy (sequence identity cutoff = 0.98). Reads were aligned to contigs via the Bioconductor package EdgeR (Robinson & Smyth, 2007) in CLC Genomics Workbench (v.8.5.1; Qiagen, Hilden, Germany) to calculate relative read recruitment (reads per kilobase of transcript per million mapped reads; RPKM) and significance (corrected for multiple comparison via false discovery rate methods; FDR). Contigs that exhibited >10-fold change (EdgeR) in read recruitment,

ΔRPKM > 100, and FDR-adjusted $p < 0.05$ between the six low LM29173 load libraries and six high LM29173 load libraries were considered significantly differentially expressed genes (DEGs). DEGs were then annotated using Blast2Go (v.4.0.7 BLASTx, $e < 1\times10^{-5}$) and functionally classified by EuKaryotic Orthologous Group, or "KOG" (Joint Genome Institute).

Eight Lake Michigan libraries were grouped by station and viral load to identify DEGs common between both stations in Lake Michigan, minimizing the effect of between-lake genetic and environmental variance. DEGs shared between libraries were defined per the following criteria: >2-fold change in expression, ΔRPKM > 10 between libraries, significantly differentially expressed with an FDR-adjusted $p < 0.05$, and consistently over or underexpressed in both Lake Michigan stations. Contigs fulfilling these criteria were annotated via BLASTx against the non-redundant (nr) database and assessed for relevance to viral infection (Altschul et al., 1990).

To identify contigs affiliated with putative immune functions, reference sequences associated with invertebrate innate immunity were collected from the Insect Innate Immunity Database (Brucker et al., 2012) or curated from NCBI protein database queries of keywords in Table 4.1 of McTaggart et al. (2009) (keywords listed in Table S4.3). Contigs homologous to these genes were identified via BLASTx ($e < 1\times10^{-5}$; Altschul et al., 1990), and the RPKM of those that were >2-fold over or underexpressed in both Lake Michigan stations (Mi27 and Mi40) were standardized to total 18s rRNA RPKM per library and depicted via web-based visualization tool, Morpheous (Broad Institute, Cambridge, MA, USA).

Quantification (RT-qPCR) of differentially expressed target genes - Whole amphipods were collected in August–September, 2014 at EPA-designated stations (United States Environmental Protection Agency, 2012, Table S4.1) in Lakes Michigan, Huron, and Superior and extracted via ZR-Duet™ DNA/RNA MiniPrep kit (Zymo Research, Irvine, CA, USA). Load of LM29173 was quantified per Hewson et al. (2013a). RNA was reverse transcribed (RT) via Superscript III (Invitrogen, Carlsbad, CA, USA per manufacturer instructions). Parallel no-RT controls were

generated using identical reaction parameters and no reverse transcriptase. cDNA was subjected to duplex RT-qPCR (quantifying both a gene of interest and a reference gene to control for organism variability) using SsoAdvanced™ Universal Probes Supermix (Bio-Rad Laboratories, Hercules, CA, USA). Amplicons were gel-purified (Zymoclean™ Gel DNA Recovery Kit; Zymo Research, Irvine, CA, USA) and cloned (pGEM®-T Easy Vector; Promega, Madison, WI, USA) using JM109 competent *E. coli* (Invitrogen, Carlsbad, CA, USA). Plasmids were extracted per Zyppy™ Plasmid Miniprep Kit instructions (Zymo Research, Irvine, CA, USA) and Sanger sequenced (Cornell University Core Laboratories Center, Ithaca, NY, USA) to confirm primer/probe specificity. Reaction parameters and primer, probe, and standard sequences are detailed in Table S4.4.

Samples were run in duplicate with congruent duplicate no-RT controls and quantified using duplicate eight-fold standard dilutions (limits of detection described in Table S4.4). Ct values, quantity, and standard deviation between technical replicates were determined via StepOnePlus software v.2.3 (Foster City, CA, USA). Valid runs were defined by reaction efficiency >94% and standard regression linearity (R^2) > 0.98. Samples were excluded if Ct standard deviation between replicates was >0.5. Quantities were corrected for total extraction and reverse transcription dilutions. Quantities of targets β-actin (ACT), ubiquitin-conjugating enzyme E2 (UBQ), and non-muscular myosin heavy chain (NMHC) were standardized by copy number of elongation factor-1α (EF1A) per reaction.

4.4 Results & Discussion **|** Investigation of amphipod transcriptomes revealed differential expression of DNA replication/repair pathways, cytoskeletal architecture, and post-translational modification associated genes in correlation with CRESS-DNA virus load. However, the degree of variability between transcriptomes limited the ability to identify over or underexpression of specific molecular pathways. It is unknown whether vertebrate and invertebrate CRESS-DNA viruses utilize similar pathways of infection, particularly in light of the considerable divergence in sequence homology and

genome architecture between groups. Despite this, DEGs in *Diporeia* transcriptomes were often homologous to DEGs in porcine circoviral infections, or were associated with putative innate immune functions. Expression of these genes varied between amphipod haplotype clades, suggesting that the transcriptional relationship with LM29173 may have a heritable component (for example, as a product of acquired resistance to a prevelant viral genotype, variation in innate immune response, etc). It remains unclear if CRESS-DNA viral load corresponds significantly with ecologically relevant changes in invertebrate physiology.

Detection of LM29173 - Prevalence and load of LM29173 was significantly greater in Lake Michigan (100%) than Lakes Huron (66.7%; Games–Howell, $p < 1 \times 10^{-8}$, Ruxton & Beauchamp, 2008) and Superior (30.8%; Games–Howell, $p < 1 \times 10^{-8}$; Ruxton & Beauchamp, 2008; Welch's ANOVA, $F_{2,26.24} = 26.4$, $p = 5.21 \times 10^{-7}$) per qPCR in individual *Diporeia* co-extracted for DNA and RNA, congruent with previous observations of the distribution of this genotype (Bistolas et al., 2017). Pilgrim et al. (2009) utilized mitochondrial COI sequences to identify sub-species genetic variation between *Diporeia* populations among Great Lakes ecosystems, ultimately delineating two clades with distinct haplotype signatures. qPCR results corroborate previous observations that LM29173 is detected in greater abundance in southern lakes haplotype clade populations (Lakes Michigan, Huron, Ontario, Erie, and the Finger Lakes), relative to northern lakes haplotype clade populations (Lake Superior; Bistolas et al., 2017). Because LM29173 was positively detected in all Lake Michigan amphipods, samples with the highest and lowest respective load of LM29173 were utilized for transcriptome preparation (Fig. 4.2; Fig. S4.1).

Transcriptome assembly and annotation of DEGs- Sequence reads from twelve *Diporeia* transcriptomes were collated (n = 14,702,859 after trimming and exclusion of rRNA-like sequences) to *de novo* assemble 82,074 contigs with a mean length of 310nt and N50 value of 290nt (Fig. S4.2; Tables S4.2 and S4.5). Despite rRNA depletion prior to sequencing, computational subtraction of rRNA reads was considerable (0.016–36.95%), but comparable to previously observed proportions

in other studies (Schmieder, Lim & Edwards, 2012; Stewart, Ottesen & DeLong, 2010), likely due to the extreme abundance and ribosomal protection of rRNAs. Less than 1.25% of all rRNA-mapped reads (90% identity, 50% coverage; SILVA rRNA database) were putatively bacterial in origin, indicating that co-infecting microbes may contribute to variation in *Diporeia* transcriptional profiles. No transcripts of non-target CRESS-DNA viruses were identified. However, putative metazoan-associated RNA viruses were identified when compared to a manually curated database of viral RNA-dependent RNA polymerase sequences (GenBank) or the non-redundant database (BLASTx; $e < 1\times10^{-5}$). Among others, these contigs were homologous to members of the *Nodaviridae*, *Nyamiviridae*, *Orthomyxoviridae*, *Peribunyaviridae*, *Phenuiviridae*, and *Rhabdoviridae*. It remains unclear if these sequences represent transient/nonpathogenic viruses or specific pathogens of *Diporeia.* Despite methodological biases favoring amplification of encapsidated RNA viruses, read recruitment to these contigs was negligible, indicating that these genotypes may have minimal relative impact on overall amphipod transcription (See Appendix I).

Due to the lack of a reference *Diporeia* genome, transcripts were conservatively assembled within haplotype clades using an isoform-sensitive algorithm, resulting in a fragmented assembly with multiple isoforms per gene. To reduce redundant read mapping, contigs were grouped into 59,317 isoform clusters prior to recruitment analysis. Library D130 (Lake Michigan site Mi40; Table S4.5) contained fewer total reads relative to other libraries, but was retained, as relative read recruitment was standardized by sequencing depth per library. The statistical package, EdgeR, detected 2,208 significantly DEGs between libraries with high and low LM29173 load among three Great Lakes stations (Table 4.1), with “low viral load” delineated by metatranscriptomes in which LM29173 was detected at the lowest quantity per site in Lake Michigan, as this genotype was positively detected in all Lake Michigan amphipods.

Correlative multidimensional scaling (MDS) analyses indicated that transcriptomes do not cluster by viral presence, viral load, station, or haplotype, likely as a result of high variability in

ontogeny and life history between organisms (Fig. S4.3). Libraries from Mi27 (Lake Michigan) contained over seven-fold more DEGs than Mi40 (Lake Michigan) or Su066 (Lake Superior) libraries. 89% of these transcripts were overexpressed in libraries with high LM29173 load (Table 4.1; Fig. 4.3) but were small, unannotated, contained no ORFs, and were therefore removed from downstream analyses. Conversely, volcano plots (Fig. 4.3) illustrate a roughly symmetrical distribution of significantly over and underexpressed genes in libraries from Mi40 and Su066 relative to viral load.

Library	Overexpressed	Underexpressed	Total
SU066	169	65	234
MI40	129	161	290
MI27	1,497	187	1,684

Table 4.1 | *Total number of over- and underexpressed contigs in transcriptomes with above average LM29173 load* - Contigs that exhibited >10-fold change (EdgeR), ΔRPKM > 100, and FDR-adjusted $p < 0.05$ were considered significantly differentially expressed genes (DEGs). DOI: 10.7717/peerj.3810/table-1

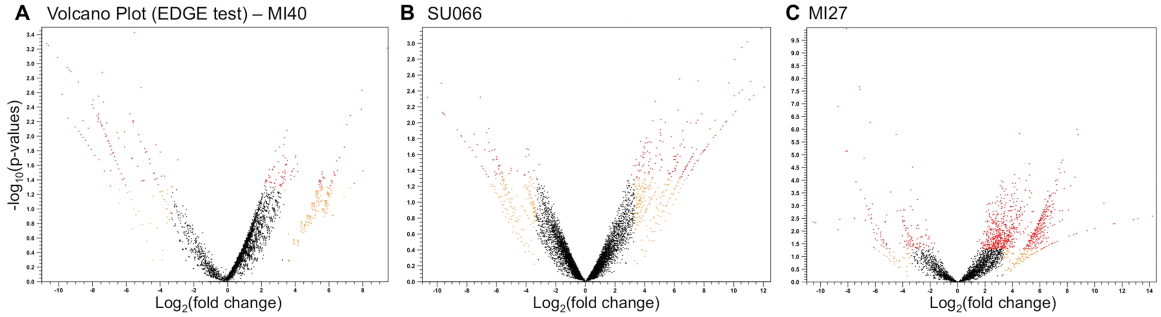

Figure 4.3 | *Volcano plots depicting the distribution of differentially expressed contigs* – Distribution of differentially expressed contigs was determined by EdgeR (Robinson & Smyth, 2007) for libraries from each of three stations: Superior 066 (Su066), Michigan 40 (Mi40), and Michigan 27 (Mi27). Orange points indicate >10-fold differentially expressed contigs (x-axis, as determined via EdgeR); red points indicate significantly differentially expressed contigs (y-axis, FDR-adjusted $p < 0.05$). DOI: 10.7717/peerj.3810/fig-3

Due to its evolutionary distance from sequenced model organisms, the *Diporeia* transcriptome remains incompletely annotated. Therefore, DEGs were broadly annotated by putative function using BLASTx via Blast2Go (v.4.0.7). Successfully identified contigs were further assigned to a euKaryotic Orthologous Group, or "KOG" classification (Joint Genome Institute, Walnut Creek,

CA, USA). Contigs that received designations of "general function prediction only" (KOG designation "R") or "function unknown" (KOG designation "S") were excluded from analysis. Among remaining functionally annotated contigs (n=696), most were involved in replication, recombination and repair (KOG designation "L", n=61), cell wall/membrane/envelope biogenesis (KOG designation "M", n=32), or post-translational modification, protein turnover, and chaperones (KOG designation "O", n=26, Fig. 4.4). These three functions were further investigated for potential relevance to viral infection.

DEGs involved in replication, recombination, and repair—KOG "L" – The proportion of DEGs involved in modulating DNA synthesis and stability may indicate that CRESS-DNA viruses alter or manipulate cellular replication pathways. This is congruent with the dynamics of circoviral infections in vertebrates, which exploit cellular DNA damage responses through a complex kinase cascade, triggering apoptosis and ultimately facilitating viral replication (Wei et al., 2016). Contigs homologous to unclassified DNA binding proteins, DNA modification enzymes, nucleases, histone structural components, and mobile elements/DNA translocases were differentially expressed. Several DEGs were responsible for chromatin remodeling, indicating a potential correlation between states of nucleosome packaging and viral load. However, many of these transcripts were associated with opposing functions (e.g., DNA methylases and demethylases) and may target different chromatin residues, rendering it difficult to determine if the presence of LM29173 leads to differential transcription.

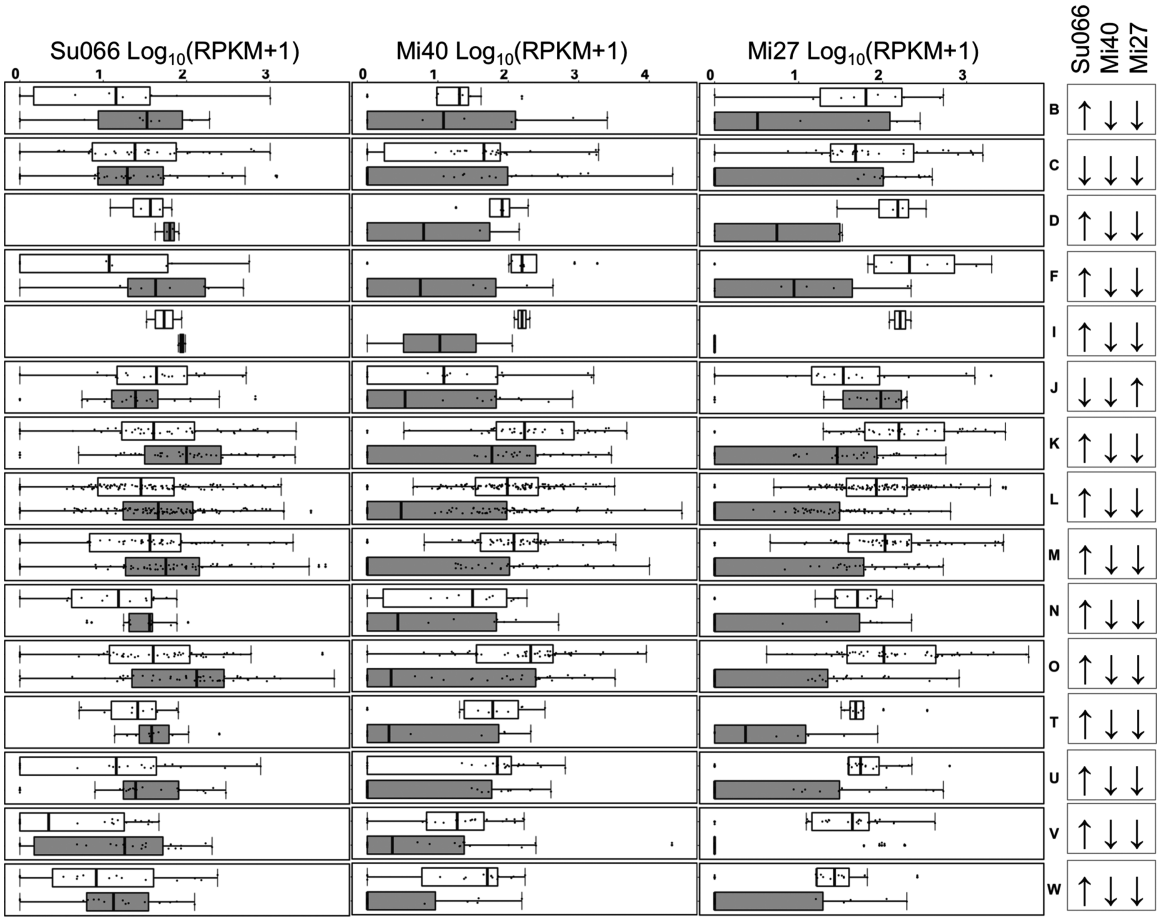

Figure 4.4 | *Average amphipod transcript expression in relation to LM29173 load* – Average expression (Log_{10}(RPKM+1)) of contigs in transcriptomes associated with above (grey) and below (white) average LM29173 load. Arrows indicate greater (↑) or reduced (↓) average transformed RPKM in transcriptomes with high LM29173 load relative to transcriptomes with low LM29173 load. Contigs are grouped by putative functional annotation (KOG, EuKaryotic Orthologous Groups), and abbreviations correspond to the following functions: (B) chromatin structure and dynamics, (C) energy production and conversion, (D) cell cycle control, cell division, chromosome partitioning, (F) nucleotide transport and metabolism, (I) lipid transport and metabolism, (J) translation, ribosomal structure and biogenesis, (K) transcription, (L) replication, recombination and repair, (M) cell wall/membrane/envelope biogenesis, (N) cell motility, (O) posttranslational modification, protein turnover, chaperones, (T) signal transduction mechanisms, (U) intracellular trafficking, secretion, and vesicular transport, (V) defense mechanisms, (W) extracellular structures.
DOI:10.7717/peerj.3810/fig-4

DEGs involved in cell wall/membrane/envelope biogenesis—KOG "M" – Several homologs of cell-surface receptors and transmembrane transporters including cubilin, calsyntenin, choline transporters, and g-protein coupled receptors were differentially expressed in transcriptomes with high LM29173 load. Contigs putatively involved in carapace biogenesis and the production of other structural/connective tissues (keratin, collagen, and elastin), as well as those involved in cell movement and intracellular transport (actin and myosin) were also significantly differentially expressed. These proteins play central roles in cell growth and replication, and differences in their transcription may be an artifact of natural variability between organisms. However, mis-regulation of these proteins is a well-documented response to many metazoan virus infections (Döhner & Sodeik, 2005; Luftig, 1982; Yan, Zhu & Yang, 2014). For example, cellular entry and trafficking of porcine circoviruses is actin and small GTPase-mediated (Misinzo et al., 2009; Yan, Zhu & Yang, 2014). Myosin is also differentially expressed in subclinical PCV-2 infections and may aid in ATP-dependent intracellular transport of viral particles to the nucleus (Arii et al., 2010; Tomás et al., 2009; Vicente-Manzanares et al., 2009; Xiong et al., 2015).

DEGs involved in post-translational modification, protein turnover, and chaperones—KOG "O" – Intracellular transporters are commonly exploited by vertebrate-associated CRESS-DNA viruses to facilitate entry into the nucleus (Cao et al., 2014; Misinzo et al., 2009). A transcript homologous to Ran (Ras-family related GTP-binding nuclear protein) was overexpressed in transcriptomes with moderate and high LM29173 load, and may be implicated in nucleocytoplasmic transport and regulation of cell cycle progression (Avis & Clarke, 1996; Sazer & Dasso, 2000). Likewise, ubiquitin-conjugating enzyme E2 (UBQ) was overexpressed in libraries with high viral load. This enzyme facilitates covalent attachment of ubiquitin to protein substrates (Liu et al., 2007), and may be exploited by viruses to mis-regulate proteolytic degradation, modify chromatin structure, activate NF-κB and other innate immune mechanisms, or advance G2/M-phase cells into S-phase (Cheng et al., 2014; Gao & Luo, 2006). For example, PCV2 encodes a protein (ORF3) that co-

localizes and interacts with E3 ubiquitin ligase, resulting in upregulation of P53 and induction of apoptotic programs, presumably benefiting viral egress (Liu et al., 2007). Knockdown of ubiquitination conjugating enzymes also stalls cells in the G2/M phase, prohibiting PCV2 from accessing S-phase DNA polymerase necessary for viral propagation (Cheng et al., 2014; Liu et al., 2007).

RT-qPCR supports a haplotype-specific relationship between LM29173 load and amphipod gene expression – Viral load correlated with opposite trends in gene expression (average log-transformed RPKM) between Lake Superior and Lake Michigan transcriptomes in all KOG categories with the exception of "chromatin structure and dynamics" (B), "translation, ribosomal structure and biogenesis" (J), and "energy production and conversion" (C; Fig. 4.4). Unlike Lake Superior libraries, Lake Michigan libraries with high viral load were associated with elevated average RPKM (Fig. 4.4). Because gene expression in organisms with high viral load may be predicated on population-specific characteristics, we identified 29 common genes differentially expressed in both Lake Michigan stations Mi27 and Mi40 (shared DEGs). However, only one contig (NMHC) was both successfully annotated and potentially affiliated with viral infection (Figs. 4.5C and 4.5F). Lake-specific transcriptional profiles confound bulk comparison of gene expression in relation to LM29173 load, and density estimation distributions of individual transcripts indicate that intermediate viral load correlates with increased expression in most KOG classes (Fig. S4.4). These patterns could indicate that CRESS-DNA virus presence has no appreciable impact on gene expression. Alternatively, because amphipods from Lakes Michigan and Superior belong to potentially phenotypically distinct clades (Pilgrim et al., 2009), these results may indicate that response to environmental and microbial stressors is haplotype-specific.

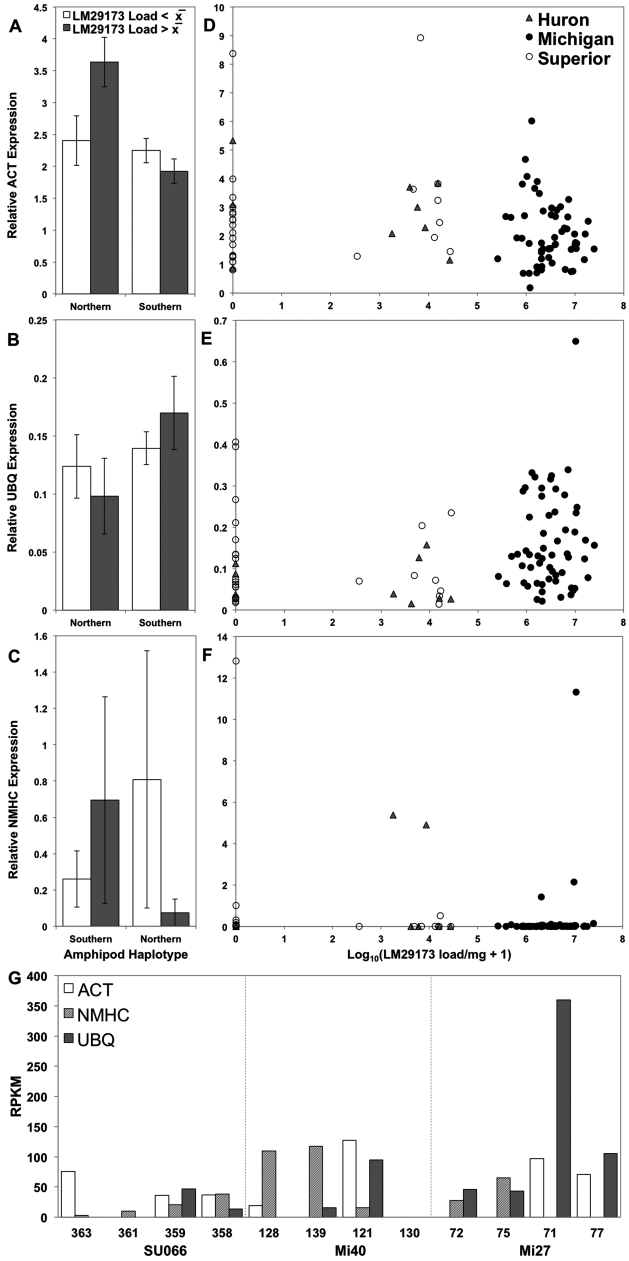

Figure 4.5 | *Relative expression of target genes ACT, UBQ, and NMHC in relation to LM29173 load* – (A–C) Relative expression of target genes β-actin (A; ACT), (B; UBQ) and (C; NMHC) in relation to expression of reference gene elongation factor-1α (EF1A) in specimens from two haplotype clusters (northern and southern) with above (grey) and below (white) average LM29173 copy number (±1 SE). (D–F) Correlation between viral load and relative expression of ACT (D), UBQ (E), and NMHC (F) in relation to EF1A reference gene expression. Quantities of target amplicons were standardized by reference gene EF1A using the following equation: $(Target_{RT} - Target_{NRT})/(EF1A_{RT} - EF1A_{NRT})$, where RT and NRT indicate samples that have been reverse transcribed via Superscript III (Invitrogen, Carlsbad, CA, USA), or not reverse transcribed (no-RT control), respectively. (G) Gene expression (reads per kilobase of transcript per million mapped reads; RPKM) of ACT, NMHC, and UBQ per transcriptome library in each of three stations: Lake Superior station 066 (SU066), Lake Michigan station 40 (Mi40) and Lake Michigan station 27 (Mi27). Libraries are ranked from left to right by increasing LM29173 load.
DOI:10.7717/peerj.3810/fig-5

RT-qPCR quantification of ACT, NMHC, and UBQ confirmed opposite trends in contig expression in correlation with above average viral load among amphipods from Lake Michigan and Superior (Fig. 4.5). Contigs DN12114c1g3i8 (230nt), DN12352c3g10i1 (1,229nt), and DN135c0g1i1 (309nt) exhibited nucleotide sequence similarity to ACT from penaeid blue shrimp (*Litopenaeus stylirostris*; BLASTx, e-value 6×10^{-50}), NMHC from freshwater amphipods (*Hyalella azteca*; e-value 3.0×10^{-8}), and UBQ from freshwater amphipods (*H. azteca*; e-value 4×10^{-52}), respectively. Relative expression of ACT, NMHC, and UBQ did not significantly correlate with viral load, suggesting that LM29173 does not likely specifically alter transcription of these genes, or the practice of whole-organism RNA extraction obscures cell-specific response(s) to this viral genotype (Fig. 4.5). However, relative expression of target genes varied in relation to amphipod population. Organisms associated with the southern haplotype clade exhibited greater average NMHC and UBQ expression, but diminished average ACT expression in concurrence with high viral copy number, relative to amphipods associated with the northern haplotype clade ($p>0.05$, Welch's t-test for all pairwise comparisons).

Expression of target genes ACT, NMHC, and UBQ was standardized to expression of contig DN11198c0g1i1 (723nt), a homolog of elongation factor 1−α (EF1A) from *H. azteca* (BLASTx, e-value 2×10^{-139}). This constitutively expressed gene has been validated as an invariant internal RT-qPCR control under experimental conditions in decapods (Leelatanawit et al., 2012), and provided adequate reference to the baseline transcriptional activity of *Diporeia*, as expression did not correlate with amphipod wet weight, lake, or LM29173 load (Fig. S4.5). Variability in ACT, NMHC, and UBQ expression may be a result of nonspecific RNA extraction, which confounds assessments of specific impacts(s) of viral presence on single tissue types or cells. Additionally, RT-qPCR cannot detect changes in the intracellular localization of myosin subunits nor the state of polymerization of actin subunits, and additional investigation via microscopy and proteomics may be warranted.

Expression of amphipod innate immunity regulators and effectors – Diporeia transcriptomes

were surveyed for homologs of genes involved in crustacean innate immunity to determine if LM29173 presence correlates with immune-specific gene expression. About 148 homologs (BLASTx e $<1\times10^{-5}$) were identified and exhibited >2-fold differential expression in both Lake Michigan station Mi27 and Mi40. Genes involved in stress response (heat shock or oxidative stress response), immune-specific signaling and post-translational modification, and immune-associated cell structure, mobility, and intracellular trafficking mechanisms were consistently overexpressed in Lake Michigan libraries with high viral load (Fig. S4.6). Correlative evidence that these immune-related genes are overexpressed in association with high viral load does not preclude the possibility of other co-occurring immune demands. Therefore, it is unclear if overexpression of these genes are a product of environmental stress, or if they are specific responses to viral infection.

It remains unclear to what extent LM29173 impacts lake-wide *Diporeia* population dynamics. However, the presence of LM29173 among stable amphipod populations and negligible changes in expression of specific amphipod disease pathways in relation to this viral genotype likely indicate that LM29173 is not solely responsible for *Diporeia* decline in the Laurentian Great Lakes. We stipulate that CRESS-DNA viruses associated with *Diporeia* may play a subtle role in altering amphipod physiology, if any. This observation corroborates data from well-characterized mammalian CRESS-DNA viruses (PCV1; Allan & Ellis, 2000; TTV; Okamoto, 2009), which often manifest asymptomatically in healthy host tissue. We speculate that LM29173, like other CRESS-DNA viruses, may evade host clearance, attenuate innate immune responses, or elicit host tolerance through post-transcriptional or translational gene regulation, ultimately establishing persistent and asymptompatic infections (Brajão de Oliveira, 2015; Okamoto, 2009). This hypothesis may explain the universal prevalence and diversity of these viruses in aquatic ecosystems, as observed by metaviromic sequencing.

4.5 Conclusions **|** In summary, while LM29173 load does not correlate with significant differential expression of specific gene pathways, transcriptional changes in genes involved in several physiological functions, including innate immunity, are detectable and specific to distinct haplotype clades. To our knowledge, this study communicates the first investigation of the transcriptional relationship between invertebrates and associated CRESS-DNA viruses in natural ecosystems. This study also provides several potential transcriptional targets for further investigation of gene/pathway-specific inquiries to determine if the bulk of these novel viruses have little effect on metazoan gene expression or physiology.

4.6 References

Allan GM, Ellis JA. 2000. Porcine circoviruses: a review. *Journal of Veterinary Diagnostic Investigation* 12(1):3-14

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215(3):403-410

Arii J, Goto H, Suenaga T, Oyama M, Kozuka Hata H, Imai T, Minowa A, AkashiH, Arase H, Kawaoka Y, Kawaguchi Y. 2010. Non-muscle myosin IIA is a functional entry receptor for herpes simplex virus-1. *Nature* 467(7317):859-862

Auer MT, Auer NA, Urban NR, Auer T. 2013. Distribution of the amphipod *Diporeia* in Lake Superior: the ring of fire. *Journal of Great Lakes Research* 39(1):33-46

Avis JM, Clarke PR. 1996. Ran, a GTPase involved in nuclear processes: its regulators and effectors. *Journal of Cell Science* 109:2423-2427

Barbiero RP, Schmude K, Lesht BM, Riseng CM, Warren GJ, Tuchman ML. 2011.Trends in *Diporeia* populations across the Laurentian Great Lakes, 1997–2009. *Journal of Great Lakes Research* 37(1):9-17

Birkett K, Lozano SJ, Rudstam LG. 2015. Long-term trends in Lake Ontario's benthic macroinvertebrate community from 1994–2008. *Aquatic Ecosystem Health & Management* 18:76-88

Bistolas KSI, Jackson EW, Watkins JM, Rudstam LG, Hewson I. 2017. Distribution of circular single-stranded DNA viruses associated with benthic amphipods of genus *Diporeia* in the Laurentian Great Lakes. *Freshwater Biology* 62(7):1220-1231

Brajão de Oliveira K. 2015. Torque teno virus: a ubiquitous virus. *Revista Brasileira de Hematologia e Hemoterapia* 37(6):357-358

Breitbart M, Benner BE, Jernigan PE, Rosario K, Birsa LM, Harbeitner RC,Fulford S, Graham C, W alters A, Goldsmith DB, Berger SA, Nejstgaard JC. 2015. Discovery, prevalence, and persistence of novel circular single-stranded DNA viruses in the ctenophores *Mnemiopsis leidyi* and *Beroe ovata*. *Frontiers in Microbiology* 6:1427

Brucker RM, Funkhouser LJ, Setia S, Pauly R, Bordenstein SR. 2012. Insect innate immunity database (IIID): an annotation tool for identifying immune genes in insect genomes. *PLOS ONE* 7:e45125

Cao J, Lin C, Wang H, Wang L, Zhou N, Jin Y, Liao M, Zhou J. 2014. Circovirus transport proceeds via direct interaction of the cytoplasmic dynein IC1 subunit with the viral capsid protein. *Journal of Virology* 89(5):2777-2791

Cheng S, Yan W, Gu W, He Q. 2014. The ubiquitin-proteasome system is required for the early stages of porcine circovirus type 2 replication. *Virology* 456–457:198-204

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

Döhner K, Sodeik B. 2005. The role of the cytoskeleton during viral infection. *Current Topics in Microbiology and Immunology* 285:67-108

Dunlap DS, Ng TFF, Rosario K, Barbosa JG, Greco AM, Breitbart M, Hewson I.2013. Molecular and microscopic evidence of viruses in marine copepods. *Proceedings of the National Academy of Sciences of the United States of Americ*a 110(4):1375-1380

Eaglesham JB, Hewson I. 2013. Widespread detection of circular replication initiator protein (*rep*)-encoding ssDNA viral genomes in estuarine, coastal and open ocean net plankton. *Marine Ecology Progress Series* 494:65-72

Eastwood JR, Berg ML, Ribot RFH, Raidal SR, Buchanan KL, Walder KR,Bennett ATD. 2014. Phylogenetic analysis of beak and feather disease virus across a host ring-species complex.

Proceedings of the National Academy of Sciences of the United States of America 111(39):14153-14158

Fahsbender E, Hewson I, Rosario K, Tuttle AD, Varsani A, Breitbart M. 2015. Discovery of a novel circular DNA virus in the Forbes sea star, *Asterias forbesi*. *Archives of Virology* 160(9):2349-2351

Kibenge FSB, Godoy MG. 2016. Aquaculture Virology. London: *Academic Press*, Elsevier.

Gao G, Luo H. 2006. The ubiquitin–proteasome pathway in viral infections. *Canadian Journal of Physiology and Pharmacology* 84:5-14

Gardner WS, Nalepa TF, Frez WA, Cichocki EA, Landrum PF. 1985. Seasonal patterns in lipid content of Lake Michigan macroinvertebrates. *Canadian Journal of Fisheries and Aquatic Sciences* 42(11):1827-1832

Guiguer KRRA, Barton DR. 2002. The trophic role of *Diporeia* (Amphipoda) in Colpoys Bay (Georgian Bay) benthic food web: a stable isotope approach. *Journal of Great Lakes Research* 28(2):228-239

Halfon E, Schito N, Ulanowicz RE. 1996. Energy flow through the Lake Ontario food web: conceptual model and an attempt at mass balance. *Ecological Modelling* 86(1):1-36

Hewson I, Eaglesham JB, Höök TO, LaBarre BA, Sepúlveda MS, Thompson PD,Watkins JM, Rudst am LG. 2013a. Investigation of viruses in *Diporeia* spp. from the Laurentian Great Lakes and Owasco Lake as potential stressors of declining populations. *Journal of Great Lakes Research* 39(3):499-506

Hewson I, Ng G, Li W, LaBarre BA, Aguirre I, Barbosa JG, Breitbart M, GrecoAW, Kearns CM, Lo oi A, Schaffner LR, Thompson PD, Hairston NG. 2013b. Metagenomic identification, seasonal dynamics, and potential transmission mechanisms of a *Daphnia*-associated single-stranded DNA virus in two temperate lakes. *Limnology and Oceanography* 58(5):1605-1620

Jackson EW, Bistolas KS, Button JB, Hewson I. 2016. Novel circular single-stranded DNA viruses among an asteroid, echinoid and holothurian (Phylum: Echinodermata) *PLOS ONE* 11(11):e0166093

Labonté JM, Suttle CA. 2013. Previously unknown and highly divergent ssDNA viruses populate the oceans. *ISME Journal* 7(11):2169-2177

Leelatanawit R, Klanchui A, Uawisetwathana U, Karoonuthaisiri N. 2012.Validation of reference genes for real time PCR of reproductive system in the black tiger shrimp. *PLOS ONE* 7(12):e52677

Liu J, Zhu Y, Chen I, Lau J, He F, Lau A, Wang Z, Karuppannan AK, Kwang J.2007. The ORF3 protein of porcine circovirus type 2 interacts with porcine ubiquitin E3 ligase Pirh2 and facilitates p53 expression in viral infection. *Journal of Virology* 81(17):9560-9567

Luftig RB. 1982. Does the cytoskeleton play a significant role in animal virus replication? *Journal of Theoretical Biology* 99(1):173-191

McTaggart SJ, Conlon C, Colbourne JK, Blaxter ML, Little TJ. 2009. The components of the *Daphnia* pulex immune system as revealed by complete genome sequencing. *BMC Genomics* 10:175

Misinzo G, Delputte PL, Lefebvre DJ, Nauwynck HJ. 2009. Porcine circovirus 2 infection of epithelial cells is clathrin-, caveolae- and dynamin-independent, actin and Rho-GTPase-mediated, and enhanced by cholesterol depletion. *Virus Research* 139(1):1-9

Okamoto H. 2009. History of discoveries and pathogenicity of TT viruses. *Current Topics in Microbiology and Immunology* 331:1-20

Pilgrim EM, Scharold JV, Darling JA, Kelly JR. 2009. Genetic structure of the benthic amphipod Diporeia (Amphipoda: Pontoporeiidae) and its relationship to abundance in Lake Superior. *Canadian Journal of Fisheries and Aquatic Sciences* 66(8):1318-1327

Robinson MD, Smyth GK. 2007. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 9(2):321-332

Rosario K, Breitbart M. 2011. Exploring the viral world through metagenomics. *Current Opinion in Virology* 1(4):289-297

Rosario K, Breitbart M, Harrach B, Segalés J, Delwart E, Biagini P, Varsani A. 2017. Revisiting the taxonomy of the family Circoviridae: establishment of the genus Cyclovirus and removal of the genus Gyrovirus. *Archives of Virology* 162(5):1447-1463

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

Rosario K, Schenck RO, Harbeitner RC, Lawler SN, Breitbart M. 2015. Novel circular single-

stranded DNA viruses identified in marine invertebrates reveal high sequence diversity and consistent predicted intrinsic disorder patterns within putative structural proteins. *Frontiers in Microbiology* 6:696

Roux S, Solonenko NE, Dang VT, Poulos BT, Schwenck SM, Goldsmith DB, Coleman ML, Breitbart M, Sullivan MB. 2016. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* 4:e2777

Ruxton GD, Beauchamp G. 2008. Time for some a priori thinking about post hoc testing. *Behavioral Ecology* 19(3):690-693

Sazer S, Dasso M. 2000. The ran decathlon: multiple roles of ran. *Journal of Cell Science* 113:1111-1118

Schmieder R, Lim YW, Edwards R. 2012. Identification and removal of ribosomal RNA sequences from metatranscriptomes. *Bioinformatics* 28(3):433-435

Soffer N, Brandt ME, Correa AMS, Smith TB, Thurber RV. 2013. Potential role of viruses in white plague coral disease. *ISME Journal* 8(2):271-283

Stewart FJ, Ottesen EA, DeLong EF. 2010. Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME Journal* 4(7):896-907

Tomás A, Fernandes LT, Sánchez A, Segalés J. 2009. Time course differential gene expression in response to porcine circovirus type 2 subclinical infection. *Veterinary Research* 41(1):12

United States Environmental Protection Agency. 2012. Quality Assurance Project Plan for the Great Lakes Water Quality Surveys 2008–2012. Washington, D.C.: United States Environmental Protection Agency, Office of Water, Office of Wetlands, Oceans and Watersheds. Appendix B.

Vicente-Manzanares M, Ma X, Adelstein RS, Horwitz AR. 2009. Non-muscle myosin II takes centre stage in cell adhesion and migration. Nature Reviews Molecular *Cell Biology* 10(11):778-790

Wei L, Zhu S, Wang J, Quan R, Yan X, Li Z, Hou L, Wang N, Yang Y, Jiang H, Liu J. 2016. Induction of a cellular DNA damage response by porcine circovirus type 2 facilitates viral replication and mediates apoptotic responses. *Scientific Reports* 6(1):39444

Wells L. 1980. Food of alewives, yellow perch, spottail shiners, troutperch, and slimy and fourhorn sculpins in southeastern Lake Michigan. Washington, D.C.: U.S. Fish and Wildlife Service. Technical Paper no. 98

Xiong D, Du Y, Wang HB, Zhao B, Zhang H, Li Y, Hu LJ, Cao J-Y, Zhong Q, Liu WL, Li MZ, Zhu XF, Tsao SW, Hutt-Fletcher LM, Song E, Zeng YX, Kieff E, Zeng MS. 2015. Nonmuscle myosin heavy chain IIA mediates Epstein–Barr virus infection of nasopharyngeal epithelial cells. Proceedings of the National Academy of Sciences of the United States of America 112:11036-11041

Yan M, Zhu L, Yang Q. 2014. Infection of porcine circovirus 2 (PCV2) in intestinal porcine epithelial cell line (IPEC-J2) and interaction between PCV2 and IPEC-J2 microfilaments. *Virology Journal* 11(1):193

CHAPTER 5

CONCLUSION

5.1 Review of aims **|** Single stranded DNA (ssDNA) viruses have been investigated since the isolation of φX174 in the mid-1920s (Sertic & Bulgakov, 1935; Székely & Breitbart, 2016). A genomic revolution emerged from target-independent, massively parallel sequencing and congruent *in silico* annotation technologies, fueling a generation of ssDNA virus discovery among non-model organisms. Without accounting for downstream ecological impacts, ssDNA viruses associated with aquatic metazoans may be hypothetically responsible for influencing community structure through mass mortality events, carbon provisioning to upper trophic levels, changes in animal reproductive rate, and mediation of zooplankton-directed macronutrient consumption or turnover (herbivory/detritivory; Munn, 2006). Largely due to their environmental ubiquity and genomic attributes allowing for artificial enrichment in viromes, circular *rep*-encoding ssDNA (CRESS-DNA) viruses have been identified in a milieu of species in the last year alone, ranging from beetles (Rosario et al, 2018) to gila monsters (Somayaji et al, 2018). Despite their detection in viromes, metagenomes, and transcriptomes, the paradigms that govern CRESS-DNA virus diversity, biogeography, distribution, ecology and evolution remain incomplete. Therefore, I aimed to describe the diversity, distribution, and ecological effects of CRESS-DNA viruses affiliated with microcrustaceans spread across biogeographic and phylogenetic spectra. Microcrustaceans constitute a large operational category of ecologically important species on both local and global scales. By capitalizing on high throughput sequencing, viral surveillance, and access to arthropods from a range of ecosystems, this dissertation intended to develop an aggregate portrait of CRESS-DNA virus consortia among microcrustacean viromes, with specific focus on defining the abiotic and biotic parameters that govern CRESS-DNA virus genomic composition and biogeography (**Chapters 1 & 2**) and the intersection between viral presence and the ecology/gene transcription of benthic freshwater amphipods from the Laurentian Great Lakes (**Chapters 2 & 3**).

5.2 Summary of results | Over thirty unique microcrustacean populations (with some overlap in species) were sampled opportunistically from populations around the world, often representing the dominant mesograzer within the respective ecosystem. The cumulative DNA viromes from these organisms were depicted via high throughput sequencing, resulting in the identification and characterization of 215 putatively novel CRESS-DNA viruses (virome composition reported in Appendix I). I therefore support the hypothesis that CRESS-DNA viruses are cosmopolitan members of microcrustacean viral consortia. Due to rapid evolution (i.e. perpetual accumulation of variants) and non-conserved genomic architecture (i.e. oscillation from sense and ambisense ORF orientation, with a proclivity for recombination), taxonomic demarcation and ecological inferences pertaining to these novel CRESS-DNA viruses posed particular challenges. However, nonalignment-based methods, including k-mer signatures (oligomer similarity), %GC content, and codon usage patterns, proved useful in comparing and characterizing these novel CRESS-DNA viral genomes. Genomic content confirmed that microcrustacean-associated CRESS-DNA viral genotypes identified computationally were similar to known invertebrate-associated ssDNA viruses, and amphipod/isopod (malacostracan) associated CRESS-DNA viruses exhibit limited codon repertoires and significant bias towards greater %GC content, providing proof-of-concept for future taxonomic delineation and the potential for host-virus pairing.

Objective 1: Evaluating parameters that modulate CRESS-DNA viral diversity or biogeography among microcrustacean communities - Sampling microcrustaceans from a range of ecosystems and genera allowed comparison of viral genotypes across metazoan phylogeny and geographic distance. CRESS-DNA viral genotypes shared greater similarities in measures of genomic composition (ranging from average nucleotide identity to relative synonymous codon usage biases) when affiliated with similar metazoan taxa or collection sites. Though purely correlative, I hypothesize that these measures of biogeography and niche characteristics (e.g. marine/freshwater differences, benthic/pelagic differences, levels of ecosystem disturbance, etc) conceivably indicate habitat-/host-driven selection. However, it is important to note the statistical interaction effects

between metazoan and biogeography, with microcrustaceans often logically limited to a specific habitat. Furthermore, microcrustaceans sharing niches and habitat types often shared similarities in variant accumulation, indicating that biogeography may play a role in the composition of microdiverse viral populations. (**Chapter 2**)

Objective 2: Biogeographic distribution of CRESS-DNA viruses among microcrustacean populations - To determine CRESS-DNA virus distribution among putative microcrustacean host populations, this study explored the biogeography and phylogeny of three viral genotypes affiliated with lacustrine amphipods from the Laurentian Great Lakes (LM29173, LM122 and LH481). Genotypes LM122 and LH481 were present in low average copy number, indicating that these may have secondary or tertiary hosts. Further investigation specifically focused on the relationship between CRESS-DNA viral genotype, LM29173, metazoan gene expression, and progression of population decline of the benthic amphipod, *Diporeia spp.,* in two of the Great Lakes (Huron and Michigan). While LM29173 prevalence and viral load was significantly greater among amphipod populations from Lakes Huron and Michigan, it was unclear if genotype distribution was a driving factor of a disease or dysbiosis state influential in population decline, or merely correlative with amphipod haplotype distribution. As LM29173 appeared specific only to a single *Diporeia spp.* haplotype, it appeared that this genotype was host-specific, a finding that juxtaposed previous viromic-based studies describing the expansive host range of CRESS-DNA viruses among arthropod populations (Delwart et al, 2012), but aligned with read-recruitment patterns among microcrustacean viromes indicating the possible specificity of genotypes among putative arthropod hosts.

Further read recruitment studies (**Chapter 2**) indicated very few CRESS-DNA viral genotypes comprised the majority of recruited reads within viromes, suggesting that, while present in a wide range of invertebrates, these viruses may have secondary hosts or represent detection of non-replicative genotypes. To confirm that CRESS-DNA viruses do utilize microcrustaceans as cellular hosts, publicly available metazoan genomes were queried, resulting in the identification of >300 CRESS-DNA virus-like endogenous sequences, providing a paleovirological record that *rep* ORF-

bearing sequences likely infected crustacean/mesozooplankton hosts at some point in evolutionary history (Metegnier et al, 2015; Thézé et al, 2014). Collectively, I submit that CRESS-DNA viruses exhibit a wide biogeographical distribution among microcrustaceans, but infection may be highly host-specific, indicating that these mesograzers may serve as reservoirs or secondary hosts for CRESS-DNA viral genotypes.

Objective 3: Interaction between CRESS-DNA viral genotype, LM29173, and microcrustacean nutritional & transcriptional profile - Because the mechanism(s) responsible for declines in *Diporeia spp.* abundance in Lakes Huron, Michigan and Ontario remained unknown, we surveyed amphipods from several populations to investigate viral distribution in relationship to amphipod nutritional quality and gene expression (eg. lipid content, C:N stoichiometry, transcriptional profile). However, despite widespread detection, this study found little evidence of physiological impact of CRESS-DNA viral exposure in microcrustacean hosts (Chapter 4). For example, viral load of LM29173 shared a small, but ultimately insignificant correlation in percent lipid and molar C:N, with likely inconsequential impacts on nutritional quality. This dissertation provides the first transcriptional profile of *Diporeia spp.,* identifying over two thousand transcripts differentially expressed in the presence of high LM29173 load (particularly those associated with DNA replication, cellular structural component biogenesis, and posttranslational modifications), though it remains unclear if any elicit a specific, acute immune response from the microcrustacean host.

5.3 Synthesis **|** The rush to sequence the global virome has heralded an age of ssDNA viral discovery, revealing the pervasive nature, convoluted ancestry, and evolutionary potential of those with circular genomes. Among arthropods, CRESS-DNA viruses appear to be everywhere at once, yet highly specific to certain hosts. Likewise, their replication appears to be extremely processive, yet infections potentially asymptomatic. Their evolution appears to be rapid, yet their genomes irreducibly small. If these genomes truly represent a single group of microcrustacean-associated viruses, these pluralistic states seem implausible. Therefore, it is important to examine the unknown

elements of eukaryote-associated ssDNA viral ecology.

Host range - In this dissertation, I posit that although CRESS-DNA viruses appear to be cosmopolitan, they may be highly host specific, indicating that arthropods serve as reservoirs of viral genetic information. These arthropods can be indiscriminate feeders, concentrating large volumes of organic matter from the water column or filtering sediment for microbial resources and potentially accruing associated viral particles. Furthermore, arthropods with overlapping niches may have overlapping diets or carry multiple, similar unicellular epibionts (Johnson et al, 2012), resulting in similar viromes. ssDNA viral capsids are resistant to an assortment of environmental stressors (Allan et al, 1994). Therefore, I predict that the widespread detection of novel arthropod-associated CRESS-DNA viruses may be a product of highly effective dispersal, long residence times, continued circulation through the water column, long duration of suspension (low sedimentation coefficient and molecular weight), and repetitive consumption and excretion without degradation (allowing for vectored transmission), rather than large host range.

This hypothesis is supported by the host specificity of amphipod-associated CRESS-DNA viral genotypes via viral surveillance and computational cross-recruitment of novel genotypes. However, this hypothesis is often difficult to integrate within the wider contemporary context of ssDNA virus representations, which dictate that ssDNA viruses of non-model systems are widely infectious and possibly even the source of zoonoses. Instead, I predict that ssDNA viruses are limited in their infectivity, and caution should be exercised when differentiating between vectoring/reservoir species and infected arthropods. Host switching may be explained by rapid rates of evolution during active infection (low fidelity replication) and predisposition for recombination – particularly via site-specific integration in host genomes with excision in a recombined state. Furthermore, among similar ssDNA viruses with linear genomes (feline parvoviruses; Modrow et al, 2013), the alteration of only three nucleotides altering three amino acid residues within the major capsid protein are capable of radically changing host range, indicating that host-switching may be commonplace among these mutable genomes.

Pathogenicity - The putative presence of CRESS-DNA viruses among so many arthropods undoubtedly raises questions about their pathogenicity and potential impact on metazoan physiology and population ecology. Due to the accessibility of viromics and rapid turnaround of sequence data, CRESS-DNA viruses are often serendipitously discovered among these non-model organisms and ascribed pathogenicity in correlation with the onset of disease symptoms. However, despite efforts to cultivate arthropod-associated CRESS-DNA viruses, artificially infect insects and crustaceans, or otherwise validate Koch's (or River's) postulates, the pathogenicity of CRESS-DNA viruses remains unknown. Therefore, quantitative description of arthropod-associated CRESS-DNA viral impacts is inconsistent, at best.

However, we may be able to glean insight from the biology of vertebrate-associated CRESS-DNA viruses. Although significant maladies have been associated with the presence of some CRESS-DNA viruses, the vast majority of highly prevalent vertebrate CRESS-DNA viruses exhibit minimally pathogenic, or "commensal" biology. For example, Torque teno, and Torque teno midi/mini anelloviruses (not CRESS-DNA viruses, but encoding an ORF with *rep*-like structure and utility) have coexisted as infectious agents of humans since their detection in the 1990s, and likely throughout human history (Hino & Miyata, 2007). While correlated with a variety of disease states, including hepatitis myopathies, cancer, lupus, and general immunosuppression, these viruses have not been implicated as the direct causative element in any (Hino & Miyata, 2007; Nishizawa et al, 1997). While model circovirus, PCV2 is a well known, agroeconomically relevant pathogen, its counterpart, PCV1 is considered nonpathogenic and differs in <20% of nucleotide identity (Hamel et al, 1998; Olvera et al, 2007).

As obligate parasites, any actively replicative virus is pathogenic by definition - requisitioning resources and potentially useful enzymes from the infected cell for the basic cost of maintenance is detrimental. However, the cost(s) of infection may be subtle, with infections serving as "commensal" or neutral bystanders. They may impose little overall impact on nonconsumptive mortality or cause only delicate changes in arthropod transcriptional profile. Likewise, the cost of

viral infection may be offset by significant mutualistic services provided to arthropod host, resulting in a net benefit to the organism (Jagdale et al, 2018; Roossinck, 2019). There is often a phenotypic advantage for a virus to preserve the health of the host (or at least the integrity of its polymerases) to assist in the replication of the virus itself. In this sense, CRESS-DNA viruses could be evolutionarily "strategically nonpathogenic." This evolutionary strategy would indicate that it would be more valuable to integrate estimates of net metabolic drawdown and immunosupression from viral infection, rather than non-consumptive mortality elicited by CRESS-DNA viruses, into models of aquatic ecosystem energy flow. However, without clear evidence of experimental infection, cell culture, or innate immune response of a representative range of CRESS-DNA virus genera, this deduction is purely speculative and a quantitative measure of this metabolic drawdown or cost of secondary infection remain unknown.

Evolution - Without significant sequencing of non-model systems to both expand and substantiate existing viral clades, it is difficult to predict the provenance of many "orphan" viruses (i.e. those representing a single species within a clade). Indeed, without a better understanding of the total diversity of these CRESS-DNA viruses, it is unclear if current cladistics are too restrictive or too inclusive. The proposed polyphyletic origin of many ssDNA viruses have led many to propose that ssDNA viruses are extant proof of a pivotal junction in biological history, supporting the hypothesis of an evolutionary continuum that includes both viruses and selfish genetic elements such as transposons, plasmids, and other cellular replicons (Krupovic et al, 2008, 2009; Gorbalenya et al, 1990; Laufs et al, 1995). In this hypothesis, capsidless genetic elements may be derived from viruses that have lost structural (i.e. capsid) proteins, whereas viruses may offer testament to the acquisition of universally conserved structural genes by plasmids or transposons to ensure transmission between possible host cells in unforgiving environments or where hosts may be scarce. Eukaryotic CRESS-DNA viruses, in particular, appear to embody the intermediates of a modular, nonlinear evolutionary trajectory.

For example, rolling circle replication is a strikingly common mechanism among circular

nucleic acids, ranging from viroids to RNA. Many CRESS-DNA viral *rep* ORFs share significant sequence similarity to plasmid *rep*-like ORFs and other mobile genetic elements (Chandler et al, 2013; Rosario et al, 2015). Rep proteins of microcrustacean-associated CRESS-DNA viruses (as well as most eukaryotic CRESS-DNA viruses) carry critical SF3 helicase domains commonly observed among algal plasmids (Krupovic et al, 2008; Gibbs & Weiller, 1999). This may indicate that these *rep* ORFs may be derived from algal plasmids or vice versa. Likewise, small ssDNA virus capsids are comprised of a distinct, widely recognized β-barrel fold, an extremely prevalent structural feature among nearly all icosahedral viruses (Kazlauskas et al, 2017; Chandler et al, 2013, Krupovic et al, 2008). It was long thought that this structural fold was simply the most parsimonious for icosahedral viruses, and the presence of these structurally related proteins in both +(ss)RNA viruses and CRESS-DNA viruses was a product of convergent evolution. However, Krupovic et al (2008, 2009) and Kazlauskas et al (2017) point out that icosahedral viruses of all sizes contain conserved secondary structures that differentiate between viral families (or even genera), leaving clear evolutionary landmarks. Therefore, conservation of these "markers" in CRESS-DNA virus capsid residues which subtly alter structure in ways unique to RNA viruses may provide evidence to support a link between +(ss)RNA viruses and CRESS-DNA viruses.

Together, these similarities in *rep* and *cap* generate three scenarios portraying the origin of CRESS-DNA viruses, where (1) RNA and DNA mobile genetic elements independently acquired capsids which evolved convergently, (2) RNA viruses gradually transitioned into DNA viruses through a reverse transcribing evolutionary intermediate, and (3) DNA selfish genetic elements acquired a capsid from a contemporary ssRNA virus (termed the "RNA-to-DNA jump" scenario). ssRNA viruses notoriously capture nonviral DNA, ranging from transposons to sheared plasmids, perhaps lending credibility to this third hypothesis (Routh et al, 2012). As previously discussed, CRESS-DNA virus *rep* endonuclease domains may facilitate recombination (Lefeuvre et al, 2009), potentially supporting the possible fusion between mobile genetic elements and ssRNA-virus encoded *cap* ORFs through a cDNA intermediate at the RNA transcript level ("copy-choice"

recombination via an RdRp template switching), or via tandem integration as endogenous viruses (Krupovic, 2012). Indeed, because eukaryotic CRESS-DNA viral *rep* ORFs appear polyphyletic, clustering closely with microbial and algal plasmid-encoded *rep* ORFs rather than within other ssDNA viral *rep* ORFs, perhaps *rep* may have been acquired multiple times (Krupovic, 2012). Evidence of this form of acquisition could be provided in the form of chimeric viruses, such as RHDV ("RNA-DNA hybrid virus;" Lassen Volcanic National Park, USA; Diemer & Stedman, 2012). This genotype contains a *rep* ORF similar to that of a circovirus and a *cap* ORF similar to that of a tombusvirus – a positive-sense ssRNA virus most similar to genotypes predicted to infect oomycetes. A similar chimeric virus genotype was identified in the course of this dissertation research within isopod epibionts from pacific intertidal nereocystis beds, sharing genomic features with circoviruses and tombusviruses (see Appendix III; Bistolas et al, 2017). This clade of chimeric viruses, tentatively named the "*Cruciviridae*" is expanding with greater recognition within aquatic viromes.

Ultimately, it is not inconceivable that aquatic invertebrates - particularly microcrustaceans – may serve as one of many evolutionary hotbeds and sources of novel ssDNA viral clades for several reasons. Microcrustaceans form close associations with unicellular eukaryotes as epibionts and consumed biomass, and have a predilection to filter (and therefore concentrate) large quantities of water, including suspended RNA and DNA virus-like particles, and may therefore serve as sites for viral recombination. Furthermore, the typically fast growth rate of invertebrates (and therefore the potential for rapid cellular replication in growth phase), and marginal innate immunity among some invertebrate species may provide a medium for intracellular recombination or gene capture, potentially evidenced by the high density of endogenized CRESS-DNA viral elements observed among arthropod genomes. Lastly, the sheer global abundance of aquatic invertebrates likely contributes to origin and dispersal of mesograzer-affiliated viruses.

5.4 Future Directions **|** This study barely scratches the surface of the potential exploration of

microcrustacean-associated CRESS-DNA viruses. The field of viromics is expanding in leaps and bounds, germinating opportunities for the detailed investigation of new viral groups in previously unexplored ecosystems. There are several ongoing attempts to slowly resolve the universal taxonomy of DNA viruses through the development of online repositories. At minimum, any additional sampling of non-model, environmental systems will aid this enterprise and help in settling ssDNA viruses within the larger context of alphasatellites, mobile genetic elements, jelly-roll encoding ssRNA viruses, etc. Major gaps in our knowledge pertain to the **(1)** basic ecology of CRESS-DNA viruses, **(2)** their relationship to infected invertebrate hosts, and **(3)** the secondary impact these infections may have on ecosystem function as a whole. However, if the CRESS-DNA viral genotypes identified in this dissertation are to be utilized to determine pathogenicity or as other targets of investigation, it is first critical that they are validated as bona fide infectious agents in microcrustacean hosts. While computationally assembled, sequenced, and surveyed throughout microcrustacean populations, it is difficult to say with absolute confidence that a single genotype represents a virion without visualizing this virion in pure culture and assessing its pathogenicity following Koch's postulates. This is a difficult task, as continuous culture of crustacean cell lines do not yet exist, and creating pure isolates of independent viruses while avoiding inadvertent experimental evolution is currently a rarity. However, cultivation of invertebrate CRESS-DNA viruses will be a boon for the field, allowing a more targeted, experimental, hypothesis-driven approach to understanding ssDNA viral clades.

Viral ecology - Despite their widespread nature, the basic tenants of CRESS-DNA viral ecology remain essentially unknown. As a function of their ssDNA genomic lability and replication mechanisms, these viruses evolve at astounding rates, with multiple genotypes existing in microdiverse ("quasispecies") swarms. Understanding the selection dynamics in a controlled setting via experimental evolution may lend insight into the descent of extant viruses. Future studies on the transmission mechanisms (vertical, horizontal, vectored) within ecosystems may further elucidate the cohesion and configuration of viral populations and the extent of their distribution. For example, if

arthropod-associated CRESS-DNA viruses are vertically transmitted, do sequential populations reflect sequential genotypic or phenotypic changes, and do associated generational bottlenecks drive selection towards specific changes? It is probable that arthropod ecology primarily governs viral ecology, but parsing the degree to which host parameters (such as sex, lifespan, diet, mobility, senescence periodicity, reproductive maturity/ontogeny, or population density) leave lasting impressions on viral ecology.

Infection dynamics - While we know the tropism of porcine circoviruses, how virions are transported to the nucleus, and how they are replicated while there, we have no comparable model for invertebrate-associated CRESS-DNA viruses. This is a significant deficit when attempting to survey invertebrate populations, as it is unclear which tissues to collect or what members of a population are most likely to be infected. Basic quantitative knowledge about infection processes is wholly unknown for any invertebrate-associated CRESS-DNA virus, including multiplicity of infection, infectivity, target receptor, intracellular transport and associated kinetics, replication site/mechanism, or burst size/basic reproductive number. Visualization of targeted genotypes via *in situ* hybridization or other methods may be a valuable first step in identifying viruses of interest. Likewise, examining change in gross morphology, histopathology, and ontogeny in correlation to viral load may preface more traditional investigations requiring cultivation, synthesis of viral proteins and associated antibodies, and analysis of host physiology on a cellular level.

Ecosystem impacts - Finally, despite their utility as molecular tools, the ability of ssDNA viruses to elicit mortality, necroses, or other deficits among invertebrates (with secondary impacts on ecosystem function) is still un-quantified, and will likely remain unresolved until Koch's postulates are directly examined. Costs of infection may extend beyond mortality, ranging from metabolic and evolutionary costs of immunity or resistance to specific pathogens, to influencing host feeding habits. For example, infection may alter microcrustacean detritivory, altering the flux of oxidized inorganic nutrients in sediment. Future studies which develop computational (ecosystem), mesocosm (population), experimental microinjection (organismal), and culture-based (cellular) methods to

experimentally test the response of CRESS-DNA virus infection on microcrustacean ecology will provide a transformative perspective on the contribution of ssDNA viruses to arthropod mortality achieving the overarching goal to quantify the intersection between viral infection and microcrustacean ecology.

5.5 Concluding remarks **|** This dissertation supports previous observations of the ubiquity of CRESS-DNA viruses in the natural world, but questions assertions of their natural ability to infect all hosts to inflict disease or drive patterns of nonconsumptive mortality in highly abundant aquatic microcrustaceans. Instead, this endeavor frames these CRESS-DNA viruses as potential evolutionary stepping-stones for the origin of many clades of ssDNA viruses, drivers of viral diversity, and exploitable resources for the investigation of ssDNA virus phylogenomics, biogeography, population dynamics, etc. I hope that these studies will be relevant to those integrating viral impacts into models of aquatic nutrient turnover and food web dynamics, those utilizing CRESS-DNA viruses as archetypes for viral evolution, and those evaluating the parameters that influence viral spatial distribution. At minimum, these discoveries should provide the basis for the identification of additional novel CRESS-DNA viruses from ecosystems and organisms beyond aquatic microcrustaceans through queries in public nucleic acid databases. This data and the depiction of the methodological challenges encountered in this study intends to inform, in some small measure, the opportunities facing the next wave of ssDNA virus hunters.

5.6 References

Allan GM, Phenix KV, Todd D, McNulty MS. 1994. Some biological and physico-chemical properties of porcine circovirus. *Zentralbl Veterinary med B*. 41(1):17-26. doi:10.1111/j.1439-0450.1994.tb00201.x.

Bistolas KSI, Besemer RM, Rudstam LG, Hewson I. 2017. Distribution and inferred evolutionary characteristics of a chimeric ssDNA virus associated with intertidal marine isopods. *Viruses*. 9(12): 361. doi:10.3390/v9120361.

Chandler M, de la Cruz F, Dyda F, Hickman AB, Moncalian G, Ton-Hoang B. 2013. Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. *Nat Rev Microbiol*. 11(8):525-38. doi:10.1038/nrmicro3067.

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 Res.* 164(1-2):114-21. doi:10.1016/j.virusres.2011.11.021.

Diemer GS, Stedman KM. 2012. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. *Biol Direct*. 7:13. doi:10.1186/1745-6150-7-13.

Gibbs MJ, Weiller GF. 1999. Evidence that a plant virus switched hosts to infect a vertebrate and then recombined with a vertebrate-infecting virus. *Proc Natl Acad Sci*. 96:8022-8027. doi:10.1073/pnas.96.14.8022.

Hamel AL, Lin LL, Nayar GP. 1998. Nucleotide sequence of porcine circovirus associated with postweaning multisystemic wasting syndrome in pigs. *J Virol*. 72(6):5262-7.

Hino S, Miyata H. 2007. Torque teno virus (TTV): current status. *Rev Med Virol.* 17(1): 45-57. doi:10.1002/rmv.524.

Johnson S, Allen M, Fylling M. 2012. Zooplankton of the Atlantic and Gulf Coasts: a guide to their identification and ecology. Baltimore: *Johns Hopkins University Press*. Project MUSE.

Kazlauskas D, Dayaram A, Kraberger S, Goldstien S, Varsani A, Krupovic M. 2017. Evolutionary history of ssDNA bacilladnaviruses features horizontal acquisition of the capsid gene from ssRNA nodaviruses. *Virology*. 504:114-121. doi:10.1016/j.virol.2017.02.001.

Krupovic M, Bamford DH. 2008. Virus evolution: how far does the double β-barrel viral lineage extend? *Nat Rev Microbiol*. 6:941–948. doi:10.1038/nrmicro2033.

Krupovic M, Ravantti JJ, Bamford DH. 2009. Geminiviruses: a tale of a plasmid becoming a virus. *BMC Evol Biol*. 9:112. doi:10.1186/1471-2148-9-112.

Krupovic M. 2012. Recombination between RNA viruses and plasmids might have played a central role in the origin and evolution of small DNA viruses. *Bioessays*. 34(10):867-70. doi:10.1002/bies.201200083.

Lefeuvre P, Lett JM, Varsani A, Martin DP. 2009. Widely conserved recombination patterns among

single-stranded DNA viruses. *J Virol*. 83(6):2697-707. doi:10.1128/JVI.02152-08.

Metegnier G, Becking T, Chebbi MA, Giraud I, Moumen B, Schaack S, Cordaux R, Gilbert C. 2015. Comparative paleovirological analysis of crustaceans identifies multiple widespread viral groups. *Mob DNA*. 6:16. doi:10.1186/s13100-015-0047-3.

Modrow S, Falke D, Truyen U, Schätzl H, 2013. Viruses with a single-stranded DNA genome. In: Molecular Virology. *Springer*, Berlin, Heidelberg: 875-918. doi:10.1007/978-3-642-20718-1_20.

Munn CB. 2006. Viruses as pathogens of marine organisms—from bacteria to whales. *J Mar Biol Assoc UK*. 86(3): 453-467. doi:10.1017/S002531540601335X.

Nishizawa T, Okamoto H, Konishi K, Yoshizawa H, Miyakawa Y, Mayumi M. 1997. A novel DNA virus (TTV) associated with elevated transaminase levels in posttransfusion hepatitis of unknown etiology. *Biochem Biophys Res Commun*. 241(1):92-7. doi:10.1006/bbrc.1997.7765.

Olvera A, Cortey M, Segales J. 2007. Molecular evolution of porcine circovirus type 2 genomes: phylogeny and clonality. *Virology*. 357(2):175–185. doi:10.1016/j.virol.2006.07.047.

Rosario K, Duffy S, Breitbart M. 2012. A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Arch Virol*. 157(10):1851-71. doi:10.1007/s00705-012-1391-y.

Rosario K, Mettel KA, Benner BE, Johnson R, Scott C, Yusseff-Vanegas SZ, Baker CCM, Cassill DL, Storer C, Varsani A, Breitbart M. 2018. Virus discovery in all three major lineages of terrestrial arthropods highlights the diversity of single-stranded DNA viruses associated with invertebrates. *PeerJ*. 11(6):e5761. doi:10.7717/peerj.5761.

Routh A, Domitrovic T, Johnson JE. 2012. Host RNAs, including transposons, are encapsidated by a eukaryotic single-stranded RNA virus. *Proc Natl Acad Sci U S A*. 109(6):1907-12. doi:10.1073/pnas.1116168109.

Sertic V, Bulgakov N. 1935. Classification et identification des typhi-phage. *C R Soc Biol*, *Paris*. 119: 1270–1272.

Somayaji V, DeNardo D, Wilson Sayres MA, Blake M, Waits K, Fontenele RS, Kraberger S, Varsani A. 2018. Genome sequence of a single-stranded DNA virus identified in gila monster feces. *Microbiol Resour Announc*. 7(7): e00925-18. doi:10.1128/MRA.00925-18.

Székely AJ, Breitbart M. 2016. Single-stranded DNA phages: from early molecular biology tools to

recent revolutions in environmental microbiology. *FEMS Microbiol Lett*. 363(6):fnw027. doi:10.1093/femsle/fnw027.

Thézé J, Leclercq S, Moumen B, Cordaux R, Gilbert C. 2014. Remarkable diversity of endogenous viruses in a crustacean genome. *Genome Biol Evol*. 6(8):2129-40. doi:10.1093/gbe/evu163.

APPENDIX I

SUMMARY OF MICROCRUSTACEAN VIROME COMPOSITION

AI.1 DNA viruses affiliated with aquatic microcrustaceans, excluding ssDNA genotypes | Bacteriophage comprise the majority of virus-like sequences in metazoan viromes. This is unsurprising, as metazoans host their own army of microbes and their 'phage. 44.4% of remaining virus-like contigs shared sequence similarity to virophage (n=19), satellite viruses (n=11), or unclassified virus-like sequences associated with taxonomic orders not associated with a known host (n=4,022). Virophage were predominantly identified within marine copepod/amphipod and freshwater brachiopod viromes, while other satellite viruses were exclusively assembled in freshwater amphipod viromes.

Despite annotation protocols, recognition of the extent of DNA viral diversity remains limited by the inability to identify viral sequences and determine their association with predicted host taxa. These functionally and taxonomically un-annotated sequences comprised 35.5-98.0% of contigs per virome, impeding sensitive and accurate characterization of community structure and evolution. It is unclear what ratio of these sequences may be viral. The magnitude of unannotated sequences highlights the value of comprehensive sequencing of non-model systems to provide comparative database queries, particularly identifying the need for investigations of viral infectivity, tropism, biogeography, and pathogenicity to aid viromic sequence annotation and host-virus pairing. Annotation of the remaining contigs provided preliminary perspectives on viral community composition contextualized by phylogenetic and biogeographic parameters and resulted in several discoveries of novel virus-like sequences.

A significant (and often overlooked) component of viromes involves metazoan-associated viruses that lack clear pathogenicity or correlation with disease/dysbiosis states, with particular deficiencies in the microcrustacean viromes of wild-caught populations relative to model or economically relevant species. Among virus-like contigs, only 2.96% were not identified as

bacteriophage, and were categorized by genome type. Many of these putative viruses may infect secondary hosts (i.e. non-microcrustacean, epibiont or gut content related cells). For example, contigs similar to Nucleocytoplasmic Large DNA Viruses (NCLDVs) were particularly common among microcrustacean-associated viromes, especially within copepod and isopod sequencing libraries, including members of the *Phycodnaviridae*, *Iridoviridae*, and *Poxviridae*. NCLDVs are often excluded from viromes via size filtration. However, it is evident from the level of host genome contamination that non-encapsidated nucleic acids is not fully digested in the process of virome pre-sequencing preparation, perhaps allowing sequencing of replicative, non-encapsidated NCLDV genomes. These NCLDVs often horizontally acquire host-like genes, and may represent annotation. Of note, representative NCDLVs are often associated with algal or protozoan hosts, rather than metazoans, and may be predominantly acquired via diet or epibiont association. Freshwater amphipods harbored a marginally greater β-diversity of putative viral groups between libraries (H'= 4.16, Shannon-Weaver; normalized to metazoan-associated DNA viral contig quantity), while marine isopods exhibited greater α-diversity within groups relative to other microcrustacean libraries (average H'=4.17, Shannon-Weaver). Overall, net DNA viral diversity did not appear to be structured by microcrustacean biogeography, taxonomy, or ecosystem, although diversity measures were obscured by limited annotation.

Paired, trimmed high-throughput sequencing reads were recruited to virus-like contigs (90% nucleotide identity over 90% length of the read; stringency imposed to prevent cross-recruitment between contigs), serving as a rough proxy for contig representation within virome libraries, though linear bias between this proxy and representation within natural populations is often subject to a variety of factors. Of the >38.5 million quality-controlled reads recruited to annotated contigs (n=7,553), only approximately 10 million were associated with unique virus-like contigs. Moreover, there was significant variation in the total percentage of reads within individual libraries recruited to non-bacteriophage virus-like contigs, ranging from 0-78.4%. Six libraries were further excluded from annotation against target ssDNA viruses, as read recruitment was insufficient for analyses.

Although there was a greater sum of unique dsDNA virus-like contigs relative to ssDNA virus-like contigs, ssDNA viruses recruited >11.7x the proportion of reads per nucleotide (3.17 reads nt^{-1}) in comparison to dsDNA virus-like contigs (0.27 reads nt^{-1}; Fig. A1). This observation is likely an artifact of comparatively larger dsDNA viral genomes. Processes essential to viromic sequencing of low biomass samples also enrich for small, circular ssDNA templates; isothermal multiple displacement amplification (MDA) with random hexamer priming was utilized prior to sequencing, employing a Φ29 DNA polymerase that preferentially amplifies ssDNA templates. While MDA nullifies quantification of ssDNA viral genomes relative to other DNA virus-like templates and inhibits measures of diversity between genomic groups, this method confirmed that ssDNA genomes are regular viral constituents across a broad spectrum of microcrustaceans viromes, although these viruses may exist at unknown densities (i.e. genomic copy number). Furthermore, MDA allows greater strength in identification and quantification of within-contig divergence and provides improved accuracy when comparing genomic structure of novel circular ssDNA viral genomes.

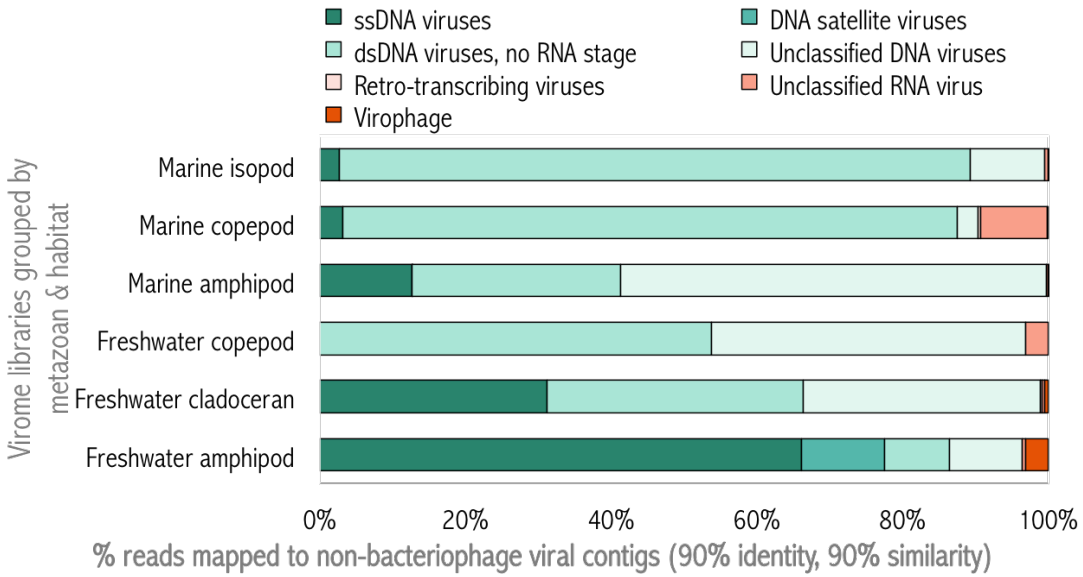

Figure AI.1 | *Read recruitment to virus-like contigs in microcrustacean viromes* - Trimmed reads were assembled *de novo* into contiguous sequences (contigs) and annotated by tBLASTx comparison to the NCBI nonredundant database. The distribution of reads mapped to annotated contigs indicate that ssDNA viruses comprise a substantial proportion of aquatic microcrustacean viral consortia.

AI.2 RNA viruses of Diporeia spp. derived from amphipod transcriptomes | Sequences resembling RNA viruses were detected among *de novo* assembled transcripts from the benthic freshwater amphipod, *Diporeia spp*. Illumina sequenced cDNA reads from 12 transcriptomes were pooled and assembled via an isoform-sensitive algorithm (Trinity on Galaxy platform per default parameters; National Center for Genome Analysis Support, Indiana University Pervasive Technology Institute) and annotated via tBLASTx comparison against the NCBI non-redundant database (Altschul et al, 1990). 15 unique contiguous sequences (contigs) shared similarity to RNA-dependent RNA polymerases (RdRps; 80%) or ORFs hypothetically encoding nucleocapsid or viral glycoproteins (20%), which often serve as hallmark genes for RNA-encoding viruses lacking an intermediate DNA stage. Specifically, negative-sense ssRNA viruses-like contigs, including members of the *Bunyaviridae*, *Rhabdoviriade*, or *Orthomyxoviridae* outnumbered *nodavirus*-like positive-sense ssRNA contigs. Congruently, Mononegavirales-like contigs exhibited virus capping methyltransferase (n=2), RdRp (n=2), nucleocapsid (N-protein; n=1), and spike glycoprotein (G-protein; n=1) specific motifs, while remaining contigs encoded RdRp motifs characteristic of the *Bunyaviridae* (n=2), *Flaviviridae* (n=1), or *Orthmyxoviridae* (n=1). These virus-like sequences may represent either endogenized/co-opted lysogens or extant viral genomes associated with *Diporeia spp*. Differentiating between the two ecoevolutionary states may remain challenging without additional investigation, as sequences flanking putative viral RdRp or structural protein encoding regions were not indicative of cellular genes, and putative ssRNA virus-like contigs were short (average length 1637.1nt), and likely represent fragments of viral genomes. However, 11 contigs were most similar (top whole-contig BLASTx hit, e-value $<10^{-5}$) to viruses associated with invertebrates, nine of which were associated specifically with arthropods including crustaceans, indicating a plausible relationship between amphipod and virus. Common non-crustacean arthropod taxa included Lepidoptera, Odonata, and other unclassified Insecta.

The majority of individual *Diporeia* transcriptomes contained ≤2 RNA virus-like contigs,

illustrated by negligible read recruitment to remaining contigs (<10x average coverage across contig length), suggesting exclusivity of several contigs to specific transcriptomes. However, read recruitment among 12 transcriptomes averaged 2,129.75 reads and contig coverage averaged 62.5x. Contig 116 encoded an ORF resembling an RdRp associated with Beihai bunya-like virus 1 (Shi et al, 2016), and unexpectedly recruited reads from all twelve libraries, ultimately exhibiting the greatest overall coverage. These recruitment patterns may indicate the cosmopolitan nature of Contig 116 among *Diporeia* from Great Lakes ecosystems, or its existence as a virome contaminant associated with collection, preservation, or processing tools. Read recruitment may be influenced by total similarity thresholds (80% nucleotide identity over 80% of read length). 72.1% of total recruited reads were non-perfect matches containing >1 site of variation when compared to the consensus sequence. Furthermore, four *de novo* assembled contigs were identified as two viral "species," via reciprocal BLASTn comparison (defined as an ANI >95% over >98% of contig length) and equivalent quantities of recruited reads (≤ 25%). Intriguingly, read recruitment was distributed such that recruitment from individual libraries was highly biased towards one genotype over the other (i.e. paired genotypes recruited reads from dissimilar libraries). Amphipods representing these libraries were collected within a single season, but were geographically distant, reiterating the rapid rate of ssRNA virus evolution and recapitulating the difficulties in classifying viral species in non-model systems.

AI.3 References

Altschul SF, Gish W, Miller W, Myers EW, Lipman, DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215:403–410. doi:10.1016/S0022-2836(05)80360-2.

Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, Qin XC, Li J, Cao JP, Eden JS, Buchmann J, Wang W1, Xu J, Holmes EC, Zhang YZ. 2016. Redefining the invertebrate RNA virosphere. *Nature*. 540(7634):539-543. doi:10.1038/nature20167.

APPENDIX II

SUPPLEMENTAL INFORMATION

AII.1 Supplementary figures and tables as referenced in Chapter 2

Supplementary figures and tables referenced as "SI Figure/Table", followed by respective number in Chapter 2.

	Database	**Access ID**
Human	NCBI - nucleotide (WGA)	NC_000001-NC_000023
Amphipod	NCBI- nucleotide (EST)	*Abludogammarus, Abyssogammarus, Abyssorchomene, Acanthogammarus, Acanthonotozoma, Allorchestes, Amathillina, Ambasiopsis, Ampelisca, Amphilochius, Amphithoe, Amphithyrus, Ampithoe, Amurocrangonyx, Anamixis, Andaniexis, Andaniopsis, Anisogammarus, Antatelson, Aora, Apherusa, Apohyale, Arrhis, Asprogammarus, Astyra, Atylus, Bactrurus, Baicalogammarus, Baikalogammarus, Barnardiorum, Barrowgammarus, Bathymedon, Bathyporeia, Boeckaxellia, Bogidiella, Brachyuropus, Brandtia, Bubocorophium, Byblis, Calliopius, Caprella, Caprogammarus, Carineogammarus, Carinogammarus, Ceratogammarus, Chaetogammarus, Chydaekata, Colomastix, Comatogammarus, Cornugammarus, Corophiomorphus, Corophium, Crangonyx, Crymostygius, Crypturopus, Cyamus, Cyllopus, Dik-524F, Dik-8179S, Dik-9172Q, Dikerogammarus, Diplacanthus, Diporeia, Dorogostaiskia, Dulichia, E.verrucosus, Echiniphimedia, Echinogammarus, Eogammarus, Eohaustorioides, Eoniphargus, Epimeria, Epimeriella, Ericthonius, Eriopisa, Eulimnogammarus, Eupronoe, Eurythenes, Eusirus, Falklandia, Fuegophoxus, Gammaracanthus, Gammarella, Gammarellus, Gammaridae, Gammaropsis, Gammarus, Garjajewia, Gmelinoides, Grandidierella, Hakonboekia, Haploops, Haustorius, Hemityphis, Hurleya, Hyale, Hyalella, Hyperia, Hyperietta, Hyperiidea, Hyperioides, Ingolfiella, Iphimediella, Jassa, Jesogammarus, Kruptus, Lepidepecreum, Leptocheirus, Lestrigonus, Leucothoe, Lilleborgia, Linevichella, Locustogammarus, Lucioblivio, Lycaea, Maarrka, Macrohectopus, Maera, Maxilliphimedia, Megalorchestia, Megaluropus, Megomaera, Melita, Melphidippa, Menigrates, Mesogammarus, Micruropus, Molina, Monoculodes, Monoliropus, Monoporeia, Nedsia, Neohela, Niphargopsis, Niphargus, Octopupilla, Odontogammarus, Ommatogammarus, Onisimus, Orchestia, Orchomenella, Orchomenyx, Oxycephalus, Pachischesis, Pallasea, Pallaseopsis, Palmorchestia, Paraceradocus, Paragarjajewia, Parallorchestes, Paramphithoe, Parapallasea, Paraphronima, Parascelus, Parathemisto, Parhyale, Pariphimedia, Paroediceros, Perotripus, Pesudoprotella, Phreatogammarus, Phronima, Phronimella, Phrosina, Phtisica, Pilbarus, Platorchestia, Plesiogammarus, Pleustes, Podocerus, Poekilogammarus, Pontogammarus, Primno, Protella, Protogeton, Protomima, Pseudomicruropus, Pseudoprotella, Pseudorchomene, Quadrimaera, Ramellogammarus, Rhipidogammarus, Salentinella, Simorhynchotus, Sinogammarus, Sinorchestia, Spasskogammarus, Spinacanthus, Stegocephalus, Stenothoe, Streetsia, Stygobromus, Synurella, Syrrhoe, Talitroides, Talitrus, Talorchestia, Tetrathyrus, Themisto, Tryphosella, Urothoe, Ventiella, Vibilia*

Brachiopod	NCBI - nucleotide (WGA)	*Daphnia pulex* (GL737708.1-GL732523.1)
Copepod	NCBI - nucleotide	*Pseudodiaptomus annandalei, Calanus helgolandicus, Calanus finmarchicus, Lepeophtheirus salmonis, Ctenopharyngodon idella*
Isopod	NCBI - nucleotide	*Eurydice pulchra, Armadillidium vulgare*
Laboratory-specific common bacteria	NCBI - nucleotide	*Pseudomonas fluorescens* (NC_016830.1), *Propionibacterium acnes* (NC_017534.1) *Malassezia globosa* (NW_001849832.1-NW_001849898.1) *Elizabethkingia* sp. (NZ_CP011059.1) *Burkholderia oklahomensis* (NZ_CP009555.1) *Klebsiella pneumoniae* (CP004000.1) *Citrobacter freundii* (NZ_CP011652.1) *Streptomyces* sp. (CM001165.1) *Bacteroides eggerthii* (ABVO01000027.1) *Bacillus cereus* (NZ_CM000726.1) *Yersinia pestis* (AL590842.1) *Staphylococcus aureus* (CP001844.2) *Erwinia carotovora* (BX950851.1) *Gluconacetobacter xylinus* (CP004360.1)
rRNA	SILVA	SSU/LSU

Table AII.1 (SI Table 2.1) | *List of databases (or genera) utilized as reference to eliminate potential contaminants from viromes* - To minimize mis-annotation and streamline viral contig identification, both reads and contigs were compared to locally curated databases, including the human genome, phylogenetically closest relevant crustacean host genome, common laboratory-specific bacterial contaminant sequences, SILVA database of small (16S/18S) and large subunit (23S/28S) rRNA, and sequences generated from via parallel sequencing of a virus-free, nuclease-free water template (no template control). Reads and contigs with significant similarity (e-value<$1x10^{-5}$) to these databases were omitted from further viral annotation. 0.12% of all assembled contigs shared sequence similarity to this no template control database. While this annotation pipeline likely excludes various viral genotypes – particularly those encoding horizontally acquired genes, or highly divergent sequences - these comparisons improve the fidelity of ssDNA virus identification. NTC (no template control) also queried: viromes prepared parallel to microcrustacean viromes with virus-free water to capture potential pipeline-specific contaminants.

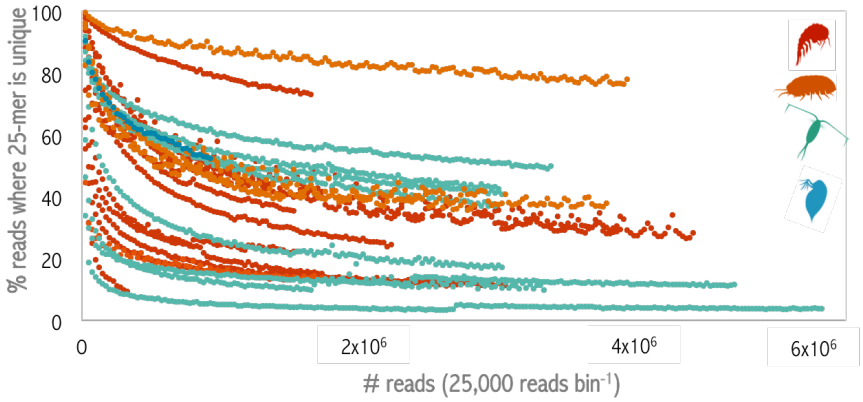

Figure AII.1 (SI Figure 2.1) | *Sequencing depth of microcrustacean viromes* – sequence saturation was assessed via quantification of unique 25-mers $read^{-1}$ $virome^{-1}$.

CRESS-DNA virus-like contigs (MCV)	215
MCVs containing a complete *rep* ORF	200
MCVs that share <95% average nucleotide ID across *rep*	173
Length of smallest MCV	580
Length of largest MCV	5748
Average nucleotide length of MCV	1963.1
Maximum number of MCV per library	33
Unique best-BLASTx hits	116
Libraries with >1 unique MCV	16
CRESS-DNA virus-like contigs per thousand viral reads	2.87
CRESS-DNA viruses derived from public databases	462

Table AII.2 (SI Table 2.2) | *Putative microcrustacean-associated CRESS-DNA virus (MCV) statistics* – Contigs were validated via comparison to locally curated databases of the nonstructural *rep* open reading frame ("ORF"; BLASTx, e-value<$1x10^{-5}$) and both contig and best BLASTx hit were secondly interrogated against IMG/vr and NCBI-nr databases. ORFs were called via GetORF v.6.6 (EMBOSS, excluding ORFs < 500nt). *Rep* ORFs were further assessed for the presence of conserved rolling circle replication motifs (RCR I-III) and viral SF3 helicase domains (Walker-A, Walker-B, motif-C), and a characteristic nonanucleotide origin of replication (ori; NANTATTAC). If multiple putative *rep*s were identified, the longest, unambiguously annotated *rep* was utilized for downstream analyses. Additional annotation of hypothetical structural genes was attempted using ORF alignment (BLASTx and tBLASTx against nr database, e-value<0.01), and HHpred remote homology predictor (v.2.0.13) against several HMM databases (PRK v.6.9, TIGRfam v.15.0, NCBI-CD v.3.16, and Pfam-A v.32.0 implemented in MPI Bioinformatics Toolkit, Max Planck Institute for Developmental Biology, Tübingen, Germany). Putative CRESS-DNA virus contig circularization was evaluated via dot-plot comparison for inverted or repeat sequences.

Amphipod (n=2 genome assemblies)	
Hyalella azteca	GCA_000764305.2
Parhyale hawaiensis	GCA_001587735.1
Brachiopod (n=5 genome assemblies)	
Daphnia magna	GCA_001632505.1
Daphnia pulex	GCA_000187875.1, GCA_900092285.1
Eulimnadia texana	GCA_002872375.1
Triops cancriformis	GCA_000981345.1
Copepod (n=10 genome assemblies)	
Acartia tonsa	GCA_900241095.1
Calanus finmarchicus	GCA_002740975.1
Calanus glacialis	GCA_002740985.1
Caligus rogercresseyi	GCA_001005125.1
Eurytemora affinis	GCA_000591075.2
Lepeophtheirus salmonis	GCA_000181255.2,GCA_001005205.1, GCA_001005235.1,GCA_001005385.1
Oithona nana	GCA_900157175.1
Decapod (n=4 genome assemblies)	
Caridina multidentata	GCA_002091895.1
Marsupenaeus japonicus	GCA_002291165.1
Penaeus monodon	GCA_002291185.1
Procambarus virginalis	GCA_002838885.1
Isopod (n=2 genome assemblies)	
Armadillidium vulgare	GCA_001887335.1
Ligia exotica	GCA_002091915.1

Table AII.3 (SI Table 2.3) | *Crustacean whole-genome assemblies* – publicly available crustaceans genomes evaluated for the presence of putative endogenous viral elements (EVEs).

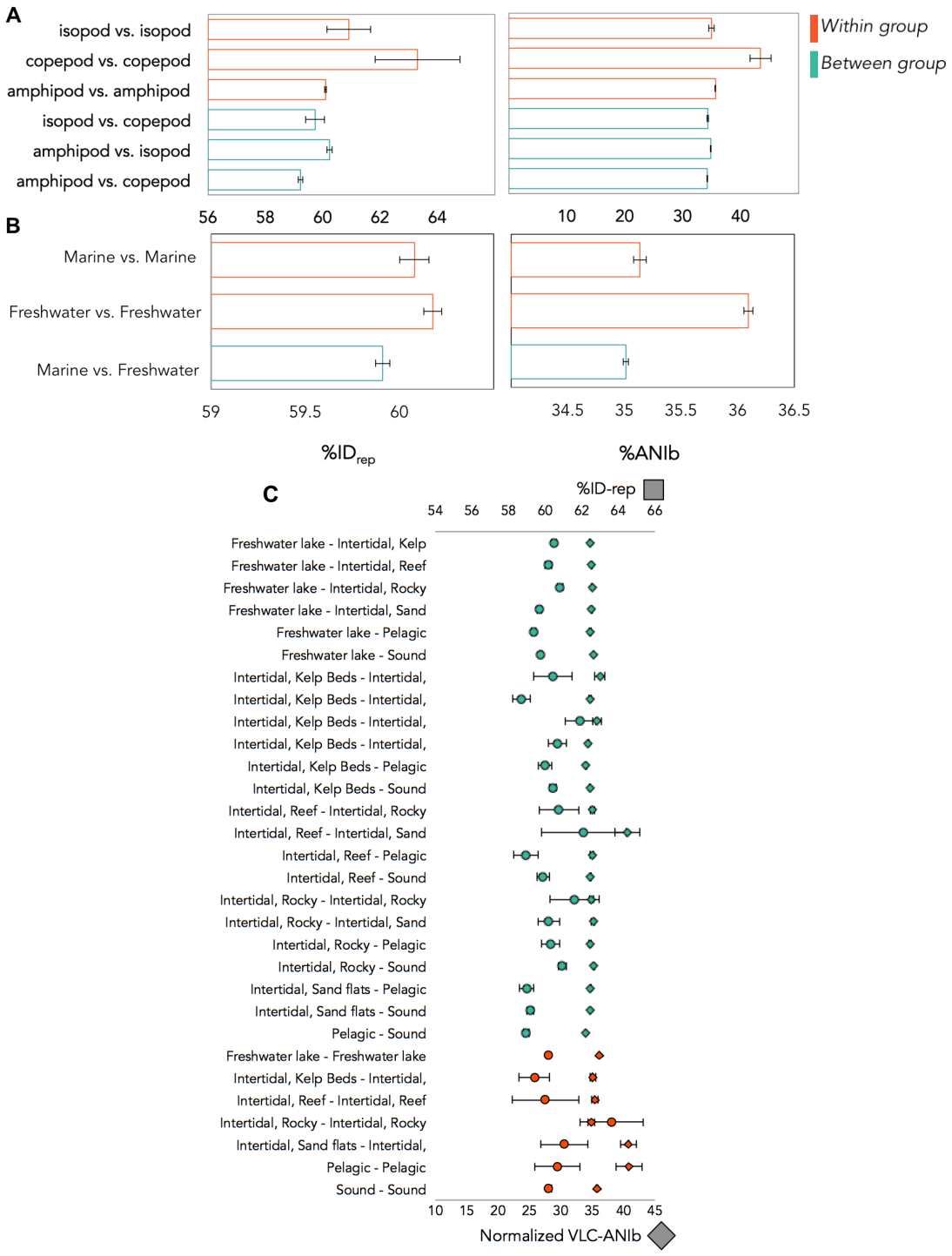

Figure AII.2 (SI Figure 2.2) | *Pairwise comparison of MCVs* – Post-hoc comparisons of MCV $\%ID_{rep}$(A, left) and ANIb (A, right) recovered from similar microcrustacean hosts versus dissimilar hosts (within-group and between-group pairwise comparison); Pairwise post-hoc comparisons of MCV genotype $\%ID_{rep}$(B, left) and ANIb (B, right) recovered from marine and lacustrine ecosystems to determine if those recovered from similar ecosystems exhibited greater pairwise genomic similarities (Average±SE). (C) Pairwise comparison of genotypes recovered from similar ecosystems.

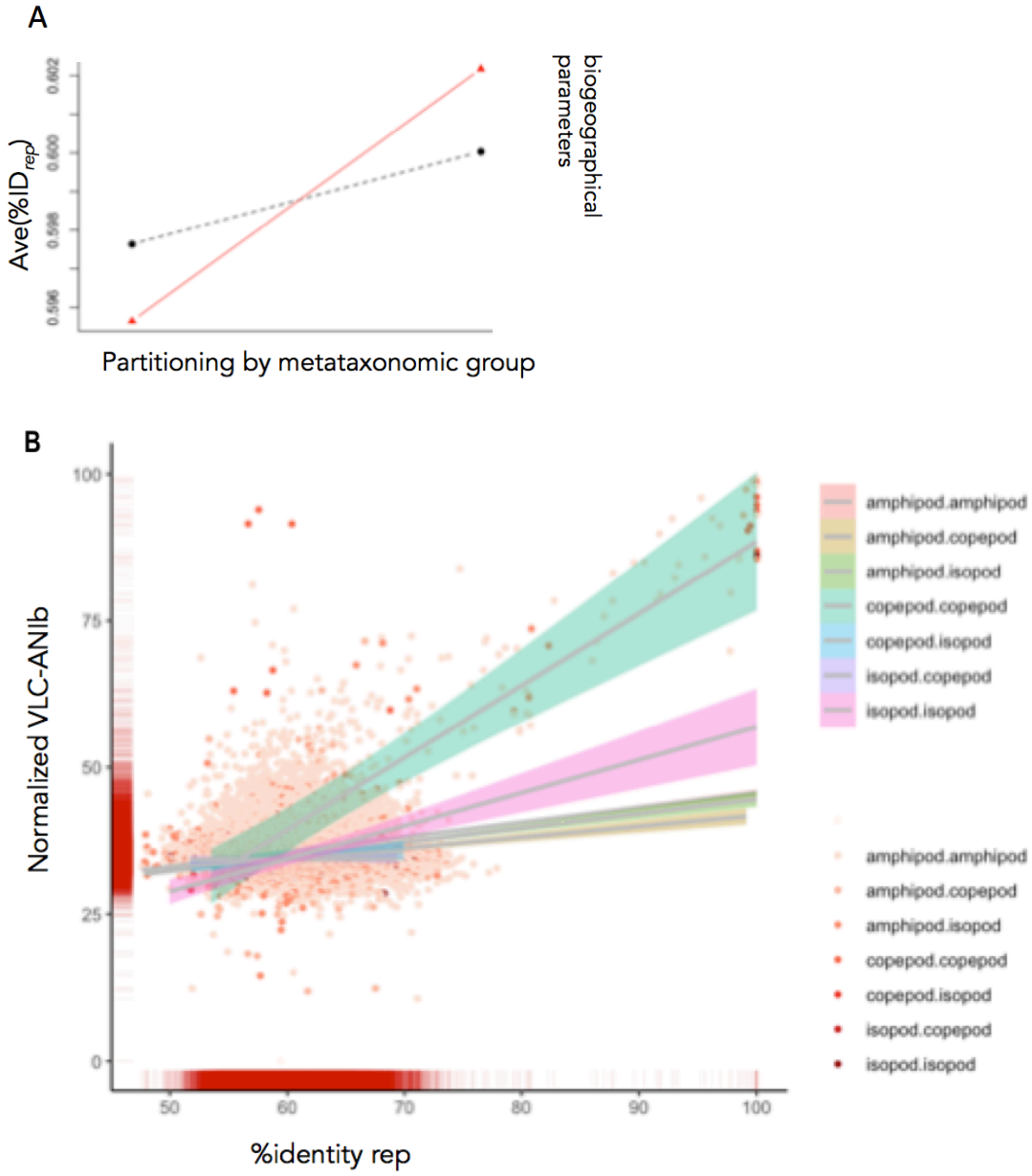

Figure AII.3 (SI Figure 2.3) | *Description of interaction between putative host phylogeny, biogeography, and viral genome composition and associated predictive power* – (A) The interaction effect between collection site variables and microcrustacean-specific variables is statistically significant, indicating that parsing the relevance of one in influencing viral genotype cannot be quantified without considering this effect of the other. This is an obvious comment, as abiotic and ecosystem-specific variables influence crustacean ecology/evolution, and only specific crustaceans inhabit specific ecosystems. (B) Scatterplot of pairwise comparisons of ANIb and $\%ID_{rep}$ within and between microcrustacean macrotaxonomic groups. Flanking density plots signify the number of observations at the given value (compressed histograms).

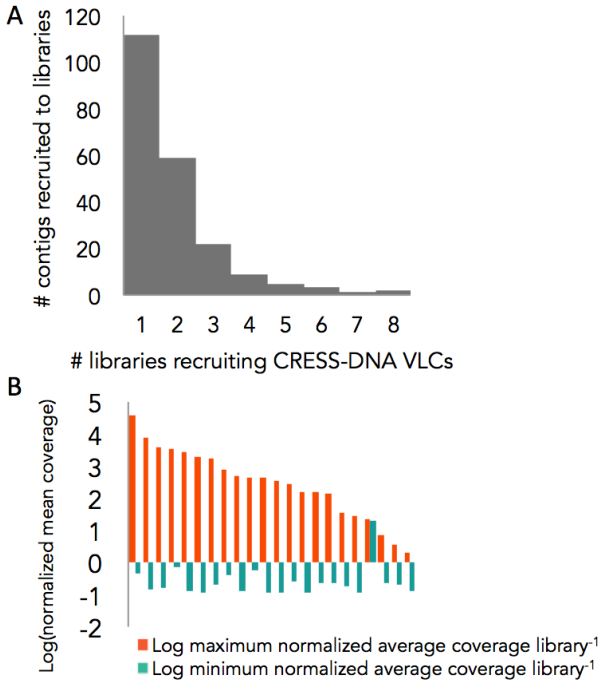

Figure AII.4 (SI Figure 2.4) | *Cross-library (cross-virome) MCV read recruitment* – (A) Histogram illustrating the number of libraries in which contigs were identified. Approximately half of all MCVs were identified in only one virome ("singleton MCVs"), while very few MCVs were identified in >5 viromes, potentially indicating specificity of MCVs; (B) comparison of coverage in contigs recruiting the maximum and minimum reads (i.e. exhibiting the maximum and minimum normalized coverage) per library, illustrating the log-fold difference in read recruitment among MCVs.

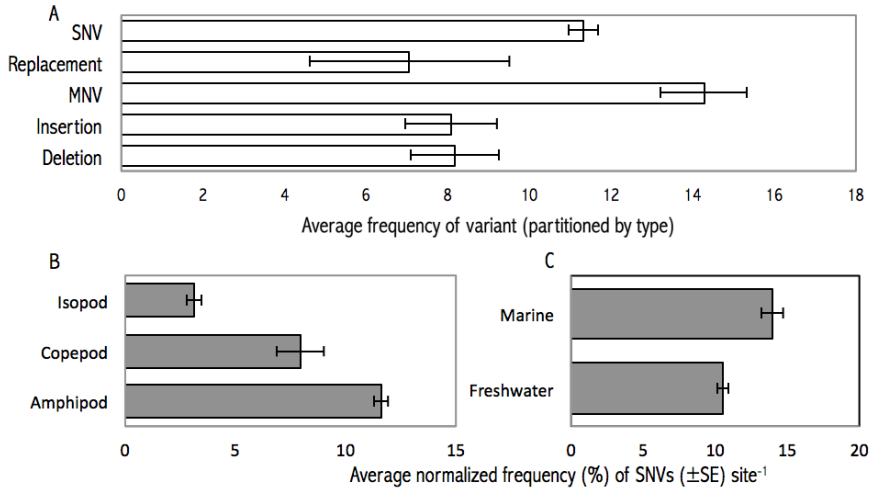

Figure AII.5 (SI Figure 2.5) | *Frequency of variants partitioned by type and metadata characteristics* – (A) Average (±SE) frequency of variant types over all putatively novel MCVs indicating greater maintenance of MNVs (multiple nucleotide variants) among sequences, despite greater observed prevalence of SNVs (single nucleotide variants); (B) Average (±SE) frequency of SNVs partitioned by association with crustacean order or broad ecosystem type (C).

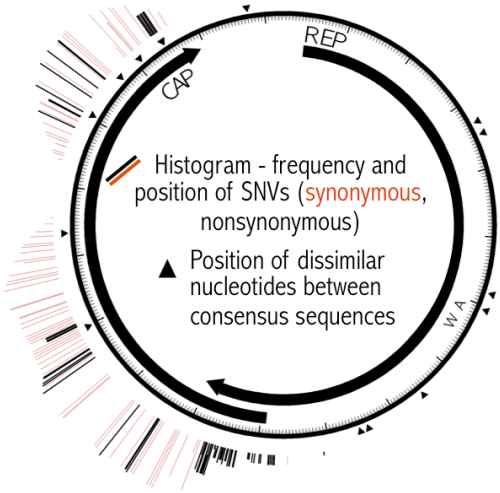

Figure AII.6 (SI Figure 2.6) | *Distribution of SNVs across the CRESS-DNA viral genotype, LM29173 associated with benthic amphipods from the Laurentian Great Lakes* – Black arrows indicate ORF position, triangles indicate sites of significant and persistent nucleotide change between the consensus sequence of Lake Michigan and Lake Huron consensus sequences of LM29173, and outer red/black histogram denote the frequency of variants, particularly among the less conserved *cap* ORF.

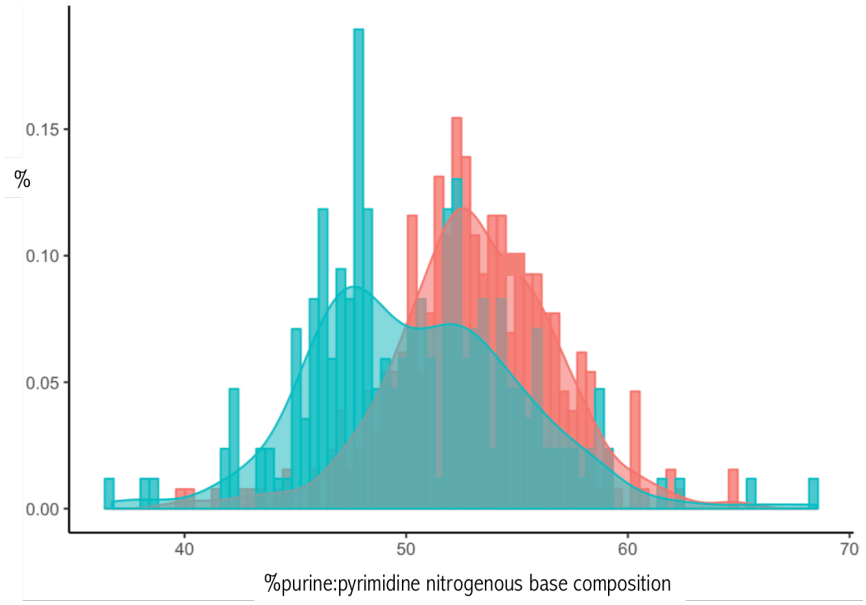

Figure AII.7 (SI Figure 2.7) | *Histogram of purine–pyrimidine distribution among coding and noncoding regions illustrating a bimodal distribution in intergenic regions* – As MCVs are single stranded, variants switching between different nitrogenous bases may alter the total composition of purine:pyrimidine bases and lability of packaged genome. Red – noncoding intergenic region (IG); Blue – coding region (*cap, rep*)

APPENDIX III

CHIMERIC CRESS-DNA VIRUS[3]

[3]Presented with minor amendment from the original published article:
Bistolas KSI, Besemer RM, Rudstam LG, Hewson I. 2017. Distribution and inferred evolutionary characteristics of a chimeric ssDNA virus associated with intertidal marine isopods. *Viruses*. 9(12): 361. doi: 10.3390/v9120361

[3]Supplementary material may be accessed at: www.mdpi.com/1999-4915/9/12/361/s1.

AIII.1 Abstract **|** Aquatic invertebrates are common reservoirs of a rapidly expanding group of circular *rep*-encoding ssDNA (CRESS-DNA) viruses. This study identified and explored the phylogenetic relationship between novel CRESS-DNA viral genotypes associated with Pacific intertidal isopods *Idotea wosnesenskii*, *Idotea resecata*, and *Gnorimosphaeroma oregonensis*. One genotype associated with *I. wosnesenskii*, IWaV278, shared sequence similarity and genomic features with *Tombusviridae* (ssRNA) and *Circoviridae* (ssDNA) genomes and was putatively assigned to the *Cruciviridae* clade comprising chimeric viruses. The complete genome of IWaV278 (3478 nt) was computationally completed, validated via Sanger sequencing, and exhibited sequence conservation and codon usage patterns analogous to other members of the *Cruciviridae*. Viral surveillance (qPCR) indicated that this virus was temporally transient (present in 2015, but not 2017), specific to *I. wosnesenskii* at a single collection site (Washington, DC, USA), more prevalent among male specimens, and frequently detected within exoskeletal structures. 18S rRNA sequences identified two alveolate protists associated with IWaV278-positive tissues and mechanical epibiont removal of ciliated exoskeletal structures eliminated viral detection, suggesting that the putative host of IWaV278 may be an epibiont of *I. wosnesenskii*. This investigation provides additional phylogenetic evidence to resolve *Cruciviridae* evolution and offers insight into the biogeography, specificity, and potential host of a crucivirus genotype.

AIII.2 Relevance | This study utilized viromes and samples described in Chapter 2, with emphasis on investigating the genomic characteristics of a putatively chimeric CRESS-DNA virus (encoding a *rep* ORF similar to a ssDNA virus and *cap* ORF similar to a ssRNA virus). This study demonstrated the transience of viral detection and potentially rapid evolution of this genotype in alignment with findings in Chapter 2. Evidence of chimerism, as well as potential affiliation with microcrustacean epibionts supports the inference that microcrustaceans may facilitate the confluence of capsidless genetic elements, promiscuously-packaging ssRNA viral capsids, and permissive cells, potentially underpinning the polyphyletic nature of ssDNA viral clades.

APPENDIX IV

VIGILANCE IN THE METHODS OF MODERN VIROMICS

AIV.1 Introduction **|** There are significant methodological challenges associated with preparing and parsing viromes associated with non-model hosts without any *a priori* knowledge. The inability to exclusively purify virus-like particles (VLPs) from composite tissue matrices requires greater sequencing efforts to capture viral diversity and computational methods to exclude cellular contamination from downstream analyses. Traditional methods of preparing metazoan-associated viromes often rely on size/weight selection (e.g. syringe filtration, cesium chloride density gradient centrifugation, etc) and nuclease digestion to separate cellular material and free nucleic acids from VLPs. However, the proportion of cellular reads generated through high-throughput sequencing (HTS) is often several orders of magnitude greater than putatively viral reads, and viral sequences average <10% of total annotated contiguous sequences (contigs) per virome. Furthermore, unannotated reads within these libraries often exceed 60-95% (Brum et al, 2015a, b; Roux et al, 2015, 2019). Within viromes prepared from whole microcrustacean samples, virus-like contigs comprised 11.9±2.4% of libraries, with ssDNA viruses only representing 2.6% of annotated virus-like contigs, despite genomic enrichment utilizing Φ29 polymerase (remaining contigs comprised microbial, host crustacean, or unannotated sequences; see Appendix II). These findings indicate that high levels of cellular contamination are routine, sequence annotation remains a key challenge in identifying viral sequences, and quantitative measures of viral diversity and abundance are often skewed. This appendix briefly outlines several methodological caveats associated with viromics that must be considered when analyzing novel viral genomes and considering downstream applications of computational assessments.

1. Utility of no template controls in HTS identification of CRESS-DNA viruses - Viral sequences have been identified in association with laboratory reagents and equipment, resulting in

the description of viruses unrelated to a targeted host. As a consequence, several viral genomes have been targeted for intense study, only to be later revealed as workflow contaminants (Kjartansdóttir et al, 2015; Laurence et al, 2014; Lusk et al, 2014, Naccache et al, 2013). For example, a novel circovirus-parvovirus hybrid DNA virus (NIH-CQV/PHV) was associated with seronegative hepatitis, but later traced to spin columns in commercial extraction kits (Naccache et al, 2014). Similarly, xenotropic murine leukemia virus–related virus (XMLV) was originally correlated with chronic fatigue syndrome (CFS) and prostate tumors in clinical studies (Ali et al, 2011; Zheng et al, 2011). XMLV resembled endogenous murine gammaretroviruses and was subsequently identified as murine contaminants of RT-PCR bioreagents. These cases highlight the necessity for both rigorous cross-validation of novel viral sequences and the development of computational techniques to exclude suspect sequences from studies of viral function and diversity. Virtually all non-computational molecular genome validation methods demand liberal use of negative controls, and are therefore capable of delineating between *bona fide* viral sequences and those derived from HTS workflows. However, many of these methods are not time or cost-effective and require genome-specific optimization, and scaling these techniques to accommodate multiple genotypes in a host-tissue associated virome is often intractable. This process may be streamlined by eliminating putative contaminant sequences in post-sequencing analysis. While contigs sharing strong homology to cellular proteins can be aligned to host/microbial genomes and removed from further analyses via computational subtraction, the inability to differentiate between relevant and contaminant viral sequences is frequently more difficult to remedy through bioinformatics methods in the absence of a comprehensive reference database of viral contaminants.

In this dissertation, we generated and maintained a local database of virus-like contaminant sequences via HTS of no-template control (NTC) libraries using nuclease-free water (used in purification, extraction, amplification reactions, etc in template-containing samples) in lieu of a template in standardized HTS workflows parallel to experimental viromes, aiding in CRESS-DNA viral validation (Illumina MiSeq 2x250bp, $1.7x10^7$ reads after trimming). 215 novel CRESS-DNA

virus-like contigs did not recruit reads from NTC libraries, confirmed via post-assembly mapping (80% similarity, 80% length). We argue that NTC libraries should be laboratory-, pipeline-, or HTS run-specific to provide a mechanism to differentiate host-associated and contaminant virus-like sequences. Few viral sequences have been well characterized as contaminants, and all are products of clinical studies with extensive and multi-laboratory verification. Predictably, no preexisting viral contaminant sequences were identified in either microcrustacean viromes or NTC libraries (e.g. NIH-CQV/PHV, XMLV). Computational subtraction with preexisting virus-like sequences is insufficient for this study: collection and preparation techniques modified for non-model organisms may introduce contaminant sequences, facilities and sequencing centers may use different combinations of reagents, vendors, handlers, and procedures, and a universal database may not reflect all permutations of the sample preparation workflow. Additionally, a contaminant in one scenario may be a significant sequence in another (e.g. silica spin-column contaminant NIH-CQV/PHV may be a relevant component of diatom viromes, where the virus may have originally been derived). Development of workflow-specific local databases through NTC-sequencing concurrent with experimental metaviromes may provide both the most conservative and accurate depiction of viral contaminants. Therefore, with progressive development in HTS platforms leading to the reduced price of HTS runs, NTC libraries and computational subtraction pipelines minimize viral genome validation costs while maintaining the integrity of CRESS-DNA viral discovery and community analyses.

2. The "great annotation problem" – In order to determine the impact of viruses on invertebrate population dynamics, it is clearly essential to first identify associated viruses. In this regard, HTS has profoundly expanded our knowledge of viral diversity. However, the majority of viromic sequences remain un-annotated "viral dark matter" (Brum et al, 2015a, b; Roux et al, 2015, 2019). This presents a challenge for detecting novel viruses and inferring their evolutionary and ecological context. Traditionally, metazoan-associated ssDNA viruses have been classified into clearly demarcated families (e.g. *Circoviridae, Parvoviridae, Genomoviridae*, etc; ICTV,

http://ictv.global/report; Simmonds et al, 2017). However, ssDNA viruses newly identified via HTS, or "e-viruses," frequently share loose homology with these well-established taxonomic groups, with few conserved genetic markers or hallmark characteristics, obscuring classification and predicted relationship with hosts. A panel convened to discuss the merits of incorporating "e-viruses" into a taxonomic hierarchy concluded that ssDNA viruses, in particular, represent an exceptionally convoluted cluster of genotypes due to variability in gene orientation and rapid accumulation of sequence variants (Rosario et al, 2012, 2018, Simmonds et al, 2017). This sequence saturation and potential for lateral gene transfer or genomic reorganization impede clear delineation between CRESS-DNA virus phylogenetic groups, suggesting that novel CRESS-DNA viruses associated with non-model hosts, including those identified in this dissertation, may by polyphyletic. CRESS-DNA virus taxonomic organization continues to undergo continuous revision, further emphasizing the necessity to continue sequencing efforts beyond agro-economically relevant systems. Characterization of these e-virsues represents a key area of growth for computational analysis of viral consortia. CRESS-DNA virus validation methods range from computational (motif conservation, coverage profiling) to molecular (genome walking/circularization, quantification of viral mRNA or protein). Additionally, epidemiological approaches to quantify the abundance/replication of CRESS-DNA viral genotypes within microcrustacean populations have been employed to infer potential ecological significance. Further strategies may range from alignment-dependent (e.g. similarity scoring), to trait-counting (e.g. motif validation, ORF calling, structure and function prediction, hidden markov model profile comparison), to alignment-free strategies (e.g. dinucleotide, oligomer, or k-mer profiling, etc). Ultimately, the growth of data repositories is essential to the exponential progress in viral discovery and may improve confidence in viral categorization.

3. When less is more – Assuming most CRESS-DNA viruses contain <6kb of DNA, with an unknown multiplicity of infection, amplification is nearly universally necessary prior to sequencing CRESS-DNA viral consortia. Isothermal multiple displacement amplification utilizing Φ29

polymerase (derived from *Podoviridae Bacillus subtilis* phage Φ29) is commonly used for high-fidelity, untargeted amplification of DNA templates using random hexamers (Binga et al, 2008; de Paz et al, 2018; Reagin et al, 2003). Φ29 polymerase is an inherently processive polymerase, catalyzing both protein-primed initiation and DNA polymerization, continually synthesizing new DNA via strand displacement (Reagin et al, 2003). Of note, Φ29 polymerase preferentially binds to ssDNA relative to dsDNA, utilizing domain TPR2 to passively unwind dsDNA rather than encoding a helicase (Morin et al, 2012). Φ29 polymerase optimally synthesizes circular templates via rolling circle replication (RCR), as these templates need to be primed only once to form multiple concatamers of the template strand. Therefore, with small, single-stranded circular genomes, CRESS-DNA viruses are often over-represented in viromes, obscuring any quantitative proxies for viral abundance via read recruitment (Yilmaz et al, 2010). These biases often remain after pooling reactions, and can result in over-amplification of single CRESS-DNA templates relative to others. Therefore, supplementary quantitative analyses, ranging from RT-qPCR to microscopy are necessary to evaluate copy number (as in chapters 3-4). In this dissertation, we capitalize on the genomic attributes of CRESS-DNA virus templates to enrich for circular, ssDNA viral genomes, aiding discovery.

4. Other consideration in experimental design and HTS of CRESS-DNA viruses – Viromics offers a key foundation for future study. However, HTS data rarely confirms the presence of active viral propagation and may systematically exclude viral taxa. These biases are numerous and must be considered in downstream analyses. For example, (1) Large viruses, including Nucleocytoplasmic large DNA viruses (NCLDVs) may be excluded via size filtration and therefore be omitted in downstream analyses of viral diversity (Kleiner et al, 2015), though these taxa are often spuriously identified even in their absence due to a proclivity for gene capture of cellular sequences. (2) ssDNA genomes, including those with non-icosahedral capsid structures, may be more labile and degrade prior to sequencing. (3) Extraction protocols may vary in efficiency or be biased towards extraction of specific genomes. (4) Small genomes and those without complex secondary structures may be

overamplified relative to others. (5) Establishing impacts of infection are often methodologically limited. While the *de facto* gold standard for pathogen validation remains fulfillment of Koch's postulates, this is often not scalable in many viruses, particularly of invertebrates (e.g. microcrustaceans) due to the challenging nature of viral isolation/propagation in the absence of established cell lines or host cultivation for non-model organisms. This approach is also inadequate to characterize those that have subtle impacts on their host, potentially including CRESS-DNA viruses (Byrd & Segre, 2016). Indeed, even confirming crustacean infectivity requires artificial infection, nucleic acid propagation, recovery of viral proteins, and/or detection of an innate immune response. (6) Low biomass samples are subject to bias in every step of virome preparation, with low nucleic acid concentration excluding or enriching for specific virus-like sequences (Chafee et al, 2015; Kleiner et al, 2015; Parkinson et al, 2012). (7) Non-ecologically relevant (contaminant) viral sequences may be introduced at multiple stages during virome preparation and sequencing. (8) Sequencing depth, normalization, read quality control, assembly method, read recruitment methods, and all other computational methods may vary among software and versions, resulting in variation in predicted viral genome and requiring molecular confirmation (Leek et al, 2010; Ramond et al, 2015; Solonenko et al, 2013; Sutton et al, 2019).

This dissertation only briefly explored the intra-ecosystem, intra-population and intra-organismal dynamics of microcrustacean-associated CRESS-DNA viruses, with abundance of CRESS-DNA viruses likely skewed by methodological caveats, such as over-amplification of circular templates by pre-sequencing polymerases, exclusion of virions >0.2μm, etc. Viromics is rarely quantitative, despite measures to lessen bias, and these caveats must be recognized to determine the implications of this study and integrate this data into the cumulative understanding of aquatic ssDNA viruses.

AIV.2 References

Ali MA, Dale JK, Kozak CA, Goldbach-Mansky R, Miller FW, Straus SE, Cohen JI. 2011.

Xenotropic murine leukemia virus-related virus is not associated with chronic fatigue syndrome in patients from different areas of the US in the 1990s. *Virol J.* 24;8:450. doi:10.1186/1743-422X-8-450

Binga EK, Lasken RS, Neufeld JD. 2008. Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. *ISME J*. 2(3):233-41. doi:10.1038/ismej.2008.10.

Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, Chaffron S, Cruaud C, de Vargas C, Gasol JM, Gorsky G, Gregory AC, Guidi L, Hingamg P, Not DLF, Ogata H, Pesant S, Poulos BT, Schwenck SM, Speich S, Dimier C, Lewis SK, Picheral M, Searson S, Tara Oceans Coordinators, Bork P, Bowler C, Sunagawa S, Wincker P, Karsenti E, Sullivan MB. 2015. Patterns and ecological drivers of ocean viral communities. *Science*. 22(348,6237). doi:10.1126/science.1261498

Brum JR, Sullivan MB. 2015. Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat Rev Microbiol*. 13:147–159. doi:10.1038/nrmicro3404.

Byrd AL, Segre JA. 2016. Adapting Koch's postulates. Science. 351(6270):224-226. doi:10.1126/science.aad6753.

Chafee M, Maignien L, Simmons SL. 2015. The effects of variable sample biomass on comparative metagenomics. *Environ Microbiol*.17(7):2239-53. doi: 10.1111/1462-2920.12668.

de Paz AM, Cybulski TR, Marblestone AH, Zamft BM, Church GM, Boyden ES, Kording KP, Tyo KEJ. 2018. High-resolution mapping of DNA polymerase fidelity using nucleotide imbalances and next-generation sequencing. *Nucleic Acids Res.* 27;46(13):e78. doi:10.1093/nar/gky296.

Kjartansdóttir, K.R., Friis-Nielsen, J., Asplund, M., Mollerup, S., Mourier, T., Jensen, R.H. Hansen TA, Rey-Iglesia A, Richter SR, Alquezar-Planas DE, Olsen PV, Vinner L, Fridholm H, Sicheritz-Pontén T, Nielsen LP, Brunak S, Willerslev E, Izarzugaza JM, Hansen AJ. Traces of ATCV-1 associated with laboratory component contamination. *Proc Natl Acad Sci USA.* 112: E925–E926. doi:10.1073/pnas.1423756112.

Kleiner M, Hooper LV, Duerkop BA. 2015. Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genomics.* 22;16:7. doi:10.1186/s12864-014-1207-4.

Laurence, M., Hatzis, C., and Brash, D.E. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS One*. 9: e97876–e97878. doi:10.1371/journal.pone.0097876.

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*.11(10):733-9. doi: 10.1038/nrg2825.

Lusk, R.W. Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PLoS One*. 9: e110808. doi:10.1371/journal.pone.0110808.

Morin JA, Cao FJ, Lázaro JM, Arias-Gonzalez JR, Valpuesta JM, Carrascosa JL, Salas M, Ibarra B. 2012. Active DNA unwinding dynamics during processive DNA replication. *Proc Natl Acad Sci USA*. 109(21):8115-20. doi:10.1073/pnas.1204759109.

Naccache SN, Greninger AL, Lee D, Coffey LL, Phan T, Rein-Weston A, Aronsohn A, Hackett J Jr, Delwart EL, Chiu CY. 2013. The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *J Virol.* 87(22):11966-77. doi:10.1128/JVI.02323-13.

Naccache SN, Hackett J Jr, Delwart EL, Chiu CY. 2014. Concerns over the origin of NIH-CQV, a novel virus discovered in Chinese patients with seronegative hepatitis. *Proc Natl Acad Sci USA*. 18;111(11):E976. doi:10.1073/pnas.1317064111.

Parkinson NJ, Maslau S, Ferneyhough B, Zhang G, Gregory L, Buck D, Ragoussis J, Ponting CP, Fischer MD. 2012. Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA. *Genome Res*. 22(1):125-33. doi:10.1101/gr.124016.111.

Ramond JB, Makhalanyane TP, Tuffin MI, Cowan DA. 2015. Normalization of environmental metagenomic DNA enhances the discovery of under-represented microbial community members. *Lett Appl Microbiol*. 60(4):359-66. doi:10.1111/lam.12380.

Reagin MJ, Giesler TL, Merla AL, Resetar-Gerke JM, Kapolka KM, Mamone JA. 2003. TempliPhi: A sequencing template preparation procedure that eliminates overnight cultures and DNA purification. *J Biomol Tech*. 14(2):143-8.

Rosario K, Duffy S, Breitbart M. 2012. A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Arch Virol*. 157(10):1851-71. doi:10.1007/s00705-012-1391-y.

Rosario K, Mettel KA, Benner BE, Johnson R, Scott C, Yusseff-Vanegas SZ, Baker CCM, Cassill DL, Storer C, Varsani A, Breitbart M. 2018. Virus discovery in all three major lineages of terrestrial arthropods highlights the diversity of single-stranded DNA viruses associated with invertebrates. *PeerJ*. e5761. doi:10.7717/peerj.5761.

Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, Kuhn JH, Lavigne R, Brister JR5, Varsani A, Amid C, Aziz RK, Bordenstein SR, Bork P, Breitbart M, Cochrane GR, Daly RA, Desnues C, Duhaime MB, Emerson JB, Enault F, Fuhrman JA22, Hingamp P, Hugenholtz P, Hurwitz BL, Ivanova NN, Labonté JM, Lee KB, Malmstrom RR, Martinez-Garcia M, Mizrachi IK, Ogata H, Páez-Espino D, Petit MA, Putonti C, Rattei T, Reyes A, Rodriguez-Valera F, Rosario K, Schriml L, Schulz F, Steward GF, Sullivan MB, Sunagawa S, Suttle CA, Temperton B, Tringe SG, Thurber RV, Webster NS, Whiteson KL, Wilhelm SW, Wommack KE, Woyke T, Wrighton KC, Yilmaz P, Yoshida T, Young MJ, Yutin Z, Allen LZ, Kyrpides NC, Eloe-Fadrosh EA. 2019. Minimum information about an uncultivated virus genome (MIUViG). *Nat Biotechnol*. 27 (1):29-37.doi: 10.1038/nbt.4306.

Roux S, Hallam SJ, Woyke T, Sullivan MB. 2015. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife*. 4:e08490. doi:10.7554/eLife.08490.

Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B, Hull R, King AM, Koonin EV, Krupovic M, Kuhn JH, Lefkowitz EJ, Nibert ML, Orton R, Roossinck MJ, Sabanadzovic S, Sullivan MB, Suttle CA, Tesh RB, van der Vlugt RA, Varsani A, Zerbini FM. 2017. Consensus statement: Virus taxonomy in the age of metagenomics. *Nat Rev Microbiol*. 15(3):161-168. doi:10.1038/nrmicro.2016.177.

Solonenko SA, Ignacio-Espinoza JC, Alberti A, Cruaud C, Hallam S, Konstantinidis K, Tyson G, Wincker P, Sullivan MB. 2013. Sequencing platform and library preparation choices impact viral metagenomes. *BMC Genomics.* 10;14:320. doi:10.1186/1471-2164-14-320.

Sutton TDS, Clooney AG, Ryan FJ, Ross RP, Hill C. 2019. Choice of assembly software has a critical impact on virome characterisation. *Microbiome*. 28;7(1):12. doi:10.1186/s40168-019-0626-5.

Yilmaz S, Allgaier M, Hugenholtz P. 2010. Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nat Methods*.7(12):943-4. doi:10.1038/nmeth1210-943.

Zheng H, Jia H, Shankar A, Heneine W, Switzer WM. 2011. Detection of murine leukemia virus or mouse DNA in commercial RT-PCR reagents and human DNAs. *PLoS One*. 6(12): e29050. doi:10.1371/journal.pone.0029050.