Genomes of Small RNA Viruses: Amendments, Discoveries, and Characterizations

Genomes of Small RNA Viruses: Amendments, Discoveries, and Characterizations

A dissertation presented

by

Minh Vong

to

The Department of Molecular and Cellular Biology

in partial fulfillment of the requirements

for the degree of

Doctorate of Philosophy

in the subject of

Biochemistry

Harvard University

Cambridge, Massachusetts

April 2018

Dissertation advisor: Professor Max Nibert By: Minh Vong

Genomes of Small RNA Viruses: Amendments, Discoveries, and Characterizations

Abstract

RNA viruses make up a disproportionate amount of emerging human diseases, largely contribute to agriculture, and are recognized for their important roles in shaping the evolution of life. Metagenomics datasets have emphasized the importance of viral genomics in virus discovery. The International Committee on Taxonomy of Viruses (ICTV) has endorsed use of metagenomics data alone in identifying new viruses. One immediate question the ICTV will have to address is how to sort through these data. The works presented here are broken into three sections and are aimed at identifying general rules that can be applied to classification and extracting further utility of metagenomics data. In Section I, I describe general rules that have been successfully applied retroactively to hypothesize potential errors or incompleteness in previously reported RNA viral sequences. In Section II, I present newly discovered small RNA viruses and subviral agents of such identified from metagenomics data and the information they contribute to an understanding of evolutionary history and virus biology. In Section III, I present development of a virus-mediated reverse genetics system that can be used to understand the biological function of viral sequences at the nucleotide level, thus providing a biological means to assess confidence in sequence differences between closely-related and concurrent infection of multiple virus species. Thus, Section I will cover amendments to previously reported RNA viral sequences have been proposed, Section II will provide how metagenomics data provide insights into virus evolution and virus biology, and Section III will provide a biological means of improving confidence in sequencing results from metagenomics data. I conclude by presenting thoughts on how the data from the three sections and their implications may shape future discussions in the age of viral metagenomics.

Table of Contents

Acknowledgements

Before I begin writing about science, I want to thank the many people who have helped, contributed, and supported my graduate training in science and leading up to it.

I want to give my deepest gratitude to my Ph.D. advisor Max Nibert for the many things he has done towards my dissertation and science training. I am extremely grateful for Max's continuous support, mentorship, and sharing of his thoughts and knowledge of virology throughout my development in science. It was a great joy and privilege to train under Max. I found excitement for science, discovery, and reasoning plentiful in Max's lab. Neither my growth as a scientist during graduate school nor any of the work described here would have been possible if not for Max's ability to teach students how to think like a research scientist.

Much of the progress described in my dissertation and my growth in conducting science research were also made possible due to the support and guidance of my dissertation advisory committee members. I am grateful for everyone having taken interest and time in learning about my work. I am also thankful for each of their help in keeping me on schedule to accomplish my planned aims and for their advice. I am super thankful for Professor Craig Hunter for his role as my committee chair. I feel super fortunate for also having both Professors Victoria D'Souza and Lee Gehrke on my committee. A lot of technical problems were solved thanks to their suggestions and their offering their labs. Much of my progress described in Chapter Six was the direct outcome of Professor D'Souza opening her lab for help. I will forever feel grateful for her generosity. I am forever thankful for their contributions to my science training and growth.

I am also thankful for the many other people who have taught me various techniques, worked with me to contribute to these works, and who I have had the pleasure of mentoring. Previous lab members who have all become great personal friends include: Delphine Depierreux,

who I enjoyed mentoring and who also ran the experiments in Chapter Two; May Yang, who showed me how to culture the protozoan used in Chapters Five and Six; Katie Smith and William Gao who were fantastic summer students I enjoyed mentoring and who also helped with experiments described in the Current & Future Directions of Chapter Four; Daniela Silva-Ayala for helpful daily comments; and Jesse Pyle for sharing his knowledge of plant RNA extraction used in Chapter Four. This work also would not be possible without our collaborators: Professor Honorine Ward and Jacob Ludington for their work in obtaining the RNA used in Chapter Three; Kate Godin, Carolina Salguero, Vincent Pham, and Nico Wagner for their help me with various parts of Chapter Six; and Professor Patricia Johnson and Brian Janssen for their kind gift and advice used in Chapter Six.

I am also very grateful for Drs. Timothy Block, Anand Mehta, and Patrick Romano for providing me with a wonderful exposure to science research experience at the Baruch S. Blumberg Institute. My start in science would not have been possible without my professors and advisors from college at the University of the Sciences in Philadelphia and I'm very grateful for all of them: Drs. William Law, Vandana Miller, Mary Beth Murray and Ms. Roxanne Evans.

Lastly, though no less important to me, I want to thank my friends and family for their support and encouragement during my graduate studies. Thank you for remaining to be my best friends since college, Kwame Lewis and Mick Sacks. Thank you, Miguel Coelho and Alan Le Goallec, for the deep friendships while at Harvard and going into our futures. Thank you to my brothers, Andy Vong and Walter Vong, and my parents, Chan Ho and William Vong, for all the lifelong encouragement.

Without anyone mentioned here I would not be writing this dissertation today. I wish to give my deepest appreciation to everyone.

Abbreviations

(-) ssRNA: negative-sense single-stranded RNA
(+) ssRNA: positive-sense single-stranded RNA
“A”: adenine
“C”: cytosine
“G”: guanine
“N”: can be either A, G, C or T (if DNA) or U (if RNA)
“T”: thymine
“U”: uracil
“W”: can either be A or U
“Y”: T or C
µg: microgram
µl: microliter
µM: micromolar
aa: amino acid
AP: adhesion protein
Arg: arginine
ATP: adenosine triphosphate
BevuMV1: Beta vulgaris mitovirus 1
bp: base pairs
cDNA: complementary DNA
CMS: cytoplasmic male sterility
CP: capsid protein
CSpV1: Cryptosporidium parvum virus 1
CyP: cysteine proteinase
ddATP: dideoxyadenosine triphosphate
DEPC: diethyl pyrocarbonate
DMSO: dimethyl sulfoxide
DNA: deoxyribonucleic acid
dNTP: deoxynucleotide triphosphate
dsRNA: double-stranded RNA
EDTA: Ethylenediaminetetraacetic acid
ERVs: endogenous retrovirus
g: gravity
GLV: Giardia lamblia virus
H1N1: hemagglutinin-1, neuraminidases-3 influenza A virus
HCl: hydrochloric acid
hr: hour
ICTV: International Committee on Taxonomy of Viruses
IDT: Integrated DNA Technologies
IVT: *in vitro* transcription
indels: insertions or deletions

IRE: internal replication enhancer
kcal: kilocalorie
kDa: kilodalton
Met: methionine
mg: milligram
min: minute
mL: milliliter
mM: millimolar
mol: mole
NaCl: sodium chloride
NERVEs: non-retroviral endogenous RNA elements
NGS: next-generation sequencing
nm: nanometer
NMIA: N-methylisatoic anhydride
NR/NT: non-redundant nucleotide
nt: nucleotides
NTP: nucleotide triphosphate
ORF: open reading frame
PAGE: polyacrylamide gel electrophoresis
PCR: polymerase chain reaction
PEG: polyethylene glycol
pI: isoelectric point
pmol: picomole
PRF: programmed ribosome frameshift
qRT-PCR: quantitative reverse-transcription polymerase chain reaction
RdRp: RNA-dependent RNA polymerase
Rf: restorer of fertility
RLM- 3' RACE: RNA ligase-mediated 3' rapid amplification of cDNA ends
RNA: ribonucleic acid
RNAi: RNA interference
rpm: revolutions per minute
RT: reverse transcriptase
RT-PCR: reverse-transcription polymerase chain reaction
SARS-CoV: severe acute respiratory syndrome coronavirus
sec: second
SHAPE: Selective 2'-hydroxyl acylation analyzed by primer extension
SRA: Sequence Read Archive
ssRNA: single-stranded RNA
STD: sexually transmitted disease
STV: southern tomato virus
Tb: terabyte
TBE: Tris/borate/EDTA
tRNA: transfer RNA

TSA: Transcriptome Shotgun Assembly
Tvag: *Trichomonas vaginalis*
TVV: Trichomonas vaginalis virus
TVVS: Trichomonas vaginalis virus satellite
UTR: untranslated region
UvURV: Ustilaginoidea virens unassigned RNA virus HNND-1
YPG media: yeast-peptone-glucose media
Z. bailii: *Zygosaccharomyces bailii*
ZbV-Z: zygosaccharomyces bailii virus Z

List of Figures

List of Tables

Chapter One: Dissertation Introduction

1.1 Dissertation Introduction

There exists over 1,400 different species of human pathogens (Taylor 2001). Viruses are the most common emerging human pathogen and RNA viruses make up more than half of all recognized viral emerging and re-emerging species (Woolhouse 2005). Further, RNA viruses have gained wide media attention as being the sources of several recent notable outbreaks and epidemics. In 2003, the positive-sense (+) ssRNA severe acute respiratory syndrome coronavirus (SARS-CoV) caused the SARS epidemic that originated from China. SARS-CoV caused over 700 reported deaths and over 8,000 illnesses in five different continents (Peiris 2003). In 2009, the negative-sense (-) ssRNA virus, hemagglutinin-1, neuraminidases-3 influenza A virus (H1N1), caused the first pandemic of the 21st century with estimated casualties of over 280,000 worldwide (Dawood 2012). In 2014-15, the (-) ssRNA ebola virus epidemic claimed over 11,000 lives while infecting more than 28,000 people in West Africa (WHO 2016). A year later in 2015-16, the (+) ssRNA zika virus epidemic affected at least 33 countries with over 4,000 reports of babies born with microcephaly in Brazil alone (Peterson 2016). Some RNA viruses have remained to be a large common problem. The retroviral RNA human immunodeficiency virus 1 (HIV) was discovered in the early 1980s (Barré-Sinoussi 1983) and is still the most serious infectious disease humanity faces. Infection by (+) ssRNA Dengue virus has gradually been increasing for the past three decades and is responsible for about 50 million annual cases throughout the tropical and subtropical world (Gubler 2002).

RNA viruses have also significantly impacted agriculture and these viruses have played monumental roles in the understanding of virology. The first discovered virus came from the mosaic disease of tobacco plants caused by the (+) ssRNA tobacco mosaic virus (Ivanofsky 1892) and the same virus also provided the first image of any virus particle (Bernal 1941). The (+) ssRNA

citrus tristeza virus was responsible for loss of almost 100 million citrus trees in the 1990s (Moreno 2008). The (+) ssRNA foot-and-mouth disease virus was responsible for the 2001 outbreak of foot and mouth disease that resulted in the slaughtering of over 3 million cattle, resulting in the over $4 billion of damages (Knight-Jones 2013). Not all virus infections of agriculture have been negative. Some viruses have been used by people because the outcomes of infection have been valuable. Infection by the (+) ssRNA tulip breaking virus produces streaks on the petals of tulips so attractive that 17th century breeders from Holland would intentionally graft infected bulbs onto new plants (Lesnaw 2000). The dsRNA virus Cryphonectria hypovirus 1 was used to successfully control the spread of a fungus that had nearly eliminated upper parts of chestnut trees in the eastern United States and Europe during the 1950s and 1960s (Hillman 2004).

RNA viruses are gaining more recognition for their impact on the evolution of life. Proteins used in present-day mammals and humans show strong homology to viral proteins. Endogenous retrovirus (ERVs) elements make up 8% of the human genome (Sharif 2012). Formation of the syncytiotrophoblast during the placenta/uterine exchange process requires the expression of syncytin, a protein derived from ERVs (Mi 2000). Further, genome sequences from various mammal species suggest that syncytin analogs have been acquired by diverse mammalian lineages on at least six independent occasions (Dupressoir 2012). Most recently in early 2018, two independent groups reported on the function of a retrovirus-derived coat protein, Arc, in the transport of proteins and RNAs between neurons (Ashley 2018; Pastuzyn 2018).

Lastly, both DNA- and RNA-based viruses are also gaining more recognition for their impact on the environment. The expression of the psbA gene, a core photosynthesis protein used by cyanobacteria, within the genomes of 88% of cyanophage species has been used to argue that these viruses may have attributed to the Great Oxidation Event of earth, when oxygen levels rose from

trace amounts to today's 21% (Sullivan 2006). Though empirical support for claims dating back billions of years may be out of reach, the view of viruses as having ecologically unimportant roles have shifted. Prior to the 1990s, viruses were viewed as having ecologically unimportant roles due to the low count of virus particles at the time (Frank 1987). Use of direct-count transmission electron microscopy (Bergh 1989) and epifluorescent counting (Marie 1999; Noble 1998) have estimated the count of virus particles at 10^{31} virus particles/ml at any given time in most environments, exceeding the number of cells by 10-100 fold. These new estimates and discoveries make the virosphere the most abundant source of nucleic acid on earth. Since these counting techniques do not provide information on the genome type of these viruses it has been assumed that most of the viruses were DNA-based bacteriophages of marine microbes, and, as a consequence many studies, particularly prior to 2003, had focused on DNA-based oceanic viruses. The first RNA virus of marine microbes was reported in 2003 and showed that the protist *Heterosigma akashiwo* can be infected and lysed by (+) ssRNA viruses (Tai 2003). This was followed by reports of other ssRNA viruses infecting various phytoplanktons, such as dinoflagellates and diatoms (Nagasaki 2004; Shirai 2008). Comparing the total RNA and DNA of viral fractions harvested from seawater, Steward et al. estimates that RNA viruses rival or exceed that of DNA viruses in abundance in coastal seawater (Steward 2008).

The virosphere's size and diversity are further uncovered by the current age of metagenomics and at a rapid pace. The virus family *Genomoviridae* has one single recognized virus by the International Committee on Taxonomy of Viruses (ICTV) but has potentially more than 120 members based solely on metagenomics data (Krupovic 2016). The *Tara* Oceans expedition of 43 different ocean sites identified more than 5,000 possible virus populations, but only 39 could be classified into the current virus groups recognized by ICTV (Brum 2015). Within a another year,

over 5Tb of metagenomics data had been generated from over 3,000 geographical areas and led to the identification of approximately 125,000 new viral genomes and a 16-fold increase in the number of identified viral genes (Paez-Espino 2016). In addition to studies using metagenomics directed at virus discovery, previously published metagenomics datasets not aimed at virus discovery also provide for rich sources of virus discovery. VirSorter is a recently developed tool that has been used to identify over 12,000 new DNA viral genomes from existing datasets (Roux 2015). More recently, our lab has also successfully identified new RNA viruses from previous transcriptomics studies of plants (Pyle 2017; Nibert 2018a; Nibert 2018b). To address this large influx of data, the ICTV has endorsed admitting new viruses solely identified from metagenomics data, departing from the historic need for experimental characterization of various features of the virus (Simmonds 2017).

The ICTV's shift to admitting new viruses solely identified by metagenomics data reemphasizes the usefulness of viral genomes in taxonomy and also a greater need to understand viral genomics, but experimental characterizations are still necessary. Indeed new viral genomics data has already complicated some historically used criteria for virus classification such as consideration of genome organization and, sometimes, host. The use of RNA-sequencing technology on 70 species of arthropods identified chuviruses that form a monophyletic cluster yet contain segmented and unsegmented species, with a potential circular form (Li 2015). Limitations of identifying new viruses solely from metagenomics data, such as claims of whether the virus's host was the species sequenced and not of the host's microbiome or meal and genome segmentation, will continue to place emphasis on the need for experimental characterization. Some claims necessarily require experimental characterization. One recent example comes from the report of the first multicomponent animal RNA virus isolated from mosquitoes (Ladner 2016). The

packaging of different genome segments within separate particles have long been an enigma given its precarious mode of transmission, so this transmission strategy was thought to be limited to hosts where cytoplasmic exchange was possible (Holmes 2016). Thus, reporting a multicomponent virus of animals will require strong experimental evidence, especially given that metagenomics data alone cannot verify such a claim.

New RNA viral genomes will add insight to the current evolutionary space that RNA viruses seem to occupy. Currently, no RNA viruses that have been identified infect archaebacteria (Adriaenssens 2018). Such a discovery will depend on both strong genetic and experimental characterizations.

Comparative studies of currently known viral genomes suggest that RNA viruses occupy a distinct evolutionary space from DNA viruses. Compared to DNA viruses, aside from having an RNA genome, RNA viruses have a much smaller genome size, narrower range of genome sizes, simpler genomes, and their genomes experience a much higher mutation rate. The genome sizes of known RNA viruses range from 2,500nts (Hillman 2013) to a maximum of about 31,000nts (Gorbalenya 2006) with a median of about 10,000nts. In contrast, the genome sizes of DNA viruses can range from about 1,200nts (Rohde 1990) to up to 2.5Mnts (Abergel 2015).

The smallest DNA virus encodes six proteins (Rohde 1990) while the smallest RNA virus encodes a single protein and lacks a capsid protein (Hillman 2013). Giant DNA viruses may also encode proteins for translation. The genome of the giant DNA tupanvirus currently has the most complete translational apparatus of all known viral genomes, encoding up to 70 tRNAs, 20 aminoacyl tRNA synthases, and 11 translation factors (Abrahão 2018). RNA viruses usually do not encode more than 12 genes.

Another defining feature of RNA viruses is that they all minimally encode a gene for either an RNA-dependent RNA polymerase (RdRp) or reverse transcriptase (RT) and it is this universal feature that can explain why RNA viral genomes experience a much higher mutation rate, and perhaps why their genomes are so limited in size. Nearly all examined RNA viruses show an overall mutation rate in the range of 10^{-2}-10^{-5} with most centering at 10^{-3} subs/site/year (Hanada 2004; Jenkins 2002). In contrast, DNA viruses have mutation rates similar to eukaryotic cells on the order of 10^{-8} to 10^{-10} subs/site/year (Holmes 2009; Fleischmann 1996). Consistently, DNA polymerases encoded by DNA viruses have proofreading functions, as is true of eukaryotic DNA polymerase, while no RNA polymerase is known to have this function. Attempts at constructing phylogeny from RdRp sequences of RNA viruses have been unsuccessful, but very short amino acid motifs can be located within the conserved palm subdomain structure (Poch 1989; Gorbalenya 2002; Shackelton 2008).

Occupying such a unique evolutionary space, governed by a limited genome size and high mutation rates, places great emphasis on the accuracy of understanding RNA viral genomes. Newly discovered and accurately described RNA viruses will deepen the understanding of the impact that RNA viruses have on life, its evolution, perhaps earth itself, and may reveal how all RNA viruses are related, thus providing updated means for classification and perhaps understanding their evolution.

This dissertation will present works related to accurately describing and discovering such viral genomes. Specifically, these works will focus on the genomes of small RNA viruses. Though there is no general consensus on the term “Small RNA virus” and the term is used flexibly, a small RNA viral genome is defined here as an RNA viral genome that is half the median genome size of RNA viruses, thus these genomes will have lengths of 5,000 (nts or bps) or less, and encoding no more

than two genes. This dissertation will be broken into three sections. In Section I, I will describe proposed amendments to previously reported genome sequences and classifications based on general understandings of the involvement of viral genomes in virus replication. In Section II, I will present validations of newly discovered small RNA viruses and subviral agents of such identified from RNA-sequencing data. In Section III, I will present how the understanding of genomes in virus families was used to develop a virus-mediated genetics system that can be used to experimentally characterize the replicative features of other virus family members at the nucleotide level.

1.2 References

Abergel C, Legendre M, Claverie JM. The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. *FEMS Microbiol Rev*. 2015 Nov;39(6):779-96.

Abrahão J, Silva L, Silva LS, Khalil JYB, Rodrigues R, Arantes T, Assis F, Boratto P, Andrade M, Kroon EG, Ribeiro B, Bergier I, Seligmann H, Ghigo E, Colson P, Levasseur A, Kroemer G, Raoult D, La Scola B. Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. *Nat Commun*. 2018 Feb 27;9(1):749.

Adriaenssens EM, et al. Taxonomy of prokaryotic viruses: 2017 update from the ICTV Bacterial and Archaeal Viruses Subcommittee. *Arch Virol*. 2018 Jan 2; doi: 10.1007/s00705-018-3723-z.

Ashley J, Cordy B, Lucia D, Fradkin LG, Budnik V, Thomson T. Retrovirus-like Gag Protein Arc1 Binds RNA and Traffics across Synaptic Boutons. *Cell*. 2018 Jan 11;172(1-2):262-274.

Barré-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, Gruest J, Dauguet C, Axler-Blin C, Vézinet-Brun F, Rouzioux C, Rozenbaum W, Montagnier L. Isolation of a Tlymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*. 1983; 220:868–871.

Bergh O, Borsheim KY, Bratbak G, Heldal M. High abundance of viruses found in aquatic environments. *Nature* 1989, 340:467-468.

Bernal JD, Fankuchen I. X-ray crystallographhic studies of plant virus preparation. *J Gen Physiol*. 1941 Sep 20;25(1):111-46.

Brum JR, et al. Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science*. 2015; 348: 1261498.

Dawood FS, Iuliano AD, Reed C, Meltzer MI, Shay DK, Cheng PY, Bandaranayake D, Breiman RF, Brooks WA, Buchy P, Feikin DR, Fowler KB, Gordon A, Hien NT, Horby P, Huang QS, Katz

MA, Krishnan A, Lal R, Montgomery JM, Mølbak K, Pebody R, Presanis AM, Razuri H, Steens A, Tinoco YO, Wallinga J, Yu H, Vong S, Bresee J, Widdowson MA. Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. *Lancet Infect Dis*. 2012 Sep;12(9):687-95.

Dupressoir A, Lavialle C, Heidmann T. From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. *Placenta*. 2012; 33: 663-671.

Fleischmann WR Jr. Chapter 43-Viral Genetics. In: Baron S, editor. Medical Microbiology. 4th edition. Galveston (TX): University of Texas Medical Branch at Galveston; 1996. Chapter 43.

Frank H, Moebus K. An electron microscopic study of bacteriophages from marine waters. *Helgol. Meeresunters.* 1987;41:385–414.

Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ. *Nidovirales*: evolving the largest RNA virus genome. *Virus Res*. 2006 Apr;117(1):17-37.

Gorbalenya AE, Pringle FM, Zeddam JL, Luke BT, Cameron CE, Kalmakoff J, Hanzlik TN, Gordon KH, Ward VK. The palm subdomain-based active site is internally permuted in viral RNA-dependent RNA polymerases of an ancient lineage. *J Mol Biol*. 2002 Nov 15;324(1):47-62.

Gubler DJ. Epidemic dengue/dengue hemorrhagic fever as a public health, social and economic problem in the 21st century. *Trends Microbiol*. 2002;10:100–103.

Hanada K, Suzuki Y, Gojobori T. A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. *Mol Biol Evol*. 2004 Jun;21(6):1074-80.

Hillman BI, Suzuki N. Viruses of the chestnut blight fungus, *Cryphonectria parasitica*. *Adv Virus Res*. 2004;63:423-72.

Hillman BI, Cai G. The family *Narnaviridae*: simplest of RNA viruses. *Adv Virus Res*. 2013;86:149-76.

Holmes EC. The Evolution and Emergence of RNA Viruses. Oxford University Press. 2009. Chapter 3.

Holmes EC. The Expanding Virosphere. *Cell Host Microbe*. 2016 Sep 14; 20(3):279-280.

Ivanofsky D. Concerning the mosaic disease of the tobacco plant. *St Petersburg Acad Imp Sci Bul* 1892; 35: 67-70.

Jenkins GM, Rambaut A, Pybus OG, Holmes EC. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J Mol Evol*. 2002 Feb;54(2):156-65.

Knight-Jones TJ, Rushton J. The economic impacts of foot and mouth disease – What are they, how big are they and where do they occur? *Prev Vet Med*. 2013 Nov 1;112(3-4):161-73.

Krupovic M, Ghabrial SA, Jiang D, Varsani A. *Genomoviridae*: a new family of widespread singlestranded DNA viruses. *Arch. Virol*. 2016; 161: 2633–2643.

Ladner JT, et al. A Multicomponent Animal Virus Isolated from Mosquitoes. *Cell Host Microbe*. 2016 Sep 14;20(3):357-367.

Lesnaw JA, Ghabrial SA. Tulip breaking: past, present, and future. *Plant Disease*. 2000; 84: 1052-1060.

Li C-X, Shi M, Tian J-H, Lin X-D, Kang Y-J, Qin X-C, Chen L-J, Xu J, Holmes EC, Zhang Y-Z. Unprecedented RNA virus diversity in arthropods reveals the ancestry of negative-sense RNA viruses. *eLife*. 2015; 4: e05378.

Marie D, Brussard CPD, Thyrhaug R, Bratbak G, Vaulot D. Enumeration of marine viruses in culture and natural samples by flow cytometry. Appl. Environ. *Microbiol*. 1999; 65:45–52.

Mi S, Lee X, Li X-P, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang X-Y, Edouard P, Howes S, Keith JC, McCoy JM. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature.* 2000; 403: 785-789.

Moreno P, Ambros S, Albiach-Marti MR, Guerri J, Pena L. Citrus tristeza virus: a pathogen that changed the course of the citrus industry. *Mol. Plant Pathol*. 2008; 9: 251–268.

Nagasaki K, Shirai Y, Takao Y, Mizumoto H, Nishida K, Tomaru Y. Comparison of genome sequences of single-stranded RNA viruses infecting the bivalve-killing dinoflagellate *Heterocapsa circularisquama*. *Appl Environ Microbiol*. 2005; 71: 8888–8894.

Nibert ML, Manny AR, Debat HJ, Firth AE, Bertini L, Caruso C. A barnavirus sequence mined from a transcriptome of the Antarctic pearlwort *Colobanthus quitensis*. *Arch Virol*. 2018a Mar 7. doi: 10.1007/s00705-018-3794-x.

Nibert ML, Vong M, Fugate KK, Debat HJ. Evidence for contemporary plant mitoviruses. *Virology*. 2018b Feb 10;518:14-24.

Noble RT, Fuhrman JA. Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. *Aquat. Microb*. *Ecol*. 1998. 14:113–118.

Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC. Uncovering Earth's virome. *Nature.* 2016 Aug 25; 536(7617):425-30.

Pastuzyn ED, Day CE, Kearns RB, Kyrke-Smith M, Taibi AV, McCormick J, Yoder N, Belnap DM, Erlendsson S, Morado DR, Briggs JAG, Feschotte C, Shepherd JD. The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. *Cell*. 2018 Jan 11;172(1-2):275-288.

Peiris JS, Yuen KY, Osterhaus AD, Stöhr K. The severe acute respiratory syndrome. *N Engl J Med*. 2003 Dec 18; 349(25):2431-41.

Petersen LR, Jamieson DJ, Powers AM, Honein MA. Zika Virus. *N Engl J Med*. 2016 Apr 21;374(16):1552-63.

Poch O, Sauvaget I, Delarue M, Tordo N. Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. *EMBO J*. 1989; 8: 3867–3874.

Pyle JD, Keeling PJ, Nibert ML. Amalga-like virus infecting Antonospora locustae, a microsporidian pathogen of grasshoppers, plus related viruses associated with other arthropods. *Virus Res*. 2017 Apr 2;233:95-104.

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

Rohde W, Randles JW, Langridge P, Hanold D. Nucleotide sequence of a circular single-stranded DNA associated with coconut foliar decay virus. *Virology*. 1990 Jun;176(2):648-51.

Shackelton LA, Holmes EC. The role of alternative genetic codes in viral evolution and emergence. *J Theor Biol* 2008; 254:128–134.

Sharif J, Shinkai Y, Koseki H. Is there a role for endogenous retroviruses to mediate long-term adaptive phenotypic response upon environmental inputs? *Philosophical Transactions of the Royal Society.* 2012; 368:20110340.

Shirai Y, Tomaru Y, Takao Y, Suzuki H, Nagumo T, Nagasaki K. Isolation and characterization of a single-stranded RNA virus infecting the marine planktonic diatom *Chaetoceros tenuissimus Meunier*. *Appl Environ Microbiol*. 2008; 74: 4022–4027.

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. Consensus statement: virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol*. 2017; 15: 161–168.

Steward GF, Culley AI, Mueller JA, Wood-Charlson EM, Belcaid M, Poisson G. Are we missing half of the viruses in the ocean? *ISME J.* 2013; 7: 672–679.

Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, Chisholm SW. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biology*. 2006;4: 1344-1357.

Tai V, Lawrence JE, Lang AS, Chan AM, Culley AI, Suttle CA. Characterization of HaRNAV, a singlestranded RNA virus causing lysis of Heterosigma akashiwo (Raphidophyceae). *J Phycol*. 2003; 39: 343–352.

Taylor LH, Latham SM, Woolhouse ME. Risk factors for human disease emergence. *Philos Trans R Soc Lond B Biol Sci*. 2001;356: 983–9.

Woolhouse ME, Gowtage-Sequeria S. Host range and emerging and reemerging pathogens. *Emerg Infect Dis*. 2005 Dec;11(12):1842-7.

World Health Organization (WHO). Ebola Virus Disease Situation Report. 2016 June 10.

Section I: Proposed changes to previously reported genome sequences

Chapter Two: Amendments to Sequence of Previously Published Genome of Zygosaccharomyces bailii virus Z

The figures in this chapter were taken from its publication:

Depierreux D, Vong M, Nibert ML. Nucleotide sequence of Zygosaccharomyces bailii virus Z: Evidence for +1 programmed ribosomal frameshifting and for assignment to family *Amalgaviridae*. *Virus Res.* 2016 Jun 2; 217: 115-24.

Special thank yous and contributions:

I want to give a very special thank you to Delphine Depierreux for doing the experiments in this chapter. Without her contribution this work would not be possible. I taught and showed Delphine all the steps of all the experiments performed, including the theory behind each step. I also mentored Delphine on the scientific process and helped troubleshoot her experiments. Max supervised the project from inception to publication.

2.1 Introduction

The Zygosaccharomyces bailii virus Z (ZbV-Z) was first identified by its dsRNA genome during a screen of 40 strains from the genus *Zygosaccharomyces*. The screen was focused on identifying yeast strains with killer activity and had identified strain *Zygosaccharomyces bailii* 412 as a positive hit (Radler 1993). Two expected dsRNAs required for conferring killer activity were present and named L (size ~4kbp) and M (size ~2kbp). The two dsRNAs were presumed related to the characterized killer system of *Saccharomyces cerevisiae*, which requires the dsRNA L-A virus, member of virus family *Totiviridae*, and its toxin-encoding M-satellite (Wickner 2013). A third uniquely sized dsRNA of about 3kbp also appeared within *Z. bailii* 412 strain and would be later followed by another group and given its name ZbV-Z (Z for *Zygosaccharomyces*) (Schmitt 1994).

Interestingly, upon characterizing the virus particles of 412, Schmitt and Neuhausen reported that the Z-dsRNA was encapsidated by a 35-kDa protein. In contrast, the other two dsRNAs were encapsidated in an 85-kDa protein, consistent with the other killer system of *Totiviridae* (Schmitt 1994). The Z-dsRNA was not yet sequenced during their report but would later be determined, without an associated publication (GenBank AF224490).

While ZbV-Z has not yet been formally classified by ICTV, several authors have made contrasting claims of its classification. Wickner et al. suggested the virus family *Totiviridae* while Liu at el. observed that it may possibly be better classified in *Amalgaviridae* (Wickner 2012; Liu 2012).

Three important observations were found upon inspection of the ZbV-Z genome. First, its genome contains two nonoverlapping ORFs, ORF1 encodes a 34-kDa protein and ORF2 encodes a RdRp. ORF2 is in-frame with ORF1 separated by ORF1 by 190nts. Many members of

Totiviridae also have these two ORFs, but most members express their ORF2 as a fusion protein via a ribosome frameshift with ORF1 (Wickner 2012). Within the ZbV-Z genome, the region between the two ORFs appears to contain a unique slippery site "CUU_UUU_CGA," where the CGN is a rare Arg codon that presumably plays a role in the programmed ribosome frameshifting (PRF). Further, this site is related to a +1 PRF site that was characterized in 2012 (Firth 2012). The second observation comes from the results of doing a NCBI PSI-BLAST showing that the ZbV-Z RdRp is more closely related to viruses from the family *Amalgaviridae*, *Partitiviridae*, and some proposed unirnavirus. The ZbV-Z RdRp sequence seems to only be distantly related to the

Table 1 NCBI PSI-BLAST search results for ZbV-Z RdRp (GenBank AAF37275): Top 10 hits based on E values

Virus	Family:genus	GenBank no.	E values
Blueberry latent virus	*Amalgaviridae*:*Amalgavirus*	ADO14116	4e–17
Blueberry latent virus	*Amalgaviridae*:*Amalgavirus*	ADO14118	5e–17
Ustilaginoidea virens unassigned RNA virus HNND-1	"Unirnavirus"	AKM52549	1e–16
Beauveria bassiana RNA virus 1	"Unirnavirus"	CEF90232	1e–16
Rhododendron virus A	*Amalgaviridae*:*Amalgavirus*	ADM36020	2e–16
Beauveria bassiana RNA virus 1	"Unirnavirus"	AKC57301	7e–16
Fragaria chiloensis cryptic virus	*Partitiviridae*:*Deltapartitivirus*	AAZ06131	6e–15
Ustilaginoidea virens RNA virus M	"Unirnavirus"	AIT56395	7e–15
Rose cryptic virus 1	*Partitiviridae*:*Deltapartitivirus*	ABY60412	1e–14

RdRp sequences of *Totiviridae* (Table 1). Further, amalgaviruses also encode two ORFs, a putative capsid protein of about 40kDa for ORF1 and an RdRp for ORF2, which has the ORF2 expressed through a +1 PRF. Further, three amalgaviruses contain +1 slippery sites similar to that of the influenza A viruses (Firth 2012). The last observation stems from the first two, if a properly placed single nucleotide were to have been included within the 5' end ZbV-Z ORF2 sequence then this ORF would be extended by enough to eliminate the 190nts gap and overlap with ORF1 while allowing use of the putative +1 slippery site. Specifically, an additional nucleotide between positions 1088 and 1100 would extend ORF2 by 221 nts upstream, allowing it to overlap ORF1 in

the +1 frame as in the plant amalgaviruses (Fig. 1A & 1B). Given these observations, it seems likely that an error had occurred during the reporting of the ZbV-Z/412 genome.

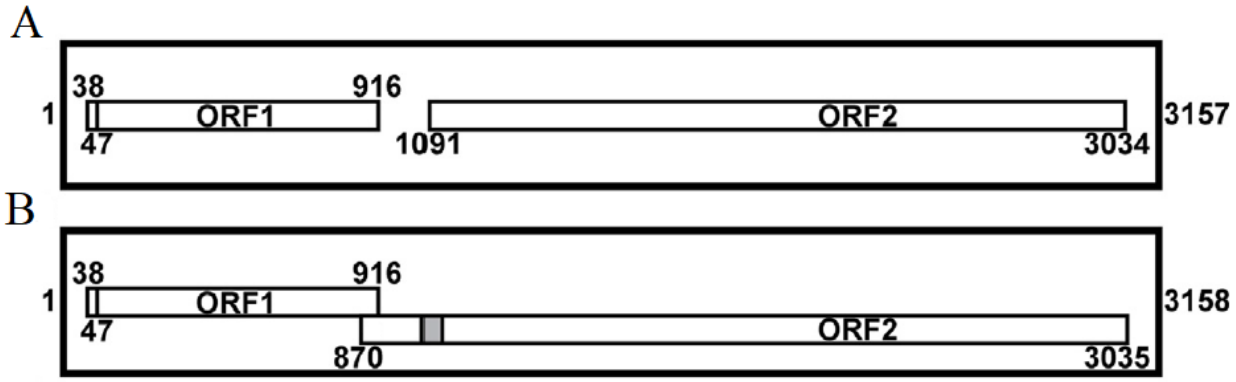


Figure 1 Genome organization of ZbV-Z. (1A) Genome organization of ZbV-Z based on GenBank AF224490. (1B) Our prediction for how a single nt insertion (within the gray-shaded area) would allow ORF2 to overlap ORF1 in the +1 frame.

We hypothesize that a single nucleotide is missing in the reported ZbV-Z genome (GenBank AF224490) between its genome positions 1088 and 1100. Thus, the previously studied *Z. bailii* 412 strain was obtained for the goal of re-sequencing the genome of ZbV-Z/412.

2.2 Materials and Methods

2.2.1 Yeast culture and harvest

A culture of *Z. bailii* 412 was obtained as a kind gift from Manfred J. Schmitt and Frank Breinig (Saarland University, Saar-brücken, Germany). *Z. bailii* 412 cells (from our lab clone DD1) were grown in liquid yeast-peptone-glucose (YPG) medium (1% yeast extract, 2% peptone, 2% glucose) at 30˚C with shaking at 250 rpm until it reached an optical density of ~1.0 as measured at 600 nm. Cells were then harvested by centrifugation at 1500g for 5 min at 4 ˚C. Resulting pellets (from 10 mL of culture each) were re-suspended in 1 mL cold deionized water, flash-frozen, and stored at −80 ˚C until use.

2.2.2 RNA purification

The yeast pellet was re-suspended in 250 µL of RNA buffer (500 mM NaCl, 10 mM EDTA, 200 mM Tris plus HCl to pH 7.5) and 750 µL of TRIzol™ LS reagent (Ambion). Equal volume of glass beads (diameter, 0.5 mm; BioSpec) was then added. Samples were incubated for 5 min at room temperature, after which 150 µL of chloroform was added and the sample was subjected to vortex agitation for 2 min. The TRIzol™ extraction protocol was then completed per the manufacturer's instructions. To enrich for dsRNA from the TRIzol™ extract of total RNA, treatment with microgranular cellulose (cellulose powder MN 301, Macherey-Nagel) was performed largely as described by Castillo et al. (Castillo 2011): Cellulose was equilibrated with STE buffer (100 mM NaCl, 50 mM Tris plus HCl to pH 7.5, 1 mM EDTA) plus 16% ethanol, followed by pelleting and removal of the buffer, and then resuspension in 1.8 mL fresh STE buffer plus ethanol. Total RNA from a TRIzol™ extraction was then added in a volume of ≤200 µL, followed by incubation with vortex agitation for 1 hr. The suspension was pushed through an empty mini-column (Promega) and the retained cellulose was washed with 2 mL fresh STE buffer plus ethanol. After placing the mini-column in an empty microtube, excess buffer was removed by microcentrigation at 13,000g for 1 min. After transferring the mini-column to a new empty microtube, 20 µL of STE buffer without ethanol was added, followed by microcentrigation at 13,000g for 1 min. The collected eluate was then reapplied to the mini-column, followed again by microcentrigation at 13,000g for 1 min to yield the final eluate containing dsRNA. Enriched dsRNA was separated and visualized by agarose gel electrophoresis.

2.2.3 Sequence determination

We used the previously reported ZbV-Z sequence (GenBank AF224490) to design oligonucleotide primers for performing RT–PCR on total RNA extracted from *Z. bailii* 412. The primers were designed to amplify four overlapping regions of the ZbV-Z/ 412 genome: positions 63–1070, 570–1580, 1473–2543, and 2194–3133 (numbered according to GenBank AF224490). The RT step was performed according to manufacturer's instructions using SuperScript III First-Strand Synthesis System (Invitrogen), except that the reaction was allowed to incubate for 45 min at 55 ◦C, followed by 15 min at 70˚C. The PCR step was performed according to manufacturer's instructions using Taq DNA Polymerase with Standard Taq Buffer (NEB), except that 34 cycles were performed with denaturation at 95˚C for 30 sec and hybridization at 50˚C for 30 sec per cycle. PCR products were separated by agarose gel electrophoresis and stained with ethidium bromide. The DNA bands were excised and purified to manufacturer's instructions using QIAquick Gel Extraction Kit (Qiagen). DNA and the same primers as used for PCR were then sent to the Dana Farber/Harvard Cancer Center DNA Resource Core for Sanger sequencing in both directions for each amplicon.

For *de novo* determination of the terminal sequences of ZbV-Z/412, RLM-3' RACE was performed on cellulose-purified dsRNA using modifications of previously described methods (Coutts 2003). The dsRNA was first denatured by incubation in 90% DMSO for 15 min at 65˚C. The denatured dsRNA was precipitated by addition of NaCl to a concentration of 150 mM, plus 2.5 volumes of 100% ethanol. The resulting RNA pellet was washed with 70% ethanol and then air-dried, followed by resuspension in ligation mix (1X T4 RNA Ligase Reaction Buffer (NEB), 1 mM DTT, 1 mM ATP, 2 U/µL RNasin (Promega), 20% PEG 8000 (NEB), 40 pmoles DNA adapter 5'phosphate-CAATACCTTCTGACCATGCAGTGACAGTCAGCATG-3'amino

modifier (IDT), and 10 units T4 RNA ligase 1 (NEB), plus DEPC-treated water treated to reach 20 µL). Ligation was allowed to proceed for 16 hr at 16˚C. Next, 80 µL of DEPC-treated water and 1 mL of TRIzol™ LS reagent were added, and the sample was incubated at room temperature for 5 min, followed by addition of 200 µL chloroform and centrifugation at 1200g for 5 min at 4˚C before recovering the aqueous phase. 10 µL 3M sodium acetate pH 5.2 and 220 µL 95% ethanol were then added, and the sample incubated on dry ice for 10 min before centrifugation at 13,000g for 20 min at 4˚C. The resulting pellet was washed with 70% ethanol and air dried, then resuspended in 10 µL DEPC-treated water. The RT step was performed as described above, but this time using outer anti-adapter primer CATGCTGACTGTCACTGCATGG. PCR amplifications were performed as described above, with each reaction not including the outer anti-adapter primer and an internal primer based on the ZbV-Z/412 sequence beginning at nt position 283 or 2850 for the amplicon corresponding to the 5' or 3' end of plus strand, respectively. The amplicon corresponding to the 5' end of the plus strand was obtained and prepared for sequencing as described above. The amplicon corresponding to the 3' end of plus strand was not obtained, however, and we therefore performed a second, nested reaction using inner anti-adapter primer CTGTCACTGCATGGTCA-GAAGG and an internal primer based on the ZbV-Z/412 sequence beginning at nt position 2884. Nested PCR included the following steps: initial denaturation at 95◦C for 2 min, followed by 24 cycles of denaturation at 95◦C for 30 sec, hybridization at 58˚C for 30 sec, and elongation at 72◦C for 30 s, followed by a final extension at 68◦C for 8 min. The amplicon was obtained in this case and prepared for sequencing as described above. Sanger sequencing was performed in both directions for both RLM-3' RACE amplicons, and the sequences were found to match 100%, in the regions of overlap, with the sequences obtained as described in the preceding paragraph.

2.2.4 Sequence-based analyses

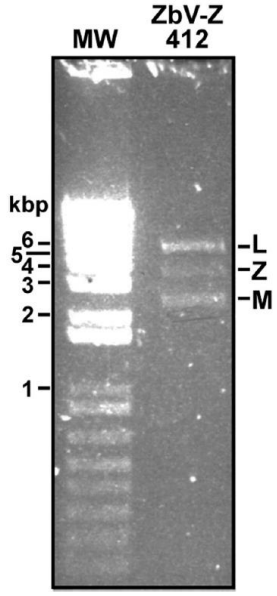
Searches of the NR database with protein sequence queries deduced from the nucleotide sequences were performed using NCBI PSI-BLAST (Schäffer 2001) as implemented with defaults at http://blast.ncbi.nlm.nih.gov/Blast.cgi. ORFs were identified in nucleotide sequences using EMBOSS getorf as implemented with defaults at http://www.bioinformatics.nl/emboss-explorer/. Molecular weight and pI values for proteins were calculated using Compute pI/MW as implemented with defaults at http://web.expasy.org/compute_pi/. Global pairwise comparisons of protein sequences were performed using Needle or Needleall as implemented at http://www.bioinformatics.nl/emboss-explorer/, with the following parameter differing from the defaults: Apply end gap penalties, Yes. As a simple convention for comparing sequences from the different viruses in the absence of certainty about translational initiation and frameshifting sites, the ORF1 and ORF2 translation product sequences began with the first Met codon in each of the two ORFs, except for the comparisons involving the ORF1 translation product of UvRV-M (GenBank KJ101567). Local pairwise comparisons of nucleotide sequences were performed with EMBOSS Matcher as implemented with defaults at http://www.ebi.ac.uk/Tools/psa/.

For phylogenetic analyses, multiple sequence alignments were performed with the RdRp sequences using MAFFT 7.27 (L-INS-i) (Katoh 2013) as implemented with defaults at http://mafft.cbrc.jp/alignment/server/. The convention of using sequences beginning with the first Met residue in the RdRp-encoding ORF was applied again here, including for Phylogenetic relationships were then determined using PhyML 3.0 (Guindon 2010) as implemented at http://www.hiv.lanl.gov/content/sequence/PHYML/interface.html with the following parameters differing from the defaults: Sequence type/model, Amino acids/LG; Proportion of invariable sites, estimated from data; Gamma shape parameter, estimated from data; Starting tree(s) optimization,

Tree topology and Branch length; Tree improvement, Best of NNI and SPR; Branch support, Approximate Likelihood Ratio Test (aLRT), SH-like supports. The results in Newick format were then submitted to TreeDyn 198.3 as implemented at http://www.phylogeny.fr/ for displaying branch support values in % and collapsing branches with support values below 50%. The output in Newick format was then opened in FigTree v1.4.0 (downloaded from http://tree.bio.ed.ac.uk/software/figtree/) for refining the phylogram for presentation. Values estimated from the data were Proportion of invariable sites, 0.018, and Gamma shape parameter, 2.445. Estimated values were proportion of invariable sites, 0.003, and Gamma shape parameter, 1.643. Alternative use of the RtREV substitution model for PhyML 3.0 yielded similar results.

2.3 Results

2.3.1 Visualization of ZbV-Z/412 dsRNA genome



A culture of *Z. bailii* 412 from Manfred J. Schmitt and Frank Breinig was used to start our lab clone DD1. DsRNA from ZbV-Z/412-DD1 was enriched and visualized by agarose gel electrophoresis according to Materials and Methods, section 1.2.2. Three bands were observed (Fig. 2), consistent in size with the expected L (~4.0 kbp), Z (~3.0 kbp), and M (~2.0 kbp) dsRNA segments carried by previously reported *Z. bailii* 412 (Radler 1993; Schmitt 1994). We then aimed to re-sequence the ZbV-Z genome.

Figure 2 Enriched dsRNA from lysate of *Z. bailii* 412

2.3.2 Redetermination of ZbV-Z/412 genome sequence

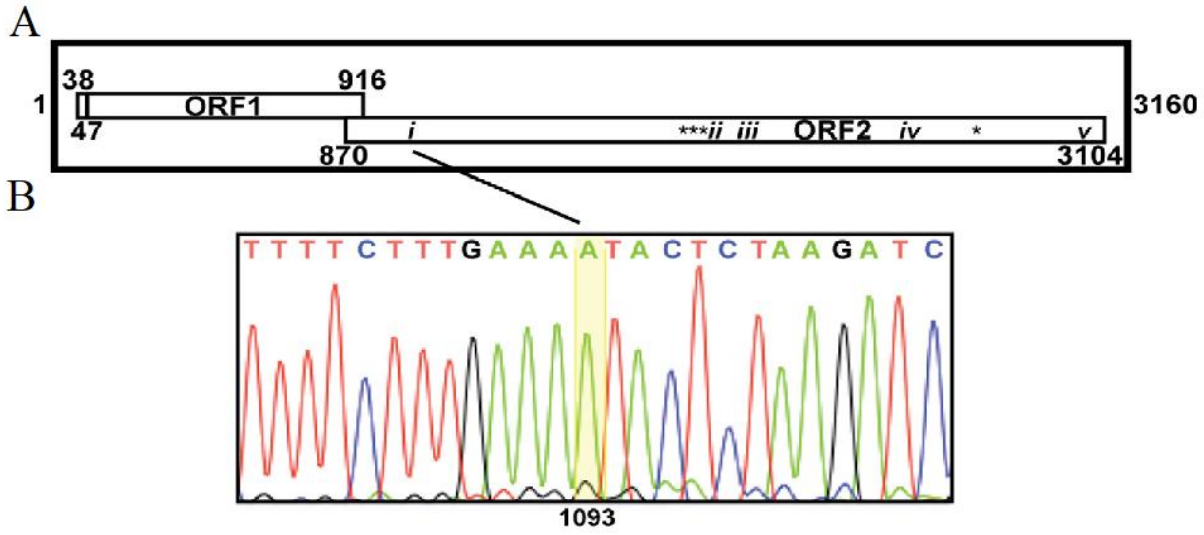


Figure 3 New genome organization of resequenced ZbV-Z. (3A) Positions of the respective indels: i-iv; Positions of four single-nucleotide substitutions: *. (3B) Sequencing electropherogram across the region of indel i in the ZbV-Z plus strand.

The full-length nucleotide sequence of ZbV-Z/412 was determined according to Materials and Methods, section 1.2.3, and it has been deposited in GenBank with accession number KU200450. The new sequencing results for ZbV-Z/412 include five insertions or deletions (indels) relative to GenBank AF224490 (Fig. 3A): (i) a 1-nt insertion after position 1092 (run of 4, not 3, As) (shown in Fig. 3B); (ii) a 2-nt deletion at positions 1811 and 1812 (run of 2, not 4, Ts); (iii) a 2-nt insertion after position 1845 (run of 6, not 4, Ts); (iv) a 3-nt insertion (TTG) after position 2331; and (v) a 1-nt deletion of position 3015 (run of 2, not 3, Gs). Indel (i) introduces a frame shift that extends ORF2 by 221 nts upstream such that ORF2 now over-laps ORF1 by 47 nts in the +1 frame (Fig. 3A & 3B). Notably, this is precisely the type of prediction that led us to undertake this study. Indel (ii) introduces a frame shift that changes the encoded protein sequence over the downstream 10 residues and indel (iii) introduces a compensatory frame shift that returns the reading frame to the same as that before indel (ii). Indel (iv) results in insertion of a new tryptophan

residue into the encoded protein sequence. Indel (v) introduces a frame shift that extends ORF2 by 45 nts downstream, changing the formerly last 7 residues of the encoded protein sequence and also extending it by 15 residues to the next downstream stop codon in the new reading frame.

In addition to these five indels, our results for ZbV-Z/412 identified four single-base substitutions relative to GenBank AF224490 at positions 1805, 1809, 1810, and 2619, yielding three further changes to the encoded protein sequence (Fig. 3A). All of these observed indels and substitutions are located within ORF2.

2.3.3 Updated genome organization of ZbV-Z/412

Our results showed that the overall length of the ZbV-Z/412 genomic plus strand to be 3160 nts, from 5'-GUAAAAGAAC to UAUGCCUUGG-3'. ORF1 spans positions 38–916, between stop codons at positions 35–37 and 917–919 (Fig. 3A). Its 5'-most AUG codon is at positions 47–49 and is in a strong sequence context for translation initiation (AGUAUGG). The protein-coding region ofORF1 is therefore predicted to span positions 47–916 and to encode a 290-aa, 34-kDa putative CP (pI 6.2). These details for ORF1 and its predicted translation product are the same as those from GenBank AF224490. ORF2 spans positions 870–3104, between stop codons at positions 867–869 and 3105–3107 and thereby overlaps ORF1 by 47 nts (positions 870–916) (Fig. 3A). Within this 47-nt over-lap is found the sequence CUU_UUU_CGA (underlines, ORF1 codon boundaries) at positions 905–913 (Fig. 3A), representing a putative +1 PRF motif per Firth et al. (Firth 2012), as described in Introduction. If a +1 PRF indeed occurs near the 3'end of this motif, then the resulting ORF1/ORF2 fusion is expected to span positions 47–910:912–3104 and to encode a 1012-aa, 119-kDa putative CP/RdRp (pI 9.2). Based on these analyses, the

untranslated regions at the ends of the ZbV-Z/412 genomic plus strand are fairly short: 46 nts at the 5' end and 77 nts at the 3' end including the ORF2 stop codon (Fig. 3A).

2.3.4 Sequence comparison and phylogenetic analyses

To begin to address the phylogeny of ZbV-Z/412, we compared its ORF2-encoded amino acid (RdRp) sequence with the dsRNA viruses that came up as hits from the NCBI PSI-BLAST (Table 1). Given the original, tentative assignment of ZbV-Z to family *Totiviridae*, we included sequences representing the type species of the five approved genera in that family (Saccharomyces cerevisiae virus L-A from genus *Totivirus*, Helminthosporium victoriae virus 190 from genus *Victorivirus*, Giardia lamblia virus from genus *Giardiavirus*, Leishmania RNA virus 1 from genus *Leishmaniavirus* ,and Trichomonas vaginalis virus 1 from genus *Trichomonasvirus*). Also, given the findings in Table 1 described above, we included sequences representing the four approved species of plant viruses in genus *Amalgavirus*, family *Amalgaviridae* (Blueberry latent virus, Rhododendron virus A, Southern tomato virus, and Vicia cryptic virus M); sequences from six strains of mono-segmented dsRNA mycoviruses that constitute proposed genus *Unirnavirus* (Alternaria longipes dsRNA virus 1 from *A. longipes* isolate HN28 (AlDRV1), Beauveria bassiana RNA virus 1 from *B. bassiana* isolates EABb-92/11-Dm (BbRV1/1) and A24 (BbRV1/2), Penicilliumjanczewskii Beauveria bassiana-like virus 1 from *P. janczewskii* isolate MUT4359 (PjBbLV1), Ustilaginoidea virens RNA virus M from *U. virens* isolate GX-1 (UvRV-M), and Ustilaginoidea virens unassigned RNA virus from *U. virens* isolate HNND-1 (UvURV) (Jiang 2015; Koloniuk 2015; Kotta-Loizou 2015; Lin 2015; Nerva 2015; Zhu 2015); and sequences representing the type species of the five approved genera in family *Partitiviridae* (White clover cryptic virus 1 from genus *Alphapartitivirus*, Atkinsonella hypoxylon virus from genus

Betapartitivirus, Penicillium stoloniferum virus S from genus *Gamma-partitivirus*, Pepper cryptic virus 1 from genus *Deltapartitivirus*, and Cryptosporidium parvum virus 1 from genus *Cryspovirus*). The results from maximum-likelihood phylogenetic analyses provided evidence that ZbV-Z/412 represents a distinct taxon relative to the other analyzed viruses, more closely related to plant amalgaviruses and unirnaviruses than to *Totiviridae* and *Partitiviridae* members (Fig. 4). Sequence identity scores from pairwise comparisons of the ORF1 and ORF2 product sequences of ZbV-Z/412 with those of plant amalgaviruses and unirnaviruses were consistent with the phylogenetic results and provided further evidence that ZbV-Z/412represents a distinct taxon (Fig. 5).

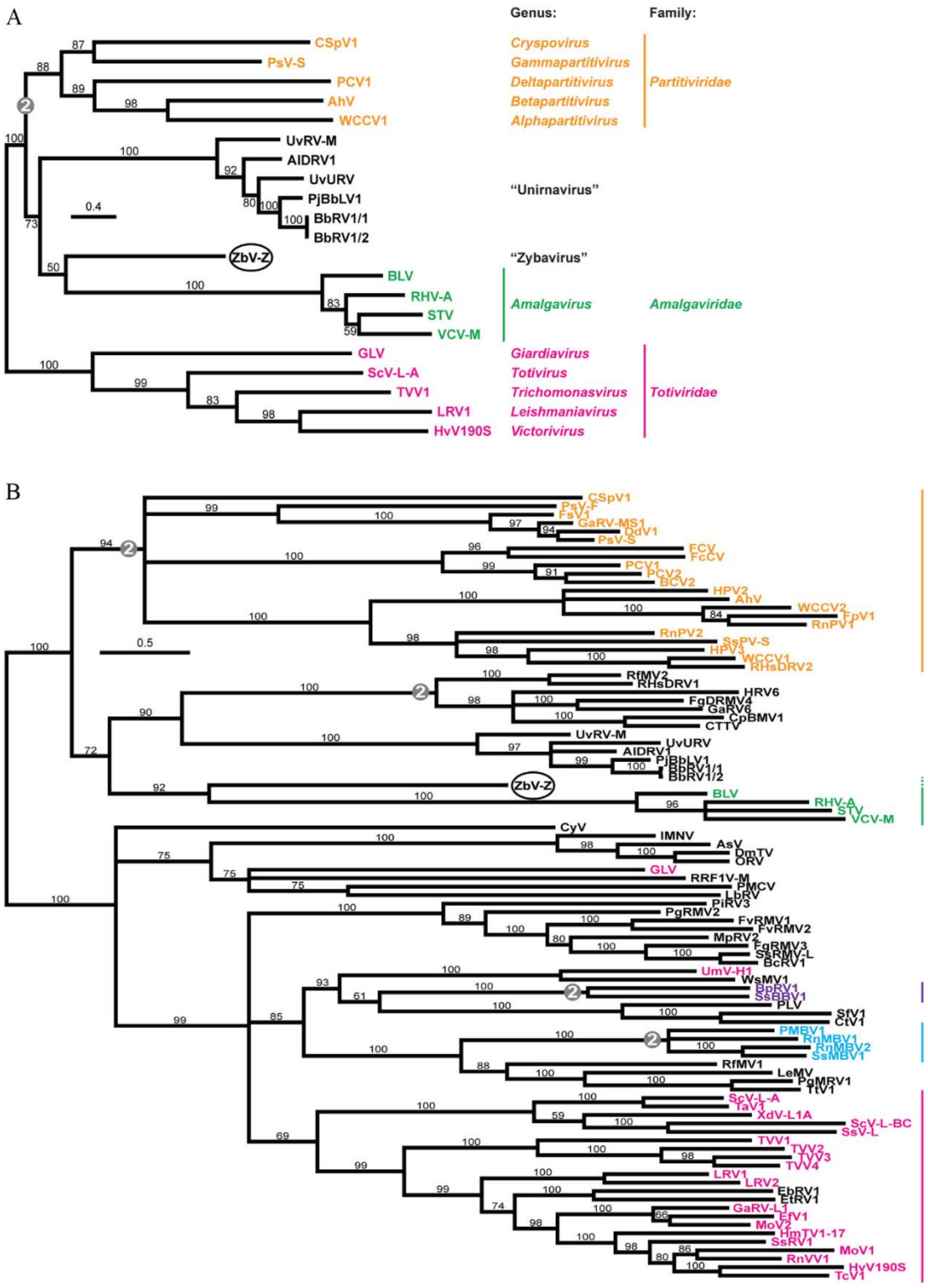


Figure 4 Phylogenetic analyses for ZbV-Z. (A) Names of approved and proposed genera are indicated. (B) Additional families in this panel are *Botybirnaviridae* (proposed; purple) and *Megabirnaviridae* (cyan). Family ranges are indicated by vertical lines, colored as in panel A. Our proposal to expand family *Amalgaviridae* to encompass ZbV-Z (proposed genus Zybavirus) is indicatedby the dotted portion of the green bar.

Thus, ZbV-Z/412 may be the prototype strain of a new species. ZbV-Z can serve as the type species of a new genus of mono-segmented dsRNA mycoviruses outside family *Totiviridae*. The provisional name "Zybavirus" is proposed for discussing this genus (Fig. 4A), as derived from the name of the prototype host *Zygosaccharomyces bailii*.

Viruses	BLV	RHV-A	STV	VCV-M	ZbV-Z	AlDRV1	BbRV1/1	BbRV1/2	PjBbLV1	UvRV-M	UvURV
BLV	*100*	45.8	43.2	41.8	18.6	19.2	19.4	19.4	17.8	18.8	17.5
RHV-A	20.9	*100*	44.7	46.5	18.2	20.3	20.9	20.2	18.2	16.3	18.0
STV	22.7	24.3	*100*	49.1	18.4	19.7	18.3	18.3	18.0	16.9	18.0
VCV-M	22.6	22.8	21.7	*100*	18.7	19.7	19.9	19.0	19.3	18.0	17.6
ZbV-Z	16.8	16.9	15.8	13.6	*100*	17.8	17.9	18.3	18.5	19.7	20.3
AlDRV1	16.0	17.9	16.1	17.6	16.1	*100*	57.2	57.2	61.0	44.7	51.7
BbRV1/1	17.2	18.1	18.8	16.5	16.7	32.2	*100*	97.6	71.9	42.4	52.0
BbRV1/2	20.0	18.3	18.6	14.9	16.4	32.7	97.8	*100*	71.9	42.7	52.3
PjBbLV1	18.0	16.4	16.2	17.2	15.8	33.0	60.8	61.4	*100*	42.4	52.8
UvRV-M	13.3	14.8	21.8	15.5	13.3	45.5	46.2	46.5	46.2	*100*	49.0
UvURV	18.1	16.4	19.1	17.5	14.7	31.5	44.1	44.7	41.8	43.0	*100*
	Amalgavirus				"Zybavirus"			"Unirnavirus"			

Figure 5 Pairwise identity scores of ORF1 and ORF2 translation products for ZbV-Z and other viruses.

2.4 Discussion

2.4.1 Taxonomic classification

Given the currently available number of viral genome sequences, a phylogenetic tree of ZbV-Z against various members of closely related RdRp-sequences from the NCBI PSI-BLAST does not clearly indicate where the family *Amalgaviridae* should be demarcated. Based on the RdRp-based phylogenetic tree in Fig. 4A, one might conclude (i) that proposed genera Zybavirus and Unirnavirus should both be assigned to family *Amalgaviridae*, (ii) that genus Zybavirus should be assigned to family *Amalgaviridae* but genus Unirnavirus should not, or (iii) that neither genus Zybavirus nor genus Unir-navirus should be assigned to family *Amalgaviridae*. In an effort to resolve this ambiguity, several additional analyses were performed.

Terminal sequences of the genomic RNA strands are often conserved among related dsRNA viruses, reflecting important roles in RNA transcription and/or replication. Among the

plant amalgaviruses, although their plus-strand 3' sequences appear more variable, their plus-strand 5' sequences are more conserved, with the consensus being 5'-GWWWWWWWWW (W = A or U). ZbV-Z/412 fits this consensus in part, with its plus-strand 5sequence being 5-GUAAAAGAAC. The consensus sequence for the plant amalgaviruses and ZbV-Z/412 combined is thus 5-GWWWWWNWW (N = potentially any base). The plus-strand 5' sequences reported to date for unirnaviruses are also more variable, perhaps because several of them are incomplete, such that they seem to add little to this analysis.

Table 2 Properties of amalgaviruses, ZbV-Z, and unirnaviruses relating to ORF2 translation by proposed PRF.

Virus	Genus	GenBank no.	ORF1 range	ORF2 range	Overlap (nt)	FS
Blueberry latent virus	*Amalgavirus*	HM029246	131–1291	930–3329	362	+1/-2
Rhododendron virus A	*Amalgavirus*	HQ128706	38–1306	696–3326	611	+1/-2
Southern tomato virus	*Amalgavirus*	EF442780	87–1268	976–3324	293	+1/-2
Zygosaccharomyces bailii virus Z	"Zybavirus"	AF224490	38–916	1091–3034	0	0
		KU200450	38–916	870–3083	47	+1/-2
Alternaria longipes dsRNA virus 1	"Unirnavirus"	KJ817371	298–1500	1428–3308	73	−1/+2
Beauveria bassiana RNA virus 1/1	"Unirnavirus"	LN610699	249–1265	1256–3097	10	−1/+2
Beauveria bassiana RNA virus 1/2	"Unirnavirus"	KM233415	212−1228	1219–3060	10	−1/+2
Penicillium janczewskii Beauveria bassiana-like virus 1	"Unirnavirus"	KT601106	≤3–951	915–2777	37	−1/+2
Ustilaginoidea virens RNA virus M	"Unirnavirus"	KJ101567	≤3–413	404–2236	10	−1/+2
Ustilaginoidea virens unassigned RNA virus HNND-1	"Unirnavirus"	KR106133	31–975	1032–2828	0	−1/+2
				936–2828	40	−1/+2

There are notable similarities within their genomes (Table 2 and Table 3). The genome lengths of all three genera fall within a range of 2890–3437 bps, their ORF1 product lengths in a range of 290–404 aa (excluding UvRV-M/GX-1 since its plus-strand sequence appears to be substantially truncated at its 5' end; see Table 2), and their predicted ORF1/ORF2product lengths in a range of 926–1077 aa (again excluding UvRV-M/GX-1). Table 3 further shows that even the uppermost values in these ranges are substantially smaller than the lowermost values of family *Totiviridae* members. Their distinctively similar genome and protein lengths are thus discrete characteristics that might be used to support the assignment of all genera to family *Amalgaviridae*.

Table 3 Additional properties of amalgaviruses, ZbV-Z, and unirnaviruses compared with *Totiviridae* family members.

Virus	Genus	GenBank no.	Genome length (bp)	Protein length (aa)	
				ORF1	ORF1/ORF2
Blueberry latent virus	*Amalgavirus*	HM029246	3431	375	1054
Rhododendron virus A	*Amalgavirus*	HQ128706	3427	404	1077
Southern tomato virus	*Amalgavirus*	EF442780	3437	377	1062
Vicia cryptic virus M	*Amalgavirus*	EU371896	3434	394	1057
Zygosaccharomyces bailii virus Z	"Zybavirus"	KU200450	3160	290	1012
Alternaria longipes dsRNA virus 1	"Unirnavirus"	KJ817371	3415	394	997
Beauveria bassiana RNA virus 1/1	"Unirnavirus"	LN610699	3218	315	926
Beauveria bassiana RNA virus 1 /2	"Unirnavirus"	KM233415	3173	315	926
Penicillium janczewskii Beauveria bassiana-like virus 1	"Unirnavirus"	KT601106	2890	317	926
Ustilaginoidea virens RNA virus M	"Unirnavirus"	KJ101567	(2714)	(137)	(745)
Ustilaginoidea virens unassigned RNA virus HNND-1	"Unirnavirus"	KR106133	2903	314	932
Saccharomyces cerevisiae virus L-A	*Totivirus*	J04692	4579	680	1505
Trichomonas vaginalis virus 1	*Trichomonasvirus*	HQ607513	4684	678	1429
Helminthosporium victoriae virus 190S	*Victorivirus*	U41345	5179	772	1607
Leishmania RNA virus 1	*Leishmaniavirus*	M92355	5284	741	1595
Giardia lamblia virus	*Giardiavirus*	L13218	6277	886	1870

There are three differences of note between the three genera. Unirnaviruses seem to have longer 5' untranslated regions and have their ORF2 in a -1 frame relative to its ORF1. And amalgaviruses have much longer regions of ORF1–ORF2 overlap than do the other viruses.

As a further examination of whether these three taxa might warrant assignment to the same family, we expanded the phylogenetic analyses to include RdRp sequences from other mono-and bi-segmented dsRNA viruses. We also added RdRp sequences representing other approved species in families *Totiviridae*, *Partitiviridae*, *Megabirnaviridae*, and Botybirnaviridae (proposed). The results provided new evidence that the RdRp of ZbV-Z/412 is most closely related to those of the plant amalgaviruses (Fig. 4B) and, combined with other findings such as shown in Table 3, that assignment of proposed genus Zybavirus to family *Amalgaviridae* appears reasonable. The results also provided new evidence that the RdRps of unirnaviruses, though related to those of plant amalgaviruses, are less related to them than is that of ZbV-Z/412 (Fig. 4B), making assignment of proposed genus Unirnavirus to family *Amalgaviridae* appear less well supported at the current stage of characterizing these viruses.

To extend the phylogenetic analyses, we attempted to compare the ORF1 product sequences of plant amalgaviruses, ZbV-Z, and unirnaviruses with those of other mono- and bi-segmented dsRNA viruses. The ORF1 product sequences, however, are more divergent than the RdRp sequences, limiting their simple utility in this regard. For example, when used as query in NCBI PSI-BLAST searches, the putative CP of ZbV-Z/412 did not identify either plant amalgavirus or unirnavirus ORF1 products as homologs. Similarly, when used as query in NCBI PSI-BLAST searches, any of the unirnavirus ORF1 products identified the other unirnavirus ORF1 products as homologs but not the ZbV-Z/412 or plant amalgavirus ORF1 products, and any of the plant amalgavirus ORF1 products identified the other plant amalgavirus ORF1 products as homologs but not the ZbV-Z/412 or unirnavirus ORF1 products. On the other hand, sequence identity scores from pairwise comparisons of the ORF1 products of these viruses are consistent with their assignment to three distinct genera (Fig. 5).

Recent studies have identified yet another distinct taxon of unclassified dsRNA mycoviruses with RdRps that are phylogenetically related to those of the plant amalgaviruses, but, in this case, viruses with two genome segments. These bi-segmented viruses include Cryphonectria parasitica bipartite mycovirus 1 from *C. parasitica* isolate 09269, Curvularia thermal tolerance virus from *C. protuberata* (CTTV, the prototype of this taxon), Fusarium graminearum dsRNA mycovirus 4 from *F. graminearum* isolate DK3, Rhizoctonia fumigata mycovirus from *R. fumigata* isolate C-314, and Rhizoctonia solani dsRNA virus 1 from *R. solani* isolate AG-1-IA-B275, and possibly also Gremmeniella abietina RNA virus 6 from *G. abietina* isolate P3-7 and Heterobasidion RNA virus 6 from *H. parviporum* isolate 195–12, for which only single genome segments have been identified to date. These viruses were also included in the analysis for Fig. 3B, which provides further evidence that they constitute a distinct taxon and that their RdRps are

most closely related to those of unirnaviruses and next most closely related to those of ZbV-Z and plant amalgaviruses. Koloniuk et al. (Koloniuk 2015) in particular have suggested that these bi-segmented mycoviruses might also warrant assignment to family *Amalgaviridae*. Based on the RdRp-based phylogenetic tree in Fig. 4B, however, assignment of this taxon to family *Amalgaviridae* does not appear to be well supported at present, similarly to the case for unirnaviruses.

2.4.2 Putative slippery sequences for +1 PRF in ZbV-Z, plant amalgaviruses, and unirnaviruses

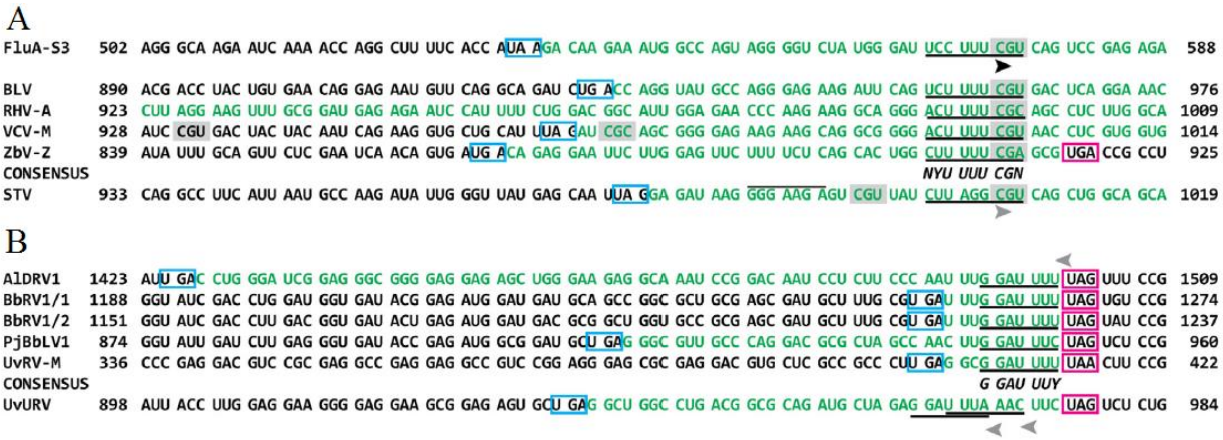


Figure 6 Amalgaviruses and ZbV-Z share a putative +1PRF motif while unirnaviruses share a putative -1PRF. Codon spacing is based on ORF1's reading frame. Cyan boxes indicate stop codon flanking 5' end of ORF1. Magenta boxes indicate stop codon for ORF2. Green letters indicate the sequences between the flanking stop codons. Putative slippery sites are underlined. Arrows indicate directin of predicted PRF. Rare Arg codon (CGN) is shaed in gray. (A) Amalgaviruses and ZbV-Z, with influenza A virus segment 3 (FluA-S3) in first row. (B) Unirnaviruses.

Our ZbV-Z genome sequence shows that ORF2 overlaps ORF1 in the +1 frame. This +1 frame organization is also seen in the four plant amalgaviruses described to date (Liu 2009; Martin 2011; Sabanadzovic 2009, 2010) (Table 2). Moreover, sequences similar to the motif for +1 PRF that was first characterized in the PA-X-encoding RNA segment of influenza A viruses (Firth 2012) are found also in the ORF1–ORF2 overlap regions of our genome sequence for ZbV-Z/412

and three of the plant amalgaviruses (Fig. 6A). Between ZbV-Z/412 and these three plant amalgaviruses, we can propose to define their consensus motif for +1 PRF as NYU_UUU_CGN (Y = T or C; N = any nucleotide; underlines, ORF1 codon boundaries), where the component sequence CGN is a rare Arg codon. This proposed consensus thus corresponds well with the motif originally defined for influenza A viruses (Fig. 6A), including the presence of a rare Arg codon at the demonstrated or proposed site of ribosomal slippage and the capacity for the P-site tRNA (anticodon 3'-AAA on codon UUU) to remain engaged after +1 slippage (moves forward to codon UUC) (Firth 2012). Southern tomato virus (STV) is notably the only characterized plant amalgavirus that has not been discussed in either the preceding paragraphs or the analysis by Firth (Firth 2012). The ORF1–ORF2 overlap region of multiple strains of STV (GenBank EU413670, KT438549, KT634055, and KT852573) lacks a sequence that strictly matches the aforementioned consensus motif for +1 PRF. It does, however, include a rare Arg codon (CGU) that is present in a somewhat similar sequence context as in the other plant amalgaviruses and ZbV-Z/412 (Fig. 6A). Of course, the major difference in STV is that the central codon in the motif is AGG, not UUU. Interestingly, however, this STV sequence might similarly allow the P-site tRNA (anticodon 3'UCC on codon AGG) to remain engaged after +1 slippage (moves forward to codon GGC, with G:U pairing in the first position). Other RNA or protein sequences that may be essential for or otherwise modulate the proposed +1 PRF activity in these viruses remain to be identified.

While the mechanism for translating RdRp from ORF2 of the viruses in proposed genus *Unirnavirus* also remains unclear, there exists contrasts between amagalaviruses and uniranaviruses. By examining uniranaviruses' genome sequences we found that in five of the six unirnavirus strains described to date, ORF2 overlaps ORF1 by 10–73 nt in the −1 frame (Table 2, Fig. 6B). This finding seems to have been overlooked in some previous reports, possibly due to

defining the upstream end of ORF2 by its first Met codon, rather than by its upstream flanking stop codon as is more informative when considering possible PRF mechanisms for translating a downstream ORF. Moreover, in the remaining one of these viruses described to date, UvURV, a single nucleotide substitution within the stop codon that currently defines the upstream end of ORF2 would bring this virus in line with the others, allowing ORF2 to overlap ORF1 by 40 nt in the −1 frame (Table 2, Fig. 6B).We therefore predict that the reported UvURV sequence (GenBank KR106133) contains at least this one error, or one or more other error with the same consequence.

This ORF1–ORF2 overlap possibly common to all unirnaviruses then indicates the strong possibility of −1 PRF as the mechanism for translating the RdRp as part of anORF1/ORF2 fusion product. The classical slippery sequence for −1 PRF is X_XXY_YYZ, where X_XX is any three of the same nucleotide, although several deviations such as GGA are tolerated; Y_YY is AAA or UUU; Z is A, C, or U; and underlines indicate codon boundaries for the upstream ORF (Firth 2012). Notably, in all six unirnavirus strains described to date, a matching sequence for this −1 PRF slippery motif is found immediately or soon before the ORF1 stop codon: G_GAU_UUU in AlDRV1, BbRV1/1, BbRV1/2, and UvRV-M; G_GAU_UUC in PjBbLV1; and U_UUA_AAC or G_GAU_UUA in UvURV (Fig. 6B). Moreover, other matching sequences are not identifiable in the region of ORF1–ORF2 overlap in any of these viruses, and in the case of BbRV1/1, BbRV1/2, and UvRV-M, the region of ORF1–ORF2 overlap is so short (only 10nt) that other possible −1 PRF slippery motifs are nonexistent. The finding of such putative −1 slippery sequences properly positioned within the ORF1–ORF2 overlap region thus strongly supports the argument for −1 PRF as the mechanism for translating the unirnavirus RdRp. Different members of family *Totiviridae* are known to use different PRF or other mechanisms for RdRp expression from ORF2 (Parent 2013), such that the different PRF mechanism proposed for unirnaviruses (−1) relative to plant

amalgaviruses and ZbV-Z/412 (+1) should not automatically consign these viruses to two different families, though it does represent evidence for their divergence.

2.5 Future Directions

Future studies include experimental characterizations of ZbV-Z. The +1PRF motif in ZbV-Z has not been experimentally characterized, nor have the three aforementioned plant amalgaviruses. Identifying *trans*-acting components, such as additional proteins, or minimal requirements of this +1PRF motif has also not been reported and such work would add to a more complete understanding of this system for both plantamalgaviruses and influenza viruses. Given that ZbV-Z infects yeast, such studies will not be limited by having to develop novel tools.

Another important question about the family *Amalgaviridae* that is still unanswered is whether any form confirmed virions. Virus particles have not been visualized to date for plant amalgaviruses, but sometimes such cryptic viruses of plants accumulate particles in small numbers that can be difficult to detect (Tzanetakis 2008). Interestingly, Krupovic et al. have recently reported that the ORF1 translation product (possible CP) of plant amalgavirus STV is homologous to the nucleocapsid proteins of certain negative-strand RNA viruses. This finding raises the possibility that the plant amalgaviruses might form filamentous nucleocapsids instead of icosahedral virus-like particles (Krupovicet 2015). Virions have also failed to be visualized to date for the members of proposed genus *Unirnavirus*. Schmitt and Neuhausen have reported the presence of virus-like particles enriched for the ∼3-kbp genome and the ∼35-kDa putative CP of ZbV-Z/412, following fractionation on sucrose gradients. These authors have moreover visualized virus-like particles from *Z. bailii* 412 that appear isometric. It is important to note, however, that *Z.bailii* 412 is also infected with a putative totivirus (ZbV-L/412) and its toxin-expressing M-

satellite RNA (Radler 1993; Schmittand 1994). We therefore consider it possible that the isometric particles shown by Schmittand Neuhausen (1994) might represent this totivirus, and not ZbV-Z/412. Our efforts to purify and characterize ZbV-Z/412 virions are in progress. We hope that by purifying ZbV-Z/412 virions, we will also be able to visualize the predicted ZbV-Z CP/RdRp fusion protein by gel and then subject it to tandem mass spectrometry to identify one or more of its constituent peptides that crosses the putative CP/RdRp junction.

2.5 References

Castillo A, Cottet L, Castro M, Sepúlveda F. Rapid isolation of mycoviraldouble-stranded RNA from *Botrytis cinerea* and *Saccharomyces cerevisiae*. *Virol. J*. 2001; 8: 38.

Coutts RHA, Livieratos IC. A rapid method for sequencing the 50- and30-termini of dsRNA viral templates using RLM-RACE. *J. Phytopathol*. 2003; 151:525–527.

Firth AE, Jagger BW, Wise HM, Nelson CC, Parsawar K, Wills NM, Napthine S, Taubenberger JK, Digard P, Atkins JF. Ribosomal frameshifting used in influenza A virus expression occurs within the sequence UCC_UUU_CGU and is in the +1 direction. *Open Biol*. 2012; 2, 120109.

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

Jiang Y, Zhang T, Luo C, Jiang D, Li G, Li Q, Hsiang T, Huang J. Prevalence and diversity of mycoviruses infecting the plant pathogen *Ustilaginoidea virens*. *Virus Res*. 2015; 195: 47–56.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol*. 2013; 30:772–780.

Koloniuk I, Hrabáková L, Petrzik K. Molecular characterization of a novel amalgavirus from the entomopathogenic fungus *Beauveria bassiana*. *Arch.Virol.* 2015; 160: 1585–1588.

Kotta-Loizou I, Sipkova J, Coutts RHA. Identification and sequence determination of a novel double-stranded RNA mycovirus from the entomopathogenic fungus *Beauveria bassiana*. *Arch. Virol*. 2015; 160: 873–875.

Krupovic M, Dolja VV, Koonin EV. Plant viruses of the *Amalgaviridae* family evolved via recombination between viruses with double-stranded and negative-strand RNA genomes. *Biol. Direct* 2015; 10: 12.

Lin Y, Zhang H, Zhao C, Liu S, Guo, L. The complete genome sequence of a novel mycovirus from *Alternaria longipes* strain HN28. *Arch. Virol.* 2015; 160:577–580.

Liu W, Chen J. A double-stranded RNA as the genome of a potential virus infecting *Vicia faba*. *Virus Genes*. 2009; 39: 126–131.

Liu H, Fu Y, Xie J, Cheng J, Ghabrial SA, Li G, Peng Y, Yi X, Jiang D. Evolutionary genomics of mycovirus-related dsRNA viruses reveals cross-family horizontal gene transfer and evolution of diverse viral lineages. *BMC Evol Biol.* 2012 Jun 20;12:91.

Martin RR, Zhou J, Tzanetakis IE. Blueberry latent virus: an amalgam of the *Partitiviridae* and *Totiviridae*. *Virus Res*. 2011; 155: 175–180.

Nerva L, Ciuffo M, Vallino M, Margaria P, Varese GC, Gnavi G, Turina M. Multiple approaches for the detection and characterization of viral and plasmid symbionts from a collection of marine fungi. *Virus. Res*. 2016 Jul 2; 219:22-38.

Radler F, Herzberger S, Schönig I, Schwarz P. Investigation of a killerstrain of *Zygosaccharomyces bailii*. *J. Gen. Microbiol*. 1993; 139: 495–500.

Sabanadzovic S, Abou Ghanem-Sabanadzovic N, Valverde RA. A novel monopartite dsRNA virus from rhododendron. *Arch. Virol.* 2010; 155: 1859–1863.

Sabanadzovic S, Valverde RA, Brown JK, Martin RR, Tzanetakis IE. Southern tomato virus: the link between the families *Totiviridae* and *Partitiviridae*. *Virus Res*. 2009; 140: 130–137.

Schäffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res*. 2001; 29: 2994–3005.

Schmitt MJ, Neuhausen F. Killer toxin-secreting double-stranded RNAmycoviruses in the yeasts *Hanseniaspora uvarum* and *Zygosaccharomyces bailii*. *J. Virol*. 1994; 68: 1765–1772.

Tzanetakis IE, Price R, Martin RR. Nucleotide sequence of the tripartite Fragaria chiloensis cryptic virus and presence of the virus in the Americas. *Virus Genes* 2008; 36: 267–272.

Wickner RB, Fujimura T, Esteban R. Viruses and prions of *Saccharomyces cerevisiae*. *Adv Virus Res*. 2013; 86:1-36.

Wickner RB, Ghabrial SA, Nibert ML, Patterson JL, Wang CC. Totiviridae. In: King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ, (Eds.). Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses. Elsevier, San Diego. 2012. pp. 639–650.

Zhu HJ, Chen D, Zhong J, Zhang SY, Gao BD. A novel mycovirus identified from the rice false smut fungus *Ustilaginoidea virens*. *Virus Genes* 2015; 51:159–162.

Section I: Proposed changes to previously reported genome sequences

Chapter Three: Determination of Complete Genome Sequence of Cryptosporidium parvum virus

The figures in this chapter were taken from its publication:

Vong M, Ludington JG, Ward HD, Nibert ML. Complete cryspovirus genome sequences from Cryptosporidium parvum isolate Iowa. *Arch Virol*. 2017 Sep; 162(9):2875-2879.

Special thank yous and contributions:

I want to give a very special thank you to Jacob Ludington and Honorine Ward. Jacob infected the Caco-2 cells with CSpV1-infected oocyst and provided the Caco-2 cells for the source of RNA. I completed the rest of the experiments, starting with RNA extraction to sequence determination. Max supervised the project from inception to publication.

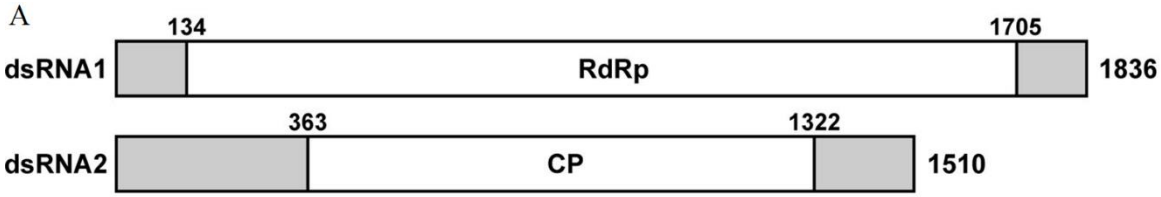
3.1 Introduction

One in ten child deaths results from diarrhea for children below the age of five years old. This results in about 800,000 deaths annual global deaths. Deaths caused by malaria for this age group only amounts to about 700,000. Most of the deaths from diarrheal disease occur in sun-Saharan Africa and south Asia (Liu 2012). In 2013, a study funded by the Bill & Melinda Foundation announced the results of their report identifying risk factors for diarrhea within this population. The study followed 22,500 children for three years and identified four main pathogens responsible for diarrhea for these infants and toddlers. Three main pathogens were the familiar *Escherichia coli, Shigella*, and enterotoxigenic *E. coli* producing heat-stable toxin. Surprisingly, the fourth pathogen identified was *Cryptosporidium* (Kotloff 2013).

Within the genus *Cryptosporidium*, *C. parvum* and *C. hominis* are the two species that are primarily responsible for human disease. Cryptosporidiosis most commonly leads to diarrhea (Checkley 2015). Many isolates of *C. parvum* and *C. hominis* are infected by a bi-segmented dsRNA virus (Nibert 2009). While these viruses are currently thought to be cryptic to their protozoan hosts, these viruses' genomes may serve useful. The viruses are persistent and transmit vertically. Transmission of the oocysts usually occurs through contaminated water sources and detection of these dsRNAs can serve as detection for trace levels of contamination or for epidemiological tracing during outbreaks. Further, one study from 2008 reported a correlation between the levels of the virus and oocyst excretion from infected calves (Jenkins 2008).

These bi-segmented dsRNA viruses belong in the genus *Cryspovirus* of family *Partitiviridae*. Their genomes consists of two separately packaged dsRNAs. The plus-strand of dsRNA1 encodes an RdRp while the plus-strand of the shorter segment, dsRNA2, encodes the viral CP (Nibert 2009; Fig 7A for proposed type species).

A



B

dsRNA1/RdRp	Iowa	KSU1	Changchun
Iowa	100	97.9	99.6
KSU1	96.1	100	97.5
Changchun	98.9	95.8	100

dsRNA2/CP	Iowa	KSU1	Changchun	1912	353	1689
Iowa	100	99.7	99.1	92.5	88.4	86.5
KSU1	97.9	100	98.7	92.2	88.1	86.2
Changchun	99.7	97.8	100	93.1	88.7	86.2
1912	91.4	91.8	91.5	100	87.5	85.9
353	87.4	86.9	87.4	86.5	100	91.2
1689	86.1	86.5	86.1	86.6	89.0	100

Figure 7 (A) Diagram of genome organization of CSpV1-Iowa. (B) Pairwise alignment scores (%). Values at the lower left of each panel are for the protein-encoding RNA sequences of each genome segment. Values at the upper right are for the deduced protein sequences.

Despite the reporting of many isolates of dsRNA viruses, few putatively complete genome sequences from both dsRNA segments have been reported, though many more partial sequences exist. Cryptosporidium parvum virus 1 (CSpV1) contains two putatively complete genome sequences of each dsRNA: KSU1 (Khramtsov 1997) and Changchun (Li 2009). In addition to the two putatively complete genome sequences of both dsRNA1 and dsRNA2 from two *C. parvum* isolates, Leoni et al. have reported the putatively complete sequences of dsRNA2 from three other species: *C. hominis*, *felis*, and *meleagridis* (Leoni 2006).

A pairwise comparison of the deduced CP sequence from the four different cryptosporidia viruses shows that they have >85% similarity (Fig. 7B), suggesting either slow evolutionary divergence of cryspoviruses in different host species or recent/ongoing exchange of the viruses among them. The dsRNA2 lengths, however, reported from *C. hominis, felis*, and *meleagridis* are

similar to each other (1481–1502 bp), but longer than those reported from *C. parvum* (1374–1375 bp). Leoni et al. also reported using rapid amplification of cDNA ends (RACE) to sequence the segment termini, whereas the reports from *C. parvum* did not.

Given these two discrepancies, we hypothesize that the sequences from the *C. parvum* viruses might be truncated at one or both termini of each genome segment. We therefore set out to determine the complete sequences of both dsRNA1 and dsRNA2 from the virus infecting *C. parvum*.

3.2 Materials and Methods

3.2.1 RNA

Oocysts from *C. parvum* isolate Iowa were purchased form Bunch Grass Farm (Idaho). These oocysts were also previously used to study the relationship between *C. parvum* fecundity and its viral load (Jenkins 2008). Oocysts were allowed to excyst *in vitro* then the released sporozoites were used to infect Caco-2 cells. At 48 hr post-infection, total RNA was extracted using TRIzol™ LS reagent (Ambion) following manufacturer's instructions, with the addition of one extra step. Equal volume of glass beads (diameter, 0.5 mm; BioSpec) was added during the cell lysis step. DsRNA was enriched from the total RNA using microgranular cellulose (cellulose powder MN 301, Macherey-Nagel) following the protocol described by Castillo et al. (Castillo 2011).

3.2.2 Sequence determination

Total RNA was used for initial sequencing. Primers designed from reported CSpV1 dsRNA1 and dsRNA2 sequences were then used for RT, SuperScript III First-Strand Synthesis

System (Invitrogen), then the RT products were subjected to PCR, Taq DNA Polymerase with Standard Taq Buffer (NEB), each according to manufacturer's instructions. The resulting amplicons were subjected to direct Sanger sequencing.

Enriched dsRNA was used for RNA-ligase-mediated rapid amplification of 3' cDNA ends (3' RLM-RACE) according to Depierreux et al. to determine the terminal sequences of both segments (Depierreux 2016). The resulting amplicons were subjected to direct Sanger sequencing.

Sequences generated from this study were named *Cryptosporidium parvum virus 1 Iowa* (CSpV1-Iowa) and were deposited in GenBank. The GenBank accession number for CSpV1-Iowa dsRNA1 is KY884720. The GenBank accession number for CSpV1-Iowa dsRNA2 is KY884721.

3.2.3 Sequence-based analyses

ORFs were identified using EMBOSS getorf. Pairwise sequence alignments were performed using EMBOSS Water or Needleall (http://www.bioinformatics.nl/emboss-explorer/). Multiple sequence alignments were performed using Clustal Omega (http://www.ebi.ac.uk/Tools/msa/). Phylogenetic analyses were performed using PhyML 3.0 (https://www.hiv.lanl.gov/content/sequence/PHYML/interface.html) with parameters Sequence type/model, Amino acids/JTT, LG, rtREV, or WAG (each yielded very similar results); Proportion of invariable sites, estimated from data; Gamma shape parameter, estimated from data; Starting tree(s) optimization, Tree topology and branch length; Tree improvement, Best of NNI and SPR; Branch support, Approximate Likelihood Ratio Test (aLRT), SH-like supports.

3.3 Results

The dsRNA1 sequence of CSpV1-Iowa (GenBank KY884720) has a length of 1836 bps while the reported CSpV1- KSU1 dsRNA1 sequences spans 1786 bps and Changchun at 1783 bps (Table 4). The plus-strand dsRNA1 sequence of CSpV1-Iowa encompasses one long ORF spanning 1572 nts (Fig. 7A). The ORF encodes a deduced protein 523 aa long (62 kDa; pI 9.45) and exhibits strong sequence similarities to viral RdRps, as expected, and shares >95% sequence similarities to CSpV1-KSU1 and Changchun (Fig. 7B). This RdRp length matches that reported for CSpV1-Changchun but is a single residue shorter than for CSpV1-KSU1, due to a 1-codon deletion corresponding to nt positions 1305–1307 in KSU1.

Table 4 Cryspovirus sequence features

Genome segment	Host species and isolate	Lengths		UTRs (nt):		GenBank acc. no.
		RNA (nt)	Protein (aa)	5′	3′	
dsRNA1	*C. parvum* KSU1	1786	524	133	78	U95995
	C. parvum Changchun	1783	523	133	78	EU183403
	C. parvum Iowa	1836	523	133	131	KY884720
dsRNA2	*C. parvum* KSU1	1374	319	247	167	U95996
	C. parvum Changchun	1375	319	247	168	EU183404
	C. hominis 1912	1492	319	346	186	DQ193518
	C. felis 353	1502	335/319 [a]	307/355 [a]	187	DQ193520
	C. meleagridis 1689	1481	319	345	176	DQ193519
	C. parvum Iowa	1510	319	362	188	KY884721

[a] The first in-frame AUG codon in *C. felis* 353 dsRNA2 (nt positions 308–310) yields a deduced CP that is 335 aa long. However, this AUG codon is not conserved in the other cryspoviruses. The first in-frame AUG codon that is conserved in all strains yields a deduced CP that is 319 aa long. Thus, in *C. felis* 353 dsRNA2, it might be the second in-frame AUG codon (nt positions 356–358) that is used for translation initiation

The dsRNA2 sequence of CSpV1-Iowa (GenBank KY884721) spans 1510 bps. This is similar to those reported from *C. hominis*, *C. felis*, and *C. meleagridis* (1481-1510 bps) but different from the reported CSpV1-KSU1 and Changchun sequences (1374-1375 bps) (Table 4). The plus-strand sequence encompasses one long ORF spanning 960 nts (Fig. 7A). The ORF encodes a deduced protein 319 aa long (37 kDa; pI 8.22), presumably the viral CP. This CP length

matches those of both CSpV1-KSU1 and Changchun, and also those of the viruses from *C. hominis*, *C. meleagridis*, and possibly *C. felis* (Table 4). This CP also shares >97% sequence similarity with the two *parvum* isolates and >85% sequence similarity with all viruses from the four different *Cryptosporidium* species (Fig. 7B).

The complete or putatively complete sequences now available for nine different cryspovirus genome segments (three dsRNA1s and six dsRNA2s) were compared via pairwise alignments. The results indicate that CSpV1-Iowa is another strain of species *Cryptosporidium parvum virus 1*, along with strains CSpV1-KSU1 and Changchun, given the high degree of conservation in both the RNA sequences (95.8–99.7% identity) and the protein sequences (97.5–99.7% identity) of these three viruses from *C. parvum* (Fig. 7B). The viruses from *C. hominis*, *C. felis*, and *C. meleagridis*, on the other hand, are somewhat more divergent in their dsRNA2 sequences (86.1–91.8% identity) and CP sequences (85.9–93.1% identity), though the degree of conservation is still high. The CP sequences were also compared by phylogenetic analysis, including the nearly complete CP sequences reported from 22 additional C. parvum isolates (Murakoshi 2016). The results again show greater divergence of the viruses from *C. hominis* (1912), *C. felis* (353), and *C. meleagridis* (1689) vs. the lesser divergence among all 25 viruses derived from C. parvum (Fig. 8). Complete genome sequences for other cryspovirus strains are needed before robust conclusions can be drawn, but this host-specific pattern of divergence suggests that exchange of viruses among some of these different host species may be infrequent. Partial CSpV1-Iowa sequences reported previously (Jenkins 2008) and the complete ones reported here are highly similar across their regions of overlap (>96%), and the observed divergence is likely explained by the different passage histories of the *C. parvum* Iowa samples obtained from different suppliers up to 20 years apart.

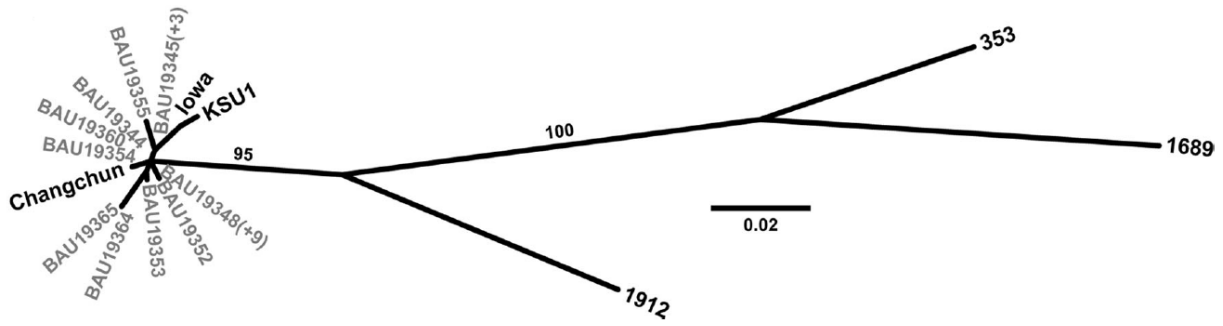


Figure 8 Maximum-likelihood phylogenetic tree using CP sequences of the indicated cryspovirus strains.

Based on the start and stop codons for dsRNA1 and dsRNA2 suggested above, the terminal untranslated regions (UTRs) in the CSpV1-Iowa plus-strand sequences span 133 and 131 nts at the 5’ and 3’ ends of dsRNA1, and 362 and 188 nts at the 5’ and 3’ ends of dsRNA2 (Table 4). The 5’-UTR of CSpV1-Iowa dsRNA1 has the same length as reported for CSpV1-KSU1 and Changchun, suggesting that all three of these sequences are complete at that end. Its 3’- UTR, on the other hand, is longer than reported for CSpV1-KSU1 and Changchun, suggesting that the latter are 3’-truncated. The 5’- and 3’-UTRs of CSpV1-Iowa dsRNA2 are each longer than reported for CSpV1-KSU1 and Changchun, suggesting that the latter sequences are truncated at both ends. The 5’- and 3’-UTRs of dsRNA2 are nearer the sizes reported from from *C. hominis*, *C. felis*, and *C. meleagridis*, but even the latter appear to be truncated by a few nucleotides based on their sequence alignments with CSpV1-Iowa (Fig. 9).

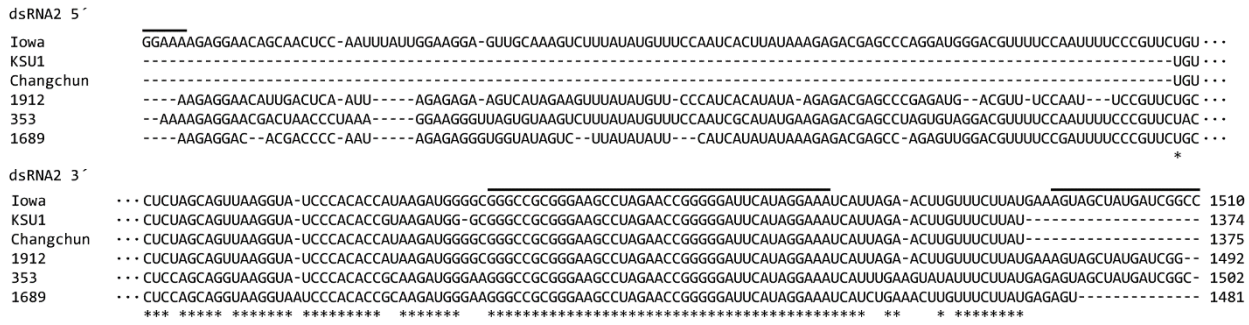


Figure 9 Multiple sequence alignment of dsRNA2

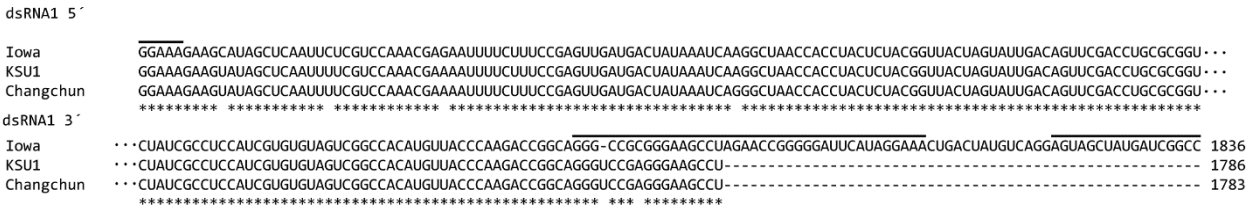


Figure 10 Multiple sequence alignment of dsRNA1

DsRNA1 and dsRNA2 of CSpV1-Iowa share several stretches of conserved plus-strand sequences, notably 5 nts at their 5’ ends and 17 nts at their 3’ ends (Fig. 9 and 10). An even longer conserved sequence (39 nts) is set back from the 3’ ends, but wholly within the 3’ UTRs of both segments, at nt positions 1767–1805 in dsRNA1 and 1431–1469 in dsRNA2. Lastly, dsRNA1 and dsRNA2 share a shorter 13-nts conserved region within their central protein-coding regions: UAAGAAAGUACCU, at nt positions 691–703 in dsRNA1 and 581–593 in dsRNA2. These motifs conserved between dsRNA1 and dsRNA2 in CSpV1-Iowa are also wholly or partly conserved in CSpV1-KSU1 and Changchun, as well as in the viruses from from *C. hominis* (1912), *C. felis* (353), and *C. meleagridis* (1689) (Fig. 9). The central motif is also conserved in both segments of numerous strains from the Murakoshi et al. study (Murakoshi 2016). These sequences conserved between dsRNA1 and dsRNA2 in multiple cryspovirus strains seem likely to be involved in specific functions, such as packaging, RdRp binding for minus- and/or plus-strand

RNA synthesis, and possibly even translation. Since partitiviruses package their two segments in separate particles, both segments might be expected to contain similar packaging signals.

Thus, we have sequence evidence to support our hypothesis that previously reported CSpV1 are likely truncated at both genome segments. Previously reported dsRNA1 sequences are likely truncated by 50nts, all likely from the 3' end. Previously reported dsRNA2 sequences are likely truncated by 135nts, 115nts from the 5' end and 20nts from the 3' end.

We report the first complete sequences of both genome segments from a single strain of type species *Cryptosporidium parvum virus 1*, in genus *Cryspovirus*, family *Partitiviridae*. We propose that this strain, CSpV1- Iowa, should be substituted for CSpV1-KSU1 as the exemplar strain of the species and genus, since the reported sequences of the latter appear to be terminally truncated. Complete genome sequences of other cryspovirus strains, from different *Cryptosporidium* species, are needed for ascertaining whether genus *Cryspovirus* should continue to contain only the single current species.

3.4 References

Castillo A, Cottet L, Castro M, Sepúlveda F. Rapid isolation of mycoviraldouble-stranded RNA from *Botrytis cinerea* and *Saccharomyces cerevisiae*. *Virol. J*. 2001; 8: 38.

Checkley W, White AC Jr, Jaganath D, et al. A review of the global burden, novel diagnostics, therapeutics, and vaccine targets for *cryptosporidium*. *Lancet Infect Dis* 2015; 15:85–94.

Depierreux D, Vong M, Nibert ML. Nucleotide sequence of Zygosaccharomyces bailii virus Z: Evidence for +1 programmed ribosomal frameshifting and for assignment to family *Amalgaviridae. Virus Res* 2016; 217:115–124.

Jenkins MC, Higgins J, Abrahante JE, Kniel KE, O'Brien C, Trout J, Lancto CA, Abrahamsen MS, Fayer R. Fecundity of *Cryptosporidium parvum* is correlated with intracellular levels of the viral symbiont CPV. *Int J Parasitol* 2008; 38:1051–1055.

Khramtsov NV, Woods KM, Nesterenko MV, Dykstra CC, Upton SJ. Virus-like, double-stranded RNAs in the parasitic protozoan *Cryptosporidium parvum*. *Mol Microbiol* 1997; 26:289–300.

Kotloff KL, Nataro JP, Blackwelder WC, et al. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *Lancet*. 2013 Jul 20; 382(9888):209-22.

Leoni F, Gallimore CI, Green J, McLauchlin J. Characterisation of small double stranded RNA molecule in *Cryptosporidium hominis, Cryptosporidium* felis and *Cryptosporidium meleagridis. Parasitol Int* 2006; 55:299–306.

Li W, Zhang N, Liang X, et al. Transient transfection of *Cryptosporidium parvum* using green fluorescent protein (GFP) as a marker. *Mol Biochem Parasitol* 2009; 168:143–148.

Liu L, Johnson HL, Cousens S, et al. Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. *Lancet* 2012; 379: 2151–61.

Murakoshi F, Ichikawa-Seki M, Aita J, et al. Molecular epidemiological analyses of Cryptosporidium parvum virus 1 (CSpV1), a symbiotic virus of *Cryptosporidium parvum*, in Japan. *Virus Res* 2016; 211:69–72.

Nibert ML, Woods KM, Upton SJ, Ghabrial SA. *Cryspovirus*: a new genus of protozoan viruses in the family *Partitiviridae*. *Arch Virol* 2015; 154:1959–1965.

Section II: Validation of novel viral genetic elements discovered from RNA-sequencing data

Chapter Four: Evidence for contemporary plant mitoviruses

The figures in this chapter were taken from its publication:

Nibert ML, Vong M, Fugate KK, Debat HJ. Evidence for contemporary plant mitoviruses. *Virology*. 2018 Feb 10; 518:14-24.

Special thank yous and contributions:

Max founded this project and completed the sequencing mining. Karine Fugate provided us with sugar beet seeds and leaves. I did the validation experiments used in Section 4.3.4. I want to give a special thank you to Jesse Pyle for showing me how he extracts RNA from plant tissue. I then had to make modifications to his method, which are the final methods described in Section 4.2.3. I want to also thank Katie Smith and William Gao for their help in the screening of mitoviruses in various beets (Section 4.6.2). I mentored, directed, and helped troubleshoot their bench experiments. Max supervised the project from inception to publication.

4.1 Introduction

Mitoviruses form the genus *Mitovirus* within the virus family *Narnaviridae*. The other currently recognized genus within this family is *Narnavirus*. All currently recognized members of *Narnaviridae* infect fungal hosts. Mitoviruses and narnaviruses are (+) ssRNA viruses consisting of a mono-segmented genome that contains a single coding sequence encoding an RNA-dependent RNA polymerase (RdRp). The genome of mitoviruses has a length that ranges from 2.1 to 4.4 kb and do not appear to be encapsidated. The RdRp coding sequence has a length that ranges from 650 to 1140 amino acids. Mitoviruses replicate in the mitochondria of their hosts whereas narnaviruses replicate in the cytoplasm (Hillman 2012). While there are currently five officially recognized species within *Mitovirus*, there are more than 90 accessions in GenBank that appear to represent other fungal mitoviruses.

Further, mitovirus sequences are further widespread. At least 175 nearly complete copies of mitovirus RdRp have been found endogenized within the mitochondrial genome of plants. These mitovirus nonretroviral endogenized RNA virus elements (NERVEs) are spread across at least 90 different plant mitochondrial genomes. Plant mitovirus NERVEs have added much interest to the evolutionary origins of mitoviruses (Marienfeld 1997; Hong 1998; Shackelton 2008; Bruenn 2015). Whether the ancestral mitovirus is of fungal or plant origin is still lively discussed, but all models share in common an intermediary actively replicating plant mitovirus stage.

Expanding on this on-going discussion, Bruenn et al. recently re-proposed that an ancestral fungal mitovirus integrated once or more into the mitochondrial genome of the common ancestor of vascular plants. Mitoviruses are currently clustered into three different clades, supporting some divergence (Hillman 2013), but plant mitovirus NERVEs, which are widely distributed in the later-branching flowering plants, seem to form a monophyletic cluster (Bruenn 2015). Interestingly,

these mitovirus NERVEs still maintain the conserved mitovirus RdRp core motifs in both sequence and order (Bruenn 2015). However, at the population level, it seems unlikely that most extant fungal mitoviruses could persist within plant mitochondria. Most extant fungal mitoviruses use the non-standard UGA codon to encode tryptophan, consistent with the genetic code used by fungal mitochondria (Nibert 2017), which would be translated as a stop codon in plant mitochondria. Thus, this non-standard UGA codon usage would be a translational barrier for fungus-to-plant transmission. A single common ancestor would likely originate from a plant mitovirus or a fungal mitovirus that lacks internal UGAs, such as that of glomeromycete mitoviruses (Kitahara 2014), whose host, the glomeromycetes, are routinely involved in endophytic interactions with many land plants, including more primitive ones (Brundrett 2002).

Further, though a promiscuous retrotransposon has been offered as an explanation for the integration event of mitoviruses into the plant mitochondria, the exact mechanism is still unknown, which also leaves open the discussion of a more precise timing of the integration event relative to the possible divergence events. Without knowing what selective pressures may induce the integration event, it still may be possible that replicating mitoviruses are still present in plant mitochondria. Replicating plant mitoviruses would also serve as an immediate ancestor for plant mitovirus NERVEs, whereas the alternative scenario for an immediate fungal mitovirus ancestor would mean such an intermediary stage is absent.

We hypothesize that replicating plant mitoviruses would serve as an immediate ancestor to plant mitovirus NERVEs.

Extant mitoviruses do not appear to produce any obvious, outward physical changes to their hosts, thus the hunt for extant plant mitoviruses should not depend on searching for physical mitovirus-induced phenotypes. However, given the increased application of whole transcriptome

sequencing and our previous success in identifying an RNA virus from these datasets (Pyle 2017), it is possible to search for extant, *bona fide*, and previously unidentified RNA viruses without depending on physical screening methods. Thus, this goal of this work is to investigate whether replicating mitoviruses are still extant in plants.

4.2 Materials and Methods

4.2.1 TSA database search

The deduced protein sequence of a mitovirus NERVE found in the mitochondrial genome of *Arabidopsis thaliana* (251 aa; GenBank P92543) was used as query in an initial tblastn search of the Transcriptome Shotgun Assembly (TSA) database for land plants (taxid 3193). Genome lengths between 2kb and 4.1kb and an apparent single long ORF, encoding a deduced protein between 674 and 821 aa long, were used to filter the hits. Replicates were also filtered out.

From previous experiences (Pyle 2017), we have learned that accessions in the TSA database are sometimes truncated at one or both termini. Raw sequence reads from which these transcript contigs had been assembled were accessed from Sequence Read Archive (SRA) database to generate refined assemblies that included 5′- and/or 3′-terminal sequence extensions.

For refining the termini of the TSA hits (Table 5), the 5′ and 3′-terminal sequences of each were used as queries to search the SRA dataset(s) from which that transcript contig had been assembled. The sequence reads identified by this search were then assembled into new contigs via CAP3 (Huang 1999) or CLC Genomics Workbench 8.0 (QIAGEN).

Table 5 Sequence features of newly identified plant mitoviruses.

Virus name	Virus abbrev.	Host source	GenBank or BioProject no	Transcript length (nt)		ORF length	UTRs 5': 3' (nt)
				Original	Refined		
Ambrosia artemisiifolia mitovirus 1	AmarMV1	Pannonia	GEZL01037418	2341	2898	821	317:115
Azolla filiculoides mitovirus 1	AzfiMV1	Stockholm1	GBTV01009554	2858	2871	795	336:145
Beta vulgaris mitovirus 1	BevuMV1[b]	C600+Roberta	JP500572	2680	2680	778	346:0
		KWS2320	PRJNA254489	na	2825	793	354:89
		KWS2320	PRJNA41497	na	2810	793	354:74
		STR06A6001	PRJNA41497	na	2786	793	340:64
		STR06B6002	PRJNA41497	na	2794	793	341:71
Cannabis sativa mitovirus 1	CasaMV1[c]	Purple Kush	JP464487	2449	2857	762	440:128
		Finola	PRJNA73819	na	2805	762	424:92
		MPC/MSU	PRJNA80055	na	2825	762	429:107
		UC-COE	PRJNA178769	na	2824	762	430:105
Dahlia pinnata mitovirus 1	DapiMV1[d]	Rio Riata	GBDN01010918	2684	2806	782	375:82
		UBC	PRJNA193277	na	2798	782	373:76
Erigeron breviscapus mitovirus 1	ErbrMV1[e]	SMMU	GDQF01116002	2804	2804	780	374:87
		YAU	PRJNA277583	na	2829	780	377:109
		YAU	PRJNA229196	na	2799	780	374:82
Humulus lupulus mitovirus 1	HuluMV1[f]	Karahanasou	LA397818	2754	2795	763	390:113
Oxybasis rubra mitovirus 1	OxruMV1	374	GEEQ. 01005055	2292	2734	763	351:91
Petunia exserta mitovirus 1	PeexMV1[g]	OPGC943	GBRT01041798	2698	2701	750	311:137
		Bern	PRJNA300556	na	2684	750	302:129
Solanum chacoense mitovirus 1	SochMV1[h]	G4	GEDG01002811	2726	2773	776	335:107
Cryphonectria [parasitica] mitovirus 1	CpMV1	NB631	L31849	2728	na	809	86:212
Ophiostoma [novo-ulmi] mitovirus 3a	OnuMV3a	Ld	AJ004930	2617	na	718	268:192
Ophiostoma [novo-ulmi] mitovirus 4	OnuMV4	Ld	AJ132754	2599	na	783	204:43
Ophiostoma [novo-ulmi] mitovirus 5	OnuMV5	Ld	AJ132755	2474	na	729	227:57
Ophiostoma [novo-ulmi] mitovirus 6	OnuMV6	Ld	AJ132756	2343	na	695	141:114

4.2.2 Sequence and phylogenetic analyses

All database searches were performed with the indicated programs as implemented with defaults at http://blast.ncbi.nlm.nih.gov/Blast.cgi. Searches of the TSA or NR/NT database with protein sequence queries deduced from nucleotide sequences were performed using tblastn. Searches of SRA data sets with nucleotide sequence queries were performed using discontiguous megablast, or occasionally blastn if further efforts were needed to extend partial contigs. Searches of the NR database with protein sequence queries deduced from nucleotide sequences were performed using blastp.

ORFs in nucleotide sequences were identified and translated using ExPASy Translate as implemented at http://web.expasy.org/translate/ or SMS Translate as implemented at http://www.bioinformatics.org/sms2/. Multiple sequence alignments of protein sequences were

performed using MAFFT 7.310 (L-INS-i) (Katoh 2013) as implemented with defaults at http://mafft.cbrc.jp/alignment/server/. Global or local pairwise alignments of RNA sequences were performed using Needle, Needleall, or Water as implemented with defaults at http://www.bioinformatics.nl/emboss-explorer/. Codon frequencies were determined using SMS Codon Usage as implemented with defaults at http://www.bioinformatics.org/sms2/.

The best-fit substitution model for each multiple sequence alignment was identified according to the Bayesian information criterion using ModelFinder (Kalyaanamoorthy 2017) as implemented with the "Find best and apply" option at https://www.hiv.lanl.gov/content/sequence/IQTREE/iqtree.html (Trifinopoulos 2016). Phylogenetic analyses were then directly performed using IQ-TREE (Nguyen 2015) and UFBoot (Minh 2013) as implemented with defaults at that same website. The results in Newick format were submitted to TreeDyn 198.3 as implemented at http://www.phylogeny.fr/ for collapsing branches with lower support values.

The following approach was used for identifying mitovirus NERVEs for phylogenetic analysis (E-values< 1e−4 considered significant). Searches of the NR/NT database for green plants (taxid:33090), via tblastn using the RdRp sequence of fern mitovirus AzfiMV1 or fungal mitoviruses CpMV1, OnuMV4, OnuMV7, or TeMV as queries, yielded no significant hits from non-flowering plants, but a merged total of 202 different significant hits from flowering plants (21 from non-eudicots, 181 from eudicots; E-values, 1e−42 to 7e−5). The query-aligned nt sequences from these hits were next downloaded and translated into aa sequences. To make the number of sequences for analysis more manageable, those with the following characteristics were discarded: sequences not clearly annotated as being genomic (mitochondrial or nuclear) in origin, translated sequences< 120 aa long, and translated sequences not encompassing the conserved GDD motif or

not aligning that motif as expected. In addition, a number of homologous sequences from plants of the same genus were noted (> 65% identity in pairwise comparisons), and the shorter one of each of these replicate pairs was also discarded. Lastly, two NERVEs with large insertions relative to all the others were discarded because in preliminary trees they mapped on very long branches well within the plant mitovirus cluster.

4.2.3 Validation studies in *Beta vulgaris* sugar beet strain VDH66156

Leaves or seeds of *B. vulgaris* sugar beet strain VDH66156 were frozen in liquid nitrogen. Mortar and pestle were used to grind the frozen tissues into a fine powder and used directly for either RNA or DNA extraction. TRIzol™ Reagent (Invitrogen) was used to extract RNA according to manufacturer's instructions. DNeasy Tissue Kit (Qiagen) was used to extract DNA according to manufacturer's instructions. RNA extract was reverse-transcribed using SuperScript III Reverse Transcriptase (Invitrogen) according to manufacturer's instructions except that 4 primer pairs specific for different, overlapping regions of BevuMV1 were used for 4 separate reactions. One µL of RT reaction was then used directly per 20-µL total volume of PCR reaction. EconoTaq PLUS (Lucigen) was used for the PCR, according to manufacturer's instructions. Separate PCR reactions were set up using the same 4 primer pairs with their respective RT reactions. For one set of controls, RNA extract was used directly for PCR with BevuMV1-specific primers. For another set of controls, DNA extract was used directly for PCR with BevuMV1-specific primers or with four other primer pairs specific for *B. vulgaris* chloroplast or mitochondrial DNA. All PCR reactions used the following conditions: one cycle of 95 °C for 30 s; 40 cycles of 95 °C for 30 s, 50 °C for 30 s, 68 °C for 1 min and 10 s; a final cycle of 68 °C for 5 min and 4 °C hold. For sequencing, RT–PCR amplicons were visualized by agarose gel electrophoresis using 1% agarose solution

(SeaKem LE) and then purified from excised gel fragments using Monarch DNA Gel Extraction Kit (New England Biolabs) according to manufacturer's instructions. Sanger sequencing was performed at the Dana-Farber/Harvard Cancer Center DNA Resource Core.

4.3 Results

4.3.1 Complete coding sequences of tentative plant mitoviruses

The top hits included 60 with E-value scores ≤1e−30, indicative of strong sequence similarities. Among these 60 were 23 that genome lengths of similar to fungal mitoviruses. Within the 23, 13 had an apparent single long ORF approximating the RdRp lengths of fungal mitoviruses. After replicates, we were left with 10 hits (E-values, 2e−95 to 1e−31) (Table 5, top).

During the course of accessing the SRA datasets for contig refinement, we also found that there are sequence reads available from other transcriptome projects on some of the same 10 plant species: three other sugar beet strains of *Beta vulgaris* (BioProjects PRJNA41497 and PRJNA254489), three other strains or sources of *Cannabis sativa* (hemp) (BioProjects PRJNA73819, PRJNA80055, and PRJNA178769), another source of *Dahlia pinnata* (a common ornamental) (BioProject PRJNA193277), another source of *Erigeron breviscapus* (a Chinese species of fleabane used in traditional medicine) (BioProjects PRJNA229196 and PRJNA277583), and another source of *Petunia exserta* (a Brazilian species of increasing use as an ornamental) (BioProject PRJNA300556) (Table 5).

After the refinements and additions, we were able to identify 20 nearly complete genome sequences of apparent mitoviruses that (i) derive from samples of 10 different plant species; (ii) are between 2684 and 2898 nts long; and (iii) each encompasses a single long ORF that is

bracketed by stop codons and encodes a deduced protein sequence between 750 and 821 aa long (Table 5, Fig. 11 & 12), for 17 of which the deduced protein sequences are unique. The lengths

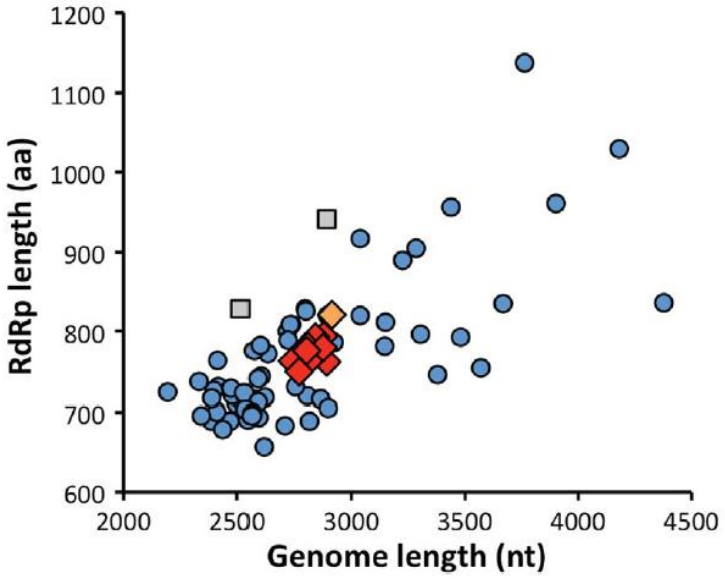


Figure 11 Scatter plot of genome and RdRp lengths. Red diamonds: apparent mitoviruses identified from flowering plants listed in Table 5. Orange diamond: apparent fern mitovirus from Table 5. Gray squares: narnaviruses. Blue circles: fungal mitoviruses.

of their UTRs, though likely missing some terminal residues in most cases, are also relatively consistent (5': 302–440 nt; 3': 64–145 nt). Further, none of the apparent plant mitoviruses contain a UGA codon, unless it is encoding a stop codon. All 21 tentative plant mitovirus sequences listed in the body of Table 5 (including the 3′-truncated coding sequence from *Beta vulgaris* strains C600 and Roberta) have been submitted to GenBank as Third-Party Annotation (TPA) sequences with accession numbers BK010422–BK010442. Additionally, the implicated hosts are all land plants and indeed all eudicot flowering plants, except for the fern *Azolla filiculoides*.

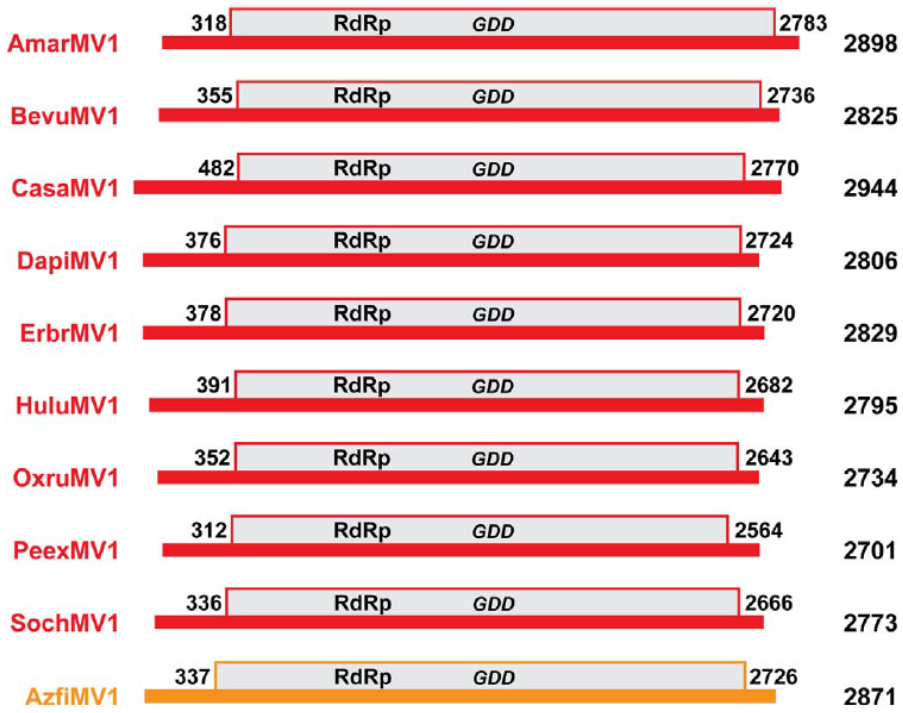


Figure 12 Scaled diagrams of apparent plant mitovirus genomes. Color-coding: red, viruses from flowering plants (*Ambrosia artemisiifolia, Beta vulgaris, Cannabis sativa, Dahlia pinnata, Erigeron breviscapus, Humulus lupulus, Oxybasis rubra, Petunia exserta, and Solanum chacoense*); orange, virus from fern (*Azolla filiculoides*).

4.3.2 Monophyletic cluster of plant mitoviruses

We next made use of phylogenetic methods to investigate the relationship between the RdRps of the tentative plant mitoviruses and those of a large collection of previously reported fungal mitoviruses. The sequences of two viruses assigned to genus Narnavirus were included as a likely out-group. Results of these analyses show that the apparent plant mitoviruses are clearly embedded alongside fungal mitoviruses within the current bounds of genus Mitovirus (Fig. 13). Moreover, the apparent plant mitoviruses are specifically associated with one particular clade of fungal mitoviruses (designated Clade II/IIb in Fig. 13) and form a monophyletic cluster within that clade. The one virus from a more primitive plant host, fern mitovirus AzfiMV1, is the most basally branching member of this cluster.

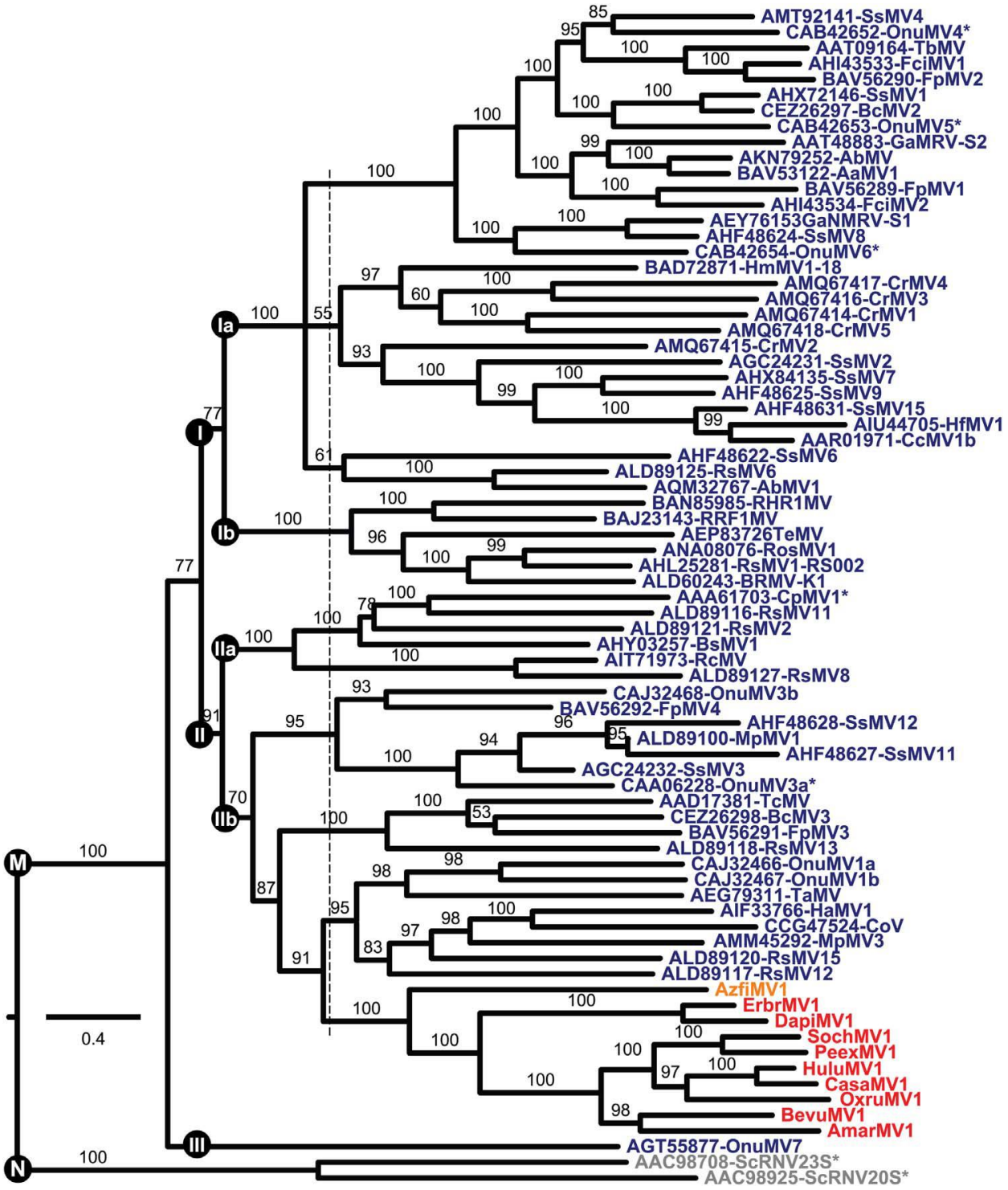


Figure 13 Phylogenetic tree of genus *Mitovirus*. Red: apparent mitoviruses identified from flowering plants listed in Table 5. Orange: apparent fern mitovirus from Table 5. Gray: narnaviruses. Blue: fungal mitoviruses.

4.3.3 Presence of apparent plant mitoviruses in different tissues

We accessed separately deposited and annotated tissue-specific SRA databases to assemble complete coding sequences in a tissue-specific manner for several of the viruses, including BevuMV1-KWS2320, CasaMV1-MPC/MSU, DapiMV1-RioRiata, PeexMV1-OPGC943, and SochMV1-G4 (Table 6). In each case, the nt sequences of the respective virus obtained from different tissues were> 99.9% identical. Thus, as probably should be expected for genuine plant mitoviruses, essentially no tissue-specific sequence differences were observed for these five viruses. This reduces the possibility that the sequences from generated from associated fungal contamination. Further, an interesting finding is that datasets from flower-related samples contain higher fractions of mitovirus sequences. This is especially notable for PeexMV1-OPGC943, which has more than 10-fold higher fractions than in any of its other four sampled tissues.

Table 6 Mitovirus sequences from different plant tissues. All mitovirus sequence are 100% identical to other mitovirus sequences derived from different tissue of same plant, with exception of CasaMV1-MPC/MSU sharing 99% identity (1 nt difference out of 2804nt).

Virus abbreviation and strain	Plant tissue	Virus-specific reads (fraction of total) reads × 1e5)	Contig length (nt)	Mean coverage (pernt position)
BevuMV1-	roots	≥ 5.1	2704	≥ 454
KWS2320	leaves	1.0	2810	579
	inflorescences	≥ 9.2	2825	≥ 1416
	seeds	≥ 34	2799	≥ 357
	seedlings	4.9	2800	410
CasaMV1-MPC/	roots	3.4	2818	137
MSU	stems	12	2824	450
	leaves	3.3	2805	177
	flower buds	9.8	2823	353
	flowers	10	2811	394
DapiMV1-	stem	5.5	2796	93
RioRiata	leaf	4.9	2806	102
	flower bud	6.7	2791	106
PeexMV1-	apical shoot	2.1	2659	43
OPGC943	trichome	2.3	2680	37
	callus	1.0	2667	14
	flowers	≥ 30	2700	≥ 740
	seedling	1.0	2664	17
SochMV1-G4	leaves	0.8	2750	84
	immature ovules	5.5	2747	1326
	mature ovules	≥ 6.5	2754	≥ 1465

4.3.4 Validation results for BevuMV1

Leaves and seeds of *Beta vulgaris* sugar beet strain VDH66156 (BioProject PRJNA219421) were obtained from the same source as for the transcriptome project where we identified evidence of an apparent plant mitovirus (Table 5). Following RNA extraction from either leaves or seeds, RT–PCR using different sets of BevuMV1-based primers yielded robust amplicons of expected sizes (Fig. 14A). The RT–PCR amplicons were then sent out for Sanger sequencing. The sequenceing results yielded 2693 nt of sequence representing the contiguous central region of the BevuMV1-VDH66156 genome and encompassing the complete coding region. Moreover, the 2693 nt of sequence obtained from either leaves or seeds are 100% identical to one another. The deduced RdRp sequence of BevuMV1-VDH66156 is ≥98.9% identical to that of the other BevuMV1 strains for which complete coding sequences were assembled. The complete coding sequence of BevuMV1-VDH66156 has been deposited at GenBank as regular accession MG721540. In contrast, no BevuMV1-specific amplicons were obtained from leaf RNA or DNA using a PCR protocol lacking reverse transcriptase (Fig. 14B & 14C), consistent with the extrachromosomal, RNA-based origin of BevuMV1. Further, the mitochondrial genome for *B. vulgaris* has been sequenced and no sequence homology to BevuMV1 can be identified in the mitochondrial genome.

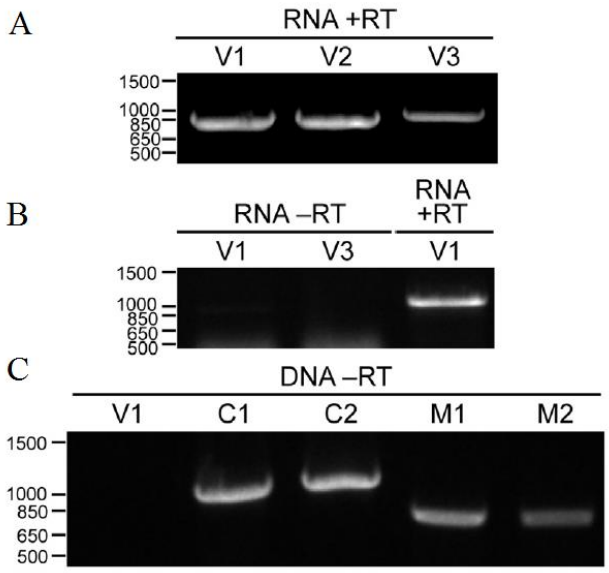


Figure 14 (A & B) RT-PCR for BevuMV1 in leaves of *B. vulgaris* strain VDH66156 using two different primer pairs specific for BevuMV1 (V1 and V3). (C) PCR for BevuMV1 in leaves of *B. vulgaris* strain VDH66156 with two primer pairs specific for *B. vulgaris* chloroplast DNA (C1 and C2) and two primer pairs specific for *B. vulgaris* mitochondrial DNA (M1 and M2).

4.3.5 Relationship of apparent plant mitoviruses to mitovirus NERVEs

Few mitovirus NERVEs are annotated in GenBank. Thus, mitovirus NERVEs were re-identified and used for phylogenetic analysis with all of the viruses from Fig. 13. The mitovirus NERVEs form a monophyletic cluster with the apparent plant mitoviruses, and apart from the fungal viruses (Fig. 15). In fact, the 79 NERVEs, all from flowering plants, are juxtaposed in the phylogram with the apparent plant mitoviruses from flowering plants. Notably, the one fern mitovirus, AzfiMV1, is also the most basal to the flowering plant cluster, as also seen for the viruses in Fig. 13. As shown in Fig. 13, fungal mitoviruses share a proximate common ancestor with the apparent plant mitoviruses and, in Fig. 15, also the mitovirus NERVEs from plant genomes.

Figure 15 Phylogenetic tree of genus *Mitovirus* and plant mitovus NERVEs. Red: apparent mitoviruses identified from flowering plants listed in Table 5. Orange: apparent fern mitovirus from Table 5. Blue: fungal mitoviruses. Green: mitovirus NERVEs

4.4 Conclusion

We provide evidence for extant, replicating plant RNA mitoviruses. We have found the sequence of these viruses from 10 different species of land plants, representing one family of ferns and four families of flowering plants, from an RNA dataset derived from transcriptome shotgun assembly. These plant mitoviruses form a monophyletic cluster embedded within fungal mitoviruses. We have located their sequences from various separate, tissue samples, with each tissue sample giving rise to >99.9% identical sequences within the same plant. We have also identified replicate sequences from other plant strains from different transcriptome projects. We have also validated these findings within both seeds and leaves from the *B. vulgaris* sugar beet strain VDH66156 by showing that it is dependent on RNA extract that must be reverse transcribed.

4.5 Discussion

The juxtaposition of flowering plant mitoviruses and flowering plant mitovirus NERVES in Fig. 15 suggests that plant mitoviruses are more likely than fungal mitoviruses to have been the immediate ancestors to most or all of the NERVEs examined here, supporting our hypothesis. There are other evolutionary congruencies between the plant hosts and their respective mitoviruses identified in this report. First, host *Azolla filiculoides* is the only representative of class Polypodiopsida (ferns), which diverged from Spermatophyta (seed plants, including flowering plants) ~400 million years ago (http://timetree.org/; Kumar 2017). In accordance, fern mitovirus AzfiMV1 is more divergent than the other plant mitoviruses (Fig. 13 and 15), which are all from flowering plants. Moreover, AzfiMV1 branches from the root of the flowering plant mitovirus cluster, suggesting that this monophyletic lineage of plant mitoviruses may have first entered plant hosts as early as the evolution of ferns. Indeed, Bruenn et al. (2015) have suggested an even earlier

possible timing, as early as the evolution of clubmosses. The identification of extant plant mitoviruses does not indicate that the last common ancestor of both fungal and plant mitoviruses originated from a plant mitovirus. While most fungal mitoviruses encode internal UGA codons, which would translate into a stop codon in plants but translates as tryptophan in most fungal mitochondria, mitoviruses of glomeromycetes lack internal UGAs (Kitahara 2014) and are found in hosts that are found in endophytic relationships with land plants (Brundrett 2002).

It appears from Figur 13 that genus *Mitovirus*, as currently recognized, is phylogenetically broad and might be usefully subdivided, as also implied by previous authors (Hillman 2013; Nibert 2017). One proposal would be for current genus *Mitovirus* to be reclassified as subfamily "Mitovirinae" in family *Narnaviridae*. Within subfamily "Mitovirinae", several new genera (possibly named "Alphamitovirus", etc.) could then be recognized, possibly three genera corresponding to Clades I, II, and III, or five genera corresponding to Clades Ia, Ib, IIa, IIb, and III, as suggested in Figure 13. The plant mitoviruses would thereby belong to the new genus corresponding to Clade II or IIb, which would include viruses from both plant and fungal hosts. The other new genera would contain viruses from only fungal hosts to date. One might wish to argue that the plant mitoviruses should instead warrant a separate genus, given their distinct host range; however, based on phylogenetic analyses such as that in Figure 13, that viewpoint would seem to argue as well that the fungal mitoviruses should be divided among a much larger number of new genera, which the current authors do not support at this time. Notably, Clades I and II suggested in Figure 13 correspond to ones previously suggested by Hillman and Cai (2013), and Clades Ia and Ib correspond to ones previously suggested by Nibert (2017), though numbered differently.

4.6 Current & Future Directions

4.6.1 Virus properties

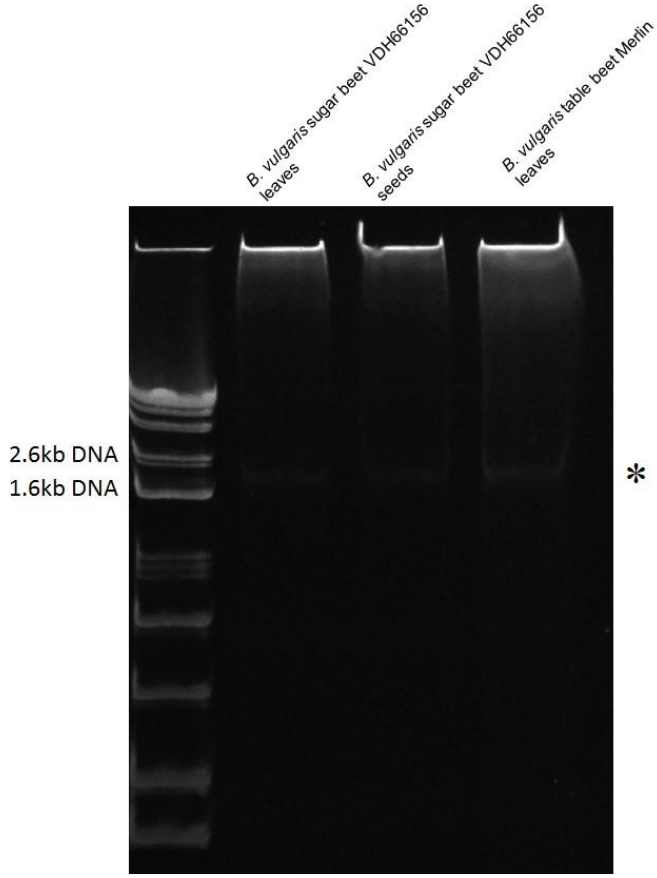


Figure 16 Enriched dsRNA MV+ *B. vulgaris* samples. Asterisks indicate migration position of dsRNA band.

We have attempted to visualize the dsRNA of BevuMV1. While we are able to visualize the dsRNA enriched from various sources, the dsRNA was about half the expected size. The sequenced genome would suggest that its dsRNA should be a bit over 2.7kb, but the dsRNA we are visualizing appears within the 1.8kb range. Both leaves and seeds from *B. vulgaris* sugar beet strain VDH66156 (Red River Valley Agricultural Research Center [USDA-ARS], Fargo, ND) and leaves from *B. vulgaris* table beet Merlin (Johnny's Seed Company; Allandale Farm) gave the same results (Fig. 16, asterisks). I used 2ml of TRIzol™ to extract about 1ml volume of liquid-

nitrogen ground sample (do not compress this powder; use uncompressed powder as 1ml volume). I resuspended the RNA with 100µl of DEPC-water. I used 30mg of MN301 cell in 1.5ml volume of 1X STE+16% ethanol to enrich RNA. I eluted RNA with 25ul of DEPC-water twice. I used 12µl and ran the sample on 0.5X TBE native gel at 100V for 60minutes (running buffer = 0.5X TBE). I removed the gel from the glass plates and stained immediately (no need to wash) with 15ml of 1X TBE+1X SYBR Green II for 35mins. I visualized without washing. Probing this dsRNA using a northern blot would help determine the identity of this band and/or other possible dsRNA bands that are too faint to currently visualize with gel electrophoresis.

We have determined the 5' terminal sequence of BevuMV1-VDH66156, but was not successful in determining the sequence of the 3' end. We did not report the 5' terminal sequence in the above publication because we plan to report both the 5' and 3' terminal sequences together as a single accession. The current accession number for BevuMV1-VDH66156 does, however, encode the full-length coding region. Further, I was only able to generate sequenced reads using an internal BevuMV1-specific primer from Sanger sequencing, whereas my previous work with CSpV1-Iowa was successful in determining sequenced reads from both directions using the same adapter-specific primers. Enriched dsRNA was used for RNA-ligase-mediated rapid amplification of 3' cDNA ends (3' RLM-RACE) according to Depierreux et al. to determine the terminal sequences of both segments (Depierreux 2016).

Other experimental characterizations of more immediate interest include isolating the mitochondria fraction from *B. vulgaris* sugar beet strain VDH66156 and assaying for the presence of BevuMV1 and showing maternal inheritance of BevuMV1.

4.6.2 Prevalence and distribution of BevuVM1

Sugarbeet is one of the two major sources of refined sugar. Sugarbeets have been extensively hybridized with table beets and Swiss chard, both of which are eaten by humans. It is derived from another important crop, the fodder beet, which is used as animal feed. Further, sugarbeets, table beets, Swiss chard, and fodder beets are all derived from the same plant species, *B. vulgaris*, with origins dating back to the Asian and Europe (Goldman 2003).

Mitoviruses are thought to enter new cells during cell division or mitochondrial exchanges. They appear to lack an extracellular lifecycle and replication is thought to be confined within the mitochondria (Hillman 2012). Though mitoviruses are largely thought to be cryptic, effects on host growth and mitochondria have been reported (Wu 2007; Xu 2015).

An economically important mitochondria-linked trait displayed in beets, as well as >150 other plant species, is cytoplasmic male sterility (CMS) (Laser 1972). Plants with CMS are functionally female and commonly used for hybridization. The first CMS sugarbeet was reported in 1945 by F.V. Owen and has become a common source of CMS for hybridizing beet plants. In beets, several mitochondrial genes have been reported to be associated with this phenotype as well as at least two nuclear genes (Mikami 2011). The two nuclear genes confer restoration of male fertility in otherwise male sterile plants and are often called restorer of fertility (Rf) genes (Owen 1945). Further, in 1981, K. Larsen reported male sterility in sugarbeet caused by the beet yellows virus but has not followed up. The weighted effects of each report on CMS is not yet fully understood (Larsen 1981).

The Owen sugar beet is the original source of CMS for CMS-hybridized table beet plants (Goldman 2003). In 1949, W.H. Gabelman launched the table beet breeding program at the University of Wisconsin-Madison. In 1959, Gabelman reported introducing the Owen CMS

cytoplasm into table beets (Stein 1959). UW-Madison has publically released inbred lines from Gabelman's program for widespread hybrid table beet production around the world. This series of inbred lines have been briefly described to include their pedigree and the list compiled together in a public report by I.L. Goldman (Goldman 1996).

Understanding the prevalence and distribution of BevuMV1 across different subspecies, cultivars, and strains of *B. vulgaris* will be helpful for goals of addressing any potential virus-host interactions. Further, such studies may also identify useful strains that can tease out the weighted effects of each CMS report and their relationship to each other. The goal of this section of the project will be to screen various *B. vulgaris* subspecies, cultivars, and strains of cultivars to address the prevalence and distribution of BevuMV1.

I started an initial screen for BevuMV1 in 22 various strains of *B. vulgaris* seeds and found that 3 strains of fodder beets, 1 strain of Swiss chard, 1 strain of table beet and 2 strains of sugarbeet were positive for BevuMV1 from the initial screen. Some of these strains also contain the putative Owen CMS locus: the 2 sugarbeet strains and 1 table beet. Interestingly, I have also identified 1 sugarbeet strain that contains the Owen CMS locus but does not appear to contain BevuMV1. However, we decided to not focus on the Owen locus for now and set it aside for its own project for a later time.

Upon further screening, we expanded the strains of *B. vulgaris* to include 3 strains of fodder beets, 3 Swiss chard, 3 sugar beets, and 8 table beets. Excitingly, one of the table beets is also a Founder table beet. A phylogenetic tree was assembled from these *B. vulgaris* strains and the original BevuMV1 TSA contig described in Table 6, *B. vulgaris* strain KWS2320 (Fig. 17). The phylogenetic analysis shows that there exists two clusters of BevuMV1 sequences, one dominated by the fodder beets and the other dominated by table beets. The current BevuMV1 sequences from

sugar beets all cluster with the table beets (Fig. 17). This apparent bottleneck is consistent with an outcome that would result from hybrids being derived from a single source. We thus decided to further divide this study into two foci. One focus will center on the lineage of table beet hybrid cultivation and sugar beets, which will possibly have a follow-up study to include the role of the Owen locus. The second focus will look at the broader cultivation of the various cultivars of *B. vulagaris* and will probe deeper back in history to include the fodder beets, founder table beets, and Swiss chards.

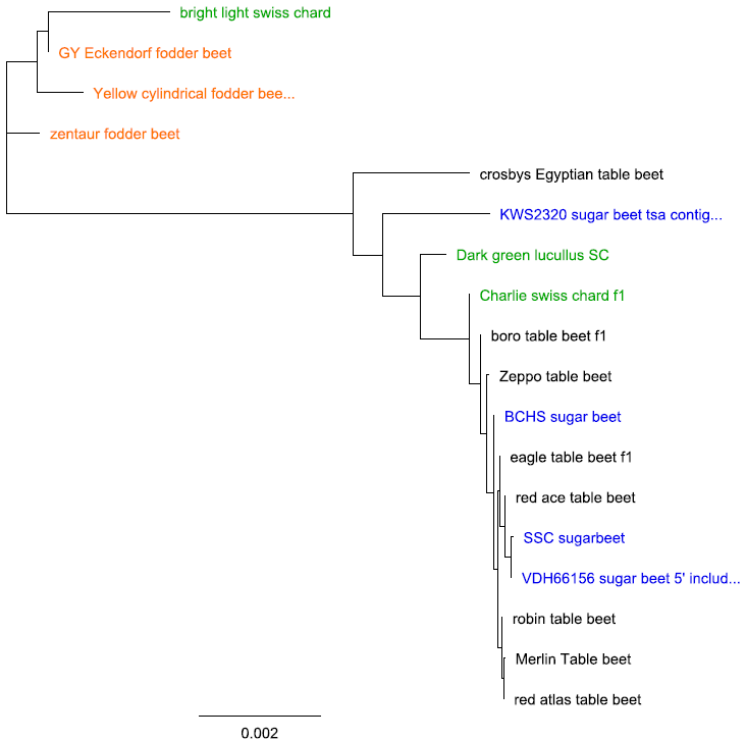


Figure 17 Phylogenetic analysis of various cultivars of *B. vulgaris*. Green: Swiss chard. Orange: fodder beet. Blue: sugar beet. Black: table beet.

Towards tracing the lineage and hybridizations of table beets, we collaborated with I.L. Goldman (Uni. Wisconsin) and obtained seeds from UW-Madison's publically released inbred lines 20 lines obtained: 14 of these lines consist of maternal and paternal pairs. In addition, we have bought an additional 30 seeds from commercial table beet hybrid strains. For this larger screen, I mentored two summer undergraduate students, Katie Smith and William Gao. Both Katie

and William have been incredibly helpful and productive. We have determined the sequence of BevuMV1 from samples. From the 30 analyzed non-UW-Madison strains, 7 genotypes have been identified. We have attributed 3 of the genotypes to a specific maternal line from UW-Madison (Table 7). Nine of the 14 parental pairs of UW-Madison lines contain ambiguities at various nucleotide positions in their BevuMV1 genomes, preventing us from grouping the female line into a single genotype. Further, upon acquiring more experience during the screens, we eventually began to shift thoughts about whether the BevuMV1-negative strains were indeed absent of

Table 7 Seven current genotypes identified in table beet hybrids. Blue: sugar beet. Green: Swiss chard. Two sugar beets fit into the group 1 genotype, which we called reference. Group 5 genotype also contains one Swiss chard.

Group	Identified UW-inbred line?	Commercial or non-UW strain	Source	Genotype
1	n/a	VDH66156	VDH	reference
		Sugar beet	Bakers Creek Heirloom	
		Red Magic (SC)	Burpee	
2	n/a	Falcon	Stokes Seeds Inc	G781A
		Falcon	Territorial Seed Co.	
3	W357A	Eagle	Harris Seeds	A844G, G1779A, U2125C
		Harrier	Gurney's Seed & Nursery Co.	
		Kestrel	Harris Seeds	
		Merlin	Johnny's Select Seeds	
		Red Ace	Osborne Seed Co.	
		Red Atlas	Osborne Seed Co.	
		Rhonda	High Mowing Organic Seeds	
		Robin	Chase Garden Seeds	
		Solo	Territorial Seed Co.	
4	W364 & 371A	Pacemaker III	Gurney's Seed & Nursery Co.	G1779A, U2125C
		Pacemaker III	UW-Madison	
		Red Cloud	Jungseed.com	
5	W425 & 427A	Action	n/a	G1779A, U2125C, A2726G
		Fire Fresh (SC)	Territorial Seed Co.	
		Vulture	Stokes Seeds	
		Vulture	Holmes Seeds	
6	n/a	Boro	High Mowing Organic Seeds	G1779A, U2125C, U2527C, A2726G
		Charlie (SC)	Osborne Seed Co.	
		Pablo	Territorial Seed Company	
		Subeto	Territorial Seed Company	
		Wodan	Kings Seeds	
		Zeppo	Osborne Seed Co.	
7	n/a	Alto	Kings Seeds	G1065R, G1779A, U2125C, U2527C, A2726G
		Taunus	Jungseed.com	

BevuMV1 infection. We think that the negative results are more likely to be false negatives because several strains, though negative at first, eventually provided unique sequences of BevuMV1 upon further troubleshooting. However we were still not able to generate sequences from all seed samples, particularly very aged samples.

Using the first sugar beet we sequenced as reference (See Materials and Methods 4.2.3), the seven genotypes we identified were (Table 7) : (1) reference (VDH66156); (2) G781A; (3) A844G, G1779A, U2125C (matched to UW line "W357A"); (4) G1779A, U2125C (matched to UW line "W364A" and "W371A"); (5) G1779A, U2125C, A2726G (matched to UW line "W425A" and "W427A"); (6) G1779A, U2125C, U2527C, A2726G; (7) G1065R, G1779A, U2125C, U2527C, A2726G. Thus, at this stage of screening, it seems that a small number of founder plants have indeed given rise to the many available commericial hybrid table beets of today. Of the seven identified genotypes, three have been traced back to UW lines.

BevuMV1 may be used in lineage tracing, though there are caveats. Consistent with lineage tracing, seeds of the same strain purchased from different seed companies were found to contain identical BevuMV1 sequences as expected. The commercially available Pacemaker III strain contains the same BevuMV1 as from UW-Madison's Pacemaker III. However, our data does not always align with the pedigree given in Goldman's 1996's description of the UW-Madison's germplasms. One discrepancy is seen in the expected maternal parent of Pacemaker III. Whereas Goldman's 1996 catalog says it should have W218 as the maternal parent and W260 as the paternal parent, the data suggests the opposite is true, if BevuMV1 is transmitted maternally.

4.7 References

Bruenn JA, Warner BE, Yerramsetty P. Widespread mitovirus sequences in plant genomes. *PeerJ*. 2015; 3:e876.

Brundrett MC. Coevolution of roots and mycorrhizas of land plants. *New Phytol*. 2002; 154: 275–304.

Depierreux D, Vong M, Nibert ML. Nucleotide sequence of Zygosaccharomyces bailii virus Z: Evidence for +1 programmed ribosomal frameshifting and for assignment to family *Amalgaviridae. Virus Res* 2016; 217:115–124.

Goldman IL. A list of germplasm releases from the University of Wisconsin table beet breeding program, 1964–1992. *HortScience* 1996; 31:880–881.

Goldman IL, Navazio JP. History and breeding of table beet in the United States. *Plant Breed. Rev*. 2003; 22:357-388.

Hillman BI, Cai G. The family *Narnaviridae*: simplest of RNA viruses. *Adv Virus Res* 2013; 86:149–176.

Hillman BI, Esteban R. Family *Narnaviridae*. In Virus Taxonomy, Ninth Report of the International Committee on Taxonomy of Viruses (Eds, King AMQ, Adams MJ, Carstens EB, Leftkowitz EJ). *Elsevier Academic Press*, San Diego. 2012; 1054-1060.

Hong Y, Cole TE, Brasier CM, Buck KW. Evolutionary relationships among putative RNA-dependent RNA polymerases encoded by a mitochondrial virus-like RNA in the Dutch elm disease fungus, *Ophiostoma novoulmi*, by other viruses and virus-like RNAs and by the *Arabidopsis* mitochondrial genome. *Virology* 1998; 246: 158–169.

Huang X, Madan A. CAP3: a DNA sequence assembly program. *Genome Res*. 1999; 9: 868–877.

Larsen K. Male sterility caused by beet yellows virus. *J Plant Disease and Protection* 2918; 88: 111-115.

Laser KD, Lersten NR. Anatomy and cytology of microsporogenesis in cytoplasmic male sterile angiosperms. *Bot Rev* 1972; 38:425–454.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol*. 2013; 30: 772–780.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 2017; 14: 587–589.

Kitahara R, Ikeda Y, Shimura H, Masuta C, Ezawa T. A unique mitovirus from *Glomeromycota*, the phylum of arbuscular mycorrhizal fungi. *Arch. Virol*. 2014; 159: 2157–2160.

Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol*. 2017; 34, 1812–1819.

Marienfeld JR, Unseld M, Brandt P, Brennicke A. Viral nucleic acid sequence transfer between fungi and plants. *Trends Genet* 1997; 13: 260–261.

Mikami T, Yamamoto MP, Matsuhira H, Kitazaki K, Kubo T. Molecular basis of cytoplasmic male sterility in beets: an overview. *Plant Genet Resources* 2011; 9: 284–287.

Minh BQ, Nguyen MAT, von Haeseler A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol*. 2013; 30: 1188–1195.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. *Mol. Biol. Evol.* 2015. 32: 268–274.

Nibert ML. Mitovirus UGA(Trp) codon usage parallels that of host mitochondria. *Virology*. 2017 Jul; 507:96-100.

Owen FV. Cytoplasmically inherited male-sterility in sugar beets. *J. Ag. Res.* 1945; 71: 423-440.

Pyle JD, Keeling PJ, Nibert ML. Amalga-like virus infecting *Antonospora locustae*, a microsporidian pathogen of grasshoppers, plus related viruses associated with other arthropods. *Virus Res*. 2017; 233: 95–104.

Shackelton LA, Holmes EC. The role of alternative genetic codes in viral evolution and emergence. *J Theor Biol* 2008; 254:128–134.

Stein H, Gabelman WH. Pollen sterility in *Beta vulgaris* associated with red pigmentation of the anthers. *J. Am. Soc. Sugar Beet Technol* 1959; X: 612-618.

Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res*. 2016; 44 (W1): W232–W235.

Wu MD, Zhang L, Li GQ, Jiang DH, Hou MS, Huang HC. Hypovirulence and doublestranded RNA in *Botrytis cinerea*. *Phytopathology* 2007; 97:1590–1599.

Xu Z, Wu S, Liu L, Cheng J, Fu Y, Jiang D, Xie J. A mitovirus related to plant mitochondrial gene confers hypovirulence on the phytopathogenic fungus *Sclerotinia sclerotiorum*. *Virus Res* 2015; 197:127–136.

Section II: Validation of novel viral genetic elements discovered from RNA-sequencing data

Chapter Five: A third clade of dsRNA satellites associated with Trichomonas vaginalis viruses (TVVs) and evidence that each satellite clade requires a different TVV species as helper virus

Some of the figures in this chapter may be used for a paper in preparation, as currently presented or in an updated form.

Special thank yous and contributions:

This project had started before I joined the lab and I want to thank May Yang for showing me how to handle the trichomonads. I determined the sequences for the two viruses in Tvag isolate UH711, the complete sequence for satellite S1-OC3, and the gel visualizing all the satellites. I also analyzed the data and created the figures, except for fig. 23, in this Chapter. Max supervised the project from inception to current stage.

5.1 Introduction

In 1986, the first virus infecting a protozoan was discovered in the obligate parasitic protozoan *Trichomonas vaginalis* (Tvag) (Wang 1991). This virus was named Trichomonas vaginalis virus (TVV) and later given its own genus, *Trichomonasvirus*, within the virus family *Totiviridae* (Goodman 2011b). Like most members of the family *Totiviridae*, TVV lacks the capability of extracellular transmission. Transmission of TVV is thought instead to occur only vertically. Despite the lack of extracellular transmission, at least half of all Tvag clinical isolates harbor TVV in the USA (Snipes 2000) with the frequency as high as 82% from some geographical locations (Weber 2003).

Like other members of *Totiviridae*, TVV is a mono-segmented dsRNA virus with a genome length around 4.7-4.9kb, depending on the TVV species (Goodman 2011a). The plus-strand encodes two overlapping ORFs. The first ORF encodes a capsid protein (CP) of about 76kDa. The second ORF encodes an RNA-dependent RNA polymerase of about 88kDa and is expressed as a fusion protein with the CP through a programmed ribosome frameshift (PRF). The TVV virions consist of its dsRNA genome encapsidated in a single layer, non-enveloped, icosahedral protein coat of diameter between 30 to 40nm with a buoyant density between 1.33 and 1.39 g/cm^3 (Ghabrial 2008).

Most recently, our lab has shown that Tvag isolates can harbor up to four different species of TVVs, named TVV1, TVV2, TVV3, and TVV4 (Goodman 2011a). Our lab has also solved the three-dimensional structure of the TVV1 virion, showing that it has a "T=2" symmetry similar to other members of *Totiviridae* (Parent 2013). Additionally, we have also provided fourteen of the twenty-one currently available TVV full-length genome sequences in the GenBank database. All the full-length TVV genomes sequenced by our lab begins with 5'-GC that is capable of folding

into a stem loop with the immediate downstream 24 nts and ends with UC-3' (Goodman 2011a). Before our contribution to the Genbank database in 2011 there were only five available TVV full-length genomes. We have also identified the newly ratified species TVV4 (Goodman 2011a), as well as propose that TVVs should be given its own genus, *Trichomonasvirus* (Goodman 2011b). Further, our lab has newly shown that the presence of TVV inside Tvag cells directly influences the disease outcome of trichomoniasis using culture endocervical cells as a model of human disease. The influence of TVV on the disease outcome of trichomoniasis is the result of the human cell's detection of the TVV dsRNA (Fichorova 2012).

While TVV replication has not yet been experimentally characterized, it is expected to be similar to those of *Totiviruses*: asymmetric (producing only plus-strand transcripts from its dsRNA genome), end-to-end (producing full-length copies), and conservative (both parental strands remain together during transcription). Further, TVV, similar to the most well-characterized totivirus, also seem to act as a helper virus that support dsRNA satellites. The dsRNA satellites of TVVs will inform a greater understanding of TVV replication, as the dsRNA satellite replication is dependent on the TVV's RdRp, and perhaps developing a genetics system to experimentally characterize TVV replication.

The limitation with using sequence information from dsRNA satellites of TVV is that, prior to this work, only three sequences have been reported and the sequences show no similarity to each other or to the TVV sequences. One dsRNA satellite was found in Tvag isolate T1 and two dsRNA satellites were found in Tvag isolate T068-II (GenBank U15991, U30166, U30167; Tai 1995; Khoshnan 1995). Further no new sequences of these satellites have been reported since their initial discovery in 1995 (Khoshnan 1995; Tai 1995), despite numerous on-going reports on TVV prevalence and TVV characterization.

In this work, we report the identification of new dsRNA satellites of TVV that show similarity to previously reported dsRNA satellites of TVV as well as a new clade of dsRNA satellites.

5.2 Materials and Methods

5.2.1 Clinical isolates

Tvag isolate UR1 was obtained from a symptomatic patient in upstate New York in 1999, first reported in 2000 (Gilbert 2000) and has been previously characterized by our lab and our collaborators in immunological and other assays (Fichorova 2012; Singh 2009). The stock of UR1 used in this study originated from the same soft-agar clone described in our previous sequencing study (Goodman 2011a). Tvag isolates OC3, OC4, OC5, UH711, and UH9 have also been described in previous reports (Fichorova 2012; Goodman 2011a). Original stocks of OC3, OC4, UH711, and UH9 were serially cloned twice in soft agar before use in this study. All isolates were grown in liquid batch culture for RNA isolation as previously described (Goodman 2011a).

5.2.2 Next-generation sequencing (NGS)

Total RNA from Tvag UR1 was obtained by TRIzol™ extraction (Life Technologies). Starting with 10 μg of RNA in 100 μL of binding buffer (10 mM Tris pH 7.5, 1 M NaCl, 1 mM EDTA), biotinylated oligonucleotides designed to hybridize to each of the Tvag rRNAs was added. The oligonucleotides for the 28S and 18S rRNAs were added at 100 pmol each and those for the 5.8S rRNA were added at 25 pmol each. The RNA/oligonucleotide mixture was then heated to 95°C for 5 min and cooled at 37°C for 30 min. Dynabeads M-280 Streptavidin (Life Technologies) were then added (2.5 mg in 250 μL equilibrated with binding buffer) and incubated for 15 min at

room temperature, followed by 2 min of magnet separation. The removed supernatant was then subjected to the same protocol again except that only the oligonucleotides for the 28S and 18S rRNAs were added at 25 pmol each. The supernatant removed from the beads from this repeat of the protocol was then treated with 6 units of RNAse-free DNAse I (Promega) for 10 min at 37°C, followed by 10 min at 75°C to inactivate the enzyme. The sample was then purified using an RNeasy Mini Kit (Qiagen), which yielded a total of ~125 ng RNA in 35 μL.

The rRNA- and DNA-depleted RNA sample (75 ng of the total) was submitted to the Biopolymers Facility at Harvard Medical School for NGS including library preparation, quality control and quantification, and sequencing by synthesis. Libraries of adapter-ligated DNA fragments were generated using the Illumina TruSeq RNA Library construction kit. The RNA sample was first re-suspended in the Elute-Prime-Fragment mix and incubated at 94°C for 8 min, followed by the remainder of the manufacturer's protocol. For assessment of library quality, samples were run on an Agilent 2100 Bioanalyzer on a High Sensitivity Chip with ladder provided. To assess library concentration, a quantitative PCR assay with SYBR green and primers to the P5 and P7 regions of the adapters was performed, using a serial dilution of bacteriophage PhiX control DNA for the standard curve. Sequencing by synthesis was then lastly performed on an Illumina HiSeq 2000 system using a Paired-End 50 flow cell. All processing and analysis of the sequencing reads were performed using the relevant tools from a licensed copy of CLC Genomics Workbench 6.5.2 (CLC bio) running locally on a Dell Precision T7600 computer with dual 6-core Intel Xeon E5-2630 processors and 64 GB of RAM.

5.2.3 Sanger sequencing

DsRNA was adapter-ligated and poly(A)-tailed as previously described (Goodman 2011a). Oligonucleotide primers were designed from the consensus sequences obtained by NGS. Satellite-specific primers and adapter-specific primers were then used for both RT–PCR (Qiagen OneStep RT–PCR kit) and direct Sanger sequencing. The sequences from both strands of each satellite were thereby determined in full, except for stretches of sequences at the satellite termini that were read only in the outward direction. To shorten these stretches read only in the one direction, we additionally determined sequences from plasmid clones. DsRNA from each satellite was reverse transcribed with SuperScript III (Life Technologies) followed by PCR amplification with Taq DNA polymerase (Platinum PCR SuperMix; Life Technologies). The amplicons were then used for TOPO PCR cloning (Life Technologies) according to the manufacturer's protocol. Insert-containing restriction digests from appropriate plasmids were then used for Sanger sequencing on both strands. In this manner, the sequence stretches read in only the outward direction were reduced to 20 nt at each satellite terminus.

5.2.4 Sequence analyses

Pairwise sequence alignments for identity scoring were obtained using EMBOSS 6.3.1: needleall as implemented at http://mobyle.pasteur.fr/. The EDNAFULL scoring matrix was used, along with a gap opening penalty of 10, a gap extension penalty of 0.5, and end gap penalties of those same values. Multiple sequence alignments were obtained using MAFFT 7.149 (Katoh 2013) with default settings as implemented at http://mafft.cbrc.jp/alignment/server/. The alignment was additionally used for phylogenetic analysis. The alignment was first analyzed with jModelTest 2.1.4 (Darriba 2012), which identified K80+G as the best nt-substitution model according to the

Bayesian information criterion. The alignment was then subjected to maximum-likelihood phylogenetic analysis with PhyML 3.0 (Guindon 2010) as implemented at http://www.hiv.lanl.gov/content/sequence/PHYML/interface.html. The starting tree was optimized according to both topology and branch length, and tree improvement was performed according to the best of NNI (nearest neighbor interchange) and SPR (subtree pruning and regrafting). Branch support was determined by the aLRT (approximate likelihood ratio test) with SH-like supports. The tree was prepared for presentation using TreeDyn 198.3 as implemented at http://www.phylogeny.fr/version2_cgi/index.cgi and a downloaded copy of FigTree 1.4.0 from http://tree.bio.ed.ac.uk/software/figtree/.

RNA folding was performed using RNAfold with default settings as implemented at http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi. The predicted stem–loop structures were then prepared for presentation using Pseudoviewer 3.0 as implemented at http://pseudoviewer.inha.ac.kr.

5.2.5 Visualization of dsRNA

Trichomonads were pelleted from a 250ml culture at stationary phase by centrifugation at 1,500 rpm at 4˚C for 30min. Two ml of TRIzol™ (Invitrogen) was used to extract total RNA from the trichomonad pellet according to manufacturer's instructions. A needle and syringe was used to break up pellet and suspend it in the TRIzol™. The RNA pellet was suspended in 2ml of 1X STE (100 mM NaCl, 50 mM Tris plus HCl to pH 7.5, 1 mM EDTA) +16% EtOH and used for dsRNA enrichment according to previously published methods (Depierreux 2016).

The dsRNA was separated on a 5% nondenaturing 0.5X TBE polyacrylamide gel electrophoresis using DEPC-treated solutions and stained with SYBR Green II in (Invitrogen) 1X TBE for 35min on shaker. The gel was visualized immediately, without wash or destain.

5.3 Results

5.3.1 NGS reveals presence of TVV1, TVV2, TVV3, and not TVV4 in Tvag isolate UR1

We used next-generation sequencing (NGS) to re-examine the TVV elements in T. vaginalis isolate UR1, which we previously found to be concurrently infected with TVV1, TVV2, TVV3, but not TVV4 (Goodman 2011a). Despite having depleted the sample of Tvag rRNAs, ~68% of all reads matched one of the three Tvag rRNA reference sequences (GenBank AF202181, U17510, and DQ029070). Still, reads matching the TVV1-UR1 reference sequence (GenBank HQ607513) made up 1.11% of the total. Reads matching the TVV2-UR1 reference sequence (GenBank HQ607514) made up 0.47% and reads matching the TVV3-UR1 reference sequence (GenBank HQ607515) made up 0.59%. In contrast, only 3 isolated reads matched any of the three published TVV4 reference sequences, consistent with our failure to find a TVV4 strain carried by Tvag UR1 in the previous study (Goodman 2011a).

The NGS reads were next subjected to contig assembly, which yielded plurality-defined consensus sequences covering nt positions 1–4,684 of TVV1-UR1 (full length relative to GenBank HQ607513); positions 1–4,674 of TVV2-UR1 (full length relative to GenBank HQ607513); and positions 4–4,845 of TVV3-UR1 (full length except for 3 nt at the plus-strand 5′ end relative to GenBank HQ607513). Coverage depth at each nt position in the consensus sequences averaged 23,228 for TVV1-UR1; 9,895 for TVV2-UR1; and 11,970 for TVV3-UR1, except at the termini. We have deposited the following new sequences in GenBank: TVV1-UR1, nt positions 11–4,676;

TVV2-UR1, nt positions 9–4,665; and TVV3-UR1, nt positions 12–4,836 (GenBank KM268108–KM268110).

Each of the mapping-derived, TVV consensus sequences could be aligned to the respective, published reference sequence without internal gaps and with >99.7% overall identity. The few differences that were present between the reference and consensus sequences numbered only 4 for TVV1-UR1, 8 for TVV2-UR1, and 7 for TVV3-UR1. None of the observed differences impinged on the ribosomal frameshifting region of each virus, and only 1–2 of the differences resulted in amino acid (aa) changes in the encoded CP protein, which were conservative in nature (Ile -> Val). The most notable difference was found within the 3′ untranslated region of TVV1-UR1, where the new consensus sequence contained a 1-nt A insertion relative to the reference sequence, within a run of A residues (A6 in the reference, A7 in the consensus). Based on these findings, published conclusions about these viruses were not substantively affected.

5.3.2 NGS reveals presence of two TVV satellites in Tvag isolate UR1: TVVS1-UR1 & TVVS1′-UR1

Within the remaining unmapped reads, 508 of the contigs had lengths >200nt and coverage depth >500. Examining these 508 contigs with BLAST (blastn/megablast), we found that 490 showed significant alignments with Tvag genes and 16 showed no significant alignments. Excitingly, the remaining 2 contigs showed significant alignments with either of the two TVV satellites (S1 and S1′; GenBank U30166 and U30167) previously reported by Khoshnan and Alderete (1995) from analysis of Tvag isolate T068-II. The lengths of these TVV satellite contigs from Tvag UR1 were 717 nts for TVVS1-UR1 and 542 nts for TVVS1′-UR1.

With this new evidence for two TVV satellites in Tvag UR1, we reverted to Sanger sequencing to corroborate their sequences and in particular to confirm that the NGS-derived sequences extended fully to the satellite termini (Methods and Materials 5.2.3 Sanger sequencing). This resulted in a sequence measured 683 nts for TVVS1-UR1 and 532 nts for TVVS1′-UR1 and have been reported to GenBank as entries KM262033 and KM262034. These newly determined reference sequences of TVVS1-UR1 and TVVS1′-UR1 were next used for mapping the original NGS reads. The results of this procedure were plurality-defined consensus sequences covering nt positions 1–683 of TVVS1-UR1 (full length relative to GenBank KM262033) and positions 1–532 of TVVS1′-UR1 (full length relative to GenBank KM262034). Reads matching the new TVVS1-UR1 reference sequence consisted of 0.33% of the total, and reads matching the new TVVS1′-UR1 reference sequence consisted of 1.76%. Coverage depth at each nt position in the satellite consensus sequences averaged 47,723 for TVVS1-UR1 and 322,758 for TVVS1′-UR1. Each of the mapping-derived consensus sequences for the TVV satellites could be aligned to the reference sequence without internal gaps and with 100% identity, leading us to report these NGS-derived sequences to GenBank under the same entries as the Sanger-derived sequences, KM262033 and KM262034.

5.3.3 A new clade of TVV satellite in Tvag concurrently infected with all four TVV species

Having discovered the presence of TVVS1 and TVVS1′ in Tvag UR1, we used Sanger sequencing to examine related satellites in other previously characterized Tvag isolates, starting with Tvag OC3, which is concurrently infected with all four known TVV species (Goodman 2011a). The results revealed the presence of not just TVVS1 and TVVS1′ in OC3, but a third TVV satellite closely related to TVVS1′. We designate the two closely related TVVS1′ sequences as

TVVS1′A-OC3 (length 639 bp) and TVVS1′B-OC3 (length 535 bp), which share only ~54% sequence identity including multiple gaps in pairwise alignments. Notably, when subjected to phylogenetic comparisons, TVVS1′A-OC3 was seen to relate more closely to TVVS1′-T068-II (616 bp) than to TVVS1′-UR1 (532 bp), and TVVS1′B-OC3 was seen to relate more closely to TVVS1′-UR1 than to TVVS1′-T068-II (Fig. 18). As a result, we have redesignated these other two S1′ satellites as TVVS1′A-T068-II and TVVS1′B-UR1.

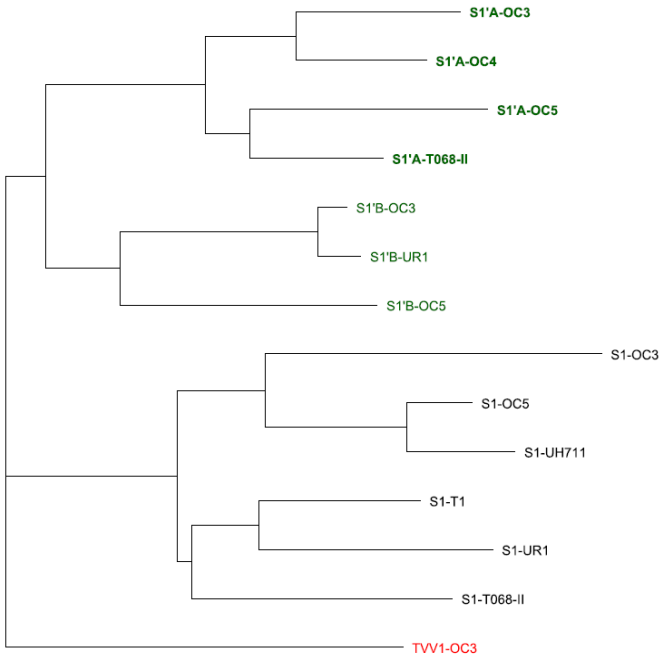


Figure 18 Phylogenetic analysis of dsRNA satellites of TVV. Bold green: TVVS1′A. Non-bold green: TVVS1′B. Red: TVV1 of Tvag isolate OC3.

5.3.4 Possible TVV-species specific association among TVV satellites

We additionally examined for TVV satellites in previously characterized Tvag isolate OC4, which carries strains of only two known TVV species, TVV1 and TVV4 (Goodman 2011a). The results revealed the presence of only TVVS1′A, again consistent with the hypothesis that TVVS1′A may require TVV4 as helper virus. Moreover, our failure to find TVVS1 and TVVS1′B

in this isolate, which does not carry a strain of TVV2 or TVV3, suggested to us that TVVS1 and TVVS1´B may require TVV2 and/or TVV3 as helper virus(es).

Tvag isolate UH711 has been previously reported to be TVV infected (Fichorova 2012), but which of the four known TVV species it carries have remained to be defined. Subsequent work in our labs, reported here for the first time, provided RT-PCR and Sanger sequencing evidence that UH711 carries strains of only two known TVV species, TVV1 and TVV3. Given the unique combination of TVV species in this isolate, we also tested it for TVV satellites, which revealed the presence of only TVVS1. The presence of TVVS1 in UH711, which carries strains of TVV1 and TVV3, but not in OC4, which carries strains of TVV1 and TVV4, is consistent with the new hypothesis that TVVS1 may specifically require TVV3 as helper virus. Moreover, our failure to find TVVS1´A and TVVS1´B in OC4 is again consistent with the hypothesis that TVVS1´A may require TVV4 as helper virus and also consistent with the new hypothesis that TVVS1´B may specifically require TVV2 as helper virus.

We also examined for TVV satellites in previously characterized Tvag isolate UH9, which contains only TVV1 (Goodman 2011a). Notably, we failed to find any of the three known satellites, TVVS1, TVVS1´A, or TVVS1´B, in this isolate, consistent with the new hypothesis that TVV1 is not sufficient to function as helper virus for any of these known satellites.

The presence of TVVS1´A in OC3, which carries strains of all four known TVV species, but not in UR1, which does not carry a TVV4 strain, suggested to us that TVVS1´A may require TVV4 as helper virus.

Lastly, a TVV-species specific association within just three of the TVV species suggests three distinct clades of dsRNA satellites. To improve on verifiying there indeed exists three clades for this current report, we lastly examined Tvag isoate OC5, which we have previously shown to

contain all three TVV species. Interestingly, Tvag OC5 also contains three distinct dsRNA satellites, though we currently do not have the terminal sequences determined for these satellites. A phylogenetic analysis of all known dsRNA satellites of TVV indeed forms three unique clusters (Fig. 18). Additionally, pairwise sequence comparisons of all the dsRNA satellites show that dsRNA satellies within a presumed clade share >60% sequence similarity compared to <50% sequence similarity between dsRNA satesllites (Fig. 19). A discussion on the fairly low sequence similarity within clades is found below.

	K/A	Tai					K/A						
	T068-SII **S1**	T1	UR1 S1	UH711 S1	OC5 S1	OC3 S1	T068-SII **S1'**	OC3 S1'A	OC4 S1'A	OC5 S1'A	UR1 S1'B	OC3 S1'B	OC5 S1'B
T068-SII S1	100	63	66	70	71	69	51	50	50	48	49	50	50
T1		100	63	61	61	66	51	51	52	52	51	51	51
UR1			100	68	61	68	50	49	50	50	51	51	50
UH711 S1				100	89	64	51	51	51	50	52	52	50
OC5 S1					100	69	50	51	50	50	50	52	53
OC3 S1						100	52	51	50	49	49	50	51
T068-SII **S1'**							100	68	73	72	58	57	60
OC3 S1'A								100	86	70	56	54	51
OC4 S1'A									100	72	57	58	59
OC5 S1'A										100	55	55	60
UR1 S1'B											100	92	67
OC3 S1'B												100	67
OC5 S1'B													100

Figure 19 Pairwise sequence comparisons of currently identified TVV-associated dsRNA satellites. K/A: satellites identified previously by Khoshnan and Alderete (1995). Tai: satellite identified previously by Tai et al (1995).

5.3.5 Visualization of dsRNA bands corresponding to sizes of TVV satellites

Use of polyacrylamide gel with SYBR Green II stain allowed visualization of enriched dsRNA bands corresponding to appropriate dsRNA TVV satellite lengths; routine imaging was not achievable with agarose gel electrophoresis with ethidium bromide staining.

Two satellite dsRNA bands were observed from samples of Tvag UR1 (Fig. 20), migrating near above the 500-bp dsRNA marker and thus consistent with the 683-bp satellite TVVS1-UR1 and the 532-bp satellite TVVS1′B-UR1. No such bands of smaller-sized dsRNA molecules were seen with samples from Tvag UH9, which Sanger sequencing was not able to detect any TVV

satellites. Gel results for Tvag isolates OC3, OC4, and UH711 were also consistent with the sequencing results, in that one smaller-sized dsRNA molecule was seen in UH711, consistent with the 652-bp satellite TVVS1-UH711; one smaller-sized dsRNA molecule was seen in OC4, consistent with the 644-bp satellite TVVS1′A-OC4; and two or more smaller-sized dsRNA molecules were seen in OC3, consistent with the satellites TVVS1′A-OC3 and TVVS1′B-OC3, with the addition of a larger dsRNA molecule consistent with the 712-bp satellite TVVS1-OC3.

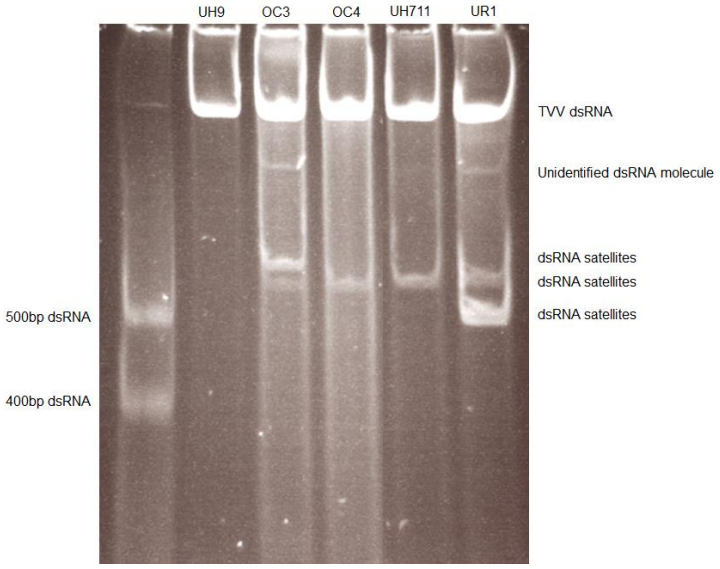


Figure 20 Non-denaturing gel loaded with enriched dsRNA from various TVV-infected Tvag strains. 1st lane: dsRNA marker. UH9: infected by only TVV1 and no satellites. OC3: infected with 4 TVV species and 3 satellites. OC4: infected with 2 TVV species and 1 satellite. UH711: infected with 2 TVV species and 1 satellite. UR1: infected with 3 TVV species and 2 satellites.

Interestingly, our gel analysis revealed a third dsRNA molecule outside of the size ranges of both the TVV genome and satellite range (marked "Unidentified dsRNA molecule" in Fig. 20). This dsRNA molecule was found in Tvag isolates OC3 and UR1. Notably, this is not the first time that such a molecule has been reported. Khoshnan et al. had reported seeing a similar dsRNA banding pattern from uncharacterized Tvag isolate 06201 (Khoshnan 1994). In their report, they isolated total dsRNA from various Tvag isolates and found that isolate 06201 had three distinctly sized dsRNAs, similar to OC3 and UR1: one dsRNA molecule corresponding to the 4.6kb of the

TVV genome, a second group of small dsRNAs corresponding to the later identified dsRNA satellites, and a third dsRNA molecule with the size inbetween TVV and the satellite, size ~1.7kb (Khoshnan 1994).

5.4 Conclusion

In this report, we have identified and determined the sequences for ten newly reported dsRNA satellites of TVV, bringing the total to thirteen, including the three previously reported satellites (Khoshnan 1995; Tai 1995). As per sequence determination, we are currently missing the terminal sequences for four of the dsRNA satellites, TVVS1'A-OC3, TVVS1'A-OC5, TVVS1'B-OC5, TVVS1-OC5, but have been able to determine the full-length sequnences of the other six we identified in this report. In the newly identified sequences, we have also identified a third clade of dsRNA satellites, which seem closely related to previously identified TVVS1´-T068-II, thus we have henced renamed TVVS1´-T068-II into TVVS1´A-T068-II to reflect this relatedness. All current dsRNA satellites to date can be clustered into the clades: TVVS1, TVVS´1A, and TVVS1´B; six dsRNAs clustering into clade TVVS1, four dsRNAs clustering into TVVS1´A, and three dsRNAs clustering into TVVS1´B (Fig. 18 & 19). DsRNA bands of expected sizes were also visualized from these Tvag isolates (Fig. 20). Tvag isolate UR1 contains two dsRNA satellites, TVVS1 and TVVS1'B, in addition to its previously recognized three TVV species, TVV1, TVV2, and TVV3. Tvag isolates OC3 and OC5 each contains three satellites, TVVS1, TVVS1'A, and TVVS1'B, in addition to its previously recognized four TVV species, TVV1-4. Tvag isolate OC4 contains TVVS1'A and TVV1 and TVV4. Tvag isolate UH711 contains TVVS1 and TVV1 and TVV3 (Table 8). The distribution of TVV species and dsRNA satellites within each Tvag isolate and the apparent clustering of dsRNA satellites into three clades may suggest a possible TVV-species specific association between dsRNA satellites clades and helper virus.

Table 8 Summary of Tvag isolates and their associated TVV-related viral agents.

Tvag isolate	TVV1	TVV2	TVV3	TVV4	S1	S1´A	S1´B
UH9	+						
OC4	+			+		+	
UH711	+		+		+		
UR1	+	+	+		+		+
OC3	+	+	+	+	+	+	+
OC5	+	+	+	+	+	+	+

5.5 Discussion

5.5.1 Conserved terminal sequences and structures

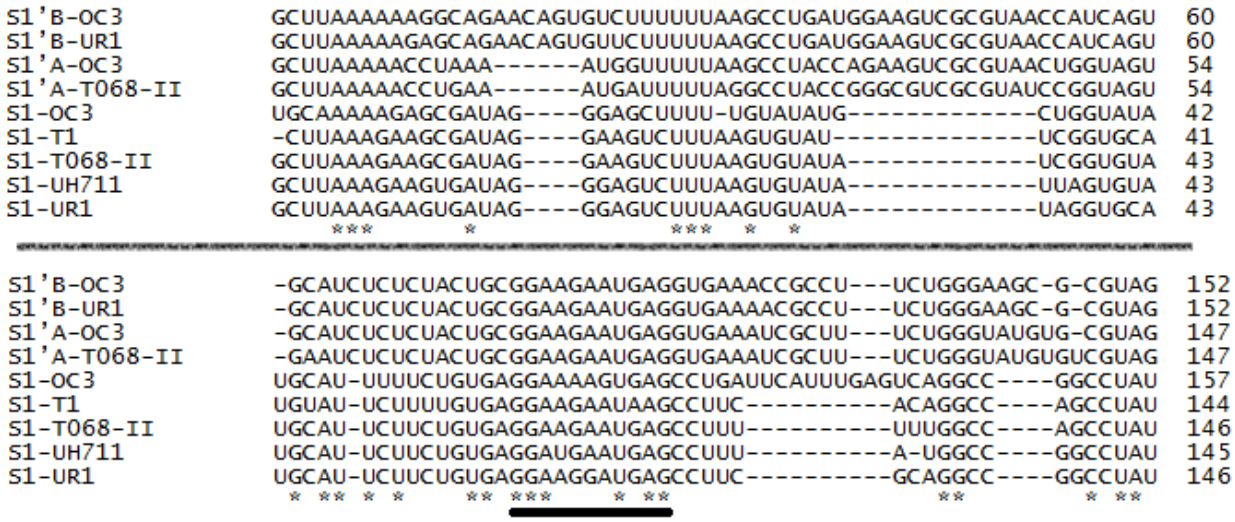


Figure 21 Multiple sequence align of full-length (nine available out of thirteen) dsRNA satellites with focus on the 5' terminal sequence and previously noted conserved domain by Khoshnan (1995).

A multiple sequence alignment of the nine full-length satellite sequences reveals a conserved 8- or 9-nt terminal sequence (G)CUUAAARA between eight of the nine satellites. TVVS1-OC3 does not share this terminal sequence. In the four S1 satellites, this sequence is (G)CUUAAAGA, whereas in the two S1´ satellites, it is GCUUAAAAA. Further, the four S1 satellites share a much longer conserved terminal sequence, (G)CUUAAAGAAGYGAUAGGRAGUCUUUAAGUGUAU (Fig. 21). Khosnan et al. pointed

out an internal sequence of GGAAGAAUGAG that they think might be conserved (1995). For notetaking purposes, all of the nine newly identified satellites in this report also contain this sequence within the same nucleotide position (underlined region of Fig. 21). With the incorporation of the additional satellite sequences the consensus seems to be GGAWRAAURAG.

A predicted stem–loop in TVVS1-T1 was previously noted by Su and Tai (Su 1996). This finding is especially notable because similar stem–loop structures have the propensity to form by the 5′-terminal plus-strand sequences of all four TVV species, again beginning at residue 1 (Goodman 2011a). Additionally, out of the five S1-satellite terminal sequences we have determined, four of these 5' terminal sequences have a tendancy to fold into a nearly identical stem-loop (Fig. 22). As previously proposed (Goodman 2011a), these 5′-terminal RNA structures seem likely to play a functional role by partially protecting the plus-strand transcripts, which are putatively uncapped, from 5′-exonuclease digestion, and could also be more directly involved in RNA packaging, replication, or translation.

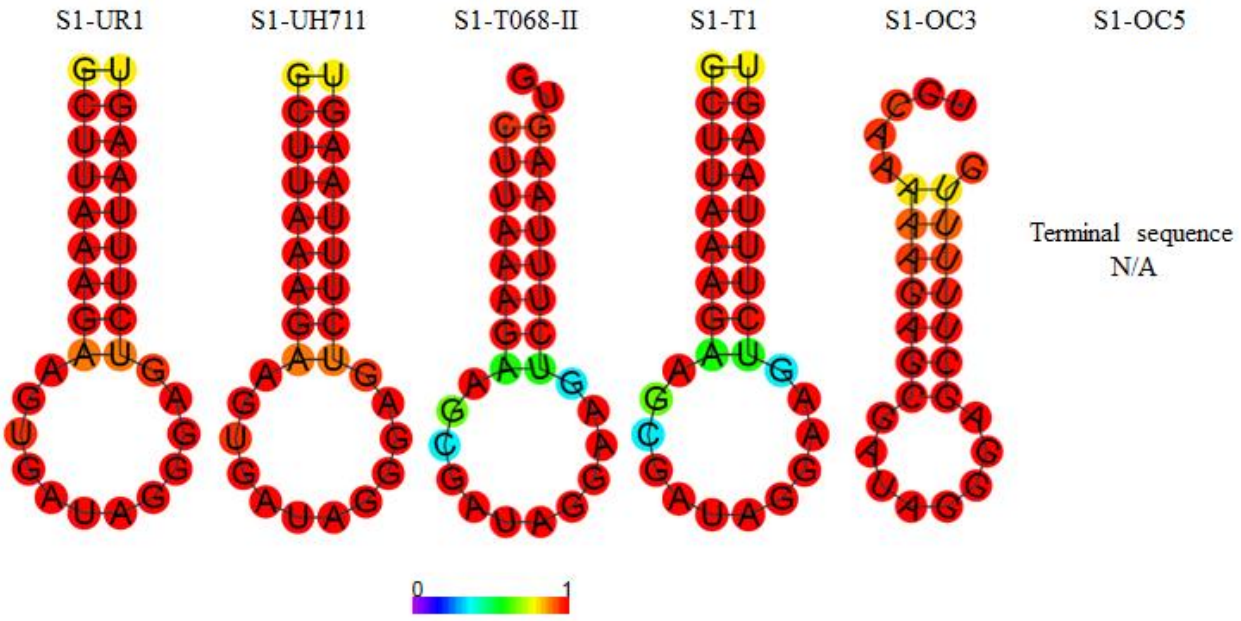


Figure 22 Stem-loop predictions in S1 satellites. Color coding represents probability of pair bonding based on minimum free-energy calculations (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi).

5.5.2 Protein-coding potential

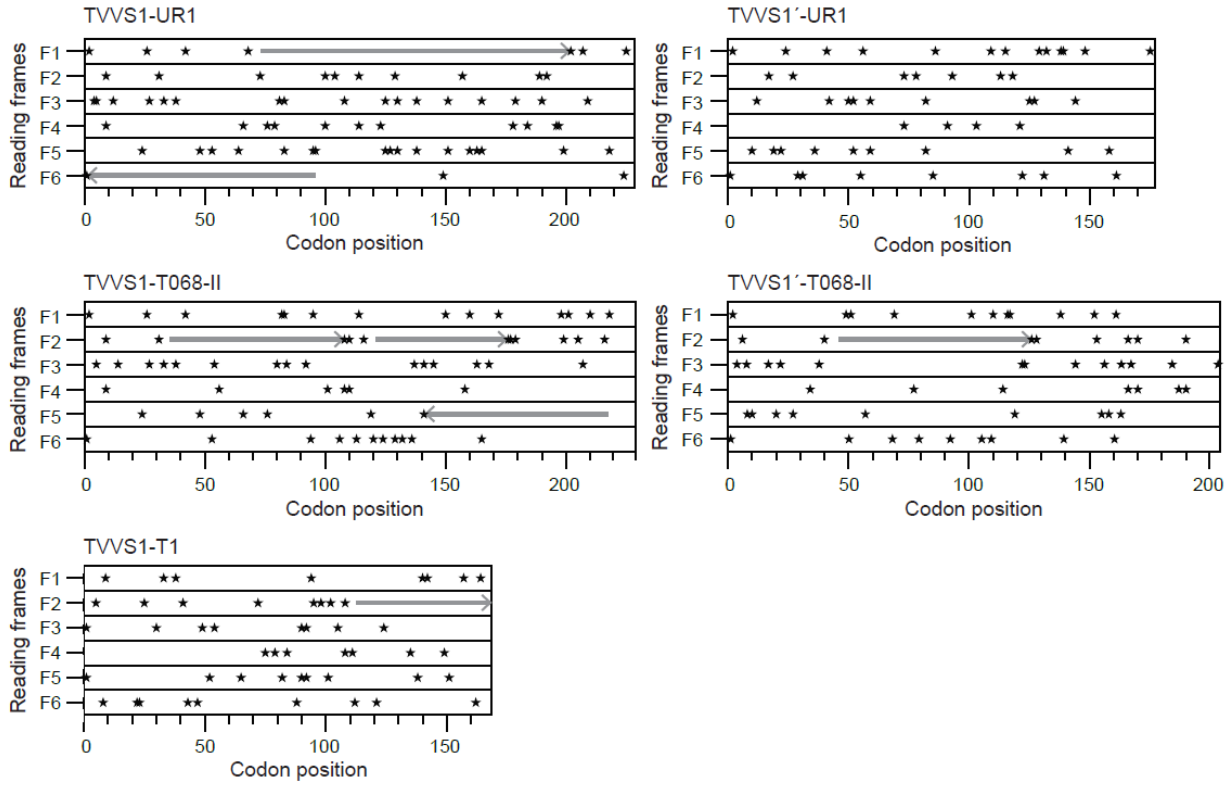


Figure 23 Potential open reading frame of satellites. Asterisks represent stop codons. Potential open reading frames >49 codons are marked by extended arrows.

Examination for stop codons in all six frames of the five satellites shows that potential open reading frames longer than 150 nt (50 aa) are uncommon and are in fact not conserved among either the S1 or the S1´ satellites (Fig. 23). As previously suggested by Khoshnan and Alderete (1995) and Tai et al. (1995), and again suggested by the findings for TVVS1-UR1 and TVVS1´-UR1 here, it therefore seems likely that neither of these TVV satellites gives rise to protein products. Further, as previously noted above, the very low overall sequence similarity from pairwise comparisons is also consistent with the lack of protein-coding capability of these satellites (Fig. 19). Thus, similarity between satellites of a putative clade is not expected to be very high if no essential proteins are encoded within the genomes. Additionally, the apparent lack of protein-coding capability among these TVV satellites presents a difficulty in easily identifying which of

the dsRNA strands is “plus sense”, in other words in this case, which of the two complementary strands is transcribed from the dsRNA satellite genome packaged inside the TVV capsid and then extruded into the cytosol for packaging into a nascent, progeny TVV capsid. Notably, previous studies have disagreed on this point: Khoshnan and Alderete (1995) have identified one of the TVV satellite strands as plus sense but Tai et al. (1995) have identified the complement as such. We favor the latter conclusion in large part because of the conserved 5′-terminal sequences and stem–loop structures discussed above and shown in Fig. 22. As a result, as well as in GenBank entries KM262033 and KM262034, we have presented the satellite sequences according to the convention that the proposed plus-strand sequences are shown and proceed left to right from 5′ to 3′ end, in this case starting with the conserved 5′-terminal sequence (G)CUUAAA in each characterized satellite, except in S1-OC3 (Fig. 24).

TVV1-UH9	GCAAAAAGAGGGAGUGAUCC
TVV1-UR1	GCAAAAAGAGGGAGUCACCC
TVV1-OC3	GCAAAAAGAGGGGGUCAUCC
TVV1-OC4	GCAAAAAGAGGGAGUGAUCC
TVV1-OC5	GCAAAAAGGAGGGAGUAGUC
TVV2-UR1	GCUUUGAAGGAGUGACGACC
TVV2-OC3	GCUUUAAAAGGAGUGACGAC
TVV2-OC5	GCUUUGAAGGAGUGACGACC
TVV3-UR1	GCUUAAAAAGCGAAGUCCAC
TVV3-OC3	GCUUAAAAGGUCUAGUCCAC
TVV3-OC5	GCUUAAAAGCUUAGUCCACU
TVV3-TAI	-CUUAAAAAGCCUAGUCCAC
TVV4-OC3	GCUUAAAGUCCCAGUGAGCU
TVV4-OC5	GCUUAAAGCCCCAGUGAGCU
S1′B-OC3	GCUUAAAAAAGGCAGAACAG
S1′B-UR1	GCUUAAAAAGAGCAGAACAG
S1′A-OC3	GCUUAAAAACCUAAAAUGGU
S1′A-T068-II	GCUUAAAAACCUGAAAUGAU
S1-OC3	UGCAAAAAGAGCGAUAGGGA
S1-T1	-CUUAAAGAAGCGAUAGGAA
S1-T068-II	GCUUAAAGAAGCGAUAGGAA
S1-UH711	GCUUAAAGAAGUGAUAGGGA
S1-UR1	GCUUAAAGAAGUGAUAGGGA

Figure 24 Identifying putative plus strand of satellites. Terminal sequences for TVVs of Tvag isolate UH711 are not available.

5.5.3 Unidentified dsRNA molecule

More experiments aimed at identifying and characterizing the unidentified dsRNA molecule are necessary before confidently drawn conclusions can be made (Fig. 20). Though serious work aimed at identifying this band was not of interest, an initial attempt has been made. This attempt, however, did not provide clean sequencing reads.

Physical properties of this unidentified dsRNA band can be used to suggest possible identities. First, this new band was visualized after an enrichment protocol for dsRNA, which we can conclude is successful because we do not see any ribosomal RNAs and we see the presence of known dsRNAs such as the TVV genome and dsRNA satellites (Fig. 20). However, circular RNA will also be enriched from the same protocol, so this band could either be linear dsRNA or circular RNA with a length of about ~1.7kb. This band is unlikely derived from the Tvag genome. While eukaryotic circular RNAs have been found to be expressed across the many eukaryotes, including protozoans, their sizes are usually less than 1k nts and are not expressed in high levels (Wang 2014; Grabowski 1981). Radioactive-labeled probes are commonly used to detect eukaryotic circular RNAs (Grabowski 1981; Memczak 2012). Thus the band's intensity suggests that it is unlike other characterized eukaryotic circular RNAs. Further, we do not see this band in all Tvag isolates (Fig. 20). This band has also only been reported in one other Tvag isolate (Khoshnan 1994). Thus, this band could either be a large linear or circular dsRNA, but unlikely derived from the host Tvag.

Next, the size of this band, ~1.7kb, also suggests that this band is a subviral agent, as opposed to an entirely different species of virus that infects Tvag. While concomitant infection of multiple virus species is found in Tvag and other hosts for viruses of family *Totiviridae*, the size of ~1.7kb would make it unlikely that this band represents another virus agent. The smallest known

RNA viruses belong to the genus *Mitovirus*. The smallest of the mitoviruses has a genome length of ~2.1kb and the shortest RdRp sequence of these viruses has a length of ~2kb (Hillman 2013). Further, subviral agents seem to be very common for virus family *Totiviridae*. The most well-characterized virus of *Totiviridae* is the L-A virus which infects *Saccharomyces cerevisiae*. The L-A virus system contains two out of the three known nucleic acid-based subviral agents that are found outside of plants. The types of subviral agents are satellite viruses, satellite nucleic acids, and defective interfering agents. The L-A system has so far been found to contain the last two types of agents. The 1.8kb dsRNA satellite called M requires the helper L-A virus for replication and encodes a preprotoxin system (Schmitt 2006). There is also a 0.5kb dsRNA that is a defective interfering agent of L-A called X. X also requires the helper L-A virus for replication and its genome sequence is derived entirely from the L-A virus's genome (Esteban 1988). Lastly, the L-A system contains yet another component. The M satellite of L-A also has its own defective interfering agent called S. The 0.8kb dsRNA known as S has a genome sequence that is derived entirely from the sequence of the M satellite (Lee 1986). Thus, this unidentified band is unlikely to be an independent virus agent and more likely be involved in the TVV system as a subviral agent. It may be a defective interfering agent to TVV, though much larger than the X molecule of L-A. So, more likely it may be either the original sequence source of the dsRNA satellites of TVV, which means that the dsRNA satellites are defective interfering agents of this unidentified band, or possibly a satellite virus if it encodes a capsid protein product.

5.6 Future Directions

As noted above, the four described TVVS1 satellites each contain a much longer conserved terminal sequence that can also fold into a stem-loop that is distinct from the other clades but

similar to each other. Additionally, while the satellite sequences do not have much similarity overall and likely lack any protein-coding potential, small stretches of conserved sequences have been noted previously and also found in the newly reported satellites here (Fig. 21). This may further support a TVV-species specific association between satellite and helper virus, particularly between TVVS1 and TVV3, but much evidence is still needed to support this claim.

First, while Tvag isolate T068-II has two identified dsRNA satellites, TVVS1-T068-II and TVVS1'A-T068-II, which of the four known TVV species are present in Tvag isolate T068-II remain(s) unknown (Khoshnan 1995). Our data suggests that Tvag T068-II may at least be infected with both species TVV3 and TVV4. Whether it should also contain species TVV1 is not yet clear. While all of our TVV-infected isolates to date contain TVV1, and originated largely from the northeastern United States of America, five independent groups from different various geographical regions have reported Tvag isolates singly infected with TVV2 or TVV3 from screening all four species or using other methods (Bessarab 2011; Fraga 2012; El-Gayar 2016; Rivera 2017; Jehee 2017). Intriguingly, Bessarab et al. reported that Tvag isolate T068-II contains three dsRNA TVV bands and have identified two of the three species as TVV2 and TVV3 while failing to identify TVV1, implying that the third species is TVV4 (Bessarab 2011). Also, Tvag isolate T1 has only been reported to contain species TVV1 and TVV2 (Bessarab 2010). Our results suggest that it may also contain TVV3, but Bessarab et al., while reporting the discovery of TVV3, failed to identify TVV3 in Tvag isolate T1. However, the results are complicated by the fact that the team also failed to detect TVV1-T1 in the same report, although they had identified TVV1-T1 just one year earlier (Bessarab 2010; Bessarab 2011). Reports of Tvag being cured of TVV from lab cultures are very rare and has not happened to our Tvag isolates (Goodman 2011b).

Thus, it remains open whether the claim of a TVV-species specific association between satellite and helper virus can be generalized for all members of a satellite clade or whether each member can rely on any species of replication. However, we still think that it is more likely that each specific satellite uses a single species for replication rather than multiple helper species for replication. Towards addressing this generalized claim among various isolates of Tvag, a TVV-mediated reverse genetics system would be useful. Such a system will more importantly allow us to test whether the aforementioned conserved sequences or stem-loops have any function with regard to satellite replication.

5.7 References

Bessarab IN, Liu HW, Ip CF, Tai JH. The complete cDNA sequence of a type II Trichomonas vaginalis virus. *Virology.* 2000 Feb 15;267(2):350-9.

Bessarab IN1, Nakajima R, Liu HW, Tai JH. Identification and characterization of a type III Trichomonas vaginalis virus in the protozoan pathogen *Trichomonas vaginalis*. *Arch Virol*. 2011 Feb;156(2):285-94.

Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* 2012; 9: 772.

Depierreux D, Vong M, Nibert ML. Nucleotide sequence of Zygosaccharomyces bailii virus Z: Evidence for +1 programmed ribosomal frameshifting and for assignment to family *Amalgaviridae. Virus Res* 2016; 217:115–124.

El-Gayar EK, Mokhtar AB, Hassan WA. Molecular characterization of double-stranded RNA virus in *Trichomonas vaginalis* Egyptian isolates and its association with pathogenicity. *Parasitol Res.* 2016 Oct;115(10):4027-36.

Esteban R, Fujimura T, Wickner RB. Site-specific binding of viral plus single-stranded RNA to replicase-containing open virus-like particles of yeast. *Proc. Natl. Acad. Sci. USA* 1988; 85:4411-4415.

Fraga J, Rojas L, Sariego I, Fernández-Calienes A. Genetic characterization of three Cuban Trichomonas vaginalis virus. Phylogeny of *Totiviridae* family. *Infect Genet Evol.* 2012 Jan;12(1):113-20.

Fichorova RN, Lee Y, Yamamoto HS, Takagi Y, Hayes GR, Goodman RP, Chepa-Lotrea X, Buck OR, Murray R, Kula T, Beach DH, Singh BN, Nibert ML. Endobiont viruses sensed by the human host—beyond conventional antiparasitic therapy. *PLoS One* 2012; 7:e48418.

Ghabrial SA. *Totiviruses*. In: Mahy BWJ, Van Regenmortel MHV (eds) Encyclopedia of Virology, vol 5, 3rd edn. *Elsevier*, San Diego. 2008; 163–174.

Gilbert RO, Elia G, Beach DH, Klaessig S, Singh BN. Cytopathogenic effect of *Trichomonas vaginalis* on human vaginal epithelial cells cultured in vitro. *Infect Immun* 2000; 68: 4200–4206.

Goodman RP, Freret TS, Kula T, Geller AM, Talkington MW, Tang-Fernandez V, Suciu O, Demidenko AA, Ghabrial SA, Beach DH, Singh BN, Fichorova RN, Nibert ML. Clinical isolates of *Trichomonas vaginalis* concurrently infected by strains of up to four *Trichomonasvirus* species (Family *Totiviridae*). *J Virol* 2011a; 85: 4258–4270.

Goodman RP, Ghabrial SA, Fichorova RN, Nibert ML. *Trichomonasvirus*: a new genus of protozoan viruses in the family *Totiviridae*. *Arch Virol* 2011b; 156: 171–179.

Grabowski PJ, Zaug AJ, Cech TR. The intervening sequence of the ribosomal RNA precursor is converted to a circular RNA in isolated nuclei of *Tetrahymena*. *Cell*. 1981 Feb;23(2):467-76.

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

Hillman BI, Cai G. The family *Narnaviridae*: simplest of RNA viruses. *Adv Virus Res*. 2013;86:149-76.

Jehee I, van der Veer C, Himschoot M, Hermans M, Bruisten S. Direct detection of Trichomonas vaginalis virus in *Trichomonas vaginalis* positive clinical samples from the Netherlands. *J Virol Methods*. 2017 Dec;250:1-5.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013; 30: 772–780.

Khoshnan A, Alderete JF. Characterization of double-stranded RNA satellites associated with the Trichomonas vaginalis virus. *J Virol* 1995; 69: 6892–6897.

Khoshnan A1, Provenzano D, Alderete JF. Unique double-stranded RNAs associated with the Trichomonas vaginalis virus are synthesized by viral RNA-dependent RNA polymerase. *J Virol*. 1994 Nov; 68(11):7108-14.

Lee M, Pietras DF, Nemeroff ME, Corstanje BJ, Field LJ, Bruenn JA. Conserved regions in defective interfering viral double-stranded RNAs from a yeast virus. *J. Virol.* 1986; 58:402-407.

Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD, Gregersen LH, Munschauer M, Loewer A, Ziebold U, Landthaler M, Kocks C, le Noble F, Rajewsky N. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*. 2013 Mar 21;495(7441):333-8.

Parent KN, Takagi Y, Cardone G, Olson NH, Ericsson M, Yang M, Lee Y, Asara JM, Fichorova RN, Baker TS, Nibert ML. Structure of a protozoan virus from the human genitourinary parasite *Trichomonas vaginalis*. *MBio*. 2013; 4(2). pii: e00056-13.

Rivera WL, Justo CAC, Relucio-San Diego MACV, Loyola LM. Detection and molecular characterization of double-stranded RNA viruses in Philippine *Trichomonas vaginalis* isolates. *J Microbiol Immunol Infect.* 2017 Oct;50(5):669-676.

Schmitt MJ, Breinig F. Yeast viral killer toxins: Lethality and self-protection. *Nature Reviews. Microbiology* 2006; 4: 212–221.

Singh BN, Hayes GR, Lucas JJ, Sommer U, Viseux N, Mirgorodskaya E, Trifonova RT, Sassi RR, Costello CE, Fichorova RN. Structural details and composition of *Trichomonas vaginalis* lipophosphoglycan in relevance to the epithelial immune function. *Glycoconj J* 2009; 26: 3–17.

Snipes LJ, Gamard PM, Narcisi EM, Beard CB, Lehmann T, Secor WE. Molecular epidemiology of metronidazole resistance in a population of *Trichomonas vaginalis* clinical isolates. *J Clin Microbiol* 2000; 38:3004–3009.

Tai JH, Chang SC, Ip CF, Ong SJ. Identification of a satellite double-stranded RNA in the parasitic protozoan *Trichomonas vaginalis* infected with T. vaginalis virus T1. *Virology* 1995; 208: 189–196.

Weber B, Mapeka TM, Maahlo MA, Hoosen AA. Double stranded RNA virus in South African *Trichomonas vaginalis* isolates. *J Clin Pathol* 2003; 56:542–543.

Wang PL, Bao Y, Yee MC, Barrett SP, Hogan GJ, Olsen MN, Dinneny JR, Brown PO, Salzman J. Circular RNA is expressed across the eukaryotic tree of life. *PLoS One.* 2014 Mar 7;9(6):e90859.

Wang AL, Wang CC. Viruses of the Protozoa. *Annu. Rev. Microbial.* 1991; 45:251-63.

Section III: Application of understanding the viral genomes of *Totiviridae*

Chapter Six: A transient virus-mediated reverse genetics system in *Trichomonas vaginalis*

The figures in this chapter may be used for a future publication.

Special thank yous and contributions:

I want to thank my dissertation advisory committee members, Professors Carig Hunter, Victoria D'Souza, and Lee Gehrke for their advice and offering their labs for help on this Chapter. I want to give a deep thank you to Carolina Salguero, Vincent Pham, and Nico Wagner from Professor D'Souza's lab for their help in troubleshooting the *in vitro* transcription and providing me with their lab-made T7 polymerase. I also want to give a deep thank you to Dr. Kate Godin from Professor D'Souza's lab for helping me with my SHAPE experiments and analysis. I also feel grateful for Professor Patricia Johnson and Dr. Brian Janssen for gifting us with pMASTER-neo and a special thanks to Dr. Janssen for helping me troubleshoot the transfection and selection protocol. With their help I was able to develop this transient and selectable transfection virus-mediated reverse genetics system. I analyzed all the data and generated all the figures for this Chapter. Max really helped me organize my thoughts and helped with troubleshooting and designing this system.

6.1 Introduction

6.1.1 The protozoan *Trichomonas vaginalis* and trichomoniasis

Trichomonas vaginalis (Tvag) is a parasitic protozoan that causes the sexually transmitted disease (STD) trichomoniasis. The disease outcomes of trichomoniasis vary from asymptomatic to severe inflammation and irritation. More serious outcomes include increased transmission and acquisition of HIV infection and induction of preterm birth in pregnant women (Cotch 1997). While trichomoniasis is curable through treatment with nitroimidazoles, Tvag resistance to metronidazole is on the rise and metronidazole treatment for pregnant women with trichomoniasis may increase the risk of preterm birth. With 276 million annual new cases of Tvag infections worldwide, Tvag infects more new individuals than *Chlamydia trachomatis, Neisseria gonorrhoeae*, and *syhphilis* combined (WHO 2012). In 2008 there were 276 million new cases of Tvag infections worldwide. In the United States, the estimated number of annual new cases of Tvag infection is estimated at 7.4 million and the prevalence of Tvag infections in inner city U.S. STD clinics ranges from 25-38%, depending on the population1. Further, the number of new infections has been increasing for nearly two decades (WHO 2001; WHO 2012). The increasing global prevalence of Tvag infections can be attributed to the recurrent nature of the disease (Fichorova 2009), its high prevalence in low resource areas, low sensitivity detection methods (Schwebke 2004), and the asymptomatic outcome of infection (Singh 2007). Around 50% of Tvag-infected women are asymptomatic. Around 70% of Tvag-infected men are asymptomatic. If the infection is left untreated, the infection may become chronic without any lasting immunity (Singh 2007).

Tvag is an obligate extracellular parasite that colonizes the mucosal epithelium of the human genitourinary tract. Given that the human host is the only natural environment for the

parasite, much about the mechanisms of pathogenesis remains unknown and has been limited to studies in *in vitro* models using cultured cervical or vaginal cells (Singh 2007). The first stage of infection is adhesion to the epithelial cells of the genitourinary tract. Four adhesion proteins, AP65, AP51, AP33, and AP23 (Alderete 1988), several cysteine proteinases (CyP) (Arroyo 1989; Mendoza-Lopez 2000), and the Tvag lipophosphoglycan (LPG) (Fichorova 2006) have been implicated in mediating Tvag adhesion to vaginal epithelial cells. After attachment, the trichomonads cause cytotoxicity towards cultured cervical and vaginal cells in a contact-dependent manner. The contact-dependent cytotoxicity seems to be dependent on parasite density and specific to host cell types. Trichomonads collected from either asymptomatic or symptomatic patients are capable of the cytopathic effects (Singh 2007), suggesting that the symptoms experienced by different patients may result from different protozoan–human interactions. One mechanism of cytotoxicity is induction of host cell apoptosis by the Tvag cysteine proteinase 30 (CyP30). This induction requires the protease activity of CyP30 and is species specific (Sommer 2005). Detection of Tvag through LPG detection produces an inflammatory response (Shaio 1995) and stimulates IL-8 production in vaginal epithelial cells (Fichorova 2006). IL-8 stimulates the production of HIV in host cells (Narimatsu 2005) and acts as chemokine for additional HIV host cells (Singh 2007). This may explain the increased susceptibility to acquiring HIV infection in patients with trichomoniasis. Tvag has to survive the host immune system. The human body is able to directly clear the trichomonads through antibody-mediated lysis (Alderete 1986a) as well as antibody-independent clearing through the alternative complement pathway (Gillin 1981). Tvag cells have been demonstrated to alter the surface expression of the highly immunogenic surface protein P270 (Alderete 1986b) and to produce CyPs that are capable of degrading immunoglobulins (Provenzano 1995). Evasion of the alternative complement pathway is also possible through action

of CyPs since CyPs have been shown capable of degrading C3 of the complement system (Alderete 1995). Though such various players in Tvag pathogenicity have been identified, how these players are regulated and controlled during infection remains poorly understood. Notably, work dating back nearly three decades may further suggest yet another participant in the pathogenicity of trichomoniasis, as described in the next part of this chapter.

6.1.2 Trichomonas vaginalis virus (TVV) and trichomoniasis

In 1986, the first virus infecting a protozoan was discovered in Tvag (Wang 1991). This virus was named Trichomonas vaginalis virus (TVV) and later given its own genus, *Trichomonasvirus*, within the virus family *Totiviridae* (Goodman 2011b). Like most members of family *Totiviridae*, TVV lacks the capability of extracellular transmission. Transmission of TVV is thought instead to occur only vertically. Despite the lack of extracellular transmission, at least half of all Tvag clinical isolates harbor TVV in the USA (Snipes 2000) with the frequency as high as 82% from some geographical locations (Weber 2003). Furthermore, our lab has shown that Tvag isolates can harbor up to four different species of TVVs, named TVV1, TVV2, TVV3, and TVV4 (Goodman 2011a). Given TVV's lack of capability for extracellular transmission, it is interesting to ponder the selective pressures maintaining TVV in the Tvag population.

The presence of TVV in Tvag is indeed associated with certain Tvag phenotypes. The expression of P270 alternates between being expressed on the surface and not being expressed on the surface even when the trichomonads are grown in axenic conditions (Alderte 1986b). This phenotype is displayed only in trichomonads infected with TVV, and loss of TVV in TVV-infected trichomonads results in the loss of the P270 phenotypic variation (Wang 1987). TVV-uninfected trichomonads never express P270 and do not have detectable levels of P270 mRNA in a northern

blot (Khoshnan 1994). Expression of P270 shows a drastic reduction in both host cell adhesion and contact-dependent cytotoxicity towards HeLa cells (Alderete 1986b). Further, the surface expression of P270 alternates with the surface expression of the four AP proteins (Alderete 1988), suggesting that TVV may regulate at least two different classes of surface proteins. Given that contact-dependent cytotoxicity is a function of parasite density, the implication of this finding is that the presence of TVV promotes survival of the trichomonads in the human body by regulating damage to the epithelial lining. Furthermore, TVV-infected trichomonads express different CyPs as well as other proteins compared to TVV-uninfected trichomonads of the same isogenic strain (Provenzano 1997). Recently, our lab has newly shown that the presence of TVV inside Tvag cells also influences the disease outcome of trichomoniasis using culture endocervical cells as a model (Fichorova 2012). The results of this study may further explain the variability in the disease outcome of trichomoniasis as well as provide a molecular explanation for the increased risk of preterm birth in pregnant women with trichomoniasis treated with metronidazole (Kigozi 2003; Klebanoff 2001). The leading hypothesis for the cause of preterm birth states that infection triggers cytokine release that induces the onset of labor (Mitchell 2991). The best cytokine predictor of preterm birth is interleukin-6 (IL-6) (El-Bastawissi 2000) and is a common virus-induced gene product (Sen 2005). Our lab has shown that detection of TVV genomic dsRNA from TVV-infected trichomonads indeed induces IL-6 production in endocervical epithelial cells through TLR-3. Exposure of cells to purified TVV virions, in the absence of Tvag cells, also stimulates the production of IL-6. TVV-uninfected trichomonads do not stimulate IL-6 production to comparably high levels. Further, cell-free supernatant from metronidazole treated TVV-infected trichomonads and not from metronidazole treated TVV-uninfected trichomonads amplified the production of IL-6 (Fichorova 2012).

The influence of protozoan viruses on the disease outcome of other mammalian parasites is also seen in leishmaniasis. Using a mouse model, Ives et al. showed that the presence of the Leishmania RNA virus 1 in *Leishmania guyanensis* increased footpad swelling of the mouse and promoted the parasite's persistence (Ives 2011). Zangger et al. also showed that the presence of Leishmania RNA virus-aethiopica in Leishmania aethiopica promotes an inflammatory response in murine macrophages (Zangger 2014). Many other mammalian parasitic protozoans have also been found to harbor protozoal viruses. A few of these protozoans are *Giardia lamblia*, *Entamoeba histolytica*, *Eimeria tenella*, and *Cryptosporidium parvum* (Banik 2014). Further, studying these viruses could reveal novel mechanisms in understanding cell host translational machinery. Work done on the Giardia lamblia virus (GLV) has provided such insight (Yu 1998).

The studies described in the preceding chapter sections highlight the importance of TVV in Tvag pathogenicity. Despite being a treatable disease, Tvag infection remains a serious human health problem given its high prevalence, its serious disease outcomes, and our current lack of understanding in trichomoniasis. The regulation of certain Tvag host proteins involved in infection is currently unknown. This regulation may be dependent on the presence of TVV. Despite this speculation, it is clear that TVV is very likely directly involved in the pathogenicity of Tvag infection. TVV may aid in the survival of Tvag as well as contribute directly to the pathogenicity of trichomoniasis. Given the importance of TVV in Tvag infection, it is necessary to complete our understanding of the basic molecular biology of TVV. Identifying novelties in the basic biology of TVV will allow for direct applications in improving treatment of trichomoniasis and will contribute to future investigations into the molecular mechanism of TVV virus-host interactions. Such studies are furthermore expected to generate new approaches and reagents for genetic studies of TVV and perhaps Tvag as well. Given these justifications, this work is focused on increasing

understanding of the basic molecular biology of TVV replication and TVV interactions with its Tvag host cells. The work described in this chapter will focus on using our understanding of *Totivirus* replication to create a genetics system in Tvag that will allow for these studies to be accomplished.

6.1.3 TVV genome and basic biology

Much of what is known about TVV comes from analysis of its full-length genome sequence. As mentioned in the previous chapter, the genome organization and genome size of TVV is similar to other members within *Totiviridae*. TVV is a mono-segmented double-stranded RNA (dsRNA) virus within genus *Trichomonasvirus*. There are currently four identified species of TVV, named TVV1, TVV2, TVV3, and TVV4. The genome length ranges between 4680 and 4940 nucleotides (nts), depending on the species (Goodman 11a). The plus-strand contains two open reading frames (ORFs). The 5' ORF encodes a single capsid protein (CP) and the 3' ORF encodes for an RNA-dependent RNA polymerase (RdRp) that overlaps the CP ORF in a different reading frame and is expressed as a Cp/RdRp fusion protein. Flanking the two ORFs are untranslated regions (UTRs). The length of the 5' UTR ranges between 295 and 363 nts, depending on the species, while the length of the 3' UTR ranges between 63 and 161 nts (Goodman 11b).

The most recent contribution towards characterizing TVV came in 2013. Our lab solved the three-dimensional structure of the TVV1 virion (Parent 2013). Additionally, our lab has provided fourteen of the twenty-one currently available TVV full-length genome sequences in the GenBank database. All the full-length TVV genomes sequenced by our lab begins with 5'-GC that is capable of folding into a stem loop with the immediate downstream 24 nts and ends with UC-3'41. Before our contribution to the Genbank database in 2011 there were only five available TVV

full-length genomes. We have also identified the newly ratified species TVV4 (Goodman 2011a), as well as propose that TVVs should be given its own genus, *Trichomonasvirus*. TVV RNA transcription is currently presumed to follow that of other members within *Totiviridae* (Goodman 2011b), however no empirical studies exist to address this assumption.

6.1.4 *Totiviridae* replication

Virus replication among members of *Totiviridae* is best characterized in the L-A virus, which infects *Saccharomyces cerevisiae* and shares the same genome organization as TVV. Replication begins with a mature virion, which consists of the dsRNA genome encapsidated by the CP and CP/RdRp fusion protein. In the first stage of replication, the plus strand is made and extruded out of the virion and into the host's cytoplasm. The plus strand is then translated by host translational machinery to generate CP and CP/RdRp proteins. The plus strand is then recognized by the newly translated CP/RdRp for encapsidation by CP and Cp/Rdp. After the plus strand is encapsidated, the plus strand serves as a template for RdRp activity and the minus strand is produced to generate a new mature virion (Wickner 2013) (Fig. 25).

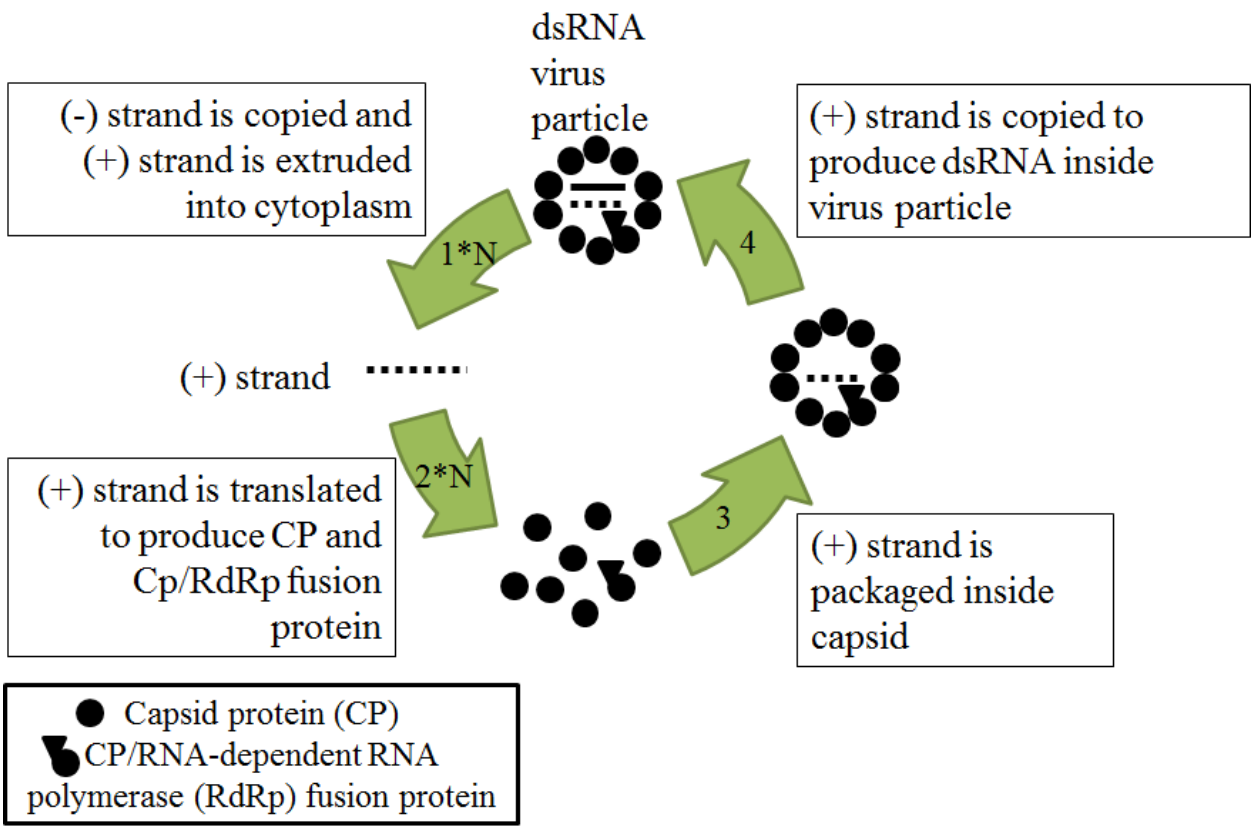


Figure 25 Virus replication of *Totiviridae* based on studies from L-A virus.

Using an *in vitro* system in which empty L-A virus particles bind and transcribe RNA when exposed to L-A plus-strand transcripts, work led by Reed Wickner mapped the regions within the L-A plus strand involved in transcription (Wickner 1989). There are three sites and each are located at the 3' end of the L-A plus strand. Two of these sites are found within the last 30 nts of the 3' terminal end and consist of (i) the very terminal sequence "AUGC" and (ii) a 25-nt long stem loop located immediately upstream of "AUGC". Eliminating either of these two sites reduces transcription to <5% (Wickner 1989). The third site largely overlaps with a site required for packaging the plus strand into virus particles and is located 400 nts upstream of the 3' end and located within the RdRp ORF. This third site is required for binding but is not essential for transcription. However, this site acts synergistically with the other two sites to enhance

transcription of the template. This third site spans 40 nts and contains a 22-nt long stem loop (Esteban 1989). Cloning these three sites into a heterologous RNA transcript allows for encapsidation of the transcript by empty L-A virus particles (Fujimura 1990). Further, the domains essential to recognition of these sites within the L-A RdRp have been mapped (Ribas 1994) and are specific to the L-A virus genome (Fujimura 1989). Empty L-A virus particles do not bind nor synthesize RNA from the plus strand of the L-BC dsRNA yeast virus (Fujimura 1989), which can be found coinfecting L-A virus-infected yeast cells. Thus, there is specificity between these three sites on the L-A plus strand and the L-A RdRp.

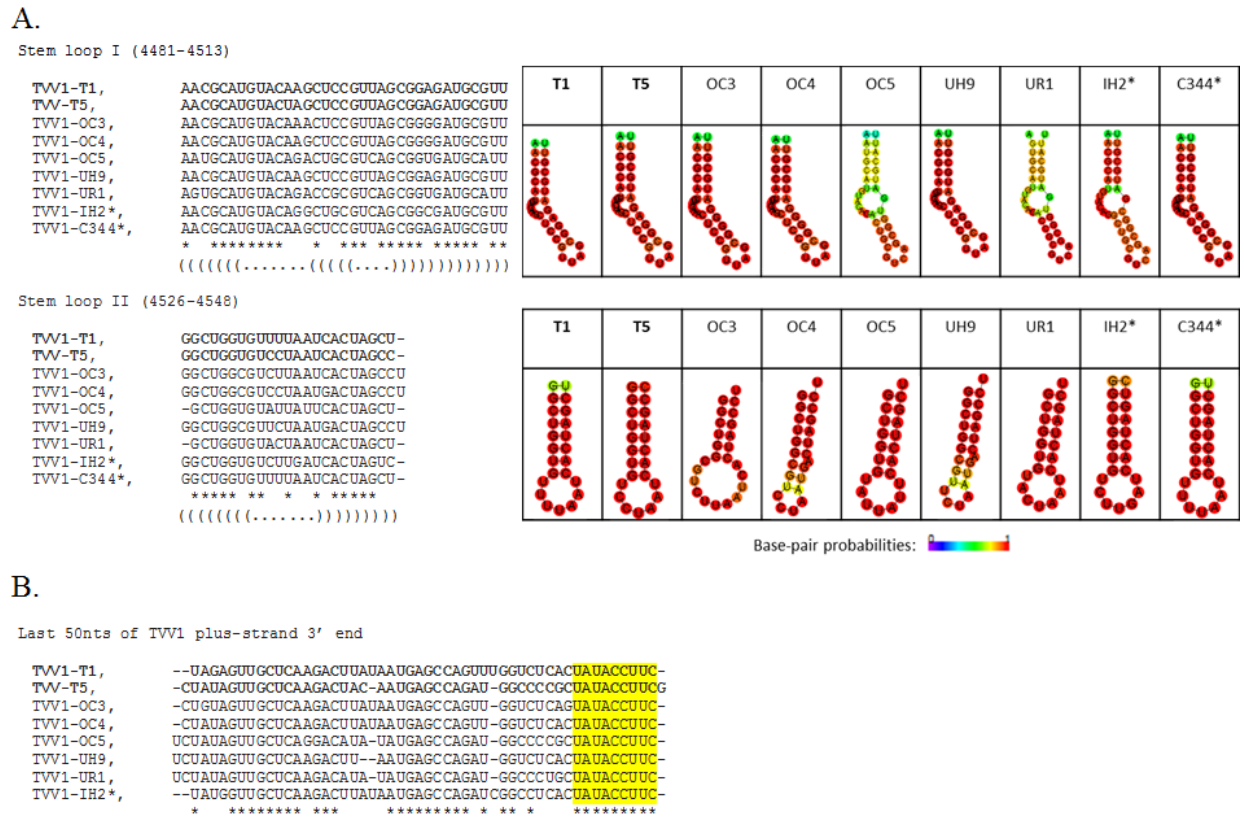


Figure 26 Identification of notable conserved structures and terminal sequences of TVV plus strand. (A) Multiple sequence alignment of TVV1 plus strand from available sequences from strains identified by us and others. Displayed is the region containing the previously reported two stem loops from Su et al. 1996. Also this region per each strain has been folded using RNAFOLD and displayed on the right side of the figure. (B) Multiple sequence alignment of TVV1 plus strand 3' terminal sequence from available strains identified by us and others.

The first prediction of conserved secondary structures in the TVV plus strand was made by a group led by J.H. Tai in 1996 (Su 1996). During this time, only two TVV strains, each from different clinical isolates of Tvag, were sequenced. Both of the strains are from species TVV1. Comparing the two sequences and using the program FOLDRNA, the group predicted two conserved stem loops located about 200nts upstream of the plus strand 3' end. Stem loop I is located between nts 4481 and 4513 and stem loop II is located between nts 4526 and 4548, thus both within the RdRp ORF. Despite only having a sample size of two, their predictions are found in all newly available sequenced TVV1 strains, within the same nt positions (Fig. 26A). These two regions from all nine currently available TVV1 sequenced strains are highly conserved and all are capable of forming a stem loop of a very similar structure. Further, mutations within the stem are often followed by another downstream complementary mutation to maintain base-pairing within the stem or prevent bulge formation. Our lab has identified a 9-nt sequence UAUACCUUC-3' that is present in the very 3' end of all our TVV1 plus strands (Fig. 26B). However not all TVV1 strains have the capability of forming a stem loop near their 3' plus-strand end, despite having highly conserved stretch of 50nts, suggesting that a 3' terminal stem loop may not be required for TVV transcription or may be located further upstream. It is plausible that the TVV1 plus-strand stem loops I and II predicted by J.H. Tai and our newly identified TVV1 plus-strand 3' terminal sequence UAUACCUUC-3' are all required for TVV1 virus replication. Developing a genetics system would allow us to test such a hypothesis, among many others as discussed in discussion.

6.1.5 A virus-mediated reverse genetics system in *Totiviridae*

The aforementioned studies of the L-A virus have mapped sites required for optimal transcription to terminal plus-strand ends and have shown that they are recognized by the L-A

RdRp. Additionally, if these sites are cloned into heterologous transcripts then the heterologous transcripts will become encapsidated by empty virus particles. Thus, it is conceivable to generate a reverse genetics system within virus-infected host cells that will use viral genomic elements, rather than cell host genetic elements, to replicate a reporter gene. The goal of this work is to develop such a system will offer a virus-specific reporter system to address virus genomic regions required for complete replication of TVV. Thus, I will develop a TVV-mediate reverse genetics system.

Such a concept is currently found in nature. The L-A virus as well as TVV can support the replication of dsRNAs satellite nucleic acids. The two most characterized satellites of L-A are called X and M and both contain the three sites required for L-A plus strand packaging and transcription. Additionally, both are encapsidated by the L-A virus and synthesized by the L-A RdRp (Wickner 2013). TVV also supports dsRNA satellites that are encapsidated by TVV virus particles and synthesized by TVV RdRp (see Chapter 5; Khoshnan 1995; Tai 1995). None of these satellite dsRNAs encode a CP nor replicate without the presence of their helper virus. Thus, their replication depends on being incorporated into the helper virus replication cycle.

Further, using the aforementioned studies on L-A virus replication, the group led by C. C. Wang established a virus-mediated genetics system using the Giardia lamblia virus (GLV) as a helper virus for propagation of an artificial transcript (Yu 1995). GLV is also a member of *Totiviridae*. The group electroporated a reporter transcript flanked by GLV plus strand 5' and 3' terminal regions into GLV-infected *G. lamblia* and GLV-uninfected *G. lamblia*. Both replication and propagation of their construct were detected only when this transcript was electroporated into GLV-infected *G. lamblia*. Truncated transcripts lacking either GLV plus strand 5' or 3' terminal region were neither replicated nor propagated when electroporated into GLV-infected *G. lamblia*.

Thus, replication of their transcript requires a helper-virus from a GLV-infected host and flanking GLV plus strand 5' and 3' terminal regions on the transcript, suggesting that replication and propagation of their transcript is GLV-dependent and, thus, dependent on transfection of a GLV-infected host. Further, the group showed that their construct was packaged in virus-like particles in the GLV-infected host (Yu 1996), suggesting that their transcript was replicated and propagated in parallel with GLV virions. This system has allowed the group to identify unique regions of the GLV genome involved in GLV replication, such as a novel IRES that extends into both sides of the start codon (Garlapati 2004).

6.1.6 Tvag and transfection systems

There are two labs that have reported establishing a transfection system in Tvag. Both labs used electroporation for their transfection but reported very different electroporation conditions, different methods, and used different Tvag strains. The first lab to publish on setting up a transfection system in Tvag is led by Patricia Johnson (Delgadillo 1997). The Johnson lab generated a transient transfection system through electroporation of plasmids containing native Tvag promoters to express exogenous genes. The second lab to report a transfection system in Tvag is led by Xichen Zhang and the lab reported electroporating transcripts flanked by TVV plus-strand 5' and 3' UTRs (Li 2012). The Zhang lab, however, did not determine the full-length genome of their TVV strain and their reported primers used for generating their construct would truncate the plus-strand 5' end by 35 nts, thus their construct did not include the aforementioned 5' stem loop found in all of our sequenced TVV genomes. They deposited their TVV sequence on GenBank (DQ528812), but it does not include the 5' and 3' UTRs. The Zhang lab has not published using their TVV system since their first report in 2012.

6.2 Materials and Methods

6.2.1 Tvag isolates

Tvag isolate UH9 was selected for initial experiments because it is infected by only one TVV species, TVV1-UH9 (Goodman 2011a), and, with this TVV isolate's three dimensional virion structure solved (Parent 2013), TVV1-UH9 is also the most well-characterized of known TVVs. It has also been used in studies involving human disease models of trichomoniasis (Fichorova 2012). Tvag UH9 was obtained from a symptomatic patient in upstate New York in 1999, first reported in 2000. The stock of Tvag UH9 used in this study originated from the same soft-agar clone described in our previous sequencing study (Goodman 2011a).

Tvag isolate PJ was also used in these experiments. Tvag PJ is not infected with any species of TVV.

All isolates were derived from a soft-agar clone and grown in liquid culture as previously described (Goodman 2011a). Trichomonads were cultured in modified Diamond's modified medium (Diamond 1957), tryptone 2%, yeast extract 1%, maltose 0.5 %, cysteine (free base) 0.1%, ascorbic acid 0.02%, KH_2PO_4 5.5 mM, K_2HPO_4 4.4 mM, pH 6.0, ferrous ammonium sulfate 178.5 nM, sulfosalycilic acid 32.3 nM, with 10% heat-inactivated horse serum (HyClone) at 37°C.

Trichmonads were counted using Bio-Rad's TC20™ Automated Cell Counter.

6.2.2 Transcript design and *in vitro* transcription

Transcripts consisted of a *neo* gene flanked by TVV1-UH9 plus-strand terminal 5' and 3' sequences. TVV1-UH9 (GenBank HQ607516) was used for generating flanking sequences. The 5' end of the transcript is flanked by 600nts of the 5' plus-strand sequence of the TVV1-UH9 genome. The 3' end of the transcript is flanked by 600nts of the 3' plus-strand sequence of the

TVV1-UH9 genome. This transcript will henceforth be called TVV-*neo* transcript for communication purposes.

An additional TVV-*neo* transcript was generated. This additional transcript lacks the 3' TVV plus-strand sequence, but is otherwise identical to TVV-*neo* transcript, and will henceforth be referred to as TVV3'-truncated-*neo* transcript for communication purposes.

The T7 RNA polymerase was a kind gift from Dr. Victoria D'Souza, for who I am very grateful for; without her help and her opening of her lab much progress would not have been possible. The i*n vitro* T7 transcription was optimized with help and advice from various members of Dr. D'Souza's lab, Carolina Salguero, Vincent Pham, Nico Wagner. Transcripts were synthesized fresh by *in vitro* transcription using T7 RNA polymerase in 200µl reaction/transfection: 0.09mg/ml of template, 40 mM $MgCl_2$, 2 mM spermidine, 80 mM Tris-HCl (pH 8.1), 2 mM of each NTP, 2 mM DTT, and 0.07 U/µl T7 RNA polymerase, incubated at 37°C for 2 hr on rotation. The RNA was purified by TRIzol™ (Invitrogen) according to manufacturer's instructions and resuspend in DEPC-water before use in transfection. The concentration of each sample was determined by measuring the optical absorbance at 260 nm. Visualization of transcripts when necessary was achieved using a denaturing MOPS gel and stained with ethidium bromide.

6.2.3 Plasmid

The transfection plasmid pMASTER-neo was a kind gift given to us by Dr. Patricia Johnson's lab. It is an updated transfection plasmid from their originally published plasmid (Delgadillo 1997). For selectable transfection, this plasmid has a *neo* gene expressed through a native Tvag β-tubulin promoter.

6.2.4 Transfection and selection

Transfection of Tvag was based on the electroporation protocol originally published from Dr. Patricia Johnson's lab (Delgadillo 1997) with modification based on correspondence with Dr. Brian Janssen, a postdoctoral scientist in Dr. Johnson's lab who I am very grateful for. For each transfection, 50 ml of Tvag (~3 x 10^8 cells) was pelleted from stationary phase by centrifugation at 1,500rpm at 4°C for 15min. Supernatant was poured off and residual volume of supernatant was left for re-suspension of pellet. Add directly to the pellet 100µl of chilled nucleic acid (100µg of pMASTER-neo or 100µg of transcript) and gently mix and re-suspend pellet by slowly pipetting up and down with wide mouth pipet tip. Volume comes to a bit over 300µl. Transfer 300µl of trichomonad and nucleic acid mixture to Bio-Rad 0.4 cm electrocuvette pre-chilled on ice. Electroporate using 350 volts, 950 µFd, and resistance set to ∞ in 0.4 cm electrocuvette using a Bio-Rad gene pulser and capacitance extender. The voltage transfer time should come out to ~100ms. Transfer trichomonads to 50ml of fresh Diamond's media pre-warmed to 37°C and let recover for 24hr. Begin four rounds of enrichment (one round of enrich: remove cell debris at bottom of culture, harvest live free-swimming trichomonads by centrifugation at 1,500rpm at 4°C for 15min, re-suspend trichomonads in fresh 50ml Diamond's media with 100µg/ml of Geneticin (G418), let recover for 24hr). Use 50µg/ml of Geneticin for enriching transfected trichomonads with transcripts, rather than the 100µg/ml of Geneticin used for plasmid transfection. After fourth round of enrichment, start passaging cells into 200µg/ml of Geneticin for selection. Use 100µg/ml of Geneticin for selecting transcript-transfected trichomonads, rather than the 200µg/ml of Geneticin used for plasmid transfection.

6.2.5 Assay for minus-strand production

Three ml of transfectants were harvested by centrifugation at 1,500rpm at 4°C for 15min. RNA was extracted from the cell pellet using TRIzol™ (Invitrogen) according to manufacturer's instructions. The RNA was then reverse transcribed using SuperScript III First-Strand Synthesis System (Invitrogen) in the presence of specific primers (*neo* plus-strand specific primer or *neo* minus-strand specific primer) according to manufacturer's instructions. The RT product was then used for separate PCR reactions with both forward and reverse *neo* gene-specific primers per reaction using Taq DNA Polymerase with standard Taq Buffer (NEB) according to manufacturer's instructions. PCR products were separated by agarose gel electrophoresis and stained with ethidium bromide.

6.2.6 Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE)

I am very grateful to have had the help of Dr. Kate Godin, postdoctoral scientist in Dr. Victoria D'Souza, who helped me with my SHAPE analysis.

TVV-*neo* transcript was used as designed as described in Materials and Methods 6.2.2 without any further modifications, such as bridges or spacers, and was generated fresh through *in vitro* T7 transcription as described in Materials and Methods 6.2.2. TVV-*neo* transcript was purified using TRIzol™ (Invitrogen) according to manufacturer's instructions before SHAPE analysis. SHAPE was performed according to previously published protocol (Wilkinson 2006). Two pmol of transcript was denatured then folded in folding solution (100 mM HEPES, pH 8.0, 6 mM MgCl2, 100 mM NaCl) for 20min at 37˚C. Reaction was split in two. One of the two RNA folded reactions was modified with five N-methylisatoic anhydride (NMIA) half-lives according to "half-life$_{min}$=360 x exp(-0.102 x temperature(˚C)"), which resulted in 4min and 16sec for my

transcript. Transcripts were pelleted by ethanol precipitation then re-suspended for primer extension. Primer extension was performed using SuperScript III First-Strand Synthesis System (Invitrogen) in the presence of TVV1-UH9 3'-specific primers conjugated with a fluorescent dye (NED, Thermo Scientific) according to manufacturer's instructions. For sequencing reactions, non-NMIA modified transcript was used for primer extension in the presence of ddATP and TVV1-UH9 3'-specific primers conjugated with a different fluorescent dye (VIC, Thermo Scientific). Transcripts were then pelleted with ethanol precipitation and re-suspended for fragment analysis using ABI 3730xl DNA Analyzer with HiDi solution. Capillary electrophoretic data was then analyzed using software QuShape (Karabiber 2013). Analyzed SHAPE data was incorporated into software RNAstructure for structure prediction (Mathews 2004).

6.3 Results

6.3.1 Transient geneticin-resistant Tvag UH9 after transfection with pMASTER-neo

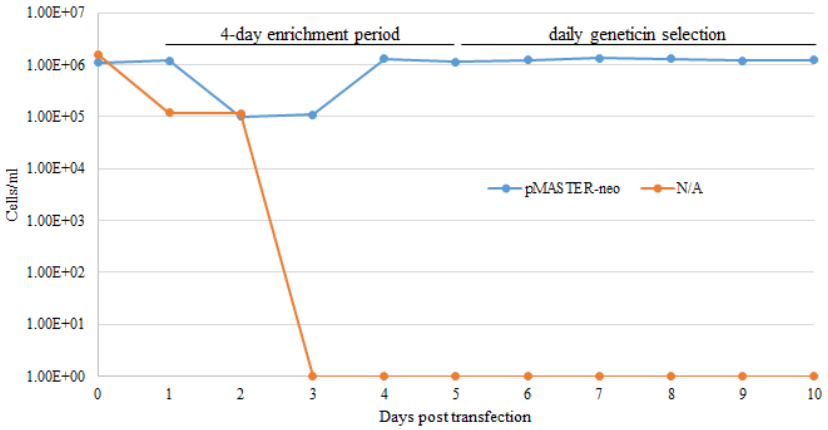


Figure 27 Geneticin resistance in pMASTER-neo-transfected Tvag UH9. Transfection protocol is described in Methods and Material 6.2.4. The four-day enrichment period consisted of treatments of 100µg/ml of Geneticin per day. Daily selection consisted of 200µg/ml of Geneticin per day.

Initial experiments focused on optimizing transfection conditions and selection of transfectants. After about a year of troubleshooting various initial cell counts, transfection buffers, concentrations of genetic material, voltages, capacitances, charge deliveries, recover conditions, and Geneticin concentrations (data not shown) I finally found the condition that worked in our lab with the help of Dr. Brian Janssen. The first successful transfection was generated according to the protocol and conditions described in Materials and Methods 6.2.4 using pMASTER-neo. This protocol will represent all transfection or electroporation conditions hereafter. Tvag UH9 transfected with pMASTER-neo survived beyond the second exposure of Geneticin whereas Tvag UH9 electroporated in the absence pMASTER-neo died after the second exposure of Geneticin (Fig. 27). The mRNA of the *neo* gene was also detected in RNA extracts from pMASTER-neo transfected Tvag UH9 but not Tvag UH9 electroporated in the absence of pMASTER-neo (Fig. 28). Further, the geneticin-resistant trichomonads survived up to 21 days before they died in the presence of Geneticin (data not shown). Thus, I am able to show transient transfection and selection of Tvag.

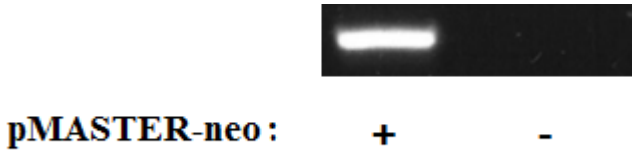


Figure 28 Transcription of pMASTER-neo after transfection into Tvag UH9. Total RNA was harvested and then assayed for *neo* using RT-PCR with *neo*-gene specific primers.

After identifying successful transfection and selection conditions, the next goal was to reproduce geneticin-resistant Tvag UH9 using the TVV-*neo* transcripts described from Materials and Methods 6.2.2.

6.3.2 Transient geneticin-resistant Tvag UH9 after transfection with TVV-*neo* transcript

Initial attempts of generating geneticin-resistant Tvag UH9 with TVV-*neo* transcript were unsuccessful, possibly due to the both quality and lack of sufficient quantity. With the help of Dr. Victoria D'Souza and her lab members, I was able to optimize a personal *in vitro* T7 RNA transcription using their home-made T7 RNA polymerase (Materials and Method 6.2.2; Fig. 29)

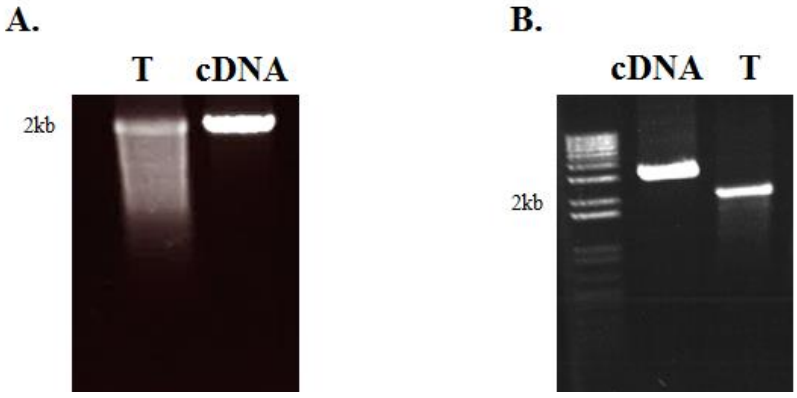


Figure 29 *In vitro* transcription (IVT) product on denaturing MOPS gel. (A) IVT product prior to optimization and help from Dr. D'Souza's lab. (B) IVT product after optimization and help from Dr. D'Souza's lab. New cDNA was also made and includes additional 1kb of plasmid sequence upstream of T7 promoter.

that successfully produced geneticin-resistant Tvag UH9. However, these experiments involving TVV-*neo* transcript also required that I used half the Geneticin concentration I was using previously while working with pMASTER-neo, but this new genetic concentration was still potent and still able to kill untransfected Tvag. TVV-infected Tvag UH9 transfected with TVV-*neo* transcript survived beyond the second exposure of Geneticin whereas Tvag UH9 electroporated in the absence of TVV-*neo* transcript died after the second exposure of Geneticin (Fig. 30). Similar to transfection with pMASTER-neo, Tvag UH9 transfected with TVV-*neo* transcript survived up to 21 days post-transfection before they died in the presence of Geneticin (data not shown). Thus, transient geneticin-resistance after transfection of Tvag with TVV-*neo* transcript requires the presence of TVV.

After successful transient transfection and selection of geneticin-resistant Tvag UH9 with TVV-*neo* transcript, the next goal was to establish an assay for mapping sites required for TVV minus-strand production.

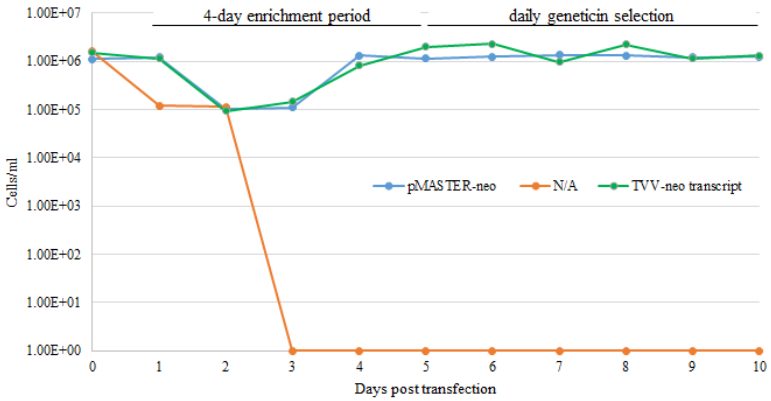


Figure 30 Geneticin resistance in TVV-*neo*-transcript -transfected Tvag UH9. Transfection protocol is described in Methods and Material 6.2.4. The four-day enrichment period consisted of treatments of 50µg/ml of Geneticin per day. Daily selection consisted of 100µg/ml of Geneticin per day.

6.3.3 Minus-strand production of TVV-*neo* transcript in TVV-infected Tvag requires presence of TVV 3' plus-strand sequence

In virus replication of *Totiviridae*, after the TVV plus strand is translated to produce CP and CP/RdRp then it will be packaged and then used for minus-strand production (Fig. 25) Given the Geneticin-resistance phenotype of the TVV-*neo* transcript transfectants, we can safely assume that the TVV-*neo* transcript has entered the TVV-infected Tvag UH9 and contains the sites for translation initiation by Tvag. We next want to assay for whether the minus-strand of the TVV-*neo* transcript was produced after transfection.

Tvag UH9 was transfected with TVV-*neo* transcript. Separately, in parallel, a fresh aliquot of Tvag UH9 was transfected with TVV3'-truncated-*neo* transcript. Both transfectants were then selected for Geneticin and interesting both survived in the presence of genetic for same duration

(data not shown). After the four rounds of enrichment and two additional days of Geneticin selection (one week post-transfection), both transfectants were harvested and assayed for minus-strand production according to Materials and Methods 6.2.5.

The plus-strand copy of the *neo* gene was detected in RNA extracts from both transfections. However, a minus-strand copy of the *neo* gene was detected only in RNA extract from the transfection with TVV-*neo* transcript. Notably, the band of the plus strand is significantly more intense than the band of the minus strand, consistent with the expected relative amounts of TVV plus strand and genomic dsRNA levels in the natural replication cycle of *Totiviridae*. RNA extract from the transfection with TVV3'-truncated-*neo* transcript did not show any detectable levels of a minus-strand copy of the *neo* gene (Fig. 31) Thus, minus-strand production of the TVV-*neo* transcript requires the last 600nts of the 3' TVV plus strand.

After providing evidence that the TVV 3' terminal end is indeed necessary for minus-strand production of the transcript and narrowing down this region to the last 600nts, we next wanted to show that the minus-strand production was indeed dependent on the presence of TVV infection.

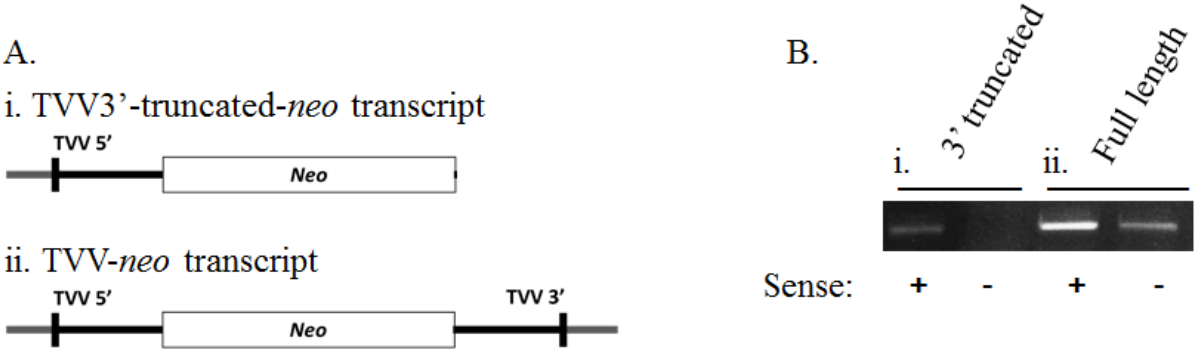


Figure 31 Sense-strand specific RT-PCR shows that TVV plus strand replication requires TVV 3' terminal sequence.

6.3.4 Minus-strand production of transcript is dependent on TVV in Tvag

Tvag isolate UH9 and Tvag isolate PJ were independently transfected with TVV-*neo* transcript, enriched for geneticin-resistance, and selected for geneticin-resistance for two days. Tvag PJ does not obtain any TVV speices. RNA extract was then harvested from both transfectants and assayed for minus-strand production. One week post-transfection, RNA extract from Tvag UH9 transfected with TVV-*neo* transcript had both copies of the plus strand and minus strand of the *neo* gene. However, RNA extract from Tvag PJ transfected with TVV-*neo* transcript showed only a copy of the plus strand of the *neo* gene (Fig. 32). Thus, minus-strand production of the TVV-*neo* transcript also requires the presence of TVV.

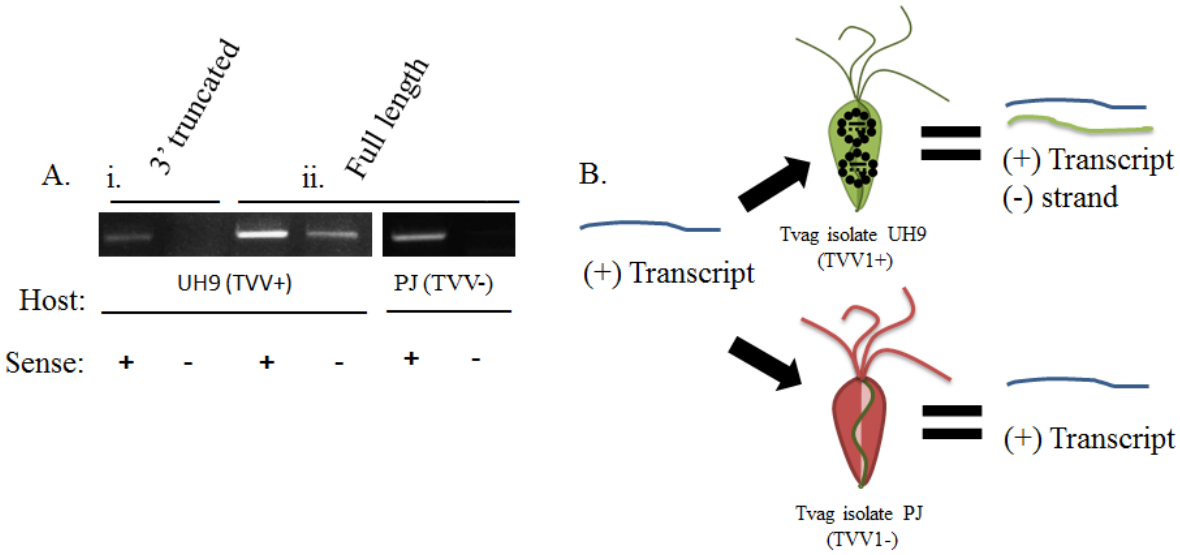


Figure 32 Sense-strand specific RT-PCR shows that TVV plus strand replication requires presence of TVV. (A) RT-PCR results. (B) Schematic of experiment and outcome.

Now having shown that (i) we have a successful assay for minus-strand production, (ii) minus-strand production of TVV-*neo* transcript is dependent on the last 600nts of the TVV 3' plus strand, (iii) and dependent on host infection by TVV, we wanted to start mapping specific sites on the TVV plus strand required for TVV minus-strand production. This will involve narrowing down

the TVV 3' 600nt sequence of our TVV-*neo* transcript to determine internal and terminal *cis*-acting sequences and structures required for TVV RdRp binding and transcription activity. Further, we also wanted to extend the TVV plus-strand sequences to include more of TVV plus-strand sequences from both directions. Doing so may create a stable TVV-mediated reverse genetics system through possible additional sites that increase translation or packaging and transcriptional activity, such as those identified in the L-A virus and GLV (Esteban 1989; Garlapati 2004).

6.3.5 SHAPE analysis identifies one of the two stem loops previously hypothesized by Su and Tai to be involved in TVV genome replication

For mapping structures of the TVV plus strand required for transcription, I am again very grateful for Dr. Victoria D'Souza for offering her lab to help me again and for Dr. Kate Godin to personally working alongside me on this project. Working directly with Dr. Godin, a postdoctoral scientist in Dr. D'Souza's lab, I analyzed the TVV 3' 600nt plus-strand region for secondary structures using selective 29-hydroxyl acylation analyzed by primer extension (SHAPE).

We have clean, strong SHAPE reactivity scores for the 3'-most 241nts of the TVV1-UH9 terminal sequence (Fig. 33A & 33B). Thus, we are able to resolve the possible secondary structures for the 3'-most 241nts of the TVV1-UH9 terminal sequence (Fig. 33C). This depicted structure represents the average fold of 230 possible best-fitted structures to my SHAPE data, each possible fold has a similar minimum free energy within -80 to -81 kcal/mol and differ only by positioning or rotation of one nucleotide to no more than ten possible base-paired potentials (data not shown). This fold depicts this 241-nt genome region alone, without regard to neighboring flanking regions, thus the edges may fold differently than depicted here. Further, this structure was folded *in vitro* and differ from the fold found *in vivo*. Confirmation of the fold is discussed below (Section 6.3.5).

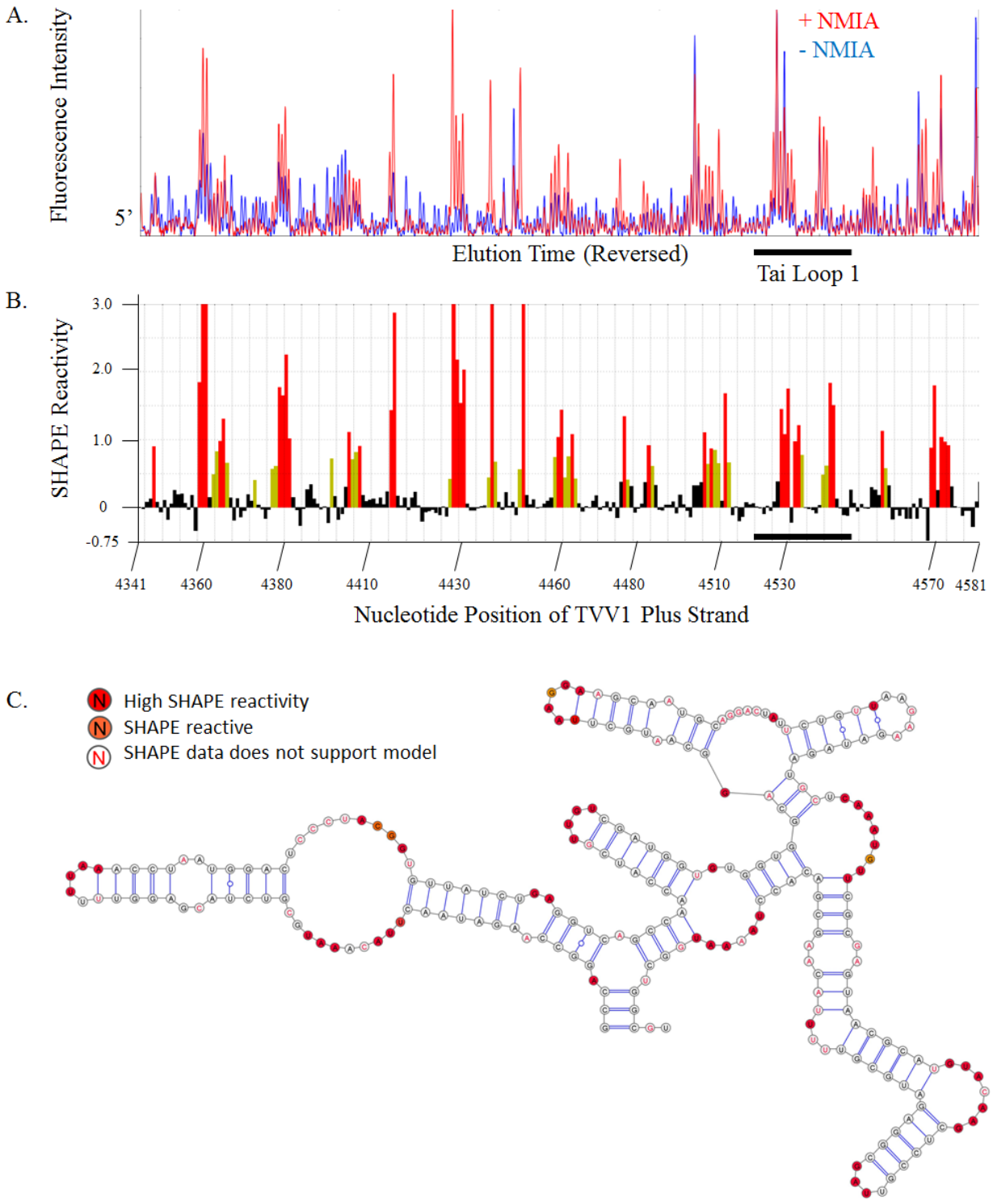


Figure 33 SHAPE analysis of 3'-most 241nts of the TVV1-UH9 terminal sequence. (A) Capillary electrophoretic data showing (+) and (−) NMIA. (B) Histogram of integrated and normalized SHAPE reactivities as a function of nucleotide position. (C) Predicted RNA secondary structure based on SHAPE analysis of 3'-most 241nts of the TVV1-UH9 terminal sequence.

Notably, this region contains the location of one of the two stem loops originally hypothesized by Su and Tai in 1996 to be involved in TVV genome replication. The second stem loop lies just outside of the 241nt region where I have clean resolution and will require a different SHAPE reaction. Our SHAPE analysis confirms with nucleotide accuracy the presence of a SHAPE reactivity pattern indicative of their stem loop that does fall within the resolved region. The stem loop fold was predicted by the software RNAstructure using the raw electropherogram data of my SHAPE reaction (Fig. 34).

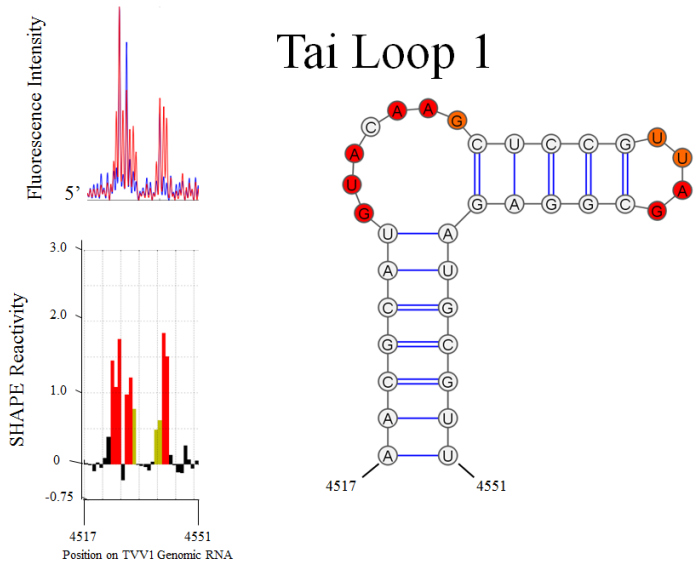


Figure 34 SHAPE analysis supports stem loop predicted by Su and Tai in 1996.

Further SHAPE analysis will need to be set up to resolve the rest of the TVV 3' plus-strand region. To test for possible function of these predicted stem loops in TVV transcription, mutant TVV-*neo* transcripts to abolish the predicted stem loops will be made and used in transfection and minus-strand production assays. SHAPE analysis on the mutant TVV-*neo* transcripts can confirm whether the stem loops have been abolished. The minus-strand production assay may have to use qRT-PCR if TVV transcription relies on the incremental effects of several sites and not dependent on having all sites present.

6.4 Conclusion

Despite trichomoniasis being a treatable disease, Tvag infection remains a serious human health problem and the importance of TVV in Tvag pathogenicity has long been accumulating. The goal of this work is to increase our understanding of the basic molecular biology of TVV through application of *Totividae* genome replication and inferences from analyzing their genome sequences.

We have developed a transient and selectable TVV-mediate reverse genetics system. We have also shown that this system can be used to assay for TVV minus-strand production. TVV minus-strand production requires both the TVV 3' 600nt sequence and the presence of TVV.

The system and these results support future experiments mapping internal and terminal *cis*-acting sequences and structures of the TVV plus strand required for TVV minus-strand production. Towards these studies, the current technical barrier seems to be the ability to reproducibly produce large amounts of high quality transcripts. Given that we have shown that the presence of TVV is required for minus-strand production, this opens up further investigations to map regions of the TVV RdRp that recognizes and binds and copies the plus strand.

6.5 Discussion

6.5.1 Transient nature of current TVV-*neo* transcript system

Possible explanations for the transient nature of the current transcript design include inefficient translation, packaging, or minus-strand production.

Detection of a 5' cap has not been detected in the TVV genome (Goodman 2011b) and the TVV genome may possibly use an IRES. Optimal translation of the GLV plus strand involves an IRES that extends into both sides of the start codon. Deletion analysis showed that efficient internal

translation initiation of the GLV plus strand requires the 253 nts upstream of the start codon, extending into the 5'-UTR, and 264 nts downstream of the start codon, extending into the ORF1 (Garlapati 2004). Our TVV-*neo* transcript contains 600nts of the TVV 5' plus-strand sequence, which includes the 324 nts of the entire TVV 5' UTR and the first 276 nts of the TVV ORF1. It may be possible that TVV also has a unique IRES that extends even further into the ORF1 for optimal translation. Such a possibility supports extending the TVV 5' plus-strand sequence of TVV-*neo* transcript.

RdRp binding to the plus strand and encapsidation of the plus strand (packaging) precedes minus-strand replication. The L-A virus has two sites on its plus strand required for minus-strand production and a third site called the internal replication enhancer (IRE) that greatly enhances replication (Esteban 1989). Functionally, it was later shown that the third site is really a site involved in packaging, specifically recognition and binding of the Rdrp to the L-A virus plus strand (Fujimura 1990). The first two sites required for L-A virus minus-strand production are located within the 30-nt terminal region of its 3' end: (i) the very terminal sequence "AUGC" and (ii) a 25-nt long stem loop located immediately upstream of "AUGC". The third site is located 400nts upstream of its 3' terminal end, within the ORF2, and contains a 23-nt stem loop and an additional 20-nt upstream sequence (Esteban 1989). The possibility that sites further upstream of 600-nt TVV 3' end may enhance packaging or minus-strand production supports extending the TVV 3' plus-strand sequence of TVV-*neo* transcript.

6.5.2 Mapping sites required for transcription of genome per each TVV species

Concurrent stable infections of host cells with two distinct viruses are known to occur for members of the *Totiviridae*. Baker's yeast *Saccharomyces cerevisiae* can be concurrently infected

with both the totiviruses L-A virus and the L-BC virus. A pathogenic fungus of corn *Ustilago maydis* can be concurrently infected by toriviruses Ustilago maydis virus H1 and H2 (Ghabrial 2008). A pathogenic fungus of pine trees *Sphaeropsis sapinea* can be concurrently infected by victoriviruses Sphaeropsis sapinea RNA virus 1 and 2. No examples of heterologous encapsidation have been reported in these cases (Wickner 2005).

Experiments testing this idea of species-specific encapsidation and replication can be accomplished using this TVV-mediated genetics system. Mapping experiments can be repeated for different TVV species to identify any differences between sequences. Further, transcripts flanked by different TVV species plus-strand sequences can be electroporated into Tvag infected with or without the species of TVV for which the flanking sequences were derived to test for a specificity associated with the TVV RdRp between species.

6.5.3 Mapping sites required for unique -2 programmed ribosome frameshift

The TVV genome contains two ORFs. The 5' ORF encodes the CP and the 3' ORF encodes the RdRp. The two ORFs overlap and are expressed in different reading frames with the RdRp expressed as a CP/RdRp fusion protein (Parent 2013). Western blot analysis using an anti-CP antibody detected a major band consistent with the size of CP and a minor band consistent with a CP/RdRp fusion protein. Further, western blot analysis using an anti-RdRp antibody detected the same minor band but not the major band (Liu 1998). The downstream RdRp ORF overlaps the CP ORF by around 120nts and is in the -1 frame in TVV2, TVV3, and TVV4. GGGCCCC is the putative slippery site for TVV2 and GGGCCCU is the putative slippery site for both TVV3 and TVV4. However, the RdRp ORF of TVV1 overlaps the CP ORF by a short span of 14nts and in the -2 frame. Our lab has also identified the TVV1 junction peptide fragment using mass

spectrometry that is consistent with a -2 programmed ribosome frameshift (PRF) (Parent 2013). Also, the putative slippery site for TVV1 is CCCUUUU (Goodman 2011a).

Mutating the slippery site from UUUAAAC to UUUUUUC in the coronavirus infectious bronchitis virus introduced a -2 or +1 PRF into an otherwise strictly -1 PRF (Brierley 1992), suggesting that a slippery site containing more uracils may cause a larger frameshift. The mechanism for this recently discovered -2 frameshift remains unknown. This -2 frameshift is currently known to be used naturally by two other viruses. The dsDNA bacteriophage Mu also utilizes a -2 PRF but contains a different putative slippery site with no uracils (Xu 2004). Arteriviruses use a -2 PRF that involves the slippery site GGUUUUU as well as a CCCANCUCC motif located 11nts downstream of the slippery site (Fang 2012). TVV1 does not contain the arterivirus's downstream motif. The efficiency of the -2 frameshift as measured by the CP:CP/RdRp ratio in TVV1 is 10-fold less than that of arterivirus (Fang 2012). Investigating the nearby genome sequences involved in TVV1 -2 PRF will advance the understanding of frameshifting. Dicistronic transcripts can be used with this established transfection system to study the TVV -2 PRF, presumably in the absence or presence of TVV.

6.5.4 Developing a knockdown system

Different groups have reported using RNA interference (RNAi) in Tvag to knockdown various genes (Mundodi 2004; Ravee 2015). RNAi has been reported functionally present in some protozoans, such as *Leishmania braziliensis* (Lye 2010), but absent in others, such as *Trypanosoma cruzi* (DaRocha 2004). The whole genome sequence of Tvag has been determined and it supports the presence of some components of RNAi, such as a Dicer-like gene and two Argonaute genes (Carlton 2007). However, other RNAi components are absent, such as Drosha-like genes as well

as exportin5 analogs. In *Giardia lamblia*, microRNAs associated with Argonaute have been identified and shown to repress target mRNA expression (Li 2011). However, whether this occurs through cleavage of target mRNA is disputed (Li 2011; Prucca 2008). MicroRNAs have also been identified in Tvag (Lin 2009). Perhaps a non-canonical RNA-targeting system may also exist in Tvag.

Recent characterization of CRISPR and CRISPR-associated genes (CRISPR-Cas) identified the effector protein C2c2 as a single-stranded RNA (ssRNA) specific endoribonuclease. C2c2 lacks homology to known DNA nuclease domains and is able to provide interference against RNA phage for *Leptotrichia shahii*. In bacteria, CRISPR-C2c2 has been programmed to cleave specific mRNAs, suggesting that this system may be used as an RNA-targeting tool (Abudayyeh 2016).

Alternatively, using our current TVV-mediated reverse genetics system, we can replace the reporter gene with a ribozyme specifically targeting the second TVV species may produce a singly-infected host. The difficulty here may be in selecting for transfectants and the transient nature of the current system.

Developing a knockdown system for Tvag would be useful for solving lingering questions about basic biology of TVV. It would be much easier to obtain purified virions from single TVV species other than TVV1 for solving their virion structure. We have solved the three dimensional structure of TVV1, but other TVV species are found in concurrent infections with another TVV species making purification of single non-TVV1 species difficult. Also, in Chapter 5 of this dissertation I presented evidence for a possible species-specific association between helper virus and dsRNA satellites of TVV. A knockdown system would allow us to test this hypothesis. We expect that certain dsRNA satellites will be knocked down following the knockdown of its helper

virus. Further isogeneic strains of Tvag with and without infection of TVV would be very useful for studies showing the effects of TVV on trichomoniasis.

6.6 Current and Future Directions

6.6.1 Making transcripts for future studies

We had started making variants of TVV-*neo* transcripts that included extending the flanking TVV 5' and 3' terminal plus-strand sequences and 5'-> 3' and 3'->5' deletion mutants of the flanking regions. We started focusing on the deletion mutants first. However, all of the mutant transcripts showed detectable levels of the minus strand immediately from the *in vitro* T7 transcription as described in Materials and Methods 6.2.2 but showed no signs of cDNA carry-over. In contrast, freshly produced non-mutant TVV-*neo* transcripts showed the presence of only the plus strand product.

Transcription of the minus strand could result after the T7 polymerase reaches the end of the cDNA and "snaps back" onto the complementary cDNA. This results in further extending the transcript by synthesizing the minus strand as the extension product (Schenborn 1985). Commonly reported causes of this snap back include the presence of a 3' overhang (Schenborn 1985) or disruptive salt concentrations in the IVT reaction (Sambrook 1990). T7 RNA polymerase is highly sensitive to salt concentrations and its activity is reduced in high salt concentrations. The salt concentration can also cause misfolding of the transcript. Triana-Alonso et al. reported T7 polymerization using an RNA template as well as the propensity of snapping back if the transcript does not properly fold, particularly its 3' terminus (Triana-Alonso 1995). Differences in the salt concentrations of the IVT may be introduced by the cDNA. The cDNAs of my mutants were made in our lab, whereas the cDNAs of the original transcripts were made in Dr. D'Souza's lab. We

were using DEPC-water whereas Dr. D'Souza's lab was using commercial molecular-grade water. Possibly byproducts of the DEPC-treatment may affect the outcome of transcript products, though the exact mechanism is not known. DEPC breaks down to carbon dioxide, water, and ethanol. Ethanol can react with trace levels of caroxulic acid to produce esters, producing a sweet aroma characteristic of DEPC-treated water. This smell was detected in all my batches of DEPC-treated water, indicating possible trace levels of esters. While there is no certainty in which part of the DEPC treatment is affecting my *in vitro* transcription, I can be certain of removing this variable by simplying using the molecular-grade water from Victoria's lab.

Since I am using PCR products for my cDNA, rather than a linearized plasmid, the 3' overhang is not as much as a concern for my IVT as the salt concentrations. I have remade my cDNAs to contain 5' overhangs. Magnesium chloride concentration has a significant effect on IVT. Indeed, changing the [MgCl2] in my original batch of full-length transcripts results in minus-strand synthesis during IVT. Given this, I decided to re-optimize my IVT conditions for my new batch of deletion constructs. I have tested [MgCl2] ranging from 5mM up to 70mM, various [NTPs], and various [cDNA]. I also obtained the same molecular-grade water from Dr. D'Souza's lab for making my cDNAs. I also remade cDNAs to contain both the hepatitis delta virus ribozyme followed by a T7 terminator sequence (Pattnaik 1992). Upon more scanning of the literature, I also tried more unique approaches, such as reducing [UTP] to 0.7mM while keeping all other NTPs at 4mM (Triana-Alonso 1995). However, nothing worked. It seems that there is a small but non-zero amount of minus-strand copy produced during the *in vitro* T7 transcription, presumably through snap-back of the T7 polymerase.

During a committee meeting discussion, Dr. Craig Hunter and Dr. Victoria D'Souza mentioned that both of their labs always runs *in vitro* T7 transcription products on a gel then gel

excises the band. Doing so would isolate and purify the transcript of interest from any runoff products containing unwanted sequences.

6.6.2 Using the TVV-mediated reverse genetics system beyond minus-strand production

The results showing that minus-strand production of TVV-*neo* transcript is dependent on host infection by TVV provides experimental evidence supporting the encapsidation of TVV-*neo* transcript within TVV virus particles. Thus, one corollary that follows from this is that such virus particles should be identified. Additionally, the presence of the minus strand copy further suggests that there should also exist a dsRNA copy of the TVV-*neo* transcript.

Given that minus-strand production is thought to occur after the transcript is packaged into virus particles, another aspect of TVV replication that can be mapped to its plus strand sequence is TVV RdRp recognition and binding. This can be assayed with an *in vitro* gel retardation assay and would corroborate data from the TVV-mediated reverse genetics system or to tease out sites involved in specifically binding or minus-strand production.

6.7 References

Abudayyeh OO, Gootenberg JS, Konermann S, Joung J, Slaymaker IM, Cox DB, Shmakov S, Makarova KS, Semenova E, Minakhin L, Severinov K, Regev A, Lander ES, Koonin EV, Zhang F. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* 2016; Aug 5;353(6299):aaf5573.

Alderete JF. Alternating phenotypic expression of two classes of *Trichomonas vaginalis* surface markers. Rev. *Infect. Dis.* 1988; 10(Suppl. 2):S408– S412.

Alderete JF, Garza GE. Identification and properties of *Trichomonas vaginalis* proteins involved in cytoadherence. *Infect Immun* 2988; 56: 26-33.

Alderete JF, Kasmala L. Monoclonal antibody to a major glycoprotein immunogen mediates differential complement-independent lysis of *Trichomonas vaginalis. Infect. Immun* 1986a; 53:697–699.

Alderete JF, Kasmala L, Metcalfe E, Garza GE. Phenotypic variation and diversity among *Trichomonas vaginalis* and correlation of phenotype with trichomonal virulence determinants. *Infect Immun* 1986b; 53:285–293.

Alderete JF, Provenzano D, Lehker W. Iron mediates *Trichomonas vaginalis* resistance to complement lysis. *Microb. Pathog* 1995; 19: 93–103.

Arroyo R, Alderete JF. *Trichomonas vaginalis* surface proteinase activity is necessary for parasite adherence to epithelial cells. *Infect Immun* 1989; 57:2991–2997.

Banik GR, Stark D, Rashid H, Ellis JT. Recent Advances in Molecular Biology of Parasitic Viruses. *Infect Disord Drug Targets* 2014; 14(3):155-67.

Brierley I, Jenner AJ, Inglis SC. Mutational analysis of the "slippery-sequence" component of a coronavirus ribosomal frameshifting signal. *J. Mol. Biol* 1992; 227, 463–479.

Carlton JM, et al. Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science* 2007; 315: 207e212.

Cotch MF, Pastorek JG 2nd, Nugent RP, et al. *Trichomonas vaginalis* associated with low birth weight and preterm delivery. The Vaginal infections and Prematurity Study Group. *Sex Transm Dis* 1997; 24:1–8.

DaRocha WD, Otsu K, Teixeira SM, Donelson JE. Tests of cytoplasmic RNA interference (RNAi) and construction of a tetracycline-inducible T7 promoter system in *Trypanosoma cruzi. Mol. Biochem. Parasitol* 2004; 133: 175e186.

Delgadillo MG, Liston DR, Niazi K, Johnson PJ. Transient and selectable transformation of the parasitic protist *Trichomonas vaginalis*. *Proc Natl Acad Sci USA* 1997; 94:4716–4720.

Diamond LS. The establishment of various trichomonads of animals and man in axenic cultures. *J Parasitol*. 1957 Aug;43(4):488-90.

El-Bastawissi AY, Williams MA, Riley DE, Hitti J, Krieger JN. Amniotic Fluid Interleukin-6 and Preterm Delivery: A Review. *Obstet Gynecol* 2000; 95(6 Pt 2):1056-64.

Esteban R, Fujimura T, Wickner RB. Internal and terminal *cis*-acting sites are necessary for *in vitro* replication of the L-A double-stranded RNA virus of yeast. *The EMBO Journal* 1989; 8: 947–954.

Fang Y, Treffers EE, Li Y, Tas A, Sun Z, van der Meer Y, de Ru AH, van Veelen PA, Atkins JF, Snijder EJ, Firth AE. Efficient -2 frameshifting by mammalian ribosomes to synthesize an additional arterivirus protein. *Proc. Natl. Acad. Sci. U. S. A.* 2012; 109:E2920 –E2928.

Fichorova RN. Impact of *T. vaginalis* infection on innate immune responses and reproductive outcome. *J Reprod Immunol* 2009; 83: 185–189.

Fichorova RN, Lee Y, Yamamoto HS, Takagi Y, Hayes GR, Goodman RP, Chepa-Lotrea X, Buck OR, Murray R, Kula T, Beach DH, Singh BN, Nibert ML. Endobiont viruses sensed by the human host—beyond conventional antiparasitic therapy. *PLoS One* 2012; 7:e48418.

Fichorova RN, Trifonova RT, Gilbert RO, Costello CE, Hayes GR, Lucas JJ, Singh BN. *Trichomonas vaginalis* lipophosphoglycan triggers a selective upregulation of cytokines by human female reproductive tract epithelial cells. *Infect. Immun* 2006; 74: 5773–5779.

Fujimura T, Esteban R, Esteban LM, Wickner RB. Portable encapsidation signal of the L-A double-stranded RNA virus of *S. cerevisiae. Cell* 1990; 62: 819–828.

Fujimura T, Wickner RB. Reconstitution of Template-dependent *in vitro* Transcriptase Activity of a Yeast Double-stranded RNA Virus. *J Biol Chem* 1989; 264(18):10872-7.

Garlapati S, Wang CC. Identification of a novel internal ribosome entry site in giardiavirus that extends to both sides of the initiation codon. *J Biol Chem* 2004; 279(5):3389-3397.

Ghabrial SA. *Totiviruses*, In BWJ Mahy, MHV Van Regenmortel (ed.), Encyclopedia of virology 3rd ed., Elsevier Academic Press, San Diego, CA. 2008; 5:163–174.

Gillin FD, Sher A. Activation of the alternative complement pathway by *Trichomonas vaginalis. Infect. Immun* 1981; 34:268–273.

Goodman RP, Freret TS, Kula T, Geller AM, Talkington MW, Tang-Fernandez V, Suciu O, Demidenko AA, Ghabrial SA, Beach DH, Singh BN, Fichorova RN, Nibert ML. Clinical isolates of *Trichomonas vaginalis* concurrently infected by strains of up to four *Trichomonasvirus* species (Family *Totiviridae*). *J Virol* 2011a; 85: 4258–4270.

Goodman RP, Ghabrial SA, Fichorova RN, Nibert ML. *Trichomonasvirus*: a new genus of protozoan viruses in the family *Totiviridae*. *Arch Virol* 2011b; 156: 171–179.

Ives A, Ronet C, Prevel F, Ruzzante G, Fuertes-Marraco S, Schutz F, Zangger H, Revaz-Breton M, Lye LF, Hickerson SM, Beverley SM, Acha-Orbea H, Launois P, Fasel N, Masina S. Leishmania RNA virus controls the severity of mucocutaneous leishmaniasis. *Science* 2011; 331: 775–778.

Karabiber F, McGinnis JL, Favorov OV, Weeks KM. QuShape: rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *RNA*. 2013 Jan;19(1):63-73.

Khoshnan A, Alderete JF. Characterization of double-stranded RNA satellites associated with the Trichomonas vaginalis virus. *J Virol* 1995; 69: 6892–6897.

Khoshnan A, Alderete JF. *Trichomonas vaginalis* with a doublestranded RNA virus has upregulated levels of phenotypically variable immunogen mRNA. *J. Virol* 1994; 68:4035– 4038.

Kigozi G, Brahmbhatt H, Wabwire-Mangen F, Wawer MJ, Serwaqdda D, Sewankambo N, Gray RH. Treatment of Trichomonas in pregnancy and adverse outcomes of pregnancy: a subanalysis of a randomized trial in Rakai, Uganda. *Am. J. Obstet. Gynecol* 2003; 189:1398–1400.

Klebanoff M, Carey J, Hauth J, Hillier SL, Nugent R, Thom E, Ernest J, Heine R, Wapner R, Trout W, Moawad A, Leveno K, Miodovnik M, Sibai B, Van Dorsten J, Dombrowski M, O'Sullivan M, Varner M, Langer O, McNellis D, Roberts J. Failure of metronidazole to prevent preterm delivery among pregnant women with asymptomatic *Trichomonas vaginalis* infection. *N. Engl. J. Med* 2001; 345:487–493.

Li W, Ding H, Zhang X, Cao L, Li J, Gong P, Li H, Zhang G, Li S, Zhang X. 2012. The viral RNA-based transfection of enhanced green fluorescent protein (EGFP) in the parasitic protozoan *Trichomonas vaginalis. Parasitol Res*. 2012; 110(3):1305-10.

Li W, Saraiya AA, Wang CC. Gene regulation in *Giardia lambia* involves a putative microRNA derived from a small nucleolar RNA. *PLoS Negl Trop Dis*. 2011; Oct;5(10):e1338.

Lin WC, Li SC, Shin JW, Hu SN, Yu XM, Huang TY, Chen SC, Chen HC, Chen SJ, Huang PJ, Gan RR, Chiu CH, Tang P. Identification of microRNA in the protist *Trichomonas vaginalis. Genomics* 2009; 93: 487e493.

Liu HW, Chu YD, Tai JH. Characterization of Trichomonas vaginalis virus proteins in the pathogenic protozoan T. vaginalis. *Arch Virol* 1998; 143:963–970.

Lye LF, Owens K, Shi H, Murta SM, Vieira AC, Turco SJ, Tschudi C, Ullu E, Beverley SM. Retention and loss of RNA interference pathways in trypanosomatid protozoans. *PLoS Pathog* 2010; 6: e1001161.

Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci U S A*. 2004 May 11; 101(19):7287-92.

Mendoza-Lopez M, Becerril-Garcia C, Fattel-Facenda LV, Avila-Gonzalez L, Ruiz-Tachiquin ME, Ortega-Lopez J, Arroyo R. CP30, a cysteine proteinase involved in *Trichomonas vaginalis* cytoadherence. *Infect. Immun* 2000; 68:4907–4912.

Mitchell MD, Branch DW, Lundin-Schiller S, Romero RJ, Daynes RA, Dudley DJ. Immunologic aspects of preterm labor. *Semin. Perinatol* 1991; 15:210–224.

Mundodi V, Kucknoor KS, Klumpp DJ, Chang TH, Alderete JF. Silencing the ap65 gene reduces adherence to vaginal epithelial cells by *Trichomonas vaginalis*. *Mol Microbiol.* 2004; 53(4): 1099–1108.

Narimatsu R, Wolday D, Patterson BK. IL-8 increases transmission of HIV type 1 in cervical explant tissue. *AIDS Res. Hum. Retrovirus* 2005; 21: 228–233.

Parent KN, Takagi Y, Cardone G, Olson NH, Ericsson M, Yang M, Lee Y, Asara JM, Fichorova RN, Baker TS, Nibert ML. Structure of a protozoan virus from the human genitourinary parasite *Trichomonas vaginalis*. *MBio*. 2013; 4(2). pii: e00056-13.

Pattnaik AK, Ball LA, LeGrone AW, Wertz GW. Infectious defective interfering particles of VSV from transcripts of a cDNA clone. *Cell*. 1992 Jun 12; 69(6):1011-20.

Provenzano D, Alderete JF. Analysis of human immunoglobulin- degrading cysteine proteinases of *Trichomonas vaginalis. Infect. Immun* 1995; 63:3388–3395.

Provenzano D, Khoshnan A, Alderete JF. Involvement of dsRNA virus in the protein composition and growth kinetics of host *Trichomonas vaginalis. Arch. Virol* 1997; 142:939 –952.

Prucca CG, Slavin I, Quiroga R, Elı´as EV, Rivero FD, et al. Antigenic variation in *Giardia lamblia* is regulated by RNA interference. *Nature* 2008; 456: 750–754.

Ravaee R, Ebadi P, Hatam G, Vafafar A, Ghahramani Seno MM. Synthetic siRNAs effectively target cystein protease 12 and α-actinin transcripts in *Trichomonas vaginalis*. *Exp Parasitol. 2015*; Oct;157:30-4.

Ribas JC, Fujimura T, Wickner RB. Essential RNA binding and packaging domains of the Gag-Pol fusion protein of the L-A double-stranded RNA virus of *Saccharomyces cerevisiae*. *The J Bio Chem* 1994; 269: 28420–28428.

Sambrook J, Fritsch EF, Maniatis T. Molecular Cloning: A Laboratory Manual, (2nd Ed.). 1990; 10:27-10.37.

Sen GC, Sarkar SN. Transcriptional signaling by double-stranded RNA: role of TLR3. *Cytokine Growth Factor Rev* 2005; 16: 1–14.

Schenborn ET, Mierendorf RC Jr. A novel transcription property of SP6 and T7 RNA polymerases: dependence on template structure. *Nucleic Acids Res*. 1985; 13:6223-36.

Schwebke JR, Burgess D. Trichomoniasis. *Clin Microbiol Rev* 2004; 17:794–803.

Shaio MF, Lin PR, Liu JY, Yang KD. Generation of interleukin-8 from human monocytes 11 in response to *Trichomonas vaginalis* stimulation. *Infect. Immun* 1995; 63: 3864–3870.

Singh BN, Lucas JJ, Fichorova RN. *Trichomonas vaginalis*: Pathobiology and Pathogenesis. In: Khan NA, et al., editors. Emerging Protozoan Pathogens. London, UK: Taylor & Francis Group 2007; 411-455.

Snipes LJ, Gamard PM, Narcisi EM, Beard CB, Lehmann T, Secor WE. Molecular epidemiology of metronidazole resistance in a population of *Trichomonas vaginalis* clinical isolates. *J Clin Microbiol* 2000; 38:3004–3009.

Sommer U, Costello CE, Hayes GR, Beach DH, Gilbert RO, Lucas JJ, Singh BN. Identification of *Trichomonas vaginalis* cysteine proteases that induce apoptosis in human vaginal epithelial cells. *J. Biol. Chem* 2005; 280: 23853–23860.

Su HM, Tai JH. Genomic organization and sequence conservation in type I Trichomonas vaginalis viruses. *Virology* 1996; 222:470–463.

Tai JH, Chang SC, Ip CF, Ong SJ. Identification of a satellite double-stranded RNA in the parasitic protozoan *Trichomonas vaginalis* infected with T. vaginalis virus T1. *Virology* 1995; 208: 189–196.

Triana-Alonso FJ, Dabrowski M, Wadzack J, Nierhaus KH. Self-coded 3′-Extension of Run-off Transcripts Produces Aberrant Products during *in vitro* Transcription with T7 RNA Polymerase. *J Biol Chem.* 1995; 17:6298-307.

Wang AL, Wang CC. Viruses of the Protozoa. *Annu. Rev. Microbial.* 1991; 45:251-63.

Wang A, Wang CC, Alderete JF. *Trichomonas vaginalis* phenotypic variation occurs only among trichomonads infected with the double-stranded RNA virus. *J Exp Med* 1987; 166:142–150.

Weber B, Mapeka TM, Maahlo MA, Hoosen AA. Double stranded RNA virus in South African *Trichomonas vaginalis* isolates. *J Clin Pathol* 2003; 56:542–543.

Wickner RB, CC Wang, JL Patterson. Family *Totiviridae*, In C. M. Fauquet, M. A. Mayo, J. Maniloff, U. Desselberger, and L. A. Ball (ed.), Virus taxonomy. Eighth report of the International Committee on Taxonomy of Viruses. Elsevier Academic Press, San Diego, CA. 2005; 571–580

Wickner RB, Fujimura T, Esteban R. Viruses and prions of *Saccharomyces cerevisiae*. *Adv Virus Res* 2013; 86:1-36.

Wilkinson KA, Merino EJ, Weeks KM. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat Protoc.* 2006; 1(3):1610-6.

World Health Organization. Global incidence and prevalence of selected curable sexually transmitted infections – 2008.2012; apps.who.int/iris/bitstream/10665/75181/1/ 9789241503839_eng.pdf?ua=1.

World Health Organization. Trichomoniasis. Global prevalence and incidence of selected curable sexually transmitted infections overview and estimates. 2001; 27-29 http://www.who.int/hiv/pub/sti/who_hiv_aids_2001.02.pdf

Xu J, Hendrix RW, Duda RL. Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. Mol. *Cell* 2004; 16:11–21.

Yu DC, Wang AL, Botka CW, Wang CC. Protein synthesis in *Giardia lamblia* may involve interaction between a downstream box (DB) in mRNA and an anti-DB in the 16S-like ribosomal RNA . *Mol. Biochem. Parasitol* 1998; 96, 151–165.

Yu DC, Wang AL, Wang CC. Amplification, expression, and packaging of a foreign gene by giardiavirus in *Giardia lamblia. J. Virol* 1996; 70: 8752-8757.

Yu DC, Wang AL, Wu CH. Wang CC. Virus-mediated expression of firefly luciferase gene in the parasitic protozoan *Giardia lamblia. Mol. Cell. Biol* 1995; 15: 4867-4872.

Zangger H, Hailu A, Desponds C, Lye LF, Akopyants NS, Dobson DE, Ronet C, Ghalib H, Beverley SM, Fasel N. *Leishmania aethiopica* Field Isolates Bearing an Endosymbiontic dsRNA Virus Induce Pro-inflammatory Cytokine Response. *PLoS Negl Trop Dis* 2014; 8(4):e2836.

Chapter Seven: Dissertation Conclusion

7.1 Dissertation Conclusion

Disease outbreaks and scientific progress from this century have greatly highlighted, expanded, and changed the view and magnitude of impact that RNA viruses play in and out of human health. Taylor et al. and Woolhouse et al. surveyed all known species of human pathogens and reported that RNA viruses make up the majority of the >1,400 pathogen species as well as majority of emerging and re-emerging species (Taylor 2001; Woolhouse 2005). Discoveries of RNA viruses in marine microbes, their ability to lyse their hosts, and measuring relative levels of RNA and DNA from seawater have begun to shift the view of RNA viruses playing ecologically unimportant roles (Tai 2003; Nagasaki 2004; Shirai 2008; Steward 2008). The size and identity of the virosphere is being rapidly uncovered due to advancements in sequencing technology. High-throughput sequencing was developed in 2000 and the next year it had already been applied to discovering new viruses from serum (Allander 2001). Now there exists well over 5Tb of metagenomics data and over 125,000 reported new viruses based (Paez-Espino 2016). While writing this dissertation and a week before submission, another 214 newly identified RNA viruses of vertebrates had been reported from metagenomics data alone (Shi 2018). The large influx and increasing use of metagenomics data for virus discovery has resulted in the International Committee on Taxonomy of Viruses (ICTV) endorsing the admission of new viruses solely identified from metagenomics data (Simmonds 2017).

Evaluating the accuracy of RNA viral genomes may be a bit trickier than that of DNA viral genomes. Given that RNA viruses experience mutation rates many folds higher than DNA viruses (Hanada 2004; Jenkins 2002; Holmes 2009; Fleischmann 1996) and the sequences of viral RNA-dependent RNA polymerases (RdRp) share so little sequence similarity between families (Poch 1989; Gorbalenya 2002; Shackelton 2008), it may be hard to separate genuine mutations from

human error. Historically, in addition to analyzing viral sequences, the ICTV has classified viruses using experimentally determined features, such as replication cycle and properties of the virus particle, but a corollary to the use of metagenomics data alone for virus identification is that the ICTV will now have to identify rules that can be used to classify the new viruses based solely on genomic sequences. Towards this end, some features of viral genomes that will likely be considered, and have been considered, are relatedness of RdRp sequences to previously characterized and established virus families and genome organizations of the virus families.

Indeed, we have analyzed the relatedness of RdRp sequences and the genome organizations of previously reported RNA viruses and, as a result, have identified accidental errors in sequence reporting and incomplete reporting of their genome sequences for the Zygosaccharomyces bailii virus Z (ZbV-Z) and Cryptosporidium parvum virus 1 (CSpV1), respectively (Chapter Two; Chapter Three). In the case of ZbV-Z, we were able to use this genomic analysis to create a hypothesis specifically targeting a single missing nucleotide within a genomic region spanning the length of just 22bps. Phylogenetic analysis have also helped us assess relatedness between viruses infecting different species within the genus *Cryptosporidium* and using this we were able to hypothesize that the current genome of the type species *Cryptosporidium parvum virus 1* is truncated in both of its genome segments.

Additional considerations that the ICTV may likely use for classifying new viruses from genomes alone are conserved terminal sequences outside of open reading frames (ORFs) and sequences encoding translational recoding mechanisms. Upon determining the truncated terminal sequence of CSpV1, we have indeed found such sequences in both the 5' and 3' untranslated regions (UTRs) that are also present in other viruses that infect other protozoan species of the genus *Cryptosporidium* (Fig. 9; Fig. 10) supporting inclusion of all these viruses under the same

virus genus *Cryspovirus*. This is particularly notable because these sequences are often involved in transcription, packaging, or translation of the viral genome. And for CSpV1 it is even more notable because cryspoviruses belong to the virus family *Partitiviridae* and members of this family are multicomponent viruses that have their bi-segmented genome segments individually packaged in separate virus particles (Nibert 2015). Analysis of the translation mechanisms in first-reported ZbV-Z genome have also contributed clues to our hypothesis that its GenBank sequence may contain an error. In fact, after doing a search for similar RdRp sequences to the ZbV-Z's RdRp sequence and identifying a possible virus family to classify the virus, it was using an analysis of its translation mechanisms that we were able to specifically propose where a single missing nucleotide error may be found in its genome. Confirmation of our hypothesis has thus rearranged its genome organization to mirror that of other virus members within the virus family *Amalgaviridae*. Further, while reporting on ZbV-Z and analyzing the translational recoding mechanisms among other closely related viruses of ZbV-Z we came across another possible sequence error in Ustilaginoidea virens unassigned RNA virus UvURV. A single nucleotide substitution within the stop codon that defines the upstream end of ORF2 would extend this ORF enough so that its genome organization will mirror that of the other reported uniranaviruses (Section 2.4.2).

Thus, understanding the information encoded by viral genomes can be used to classify newly identified viruses based on metagenomics data alone. The ability to classify viral genomes from metagenomics data has expanded our knowledge of the virosphere and in doing so it will inform our understanding of viral evolution through time (Shi 2018). One additional layer of information that can be derived from metagenomics data alone and may be worth the ICTV's consideration in classification is information of virus-host associations. Indeed, we have used the

understanding of viral genome organization and RdRp sequence similarities to discover new RNA viruses from metagenomics data (Chapter Four). While it is true that RNA viruses tend to exhibit more cross-species transmission than DNA viruses, RNA viruses are still limited in the hosts they can infect and successfully replicate in (Holmes 2009). The most recent study to date addressing the long-term evolutionary relationship between RNA viruses of vertebrates and their hosts was recently reported by Shi et al. (2018). After identifying over 200 new vertebrate RNA viruses through a metagenomics sample of 186 vertebrate species of fish, amphibians, and reptiles, including their basally branching species, such as jawless fish diverging back 500 million years ago, the team then assembled phylogenetic trees to describe the relationship between these viruses and other previously known viruses infecting mammals and birds. Interestingly, their results showed that the evolution and divergence of these vertebrate RNA viruses largely mirrored the evolutionary divergence of their hosts. Occasional cross-species transmission were identified, but each vertebrate class was still dominated by its own set of RNA viruses (Shi 2018). Thus, host or host range may be a less-forward criteria to use in the taxonomy of viruses, especially as more viruses are being identified, but this seems to be a warranted discussion for the ICTV.

Indeed, we have expanded the presumed host range for members of genus *Mitovirus*, using the aforementioned analyses of viral genomes and metagenomics data, thus better informing the evolutionary history of these viruses as well as possibly having to reconsider classification of these viruses (Chapter Four). In Chapter Four of this dissertation, we have provided strong evidence for mitoviruses infecting plants. This shifts the current thinking about the evolutionary history of mitoviruses with regards to the assumption that mitoviruses have all gone extinct in plant hosts and are thus only presently found to infect fungi (Bruenn 2015). Further, phylogenetic analysis of newly identified plant mitoviruses suggests that these viruses cluster into currently recognized

Clade II of genus *Mitovirus*, though having a more complete identification of all mitoviruses may reveal of a new cluster elsewhere or further spread this cluster. A proposed change to current mitovirus classification is also presented in Chapter Four to accommodate these new plant viruses within the currently broad distribution of these viruses. Thus, here we have put priority on RdRp sequence similarity and genome organization, but have considered host range for classification of these new mitovirus-related viruses of plants.

A grand, unifying model of the phylogenetic relationships between all RNA viruses will largely influence how the ICTV classifies newly identified viruses from metagenomics data alone. Despite all RNA viruses encoding either an RNA-dependent RNA polymerase or reverse transcriptase (RT), such phylogenetic trees have been unsuccessful, though as mentioned earlier there is enough sequence similarities for viruses within families to permit significant alignments. Conserved short amino acid motifs can be located within the conserved palm subdomain structure (Poch 1989; Gorbalenya 2002; Shackelton 2008), suggesting that searching outside of nucleic acid sequence similarities may provide useful insights towards understanding how all RNA viruses may be related. Even still, there is no strong evidence, yet, to suggest that all RNA viruses may even derive from a single common ancestor. Multiple ancestral RNA viruses can also explain our current inability to relate all RNA viruses. However, expansion of the diversity of known RNA virus families may confirm that all RNA viruses fill a continual phylogenetic spectrum, thus making it possible to understand how all RNA viruses are related without the development of new phylogenetic strategies. Expanding the diversity of known RNA viruses from >200 invertebrates by >1,400 new RNA viruses seem to suggest this outcome (Shi 2016).

Use of higher-order similarities may be so broadly applicable as to then become too difficult to use with regards to taxonomy, such as structures so common that they are found in both

RNA and DNA viruses, like the canonical jelly-fold protein structure of many capsid proteins of icosahedral viruses (Coulibaly 2005) or the palm subdomain of both RNA and DNA polymerases (Koonin 2006). Conversely, looking at broad ranging evolutionary pressures at the single-nucleotide level will also be difficult to use for taxonomy between virus families. One such finding, though useful for understanding the evolutionary relationship between virus and host but not relevant for taxonomy, is the low frequency of CG dinucleotide pairs (Takata 2017). Perhaps some intermediary level of secondary information may be useful. Using a rolling window of about 10 amino acids and mapping similarity scores across entire ORFs within these windows, Bruenn reports being able to identify regions of six conserved motifs between RdRps of different RNA virus families (Bruenn 2003). But it is not yet clear how higher-order information can be applied towards phylogeny or if these higher-order structures evolved through convergent evolution, though use of this information may help guide inter-family relatedness when nucleotide sequence similarity is insufficient. Thus, perhaps a grand phylogenetic tree of all RNA viruses will not be generated in the immediate future and looking for higher-level relatedness between RNA virus families should be saved for a more distant future.

Understanding deeper relationships between the genome sequences of viruses within the same virus family, however, is possible by looking beyond RdRp sequences or genome organizations. As mentioned above, determining the terminal end sequences of bi-segmented CSpV1's genome segments revealed conserved stretches of sequences that were also seen in the UTRs of other cryspoviruses (Fig. 9; Fig. 10). Indeed such terminal sequences are strongly shared for viruses of the same genus, yet different between species. One example is that of the genus *Trichomonasvirus*, which contains four virus species, each containing terminal sequences very different from viruses in other genera of the same family but similar to other species within the

genus, yet unique between species (Goodman 2011). These conserved sequences are particularly useful in describing sequences of subviral agents, such as nucleic acid satellites that do not encode for any proteins and thus lack any notable ORFs.

Indeed, this was the case in Chapter Five (Fig. 21; Fig. 22). After surprisingly have identified sequences similar to previously reported dsRNA satellites of Trichomonas vaginas virus (TVV) from our own metagenomics dataset, we further expanded this search in other hosts infected with different TVV species. In doing so, we identified a new clade of dsRNA satellites, expanding the reported clades from two to three, as well as a possible TVV-species specific association for the satellite clades (Fig. 18; Fig. 19). None of these satellites seem to encode any notable ORFs (Fig. 23), so identifying the plus-sense strand is tough and indeed has been disagreed upon by previous groups (Khoshnan 1995; Su 1996). Given the lack of notable ORFs in the satellites, the plus strand is taken as the strand that will be transcribed from its dsRNA genome. However, given that these dsRNA satellites require the TVV helper virus's RdRp for replication, sequences and motifs required for TVV plus-strand replication may also be present in the dsRNA satellites. Indeed, we noticed that in each of the satellites, except for one, one of their strands shares the same 5' terminal sequence as that of the 5' UTR terminal sequence of the plus strands for three of the four TVV helper virus species (Fig. 24). These terminal sequences between the satellites also agree with clade assignments based on genome-wide sequence similarities (Fig. 18; Fig. 21). Thus, analyzing differences within conserved terminal sequences outside of ORFs can be used to understand finer distinctions between species, therefore extending phylogeny to yet a deeper level below species, such as clades.

Use of finer nucleotide differences to tease apart differences between species or clades will at least require confidence in the accuracy of the sequencing data. Indeed, while screening for the

Beta vulgaris mitovirus 1 (BevuMV1) among various cultivars and strains of *Beta vulgaris* (beet) we have identified seven genotypes, where differences between these genotypes are all found within just five nucleotide positions out of a genome size of about 3k nts (Table 7). Increasing the sample size or use of enrichment protocols will improve sensitivity measurements for low abundant genomes or detection of rare and real mutations from a metagenomics dataset. The confidence in the accuracy of sequencing output will also benefit from current improvements in sequencing methods and technologies (Saliba 2014). Single cell RNA-seq is particularly relevant to problems we have come across. While screening for BevuMV1 using seeds of heirloom crops, we have come across ambiguities in the same aforementioned nucleotide positions that would identify the seven genotypes (Chapter Four).

Biological means to assess sequence accuracies and the significance of sequence differences will also increase confidence in evaluating true positive sequence reads from any sequencing error. One such example of likely non-significant sequence differences located within a potentially functionally significant region comes from some analysis done in Chapter Six. After expanding the number of available TVV1 genome sequences from various strains and analyzing for previously reported conserved structures, we found two stem loops originally reported by Su et al. (1996) in all of the new genomes (Fig. 26A). One of the stem loops was unique. The sequences weren't entirely conserved between the strains, but the secondary structures seemed conserved. Their stem loops have stems of the same length and all contain an identical, unique, and asymmetric bulge, and their loops are all made with the same number of nucleotides with only one nucleotide difference in the loop. When there was a sequence difference located within the stem region, there was another complementary mutation to maintain the base-pairing potential in the stem (Fig. 26A). It is likely that this stem loop is involved in some function, but has yet to be

tested. The four known species of TVV may concomitantly infect the same trichomonad and which of the sequence differences present in the terminal ends of the TVV genomes are important for functional demarcation of the species and act in species-specific recognition by its own RdRp is still unknown and has yet to be tested. But it is presumed that heterogeneous encapsidation or replication is unlikely (Goodman 2011). Identifying a function behind specific nucleotides will be a biological means of increasing confidence in true positive sequence reads.

In Chapter Six of this dissertation, we have developed such a tool. We have developed a transient and selectable TVV-mediate reverse genetics system. We have also shown that this system can be used to assay for TVV minus-strand production. TVV minus-strand production requires both the TVV 3' 600nt sequence and the presence of TVV (Chapter Six). Using this system will allow to us assign function to specific sequences and assess the impact that single nucleotide differences will have on these sequence regions. Thus, we will add biologically relevant confidence towards answering questions such as "Why does this sequence region have the identity that it has?" and "What variations from the standard are tolerable and which will significantly affect functional outcomes?"

In answering these questions, we will have more confidence in using sequences to identify viruses from metagenomics data alone, improve our ability to classify these newly identified viruses, and eventually move towards understanding the biological impacts of single-nucleotide differences between viruses of the same species.

7.2 References

Allander T, Emerson SU, Engle RE, Purcell RH, Bukh J. A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. *Proc Natl Acad Sci U S A*. 2001 Sep 25;98(20):11609-14.

Fleischmann WR Jr. Chapter 43-Viral Genetics. In: Baron S, editor. Medical Microbiology. 4th edition. Galveston (TX): University of Texas Medical Branch at Galveston; 1996. Chapter 43.

Bruenn JA. A structural and primary sequence comparison of the viral RNA-dependent RNA polymerases. *Nucleic Acids Res.* 2003 Apr 1;31(7):1821-9.

Bruenn JA, Warner BE, Yerramsetty P. Widespread mitovirus sequences in plant genomes. *PeerJ*. 2015; 3:e876.

Coulibaly F, Chevalier C, Gutsche I, Pous J, Navaza J, Bressanelli S, Delmas B, Rey FA. The birnavirus crystal structure reveals structural relationships among icosahedral viruses. *Cell*. 2005 Mar 25; 120(6):761-72.

Goodman RP, Freret TS, Kula T, Geller AM, Talkington MW, Tang-Fernandez V, Suciu O, Demidenko AA, Ghabrial SA, Beach DH, Singh BN, Fichorova RN, Nibert ML. Clinical isolates of *Trichomonas vaginalis* concurrently infected by strains of up to four *Trichomonasvirus* species (Family *Totiviridae*). *J Virol* 2011; 85: 4258–4270.

Gorbalenya AE, Pringle FM, Zeddam JL, Luke BT, Cameron CE, Kalmakoff J, Hanzlik TN, Gordon KH, Ward VK. The palm subdomain-based active site is internally permuted in viral RNA-dependent RNA polymerases of an ancient lineage. *J Mol Biol*. 2002 Nov 15;324(1):47-62.

Hanada K, Suzuki Y, Gojobori T. A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. *Mol Biol Evol*. 2004 Jun;21(6):1074-80.

Holmes EC. The Evolution and Emergence of RNA Viruses. Oxford University Press. 2009. Chapter 6.

Jenkins GM, Rambaut A, Pybus OG, Holmes EC. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J Mol Evol*. 2002 Feb;54(2):156-65.

Koonin EV, Senkevich TG, Dolja VV. The ancient Virus World and evolution of cells. *Biol Direct*. 2006 Sep 19; 1:29.

Khoshnan A, Alderete JF. Characterization of double-stranded RNA satellites associated with the Trichomonas vaginalis virus. *J Virol* 1995; 69: 6892–6897.

Nagasaki K, Shirai Y, Takao Y, Mizumoto H, Nishida K, Tomaru Y. Comparison of genome sequences of single-stranded RNA viruses infecting the bivalve-killing dinoflagellate *Heterocapsa circularisquama*. *Appl Environ Microbiol*. 2005; 71: 8888–8894.

Nibert ML, Woods KM, Upton SJ, Ghabrial SA. *Cryspovirus*: a new genus of protozoan viruses in the family *Partitiviridae*. *Arch Virol* 2015; 154:1959–1965.

Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC. Uncovering Earth's virome. *Nature.* 2016 Aug 25; 536(7617):425-30.

Poch O, Sauvaget I, Delarue M, Tordo N. Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. *EMBO J.* 1989; 8: 3867–3874.

Saliba AE, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.* 2014 Aug;42(14):8845-60.

Shackelton LA, Holmes EC. The role of alternative genetic codes in viral evolution and emergence. *J Theor Biol* 2008; 254:128–134.

Shi M, Lin XD, Chen X, Tian JH, Chen LJ, Li K, Wang W, Eden JS, Shen JJ, Liu L, Holmes EC, Zhang YZ. The evolutionary history of vertebrate RNA viruses. *Nature*. 2018 Apr 4. doi: 10.1038/s41586-018-0012-7.

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

Shirai Y, Tomaru Y, Takao Y, Suzuki H, Nagumo T, Nagasaki K. Isolation and characterization of a single-stranded RNA virus infecting the marine planktonic diatom *Chaetoceros tenuissimus Meunier*. *Appl Environ Microbiol*. 2008; 74: 4022–4027.

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. Consensus statement: virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol*. 2017; 15: 161–168.

Steward GF, Culley AI, Mueller JA, Wood-Charlson EM, Belcaid M, Poisson G. Are we missing half of the viruses in the ocean? *ISME J.* 2013; 7: 672–679.

Su HM, Tai JH. Genomic organization and sequence conservation in type I Trichomonas vaginalis viruses. *Virology* 1996; 222:470–463.

Tai V, Lawrence JE, Lang AS, Chan AM, Culley AI, Suttle CA. Characterization of HaRNAV, a singlestranded RNA virus causing lysis of Heterosigma akashiwo (Raphidophyceae). *J Phycol*. 2003; 39: 343–352.

Taylor LH, Latham SM, Woolhouse ME. Risk factors for human disease emergence. *Philos Trans R Soc Lond B Biol Sci.* 2001;356: 983–9.

Takata MA, Gonçalves-Carneiro D, Zang TM, Soll SJ, York A, Blanco-Melo D, Bieniasz PD. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. *Nature*. 2017 Oct 5; 550(7674):124-127.

Woolhouse ME, Gowtage-Sequeria S. Host range and emerging and reemerging pathogens. *Emerg Infect Dis*. 2005 Dec;11(12):1842-7.