EVOLUTION OF FAST-EVOLVING VIRUSES AT VARIOUS TIMESCALES

By

LELE ZHAO

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Microbial Biology

Written under the direction of

Siobain Duffy

And approved by

New Brunswick, New Jersey

JANUARY, 2019

ABSTRACT OF THE DISSERTATION

Evolution of fast-evolving viruses at various timescales

By LELE ZHAO

Dissertation Director:

Siobain Duffy

In an attempt to combat emerging and re-emerging viral pathogens and to understand the deep evolutionary origins of viruses, a vast amount of research effort is being devoted to the study of virus evolution. Of the many viral entities on this planet, fast-evolving viruses are most problematic and most intriguing. Using experimental evolution and comparative computational methods, this dissertation addresses questions in virus evolution of fast-evolving viruses at various timescales. A model RNA virus, the *Pseudomonas* bacteriophage phi6, was used for experimental evolution studies on genetic diversity. Starting from an overnight growth on an agar plate, three phi6 genotypes were assessed for their mutational frequency on a novel host. It was observed that extant host range mutations are epistatically constraining subsequent host range mutational neighborhoods during sequential host shift events. Next, genotypic generalist and ecological generalist were compared to specialist during long term experimental evolution to see if their evolutionary advantages can be attributed to higher genetic diversity. Ecological history was the major determinant of population genetic diversity in this study, in which selection on novel hosts purged generalist populations' genetic diversity and the specialist ecology was better at maintained higher levels of genetic diversity. Moving on to a larger timescale and to the fast-evolving single-stranded DNA viruses, 926 replication-associated protein (Rep) sequences of circular Rep-encoding ssDNA (CRESS DNA) viruses were collected from GenBank RefSeq database and used to estimate a Rep specific amino acid substitution matrix to model the evolutionary patterns of this homologous protein. Amidst the many recent taxonomic revisions in the CRESS DNA viruses, we contributed a novel matrix and a complete Rep tree as the accurate backbone for future classifications. Finally, the longest timespan covered in this dissertation was reached in the final chapter, on paleovirology of CRESS DNA viruses. Endogenous Rep sequences were found in 163 unique species, 24 different eukaryotic phyla across the tree of life using a relaxed tblastn search in the non-redundant eukaryotic nucleotide database. Apart from expansion on previous findings, genomovirus Reps were shown to exclusively group with endogenous sequences from fungal pathogens, where the only characterized genomovirus was isolated, suggesting potential hosts for uncharacterized members of the family. This dissertation uses many techniques to study dynamics in virus evolution and advances our ability to study both RNA and CRESS DNA viruses.

ACKNOWLEDGEMENT AND DEDICATION

I would like to thank my advisor Dr. Siobain Duffy, for being a wonderful mentor and friend to me both in science and in life. She is smart, caring, fair, strong, and many more, and I always see her as a role model for when I run my own lab. Her attitude toward scientific research displayed through her obsession for answers, enthusiasm for data and effectiveness in communication has made my Ph.D. studies an unforgettable and enjoyable experience. She taught me about balancing my passion for research and priorities in life. If I had a chance to go back in time to when I started at Rutgers, I would change nothing about my time spent in the Duffy Lab.

I would like to thank my dissertation committee members. First, I would like to thank Dr. Ning Zhang and Dr. Bradley Hillman for their inspirational inputs to my projects and support to my academic career. Second, I would like to thank Dr. Marco Vignuzzi for kindly agreeing to serve as my outside member and bringing his expertise in virus evolution to my dissertation from the Institut Pasteur. Third, I would like to thank Dr. Kay Bidle for being on my dissertation proposal committee and giving insightful suggestions to my project proposals.

This dissertation was made possible by various sources of support and funding. I would like to thank the Department of Biochemistry and Microbiology and the Microbial Biology program at Rutgers University. Thanks for awarding me a Teaching Assistantship in my third year and the Douglas E. Eveleigh Travel Award in 2016. I appreciate the support from our program director Dr. Gerben Zylstra and everyone in the department for making being in the program a great experience. I would also like to thank the Department of Ecology, Evolution and Natural Resources. Marsha and Georgie were wonderful and always had answers to my questions and concerns. I would like to thank the National Science Foundation for supporting my studies and research (grant number: DEB 1240049 and DEB 1453241). And finally, I would like to thank

Rutgers School of Graduate Studies for awarding me multiple Teaching and Graduate Assistant Professional Development Funds to support my research and conference travel.

Many people helped me during the process of my projects. First, I would like to thank our collaborators at the University of South Florida, Dr. Mya Breitbart and Dr. Karyna Rosario. They are very intelligent scientists who work tirelessly in research and outreach. I appreciate their valuable input in many aspects of my work. Second, I would like to thank the Genome Cooperative at the School of Environmental and Biological Sciences, especially Nicole Wagner and Ehud Zel-Zion, our Illumina MiSeq experts. Without them, our lab would be spending many times more money on sequencing for my experiments. Third, I would like to express my gratitude to Dr. Jeremy Draghi who kindly contributed a valuable section to my first publication; to Dr. David Swofford who kindly instructed me with validation methods in my fourth chapter; to Dr. Gael Kurath whose words made me panic less.

I would like to thank my lab mates at the Duffy lab for all the peer mentoring, heated discussions and memorable dinners. I would like to especially thank LaShanda Williams, Natasia Jacko, Willard Mbewe and Dragoş Stemate. I appreciate the company and many long conversations with LaShanda who made me think more and become a more considerate person. I thank Natasia for always being there when I need her or when the lab needs her. I thank Willard for bringing his personality to the group and being a great friend. I thank Dragoş for teaching me how to be a mentor.

Finally, I would like to thank my friends and family. Thanks to my friends here at Rutgers for all the support and comfort throughout the past five years. Thanks to my friends in China for all the love and support no matter when I needed them. Thanks to my parents, Qizu Zhao and Yongfang Wang, for supporting me while I pursue my dreams. Thanks to my sister, Mingming Zhao, for taking care of our parents while I am not there. Thanks to my husband, Chen Zhao, for being the best.

I dedicate this dissertation to my family, the most important part of my life.

PREFACE

Chapter 1 has been accepted as Lele Zhao, Mansha Seth Pasricha, Dragoş Stemate, Alvin Crespo-Bellido, Jacqueline Gagnon, Jeremy Draghi, Siobain Duffy. 2018. Existing host range mutations constrain further emergence of RNA viruses. Journal of Virology. Lele Zhao participated in writing the manuscript and is directly responsible for those experiments dealing with mutational neighborhood mapping of phi6-WT and phi6-E8G, PA mutation frequency assays, fitness assays, Sanger and Illumina sequencing of phi6-WT and phi6-E8G, and deep sequencing data analysis; Figures 1-6; and Tables 1-4.

Chapter 2 is being prepared for publication as Lele Zhao and Siobain Duffy. 2018. We found that neither genetic nor ecological generalists maintain higher population genetic diversity, which is probably due to ecological history being the key determinant of population genetic diversity. Lele Zhao participated in writing the manuscript and is directly responsible for experimental evolution, phenotypic and fitness assays, Illumina sequencing sample preparation and data analysis.

Chapter 3 has been accepted as Lele Zhao, Karyna Rosario, Mya Breitbart, Siobain Duffy. 2018. Eukaryotic Circular Rep-Encoding Single-Stranded DNA (CRESS DNA) Viruses: Ubiquitous Viruses With Small Genomes and a Diverse Host Range. Advances in Virus Research. Lele Zhao participated in writing the manuscript; Figure 4; and Table 1-3.

Chapter 4 is being prepared for publication as Lele Zhao, Erik Lavington, Siobain Duffy. 2018. We estimated a CRESS DNA (circular replication-associated-protein-encoding single stranded DNA) virus specific amino acid substitution matrix and built a Rep (replication-associated protein) genealogy for CRESS DNA viruses. Lele Zhao participated in writing the manuscript and is directly responsible for those experiments dealing with matrix estimation and evaluation, and genealogy construction.

Chapter 5 is being prepared for publication as Lele Zhao, Erik Lavington, Siobain Duffy. 2018. We found CRESS DNA viral Reps in 163 species inside 24 eukaryotic phyla. Many identified eukaryotic species suggest past or current hosts of many uncharacterized CRESS DNA viruses. Lele Zhao participated in writing the manuscript and is directly responsible for tblastn searches, and phylogenetic tree construction.

TABLE OF CONTENTS

ABSTRATC OF THE DISSERTATION ii

ACKNOWLEDGEMENT AND DEDICATION iv

PREFACE vii

TABLE OF CONTENTS ix

TABLE OF TABLES xiii

TABLE OF ILLUSTRATIONS xiv

Introduction 1

The model system phi6 3

CRESS DNA viruses 5

Overview of the chapters 6

References 11

Chapter 1 19

Existing host range mutations constrain further emergence of RNA viruses

Abstract 19

Importance 20

Introduction 21

Results 22

Discussion 40

Material and Methods 44

Acknowledgements 48

References 49

Chapter 2 54

Gauging genetic diversity of generalists: a test of genetic and ecological generalism with phi6 experimental evolution

Abstract 54

Introduction 55

Material and Methods 57

Results 61

Discussion 73

Acknowledgements 76

References 77

Chapter 3 81

Eukaryotic Circular Rep-Encoding Single-Stranded DNA (CRESS DNA) viruses: ubiquitous viruses with small genomes and a diverse host range

Abstract 81

Defining CRESS DNA Viruses 82

Discovery of Eukaryotic CRESS DNA Viruses 84

Unity and Diversity 88

Ecology of Eukaryotic CRESS DNA Viruses 107

Evolution of Eukaryotic CRESS DNA Viruses 123

Conclusions 128

Acknowledgements 129

References 130

Chapter 4 153

The genealogy of the replication associated protein of circular rep-encoding single-stranded DNA viruses (CRESS DNA viruses)

Abstract 153

Introduction 155

Material and Methods 158

Results 163

Discussion 179

Acknowledgements 191

References 192

Chapter 5 201

Truly ubiquitous, CRESS DNA viruses scattered across the eukaryotic tree of life

Abstract 201

Introduction 202

Material and Methods 205

Results 206

Discussion 223

Acknowledgements 231

References 232

Appendices 236

LIST OF TABLES

TABLE 1.1 PA host range mutations detected on P3 with Sanger sequencing. 26

TABLE 1.2 Host range mutational neighborhoods inferred from non-synonymous mutations in P3 from previous publications. 27

TABLE 1.3 PA host range mutations of phi6-WT detected on the Small segment with Sanger sequencing. 33

TABLE 1.4 Pairwise unique SNPs above 0.1% frequency in phi6 populations detected through deep sequencing using VarScan. 36

TABLE 3.1 Basic information on eukaryotic CRESS DNA viral families. 83

TABLE 3.2 Familes, genera, number of classified species, and NCBI GenBank genome accession number for each genus' type species. 96

TABLE 3.3 Representative studies where eukaryotic CRESS DNA viruses were found (2010-recent). 108

TABLE 4.1 Best model chosen by ProtTest3 for each tested ssDNA protein alignment. 178

TABLE 5.1 Summary of results from tblastn using viral Rep queries. 207

LIST OF ILLUSTRATIONS

FIGURE 1.1 2D schematic representing mutational neighborhoods of phi6 P3. 24

FIGURE 1.2 Change in Shannon entropy in Medium segment of phi6-WT, phi6-E8G and phi6-G515S. 29

FIGURE 1.3 Change in Shannon entropy in Small Segment of phi6-WT, phi6-E8G and phi6-G515S. 30

FIGURE 1.4 Change in Shannon entropy in Large segment of phi6-WT, phi6-E8G and phi6-G515S. 31

FIGURE 1.5 PA host range mutation frequency of phi6-WT(4hrs), phi6-WT, phi6-E8G, phi6-G515S. 37

FIGURE 1.6 Relative fitness of host range mutants on PP. 39

FIGURE 2.1 Heat shock tolerance mutation frequency of all lineages. 63

FIGURE 2.2 Number of SNPs above 1% across all lineages of all passaging schemes. 66

FIGURE 2.3 Top 10 SNPs within all Day 30 populations. 68

FIGURE 2.4 Relative fitness of populations on P. 70

FIGURE 2.5 Relative fitness of populations on T and E. 72

FIGURE 3.1 Abutilon mosaic virus symptoms in abutilon, from Hawaii. 85

FIGURE 3.2 Rolling circle replication in CRESS DNA viruses. 90

FIGURE 3.3 Maize streak virus particles. 93

FIGURE 3.4 Maximum likelihood tree of eukaryotic CRESS DNA virus Rep protein sequences. 106

FIGURE 3.5 Symptoms of psittacine beak and feather disease in a Red Rump Parrot. 118

FIGURE 4.1 Comparison of –ln scores of matrices fitting the test set. 165

FIGURE 4.2 Pairwise Pearson correlation of amino acid substitution rates of matrices. 166

FIGURE 4.3 Comparison of amino acid frequencies among the matrices. 170

FIGURE 4.4 Comparison of four matrices. 171

FIGURE 4.5 Maximum likelihood tree built using CRESS matrix. 173

FIGURE 4.6 Unrooted maximum likelihood tree of CRESS Rep endonuclease domain. 174

FIGURE 4.7 Unrooted maximum likelihood tree of CRESS Rep helicase domain. 175

FIGURE 4.8 Unrooted maximum likelihood trees with equal numbers of *Begomovirus* and *Mastrevirus* Reps built with CRESS matrix. 184

FIGURE 4.9 Unrooted maximum likelihood tree built with 123 recombinant sequences removed from the 926 CRESS Rep dataset. 186

FIGURE 4.10 Unrooted maximum likelihood tree built with the full 926 CRESS Rep dataset. 187

FIGURE 4.11 Unrooted maximum likelihood trees with Geminivirus and Genomovirus Reps built with CRESS matrix. 188

FIGURE 4.12 Unrooted maximum likelihood trees with Geminivirus and Genomovirus Reps built with CRESS matrix. 189

FIGURE 5.1 Distribution of endogenous Reps across eukaryotic phyla. 208

FIGURE 5.2 Geminivirus Reps and endogenous sequences midpoint-rooted maximum likelihood tree. 211

FIGURE 5.3 Circovirus Reps and endogenous sequences midpoint-rooted maximum likelihood tree. 212

FIGURE 5.4 Bacilladnavirus Reps and endogenous sequences midpoint-rooted maximum likelihood tree. 215

FIGURE 5.5 Nanovirus Reps and endogenous sequences midpoint-rooted maximum likelihood tree. 216

FIGURE 5.6 Genomovirus Reps and endogenous sequences midpoint-rooted maximum likelihood tree. 217

FIGURE 5.7 Smacovirus Reps and endogenous sequences midpoint-rooted maximum likelihood tree. 218

FIGURE 5.8 Alphasatellites Reps and endogenous sequences midpoint-rooted maximum likelihood tree. 219

Introduction

Viruses are the most abundant biological entities with extraordinary diversity (Breitbart and Rohwer, 2005; Koonin et al., 2006). Viruses have diverse morphology, from icosahedral to spindle shapes (Prangishvili et al., 2006). They have a wide host range, being able to infect all forms of life, even themselves (La Scola et al., 2008; Raoult and Forterre, 2008). Apart from variable genomic size, their genomic composition also comes in different flavors: single-stranded RNA (positive or negative sense), double-stranded RNA, single-stranded DNA, double-stranded DNA, and RNA/DNA hybrid (Koonin and Dolja, 2014; Raoult et al., 2004; Roux et al., 2013). Their successful existence could be attributed to both evolvability and robustness in the face of environmental challenges, whether from the host immune system or environmental instability (Wasik and Turner, 2013). Many RNA and ssDNA viruses mutate at rates orders of magnitudes faster than prokaryotic or eukaryotic cells, which makes them effective competitors to complex host immune defense in the coevolution of virus and host (Lobo et al., 2009; Lukashov and Goudsmit, 2001; Obbard and Dudas, 2014). All of these unique features and rapid dynamics make viral evolution an irresistible area of study.

Virus evolution, especially as studied in molecular epidemiology, is among the most popular topics in biomedical research. Increasing numbers of different viral pathogens and their epidemics are coming into proximity or even becoming a part of people's lives (Cleaveland et al., 2001; Taylor et al., 2001). Spillovers events of avian and swine influenza have drawn much precaution and speculations on potential risk factors in these zoonotic viruses and their evolutionary trajectory to pandemic emergence in the human population (Lipsitch et al., 2016). Taking an evolutionary perspective has also inspired the study of HIV, as it provides new sights into HIV pathogenesis such as the heritability and transmission of specialized virus individuals and constrained evolution (Fraser et al., 2014). The largest and longest Ebola outbreak lasting from 2013-2016 in West Africa resulted in approximately 29,000 cases and some 11,000 deaths

(Geoghegan and Holmes, 2018). The high mortality rate alarmed virologists to study the key substitutions seemingly underlying adaptation (Diehl et al., 2016) and ponder about virulence evolution (Sofonea et al., 2018).

With the mechanisms supporting evolvability mentioned previously, viruses make excellent model systems for conducting basic evolution research. Popular model viruses for experimental evolution include phi6, phiX174, Qβ and VSV, where topics such mutations' effect on fitness (Lee et al., 1997; Sanjuán, 2010; Sanjuan et al., 2004b), parallel evolution (Bollback and Huelsenbeck, 2009; Pepin et al., 2008; Wichman and Brown, 2010), the evolution of sex (Turner and Chao, 1998; Whitlock et al., 2016), adaptive landscapes (Burch and Chao, 1999; Cervera et al., 2016a, b), host range expansion (Turner et al., 2012; Turner et al., 2010; Wasik et al., 2016), and epistasis (Lalic and Elena, 2013, 2015; Sanjuan et al., 2004a) have been studied. However, culture-based methods for studying viruses preclude the study of huge swaths of viral diversity that is as yet uncultured. Indeed, metagenomics and high-throughput sequencing techniques have added a new dimension to virology. With metagenomic discoveries, viral ecologists could truly appreciate the earth's unknown virome. Early viromes were composed of 60-95% "viral dark matter", unlike any sequence documented in the database (Brum and Sullivan, 2015; Mizuno et al., 2013; Reyes et al., 2012; Youle et al., 2012), demonstrating the biological diversity remaining to be understood in viruses. As complete genomes of novel viruses from metagenomic sequencing studies are becoming more frequently validated and deposited to the databases, the International Committee on Taxonomy of Viruses has affirmed they are of equal significance as isolated and characterized virions and should be classified into existing and future virus families (Simmonds et al., 2017). Both wet lab studies and genomic comparisons are essential techniques for understanding viral evolution.

The model system phi6

The model system in this dissertation (Chapters 1 and 2) is the segmented dsRNA *Pseudomonas* bacteriophage phi6. Phi6 is the type species of the family *Cystoviridae*, and it has many features that render it an exceptionally well-suited phage for studying viral evolution. This phage was first purified and described by Dr. Anne Vidaver in 1973 (Vidaver et al., 1973). As Dr. Vidaver studied bacterial plant pathogens, she used filters instead of lipid-dissolving chloroform to purify pathogens, which fortunately preserved the lipid-coated phi6. Phi6 was the second lipid containing phage discovered, after phage PM2 (Espejo and Canelo, 1968). Phi6 has a double-stranded RNA (dsRNA) genome, segmented into three “chromosomes”. DsRNA viruses primarily infect fungi and plants; phi6’s family *Cystoviridae* contains the only prokaryote infecting dsRNA viruses. Phi6 is distantly related to eukaryotic infecting dsRNA viruses, and cystoviruses are the only phages with segmented genomes (Mindich, 1988). Related dsRNA viruses include rotavirus, a pathogen that causes diarrhea, which is a leading cause of infant and children death worldwide (Jiang et al., 2010). *Cystoviridae* is also the only family of bacteriophage that has a lipid envelope, a feature much more common in animal-infecting viruses. All other known lipid-containing phage only have lipid as an inner layer under the protein capsid (Laurinavicius et al., 2004). Although phi6 uses the lipid envelope to fuse with gram-negative *Pseudomonas* bacteria outer membrane during infection (Bamford et al., 1987), it produces a peptidoglycan-dissolving enzyme in order to get through the cell wall (Daugelavicius et al., 2005; Mindich and Lehman, 1979a).

Apart from its use in evolutionary biology, phi6 has been used as a model system for studying genome packaging of dsRNA and negative sense single-stranded RNA viruses (Mindich, 1999; Qiao et al., 1995). Similarities have been observed in this phage family and reoviruses (rotavirus, bluetongue virus) in virion architecture, genome packaging, replication and transcription (Bamford, 2002), therefore results from phi6 studies have clear implications in understanding eukaryote infecting viruses. Possibly the protagonist in a domain-level host switching event, members of *Cystoviridae* are uniquely suited to be phage model systems for eukaryotic viruses.

Phi6's small segment is 2949bp in length, containing coding sequences for proteins: major outer capsid protein P8 (Bamford et al., 1993), morphogenesis protein P12 (Johnson and Mindich, 1994; Sinclair et al., 1975), major envelope protein P9 (Johnson and Mindich, 1994; Mindich et al., 1976) and lytic enzyme P5 (Caldentey and Bamford, 1992). The medium segment is 4065bp in length, containing coding sequences for proteins: envelope membrane protein P10 (Johnson and Mindich, 1994; Mindich and Lehman, 1979b), fusogenic protein P6 (Stitt and Mindich, 1983), host attachment protein or spike protein P3 (Etten et al., 1976), and the non-essential envelope membrane protein P13 (Mindich et al., 1992). And the large segment is 6374bp in length, containing coding sequences for proteins: minor capsid protein/packaging factor P7 (Juuti and Bamford, 1995, 1997), RNA-dependent RNA polymerase P2 (Makeyev and Bamford, 2000), packaging NTPase P4 (Pirttimaa et al., 2002), major capsid protein P1 (Qiao et al., 2003).

Phi6 has many readily available permissive and novel hosts that can be easily cultured in the laboratory. The ones relevant for this dissertation are *Pseudomonas syringae* pathovar *phaseolicola*, *P. syringae* pathovar *tomato*, *P. syringae* pathovar *atrofaciens*, *P. syringae* pathovar *glycinea* and *P. pseudoalcaligenes* East River isolate A (ERA). The ATCC strain of phi6 in this dissertation can infect *P. syringae* pv. *phaseolicola*, its typical laboratory host, without additional mutations, but the other hosts generally all require host range mutations to infect.

Bacteriophages like phi6 are powerful experimental evolution systems because of their short generation time, easy-to-monitor life cycles, high mutation rates, and controllable population size. Phages are nonpathogenic to humans making them ideal for outreach and undergraduate research projects (Hatfull et al., 2006). As there is debate over whether host-shifting experiments and other "gain-of-function" research should be conducted on emergent human pathogens, *Pseudomonas* phage phi6 offers a safe, uncontroversial system in which to test evolutionary theories. Its protein sequence and structural similarities to eukaryotic viruses means that results obtained with this phage system may be generalizable to viruses of multicellular organisms (Bamford, 2003; Butcher et al., 2001; Katz et al., 2012; Mindich, 2004). Some of the successful uses of phi6 experimental evolution include creating a “known phylogeny” to test the accuracy of the phylogenetic reconstruction (Turner et al., 2012), to study viral emergence (Dennehy et al., 2010; Duffy et al., 2007), the evolution of thermotolerance (Dessau et al., 2012; Goldhill et al., 2014; McBride et al., 2008), host range expansion (Ferris et al., 2007; Ford et al., 2014a) and the effect of coinfection on evolution (Montville et al., 2005; Turner et al., 1999).

CRESS DNA viruses

Similar to RNA viruses, single-stranded DNA (ssDNA) viruses also have small genomes and high mutation rates (Duffy et al., 2008; Sanjuán et al., 2010). They attracted scientific focus because many are pathogens of humans, companion animals, livestock and crops. SsDNA viruses with linear genomes include Erthythrovirus B19, which causes fifth disease in children (Flower and Macmahon 2017), and canine parvovirus, which arose by a host shift event from cat populations in the mid-1970s and quickly spread to be endemic globally (Parrish, 1990, 1991; Siegl, 1984). There are many pathogenic circular ssDNA viruses in plants, for example, the whitefly-transmitted tomato yellow leaf curl virus, the leafhopper-transmitted maize streak virus and the aphid-transmitted banana bunchy top virus. All these viruses cause some form of deformation in plants and fruits, reducing yield (Abraham et al., 2012; Cathrin and Ghanim, 2014; Hooks et al., 2008). Another group of circular ssDNA viruses infect animals, which includes porcine circoviruses that cause postweaning multisystemic wasting syndrome in piglets (Allan et al., 1999; Harding and Clark, 1997; Nayar et al., 1997) and beak and feather disease viruses affecting many bird species (Ritchie et al., 1989; Schoemaker et al., 2000). These plant and animal viruses heavily impact food security and animal health across the globe (Bull et al., 2006; dos Santos et al., 2012; Legg and Fauquet, 2004; Lima et al., 2015).

With the commercialization of the processive phi29 polymerase, scientists are able to enrich for previously under-detected circular ssDNA viruses through rolling circle amplification (Li et al., 2010; Rosario et al., 2009, 2012b). As the number of circular ssDNA viruses deposited in databases is exponentially increasing, a majority are identified to encode a homologous replication-associated protein (Rep) required for rolling circle replication and the group was named circular Rep-encoding ssDNA (CRESS DNA) viruses (Rosario et al., 2012a). As the group grows, taxonomic revisions have been published and novel families have been created, which provided an initial structure for the first review of CRESS DNA viruses (Chapter 3). This review discusses the unity and diversity, ecology, and evolution of the eukaryotic CRESS DNA viruses. This comprehensive primer will undoubtedly be rapidly out of date as the discovery and systematization of CRESS DNA viruses continues at a fast pace.

Overview of the chapters

This dissertation focuses on evolution of fast-evolving viruses at short to very long timescales. The experimental evolution chapters study evolution overnight, or over thirty days of passaging, while the computational chapters study evolutionary relatedness over thousands to millions of years.

Mutational neighborhoods determine the pathways of evolution: different mutational neighborhoods facilitate different evolutionary trajectories (Qu et al., 2012), and thus evolvability (Burch and Chao, 2000). Other virologists have demonstrated how codon usage constrains mutational neighborhoods (Lauring et al., 2012) and bacteriologists have shown that the mutational neighborhood constrains adaptation in the *Pseudomonas aeruginosa* (Hall et al., 2010). **The first chapter shows the epistatic consequences of a host range mutation on mutational neighborhoods during overnight growth.** This study (accepted by Journal of

Virology) compares three *Pseudomonas* bacteriophage phi6 genotypes to demonstrate that existing host range mutations constrain subsequent host range expansion. As previous studies have shown in the phi6 system, the ability to infect a novel host can be gained through single non-synonymous changes (Duffy et al., 2006; Ferris et al., 2007; Ford et al., 2014). It is relatively easy for RNA viruses, which have per-site mutation rates from 10^{-3}-10^{-6} and small genomes (Duffy et al., 2008; Sanjuán et al., 2010), to infect novel hosts (Lanciotti et al., 1999; Parrish et al., 2008; Streicker et al., 2010). We show that a "wild-type" phi6 genotype (ATCC strain) has a high host range mutation frequency of 10^{-4}-10^{-5} on a novel host *P. syringae* pv. *atrofaciens* (PA). Since the average mutation rate of phi6 is approximately $2.7x10^{-6}$ per site per round of replication (Chao et al., 2002), this indicates that wild-type phi6 has a substantial host range mutational neighborhood for PA. Two other genotypes with host range mutants that allow infection of *P. syringae* pv. *tomato* (phi6-E8G and phi6-G515S) had much more restricted neighborhoods, with a mutation frequency difference of 10-100 folds to the wild-type. After mapping in detail the PA host range mutational neighborhood of wild-type phi6, phi6-E8G and phi6-G515S through clonal Sanger sequencing and population Illumina sequencing, we quantified the constrained neighborhoods for the isogenic mutants. While wild-type phi6 has at least 16 different ways (single mutations) to get into PA, phi6-E8G has two ways and phi6-G515S has one way. This result is directly applicable to the study of sequential host shift in emerging infectious diseases. Many devastating zoonotic viruses such as MERS-CoV, HIV, and canine parvovirus have a track record of sequential host shifts in their ecological history. Our overnight growth of the viral populations, although guaranteeing a sufficient amount of population genetic diversity, does not take into account the potential effects that many generations' adaptation to intermediate hosts might have. Regardless, we have shown the powerful influences of pre-existing mutations carrying imprints of recent ecological history on subsequent evolutionary trajectories.

This study also compared clonal Sanger sequencing and population-based Illumina sequencing. While population sequencing provides a quick, cost-effective and full view of the genomic hotspots for host range expansion, clonal sequencing provided both a roster of the most fit and frequent mutants to be isolated, and unambiguously demonstrated the ability of many single mutations to cause the expanded host range.

The second chapter tests whether generalist populations maintain higher genetic diversity with a ~150 generation evolution experiment. We compared population genetic diversity of a genotypic generalist (phi6-E8G) against genotypic specialist (wild-type phi6), and ecological generalist against ecological specialist. Generalism is a relative term, and generalist viruses are those that can infect more hosts (the one-dimensional resource niche) compared to specialists. Expanded host range strains of phi6 are genetic generalists compared to wildtype phi6, and genetic generalists can be grown in either a single host (ecological specialist) or a more heterogeneous passaging scheme of two hosts (ecological generalists). Generalists have been observed to be more evolvable than specialists (Dennehy et al., 2013; Kassen, 2002; Ketola et al., 2013); in fact, most emerging and re-emerging pathogens are most likely to be host generalists (Woolhouse and Gowtage-Sequeria, 2005). This evolvability suggests generalists have higher genetic diversity than specialists. In viruses, genetic diversity is a frequent proxy for fitness (Arenas et al., 2016; Rodriguez-Roche et al., 2016), evolvability (Mas et al., 2010; Mas et al., 2004) and virulence (Lauring and Andino, 2010; Pita and Roossinck, 2013). Computational simulations have suggested that ecological generalists maintain higher population genetic diversity (Haydon and Woolhouse, 1998; Morand et al., 1996; Nowak et al., 1991), but there have been no experimental tests of this link. Some experimental results have shown that ecological specialists have reduced levels of genetic diversity in bacterial populations (Buckling et al., 2003). To experimentally address this question, I passaged different phi6 populations under four ecological schemes: a specialist genotype in a specialist ecology, a generalist genotype in a specialist ecology and a generalist genotype in two different generalist ecology involving alternating host passaging. Each of these four treatments had four replicates, and the diversity of the 16 lineages was measured through Illumina sequencing and SNP calling. After controlling for sequencing noise, I compare the relative number and frequency of SNPs among the populations. The results show genotypic generalists do not have any advantage in population genetic diversity compared to genotypic specialist under the same passage scheme. Generalist ecology also generated less population genetic diversity than specialist ecology. We attribute this to adaptation to novel hosts in the alternating host passaging scheme; the selective process purged population genetic diversity in the populations experiencing generalist ecology. Therefore, we came to the conclusion that ecological history is a big determinant of population genetic diversity.

The third chapter is a review of CRESS DNA viruses, accepted for publication at Advances in Virus Research, that covers the changing systematics, ecology and shared evolutionary history of this diverse and increasingly numerous group. Within the last decade, the number of CRESS DNA viruses has more than doubled. The number of families increased from three (*Circoviridae*, *Geminiviridae*, and *Nanoviridae*) to six (*Bacilladnaviridae*, *Genomoviridae*, and *Smacoviridae*), not including two satellite families (*Alphasatellitidae* and *Tolecusatellitidae*) that group with the CRESS DNA viruses in Rep sequences. What was previously thought to be a group existing in low abundance in nature now appears to be ubiquitous – on all continents and in all large bodies of water.

The fourth chapter is on protein evolution of the CRESS DNA virus Rep, spanning a much longer evolutionary time than the experimental chapters. Despite substantial improvements in CRESS DNA virus systematics, it is still not clear how these groups are related to each other, and their evolutionary history is poorly described. We constructed a genealogy of the Rep protein, the only homologous ORF among these diverse viruses. The genealogy was constructed using a carefully validated Rep-specific amino acid substitution matrix (CRESS matrix), inspired by the contribution of other specific matrices, such as rtREV and FLU, to accurate viral phylogenetic resolution. Not only was the CRESS DNA matrix the best-fitting for the CRESS DNA viral Rep alignment, almost all CRESS DNA viral family capsid protein alignments preferred CRESS matrix to other matrices. Even the NS1/Rep protein alignment of linear ssDNA parvoviruses preferred CRESS matrix. This indicates that the CRESS matrix has captured ssDNA viral evolutionary patterns and will be widely applicable beyond our Rep genealogy.

The fifth chapter delves into paleovirology, finding evidence of viruses that might have integrated into host genomes over millions of years ago. Without a clear understanding of the mechanism by which CRESS DNA viruses might integrate into eukaryotic genomes, sequences homologous to CRESS DNA viruses are frequently found in various plants and animal hosts. An exhaustive blast search with relaxed criteria through the eukaryotic nucleotide databases for CRESS DNA viral Reps revealed endogenous Rep sequences in 163 unique species from 24 different phyla in the eukaryotic tree of life. We confirmed previous findings such as geminivirus-like sequences found in *Dioscorea* and *Nicotiana* species and were able to greatly expand the number of species with endogenous Rep sequences compared to previous searches. Endogenous sequences suggested potential past and present susceptible host of viruses. This is especially important to the CRESS DNA viruses, which are often isolated without a known host. The *Genomoviridae* Rep and endogenous sequence tree clearly showed that fungal plant pathogens to be a major group of hosts for uncharacterized genomoviruses, aligning well with the only characterized member of genomovirus being isolated also on a fungal plant pathogen.

This dissertation threaded questions in virus evolution using time, a dimension that is more prominent and revealing for fast-evolving viruses. I studied significant and defining interactions within several generations of evolution and telltale historical imprints of millions of years of evolution. The following chapters will relate in detail the findings from each project.

References

Abraham, A.D., Varrelmann, M., Josef Vetten, H., 2012. Three Distinct Nanoviruses, One of Which Represents a New Species, Infect Faba Bean in Ethiopia. Plant Disease 96, 1045-1053.

Allan, G.M., Mc Neilly, F., Meehan, B.M., Kennedy, S., Mackie, D.P., Ellis, J.A., Clark, E.G., Espuna, E., Saubi, N., Riera, P., Bøtner, A., Charreyre, C.E., 1999. Isolation and characterisation of circoviruses from pigs with wasting syndromes in Spain, Denmark and Northern Ireland. Veterinary Microbiology 66, 115-123.

Arenas, M., Lorenzo-Redondo, R., Lopez-Galindez, C., 2016. Influence of mutation and recombination on HIV-1 in vitro fitness recovery. Molecular phylogenetics and evolution 94, 264-270.

Bamford, D.H., 2002. Those magnificent molecular machines: logistics in dsRNA virus transcription. EMBO reports 3, 317-318.

Bamford, D.H., 2003. Do viruses form lineages across different domains of life? Research in microbiology 154, 231-236.

Bamford, D.H., Romantschuk, M., Somerharju, P.J., 1987. Membrane fusion in prokaryotes: bacteriophage phi 6 membrane fuses with the Pseudomonas syringae outer membrane. The EMBO journal 6, 1467-1473.

Bamford, J.K., Bamford, D.H., Li, T., Thomas, G.J., Jr., 1993. Structural studies of the enveloped dsRNA bacteriophage phi 6 of Pseudomonas syringae by Raman spectroscopy. II. Nucleocapsid structure and thermostability of the virion, nucleocapsid and polymerase complex. J Mol Biol 230, 473-482.

Bollback, J.P., Huelsenbeck, J.P., 2009. Parallel Genetic Evolution Within and Between Bacteriophage Species of Varying Degrees of Divergence. Genetics 181, 225-234.

Breitbart, M., Rohwer, F., 2005. Here a virus, there a virus, everywhere the same virus? Trends in Microbiology 13, 278-284.

Brum, J.R., Sullivan, M.B., 2015. Rising to the challenge: accelerated pace of discovery transforms marine virology. Nature Reviews Microbiology 13, 147.

Buckling, A., Wills, M.A., Colegrave, N., 2003. Adaptation limits diversification of experimental bacterial populations. Science (New York, N.Y.) 302, 2107-2109.

Bull, S.E., Briddon, R.W., Sserubombwe, W.S., Ngugi, K., Markham, P.G., Stanley, J., 2006. Genetic diversity and phylogeography of cassava mosaic viruses in Kenya. The Journal of general virology 87, 3053-3065.

Burch, C.L., Chao, L., 1999. Evolution by Small Steps and Rugged Landscapes in the RNA Virus ϕ6. Genetics 151, 921-927.

Burch, C.L., Chao, L., 2000. Evolvability of an RNA virus is determined by its mutational neighbourhood. Nature 406, 625-628.

Butcher, S.J., Grimes, J.M., Makeyev, E.V., Bamford, D.H., Stuart, D.I., 2001. A mechanism for initiating RNA-dependent RNA polymerization. Nature 410, 235-240.

Caldentey, J., Bamford, D.H., 1992. The lytic enzyme of the Pseudomonas phage φ6. Purification and biochemical characterization. Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology 1159, 44-50.

Cathrin, P.B., Ghanim, M., 2014. Recent advances on interactions between the whitefly Bemisia tabaci and begomoviruses, with emphasis on Tomato yellow leaf curl virus, Plant Virus–Host Interaction, pp. 79-103.

Cervera, H., Lalić, J., Elena, S.F., 2016a. Effect of Host Species on Topography of the Fitness Landscape for a Plant RNA Virus. Journal of virology 90, 10160.

Cervera, H., Lalić, J., Elena, S.F., 2016b. Efficient escape from local optima in a highly rugged fitness landscape by evolving RNA virus populations. Proceedings of the Royal Society B: Biological Sciences 283, 20160984.

Chao, L., Rang, C.U., Wong, L.E., 2002. Distribution of Spontaneous Mutants and Inferences about the Replication Mode of the RNA Bacteriophage φ6. Journal of virology 76, 3276-3281.

Cleaveland, S., Laurenson, M.K., Taylor, L.H., 2001. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 356, 991-999.

Daugelavicius, R., Cvirkaite, V., Gaidelyte, A., Bakiene, E., Gabrenaite-Verkhovskaya, R., Bamford, D.H., 2005. Penetration of enveloped double-stranded RNA bacteriophages phi13 and phi6 into Pseudomonas syringae cells. Journal of virology 79, 5017-5026.

Dennehy, J.J., Duffy, S., O'Keefe, K.J., Edwards, S.V., Turner, P.E., 2013. Frequent Coinfection Reduces RNA Virus Population Genetic Diversity. Journal of Heredity 104, 704-712.

Dennehy, J.J., Friedenberg, N.A., McBride, R.C., Holt, R.D., Turner, P.E., 2010. Experimental evidence that source genetic variation drives pathogen emergence. Proceedings of the Royal Society B: Biological Sciences 277, 3113-3121.

Dessau, M., Goldhill, D., McBride, R., Turner, P.E., Modis, Y., 2012. Selective pressure causes an RNA virus to trade reproductive fitness for increased structural and thermal stability of a viral enzyme. PLoS Genet 8, e1003102.

Diehl, W.E., Lin, A.E., Grubaugh, N.D., Carvalho, L.M., Kim, K., Kyawe, P.P., McCauley, S.M., Donnard, E., Kucukural, A., McDonel, P., Schaffner, S.F., Garber, M., Rambaut, A., Andersen, K.G., Sabeti, P.C., Luban, J., 2016. Ebola Virus Glycoprotein with Increased Infectivity Dominated the 2013-2016 Epidemic. Cell 167, 1088-1098.

dos Santos, H.F., Knak, M.B., de Castro, F.L., Slongo, J., Ritterbusch, G.A., Klein, T.A.P., Esteves, P.A., Silva, A.D., Trevisol, I.M., Claassen, E.A.W., Cornelissen, L.A.H.M., Lovato, M., Franco, A.C., Roehe, P.M., Rijsewijk, F.A.M., 2012. Variants of the recently discovered avian gyrovirus 2 are detected in Southern Brazil and The Netherlands. Veterinary Microbiology 155, 230-236.

Duffy, S., Burch, C.L., Turner, P.E., 2007. Evolution of host specificity drives reproductive isolation among RNA viruses. Evolution 61, 2614-2622.

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

Duffy, S., Turner, P.E., Burch, C.L., 2006. Pleiotropic Costs of Niche Expansion in the RNA Bacteriophage φ6. Genetics 172, 751-757.

Espejo, R.T., Canelo, E.S., 1968. Properties of bacteriophage PM2: a lipid-containing bacterial virus. Virology 34, 738-747.

Etten, J.V., Lane, L., Gonzalez, C., Partridge, J., Vidaver, A., 1976. Comparative properties of bacteriophage phi6 and phi6 nucleocapsid. Journal of virology 18, 652-658.

Ferris, M.T., Joyce, P., Burch, C.L., 2007. High Frequency of Mutations That Expand the Host Range of an RNA Virus. Genetics 176, 1013-1022.

Ford, B.E., Sun, B., Carpino, J., Chapler, E.S., Ching, J., Choi, Y., Jhun, K., Kim, J.D., Lallos, G.G., Morgenstern, R., Singh, S., Theja, S., Dennehy, J.J., 2014. Frequency and fitness consequences of bacteriophage phi6 host range mutations. PLoS One 9, e113078.

Fraser, C., Lythgoe, K., Leventhal, G.E., Shirreff, G., Hollingsworth, T.D., Alizon, S., Bonhoeffer, S., 2014. Virulence and Pathogenesis of HIV-1 Infection: An Evolutionary Perspective. Science (New York, N.Y.) 343, 1243727.

Geoghegan, J.L., Holmes, E.C., 2018. The phylogenomics of evolving virus virulence. Nature Reviews Genetics, 1.

Goldhill, D., Lee, A., Williams, E.S., Turner, P.E., 2014. Evolvability and robustness in populations of RNA virus Phi6. Front Microbiol 5, 35.

Hall, A.R., Griffiths, V.F., MacLean, R.C., Colegrave, N., 2010. Mutational neighbourhood and mutation supply rate constrain adaptation in Pseudomonas aeruginosa. Science 277, 643-650.

Harding, J.C.S., Clark, E.G., 1997. Recognizing and diagnosing postweaning multisystemic wasting syndrome (PMWS). Swine Health Prod. 5, 201-203.

Hatfull, G.F., Pedulla, M.L., Jacobs-Sera, D., Cichon, P.M., Foley, A., Ford, M.E., Gonda, R.M., Houtz, J.M., Hryckowian, A.J., Kelchner, V.A., Namburi, S., Pajcini, K.V., Popovich, M.G., Schleicher, D.T., Simanek, B.Z., Smith, A.L., Zdanowicz, G.M., Kumar, V., Peebles, C.L., Jacobs, W.R., Jr., Lawrence, J.G., Hendrix, R.W., 2006. Exploring the Mycobacteriophage Metaproteome: Phage Genomics as an Educational Platform. PLoS Genet 2, e92.

Haydon, D.T., Woolhouse, M.E.J., 1998. Immune Avoidance Strategies in RNA Viruses: Fitness Continuums arising from Trade-offs between Immunogenicity and Antigenic Variability. Journal of Theoretical Biology 193, 601-612.

Hooks, C.R.R., Wright, M.G., Kabasawa, D.S., Manandhar, R., Almeida, R.P.P., 2008. Effect of banana bunchy top virus infection on morphology and growth characteristics of banana. Annals of Applied Biology 153, 1-9.

Jiang, V., Jiang, B., Tate, J., Parashar, U.D., Patel, M.M., 2010. Performance of rotavirus vaccines in developed and developing countries. Human vaccines 6, 532-542.

Johnson, M.D., 3rd, Mindich, L., 1994. Isolation and characterization of nonsense mutations in gene 10 of bacteriophage phi 6. Journal of virology 68, 2331-2338.

Juuti, J.T., Bamford, D.H., 1995. RNA Binding, Packaging and Polymerase Activities of the Different Incomplete Polymerase Complex Particles of dsRNA Bacteriophage φ6. Journal of Molecular Biology 249, 545-554.

Juuti, J.T., Bamford, D.H., 1997. Protein P7 of phage φ6 RNA polymerase complex, acquiring of RNA packaging activity by In Vitro assembly of the purified protein onto deficient particles11Edited by J.Karn. Journal of Molecular Biology 266, 891-900.

Kassen, R., 2002. The experimental evolution of specialists, generalists, and the maintenance of diversity. Journal of Evolutionary Biology 15, 173-190.

Katz, G., Wei, H., Alimova, A., Katz, A., Morgan, D.G., Gottlieb, P., 2012. Protein P7 of the Cystovirus φ6 Is Located at the Three-Fold Axis of the Unexpanded Procapsid. PLoS ONE 7, e47489.

Ketola, S., Lehtinen, J., Rousi, T., Nissinen, M., Huhtala, H., Konttinen, Y.T., Arnala, I., 2013. No evidence of long-term benefits of arthroscopic acromioplasty in the treatment of shoulder impingement syndrome: Five-year results of a randomised controlled trial. Bone and Joint Research 2, 132-139.

Koonin, E.V., Dolja, V.V., 2014. Virus World as an Evolutionary Network of Viruses and Capsidless Selfish Elements. Microbiology and Molecular Biology Reviews 78, 278.

Koonin, E.V., Senkevich, T.G., Dolja, V.V., 2006. The ancient Virus World and evolution of cells. Biology Direct 1, 29.

La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., Merchat, M., Suzan-Monti, M., Forterre, P., Koonin, E., Raoult, D., 2008. The virophage as a unique parasite of the giant mimivirus. Nature 455, 100.

Lalic, J., Elena, S.F., 2013. Epistasis between mutations is host-dependent for an RNA virus. Biology letters 9, 20120396.

Lalic, J., Elena, S.F., 2015. The impact of high-order epistasis in the within-host fitness of a positive-sense plant RNA virus. J Evol Biol 28, 2236-2247.

Lanciotti, R.S., Roehrig, J.T., Deubel, V., Smith, J., Parker, M., Steele, K., Crise, B., Volpe, K.E., Crabtree, M.B., Scherret, J.H., Hall, R.A., MacKenzie, J.S., Cropp, C.B., Panigrahy, B., Ostlund, E., Schmitt, B., Malkinson, M., Banet, C., Weissman, J., Komar, N., Savage, H.M., Stone, W., McNamara, T., Gubler, D.J., 1999. Origin of the West Nile virus responsible for an outbreak of encephalitis in the northeastern United States. Science (New York, N.Y.) 286, 2333-2337.

Laurinavicius, S., Kakela, R., Bamford, D.H., Somerharju, P., 2004. The origin of phospholipids of the enveloped bacteriophage phi6. Virology 326, 182-190.

Lauring, Adam S., Acevedo, A., Cooper, Samantha B., Andino, R., 2012. Codon Usage Determines the Mutational Robustness, Evolutionary Capacity, and Virulence of an RNA Virus. Cell Host & Microbe 12, 623-632.

Lauring, A.S., Andino, R., 2010. Quasispecies Theory and the Behavior of RNA Viruses. PLoS Pathogens 6, e1001005.

Lee, C.H., Gilbertson, D.L., Novella, I.S., Huerta, R., Domingo, E., Holland, J.J., 1997. Negative effects of chemical mutagenesis on the adaptive behavior of vesicular stomatitis virus. Journal of virology 71, 3636-3640.

Legg, J.P., Fauquet, C.M., 2004. Cassava mosaic geminiviruses in Africa. Plant molecular biology 56, 585-599.

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

Lima, E.d., Cibulski, S.P., Santos, H.F., Teixeira, T.F., Varela, A.P., Roehe, P.M., Delwart, E., Franco, A.C., 2015. Genomic Characterization of Novel Circular ssDNA Viruses from Insectivorous Bats in Southern Brazil. PLoS One 10, e0118070.

Lipsitch, M., Barclay, W., Raman, R., Russell, C.J., Belser, J.A., Cobey, S., Kasson, P.M., Lloyd-Smith, J.O., Maurer-Stroh, S., Riley, S., Beauchemin, C.A., Bedford, T., Friedrich, T.C., Handel, A., Herfst, S., Murcia, P.R., Roche, B., Wilke, C.O., Russell, C.A., 2016. Viral factors in influenza pandemic risk assessment. eLife 5, e18491.

Lobo, F.P., Mota, B.E.F., Pena, S.D.J., Azevedo, V., Macedo, A.M., Tauch, A., Machado, C.R., Franco, G.R., 2009. Virus-Host Coevolution: Common Patterns of Nucleotide Motif Usage in Flaviviridae and Their Hosts. PLOS ONE 4, e6282.

Lukashov, V.V., Goudsmit, J., 2001. Evolutionary Relationships among Parvoviruses: Virus-Host Coevolution among Autonomous Primate Parvoviruses and Links between Adeno-Associated and Avian Parvoviruses. Journal of virology 75, 2729-2740.

Makeyev, E.V., Bamford, D.H., 2000. Replicase activity of purified recombinant protein P2 of double-stranded RNA bacteriophage phi6. The EMBO journal 19, 124-133.

Mas, A., Lopez-Galindez, C., Cacho, I., Gomez, J., Martinez, M.A., 2010. Unfinished stories on viral quasispecies and Darwinian views of evolution. J Mol Biol 397, 865-877.

Mas, A., Ulloa, E., Bruguera, M., Furcic, I., Garriga, D., Fabregas, S., Andreu, D., Saiz, J.C., Diez, J., 2004. Hepatitis C virus population analysis of a single-source nosocomial outbreak reveals an inverse correlation between viral load and quasispecies complexity. The Journal of general virology 85, 3619-3626.

McBride, R.C., Ogbunugafor, C.B., Turner, P.E., 2008. Robustness promotes evolvability of thermotolerance in an RNA virus. BMC Evolutionary Biology 8, 231.

Mindich, L., 1988. Bacteriophage ϕ6: a Unique Virus Having a Lipid-Containing Membrane and a Genome Composed of Three dsRNA Segments, in: Maramorosch, K., Murphy, F.A., Shatkin, A.J. (Eds.), Advances in Virus Research. Academic Press, pp. 137-176.

Mindich, L., 1999. Precise Packaging of the Three Genomic Segments of the Double-Stranded-RNA Bacteriophage φ6. Microbiology and Molecular Biology Reviews 63, 149-160.

Mindich, L., 2004. Packaging, replication and recombination of the segmented genomes of bacteriophage Φ6 and its relatives. Virus research 101, 83-92.

Mindich, L., Lehman, J., 1979a. Cell Wall Lysin as a Component of the Bacteriophage ø6 Virion. Journal of virology 30, 489-496.

Mindich, L., Lehman, J., 1979b. Cell wall lysin as a component of the bacteriophage phi 6 virion. Journal of virology 30, 489-496.

Mindich, L., Qiao, X., Onodera, S., Gottlieb, P., Strassman, J., 1992. Heterologous recombination in the double-stranded RNA bacteriophage phi 6. Journal of virology 66, 2605-2610.

Mindich, L., Sinclair, J.F., Cohen, J., 1976. The morphogenesis of bacteriophage φ6: Particles formed by nonsense mutants. Virology 75, 224-231.

Mizuno, C.M., Rodriguez-Valera, F., Kimes, N.E., Ghai, R., 2013. Expanding the marine virosphere using metagenomics. PLoS genetics 9, e1003987.

Montville, R., Froissart, R., Remold, S.K., Tenaillon, O., Turner, P.E., 2005. Evolution of Mutational Robustness in an RNA Virus. PLoS Biology 3, e381.

Morand, S., Manning, S.D., Woolhouse, M.E.J., 1996. Parasite-Host Coevolution and Geographic Patterns of Parasite Infectivity and Host Susceptibility. Proceedings: Biological Sciences 263, 119-128.

Nayar, G.P., Hamel, A., Lin, L., 1997. Detection and characterization of porcine circovirus associated with postweaning multisystemic wasting syndrome in pigs. The Canadian Veterinary Journal 38, 385-386.

Nowak, M.A., Anderson, R.M., McLean, A.R., Wolfs, T.F., Goudsmit, J., May, R.M., 1991. Antigenic diversity thresholds and the development of AIDS. Science (New York, N.Y.) 254, 963-969.

Obbard, D.J., Dudas, G., 2014. The genetics of host-virus coevolution in invertebrates. Current opinion in virology 8, 73-78.

Parrish, C.R., 1990. Emergence, Natural History, and Variation of Canine, Mink, and Feline Parvoviruses, in: Maramorosch, K., Murphy, F.A., Shatkin, A.J. (Eds.), Advances in Virus Research. Academic Press, pp. 403-450.

Parrish, C.R., 1991. Mapping specific functions in the capsid structure of canine parvovirus and feline panleukopenia virus using infectious plasmid clones. Virology 183, 195-205.

Parrish, C.R., Holmes, E.C., Morens, D.M., Park, E.C., Burke, D.S., Calisher, C.H., Laughlin, C.A., Saif, L.J., Daszak, P., 2008. Cross-species virus transmission and the emergence of new epidemic diseases. Microbiology and molecular biology reviews : MMBR 72, 457-470.

Pepin, K.M., Domsic, J., McKenna, R., 2008. Genomic evolution in a virus under specific selection for host recognition. Infection, Genetics and Evolution 8, 825-834.

Pirttimaa, M.J., Paatero, A.O., Frilander, M.J., Bamford, D.H., 2002. Nonspecific Nucleoside Triphosphatase P4 of Double-Stranded RNA Bacteriophage φ6 Is Required for Single-Stranded RNA Packaging and Transcription. Journal of virology 76, 10122-10127.

Pita, J.S., Roossinck, M.J., 2013. Mapping Viral Functional Domains for Genetic Diversity in Plants. Journal of virology 87, 790-797.

Prangishvili, D., Forterre, P., Garrett, R.A., 2006. Viruses of the Archaea: a unifying view. Nature Reviews Microbiology 4, 837.

Qiao, X., Casini, G., Qiao, J., Mindich, L., 1995. In vitro packaging of individual genomic segments of bacteriophage phi 6 RNA: serial dependence relationships. Journal of virology 69, 2926-2931.

Qiao, X., Qiao, J., Mindich, L., 2003. Analysis of specific binding involved in genomic packaging of the double-stranded-RNA bacteriophage phi6. Journal of bacteriology 185, 6409-6414.

Qu, B.Y., Suganthan, P.N., Liang, J.J., 2012. Differential Evolution With Neighborhood Mutation for Multimodal Optimization. Evolutionary Computation, IEEE Transactions on 16, 601-614.

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., Claverie, J.M., 2004. The 1.2-megabase genome sequence of Mimivirus. Science (New York, N.Y.) 306, 1344-1350.

Raoult, D., Forterre, P., 2008. Redefining viruses: lessons from Mimivirus. Nature Reviews Microbiology 6, 315.

Reyes, A., Semenkovich, N.P., Whiteson, K., Rohwer, F., Gordon, J.I., 2012. Going viral: next-generation sequencing applied to phage populations in the human gut. Nature Reviews Microbiology 10, 607.

Ritchie, B.W., Niagro, F.D., Lukert, P.D., Steffens, W.L., Latimer, K.S., 1989. Characterization of a new virus from cockatoos with psittacine beak and feather disease. Virology 171, 83-88.

Rodriguez-Roche, R., Blanc, H., Bordería, A.V., Díaz, G., Henningsson, R., Gonzalez, D., Santana, E., Alvarez, M., Castro, O., Fontes, M., Vignuzzi, M., Guzman, M.G., 2016. Increasing

Clinical Severity during a Dengue Virus Type 3 Cuban Epidemic: Deep Sequencing of Evolving Viral Populations. Journal of virology 90, 4320-4333.

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

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

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

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

Sanjuán, R., 2010. Mutational fitness effects in RNA and single-stranded DNA viruses: common patterns revealed by site-directed mutagenesis studies. Philosophical Transactions of the Royal Society of London B: Biological Sciences 365, 1975-1982.

Sanjuan, R., Moya, A., Elena, S.F., 2004a. The contribution of epistasis to the architecture of fitness in an RNA virus. Proc Natl Acad Sci USA 101, 15376-15379.

Sanjuan, R., Moya, A., Elena, S.F., 2004b. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc Natl Acad Sci U S A 101, 8396-8401.

Sanjuán, R., Nebot, M.R., Chirico, N., Mansky, L.M., Belshaw, R., 2010. Viral Mutation Rates. Journal of virology 84, 9733-9748.

Schoemaker, N.J., Dorrestein, G.M., Latimer, K.S., Lumeij, J.T., Kik, M.J., van der Hage, M.H., Campagnoli, R.P., 2000. Severe leukopenia and liver necrosis in young African grey parrots (Psittacus erithacus erithacus) infected with psittacine circovirus. Avian diseases 44, 470-478.

Siegl, G., 1984. Canine Parvovirus, in: Berns, K.I. (Ed.), The Parvoviruses. Springer US, Boston, MA, pp. 363-388.

Simmonds, P., Adams, M.J., Benkő, M., Breitbart, M., Brister, J.R., Carstens, E.B., Davison, A.J., Delwart, E., Gorbalenya, A.E., Harrach, B., Hull, R., King, A.M.Q., Koonin, E.V., Krupovic, M., Kuhn, J.H., Lefkowitz, E.J., Nibert, M.L., Orton, R., Roossinck, M.J., Sabanadzovic, S., Sullivan, M.B., Suttle, C.A., Tesh, R.B., van der Vlugt, R.A., Varsani, A., Zerbini, F.M., 2017. Virus taxonomy in the age of metagenomics. Nature Reviews Microbiology 15, 161.

Sinclair, J.F., Tzagoloff, A., Levine, D., Mindich, L., 1975. Proteins of bacteriophage phi6. Journal of virology 16, 685-695.

Sofonea, M.T., Aldakak, L., Boullosa, L., Alizon, S., 2018. Can Ebola virus evolve to be less virulent in humans? J Evol Biol 31, 382-392.

Stitt, B.L., Mindich, L., 1983. The structure of bacteriophage phi 6: protease digestion of phi 6 virions. Virology 127, 459-462.

Streicker, D.G., Turmelle, A.S., Vonhof, M.J., Kuzmin, I.V., McCracken, G.F., Rupprecht, C.E., 2010. Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. Science (New York, N.Y.) 329, 676-679.

Taylor, L.H., Latham, S.M., Woolhouse, M.E., 2001. Risk factors for human disease emergence. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 356, 983-989.

Turner, P.E., Burch, C.L., Hanley, K.A., Chao, L., 1999. Hybrid frequencies confirm limit to coinfection in the RNA bacteriophage phi6. Journal of virology 73, 2420-2424.

Turner, P.E., Chao, L., 1998. Sex and the Evolution of Intrahost Competition in RNA Virus φ6. Genetics 150, 523-532.

Turner, P.E., McBride, R.C., Duffy, S., Montville, R., Wang, L.-S., Yang, Y.W., Lee, S.J., Kim, J., 2012. Evolutionary genomics of host-use in bifurcating demes of RNA virus phi-6. BMC Evolutionary Biology 12, 153-167.

Turner, P.E., Morales, N.M., Alto, B.W., Remold, S.K., 2010. Role of Evolved Host Breadth in the Initial Emergence of an Rna Virus. Evolution 64, 3273-3286.

Vidaver, A.K., Koski, R.K., Van Etten, J.L., 1973. Bacteriophage phi6: a Lipid-Containing Virus of Pseudomonas phaseolicola. Journal of virology 11, 799-805.

Wasik, B.R., Muñoz-Rojas, A.R., Okamoto, K.W., Miller-Jensen, K., Turner, P.E., 2016. Generalized selection to overcome innate immunity selects for host breadth in an RNA virus. Evolution 70, 270-281.

Wasik, B.R., Turner, P.E., 2013. On the Biological Success of Viruses. Annual Review of Microbiology 67, 519-541.

Whitlock, A.O.B., Peck, K.M., Azevedo, R.B.R., Burch, C.L., 2016. An Evolving Genetic Architecture Interacts with Hill–Robertson Interference to Determine the Benefit of Sex. Genetics 203, 923-936.

Wichman, H.A., Brown, C.J., 2010. Experimental evolution of viruses: Microviridae as a model system. Philosophical Transactions of the Royal Society of London B: Biological Sciences 365, 2495-2501.

Woolhouse, M.E.J., Gowtage-Sequeria, S., 2005. Host Range and Emerging and Reemerging Pathogens. Emerging Infectious Diseases 11, 1842-1847.

Youle, M., Haynes, M., Rohwer, F., 2012. Scratching the surface of biology's dark matter, Viruses: essential agents of life. Springer, pp. 61-81.

Chapter 1

Existing host range mutations constrain further emergence of RNA viruses

Abstract

RNA viruses are capable of rapid host shifting, typically due to a point mutation that confers expanded host range. As additional point mutations are necessary for further expansions, epistasis among host range mutations can potentially affect the mutational neighborhood and frequency of niche expansion. We mapped the mutational neighborhood of host range expansion using three genotypes of the dsRNA bacteriophage phi6 (wildtype and two isogenic host range mutants) on the novel host *Pseudomonas syringae* pv. *atrofaciens* (PA). Sanger sequencing of fifty PA mutant clones for each genotype and population Illumina sequencing both revealed the same high frequency mutations allowing infection of PA. Wildtype phi6 had at least nine different ways of mutating to enter the novel host, eight of which are in p3 (host attachment protein gene), and 13/50 clones had unchanged p3 genes. However, the two isogenic mutants had dramatically restricted neighborhoods: only one or two mutations, all in p3. Deep sequencing revealed that wildtype clones without mutations in p3 likely had changes in p12 (morphogenic protein), a region that was not polymorphic for the two isogenic host range mutants. Sanger sequencing confirmed that 10/13 of the wildtype phi6 clones had nonsynonymous mutations in p12 and two others had point mutations in p9 and p5 – none of these genes had previously been associated with host range expansion in phi6. We demonstrate, for the first time, epistatic constraint in an

RNA virus due to host range mutations themselves, which has implications for models of serial host range expansion.

Importance

RNA viruses mutate rapidly and frequently expand their host ranges to infect novel hosts, leading to serial host shifts. Using an RNA bacteriophage model system (*Pseudomonas* phage phi6), we studied the impact of pre-existing host range mutations on another host range expansion. Results from both clonal Sanger and Illumina sequencing show extant host range mutations dramatically narrow the neighborhood of potential host range mutations compared to wildtype phi6. This research suggests that serial host shifting viruses may follow a small number of molecular paths to enter additional novel hosts. We also identified new genes involved in phi6 host range expansion, expanding our knowledge of this important model system in experimental evolution.

Introduction

Emerging and re-emerging viruses that host shift to infect new species pose significant economic and health costs to humans, animals, plants and our ecosystems (1-3). While ecological exposure is an essential part of emergence on a novel host (2), spillover infection of the novel host typically requires a host range mutation – the genetic component of host range expansion (4). These exaptive host range mutations must exist in the viral population prior to contact with the novel host, as part of the virus' standing genetic diversity (5, 6). The exact mutations and mechanisms of host shifting are intensively studied in emerging zoonotic viruses such as influenza, SARS-CoV, and Ebola virus (7, 8).

Given the high mutation rates (9), potentially large population sizes and fast replication of many emergent RNA viruses (10), they are capable of generating and maintaining substantial genetic variation (11, 12). This variation fuels adaptation, and selective sweeps leave genetic marks of past ecological history in viral genomes. These fixed mutations can alter the fitness landscape and constrain evolutionary trajectories of viruses due to epistatic interactions between mutations (13). Virus evolution is known to be shaped by epistasis, detected by both laboratory experimentation and phylogenetic analysis (14-16), and increased understanding of epistasis promises to improve our predictions of why some viral emergence events are more successful than others (17).

Some emergent viruses experience several hosts, often due to serial emergence events (18). MERS-CoV is proposed to have host jumped from its natural reservoir (bats) into camels, then later spilling over to the human population (19). Similarly, canine parvovirus jumped from infecting cats to raccoons and then jumped again to infect dogs (20). Influenza strains have also serially shifted hosts (*e.g.*, H3N8 originated from avian hosts infecting horses, and then shifting to dogs (21)). This kind of serial emergence allows for the possibility of host range mutations themselves to play a significant role in shaping the landscape of further emergence – one of the legacies of previous host use (7, 22). We used the model RNA virus, *Pseudomonas* dsRNA bacteriophage phi6, to investigate the role of extant host range mutations on further host range expansion.

Phi6 has been a popular model for understanding host range mutations and their fitness effects (5, 23, 24). However, all previous studies have exclusively looked at a wildtype genotype, replicating in its reservoir host, instead of investigating the interactions of multiple host range mutations during frequent host shifting, or serial emergence. In this study, we mapped the host range mutational neighborhoods of wildtype phi6 and two isogenic host range mutants (E8G in P3; G515S in P3) emerging in novel host *P. syringae* pv. *atrofaciens* (PA). Significant epistatic constraint was observed with both host range mutants in clonal and Illumina sequencing – only one or two mutations were found that allowed infection of PA. These mutations were a subset of the large mutational neighborhood of PA host range mutations available to wildtype phi6. Additionally, we have identified host range associated genes other than the canonical site of host range mutations, three genes on the Small segment, not previously implicated in host-shifting. Our work supports using deep sequencing to map mutational neighborhoods in future studies, though both deep sequencing and the more labor-intensive characterization of clones complemented each other. This work provides a panoramic view of host range mutational neighborhoods in phi6, while demonstrating a significant constraint imposed by host range mutation in a fast-evolving RNA virus.

Results

Mapping P3 PA mutational neighborhood

We found that PA host range mutational neighborhood is highly genotype-dependent. Fifty host range mutant plaques were isolated for each of the three genotypes (phi6-WT, phi6-E8G and phi6-G515S; TABLE 1.1) and their p3 genes for the phi6 attachment protein were Sanger sequenced**.** We only sequenced the p3 gene because P3 is the only highly accessible protein on the outside of the virion (25) and the only protein associated with phi6 host range in all previous studies (5, 23, 24). Thirty-five out of the fifty sequenced phi6-WT p3 sequences had single nonsynonymous mutations (seven unique mutations identified), two had double mutations, and thirteen had no detectable mutations on p3 gene. Forty-eight of the fifty phi6-E8G host range mutants contained one of the two single mutations present in the phi6-WT clones (A133V, S299W), the remaining two clones had double mutations: A133V and an additional nonsynonymous mutation. All 50 phi6-G515S host range mutants had the A133V mutation; 43 as a single mutation, five as double mutants and two as triple mutants. The nonsynonymous mutation A133V was the most frequent in the three tested populations: 24% in phi6-WT isolates, 96% in phi6-E8G isolates, and 100% in phi6-G515S isolates, making this the most prevalent mutation conferring infection of PA. In addition, there was a noticeable drop in diversity of single host range mutations from the phi6-WT population compared to the phi6-E8G and phi6-G515S populations, consistent with epistatic constraint on mutational neighborhood by host range mutations. We have summarized these P3 host range mutational neighborhoods on PA in a two-dimensional schematic (FIGURE 1.1).

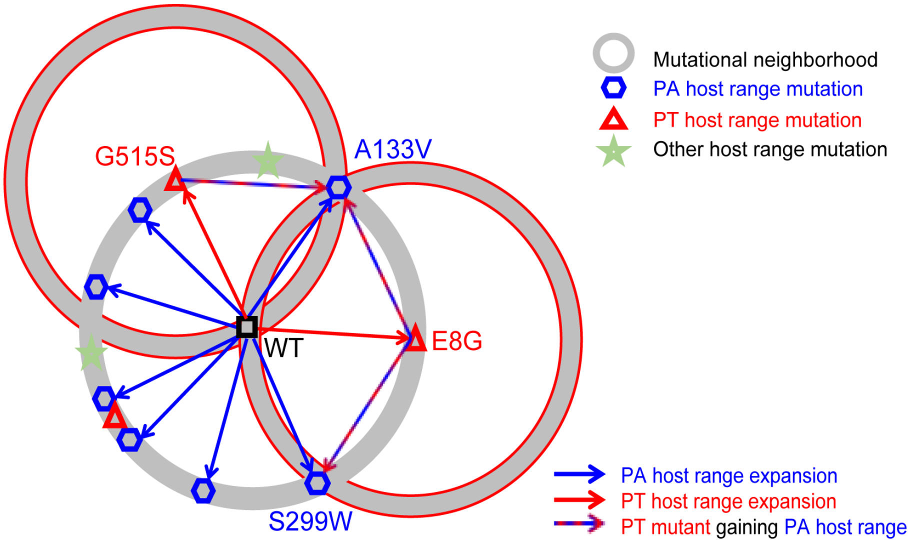


FIGURE 1.1. 2D schematic representing mutational neighborhoods of phi6 P3. Circles represent P3 mutational neighborhoods of the mutants, which are the centers of circles. The geometric shapes are known P3 mutants. It is assumed that they are distributed on the mutational neighborhood (circles) in a non-random way. The arrows are events, such as host range expansion.

Several published works have investigated the mutational neighborhoods of phi6 p3 during expansion of host range (TABLE 1.2). Two sites were favored by host range expansion events onto *P. pseudoalcaligenes* and *P.syringae* pv *glyclinea*: the 8th and 554th amino acid of attachment protein P3 (66/81 and 18/39 of isolated mutants, respectively, (5, 24)). However, the PA mutational neighborhood does not include these frequent sites of mutation to other hosts. There is some overlap with previous studies, for instance N146S (5) and A133V (23), but this suggests that phi6 may interact differently with host PA during attachment than with other *Pseudomonas* species or *P. syringae* pathovars. The absence of host range mutations in p3 for thirteen of the phi6-WT PA isolates was unexpected, since 97% of previously independently isolated host range mutants had nonsynonymous mutations in p3, with no other sites in the phi6 genome identified as causing the expanded host range in the remaining 4/118 (5, 23, 24). This motivated a more in-depth approach: deep sequencing to map the entire mutational neighborhood.

TABLE 1.1. PA host range mutations detected on P3 with Sanger sequencing. Numbers under mutations are amino acid positions.

	Mutation(s)			WT	E8G	G515S
None				13		
Singles	D35A			3		
	A133V			12	46	43
	Q140R			1		
	K144R			11		
	N146K/S			1/4		
	S299W			3	2	
Doubles	A133V	K144R			1	
		A324A				1
		E366E				1
		Q436Q				1
		L461L			1	
		S515G				1
		V606V				1
	S299W	S628A		1		
	V326F	*L147L*		1		
Triples	A133V	S515G	V531A			1
			T427T			1

TABLE 1.2. Host range mutational neighborhoods inferred from non-synonymous mutations in P3 from previous publications. Amino acids are in upper case letters with specific site in protein. Numbers indicate occurrence out of isolates sequenced. “Total” is the number of wild type isolates studied. *P. syringae* pv. *atrofaciens* and *P. syringae* pv. *glycinea* are closely related to the original host *P. syringae* pv. *phaseolicola*. While P. *pseudoalcaligenes* ERA is distantly related to the original host. (glycinea data from Ferris et al. 2007, ERA data from Ford et al. 2014)

P3 aa mutation	*P. syringae* pv. *atrofaciens*	*P. syringae* pv. *glycinea*	*P. pseudoalcaligenes* ERA
G5S		2	
E8K/G/D/A		6	52
D35A	3		
Q130R			1
A133V	12		
Q140R	1		
K144R	11		
D145G		3	
N146K/S	5	6	
E178D		2	
S299W	3		
V326F	1		
P339H		1	
T516A		4	
D533A		1	
D535N		1	
D554G/A/V/N		11	3
L555F		1	
Double/Triples	1	1	10
None	13	1	3
Total	50	40	69

Deep sequencing of phi6 populations

Each phi6 population was raised to high titer on its most recent host (phi6-WT on PP and phi6-E8G, phi6-G515S both on PT) and plated on PA to obtain a lysate made of ~400 host range mutant plaques. All population lysates from before and after PA host range expansion were sequenced. Change in Shannon entropy was calculated to determine the sites that became more or less variable after overnight growth on PA. The signals of increased variation in the p3 gene matched Sanger sequencing results; all single mutations identified in the three genotypes underwent noticeable entropy change after gaining PA host range (FIGURE 1.2). We also found several sites in P3 that may have evaded detection by clonal sampling; including amino acid 247 of phi6-WT and amino acid 35 of phi6-G515S. Deep sequencing also revealed sites of high entropy change in other genes in phi6-WT that could be additional genes controlling host range, and suggesting targets for sequencing in the phi6-WT clones that did not contain p3 mutations (FIGURE 1.3, 4). Our results suggested that non-structural protein genes p12 (encoding the morphogenic protein) and p9 (encoding the major membrane protein) were the most probable sites of additional host range mutations; both genes are involved with viral nucleocapsid vesiculation of the host inner membrane (26, 27). Results from deep sequencing the p3 gene and from the entire genome further confirmed the constrained neighborhood of host range mutants phi6-E8G and phi6-G515S revealing fewer possibilities for PA mutations in p3, and none elsewhere in the genome.

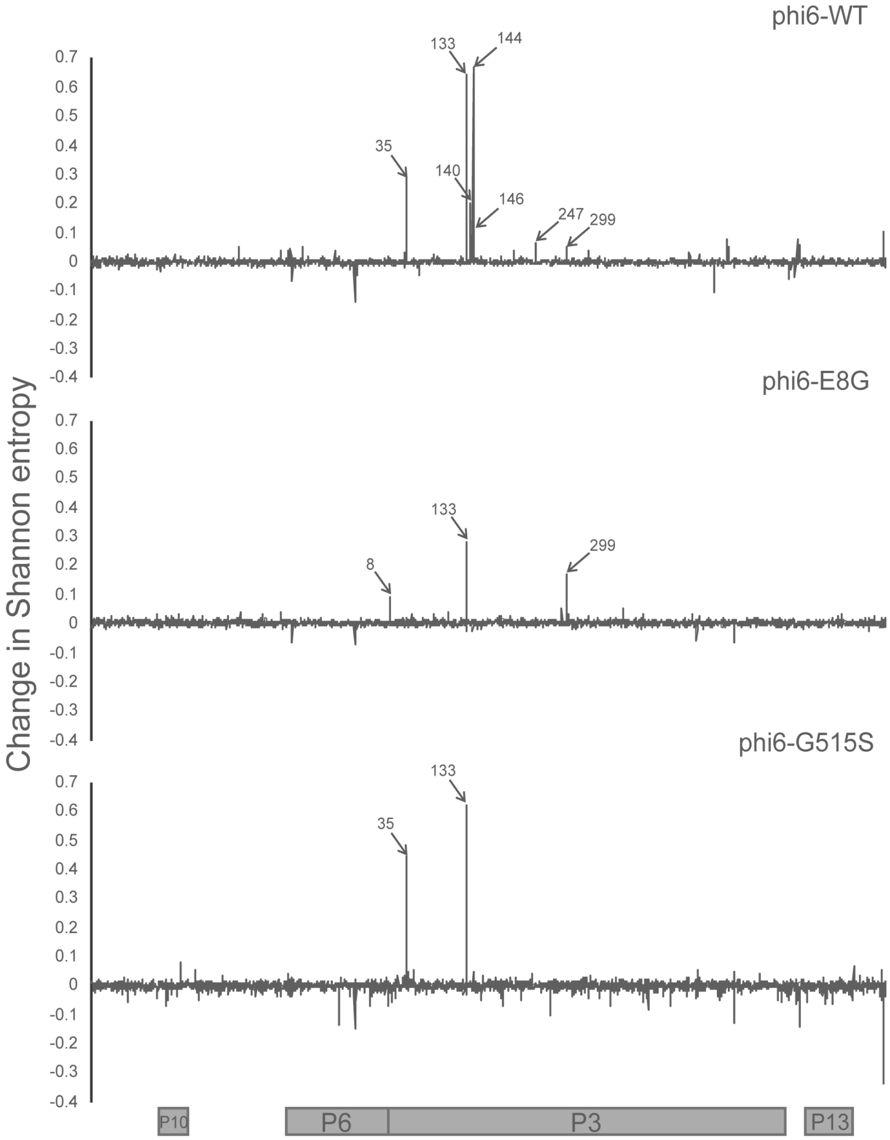


FIGURE 1.2. Change in Shannon entropy in Medium segment of phi6-WT, phi6-E8G and phi6-G515S. Positions labeled correspond to amino acid position in P3. Coding regions of the Medium segment are aligned to the graphs. The x-axis corresponds to the nucleotide positions on the Medium segment.

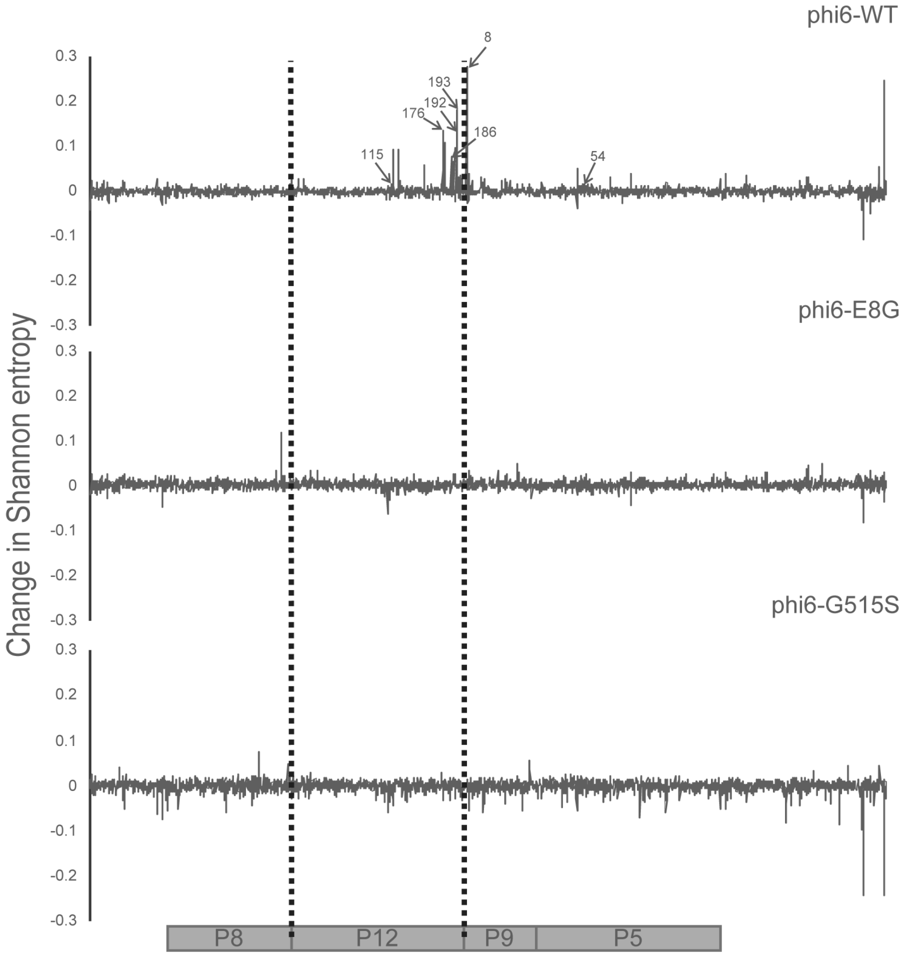


FIGURE 1.3. Change in Shannon entropy in Small Segment of phi6-WT, phi6-E8G and phi6-G515S. Coding regions of the Small segment are aligned to the graphs. Positions labeled correspond to amino acid positions in aligned genes. The x-axis corresponds to the nucleotide positions on the Small segment.

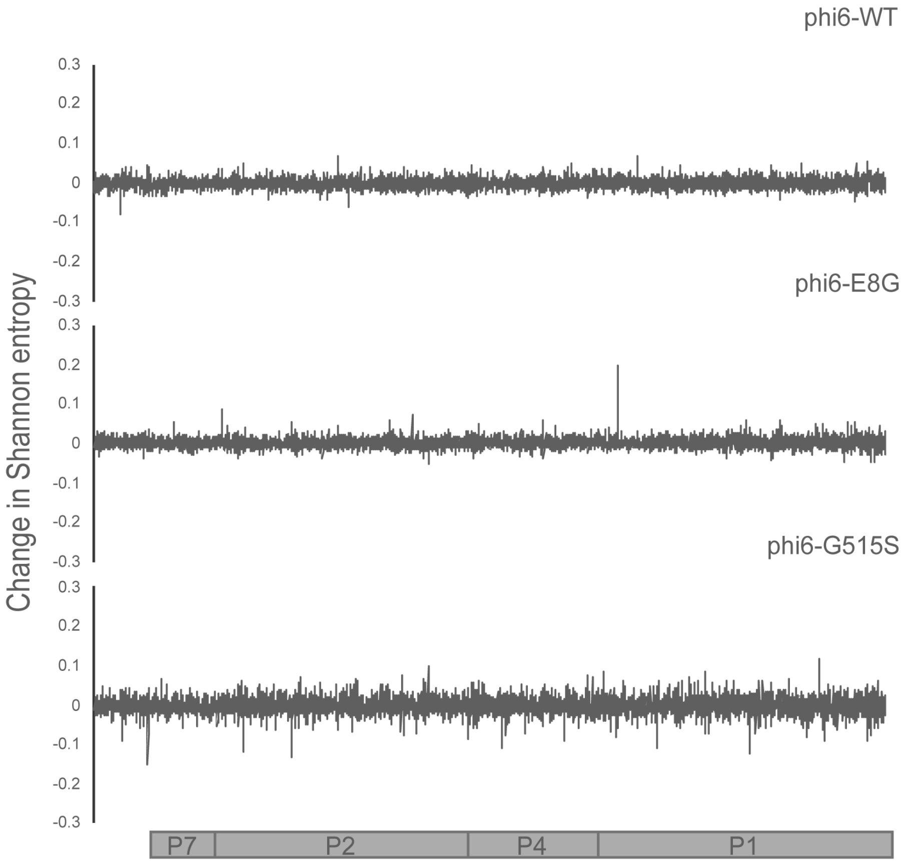


FIGURE 1.4. Change in Shannon entropy in Large segment of phi6-WT, phi6-E8G and phi6-G515S. Coding regions of the Large segment are aligned to the graphs. The x-axis corresponds to the nucleotide positions on the Large segment.

Non-p3 phi6-WT mutant sequencing

We amplified and Sanger sequenced the Small segment of all phi6-WT isolates that did not show mutation in p3 (TABLE 1.3). Ten of the thirteen isolates contained a single nonsynonymous mutation in p12. One contained a single nonsynonymous mutation in p9, another contained a nonsynonymous mutation in p5. The final mutant had a single synonymous mutation in p9. This clone was then fully Sanger sequenced, but no nonsynonymous mutations were identified. These results matched many of the sites with the highest change in Shannon entropy we observed on the Small segment of deep sequenced phi6-WT populations, and strongly suggest that mutations in these non-structural genes can affect phi6 host range.

TABLE 1.3. PA host range mutations of phi6-WT detected on the Small segment with Sanger sequencing. Numbers under mutation are amino acid positions.

gene	mutation	# of isolates
p12	K115T	1
	F176L	3
	V186L	1
	K192R	1
	D193G/A	3/1
p9	*P4P*	1
	Q8R	1
p5	K54R	1
Total		13

Estimation of PA mutational neighborhood of phi6-WT

While 49/50 phi6-WT isolates contained a non-synonymous mutation that can be correlated with PA expansion, we have not likely exhausted the complete neighborhood of PA mutations available to this genotype. Sanger sequencing detected 16 of these mutations, but the large number of mutations only observed once suggests there are other PA host range mutations that either occur at low mutation frequency or are of low fitness. Using a jackknife estimator developed for the species number problem (28) and the frequency of the PA mutations identified (TABLE 1.1 and 1.3), the most likely estimate of the true mutational number is 67 (95% confidence interval: 35-99). A similar analysis could not be conducted for the two isogenic mutants because the frequencies of their one or two possible PA mutations is inappropriate for this analysis, and suggests that we have well described their limited mutational neighborhoods.

SNPs of host range expanded phi6 populations

In addition to the sites of highest entropy change, we called SNPs present in the deep sequenced pairs of populations. The counts and details of unique synonymous and nonsynonymous SNPs are summarized in TABLE 1.4 and Appendix 1.1-1.6 in the supplemental material. An increase in detectable polymorphism was observed for phi6-E8G (paired t-test p= 0.001) after host shifting on to PA, but no significant change in SNP numbers was observed for phi6-WT and phi6-G515S (paired t-test p= 0.38, 0.40). The numbers of SNPs detected in the phi6-E8G population grown on PA were also significantly higher than that for the phi6-WT and phi6-G515S populations grown on PA (paired t-test p= 0.03, 0.0001 respectively). Gene p2 on the Large segment, coding for the RNA-dependent RNA polymerase, appears to maintain a constant, high level of diversity. The surprisingly large number of low-frequency SNPs for phi6-WT raised on PP demonstrated the potential of a dsRNA virus with a 13Kb length genome that grows ~5 generations in overnight plaque growth to generate substantial genetic diversity. This may also be the reason phi6-WT is more able to readily infect PA with a high PA host range mutation frequency (FIGURE 1.5). The high mutation frequency of phi6-WT was apparent even after an abbreviated 4 hrs of incubation, which resulted in 100- to 1000-fold lower population sizes in the lysates (FIGURE 1.5).

TABLE 1.4. Pairwise unique SNPs above 0.1% frequency in phi6 populations detected through deep sequencing using VarScan. S for synonymous changes, NS for non-synonymous changes.

		WT				E8G				G515S			
		PP		PA		PT		PA		PT		PA	
		S	NS	S	NS	S	NS	S	NS	S	NS	S	NS
Non-Coding		14		20		4		29		7		8	
Coding		23	44	19	54	8	20	39	67	7	24	12	16
Small	P8	2	4	0	1	0	1	2	3	0	0	2	0
	P12	0	1	8	14	0	1	4	5	1	0	1	4
	P9	0	3	3	7	0	1	4	5	0	0	1	0
	P5	2	4	0	1	0	1	2	3	1	1	0	2
Medium	P10	0	1	0	2	0	0	1	1	1	0	0	1
	P6	0	2	1	1	1	0	1	6	1	3	1	1
	P3	7	13	5	17	0	2	11	21	0	3	6	6
	P13	0	1	0	1	0	0	1	1	0	0	0	0
Large	P7	2	3	1	0	2	2	0	1	1	2	0	0
	P2	5	7	1	4	4	6	3	10	1	7	0	1
	P4	0	3	0	3	1	4	3	5	0	4	0	1
	P1	5	2	0	3	0	2	7	6	1	4	1	0

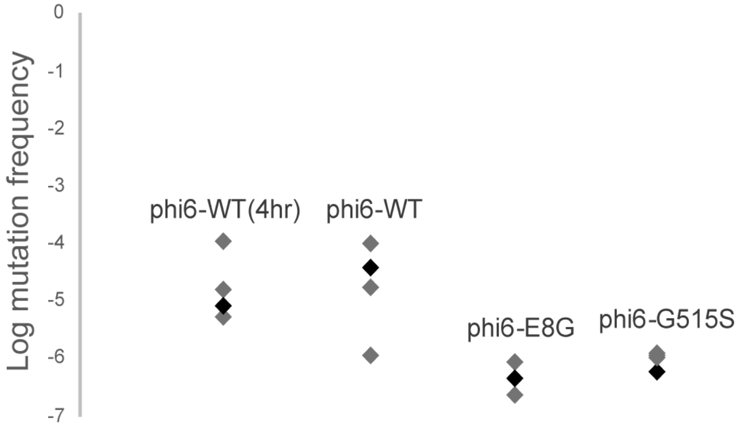


FIGURE 1.5. PA host range mutation frequency of phi6-WT(4hrs), phi6-WT, phi6-E8G, phi6-G515S. Values are measured from four purified single plaques. Plaques used for mutational neighborhood mapping are in black.

We compared the SNP results for phi6-WT to the estimated number of PA mutations obtained by jackknifing analysis. The total number of unique SNPs for phi6-WT grown on PA is 54, which is less than the estimated 67, and includes SNPs in genes such as the RNA-dependent RNA polymerase P2, which are unlikely to confer a changed host range. The nonsynonymous SNPs in the four genes now implicated in PA host range expansion (p3 on the Medium segment, p12, p9 and p5 on the Small segment) total to 39. This is within the 95% confidence interval predicted for the estimate of the complete mutational neighborhood, and suggests that some of these detected SNPs may be part of the PA host range mutational neighborhood.

Relative fitness of naturally occurring phi6 PA host range mutants

A single mutation (A133V) was shared across the three genotypes, which prompted us to look at its fitness effects in the three genetic backgrounds. We used paired growth assays to measure the relative fitness of host range mutants on their shared, original host PP. Phi6-WT, the ancestor of all the tested strains, was the common competitor for all mutant genotypes and therefore has the relative fitness of 1 (FIGURE 1.6). Relative fitness was not affected when phi6-WT obtained A133V mutation on p3 (two-tailed one sample t-test, p= 0.66). However, when phi6-E8G and phi6-G515S gained the A133V mutation, each significantly increased their PP fitness on PP (Tukey’s HSD adjusted p= 0.000011, 0.011, respectively). K144R, on the other hand, was not beneficial in two genotypes on the PP host (p< 0.005).

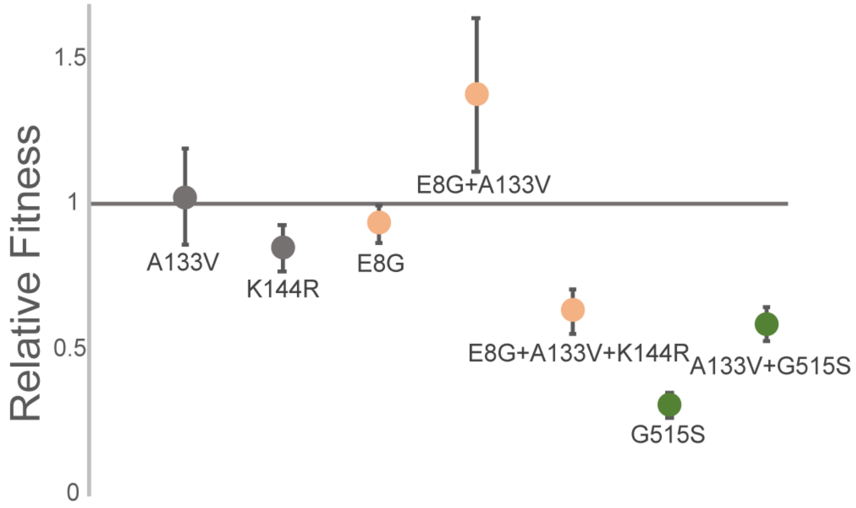


FIGURE 1.6. Relative fitness of host range mutants on PP. Same color genotypes share the same genetic background (phi6-WT: grey; phi6-E8G: peach; phi6-G515S: green). Values are averages of six replicates each, error bars are standard deviations.

Discussion

Epistasis plays a large role in viral evolution, in part because viral proteins are often highly interactive, multifunctional, and many viral genomes are of limited size (29-31). The constraining effects of larger and smaller beneficial mutational neighborhoods were elegantly demonstrated in phi6 by Burch and Chao, who noted that the high mutation rate of phi6 was not sufficient to allow constrained genotypes to traverse a rugged fitness landscape (Burch and Chao, 2000). The ruggedness of viral fitness landscapes has been demonstrated for many viruses including HIV (30), Influenza virus A (32, 33), and Ebola virus (34), and the phenomenon of mutational neighborhoods constraining evolutionary trajectories has been demonstrated in cellular organisms as well (35-37). Many of these studies involved prolonged experimental evolution, whereas our study used a very narrow window of lethal selection: a single night's selection on a novel host. Nevertheless, we detected strong epistatic constraint from single amino acid changes, and in an ecologically realistic scenario for a host-shifting RNA or DNA virus.

Evolvability may vary over evolutionary history (38), and we have only characterized one mutational step (PA mutation frequency) for these three phage genotypes. The genotypes have differences in mutational supply affected by the size of the PA mutational neighborhood more than overall population size (35). Moderate differences in population size are a concern because phi6-WT was grown on the highly productive host PP and the two mutants were grown on PT, a host on which they are less productive, but an abbreviated incubation for phi6-WT still produced the same high frequency of PA mutations (p=0.35). The two mutants could not be grown on PP without risking reversion of their PT host range mutations; G515S is highly deleterious on PP (23) and populations grown on PP quickly revert back to glycine at this position (unpublished data from our lab). Nonetheless, the library preparation for Illumina sequencing involved the same amount of RNA, isolated from 24 hrs growth of twice-purified plaque freezer stocks, enforcing a similar population size of genomes sampled by sequencing. Phi6-WT had double the number of SNPs in high titer lysates in deep sequencing following double-plaque purification, demonstrating that constrained mutational neighborhood for the two host range mutant genotypes played a large role in the reduced evolvability on PA.

Further, our results do not necessarily mean that phi6-E8G and phi6-G515S are trapped on their fitness landscapes with regard to host range expansion on PA. If these mutants were allowed to evolve further (on the original host or within their novel host range), it is difficult to predict if their evolved descendants would face identical constraints when infecting the PA host – which more accurately reflects the serial host jumping we have observed for mammalian viruses (39). It is easy to imagine that a phi6 P3 protein, somewhat destabilized by the addition of one host range mutation, cannot tolerate further destabilization while retaining its structure and function. The fitness benefits of A133V on the original host (PP) for both phi6-E8G and phi6-G515S may indicate that this is one of few (or the only) PA host range mutation in P3 that improves the mutants' P3 structure and function. Other compensatory mutations acquired over evolutionary time could stabilize the P3 of descendants of phi6-E8G and phi6-G515S, creating larger mutational neighborhoods for PA host range expansion. However, if growth on PT in and of itself affects PA host shifting, for example, through "maternal effects" -- epigenetic effects due to the host used to generate a high titer phi6 lysate, which could cause further differences in PA plaquing efficiency among phi6 strains grown on different hosts (23), then that constraint may continue to affect even derived genotypes of the two isogenic mutants.

Specific mutations found in our study

The most common means for a mutant to adapt is through additional (potentially compensatory) mutations rather than reversion of mutation (40). We observed three clones in the phi6-G515S population reverting (S515G) while fixing a PA host range mutation. This suggests that the G515S mutation is a relatively deleterious mutation to maintain in the genome when it is not grown on PT, which is bolstered by the low fitness of phi6-G515S on PP (FIGURE 1.6).

While the Small segment has not previously been associated with host range, one of our PA host range mutations (P12:F176L) was previously observed in a phi6 evolution experiment on a different novel host (41). Changes in the Small segment-encoded protein P5 are known to affect phi6 thermal niche expansion (42), but this is the first time that membrane protein P9, enveloped lytic protein P5 and membrane morphogenic protein P12 (which is not found within the virion) were associated with host range. Across diverse viruses it is not uncommon for a variety of proteins in the envelope (such as P9) to interact with host receptor proteins (6, 43). It is more rare for non-structural genes that are not on the exterior of the virion to be a host range determinant, but there are examples of this: PB2 of avian influenza (44, 45) and in picornaviruses (46, 47). While P5 plays a role in both phi6 entry and egress (48-50), P12 is solely associated with egress from cells (forming membranes from the host cytoplasmic membrane around completed nucleocapsids) and is not detected in phi6 virions (27, 51). P12 was a hotspot of change in entropy for phi6-WT and 10 of the 13 clones that did not have a mutation in P3 had one of five nonsynonymous mutations in P12, which strongly suggests that this non-structural protein has a role in host range expansion to PA. Given our current understanding of P12's role in the phi6 life cycle (26), this would require phi6-WT to be able to attach and infect PA but fail to show its infection through a plaque assay – and then mutations in a number of genes could boost the infectivity of phi6 to cause successful plaque formation. As phage host range is a difficult and debated phenotype to measure (52), and plaque formation is known to be affected by many genetic and environmental factors (https://www.dairyscience.info/index.php/enumeration-of-lactococcal-bacteriophages/factors-affecting-plaque-formation.html), this could well be the scenario for phi6-WT on hosts closely related to its original host PP, which express the same attachment site: the type IV pilus (53). It is known that DNA phage can attach to more hosts than they can productively infect, often due to successful host defense mechanisms such as CRISPR-

Cas, restriction endonucleases and suicide of the infected cell prior to phage maturation (54); our RNA phage which are not known to trigger abortive infection or be susceptible to these defenses may still enter hosts they cannot productively exit. This suggests interesting follow up experiments with phi6-WT and hosts considered outside of its current host range. While spot plating is considered a sensitive method for detecting phage host range (55), lack of visible plaques does not definitively mean that phage cannot productively infect a given host at a low level or at a slow pace (56). It further suggests some of the host range mutation observed here may be better categorized as mutations that aid in spread among novel hosts rather than the attachment mutations that allow for a spillover infection, which are often considered separate steps in the emergence of a virus on a novel host (57). On a practical level, one or more of the PA-host range associated mutations in p12, p9 or p5 may prove useful as a Small segment marker for genetic crosses in phi6. The existing mutational markers on the Small segment are an easily reverted temperature sensitivity and an unstable genetic insertion (58, 59).

Contrasting clonal sequencing and population deep-sequencing

Next generation sequencing is now more frequently applied to microbial experimental evolution studies, changing how microbial populations are monitored and analyzed, focusing mostly on relative variant frequencies and their fitness effects (60-63). However, when determining population diversity structure, many studies still use cloning for isolate sequencing, and examining chromatograms to describe nucleotide polymorphism (64-66). Increasingly, studies have exclusively used deep-sequencing for population SNP detection (67, 68). In this study, both clonal sequencing and population deep sequencing had merits and shortcomings. Clonal mapping of mutational neighborhoods with 50 clones involved relatively small sample sizes but allowed unambiguous identification of single, double, and triple mutant combinations. Illumina sequencing of populations provided a more reassuringly complete picture of the mutational neighborhood – highly consistent with that of the clonal sequencing – but it might recover hitchhiking mutations that might not be responsible for the phenotype of interest, and cannot assign combinations of mutations from different segments to a single genome.

Using the change in Shannon entropy provided more accurate data analysis because it allowed us to cancel out a large amount of the noise produced by potential sequencing errors, and is ideal for our study's purpose, which is contrasting populations before and after a challenge. However, we found we could not rely on entropy signals as estimates for mutations frequency or abundance. Although the top three most frequently observed P3 mutations in phi6-WT showed the largest change in entropy, this pattern does not apply to other observed host range mutations (e.g. phi6-WT P3: position 140, phi6-G515S P3: position 35, phi6-WT P9: position 8). Population deep sequencing provided an excellent snapshot of the mutational neighborhood, but it prevents many downstream analyses (including any further experimentation with clones of interest). However, it is cheaper and faster than clonal isolation and will serve the needs of many researchers, especially in studies of host range mutations for emerging disease surveillance.

Material and Methods

Strains and culture conditions

Wildtype phi6 (ATCC no. 21781-B1) and its standard laboratory host: *P. syringae* pathovar *phaseolicola* (PP) strain HB10Y (ATCC no. 21781) were originally obtained from American Type Culture Collection (ATCC, Bethesda, MD). These, along with novel hosts *P. syringae* pathovar *atrofaciens* (PA), *P. syringae* pathovar *tomato* (PT), *P. pseudoalcaligenes* East River isolate A (PE) were streaked from glycerol stocks originally obtained from G. Martin (Cornell University, Ithaca, NY), and L. Mindich (Public Health Research Institute, Newark, NJ) as described in previous studies (23, 69). Previously isolated isogenic host range mutants phi6-E8G and phi6-G515S (with E8G and G515S mutation on the host attachment protein P3, respectively) (23) were used to examine genome-wide mutations on host expansion. Both host range mutants can infect PP, PT and PE. Bacteria were grown in LC media (LB broth pH 7.5), 25°C. Phages were grown with bacteria in 3mL 0.7% agar top layer on 1.5% agar plates as previously described (23).

Mutational neighborhood mapping

Twice-plaque-purified phi6-WT, phi6-E8G, and phi6-G515S were raised to high-titer lysates on their respective hosts (i.e. phi6-WT was grown on PP while phi6-E8G and phi6-G515S on PT), and titered on their respective hosts. All lysates were tested for existent PA host range by spot plating approximately 10^4-10^5 plaque forming units (pfu) on a lawn of PA before plating to select for host range mutants. At least 10^6 pfu of phage were plated on novel host PA to isolate one host range mutant per lysate. 50 single plaques were isolated from each lysate by plating on PA. Five of the phi6-G515S PA host range mutant plaques were isolated by the Biotechnology class of Spring 2016 at South Brunswick High School. All 150 plaques were stored in 40% glycerol at -20°C as freezer stock and generated into high-titer lysates again for further analyses.

PA mutation frequency assays

Four independent clones of twice-plaque purified phi6-WT, phi6-E8G, and phi6-G515S plaques were raised to high-titer lysates (22-24 hours overnight incubation) on host PP, PT and PT respectively. After measuring titers on these hosts, these high titer lysates were titered on PA to assess the PA mutation frequency within the population standing genetic diversity. The four clones of phi6-WT were also incubated for a shorter length of time (4 hours) to assess the PA mutation frequency in smaller populations that had not had as many chances for mutations to accumulate. One of the four clones tested for each genotype was the source of the 50 clones.

Fitness assays

Twice purified ancestral plaque freezer stocks (phi6-WT, phi6-E8G, phi6-G515S), the only E8G+A133V+K144R plaque and one representative from the isolated mutant plaques containing mutations A133V, K144R, E8G+A133V, and A133V+G515S were arbitrarily chosen to perform paired growth assays. Equal amounts of host range mutant were mixed with a common competitor (phi6-WT) in paired growth assays (PGA) (70) to test the mutant's relative fitness on PP. Ratios of host range mutant and common competitor (CC) in the mixtures were obtained by counting the pfu of the initial mix (Day0) and after 24 hours of growth (Day1). The relative fitness to the common competitor was calculated using the following formula.

$$Relative\ fitness = 10^{(\log_{10}(Day1\ mutant/Day1\ CC)-\log_{10}(Day0\ mutant/Day0\ CC))}$$

To distinguish the different genotypes, two hosts were mixed (20:1 PP:PA) to generate the bacterial lawn as phi6-WT can only infect PP, which creates turbid plaques while the mutants can infect both hosts creating clear plaques. Statistical analyses of fitness data, including ANOVA and Tukey's honestly significant difference (HSD) tests, were performed in R (71).

Sanger sequencing

One microliter of the host range mutant glycerol stock was plated on a lawn of PA to generate high-titer lysates. Viral RNA was extracted from these lysates using QiaAmp Viral RNA Mini Kit (Qiagen, Valencia CA) per manufacturer guidelines. RT-PCR was conducted using SuperScript II Reverse Transcriptase (Invitrogen, now Thermo Fisher Scientific, MA) with random hexamers and KAPA Taq DNA Polymerase (Kapa Biosystems, now Roche) with primers that amplified the regions encoding P3 (host attachment) and P6 (membrane fusion) on the Medium segment. Amplified PCR products were cleaned up using EXO-SapIT (US Biological, Swampscott, MA) and Sanger sequencing was performed by Genewiz, Inc. (South Plainfield, NJ). Sequencing results were aligned and mutations identified with Sequencher 4.10.1.

Mutation estimation

We used a jackknife algorithm to estimate the true number of PA host range mutations in phi6-WT's mutational neighborhood (28). As implemented in the R package 'SPECIES' (72), this nonparametric method extrapolates the total number of types from the properties of subsamples of the observed sample (the Sanger sequencing results).

Library preparation

Phi6-WT, phi6-E8G and G515S were raised on their most recent host to obtain high-titer lysates, as described above. Each high-titer lysate was diluted and plated on PA to obtain a plate comprised of ~400 plaques each. Each plaque was estimated to contain at least 10^6 pfu. These plates were harvested to make lysates of phi6-WT, phi6-E8G and phi6-G515S capable of infecting PA. Viral RNA extracted using QiaAmp Viral RNA Mini Kit (Qiagen, Valencia CA) were purified by 1% low melt agarose gel electrophoresis (IBI Scientific, IA) and Gelase digestion following manufacturer instructions (GELase™, Lucigen). Individual RNA samples at a final concentration of ~15 ng/uL were prepared into Illumina RNA libraries using TruSeq RNA Library Prep Kit (Illumina, CA). Single-ended 150-cycle deep sequencing was performed on Illumina MiSeq housed in Foran Hall, Rutgers University (SEBS Genome Cooperative).

NGS data analysis

Raw reads were trimmed and filtered with cutadapt 1.12 (Q score cutoff: 30, minimum length cutoff: 75bp, adapters and terminal Ns of reads removed) (73). Then, the reads were mapped to the *Pseudomonas* bacteriophage phi6 genomes (reference sequences derived from the Illumina sequencing of phi6-WT, phi6-E8G and phi6-G515S (also confirmed with Sanger sequencing)), using BWA-MEM with default settings (74). Although approximately 33.75%-66.88% of the NGS reads mapped to *Pseudomonas* host genome, all virus genome positions had above 1000X reads coverage, with the exception of the Large segment of phi6-G515S, which had 322X to 12,058X coverage. Additional file conversion was performed using SAMtools (75). Genome nucleotide counts by position were counted with Integrative Genomics Viewer IGVTools (count options: window size 1 and --bases) (76). Whole genomes variant calling was performed using

VarScan (77). Shannon entropy was calculated for each position of the genome with the following equation:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

where n=4 for 4 nucleotides, $P(x_i)$ is the proportion of a single nucleotide over all nucleotides read at that position. Change in Shannon entropy was calculated by subtracting values before growth on PA (X_b) from values after growth on PA (X_a). Shannon entropy shows the polymorphism at each site in the genome through absolute base coverage from deep sequencing. Change in Shannon entropy shows the change in polymorphic base composition at each position in the genome.

$$\Delta\mathrm{H} = H(X_a) - H(X_b)$$

When comparing levels of polymorphism between populations directly, the SNPs in each protein-coding gene were considered as independent observations for a paired t-test (Microsoft Excel, Redmond, WA). Raw sequencing data are available through BioProject accession number PRJNA485986.

Acknowledgements: This research was funded by the National Science Foundation DEB 1453241 (to SD). We thank Nicole Wagner of the SEBS Genome Cooperative and Natasia Jacko for expert technical assistance. German Lagunas Robles and the students of South Brunswick High School’s 2016 Biotechnology class isolated several of the clones used in this study.

References

1. **Woolhouse ME, Haydon DT, Antia R.** 2005. Emerging pathogens: the epidemiology and evolution of species jumps. Trends Ecol Evol **20:**238-244.
2. **Elena SF, Bedhomme S, Carrasco P, Cuevas JM, de la Iglesia F, Lafforgue G, Lalić J, Pròsper À, Tromas N, Zwart MP.** 2011. The Evolutionary Genetics of Emerging Plant RNA Viruses. Molecular Plant-Microbe Interactions **24:**287-293.
3. **Marston HD, Folkers GK, Morens DM, Fauci AS.** 2014. Emerging Viral Diseases: Confronting Threats with New Technologies. Science Translational Medicine **6:**253ps210.
4. **Hui EK.** 2006. Reasons for the increase in emerging and re-emerging viral infectious diseases. Microbes Infect **8:**905-916.
5. **Ferris MT, Joyce P, Burch CL.** 2007. High Frequency of Mutations That Expand the Host Range of an RNA Virus. Genetics **176:**1013-1022.
6. **Lounkova A, Kosla J, Prikryl D, Stafl K, Kucerova D, Svoboda J.** 2017. Retroviral host range extension is coupled with Env-activating mutations resulting in receptor-independent entry. Proc Natl Acad Sci USA **114:**E5148-E5157.
7. **Pepin KM, Lass S, Pulliam JRC, Read AF, Lloyd-Smith JO.** 2010. Identifying genetic markers of adaptation for surveillance of viral host jumps. Nat Rev Microbiol **8:**802-813.
8. **Aguas R, Ferguson NM.** 2013. Feature Selection Methods for Identifying Genetic Determinants of Host Species in RNA Viruses. PLOS Computational Biology **9:**e1003254.
9. **Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R.** 2010. Viral mutation rates. J Virol **84:**9733-9748.
10. **Wasik BR, Turner PE.** 2013. On the Biological Success of Viruses. Annual Review of Microbiology **67:**519-541.
11. **Domingo E, Sheldon J, Perales C.** 2012. Viral Quasispecies Evolution. Microbiology and Molecular Biology Reviews **76:**159-216.
12. **Pennings PS, Kryazhimskiy S, Wakeley J.** 2014. Loss and recovery of genetic diversity in adapting populations of HIV. PLoS Genet **10:**e1004000.
13. **Elena SF, Solé RV, Sardanyés J.** 2010. Simple genomes, complex interactions: Epistasis in RNA virus. Chaos: An Interdisciplinary Journal of Nonlinear Science **20:**026106.
14. **Kryazhimskiy S, Dushoff J, Bazykin GA, Plotkin JB.** 2011. Prevalence of epistasis in the evolution of influenza A surface proteins. PLoS Genet **7:**e1001301.
15. **Lalic J, Elena SF.** 2012. Magnitude and sign epistasis among deleterious mutations in a positive-sense plant RNA virus. Heredity (Edinb) **109:**71-77.
16. **Akand EH, Downard KM.** 2018. Identification of epistatic mutations and insights into the evolution of the influenza virus using a mass-based protein phylogenetic approach. Mol Phylogenet Evol **121:**132-138.

17. **Holmes EC, Rambaut A.** 2004. Viral evolution and the emergence of SARS coronavirus. Philos Trans R Soc Lond B Biol Sci **359:**1059-1065.

18. **Parrish CR, Holmes EC, Morens DM, Park EC, Burke DS, Calisher CH, Laughlin CA, Saif LJ, Daszak P.** 2008. Cross-species virus transmission and the emergence of new epidemic diseases. Microbiol Mol Biol Rev **72:**457-470.

19. **Corman VM, Ithete NL, Richards LR, Schoeman MC, Preiser W, Drosten C, Drexler JF.** 2014. Rooting the phylogenetic tree of middle East respiratory syndrome coronavirus by characterization of a conspecific virus from an African bat. J Virol **88:**11297-11303.

20. **Allison AB, Harbison CE, Pagan I, Stucker KM, Kaelber JT, Brown JD, Ruder MG, Keel MK, Dubovi EJ, Holmes EC, Parrish CR.** 2012. Role of multiple hosts in the cross-species transmission and emergence of a pandemic parvovirus. J Virol **86:**865-872.

21. **Parrish CR, Murcia PR, Holmes EC.** 2015. Influenza virus reservoirs and intermediate hosts: dogs, horses, and new possibilities for influenza virus exposure of humans. J Virol **89:**2990-2994.

22. **Bedhomme S, Hillung J, Elena SF.** 2015. Emerging viruses: why they are not jacks of all trades? Current Opinion in Virology **10:**1-6.

23. **Duffy S, Turner PE, Burch CL.** 2006. Pleiotropic Costs of Niche Expansion in the RNA Bacteriophage φ6. Genetics **172:**751-757.

24. **Ford BE, Sun B, Carpino J, Chapler ES, Ching J, Choi Y, Jhun K, Kim JD, Lallos GG, Morgenstern R, Singh S, Theja S, Dennehy JJ.** 2014. Frequency and Fitness Consequences of Bacteriophage Φ6 Host Range Mutations. PLoS One **9:**e113078.

25. **Stitt BL, Mindich L.** 1983. The structure of bacteriophage phi 6: protease digestion of phi 6 virions. Virology **127:**459-462.

26. **Johnson MD, Mindich L.** 1994. Plasmid-directed assembly of the lipid-containing membrane of bacteriophage phi 6. J Bacteriol **176:**4124-4132.

27. **Laurinavicius S, Kakela R, Bamford DH, Somerharju P.** 2004. The origin of phospholipids of the enveloped bacteriophage phi6. Virology **326:**182-190.

28. **Burnham KP, Overton WS.** 1979. Robust Estimation of Population Size When Capture Probabilities Vary Among Animals. Ecology **60:**927-936.

29. **Betancourt A.** 2010. Lack of Evidence for Sign Epistasis Between Beneficial Mutations in an RNA Bacteriophage. Journal of Molecular Evolution **71:**437-443.

30. **da Silva J, Coetzer M, Nedellec R, Pastore C, Mosier DE.** 2010. Fitness Epistasis and Constraints on Adaptation in a Human Immunodeficiency Virus Type 1 Protein Region. Genetics **185:**293-U430.

31. **Lalic J, Elena SF.** 2015. The impact of high-order epistasis in the within-host fitness of a positive-sense plant RNA virus. J Evol Biol **28:**2236-2247.

32. **Sanjuan R, Moya A, Elena SF.** 2004. The contribution of epistasis to the architecture of fitness in an RNA virus. Proc Natl Acad Sci USA **101:**15376-15379.

33. **Gong LI, Suchard MA, Bloom JD.** 2013. Stability-mediated epistasis constrains the evolution of an influenza protein. Elife **2:**e00631.

34. **Ibeh N, Nshogozabahizi JC, Aris-Brosou S.** 2016. Both Epistasis and Diversifying Selection Drive the Structural Evolution of the Ebola Virus Glycoprotein Mucin-Like Domain. J Virol **90:**5475-5484.

35. **Hall AR, Griffiths VF, MacLean RC, Colegrave N.** 2010. Mutational neighbourhood and mutation supply rate constrain adaptation in Pseudomonas aeruginosa. Proceedings of the Royal Society B-Biological Sciences **277:**643-650.

36. **Roles AJ, Rutter MT, Dworkin I, Fenster CB, Conner JK.** 2016. Field measurements of genotype by environment interaction for fitness caused by spontaneous mutations in Arabidopsis thaliana. Evolution **70:**1039-1050.

37. **Jerison ER, Kryazhimskiy S, Mitchel JK, Bloom JS, Kruglyak L, Desai MM.** 2017. Genetic variation in adaptability and pleiotropy in budding yeast. Elife **6:**e27167.

38. **Barrick JE, Kauth MR, Strelioff CC, Lenski RE.** 2010. Escherichia coli rpoB Mutants Have Increased Evolvability in Proportion to Their Fitness Defects. Mol Biol Evol **27:**1338-1347.

39. **Turner PE, Morales NM, Alto BW, Remold SK.** 2010. Role of Evolved Host Breadth in the Initial Emergence of an Rna Virus. Evolution **64:**3273-3286.

40. **Crill WD, Wichman HA, Bull JJ.** 2000. Evolutionary reversals during viral adaptation to alternating hosts. Genetics **154:**27-37.

41. **Turner PE, McBride RC, Duffy S, Montville R, Wang L-S, Yang YW, Lee SJ, Kim J.** 2012. Evolutionary genomics of host-use in bifurcating demes of RNA virus phi-6. BMC Evol Biol **12:**153.

42. **Goldhill D, Lee A, Williams ES, Turner PE.** 2014. Evolvability and robustness in populations of RNA virus Phi6. Frontier Microbiology **5:**35.

43. **Takeuchi Y, Akutsu M, Murayama K, Shimizu N, Hoshino H.** 1991. Host range mutant of human immunodeficiency virus type 1: modification of cell tropism by a single point mutation at the neutralization epitope in the env gene. J Virol **65:**1710-1718.

44. **Subbarao EK, London W, Murphy BR.** 1993. A single amino acid in the PB2 gene of influenza A virus is a determinant of host range. J Virol **67:**1761-1764.

45. **Min JY, Santos C, Fitch A, Twaddle A, Toyoda Y, DePasse JV, Ghedin E, Subbarao K.** 2013. Mammalian adaptation in the PB2 gene of avian H5N1 influenza virus. J Virol **87:**10884-10888.

46. **Yin FH, Lomax NB.** 1983. Host range mutants of human rhinovirus in which nonstructural proteins are altered. J Virol **48:**410-418.

47. **Pacheco JM, Henry TM, O'Donnell VK, Gregory JB, Mason PW.** 2003. Role of nonstructural proteins 3A and 3B in host range and pathogenicity of foot-and-mouth disease virus. J Virol **77:**13017-13027.

48. **Bamford DH, Romantschuk M, Somerharju PJ.** 1987. Membrane fusion in prokaryotes: bacteriophage phi 6 membrane fuses with the Pseudomonas syringae outer membrane. The EMBO journal **6:**1467-1473.

49. **Caldentey J, Bamford DH.** 1992. The lytic enzyme of the Pseudomonas phage φ6. Purification and biochemical characterization. Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology **1159:**44-50.

50. **Pei J, Grishin NV.** 2005. The P5 protein from bacteriophage phi-6 is a distant homolog of lytic transglycosylases. Protein Sci **14:**1370-1374.

51. **Sinclair JF, Tzagoloff A, Levine D, Mindich L.** 1975. Proteins of bacteriophage phi6. J Virol **16:**685-695.

52. **Hyman P, Abedon ST.** 2010. Bacteriophage Host Range and Bacterial Resistance, p 217-248, Advances in Applied Microbiology doi:10.1016/s0065-2164(10)70007-1. Academic Press.

53. **Roine E, Raineri DM, Romantschuk M, Wilson M, Nunn DN.** 1998. Characterization of Type IV Pilus Genes in Pseudomonas syringae pv. tomato DC3000. Molecular Plant-Microbe Interactions **11:**1048-1056.

54. **Labrie SJ, Samson JE, Moineau S.** 2010. Bacteriophage resistance mechanisms. Nat Rev Microbiol **8:**317-327.

55. **Xie Y, Wahab L, Gill JJ.** 2018. Development and Validation of a Microtiter Plate-Based Assay for Determination of Bacteriophage Host Range and Virulence. Viruses **10:**189.

56. **Abedon ST, Yin J.** 2009. Bacteriophage plaques: theory and analysis. Methods Mol Biol **501:**161-174.

57. **Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM.** 2005. Superspreading and the effect of individual variation on disease emergence. Nature **438:**355-359.

58. **Mindich L, Lehman J.** 1983. Characterization of phi 6 mutants that are temperature sensitive in the morphogenetic protein P12. Virology **127:**438-445.

59. **Turner PE, Burch CL, Hanley KA, Chao L.** 1999. Hybrid Frequencies Confirm Limit to Coinfection in the RNA Bacteriophage φ6. J Virol **73:**2420-2424.

60. **Acevedo A, Brodsky L, Andino R.** 2014. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature **505:**686-690.

61. **Levy SF, Blundell JR, Venkataram S, Petrov DA, Fisher DS, Sherlock G.** 2015. Quantitative evolutionary dynamics using high-resolution lineage tracking. Nature **519:**181-186.

62. **Tenaillon O, Barrick JE, Ribeck N, Deatherage DE, Blanchard JL, Dasgupta A, Wu GC, Wielgoss S, Cruveiller S, Medigue C, Schneider D, Lenski RE.** 2016. Tempo and mode of genome evolution in a 50,000-generation experiment. Nature **536:**165-170.

63. **Morley VJ, Turner PE.** 2017. Dynamics of molecular evolution in RNA virus populations depend on sudden versus gradual environmental change. Evolution **71:**872-883.

64. **Arribas M, Cabanillas L, Kubota K, Lazaro E.** 2016. Impact of increased mutagenesis on adaptation to high temperature in bacteriophage Qbeta. Virology **497:**163-170.

65. **Baym M, Lieberman TD, Kelsic ED, Chait R, Gross R, Yelin I, Kishony R.** 2016. Spatiotemporal microbial evolution on antibiotic landscapes. Science **353:**1147-1151.

66. **Presloid JB, Mohammad TF, Lauring AS, Novella IS.** 2016. Antigenic diversification is correlated with increased thermostability in a mammalian virus. Virology **496:**203-214.

67. **Cuevas JM, Willemsen A, Hillung J, Zwart MP, Elena SF.** 2015. Temporal dynamics of intrahost molecular evolution for a plant RNA virus. Mol Biol Evol **32:**1132-1147.

68. **Pauly MD, Lauring AS.** 2015. Effective lethal mutagenesis of influenza virus by three nucleoside analogs. J Virol **89:**3584-3597.

69. **Duffy S, Burch CL, Turner PE.** 2007. Evolution of host specificity drives reproductive isolation among RNA viruses. Evolution **61:**2614-2622.

70. **Chao L.** 1990. Fitness of RNA virus decreased by Muller's ratchet. Nature **348:**454-455.

71. **R Development Core Team.** 2010. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

72. **Wang J-P.** 2011. SPECIES: An R Package for Species Richness Estimation. Journal of Statistical Software **40:**1-15.

73. **Martin M.** 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal **17:**10-12.

74. **Li H.** 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint**:**arXiv:1303.3997.

75. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.** 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25:**2078-2079.

76. **Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP.** 2011. Integrative Genomics Viewer. Nature biotechnology **29:**24-26.

77. **Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK.** 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res **22:**568-576.

Chapter 2

Gauging genetic diversity of generalists: a test of genetic and ecological generalism with phi6 experimental evolution

Abstract

It has been suggested that generalist viruses, those that infect a comparatively larger host range, are more likely to emerge on new hosts. Viral genetic diversity is a measure of evolvability and has been linked to both virulence and emergence potentials. However, the literature is mixed on whether infecting a larger number of hosts (heterogeneous environment) leads to higher genetic diversity, or whether diversity is better maintained in a homogeneous environment, similar to the lifestyle of a specialist virus. Using experimental evolution with bacteriophage phi6, we directly tested whether genetic generalism (carrying a mutation that confer expanded host range) or environmental generalism (growing on multiple hosts) leads to RNA virus population showing more genetic and phenotypic variation. Sixteen evolved viral lineages were deep sequenced to provide genetic evidence for population diversity. When evolved on a single host, specialist and generalist genotypes both maintained the same level of diversity (measured by the number of SNPs above 1%, p=0.81). However, the generalist genotype evolved on a single host had higher SNP levels than generalist lineages under two heterogeneous host passaging schemes (p=0.001, p<0.001). RNA viruses' response to selection in novel hosts appears to reduce the populations' genetic diversity compared to those evolving in a single host to which the virus is already well-adapted.

Introduction

Measures of genetic diversity, including nucleotide diversity, phylogenetic distance, and quasispecies size, are frequently used as proxies for viral evolvability, making them useful metrics for understanding viral emergence (Bordería et al., 2011; Day, 2015; Dutta et al., 2008). The evolvability associated with high levels of viral genetic diversity has been observed in both laboratory and clinical settings. For instance, more diverse populations of Influenza, Hepatitis C and HIV-1 are more likely to evade antiviral therapy (Mas et al., 2010). Quasispecies complexity – a measure of genetic diversity – has been suggested to be predictive of evolvability for individual pathogens (Mas et al., 2004). Genetic diversity is also positively correlated with viral fitness (Arenas et al., 2016) and virulence in emerging viral pathogens (Lauring and Andino, 2010; Pita and Roossinck, 2013). Understanding viral population genetic diversity is therefore a crucial component in the studies of fast-evolving viruses. However, most related research focuses on clinical samples and disease outcomes (Fusaro et al., 2016; Hasing et al., 2016), and there has been a lack of basic research on genetic diversity in viral populations, where controlled experimental evolution designs are involved.

Specialism and generalism are relative terms in evolution – a generalist is able to exist and thrive in more environments than a specialist. For viruses, specialism usually restricts a virus to a single resource niche, while host generalism typically means infecting more hosts than a specialist. The wider host range is usually caused by a genetic difference, resulting in generalist genotypes. In the case of RNA viruses, often a single amino acid change can lead to a more generalist genotype (Duffy et al., 2006; Hillung et al., 2014). Specialist and generalist genotypes can each outcompete the other under different ecological conditions: the former will thrive under constant environmental selection pressure, while the latter performs better given fluctuating environmental changes (Kassen, 2002); "evolutionary rescue" of specialist could also appear when host heterogeneity favors the adaptation of specialists (Bono et al., 2015). How exactly generalists cope with ecological changes is an open area of research, and one way could be by maintaining higher levels of genetic diversity. Indeed, mathematical models imply that generalists should have higher levels of genetic diversity (Haydon and Woolhouse, 1998; Morand et al., 1996; Nowak et al., 1991), but these predictions have not been tested in viruses. Successful specialists are hypothesized to maintain less genetic diversity (Liberman and Feldman, 1986), which has been experimentally verified (Buckling et al., 2003), but the converse, for generalists, has not yet been observed.

While experimental studies that could resolve the relationship between generalism and genetic diversity have mixed results (Kassen, 2002), it remains conventional wisdom that generalist pathogens are more prepared to survive and reproduce in new environments (Dennehy et al., 2013). In fact, emerging and re-emerging pathogens are most likely host generalists (Woolhouse and Gowtage-Sequeria, 2005). Directly and indirectly selected generalist virus populations infect novel hosts with higher initial fitness than specialist populations (Turner et al., 2010), and bacterial generalists have an increased ability to invade novel environments (Ketola et al., 2013). Although these studies did not directly assess the level of genetic diversity in these evolvable generalists, these generalists all displayed features representative of higher genetic diversity such as emergence potential and higher fitness. In contrast, intensive sequencing of dengue virus as it infected multiple tissue types showed reduced genetic diversity compared to those in a single tissue (Lequime et al., 2016). Further complicating the relationship between generalism and genetic diversity is the theoretical prediction that using one resource (ecological specialism) or multiple recourses (ecological generalism) should select for different arrays of mutations, but not necessarily different levels of overall population diversity (Aguirre et al., 2009).

As the literature is mixed with predictions of the relationship between generalism and population genetic diversity, we designed an experiment to directly address this relationship. Because ecological generalist populations experience changing environmental selective pressures provided by diverse hosts, genetic diversity may be maintained at higher levels than ecological specialists. Although there have been a few studies of genetic diversity of generalists, (*e.g*., (Saito and Tojo, 2016), none have been in a controlled experiment where comparison to an isogenic specialist is possible. Using the RNA *Pseudomonas* bacteriophage phi6 as a model system, which has a high mutation rate and can have large population sizes (Elena, 2016), we examined whether generalist genotypes evolving in a single environment (ecological specialism) generated similar levels of diversity as specialist genotypes, and whether multiple host environments (ecological generalism) led to higher levels of genetic diversity. This will determine whether standing genetic diversity is a pleiotropic effect of the host range mutation conferring a generalist genotype (equal diversity for generalists evolved in a single and multiple hosts) or if ecological history is the more significant determinant of population genetic diversity. The 13kb genomes of phi6 allowed significant read depth even when multiplexing on MiSeq flow cells, and we were able to directly compare the number of SNPs, as an indicator of genetic diversity, from experimentally evolved populations. Our results showed that genetic generalism does not produce higher levels of population genetic diversity. Combining results from deep sequencing and phenotypic assays, we observed selection pressure from a generalist ecological purging instead of promoting population genetic diversity.

Material and Methods

Strains and culture conditions

Phi6 is a lipid enveloped dsRNA *Pseudomonas* bacteriophage with three segments: Small, Medium and Large (Semancik et al., 1973; Vidaver et al., 1973). The starting genotypes of phi6 are twice-plaque purified wild-type phi6 (specialist, strain #21781-B1, American Type Culture Collection, Bethesda, MD), and its isogenic host range mutant (generalist) isolated previously, with a single nonsynonymous mutation in the p3 attachment protein gene, E8G (Duffy et al., 2006). The specialist can infect *P. syringae* pathovar *phaseolicola* (P) strain HB10Y (ATCC no.

21781) and three other permissive hosts (*P. syringae* pv. *persicae*, *savastanoi*, and *tagetis* (Duffy et al., 2006)), while the generalist can infect two additional hosts (*P. syringae* pathovar *tomato* (T), *P. pseudoalcaligenes* East River isolate A (E)). One additional strain neither the specialist nor generalist could readily infect, *P. syringae* pathovar *atrofaciens* (A), was also used for host range mutation frequency assays. All of these additional *Pseudomonas* strains were streaked from glycerol stocks originally obtained from G. Martin (Cornell University, Ithaca, NY), and L. Mindich (Public Health Research Institute, New Jersey Medical School, Rutgers University, Newark, NJ) as described in previous studies (Duffy et al., 2007; Duffy et al., 2006). Bacterial overnight cultures were grown to exponential growth in LC media (LB broth, pH 7.5). Phages were raised by mixing with either 200μL PP, 50μL PT, or 5μL PE bacterial culture in 3mL 0.7% agar top layer poured on 1.5% agar LC plates. All microbial growth were incubated at 25°C. Phage lysates were stored in 40% glycerol at -80°C for additional assays as described previously (Duffy et al., 2006).

<u>Experimental evolution</u>

The phi6 specialist (S) and phi6 generalist (G) were experimentally evolved for 30 days. Specialist and generalist single plaques were first raised on their most recent host (S on P, G on T) to high titer lysates. Then the two populations were serial passaged only on P for 30 passages (S/P, G/P to designate passaging on *P. syringae* pv. *phaseolicola*). Additionally, the generalist population was passaged on two alternating host schemes: on hosts P and T (G/PT) and hosts P and E (G/PE). All passaging schemes were carried out with four replicates lineages. Comparing S/P and G/P populations allows the effect of a generalist genotype to be examined with the same ecological history. Passaging on one or alternating hosts (G/P, G/PT, G/PE) would show the effects of different ecological history on population genetic diversity. During each passage, approximately 350 plaques, each containing around 10^6 pfu (plaque forming unit), of phage were harvested, diluted and plated from previous day's incubation. Incubation time for each passage was 24 hours, which allows for ~5 generations' growth (Burch and Chao, 1999), therefore the 16 lineages were evolved for a total of ~150 generations. The bottleneck size was chosen to minimize the effect of drift and the potential for overlapping plaques, which creates competition for resources and allow gene exchange.

Phenotypic assays

Every three passages, the four G/P lineages were tested for the potential loss of T host range. Diluted phage lysates were plated with both P (200µL) and T (10µL) on the plate. A phage unable to infect T would show a turbid plaque, and one able to infect both would show a clear plaque. The numbers of clear and turbid plaques were counted. Every ten passages, all phage lineage high titer lysates (at least 10^6 pfu of phage) were plated on A to obtain the population's mutation frequency in the novel host A. These lysates were also titered on host P before and after being tested for heat shock tolerance by placing lysates in heat blocks for 5 mins at 50°C.

Fitness assays

Genotype S and G lysates from Day 0, G/PT and G/PE lysates from Day 29 and all lysates from Day 30 were tested for their relative fitness on P (descendants of the generalist were also tested on hosts T and E). Approximately 1000 pfu of phage were mixed with equal pfu of a common competitor in paired growth assays (Chao, 1990; Chao et al., 1997) to demonstrate the population's relative fitness on designated host. Six technical replicates were tested for each population. A common competitor selected in our lab for fitness assays on host T (TCC, able to infect P, T, and A) was used as common competitor in P and T fitness assays, while another common competitor (ECC, able to infect P, E, and A) was used in E host fitness assays. Ratios of phage population and common competitor in the mixtures were obtained by counting the pfu of each phage in the initial mix after 24 hours of growth. In order to distinguish the different phages, P (200µL) and A (10µL) were mixed in the bacterial lawn. While both common competitors can infect P and A, none of the passaged populations can readily infect A with 1000 pfu. Therefore, we were able to distinguish the clear plaques formed by common competitor phages, and turbid plaques formed by the populations from our study. The relative fitness to the common competitor was calculated using the following formula:

$$Relative\ fitness = 10^{\,(\log_{10}(Day1\ population/Day1\ CC)-\log_{10}(Day0\ population/Day0\ CC))}$$

Statistical analyses of fitness data, including ANOVA and Tukey's honestly significant different (HSD) tests, were performed in R (R Development Core Team, 2010).

Library preparation

Viral RNA was extracted from G/P, G/PT and G/PE lysates from Day 29, all lysates from Day 30 and S and G lysates from Day 0 using QiaAmp Viral RNA Mini Kit (Qiagen, Valencia CA). Viral RNA were purified by 1% low melt agarose gel electrophoresis (100V, 40mins) (IBI Scientific, IA) and Gelase digestion following manufacturer instructions (GELase™,Lucigen). Individual RNA samples at a final concentration of ~15 ng/uL were prepared into Illumina RNA libraries using Illumina TruSeq RNA Library Prep Kit (Illumina, CA) (Zhao et al., 2018). Viral RNA libraries were divided into four batches, each containing one Day 30 lineage from each passaging scheme, and Day29 from the corresponding G/S, G/PT, and G/PE lineages. Single-end 150-cycle sequencing was done on Illumina MiSeq housed in Foran Hall, Rutgers University (SEBS Genome Cooperative).

NGS data analysis

Raw reads were trimmed and filtered with cutadapt 1.12 (Q score cutoff: 30, minimum length cutoff: 75bp, adapters and terminal Ns of reads removed) (Martin, 2011). The cleaned up reads were either mapped to phi6 specialist and generalist genotype reference genomes, derived from Day 0 Illumina sequencing results and confirmed with Sanger sequencing, using bwa mem with default settings (Li, 2013; Li and Durbin, 2010). Although 5.74%-59.09% (median: 38.26%) of the NGS reads mapped to *Pseudomonas* host genome, S/P, G/P and G/PT population samples produced above 1000X coverage for 91%-97% of the genome. G/PE population samples had above 1000X coverage for 80%-97% of genome because 5 samples had lower than 1000X Large segment coverage. Additional file format conversion was performed using SAMtools (Li et al., 2009). Genome coverage data produced by Integrative Genomics Viewer IGVTools (count options: window size 1 and --bases) (Robinson et al., 2011). Whole genomes variant calling was performed using VarScan (default parameters with adjusted minimum coverage 500, minimum variant frequency 0.001, p-value threshold 0.05) (Koboldt et al., 2012). SNP data from VarScan were further analyzed and visualized in R (R Development Core Team, 2010) and Excel (Microsoft, Redmond, WA).

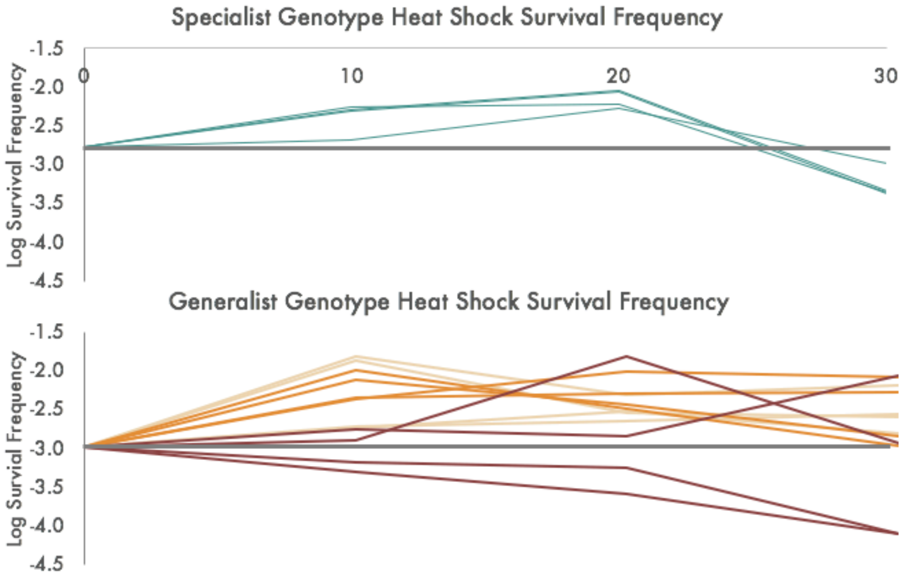
Results

Phenotypic assays

To test the effect of the generalist genotype on accumulated genetic diversity, it is important to know the G/P populations do not revert their host range mutation while being passed on single host. We monitored the loss of host range in G/P lineages by enumerating the proportion of phages unable to infect T. Results showed that the generalist genotype being passaged on single host P had a relatively stable phenotypic reversion rate with a mean of 1.1% and median of 0.7%. These rates were summarized from 10 measurements (every three passages for 30 passages) from 4 lineages. This showed that G/P populations were stable generalists throughout the experiment.

Frequency of mutations in the population to jump into a novel host or to withstand heat shock are good indicators of population genetic diversity. We measured the mutation frequency for novel host A host range expansion by plating high titer phage lysate from the passaged populations on a host A bacterial lawn. After sampling the lineages at Day 10, Day 20, and Day 30, we found that mutation frequency is dependent on initial genotype and relatively invariant over the 30 days: S/P populations, and the S ancestor, had significantly ($p=0.002$) higher mutation frequencies (mean: $1.21x10^{-4}$, range: $5.99x10^{-6}$-$6.65x10^{-4}$) than any population with a G genotype (mean: $2.68x10^{-}$

^{7},range: 0-3.71x10^{-7}) (Appendix 2.29). In contrast, heat shock tolerance survival rates fluctuated for the populations over the 30 days. Although the four lineages of S/P displayed similar survival of heat shock tolerance through experimental evolution, all experienced an uniform increase and then a drop below Day 0 values at the end of passages (FIGURE 2.1.1). The twelve populations derived from the generalist genotype showed more variation in heat shock tolerance. All G/P and G/PT lineages maintained or increased survival during experimental evolution, but two G/PE steadily lost heat shock tolerance during passaging (FIGURE 2.1.2).



Specialist Genotype Heat Shock Survival Frequency

Generalist Genotype Heat Shock Survival Frequency

FIGURE 2.1 Heat shock tolerance mutation frequency of all lineages.

2.1.1(top) Heat shock tolerance mutation frequency of four S/P lineages. Data collected every 10 days during the 30-day passaging. The S ancestor’s survival is shown as the grey horizontal line.

2.1.2(bottom) Heat shock tolerance mutation frequency of four G/P lineages (beige), four G/PT lineages (orange), four G/PE lineages (maroon). Data collected every 10 days during the 30-day passaging. The G ancestor’s survival is shown as the grey horizontal line.

Genetic diversity comparison

Population genetic diversity is well-depicted by SNPs within the genomes of the population, so we deep-sequenced Day 30 phage populations to quantify the number of SNPs. We also sequenced the Day 29 lysate of G/P lineages as internal quality control and Day 29 lysate of G/PT and G/PE to account for "maternal effect" for these lipid enveloped bacteriophages (Duffy et al., 2006). Because of host alternating passaging, G/PT and G/PE populations were grown on P on Day 29, and grown on T and E on Day 30, respectively. All raw numbers of SNPs called by VarScan were normalized by sequencing batch, which included all samples that shared the same MiSeq lane. Since our SNP cutoff was arbitrarily set to 0.1% frequency, we compared the number of SNPs above 0.1% for all sequenced populations. While most populations are indistinguishable from each other, SNPs called from G/P-29 was significantly higher than that of G/PT-30 (adj p = 0.04), and G/PE-29 (adj p = 0.001). This is statistical evidence that ecological specialism leads to higher diversity than two alternating host passaging schemes that mimicked ecological generalism. Within a single analysis, SNPs called from G/PE-29 were significantly lower than that of G/PE-30 (adj p = 0.03), indicating a loss of diversity when the Day 29 lysate was plated on host P, which showed positive selection on well adapted host P and diversification on the less adapted host E. Similar patterns were evident when we took a closer look at population SNPs with frequencies between 0.1% and 1%. Most populations were still insignificantly different from each other, and G/P-29 populations still had more SNPs than G/PT-30 (adj p = 0.08) and G/PE-29 (adj p = 0.005). In this narrow band of SNP frequency, the G/PE-29 populations had fewer SNPs than the G/PE-30 populations (adj p = 0.03), but this reflect both lower frequency SNPs being less frequent after growth on P, and that higher frequency SNPs (above 1%) move into the category of between 0.1% and 1% when they become less frequent after passage on E. SNPs with frequencies between 0.1% to 1% takes up 79.7±7.5% of all SNPs above 0.1%, therefore it is highly likely that the SNPs between 0.1% to 1% frequency are going to dictate the patterns of all

SNPs within these population samples. In fact, when we looked only at the number of higher frequency SNPs (above 1% frequency), we found that S/P populations have significantly more SNPs than G/PT and G/PE populations (adj p = 0.0002-0.02) (FIGURE 2.2). G/P populations continued to have significantly more SNPs than G/PE populations (adj p = 0.0007-0.002) and qualitatively more average SNPs than G/PT populations (adj p = 0.06-0.13). There was no significant difference in SNP counts above 1% between S/P and G/P populations, G/P-29 and G/P-30 populations, G/PT -29 and G/PT-30 populations, G/PE-29 and G/PE-30 populations, and G/PT and G/PE populations (both Day 29 and Day 30, FIGURE 2.2).

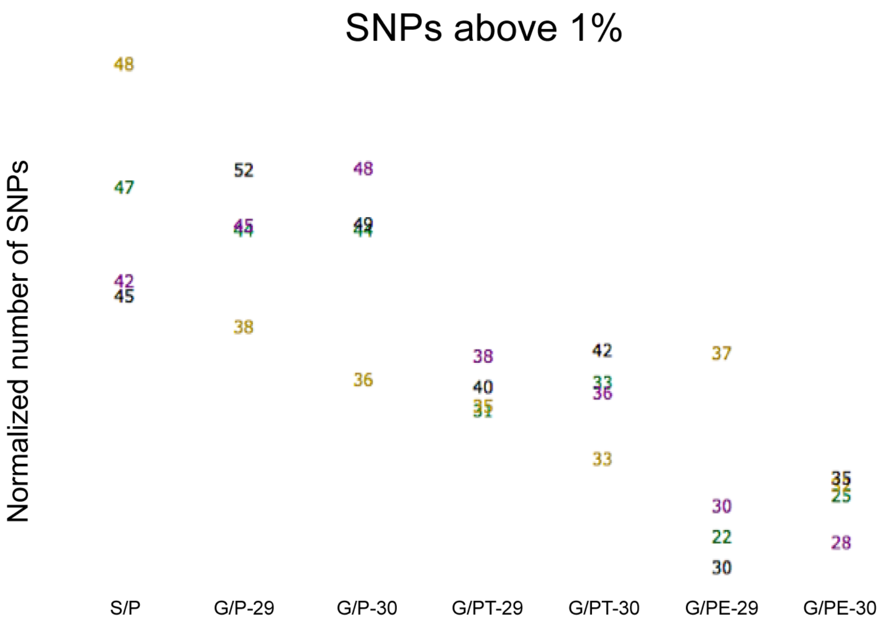


FIGURE 2.2. Number of SNPs above 1% across all lineages of all passaging schemes. Each data point is shown as the raw SNP count, and the position of the data point are normalized within samples sharing the same sequencing lane. Four colors indicate four different sequencing batches – the 1st replicate of each passaging scheme was run in one MiSeq flow cell, and so forth.

While counts at different SNP frequency cutoffs are essential for comparisons between treatments, there is value in looking at the individual SNPs. We looked at the top 10 SNPs in each lineage from Day30 (FIGURE 2.3). These most frequent SNPs were fairly stable, even in alternate host passaging. The Day 29 top 10 SNPs had an average of 8.75 overlap with Day 30 top 10 SNPs for all tested lineages; when there was an absence of SNP from Day30 compared to Day29, it was mostly due to the SNP being ranked just below the top 10. However, there were relatively few shared SNPs among replicates or between treatments. Just six examples of parallel SNPs were observed among these most frequent SNPs. Two S/P lineages shared the same SNP resulting in a nonsynonymous amino acid change S166G (12.87% and 7.83%) in P6 on the Medium segment (FIGURE 2.3). One lineage each from S/P and G/P shared the same non-coding region SNP, a830g (11.28% and 9.99%), on the Medium segment. Two G/PT lineages had the same SNP causing a nonsynonymous change Q130R (98.15% and 79.01%) in P3 on the Medium segment. Two G/PE lineages shared one SNP causing a nonsynonymous change G101C (99.73% and 99.61%) in P5 on the Small segment. And another two had one non-coding region SNP t869c (50.05% and 5.83%) on the Medium segment. One lineage each from S/P and G/PE had the same T65A (4.64% and 5.81%) in P13 on the Medium segment. As these parallel SNP frequencies and FIGURE 2.3 show, only ecological generalism (G/PT and G/PE) brought SNPs to very high frequency (highest frequency SNPs: 79.01%-99.81%). This indicates stronger positive selection occurring in the alternating host passages than in the lineages evolved only on host P. The highest frequency in single host specialist ecology passaged lineages was 54.12%, falling short of a completed selective sweep in the specialist populations.

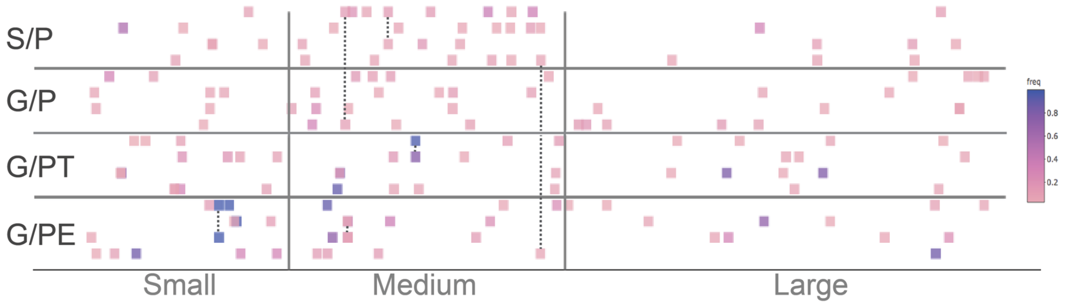


FIGURE 2.3. Top 10 SNPs within all Day 30 populations. Each square is a SNP placed according to its genomic position on the segmented phi6 genome. Color indicate frequency of the SNP. Each passaging scheme showed four population sequences, one from each lineage. From left to right and connected by dashed lines (upper case for amino acids within the indicated protein, lower case letters for non-coding region nucleotides): P5 G101C on Small segment; a830g, t869c, P6 S166G, P3 Q130R, P13 T65A on Medium segment. All other specific SNPs shown are listed in Appendix 2.1-2.28.

Evolved population fitness

After the 30-day experimental evolution, we measured the sixteen populations' fitness on the *Pseudomonas* hosts each could infect. All Day 30 populations, Day 29 of G/PT and G/PE, and Day 0 S and Day 0 G were mixed and grown with a common competitor to show their relative fitness on their shared host P (FIGURE 2.4). Because these were population relative fitness values instead of isolating and comparing single genotypes, the standard deviations among the six technical replicates were considerably large. There was no difference in fitness on host P for both G/PT and G/PE from Day 29 or Day 30 (G/PT: adj p = 0.96; G/PE: adj p = 0.67). Therefore, we combined all measurements from G/PT and all measurements from G/PE for subsequent P relative fitness comparison. G/PT populations had lower relative fitness than S/P (adj p = 0.03), G/P (adj p = 0.04), and G/PE (adj p = 0.008), while all other treatments were statistically indistinguishable.

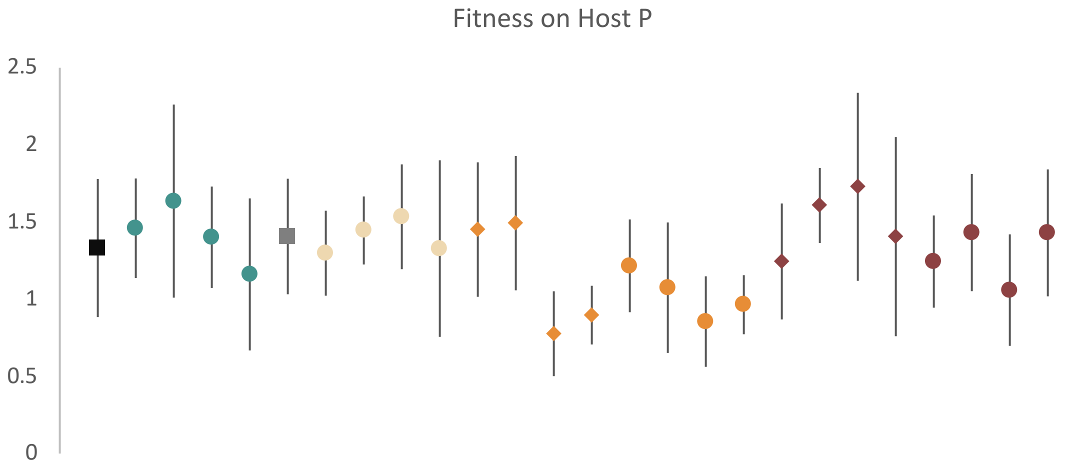


FIGURE 2.4. Relative fitness of populations on P. All Day 0 are squares, Day 30 are circles, Day 29 are diamonds. Specialist Day 0 population is black, Generalist Day 0 population is grey, S/P populations are teal, G/P populations are beige, G/PT populations are orange, G/PE populations maroon. The point values in the graphs are averages of six replicates, error bars are standard deviations.

As for the Generalist populations that are able to infect other *Pseudomonas* hosts, we tested their relative fitness on hosts T and E separately (FIGURE 2.5). There was no difference in T fitness between Day 29 and Day 30 populations for G/PT (adj $p = 0.97$) or G/PE (adj $p = 0.36$), therefore we compared the T fitness among the four passaged populations and Day 0 Generalist population. G/PT populations had a higher T fitness than Day 0 generalist population, G/S and G/PE populations (adj $p < 0.001$). Intriguingly, G/PE had higher average fitness on T than G/S (adj $p = 0.09$), even though both treatments excluded host T in their passaging. When measuring fitness on host E, yet again there was no difference in E fitness between Day 29 and Day 30 populations of G/PT (adj $p = 1$) and G/PE (adj $p = 0.59$), so we compared E fitness among the four passaged populations and Day 0 Generalist population. G/PE populations had significantly higher PE fitness values than all other populations (adj $p < 0.001$), which all had similar fitness (adj $p = 0.9$-1). In both cases, we see an increase in relative fitness for the lineages passaged on the corresponding host, but surprisingly also saw G/PE populations gain fitness on unselected host T.

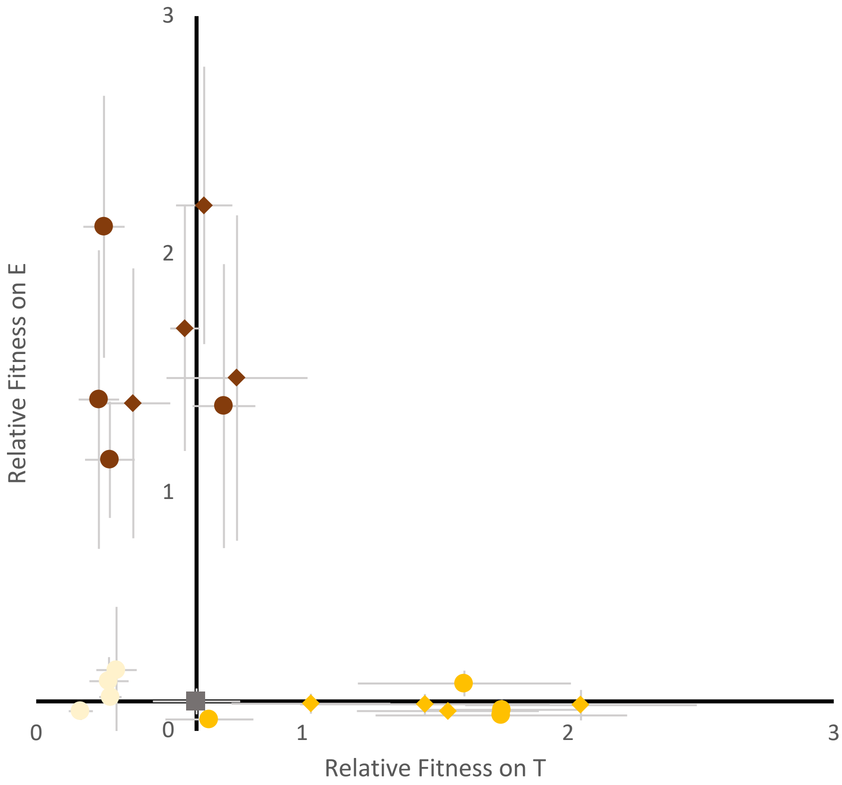


FIGURE 2.5. Relative fitness of populations on T and E. All Day 0 are shown as squares, Day 29 are diamonds and Day 30 are circles. The generalist ancestor is grey, G/P populations are beige, G/PT populations are orange, G/PE populations maroon. The graph axes are place on the coordinates of the generalist ancestor. The point values in the graphs are averages of six replicates, error bars are standard deviations.

Discussion

Elevated mutation rates and large population sizes allow RNA viruses to maintain higher levels of population genetic diversity compared to prokaryotes and eukaryotes, ensuring an ample supply of variation for potential acquisition of and adaptation to new niches. For this reason, virologists often use viral genetic diversity as a measure of evolvability, virulence and potential for emergence (Bordería et al., 2011; Day, 2015; Lauring and Andino, 2010; Pita and Roossinck, 2013). With experimental evolution, we directly compared the effect of generalist genotype and generalist ecology on RNA virus population genetic diversity. We found that neither kind of generalism led to larger population diversity. Although previous studies have shown genetic diversity to show positive correlation with host range size among divergent species sharing a common ancestor (Schneider and Roossinck, 2000), we did not observe an expanded host range mutant of wildtype phi6, repeatedly grown on the ancestral host, to had a larger amount of genetic diversity compared to wildtype phi6 in the same passaging scheme. It was also unexpected that having a generalist ecology did not lead to more genetic diversity, though others have noted that some ecological histories constrain organisms from finding the optimal genotype in variable environment (Jasmin and Kassen, 2007). Both specialist and generalist viral populations are thought to be under continuous selective pressure for higher fitness, and to combat host defenses (Kassen, 2002). However, generalist viruses are known to face more selective pressures compared to specialists as they infect different hosts with different fitness optima (Elena et al., 2009). As long as selection is present, it will purge genetic diversity within the population; even if a specialist achieves the optimal genotype, mutations might be purged by selection processes such as clonal interferences or drift (Elena et al., 2003). As phi6 is not capable of homologous recombination, diversity is reduced with each selective sweep (Amos and Harwood, 1998; Wootton et al., 2002). In our result from SNPs above 1%, the response to selection on hosts T and E in the alternating passages was shown by high frequency SNPs (2 mutations above 90% for two out of the four PT lineages; 6 mutations above 90% for all of the PE lineages), and a commensurate loss of diversity elsewhere in the genome.

The importance of ecological history in virus evolution

While the specialist was always more capable of infecting host A than the generalist, none of the populations showed changes in mutation frequency over the 30 days of evolution, and there was little variance among the replicate populations. This strong constraint of genotype on further host range expansion was previously demonstrated by our lab (Zhao et al., 2018), but over a much shorter time span of one passage. These results extend on our previous work to show that 30 days' passage, and even the fixation of other mutations in the host attachment gene P3 (*e.g.*, Q130R in G/PT) do not change the epistatic constraint of the host range mutation in the generalist on further range expansion to host A. This is evidence of past host-shifting ecological history affecting subsequent virus evolution. No such epistatic constraint was observed in heat shock survival. The replicates of the two specialist ecology treatments (S/P and G/P) behaved similarly within each treatment, but the replicates of the two generalist ecology treatments (G/PT and G/PE) showed much higher variances within treatment. The amount of variance positively correlates with selective pressure, since host T is closely related to P (lower pressure), and host E is distantly related to P (another species, higher pressure).

One form of simplification is that viruses are often exclusively passaged on novel hosts (Ciota et al., 2014; Cuevas et al., 2009), which speeds up adaptive molecular evolution (Pepin et al., 2008; Wasik et al., 2016). When hosts are alternated, seldom were natural host involved (Coffey and Vignuzzi, 2011), it is common for both of the hosts to be novel to the virus (e.g.,(Turner and Elena, 2000; Turner et al., 2010)). However, despite the significant results from these experimental designs, their results may be less applicable to real world scenarios. As many studies of viral spillover and emergence have shown evidence of complex ecological host shifts, passaging schemes involving both the novel host and the original, or reservoir, host are more realistic than having the pathogen exclusively replicate in the novel host (Allison et al., 2013; Troupin et al., 2016). Therefore, we chose to alternate our generalist lineages between the original host and one other permissive host. Despite the frequent infection of original host P slowing the speed of molecular change in the alternating host passage populations, selective sweeps still occurred in the 30 days.

Given that genotypic constraints are prevalent in virus evolutionary trajectories and complex protein interactions may result from genotypic change, genetic generalism may not in itself be linked with evolvability. Instead, ecological history appears to be the larger determinant of standing genetic diversity when evaluating fast evolving RNA virus populations. Although the level of genetic diversity might be similar, populations under hard and soft selection might experience distinct evolutionary dynamics given the contingent, historical events (Van Tienderen, 1991). And the composition and structure of population genetic diversity can be dependent on ecological aspects, as it was demonstrated computationally that selection favors a specific form of variability that maximizes evolvability under that specific environments (Draghi and Wagner, 2007). In fact, that accessible environments determine what kind of alleles are selected and maintained has been proven experimentally (Bono et al., 2016).

Future directions

The comparison between different selection pressures are not uncommon in experimental evolution, and it has been shown that populations evolving in benign environments are likely to harbor more adaptive solutions than those in harsh, higher selection pressure environments (Pepin and Wichman, 2008). It has also been observed that eliminating the initial selective pressure of a novel host or environment through pre-adaptation can simplify interpretation of subsequent evolution experiments (Domingo-Calap et al., 2009; Pepin et al., 2008; Pepin and Wichman, 2008).

Our specialist and generalist ancestors were pre-adapted to their original host P, but we did not pre-adapt the generalist to the novel hosts. Our experimental design is strongest when the two genotypes of the ancestors were as similar as possible and pre-adaptation of the generalist to the novel hosts would have made it more distantly related to the specialist when adaptive mutations sweep through the populations. However, future work aiming to only examine ecological generalism might be well-served by pre-adapting viruses to all hosts that will be used, to see if evolving on one established host or multiple established hosts has any effect on standing genetic diversity. On the other hand, our experimental design mimics how viruses expand their host range in nature. It is unlikely that a generalist genotype, newly arisen through mutation, is near the peak of multiple host fitness landscapes. At least initially generalist genotypes experiencing novel generalist ecologies likely always experience selection and its purging of genetic diversity relative to specialist ecologies. At times strong selection experienced by viruses in novel hosts shows a contrast in the application of the term generalist to viruses compared to cellular organisms like bacteria and phytophagous insects. Bacteria often have operons to toggle their metabolic enzymes, which allows the same genotype to have different proteins expressed when consuming a single or different combination of carbon sources (Dandekar et al., 2015). Similarly, generalist insects are not arising de novo in populations as much as already established species or sub-species that have had a long evolutionary time with their host range of plants (Ali and Agrawal, 2012). It would be worthwhile to test for the effect of these kinds of generalism on standing genetic diversity.

Acknowledgements: This research was funded by the National Science Foundation DEB 1453241 (to SD). The sequencing was partially funded by the Graduate School of New Brunswick, Rutgers the State University of New Jersey (award to LZ).

References

Aguirre, J., Lazaro, E., Manrubia, S.C., 2009. A trade-off between neutrality and adaptability limits the optimization of viral quasispecies. J Theor Biol 261, 148-155.

Ali, J.G., Agrawal, A.A., 2012. Specialist versus generalist insect herbivores and plant defense. Trends in plant science 17, 293-302.

Allison, A.B., Kohler, D.J., Fox, K.A., Brown, J.D., Gerhold, R.W., Shearn-Bochsler, V.I., Dubovi, E.J., Parrish, C.R., Holmes, E.C., 2013. Frequent cross-species transmission of parvoviruses among diverse carnivore hosts. J Virol 87, 2342-2347.

Amos, W., Harwood, J., 1998. Factors affecting levels of genetic diversity in natural populations. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 353, 177.

Arenas, M., Lorenzo-Redondo, R., Lopez-Galindez, C., 2016. Influence of mutation and recombination on HIV-1 in vitro fitness recovery. Molecular phylogenetics and evolution 94, 264-270.

Bono, L.M., Gensel, C.L., Pfennig, D.W., Burch, C.L., 2015. Evolutionary rescue and the coexistence of generalist and specialist competitors: an experimental test. Proceedings of the Royal Society B: Biological Sciences 282.

Bono, L.M., Smith, L.B., Pfennig, D.W., Burch, C.L., 2016. The emergence of performance trade-offs during local adaptation: insights from experimental evolution. Molecular ecology 26, 1720-1733.

Bordería, A.V., Stapleford, K.A., Vignuzzi, M., 2011. RNA virus population diversity: implications for inter-species transmission. Current Opinion in Virology 1, 643-648.

Buckling, A., Wills, M.A., Colegrave, N., 2003. Adaptation limits diversification of experimental bacterial populations. Science (New York, N.Y.) 302, 2107-2109.

Burch, C.L., Chao, L., 1999. Evolution by small steps and rugged landscapes in the RNA virus phi6. Genetics 151, 921-927.

Chao, L., 1990. Fitness of RNA virus decreased by Muller's ratchet. Nature 348, 454-455.

Chao, L., Tran, T.T., Tran, T.T., 1997. The Advantage of Sex in the RNA Virus φ6. Genetics 147, 953-959.

Ciota, A.T., Payne, A.F., Ngo, K.A., Kramer, L.D., 2014. Consequences of in vitro host shift for St. Louis encephalitis virus. Journal of General Virology 95, 1281-1288.

Coffey, L.L., Vignuzzi, M., 2011. Host Alternation of Chikungunya Virus Increases Fitness while Restricting Population Diversity and Adaptability to Novel Selective Pressures. Journal of Virology 85, 1025-1035.

Cuevas, J.M., Moya, A., SanjuÁN, R., 2009. A genetic background with low mutational robustness is associated with increased adaptability to a novel host in an RNA virus. Journal of Evolutionary Biology 22, 2041-2048.

Dandekar, T., Fieselmann, A., Fischer, E., Popp, J., Hensel, M., Noster, J., 2015. Salmonella-how a metabolic generalist adopts an intracellular lifestyle during infection. Frontiers in cellular and infection microbiology 4, 191-191.

Day, T., 2015. Information entropy as a measure of genetic diversity and evolvability in colonization. Molecular ecology 24, 2073-2083.

Dennehy, J.J., Duffy, S., O'Keefe, K.J., Edwards, S.V., Turner, P.E., 2013. Frequent Coinfection Reduces RNA Virus Population Genetic Diversity. Journal of Heredity 104, 704-712.

Domingo-Calap, P., Cuevas, J.M., Sanjuán, R., 2009. The Fitness Effects of Random Mutations in Single-Stranded DNA and RNA Bacteriophages. PLOS Genetics 5, e1000742.

Draghi, J., Wagner, G.P., 2007. Evolution of Evolvability in a Developmental Model. Evolution 62, 301-315.

Duffy, S., Burch, C.L., Turner, P.E., 2007. Evolution of Host Specificity Drives Reproductive Isolation Among RNA Viruses. Evolution 61, 2614-2622.

Duffy, S., Turner, P.E., Burch, C.L., 2006. Pleiotropic Costs of Niche Expansion in the RNA Bacteriophage φ6. Genetics 172, 751-757.

Dutta, R.N., Rouzine, I.M., Smith, S.D., Wilke, C.O., Novella, I.S., 2008. Rapid Adaptive Amplification of Preexisting Variation in an RNA Virus. Journal of Virology 82, 4354-4362.

Elena, S.F., 2016. Evolutionary transitions during RNA virus experimental evolution. Philosophical Transactions of the Royal Society B: Biological Sciences 371.

Elena, S.F., Agudelo-Romero, P., Lalić, J., 2009. The evolution of viruses in multi-host fitness landscapes. Open Virol J 3, 1-6.

Elena, S.F., CodoÑEr, F.M., SanjuÁN, R., 2003. Intraclonal variation in RNA viruses: generation, maintenance and consequences. Biological Journal of the Linnean Society 79, 17-26.

Fusaro, A., Tassoni, L., Milani, A., Hughes, J., Salviato, A., Murcia, P.R., 2016. Unexpected Interfarm Transmission Dynamics during a Highly Pathogenic Avian Influenza Epidemic. 90, 6401-6411.

Hasing, M.E., Hazes, B., Lee, B.E., Preiksaitis, J.K., Pang, X.L., 2016. A next generation sequencing-based method to study the intra-host genetic diversity of norovirus in patients with acute and chronic infection. BMC Genomics 17, 1-11.

Haydon, D.T., Woolhouse, M.E.J., 1998. Immune Avoidance Strategies in RNA Viruses: Fitness Continuums arising from Trade-offs between Immunogenicity and Antigenic Variability. Journal of Theoretical Biology 193, 601-612.

Hillung, J., Cuevas, J.M., Valverde, S., Elena, S.F., 2014. Experimental Evolution of an Emerging Plant Virus in Host Genotypes That Differ in Their Susceptibility to Infection. Evolution 68, 2467-2480.

Jasmin, J.-N., Kassen, R., 2007. Evolution of a single niche specialist in variable environments. Proceedings. Biological sciences 274, 2761-2767.

Kassen, R., 2002. The experimental evolution of specialists, generalists, and the maintenance of diversity. Journal of Evolutionary Biology 15, 173-190.

Ketola, S., Lehtinen, J., Rousi, T., Nissinen, M., Huhtala, H., Konttinen, Y.T., Arnala, I., 2013. No evidence of long-term benefits of arthroscopic acromioplasty in the treatment of shoulder impingement syndrome: Five-year results of a randomised controlled trial. Bone and Joint Research 2, 132-139.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., Wilson, R.K., 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22, 568-576.

Lauring, A.S., Andino, R., 2010. Quasispecies Theory and the Behavior of RNA Viruses. PLoS Pathogens 6, e1001005.

Lequime, S., Fontaine, A., Ar Gouilh, M., Moltini-Conclois, I., Lambrechts, L., 2016. Genetic Drift, Purifying Selection and Vector Genotype Shape Dengue Virus Intra-host Genetic Diversity in Mosquitoes. 12, e1006111.

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN].

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 26, 589-595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) 25, 2078-2079.

Liberman, U., Feldman, M.W., 1986. Modifiers of mutation rate: A general reduction principle. Theoretical Population Biology 30, 125-142.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10.

Mas, A., Lopez-Galindez, C., Cacho, I., Gomez, J., Martinez, M.A., 2010. Unfinished stories on viral quasispecies and Darwinian views of evolution. Journal of molecular biology 397, 865-877.

Mas, A., Ulloa, E., Bruguera, M., Furcic, I., Garriga, D., Fabregas, S., Andreu, D., Saiz, J.C., Diez, J., 2004. Hepatitis C virus population analysis of a single-source nosocomial outbreak reveals an inverse correlation between viral load and quasispecies complexity. The Journal of general virology 85, 3619-3626.

Morand, S., Manning, S.D., Woolhouse, M.E.J., 1996. Parasite-Host Coevolution and Geographic Patterns of Parasite Infectivity and Host Susceptibility. Proceedings: Biological Sciences 263, 119-128.

Nowak, M.A., Anderson, R.M., McLean, A.R., Wolfs, T.F., Goudsmit, J., May, R.M., 1991. Antigenic diversity thresholds and the development of AIDS. Science 254, 963-969.

Pepin, K.M., Domsic, J., McKenna, R., 2008. Genomic evolution in a virus under specific selection for host recognition. Infection, Genetics and Evolution 8, 825-834.

Pepin, K.M., Wichman, H.A., 2008. Experimental evolution and genome sequencing reveal variation in levels of clonal interference in large populations of bacteriophage φX174. BMC Evolutionary Biology 8, 85.

Pita, J.S., Roossinck, M.J., 2013. Mapping Viral Functional Domains for Genetic Diversity in Plants. Journal of Virology 87, 790-797.

R Development Core Team, 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P., 2011. Integrative Genomics Viewer. Nature biotechnology 29, 24-26.

Saito, R., Tojo, K., 2016. Comparing spatial patterns of population density, biomass, and genetic diversity patterns of the habitat generalist mayfly Isonychia japonica Ulmer (Ephemeroptera:Isonychiidae) in the Chikuma–Shinano River basin. Freshwater Science 35, 724-737.

Schneider, W.L., Roossinck, M.J., 2000. Evolutionarily Related Sindbis-Like Plant Viruses Maintain Different Levels of Population Diversity in a Common Host. Journal of Virology 74, 3130-3134.

Semancik, J.S., Vidaver, A.K., Van Etten, J.L., 1973. Characterization of segmented double-helical RNA from bacteriophage phi6. Journal of molecular biology 78, 617-625.

Troupin, C., Dacheux, L., Tanguy, M., Sabeta, C., Blanc, H., Bouchier, C., Vignuzzi, M., Duchene, S., Holmes, E.C., Bourhy, H., 2016. Large-Scale Phylogenomic Analysis Reveals the Complex Evolutionary History of Rabies Virus in Multiple Carnivore Hosts. PLOS Pathogens 12, e1006041.

Turner, P.E., Elena, S.F., 2000. Cost of host radiation in an RNA virus. Genetics 156, 1465-1470.

Turner, P.E., Morales, N.M., Alto, B.W., Remold, S.K., 2010. Role of evolved host breadth in the initial emergence of an RNA virus. Evolution 64, 3273-3286.

Van Tienderen, P.H., 1991. Evolution of Generalists and Specialist in Spatially Heterogeneous Environments. Evolution 45, 1317-1331.

Vidaver, A.K., Koski, R.K., Van Etten, J.L., 1973. Bacteriophage phi6: a Lipid-Containing Virus of Pseudomonas phaseolicola. J Virol 11, 799-805.

Wasik, B.R., Munoz-Rojas, A.R., Okamoto, K.W., Miller-Jensen, K., Turner, P.E., 2016. Generalized selection to overcome innate immunity selects for host breadth in an RNA virus. Evolution 70, 270-281.

Woolhouse, M.E.J., Gowtage-Sequeria, S., 2005. Host Range and Emerging and Reemerging Pathogens. Emerging Infectious Diseases 11, 1842-1847.

Wootton, J.C., Feng, X., Ferdig, M.T., Cooper, R.A., Mu, J., Baruch, D.I., Magill, A.J., Su, X.-z., 2002. Genetic diversity and chloroquine selective sweeps in Plasmodium falciparum. Nature 418, 320-323.

Zhao, L., Seth Pasricha, M., Stemate, D., Crespo-Bellido, A., Gagnon, J., Duffy, S., 2018. Existing host range mutations constrain further emergence of RNA viruses. bioRxiv.

Chapter 3

Eukaryotic Circular Rep-Encoding Single-Stranded DNA (CRESS DNA) viruses: ubiquitous viruses with small genomes and a diverse host range

Lele Zhao, Karyna Rosario (University of South Florida), Mya Breitbart (University of South Florida), Siobain Duffy

Abstract

While single-stranded DNA (ssDNA) was once thought to be a relatively rare genomic architecture for viruses, modern metagenomics sequencing has revealed circular ssDNA viruses in most environments and in association with diverse hosts. In particular, circular ssDNA viruses encoding a homologous replication-associated protein (Rep) have been identified in the majority of eukaryotic supergroups, generating interest in the ecological effects and evolutionary history of circular Rep-encoding ssDNA viruses (CRESS DNA) viruses. This review surveys the explosion of sequence diversity and expansion of eukaryotic CRESS DNA taxonomic groups over the last decade, highlights similarities between the well-studied geminiviruses and circoviruses with newly identified groups known only through their genome sequences, discusses the ecology and evolution of eukaryotic CRESS DNA viruses, and speculates on future research horizons.

Defining CRESS DNA viruses

The term CRESS DNA viruses was coined in 2012 to refer to a group of single-stranded DNA (ssDNA) viruses encoding a replication-associated protein (Rep) that appears to be descended from a common ancestor. CRESS DNA stands for circular, Rep-encoding ssDNA and encompasses both prokaryotic and eukaryotic viruses, although the Reps of each of these groups have distinct characteristics (Koonin and Ilyina, 1993). Most ssDNA viruses are CRESS DNA viruses. Eleven out of thirteen ssDNA virus families established by the International Committee on the Taxonomy of Viruses (ICTV, http://ictv.global/report) contain circular genomes, with *Parvoviridae* and *Bidnaviridae* being the only exceptions. Seven of these eleven circular ssDNA virus families infect eukaryotic organisms (see Confirmed and potential host range and pathogenesis of eukaryotic CRESS DNA viruses, below), and among these only members of the *Anelloviridae* family do not encode a homologous Rep. The Rep of the six families of eukaryotic CRESS DNA viruses (TABLE 3.1; *Bacilladnaviridae*, *Circoviridae*, *Geminiviridae*, *Genomoviridae*, *Nanoviridae*, *Smacoviridae*) is distantly related to the Rep of bacterial (*Microviridae* and *Inoviridae*) and archaeal (*Pleolipoviridae*) CRESS DNA viruses; however, the Reps of these groups have distinct evolutionary histories (Koonin and Ilyina, 1993, Krupovic, 2013, Koonin et al., 2015, Rosario et al., 2012b). This review focuses on eukaryotic CRESS DNA viruses, for which an unforeseen diversity and distribution has been recognized and continues to expand.

TABLE 3.1. Basic information on eukaryotic CRESS DNA viral families (http://ictv.global/report, King et al. 2012)

	Geminiviridae	***Circoviridae*** (Breitbart et al., 2017)	***Nanoviridae***	***Genomoviridae*** (Varsani and Krupovic, 2017)	***Smacoviridae*** (Varsani and Krupovic, 2018)	***Bacilladnaviridae*** (Kazlauskas et al., 2017)
Genome Size Range	1 or 2 segments 2.7-3.0kb per segment,	1.7-2.1kb	6-10 segments ~1kb per segment	2-2.3kb	2.3-2.8kb	4.5-6kb
Ambisense	Yes	Yes	N/A	Yes	Yes	Yes
Capsid Morphology & Size	Twin iscosahedra T=1 ~22x38nm	Icosahedral T=1 15-25nm	Icosahedral 18-20nm	Icosahedral 20-22nm (SsHADV-1)	Unknown	Icosahedral 33-38nm
Established Hosts	Plants	Animals	Plants	Fungi	None	Diatoms
Species Threshold (when a novel sequence is below this pairwise percent nucleotide identity across the whole genome to any established species, it represents a new species)	*Becurtovirus*: 80% (Varsani et al., 2014b) *Begomovirus*: 91% (Zerbini et al., 2017) *Capulavirus*: 77% (Varsani et al., 2017) *Curtovirus*: 77% (Varsani et al., 2014a) *Grablovirus*: 80% (Varsani et al., 2017) *Mastrevirus*:78% (Muhire et al., 2013)	80% (Rosario et al., 2017)	75% (King et al., 2012)	78%	77%	75% pairwise identity in **Rep protein** (note this is not nucleotide percent identity, nor is it genome-wide)

Discovery of Eukaryotic CRESS DNA viruses

Although the first eukaryotic CRESS DNA viruses were only identified as such in the 1970's, symptoms consistent with CRESS DNA viral infection were described over a millennium ago. The common yellowing symptom of CRESS DNA viral infection of euphorbia leaves was the inspiration for a Japanese poet in 752AD (Saunders et al., 2003), though the symptoms are not distinct enough for a definitive retrospective diagnosis. More definitively, plants with symptoms caused by geminiviruses were first observed more than 100 years ago. Abutilon mosaic virus is now known to be the cause of a pleasing mosaic pattern on the ornamental abutilon plant (FIGURE 3.1), which was highly coveted in Europe in the mid-1800s (reviewed in Wege et al., 2000). In animal hosts, the symptoms of psittacine beak and feather disease (now known to be caused by a circovirus) may have first been observed in Australia in 1888. An avid birder described parakeets which failed to grow back their feathers after their molt, leading to bizarre-looking, bald birds (Ashby, 1921). Eukaryotic CRESS DNA viruses very likely originated much longer ago than these human recordings imply, as several CRESS DNA viral genes have been found in the germ line of eukaryotic lineages that diverged at least a million years ago (see endogenized CRESS DNA viruses, below). Therefore, both historical records and modern molecular analyses support the ancient origin of eukaryotic CRESS DNA viruses.

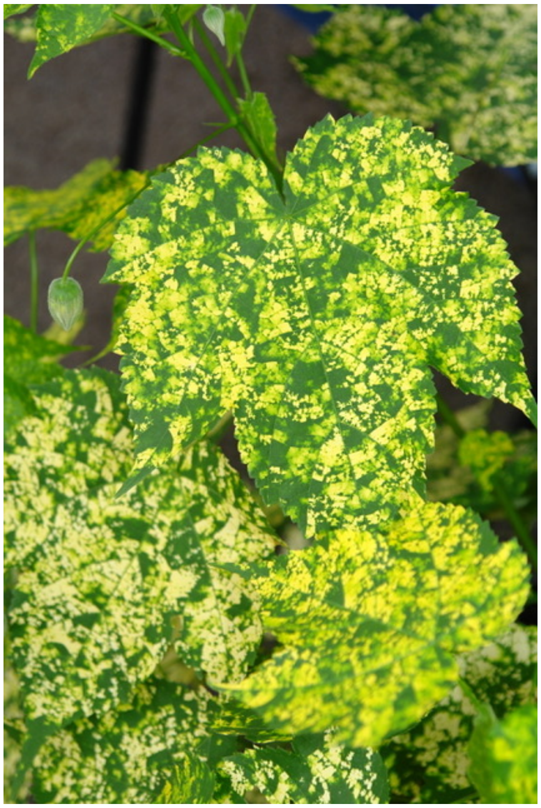


FIGURE 3.1. Abutilon mosaic virus symptoms in abutilon, from Hawaii. Scot Nelson, Public Domain (Wikimedia Commons)

Throughout history, the effects of viral infection were observed long before humans could isolate and identify the etiological agent, and CRESS DNA viruses were no exception. It wasn't until 1977 that scientists identified the first eukaryotic virus containing a circular ssDNA genome, bean golden mosaic virus (*Geminiviridae*, Goodman, 1977a). Despite the earlier discovery of ssDNA viruses infecting bacteria (Sertic and Bulgakov, 1935) and mammals (Crawford, 1966), the overwhelming majority of DNA viruses were considered to be double-stranded at the time. Five years after the description of geminiviruses, porcine circovirus was the first animal-infecting circular ssDNA virus described (Tischer et al., 1982). Following these findings, the known diversity of geminiviruses rapidly increased, with 63 species identified by 1995 (Murphy et al., 1995). Largely these efforts were the result of pathologists determining the causative agent of economically important crop diseases. In contrast to the geminiviruses, awareness of other groups of eukaryotic CRESS DNA viruses did not expand significantly in terms of either diversity or detection until the 2000s -- only 3 circovirus species and 4 nanovirus species were recognized by 1999 (Regenmortel et al., 2000). Due to their lack of affiliation with human disease, CRESS DNA viruses were largely ignored by biomedical funding agencies, which instead directed efforts towards known pathogenic dsDNA, RNA, and retro-transcribing viruses. As a result, eukaryotic CRESS DNA viruses only recently gained recognition commensurate with their ubiquity, diversity, and impact.

The application of phi29 DNA polymerase and random hexamers (rolling circle amplification; RCA) for whole genome amplification of circular DNA templates (Dean et al., 2001) represented a turning point in the study of CRESS DNA viruses (reviewed in Rosario et al., 2012b). This method was so efficient that it was quickly adapted to clone and sequence complete geminivirus genomes, revolutionizing methodological approaches for detection of plant pathogens (Haible et al., 2006, Inoue-Nagata et al., 2004, Wyant et al., 2012). During this same time period, RCA was also being utilized in both clinical and environmental settings to obtain sufficient DNA concentrations for next-generation sequencing (Angly et al., 2006, Breitbart and Rohwer, 2005, Lasken and Egholm, 2003). Although the discovery of circular ssDNA viruses was not the intention of these endeavors, application of this method serendipitously resulted in the detection of a diversity of CRESS DNA viruses in unsuspected organisms and disparate environments (reviewed in Rosario et al., 2012b). Since RCA leads to a gross overrepresentation of viruses with circular genomes, this method cannot be used for quantitative analyses of viral communities (Kim and Bae, 2011, Roux et al., 2016). RCA is currently used in two distinct ways for discovering viruses with small circular genomes. The incorporation of RCA into standard metagenomics pipelines enables assembly of complete CRESS DNA genomes from both individual organisms and complex environmental communities (Li et al., 2010a, Rosario et al., 2009). However, caution must be used with assembly-based methods since RCA may lead to chimeric sequences (Tu et al., 2015). To verify assembled genomes, inverse PCR with abutting primers is recommended (Rosario et al., 2009). Alternatively, numerous studies have directly recovered unit length CRESS DNA viral genomes by applying RCA followed by restriction enzyme digestion to clone single genomes of viruses, circumventing potential assembly errors such as chimeric genomes (Inoue-Nagata et al., 2004, Rosario et al., 2012).

RCA has been instrumental in the recognition of the ubiquity and diversity of CRESS DNA viruses. The number of CRESS DNA viruses identified using this technique now far exceeds the number of well-characterized viral isolates obtained using classical methods. This is in stark contrast to historical situations where symptoms of viral infection were recognized long before the etiological agent was identified. The explosion of CRESS DNA viral discovery through sequence-based methods has divorced viral sequences from much of their ecological context, including fundamental aspects such as host range.

We have now passed through the looking glass, where we are increasingly aware of the existence of incredible numbers of distinct viral species, without having any sense of the impact of these viruses.

Unity and Diversity

While some eukaryotic CRESS DNA viruses have up to 10 open reading frames (ORFs), even the most compact genomes have two ORFs: one encoding the Rep and one encoding a capsid protein (CP). The conserved Rep serves as the anchor for this group of viruses, but the CP is highly divergent. Beyond the Rep and CP, protein content differs dramatically among CRESS DNA viruses.

Molecular biology of Rep

All eukaryotic CRESS DNA viruses encode a distinctive homologous Rep that is presumably conserved due to its essential function in viral genome replication through rolling circle replication (RCR). In fact, the Rep is often the only gene with homology among the divergent eukaryotic CRESS DNA viruses, and thus has been used extensively for phylogenetic analyses and higher-level taxonomic classification (Simmonds et al., 2017). The RCR mechanism (FIGURE 3.2) employed by eukaryotic CRESS DNA viruses (reviewed in Rosario et al., 2012b) has been elucidated based on *in vitro* studies performed with members representing only three of the eukaryotic CRESS DNA viral families, including the *Geminiviridae* (Hanley-Bowdoin et al., 2013, Jeske et al., 2001, Laufs et al., 1995), *Circoviridae* (Faurez et al., 2009, Steinfeldt et al., 2006), and *Nanoviridae* (Timchenko et al., 2000, Timchenko et al., 1999). In addition, the structure of several representative Reps from these three families has been solved (Campos-Olivas et al., 2002, Vega-Rocha et al., 2007a, Vega-Rocha et al., 2007b). The Rep binds to iterative sequences near an origin of replication (*ori*) distinguished by a conserved nonanucleotide motif at the apex of a hairpin structure. The Rep then nicks the covalently closed virion strand of the double-stranded replicative form of the viral genome within the nonanucleotide motif. After the exposed 3’-OH end is primed by a host polymerase, leading-strand synthesis proceeds until the Rep rejoins the virion strand at the initial nicking site.

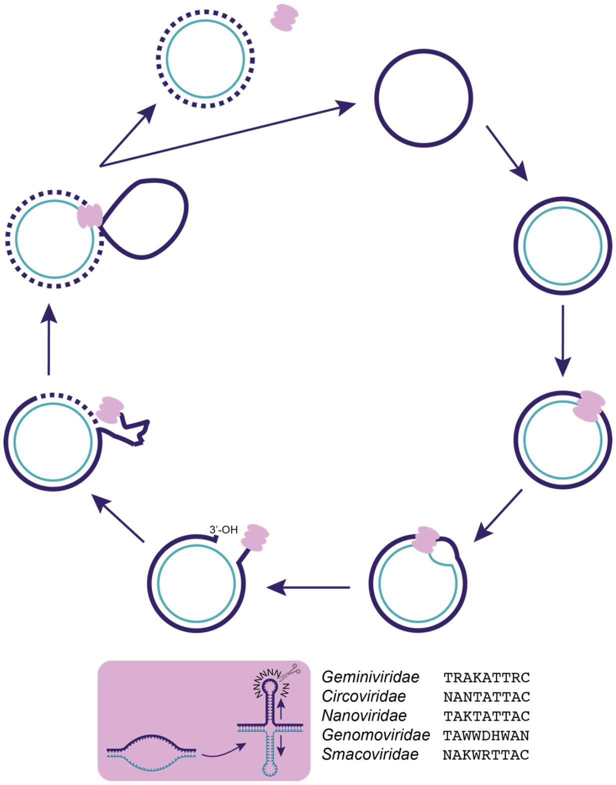


FIGURE 3.2. Rolling circle replication in CRESS DNA viruses. The single stranded viral genome (thick purple circle) is converted to a double-stranded replicative form (the complementary strand is shown as the thinner blue circle) by the use of host factors and DNA polymerase. During initiation, the viral replication-associated protein (Rep, light purple) binds to the replicative form near the origin of replication (a stem-loop hairpin). The Rep's endonuclease domain is predicted to introduce a nick between position 7 and 8 of the nonanucleotide motif (degenerate motifs for five eukaryotic CRESS DNA viral families given). The nick site is ready for leading-strand synthesis at the 3'-OH end by host DNA polymerases, while the 5' end is covalently bonded to the Rep protein. During replication, the initial virion strand (thick purple line) is displaced by the newly synthesized positive strand (dashed thick purple line). When replication is complete, Rep catalytically joins the nicks, allowing the initial virion strand to be released from the double stranded replicative form.

The Reps of established and metagenomically identified eukaryotic CRESS DNA viruses stably contain a distinctive two functional domain organization, containing a HUH (His-hydrophobe-His) endonuclease domain towards the N-terminus and superfamily 3 (SF3) helicase domain at the C-terminus (Ilyina and Koonin, 1992, Koonin, 1993). Each of the eukaryotic CRESS DNA viral Rep domains is characterized by conserved motifs important for RCR (reviewed by (Rosario et al., 2017). The HUH endonuclease domain is characterized by RCR motifs I through III, which are important for RCR initiation and termination. The SF3 helicase domain contains Walker A, Walker B and motif C motifs that presumably allow the Rep to act as a replicative helicase during RCR elongation (Gorbalenya et al., 1990, Gorbalenya and Koonin, 1993). The Rep SF3 helicase domain also commonly contains a fourth motif corresponding to a catalytic arginine finger (Kazlauskas et al., 2017). This motif is conserved in various AAA+ family ATPases that may fuel helicase activity (Nagy et al., 2016). This arginine finger has been identified in Reps encoded by members of *Bacilladnaviridae*, *Circoviridae*, *Nanoviridae*, and *Smacoviridae*, but not in *Geminiviridae* and *Genomoviridae* (Kazlauskas et al., 2017).

Intron-containing Reps

Many eukaryotic CRESS DNA viruses are known or predicted to express both spliced and non-spliced forms of the Rep. In mastreviruses (*Geminiviridae*), RepA (non-spliced form) and Rep (spliced form) are identical for the first ~200 N-terminal residues and are multifunctional proteins involved in virus genome replication, transcription, and gene regulation (Hefferon et al., 2006, Muñoz-Martín et al., 2003, Fondong, 2013). RepA has also been identified in other geminivirus genera including, *Becurtovirus*, *Capulavirus* and *Grablovirus* (Varsani et al., 2014b, Varsani et al., 2017). The majority of the genomes representing the newly denoted family *Genomoviridae* also contain both RepA and the spliced Rep (Conceicao-Neto et al., 2015, Steel et al., 2016). Some members of the family *Circoviridae*, specifically porcine circoviruses (PCV), contain Rep and Rep’, which is a spliced isoform of Rep with the C-terminus truncated and expressed in a different frame (Steinfeldt et al., 2006). Both proteins perform the nicking and joining activities during RCR and have sequence similarity to geminivirus and nanovirus Reps. One proposed mechanism for first appearance of spliced Reps in eukaryotic CRESS DNA viruses involves the endonuclease function of the Rep itself: site-specific endonucleases have been found in intron homing processes (Belfort and Perlman, 1995), implying that the Rep protein may have been involved in the acquisition of an intron (Gibbs et al., 2006).

Capsid proteins

All CRESS DNA virus groups that have been visualized by electron microscopy have icosahedral capsids, which are encoded by a capsid protein (CP). The CP protects the genome and is needed for the virus to move between individual hosts. The known capsids of CRESS DNA viruses vary in size (TABLE 3.1), and in the case of *Geminiviridae* in shape: the geminiviruses derive their name from their twinned icosahedral capsid (FIGURE 3.3), which packages a single genomic segment (Goodman, 1977b, Harrison et al., 1977).

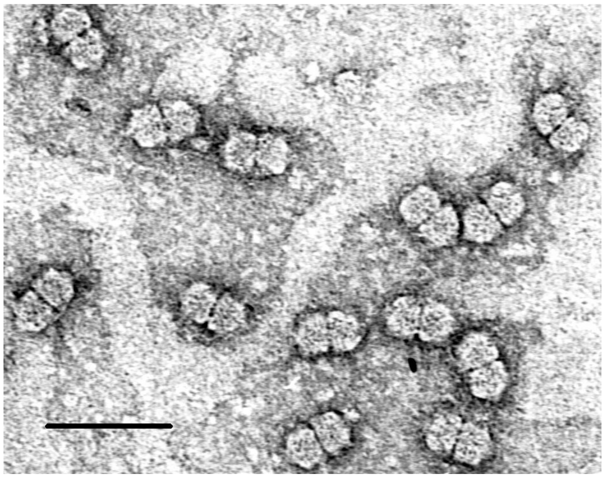


FIGURE 3.3. Maize streak virus particles. Kassie Kasdorf, Public Domain (Wikimedia Commons)

There is evidence for recombination among CRESS DNA viral CPs (e.g., the genus *Curtovirus* contains viruses descended from a recombinant begomovirus with a CP from a mastrevirus, which changed the vector specificity, Rybicki, 1994); however, some CP genes appear to have evolutionary linkages to those of RNA viruses (Kazlauskas et al., 2017, Krupovic et al., 2009, Lefeuvre et al., 2009, Roux et al., 2013). The capsid proteins of geminiviruses are structurally similar to that of a ssRNA plant virus, satellite tobacco necrosis virus, which is suggestive of a shared evolutionary history (Krupovic et al., 2009). The *Bacilladnaviridae* capsids are also suggested to be structurally similar to another group of ssRNA viruses, the nodaviruses (Kazlauskas et al., 2017) and the *Circoviridae* capsid proteins may have yet another common ancestor with capsids of RNA viruses (Gibbs and Weiller, 1999). In 2012, a RNA-DNA hybrid virus (RDHV) encoding a eukaryotic CRESS DNA viral Rep and a CP from unclassified ssRNA viruses similar to *Tombusviridae* was discovered from Boiling Spring Lake (Diemer and Stedman, 2012). This unique CRESS DNA virus group, named the cruciviruses (Quaiser et al., 2016), has been growing steadily since its initial discovery (Bistolas et al., 2017a, Dayaram et al., 2016, Hewson et al., 2013, Krupovic et al., 2015, Steel et al., 2016). The frequency of discovery of these viruses with a seeming RNA-DNA hybrid evolutionary history suggests that these recombination events are not vanishingly rare, and the mechanisms that could create these hybrids will continue to be explored in the coming years (Stedman, 2015). Capsid protein sequence diversity has not only come from RNA viruses, newly sequenced eukaryotic CRESS DNA viral genomes carry capsid protein sequences similar to ORF1, the putative capsid protein coding gene of the ssDNA *Anelloviridae* (Lamberto et al., 2014).

In contrast to the homologous Rep protein shared among the eukaryotic CRESS DNA viruses, the CP can be highly divergent and is sometimes even unrecognizable by sequence similarity (e.g., Yoon et al., 2011). In public databases, some researchers have annotated the non-Rep-encoding ORF of CRESS DNA viruses as a CP by default; however, caution should be applied in interpreting and propagating these annotations without independent evidence that the ORF actually represents a capsid. One useful tool for identifying structural proteins within novel CRESS DNA viral genomes is the prediction of disorder patterns, which are conserved among CRESS DNA viral capsid proteins and can be used to complement similarity-based searches (Rosario et al., 2015a). While great strides have been made in understanding the evolution of eukaryotic CRESS DNA viruses based on the unifying Rep, deciphering the evolutionary histories of the CPs is clearly a more complex task that needs to be the subject of future studies.

Current taxonomy

Prior to 2015, the ICTV recognized three families of CRESS DNA viruses, namely *Geminiviridae* (approved in 1993), *Circoviridae* (approved in 1993), and *Nanoviridae* (approved in 2002). The number of CRESS DNA viral families has recently doubled with the addition of *Genomoviridae* (approved in 2015), *Bacilladnaviridae* (approved in 2017), and *Smacoviridae* (approved in 2017). The number of CRESS DNA viral genera increased more than five-fold (from 6 to 31) during the past five years and is currently distributed as follows (TABLE 3.2): *Geminiviridae* (9 genera), *Circoviridae* (2 genera), *Nanoviridae* (2 genera), *Genomoviridae* (9 genera), *Smacoviridae* (6 genera), and *Bacilladnaviridae* (3 genera). The majority of this increase is due to the advances in viral metagenomics (Simmonds et al., 2017), and one of the six families (*Smacoviridae*) has no cultured representatives. A recent exemplary geminivirus case study highlights the revolutionary changes metagenomics research has had on CRESS DNA virology (Claverie et al., 2018).

TABLE 3.2. Familes, genera, number of classified species, and NCBI GenBank accession number for each genus' type species.

Family	Genus	Species #	Type Species	Accession Number
Bacilladnaviridae	*Diatodnavirus*	1	Chaetoceros diatodnavirus 1	AB781089
	Kieseladnavirus	1	Avon-Heathcote Estuary associated kieseladnavirus	AQA27298
	Protobacilladnavirus	7	Chaetoceros protobacilladnavirus 1	AB193315
Circoviridae	*Circovirus*	29	Porcine circovirus 1	AF071879
	Cyclovirus	45	Human associated cyclovirus 8	KC771281
Geminiviridae	*Becurtovirus*	2	Beet curly top Iran virus	EU273818
	Begomovirus	388	Bean golden yellow mosaic virus	DNA-A: M88686 DNA-B: M88687
	Capulavirus	4	Euphorbia caput-medusae latent virus	KT214376
	Curtovirus	3	Beet curly top virus	M24597
	Eragrovirus	1	Eragrostis curvula streak virus	FJ665631
	Grablovirus	1	Grapevine red blotch virus	KC896623
	Mastrevirus	37	Maize streak virus	AF329878
	Topocuvirus	1	Tomato pseudo-curly top virus	X84735
	Turncurtovirus	2	Turnip curly top virus	GU456685
	Unclassified	2		
Genomoviridae	*Gemycircularvirus*	43	Sclerotinia gemycircularvirus 1	GQ365709
	Gemyduguivirus	1	Dragonfly associated gemyduguivirus 1	JX185428
	Gemygorvirus	5	Staling associated gemygorvirus 1	KF371632
	Gemykibivirus	16	Dragonfly associated gemykibivirus 1	JX185430
	Gemykolovirus	2	Pteropus associated gemykolovirus 1	LK931484
	Gemykrogvirus	3	Bovine associated gemykrogvirus 1	LK931484
	Gemykroznavirus	1	Rabitt associated gemykronznavirus 1	KF371631
	Gemytondvirus	1	Ostrich associated gemytondvirus 1	KF371630
	Gemyvongvirus	1	Human associated gemyvongvirus 1	KP974693
Nanoviridae	*Babuvirus*	3	Banana bunchy top virus	Rep: S56276 CP: L41574 L41575 L41576 L41577 L41578
	Nanovirus	8	Subterranean clover stunt virus	Rep: AJ290434 CP: U16734 U16730 U16732 U16733 U16736
	Unassigned	1		
Smacoviridae	*Bovismacovirus*	3	Bovine associated bovismacovirus 1	JN634851
	Cosmacovirus	1	Bovine associated cosmacovirus 1	KT862228
	Dragsmacovirus	1	Dragonfly associated dragsmacovirus 1	KM598410
	Drosmacovirus	3	Camel associated drosmacovirus 1	KM573769
	Huchismacovirus	7	Human associated huchismacovirus 1	KP233180
	Porprismacovirus	27	Chimpanzee associated porprismacovirus 1	GQ351272

Geminiviridae

Geminiviruses are a group of plant-infecting viruses, encapsidated in twinned icosahedral capsids -- the structural feature that gave them the name gemini (Hanley-Bowdoin et al., 2013). *Geminiviridae* is currently the most speciose family of all viruses. Members of this viral family infect monocotyledons (monocots) and dicotyledons (dicots). Geminiviruses encompass both monopartite (six ORFs) and bipartite genomes (one segment, DNA-A, with five ORFs which is homologous to the monopartite genome, the other, DNA-B, with 2 ORFs, Rey et al., 2012). The monopartite genomes and DNA-As both encode a coat protein (CP) in the sense orientation and four ORFs in anti-sense: the Rep protein, the replication-enhancer protein (REN) and transactivating protein (TraP) and an overlapping C4 protein which affects virulence (Hanley-Bowdoin et al., 2013). Monopartite genomes also encode a partially overlapping pre-coat protein that functions as a movement protein within the plant; bipartite geminiviruses have a DNA-B encoding a movement protein and a nuclear shuttle protein (Ho et al., 2014). Bipartite begomoviruses typically dominate in the New World (the Americas), while monopartite begomoviruses are the vast majority in the Old World (Ho et al., 2014); however, there are an increasing number of exceptions to this trend (e.g., Inoue-Nagata et al., 2016, Rosario et al., 2015b).

The largest genus within *Geminiviridae* is *Begomovirus*, which contains over three quarters of the current classified geminivirus species. Begomoviruses infect dicots and are transmitted by the whitefly *Bemisia tabaci*. *B. tabaci* is phloem-feeding and some biotypes have a very broad plant host range, including ornamental, vegetable, grain, legume, and cotton plants (De Barro et al., 2011). Possibly due to the very large number of identified species of begomoviruses, they comprise 90% of the viruses known to be transmitted by whiteflies (Jones, 2003). Species belonging to the second largest genus, *Mastrevirus*, are transmitted by leafhoppers (order Hemiptera, family Cicadellidae) and infect both monocot and dicot plants. The geminiviruses are usually transmitted by the vectors in a circulative, persistent and non-propagative manner (Blanc et al., 2014) – there is very little evidence that they replicate in their vectors despite being capable of transmission long after their initial acquisition (Czosnek et al., 2017). The long residence time of geminiviruses in their vectors has facilitated discovery efforts, as researchers have targeted insect vectors and their predators as a method for surveying viral diversity circulating among plants in a given region (vector-enabled metagenomics, (Dayaram et al., 2013b, Ng et al., 2011a, Rosario et al., 2012a)).

Recent revisions to the *Geminiviridae* family established the new genera *Becurtovirus*, *Eragrovirus*, *Turncurtovirus* (Varsani et al., 2014b), *Capulavirus* and *Grablovirus* (Varsani et al., 2017). Geminiviruses are well studied in molecular virology because of the devastating effects of the well-characterized genera in important crops across temperate and tropic regions (Seal et al., 2006b). Frequent emergence and re-emergence of the geminiviruses in crops have attracted much research focus (Anderson et al., 2004), and the threat of unidentified geminiviruses emerging in crops is ever imminent (Claverie et al., 2018). Geminiviruses, as eukaryotic CRESS DNA viruses, have several evolutionary advantages that favor emergence (see Evolution, below), but a key ecological factor in the spread of begomoviruses specifically is the spread and abundance of their vectors (Seal et al., 2006a). Climate change and the increase of international trade have given rise to the expansion and invasion of vector populations to previously naïve parts of the world (for begomoviruses, the polyphagous *B. tabaci* Middle East-Asia minor 1), which both transport viruses to new areas and promote virus transmission in their invaded range (Varma et al., 2011).

Circoviridae

The family *Circoviridae* is notable because it contains the smallest known animal viruses, with nearly all members having genome sizes under 2kb. Prior to the recent taxonomic revisions of the eukaryotic CRESS DNA viruses, *Circoviridae* was the catch-all for the animal infecting circular ssDNA viruses – including viruses affecting birds and mammals. As more and more diverse CRESS DNA sequences were identified, some of the sequences initially thought to represent circoviruses or 'circo-like' viruses based on low amino acid level similarities to the Rep of members of the *Circoviridae*, were assigned to other new CRESS DNA viral families. Notably, taxa without a Rep ORF were reassigned *Anelloviridae* (Rosario et al., 2017). Despite this culling, there is still tremendous diversity associated with *Circoviridae*, which is now composed of the genera *Circovirus* (established 1993) and *Cyclovirus* (established 2015). Members of the *Circoviridae* have been found in numerous vertebrate and invertebrate organisms (e.g., birds, fish, mammal, insects). Interestingly, members from the genus *Circovirus* seem to be mainly restricted to vertebrate hosts, whereas cycloviruses have been identified in both vertebrates and invertebrates (Rosario et al., 2017).

Although cycloviruses have been only identified through molecular analysis and no definitive host has not been identified for this genus, some members of the *Circovirus* genus are well-recognized pathogens in animals. Porcine circovirus 2 (PCV2) causes porcine circovirus-associated disease (PCAD). This PCAD includes the post-weaning multisystemic wasting syndrome that causes devastating economical losses in the commercial hog industry, prompting widespread vaccination against a dominant strain of porcine circovirus 2 (Alarcon et al., 2013, Allan et al., 2012, Meng, 2013). Other circoviruses are known to infect livestock or human companion animals, for example, beak and feather disease virus (BFDV) (Harkins et al., 2014), and canine circovirus (Kapoor et al., 2012) .

Each genus in the family *Circoviridae* is distinguished by the genome organization. All genomes in the genus *Circovirus* encode the Rep protein in virion sense (positive sense), and CP in anti-sense, while the genomes in *Cyclovirus* have the opposite orientation: encoding the Rep in anti-sense and the CP in the virion sense predicted strands (the virion sense prediction is based on the nonanucleotide sequence in the origin of replication, FIGURE 3.2). The replication and transcription processes are thought to differ between two genera of *Circoviridae* due to their difference in genome organization (Rosario et al., 2017). While all genomes contain these two ORFs, a third, overlapping ORF has been experimentally verified in some species, and others have potential additional ORFs (Bassami et al., 1998, Hamel et al., 1998).

Cycloviruses are found in a diverse range of samples: squirrels (Sato et al., 2015), cats (Zhang et al., 2014), humans (Li et al., 2010a), goats (Li et al., 2011), horses (Li et al., 2015a), bats (Wu et al., 2016), cows (Li et al., 2011), sheep (Li et al., 2010a), chickens (Li et al., 2011), cockroaches (Padilla-Rodriguez et al., 2013) and dragonflies (Rosario et al., 2012a). However, it is important to note that while cycloviruses are found in association with all of these animals, the host range of all members of the genus remains unresolved without definitive experiments. Even the Dragonfly cyclovirus that researchers are nearly certain infects an insect, may not infect the dragonflies from which it was isolated – the virus was present in the dragonfly gut. Dragonflies eat a variety of insects and the virus could have been infecting an insect the dragonfly had eaten (Rosario et al., 2011). Among eukaryotic CRESS DNA viruses, only the cycloviruses were thought to be associated with invertebrates (Tijssen et al., 2016), but more recent work has shown that Rep sequences similar to members of the *Smacoviridae* and *Genomoviridae* are also associated with invertebrates (Rosario et al., 2018). Further complicating establishing definitive host range, phylogenetic analysis of cycloviruses do not show sequences clustering according to their host of isolation, making inferences about host use unproductive (Rosario et al., 2017).

Nanoviridae

Family *Nanoviridae* contains two genera: *Babuvirus* and *Nanovirus*. All viruses in this family are multipartite, meaning they maintain their genomes in multiple segments of circular positive sense ssDNA that independently package into multiple, separate capsids. Genus *Babuvirus* contain three species, with either six or nine segments to their genome, that are known to infect tropical crops including bananas, abaca, taro and cardamom (Stainton et al., 2015). Genus *Nanovirus* contains eight species, seven of which have eight segments in their genomes, and one, Subterranean clover stunt virus, has six segments. All species from genus *Nanovirus* naturally infect legumes. Individual genomic segments are around 1 kb in size, usually carrying one identifiable ORF per segment (Sharman et al., 2008, Sicard et al., 2013). Similar to ORFs maintained in begomoviruses, nanoviruses usually encode a replication-associated protein, a capsid protein, a movement protein, a replication enhancer protein, and a nuclear shuttle protein, as well as one or more proteins of unknown function (Sicard et al., 2013). Nanoviruses are transmitted by aphids, usually causing stunt symptoms in infected plants, leading to agricultural losses throughout the tropics. Like the other plant infecting eukaryotic CRESS DNA virus family, *Geminiviridae*, the nanoviruses are also transmitted in a persistent, circulative and non-propagative manner (Blanc et al., 2014).

The intriguing genomic compartmentalization of multipartite viruses has motivated scientists to study the cost of maintaining such a lifestyle. Due to their independent packaging, each segment must be independently transmitted to a new host for successful infection, and bottlenecks both in movement within plants and in aphid transmission pose obstacles to efficient infection of new hosts (Gallet et al., 2018). Further, instead of observing all segments at equal frequency, which is theoretically most efficient, researchers found that each virus maintains different segments at different frequencies, named the setpoint genome formula. Moreover, this setpoint genome formula varies with different hosts and has been proposed to be potentially beneficial for multipartite viruses (Sicard et al., 2013).

Genomoviridae

The first member of what would eventually be described as the family *Genomoviridae* is *Sclerotinia sclerotiorum* hypovirulence associated DNA virus 1 (SsHADV-1), isolated in 2010 (Yu et al., 2010). The genetic similarity between geminiviruses and SsHADV-1 was obvious from its isolation, and the family derives its name from ‘geminivirus-like with no movement protein'. SsHADV-1 is the first and only known fungal infecting ssDNA virus (i.e., mycovirus), with all other identified mycoviruses containing double-stranded or single-stranded RNA genomes (Yu et al., 2010). In contrast to geminiviruses, SsHADV-1 purified virions and viral DNA, both dsDNA and ssDNA, are infectious to the fungal host (Yu et al., 2013). The virus-host pair SsHADV-1 and *S. sclerotiorum* have been proposed as a potential tool for genetic studies because of easy manipulation in PEG-mediated protoplast transfection assays (Yu et al., 2013). If its host range can be widened experimentally, SsHADV-1 may also be applicable as a biological control measure by inducing hypovirulence in plant pathogenic fungi (Yu et al., 2010).

The classification of *Genomoviridae* viral genomes is based on maximum likelihood phylogenetic analysis on the Rep protein sequence alone. Five clades and 4 single branches are displayed in the phylogenetic tree – the group with SsDHAV-1 and eight other genera (Varsani and Krupovic, 2017). As SsHADV-1 has been characterized in the lab, it serves as the type species for genus *Gemycircularvirus*, which is named for Gemini-like myco-infecting circular virus (Rosario et al., 2012a). It currently contains 43 species, with other sequences assigned to *Gemycircularvirus* found associated with mammals, birds, insects, plants, fungi, sediments and sewage samples, but without definitive hosts (Steel et al., 2016, Sikorski et al., 2013c, Dayaram et al., 2012, Dayaram et al., 2015b, Male et al., 2015, Kraberger et al., 2013, Yu et al., 2010, Kraberger et al., 2015a). The established tradition for geminiviruses is to create genus names from abbreviations of the type species of each genus, and its host and symptoms – as in Bean golden mosaic virus - *Begomovirus*. Lacking this information, genomovirus genera all use the "gemy" prefix (Gemini-like, myco-infecting) with words from different languages to emphasize the circularity of the genomes (Table 2).

Smacoviridae

Another novel family discovered largely through metagenomic sequencing is *Smacoviridae* (smaco stands for small circular, Varsani and Krupovic, 2018). Although no member of this family has been cultured and smacoviruses have only been identified in fecal matter or dragonflies, the genomes of classified species have all been verified through PCR and Sanger sequencing. Smacovirus genera were established based on Rep phylogenetic analysis, with ≥40% Rep protein sequence identity required for members of the same genus. *Smacoviridae* is divided into six genera: *Bovismacovirus*, *Drosmacovirus*, *Huchismacovirus*, *Porprismacovirus*, *Cosmacovirus*, and *Dragsmacovirus*. The CPs of *Smacoviridae* are shared within the family, but not related to other CRESS DNA viruses (Varsani and Krupovic, 2018).

Bacilladnaviridae

Bacilladnaviridae contains 3 genera: *Protobacilladnavirus*, *Diatodnavirus*, and *Kieseladnavirus*. The first officially classified member of the *Bacilladnaviridae* family was the *Chaetoceros salsugineum* DNA virus 01 (CsalDNAV01), the first diatom-infecting DNA virus and only the second known diatom-infecting virus (the first was *Rhizosolenia setigera* RNA virus, Nagasaki et al., 2005). The sediment samples containing CsalDNAV01 were collected in 2003, but in the following years, another nine bacilladnaviruses that infected other abundant *Chaetoceros* species and *Thalassionema* species were identified (Kimura and Tomaru, 2013, Kimura and Tomaru, 2015, Tomaru et al., 2008, Tomaru et al., 2011a, Tomaru et al., 2011b, Tomaru et al., 2012, Tomaru et al., 2013, Toyoda et al. 2012). As diatoms are an important player in marine and freshwater ecology, these host-virus systems are expected to provide insights into diatom-blooming dynamics, and should be more intensively researched in the coming years.

Several characterized *Bacilladnaviridae* viruses have double-stranded DNA genomic segments of varying size and location, the properties of which are unknown (Kimura and Tomaru, 2015, Tomaru et al., 2013), which is unique thus far among eukaryotic CRESS DNA viruses. The CPs of bacilladnaviruses show sequence and structure similarity to the CPs of nodaviruses, a group of ssRNA virus. Bacilladnaviruses have larger genomes sizes than other eukaryotic CRESS DNA viruses (~4.5-6kb, Kimura and Tomaru, 2015), but similar to the genome size of nodaviruses.

This larger genome size requires larger internal volume and the virus particle size (33-38 nm) of the nodavirus capsid structure accommodates the genome size (Kazlauskas et al., 2017).

Evolutionary relationships among the eukaryotic CRESS DNA viral families

The homologous Rep protein of all eukaryotic CRESS DNA viruses allows for a straightforward analysis of the evolutionary history of this one ORF among all the diverse families mentioned above. However, ssDNA viruses frequently recombine, which interferes with accurate resolution of phylogenetic relationships (Martin et al., 2011). A recent publication accounted for the pervasive recombination in the Rep gene among unclassified eukaryotic CRESS DNA viruses, producing a dataset free of detectable recombination to build a robust Rep genealogy (Kazlauskas et al., 2018). FIGURE 3.4 shows our own analysis conducted with that recombinant-free dataset with a novel substitution matrix, which shows the same broad patterns as the previous publication. It reaffirms that the *Genomoviridae* Rep is closely related to that of *Geminiviridae*, and shows their reciprocal monophyly. *Nanoviridae* sequences form a clade with Reps from a group of satellites (discussed in Confirmed and potential pathogenesis of eukaryotic CRESS DNA viruses, below). This clade forms a larger monophyletic group with the *Smacoviridae* (and some unclassified Rep sequences), indicating that nanovirus Reps are the closest classified relatives to the smacoviruses. The recently established *Bacilladnaviridae* is roughly equally distant from all other named families of eukaryotic CRESS DNA viruses, foiling attempts to speculate with which other group of eukaryotic CRESS viruses it might share a recent common ancestor. Indeed, *Bacilladnaviridae* is fairly distant even from the unclassified Rep sequences (shown in black), suggesting either its sister taxon has yet to be sampled or that its closer relatives were unsuccessful over evolutionary time. Finally, it is important to note the preponderance of black taxa on the tree, and the groups they assemble into. While Rep similarity is not the sole determinant of eukaryotic CRESS DNA viral taxonomy, this is an indication that there are further cohesively evolving groups to be systematized within the eukaryotic CRESS DNA viruses. These candidate groups have been given working names for the time being [see (Kazlauskas et al., 2018)]. The diversity of these viruses may not be completely sampled, but we know that the six named families cover just over half of the Rep diversity we do know about.

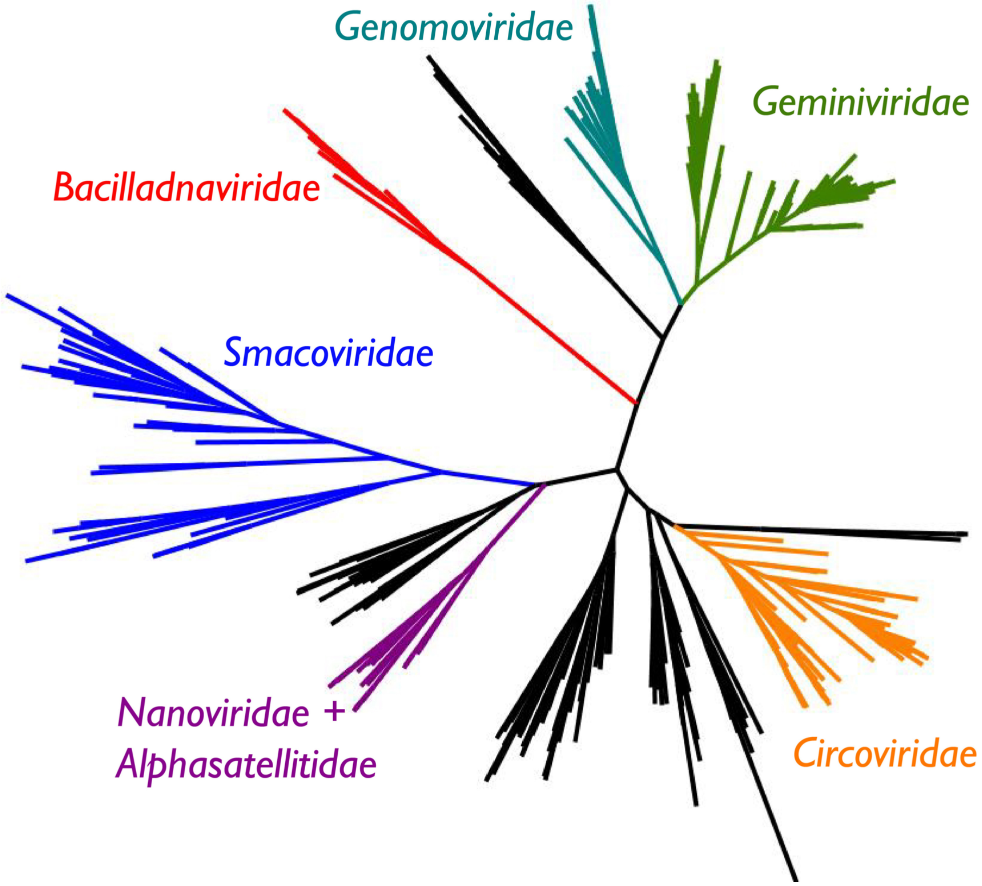


FIGURE 3.4. Maximum likelihood tree of eukaryotic CRESS DNA virus Rep protein sequences. This tree was built using sequences from a published recombination-free dataset containing sequences from classified species (the six recognized families, shown in bright colors) and unclassified eukaryotic CRESS DNA genomes shown in black (Kazlauskas et al., 2018). The 382 sequences were aligned in MUSCLE (Edgar, 2004) and then trimmed with TrimAl [conserving 30% of alignment, (Capella-Gutierrez et al., 2009)]. The phylogenetic tree was built in PhyML (Guindon and Gascuel, 2003) using a substitution matrix we developed for eukaryotic CRESS DNA genomes ("CRESS", see Chapter 4)+G+F.

Ecology of Eukaryotic CRESS DNA Viruses

Distribution and sampling

We are not sure how close to the tip of the iceberg virologists are in terms of uncovering CRESS DNA viral diversity, but recent global efforts have more than doubled the number of eukaryotic CRESS DNA viral species in GenBank over the last 10 years (let alone the new strain sequences that add to our appreciation of the diversity within some eukaryotic CRESS DNA viral species). These abundant sequences come from a huge range of hosts and environments, and include both opportunistic sampling and intentional surveys.

While we have tried to be comprehensive in our survey of hosts and environments where researchers have identified eukaryotic CRESS DNA viral genomes (TABLE 3.3), our summary will be out of date soon after publication as further genomes are detected. Additionally, a list of places where eukaryotic CRESS DNA viruses have been detected suffers from the bias of not having a companion list of environments where researchers tried and failed to detect eukaryotic CRESS genomes. While the literature truly makes these viruses appear ubiquitous, the difficulty of proving a negative result means that a lack of CRESS DNA viruses would be hard to report. However, we note that there is a strong place in eukaryotic CRESS DNA viral discovery for negative controls; sequencing-based studies can be compromised by contaminated reagents (Salter et al., 2014) and CRESS DNA viral genomes have been found in commercially available DNA isolation spin columns (Naccache et al., 2013).

TABLE 3.3. Representative studies where eukaryotic CRESS DNA viruses were found (2010-recent).

Sampling from Wildlife	Animal	Location	Reference
feces	bats	Texas and California, USA	(Li et al., 2010b)
		Brazil	(Lima et al., 2015)
		Kingdom of Tonga	(Male et al., 2016)
		Europe	(Kemenesi et al., 2018)
		China	(Ge et al., 2011)
	chimpanzees	Africa	(Blinkova et al., 2010; Li et al., 2010a)
	wild rodents	California, USA	(Phan et al., 2011)
	mammal and bird	New Zealand	(Sikorski et al., 2013c)
	700-y-old frozen caribou	Northwest Territories	(Ng et al., 2014)
	carnivores	Portugal	(Conceicao-Neto et al., 2015)
abdomen	dragonflies	Kingdom of Tonga	(Rosario et al., 2011)
		Australia, New Zealand and US	(Dayaram et al., 2013b)
		Puerto Rico	(Rosario et al., 2013)
		Florida, USA, Puerto Rico, Kingdom of Tonga, Bulgaria, Germany, Austria, Finland, and Hungary	(Rosario et al., 2012a)
	Florida woods cockroach	Florida, USA	(Padilla-Rodriguez et al., 2013)
	dragonfly and damselfly	Arizona and Oklahoma, USA	(Dayaram et al., 2015b)
insect body	mosquitoes	California, USA	(Ng et al., 2011b)
		China	(Xia et al., 2018).
	whiteflies	Brazil	(Nakasu et al., 2017)
	ticks	Pennsylvania, USA	(Waits et al., 2018)
intestinal contents	urban wild rats	Berlin	(Sachsenroder et al., 2014)
	squirrels	Japan	(Sato et al., 2015)
rectal swabs	European badgers	The Netherlands	(van den Brand et al., 2012)
hepatopancreas tissue	shrimp	Gulf of Mexico	(Ng et al., 2013)
desiccated tissue and nesting material	chick	New Zealand	(Sikorski et al., 2013b)
larvae	dragonfly	New Zealand	(Dayaram et al., 2014)
liver, digestive tract (gut), and fecal samples	mink	China	(Lian et al., 2014)
GI tract sample	Alaskan black-capped chickadee	Alaska, USA	(Hanna et al., 2015)
lung, liver, spleen, kidney and intestinal contents	shrews and rodents	Zambia	(Sasaki et al., 2015)
pharyngeal and anal swab	bats	China	(Wu et al., 2016)
cloacal swab	Eurasian Jay	Hungary	(Kaszab et al., 2018)
tissue	terrestrial arthropods (Hexapoda, Euchelicerata, Myriapoda)	Kenya, USA, Puerto Rico, Victoria British Columbia, Dominican Republic, Saint Barthelemy, Guadeloupe, Nevis	(Rosario et al., 2018)
Sampling from Livestock	Animal	Location	Reference
feces	pigs	USA	(Cheung et al., 2014; Cheung et al., 2013; Shan et al., 2011)
		Germany	(Sachsenroder et al., 2012)
		New Zealand	(Sikorski et al., 2013a)
		Korea	(Kim et al., 2014)
		Madagascar and Cameroon	(Garigliany et al., 2014)
	febrile and anorexic calves	Korea	(Kim et al., 2012)
	turkey with diarrhea	Hungary	(Reuter et al., 2014)
	dromedaries	Dubai	(Woo et al., 2014)
	swine with diarrheal disease	--	(Cheung et al., 2015)

		Japan	(Oba et al., 2017)
	domestic and wild animals	New Zealand	(Steel et al., 2016)
	healthy chicken	Brazil	(Lima et al., 2017b)
blood	healthy cattle	Germany	(Lamberto et al., 2014)
upper respiratory samples	dromedary camels	United Arab Emirates	(Li et al., 2017)
meat products (muscle tissue)	chicken, cow, goat, sheep and camels	USA, Pakistan, Nigeria	(Li et al., 2010a; Li et al., 2011)
plasma	cattle	Northeast China	(Wang et al., 2018)
Sampling from Other Animals	Animal	Location	Reference
feces	cats	California, USA	(Zhang et al., 2014)
	non-human primates	San Francisco zoo, USA	(Ng et al., 2015)
serum samples	dogs	USA	(Kapoor et al., 2012)
blood sample	laboratory rats	China	(Li et al., 2015b)
liver, spleen and bursa of Fabricius	pigeons	Poland	(Stenzel et al., 2014)
liver and intestine	one dead canine pup	Italy	(Decaro et al., 2014)
nasal secretion	diseased horses	USA	(Li et al., 2015a)
fecal, nasopharyngeal secretion and blood	pandas	China	(Zhang et al., 2017)
fresh tissue samples	dogs	Thailand	(Piewbang et al., 2018)
Human sample	Health Notes	Location	Reference
feces	healthy or non-polio-infected acute flaccid paralysis	South Asian, Nigeria, Tunisia, USA	(Li et al., 2010a)
	healthy or diarrhea or vomiting symptoms	Africa	(Garigliany et al., 2014)
	prolonged diarrhea	Brazil	(Castrignano et al., 2013)
	unexplained diarrheal disease	USA and France	(Ng et al., 2015)
	children with diarrhea of unknown etiology	Peru	(Phan et al., 2016)
serum and cerebrospinal fluid	patients with unexplained paraplegia	Malawi	(Smits et al., 2013)
	immunocompromised patients	Italy	(Macera et al., 2016)
cerebrospinal fluid and feces	patients with acute central nervous system infections	Vietnam	(Tan et al., 2013)
cerebrospinal fluid	patients with unexplained encephalitis	Sri Lanka	(Phan et al., 2015)
brain and serum	multiple sclerosis samples	Germany	(Lamberto et al., 2014)
blood	HIV-positive	France	(Uch et al., 2015)
plasma specimens	US blood donors and African bush hunters	USA and Africa	(Li et al., 2010a)
nasopharyngeal aspirates	children with acute lower respiratory tract infections	Chilean	(Phan et al., 2014)
respiratory secretion	a febrile traveler	(returning from Singapore)	(Cui et al., 2017)
Sampling from Plants	Origin	Location	Reference
grapevines	grape	New York, USA	(Krenz et al., 2012)
		California, USA	(Rwahnih et al., 2013)
		South Korea	(Al Rwahnih et al., 2017)
leaves	fava bean	Morocco	(Abraham et al., 2010)
		Ethiopia	(Abraham et al., 2012)
	mulberry	China	(Lu et al., 2015)
		Shanxi, China	(Ma et al., 2015)
	pea	Germany	(Grigoras et al., 2010a,b)
	cassava infected with fungi	Ghana	(Dayaram et al., 2012)
	citrus	Italy	(Loconsole et al., 2012)
	tomato, bean, weeds and ornamental plants	Morocco, Spain and Italy	(Belabess et al., 2015)
	Bromus hordeaceus and *Trifolium resupinatum*	New Zealand and France	(Kraberger et al., 2015b)
	Nicotiana tobacum	China	(Yang et al., 2015)
	soybean	USA	(Marzano and Domier, 2016)
	Poaceae plants	Kingdom of Tonga	(Male et al., 2015)
Plant samples	fava bean	Tunisia	(Kraberger et al., 2018)
		Denmark	(Gaafar et al., 2018)
	Hypericum japonicum	Vietnam	(Du et al., 2014)
	alfalfa and *E. caput-medusae*	France, Spain	(Bernardo et al., 2016)
	Sophora alopecuroides L.	Iran	(Heydarnejad et al., 2017)

	wild rice plants	Australia	(Kraberger et al., 2017)
	pea	The Netherlands	(Gaafar et al., 2017)
	Passiflora sp.	Brazil	(Fontenele et al., 2018)
	symptomatic tomato plants	Argentina	(Medina et al., 2018)
leafy or woody tissue	apple, pear and grapevine	Brazil	(Basso et al., 2015)
Freshwater, marine, sediment and aquatic eukaryote samples		Location	Reference
water		Japan	(Kimura and Tomaru, 2013)
			(Tomaru et al., 2013)
		sewage oxidation pond, New Zealand	(Kraberger et al., 2015a)
		Florida, USA	(McDaniel et al., 2014)
		Tara Oceans	(Roux et al., 2016)
		Lake Superior, Lake Erie, Lake Michigan, USA	(Roux et al., 2016)
		Lake Bourget, Lake Pavin, France	(Roux et al., 2012)
sediment		Japan	(Kimura and Tomaru, 2015)
		Arctic shelf seafloor	(Nguyen et al., 2017)
water and sediment		Japan	(Tomaru et al., 2012; Toyoda et al., 2012)
diatoms		Japan	(Tomaru et al., 2011a,b)
estuarine mollusc		New Zealand	(Dayaram et al., 2013a)
copecods		Florida, USA	(Dunlap et al., 2013)
net plankton		Atlantic and Pacific Ocean	(Eaglesham and Hewson, 2013)
benthic amphipod *Diporeia* spp.		North American lakes	(Hewson et al., 2013)
benthic and bank river sediments		New Zealand	(Kraberger et al., 2013)
deep-sea vents		Japan	(Yoshida et al., 2013)
Red benthic mat dominated by filamentous cyanobacteria		a freshwater pond on McMurdo Ice Shelf, Antarctica	(Zawar-Reza et al., 2014)
white plague coral		US Virgin Islands	(Soffer et al., 2014)
zooplankton (ctenophores)		coastal Georgia	(Breitbart et al., 2015)
molluscs and benthic sediments		Avon-Heathcote estuary in New Zealand	(Dayaram et al., 2015a)
Forbes sea star with sea star wasting disease		Rhode Island	(Fahsbender et al., 2015)
marine invertebrates		USA	(Rosario et al., 2015a)
aquatic invertebrates, insect larvae, benthic sediments, lake water		Lake Sarah, New Zealand	(Dayaram et al., 2016)
asteroid, echinoid, and holothurian tissues		North America	(Jackson et al., 2016)
sewage		Florida, USA	(Pearson et al., 2016)
sewage and reclaimed water		Brazil	(Castrignano et al., 2017)
amphipods		North American lakes	(Bistolas et al., 2017b,c)
eel and sichel		Hungary	(Borzak et al., 2017)
Sampling from Other Environments		Location	Reference
fungal		China	(Yu et al., 2010)
		Australia	(Mu et al., 2017)
air		Korea	(Whon et al., 2012)
soil		Scotland	(Reavy et al., 2015)

These many references evince the immense efforts that virologists worldwide have made to sample and identify the potential eukaryotic CRESS DNA virus genomes. While the earth's virome may still be largely unknown, these efforts have better defined eukaryotic CRESS DNA viral diversity and prevalence, leading to improved systematics and a sense of their relatedness (see FIGURE 3.4, above). However, it has been increasingly frustrating that as genetic knowledge of these viruses increase, we lack commensurate information about these viruses' phenotype.

Confirmed and potential host range and pathogenesis of eukaryotic CRESS DNA viruses

The three families of eukaryotic CRESS DNA viruses that have been established for the longest time contain well-studied pathogens of plants and animals, though not all members of these families cause disease in all (or any) of the hosts they productively infect. Geminiviruses and nanoviruses infect plants, while circoviruses infect both vertebrates (mammals and birds) and invertebrates.

Plant infections

The speciose family *Geminiviridae* has members that can infect a wide range of plant hosts, from many plant families including Acanthaceae (e.g., Chinese violet), Amaranthaceae (e.g., sugar beet), Apocynaceae (e.g., golden trumpet), Asteraceae (e.g. zinnia), Bignoniaceae (e.g., yellow trumpetbush), Brassicaceae (e.g., cabbage), Capparaceae (e.g., spider flower), Caprifoliaceae (e.g., honeysuckle), Caricaceae (e.g. papaya), Convolvulaceae (e.g., sweet potato), Cucurbitaceae (e.g., squash), Cyperaceae (e.g., Nees weeping lovegrass), Dioscoreaceae (e.g., yam), Euphorbiaceae (e.g., cassava), Fabaceae (e.g., soybean), Gentianaceae (e.g., lisianthus), Lamiaceae (e.g., mint), Linderniaceae (e.g., false pimpernel), Malvaceae (e.g., cotton), Meliaceae (e.g., chinaberry tree), Moraceae (e.g., mulberry), Nyctaginaceae (e.g., red boerhavia), Oleaceae (e.g., Arabian jasmine), Onagraceae (e.g., Mexican primrose-willow), Oxalidaceae (e.g., pink woodsorrel), Papaveraceae (e.g., opium poppy), Passifloraceae (e.g., passionfruit), Phyllanthaceae (e.g., star gooseberry), Plantaginaceae (e.g., ribwort plaintain), Poaceae (e. g., maize), Polygalaceae (e.g., dainty butterfly bush), Rosaceae (e.g., rose), Rubiaceae (e.g., *Hedyotis uncinella*), Rutaceae (e.g., lemon), Sapindaceae (e.g., *Deinbollia borbonica*), Solanaceae (e.g, tomato), Urticaceae (e.g., ramie), Verbenaceae (e.g., pigeon berry), and Vitaceae (e.g., grapevine). The true host range of geminiviruses may be even larger, since there has been a sampling bias towards identifying symptomatic crop viruses and families without cultivated species that attract geminivirus insect vectors may also be susceptible to infection.

When infections are symptomatic, geminivirus cause yellowing of leaves (streaking, mosaicism (FIGURE 3.1), mottling) and distortions of the leaves (crumpling, curling, stunting, etc.), both of which interfere with the host plant's ability to conduct photosynthesis (Inoue-Nagata et al., 2016). Characterized geminiviruses are named for their plant host of isolation and symptoms, although many of these viruses infect more than one host and pose a series of symptoms, such that the species name may not reflect its predominant host or most typical symptoms in nature (Brown et al., 2015). Many affected plant species are economically and agriculturally important crops, so geminivirus infections can lead to economic losses and famine. For example, cassava is the third most important staple food for people living in the tropics, and more than 800 million people in Africa, Asia and Latin America depend on this plant for food and income (Legg et al., 2015). Cassava can withstand unfavorable soil conditions and drought, ensuring food security in marginal agricultural areas (Thresh and Cooter, 2005). In sub-Saharan Africa, cassava production is limited by a number of begomoviruses that cause cassava mosaic disease – the yellowing and distortion of the plant leaves prevents efficient starch production, and stunted tubers have sharply reduced yield compared to uninfected cassava (Alabi et al., 2011). The routine concern of cassava mosaic disease can worsen when the viruses evolve quickly (see Evolution, below), for instance when a recombinant between African cassava mosaic virus and East African mosaic virus

(EACMV-UG) emerged, causing >90% crop loss in Uganda in 1997, leading some to starve in the resulting famine (Zhou et al., 1997).

Eighty-eight percent of classified geminiviruses are begomoviruses (TABLE 3.2), leading to an understandable bias in the literature towards this genus. Much of the work on geminivirus pathogenicity has been done in begomovirus-dicot host systems, including the model Arabidopsis (Hanley-Bowdoin et al., 2013), but many discoveries have also been made through field surveys. For instance, the symptoms of geminivirus infections can be altered by the presence of satellites. To date, four groups of geminivirus associated satellites have been described, including alphasatellites, betasatellites, gammasatellites, and deltasatellites. The majority of described satellite species are found in association with monopartite begomoviruses (Zerbini et al., 2017). However, alphasatellites, betasatellites, and deltasatellites have been found with bipartite begomoviruses (Fiallo-Olivé et al., 2012, Lozano et al., 2016). Notably, alphasatellites have also been found with mastreviruses (Kumar et al., 2014), and thus geminivirus associated satellites are not limited to members of the genus *Begomovirus*. All four kinds of satellites are circular ssDNA with a stem-loop origin of replication (see Molecular biology of Rep, above).

The protein-coding alphasatellites and betasatellites are the best studied geminivirus-associated satellites. Alphasatellites are ~1300 nt, roughly half the length of a geminivirus genome (or genomic segment in the case of the bipartite begomoviruses). They encode their own Rep protein and autonomously replicate. Their Rep protein is closely related to that of nanoviruses, and the nonanucleotide in their stem-loop origin of replication matches that of nanoviruses as well (Rosario et al., 2012b). However, when coinfecting with either a geminivirus or nanovirus, an alphasatellite can be encapsidated and transmitted. The roles that alphasatellites play in the infection dynamics have not been definitively determined, with some research suggesting they reduce begomovirus titer, thus prolonging the length of infection (Idris et al., 2011), while others suggest they help suppress silencing (Nawaz-ul-Rehman et al., 2010). Recently alphasatellites have been classified in to a virus family, *Alphasatellitidae* (included with *Nanoviridae* in FIGURE 3.4), with two genera reflecting their associated hosts: *Geminialphasatellitinae* and *Nanoalphasatellitinae* (Briddon et al., 2018). The most frequently encountered satellites, at least in the Old World, are betasatellites, which hijack both the capsids and replication initiation processes of their coinfecting geminiviruses (Sattar et al., 2013). A betasatellite is also usually half the size of a geminivirus genome (~1300nt) and has a stem-loop that is recognized by a geminivirus Rep [they encode the predominant begomovirus nonanucleotide (Rosario et al., 2016)]. It encodes a single protein, bC1, which can affect symptom severity and affect host silencing (functions thoroughly reviewed/listed in (Sattar et al., 2013, Zhou, 2013)). Geminivirus betasatellites are currently classified within the family *Tolecusatellitidae*, genus *Betasatellite*.

The remaining two groups of satellites are comparatively poorly studied and encompass non-protein encoding satellite molecules. Gammasatellites and deltasatellites are smaller (~700nt) than alphasatellites and betasatellites and seem to be largely restricted to the New World (Fiallo-Olivé et al., 2012, Fiallo-Olivé et al., 2016, Lozano et al., 2016, Rosario et al., 2016). Deltasatellites have been isolated from plants infected with either monopartite or bipartite begomoviruses (Fiallo-Olivé et al., 2012, Lozano et al., 2016). On the other hand, gammasatellites were discovered from the begomovirus whitefly vector through molecular analysis and, thus, have not been associated with a particular geminivirus (Rosario et al., 2016). The effects of these non-protein coding satellites on begomovirus infection are not yet known, though deltasatellites may decrease begomovirus accumulation in the host plant (Fiallo-Olivé et al., 2016). The nonanucleotide in the origin of replication for all identified gammasatellites and deltasatellites matches that of the begomoviruses (Fiallo-Olivé et al., 2016, Rosario et al., 2016). Gammasatellites were named based on the Greek alphabetical order since they were discovered after alpha and betasatellites. However, the first report of deltasatellites was almost simultaneous to that of gammasatellites and their name comes from their apparent derivation from betasatellites (delta, in the sense of a change from betasatellites). Currently there is no formal taxonomic classification for either gammasatellites or deltasatellites. Since both groups refer to small, non-protein coding satellites, it remains to be determined if both groups will be merged or will remain separate. Further phylogenetic analyses and biological characterization of these small satellites should shed light on this issue.

Nanovirus pathogenicity has been understudied compared to geminiviruses. They are known to infect a smaller number of plant families: Arecaceae (coconut), Caricaceae (papaya), Fabaceae (cow vetch, fava bean, pea, sophora root, subterranean clover), Musaceae (abaca, banana), Solanaceae (tobacco) and Zingiberaceae (cardamom), though they have been found to be most problematic in leguminous hosts (Fabaceae). This known host range has the same caveat that symptomatic crops have disproportionately been sampled and uncultivated plants may also be susceptible to nanoviruses. The most pronounced symptoms of nanovirus infection are stunting, yellowing and leaf rolling, sometimes followed by necrosis (Abraham et al., 2012). Members of the genus *Nanovirus* are transmitted by two aphid vectors, *Aphis craccivora* (Koch) and *Acyrthosiphou pisian* (Harris), and alphasatellites can also associate with the already multipartite nanovirus genomes (Kraberger et al., 2018). Nanoviruses have been reported in countries throughout North and Eastern Africa, Europe, the Middle East, China, Japan and Australia (Abraham et al., 2010, Abraham et al., 2012, Babin et al., 2000, Gaafar et al., 2017, Grigoras et al., 2010a, Grigoras et al., 2014, Kraberger et al., 2018, Kumari et al., 2009, Makkouk and Kumari, 2009, Vetten, 2008).

The symptoms of banana bunchy top virus, the type species of genus *Babuvirus* was first reported in Fiji in 1889 (Magee, 1927, Stover, 1972). This virus, which is transmitted by the banana aphid (*Pentalonia nigronervosa*, (Magee, 1927)), has been recognized as a pathogen that affects banana production worldwide (Dale, 1987). Infected plants show significant reduction in size, and produce stunted fruits or fail to fruit (Hooks et al., 2008). Efforts to control banana bunchy top virus are complicated by asymptomatic infections, the virus' ability to infect other uncultivated hosts and long incubation periods, but molecular surveillance has enabled earlier identification of infected plants (Allen, 1978, Dale and Harding, 1998, Hooks et al., 2008).

Animal infections

The family *Circoviridae*, genus *Circovirus* contains well-studied pathogens like PCV2 and BFDV. Since cycloviruses are largely known only from metagenomics efforts, and their hosts are unknown, it is premature to discuss their pathogenicity. The host-virus interactions of porcine circoviruses in particular are well studied, starting with PCV1. Although PCV1 is not pathogenic in its typical porcine hosts, it was characterized as contaminant in a pig kidney cell line (PK-15) and studied for its unusual genomic architecture (Tischer et al., 1974). PCV2 contributes to several virulent diseases of pigs, notably post-weaning multisystemic wasting syndrome in piglets – a fatal disease that became epidemic in the late 1990s (Allan et al., 1999, Harding and Clark, 1997, Nayar et al., 1997). Retrospective studies found PCV2-antibodies in serum from Belgium as early as 1969 (Sanchez et al., 2001). Subsequently it was recognized that PCV2 is the leading causative agent for a collection of syndromes known as PCAD, including respiratory diseases, enteric diseases (porcine dermatitis and nephropathy syndrome) and reproductive problems (Opriessnig et al., 2007). PCV2 efficiently spreads horizontally through contact with respiratory, oral, urinary secretions and feces (Magar et al., 2000, Gillespie et al., 2009, Rose et al., 2012), and in rare cases, can transmit vertically from mother to piglets (Maldonado et al., 2005, Shen et al., 2010). PCV2 infects components of the immune system (Choi and Chae, 1999, Vincent et al., 2003), leading to depleted levels of lymphocytes in infected animals. PCV2 is now endemic globally and antibodies are found in up to 100% of pigs (Walker et al., 2000), but a much smaller percentage show symptoms of PCAD. This is likely because PCAD seems polymicrobial, and requires co-infection by PCV2 and another microbe – an RNA virus, another ssDNA virus (including porcine parvovirus, which has a linear genome), or even bacteria (Rose et al., 2012).The initial epidemics of PCV2 were caused by one strain (PCV2a) and widespread vaccination was implemented to prevent piglets from succumbing to PCAD. In the wake of this successful intervention, the prevalent genotype shifted to PCV2b, which is thought to be less virulent (Rose et al., 2012). Vaccination against PCV2b has led to another strain replacement, with PCV2d (Opriessnig et al., 2017).

The etiological agent for psittacine beak and feather disease is BFDV. Infected birds can exhibit many kinds of symptoms: peracute, acute, chronic and subclinical, depending on their age. Neonates and fledglings (young birds) typically show peracute and acute symptoms with high mortality rates (Doneley, 2003, Ritchie et al., 1989, Schoemaker et al., 2000). In chronic cases in more mature birds, beak and feather deformities are observed (FIGURE 3.5) and most birds become immunocompromised and become susceptible to secondary infections (Pass and Perry, 1984, Ritchie et al., 1989). Chronic symptoms include lethargy, depression, and diarrhea, which helps shed BFDV virions (Fogell et al., 2016). There is currently no cure, treatment or vaccine available for psittacine beak and feather disease (Regnard et al., 2017, Robino et al., 2014). Just as for PCV2, BFDV can transmit horizontally through contact with infected secretions, for instance on nesting material (Gerlach, 1994, Ritchie et al., 2003), and rarely transmit vertically (Rahaus et al., 2008, Todd, 2004). BFDV is endemic to Australia and has become a global concern due to legal and illegal trades of psittacine species (parrots, cockatoos, parakeets, Raidal et al., 2015). Phylogenetic analyses have confirmed the historical record, and shown that BFDV originated from Australia and then spread to the rest of the world (Harkins et al., 2014, Pass and Perry, 1984, Raidal et al., 2015). BFDV can infect at least 60 species within the order Psittaciformes (Harkins et al., 2014), and it is considered capable of emerging in other parrot species, including many “exotic” endangered species (Raidal et al., 2015, Sarker et al., 2015b, Sarker et al., 2015a).

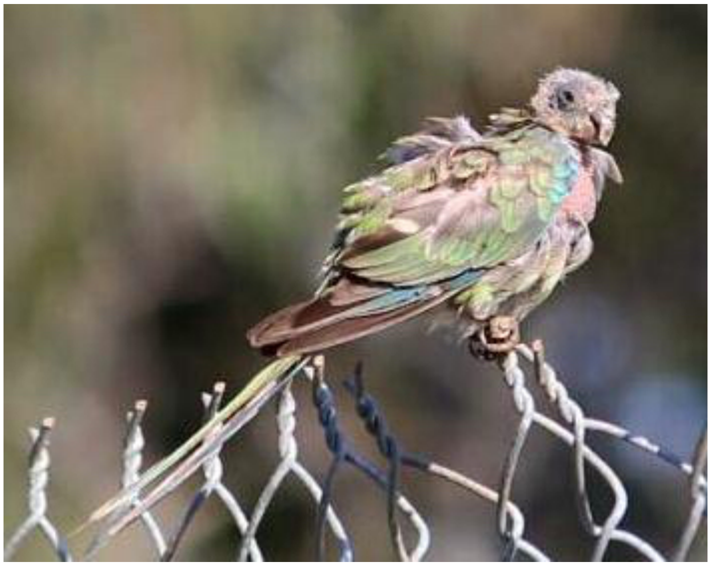


FIGURE 3.5. Symptoms of psittacine beak and feather disease in a Red Rump Parrot. Public Domain (Wikimedia Commons)

Potential pathogens

Because of the lack of biological characterization among the newly identified CRESS DNA viruses, it is difficult to identify their roles in the biosphere. While viruses must use host resources to replicate, not all viruses significantly impact the fitness of their hosts, and the effects of some viral infections help their hosts survive and reproduce (Roossinck, 2011). Therefore, even if a eukaryotic CRESS DNA virus found in association with a host truly infects that host, it may not cause detectable disease symptoms. Indeed, one of the consequences of sequencing-based surveys of ecosystems is that a large number and diversity of viruses are uncovered in healthy hosts – the methods are ideal for detecting benign or latent viruses (Roossinck et al., 2010).

It is nevertheless intriguing to many researchers that sequenced eukaryotic CRESS DNA viruses may impact health and disease, especially in humans (Zhao et al. 2017). CRESS DNA viruses have been discovered in association with several cases of human diarrhea, though these viruses may not be linked to illness and may be present in human waste because they infect food ingested by the study subjects (TABLE 3.3). Other studies have found eukaryotic CRESS DNA viral genomes isolated from human cerebrospinal fluids. The tentatively named cyclovirus-Vietnam (CyCV-VN) was first found in patients with acute central nervous system infections in Vietnam, but was then also detected in fecal samples from humans, pigs and poultry (Tan et al., 2013). That same year, human cyclovirus VS5700009 (closely related to CyCV-VN (Sasaki et al., 2015)) was independently found in the cerebrospinal fluid of patients with unexplained paraplegia in Malawi (Smits et al., 2013). Others have subsequently screened for CyCV-VN in cerebrospinal fluid of diseased patients without finding it: in northern Vietnam, Cambodia, Nepal and The Netherlands (Le et al., 2014). CyCV-VN still exists in association with humans – it was found in healthy children’s feces and pig feces in Africa (Garigliany et al., 2014) and in the blood but not cerebrospinal fluid samples of immunodeficient Italian men (Macera et al., 2016). There is still a lack of evidence for any disease-causing properties and follow up studies concerning CyCV-VN, and many aspects of Koch's postulates remain unfulfilled for this potential human pathogen.

Outside of looking for human (and other animal) pathogens, researchers would like to understand the current role and potential of eukaryotic CRESS DNA viruses in our ecosystems, such as the ocean. As some eukaryotic CRESS DNA viruses can infect diatoms, they could possibly be used as biological control agents to stop the onset of some harmful algal blooms. A number of bacilladnaviruses were found infecting the most abundant genus of diatoms, *Chaetoceros* (Kimura and Tomaru, 2013, Kimura and Tomaru, 2015, Nagasaki et al., 2005, Tomaru et al., 2008, Tomaru et al., 2011a, Tomaru et al., 2011b, Toyoda et al., 2012, Tomaru et al., 2013), which means these viruses could potentially be used worldwide to cure locations of toxic, anoxic algal overgrowth. Since so many eukaryotic CRESS DNA viruses have been discovered in association with aquatic invertebrates (ctenophores, sea stars, sea urchins, etc.), these viruses may play important roles in food web dynamics and biogeochemistry in aquatic systems (Rosario et al., 2015a).

Endogenized eukaryotic CRESS DNA viruses

Viruses have played many roles in eukaryotic evolution, but the age of genomic sequencing has revealed that eukaryotic genomes have more abundant and more diverse endogenized viral sequences than previously thought. Some viral genomic architectures lend themselves to frequent endogenization, such as retroviruses, which comprise ~8% of the human genome. Human endogenized retroviruses are genomic fossils, which have evolved at the slower rate of human evolution since their integration, and thus provide good information on the deep evolutionary history of retroviruses. Genomic fossils provide information on host jumps and host-virus interactions including arms races (Hayward and Katzourakis, 2015). There have been some documented benefits to endogenized retroviruses as well, as eukaryotes have incorporated the viral genes and proteins into their functional systems. The best known example of an endogenized retroviral protein evolving to serve a critical function for the host would be the cooption of syncitin for host placenta morphogenesis, something that has occurred multiple independent times in the evolution of mammals (Mi et al., 2000). While some viruses routinely integrate into their hosts' genomes, and their occasional invasion of the germ line would be mechanistically understandable (retroviruses, some large DNA viruses), genomic fossils are being discovered in eukaryotes that are related to all types of virulent viruses, including RNA and ssDNA viruses. These sequences stand in opposition to our current understanding of these viruses' replication and inability to integrate into host genomes. Regardless they exist, even if they are the product of very small chance events [see several hypothetical mechanisms in (Krupovic and Forterre, 2015)], given many chances over the long arc of evolutionary time. The study of endogenized viruses comprises a major part of paleovirology, which not only allows us to know more about the origin and evolutionary history of viruses, but also how virus integration may have affected the evolutionary history of their hosts (Feschotte and Gilbert, 2012).

Many endogenized partial eukaryotic CRESS DNA viral genomes (most often a sequence with homology to Rep) have been founds in all eukaryotic supergroups (Dennis et al., 2018a, Kryukov et al., 2018). One of the earliest examples was a geminivirus Rep homolog in several tobacco species genomes, suggesting an integration event in a common ancestor more than a million years ago (Bejarano et al., 1996, Ashby et al., 1997). Other groups have found multiple cases of circovirus-like sequences in animal host genomes (Belyi et al., 2010, Katzourakis and Gifford, 2010). Comprehensive searches in eukaryotic genomes have found endogenized CRESS DNA virus-like sequences inside genomes of plants, fungi, animals and protists, suggesting these may have been hosts for eukaryotic CRESS DNA viruses in the past (Liu et al., 2011). More recent studies have found many endogenized circovirus-like elements inside host genomes (Dennis et al., 2018a). A thorough scan of ~4000 eukaryotic genomes with all non-retroviral virus sequences found sequences homologous to ssDNA viruses endogenized in all eukaryotic supergroups, but nearly half of the hits were in plant genomes (Kryukov et al., 2018). It is not yet known how active these virus-like sequences are within their hosts, and if they affect their hosts' fitness. Some of the more conserved sequences suggest that either the endogenization event was relatively recent, or selection has maintained the function of the sequence (Filloux et al., 2015). One study has reported geminivirus-like endogenized sequences in yam that are active, producing small RNA transcripts (Filloux et al., 2015).

Since so many eukaryotic CRESS DNA viruses do not have a definitive host, endogenized homologous sequences can give researchers an idea of what kind of host these viruses used to or could still infect (Aiewsakun and Katzourakis, 2015). These genomic fossils may point researchers towards fruitful hosts for isolating viruses that can be brought into the lab and characterized. This has already begun for those studying endogenous circoviral elements (Dennis et al., 2018b). Researchers have found endogenous sequences in host genomes cluster with exogenous contemporary viruses infecting similar hosts such as birds compared to mammals or fish (Dennis et al., 2018a). Endogenized cyclovirus elements have further confirmed their potential infectivity of insects, as endogenous sequences are similar to those found associated with arthropods (Dayaram et al., 2013b, Dayaram et al., 2014, Rosario et al., 2011, Rosario et al., 2012a, Rosario et al., 2018).

In addition to informing host range, these genomic fossils help provide depth to evolutionary studies of eukaryotic CRESS DNA viruses. The integrated Rep sequence in tobacco plants, for instance, indicates that gemini-like viruses were present in South America (where the tobacco plants diversified) at least 1.8 million years ago (Lefeuvre et al., 2011). This contradicts what researchers assumed from the modern distribution of whitefly-transmitted geminiviruses, where the begomoviruses in the New World appear to be descended from the more diverse Old World begomoviruses (Nawaz-ul-Rehman and Fauquet, 2009). These genomic fossils can inform where and when the ancestors of all circulating related viruses existed, even if the rapid evolution of eukaryotic CRESS DNA viruses and coalescence have eliminated all traces of that history from their current sequences. As more eukaryotic genomes are sequenced, the greater the opportunity will be to identify more endogenized eukaryotic CRESS DNA virus-like elements, and eukaryotic CRESS DNA paleovirology is likely to grow in the upcoming decade.

Evolution of Eukaryotic CRESS DNA Viruses

Eukaryotic CRESS DNA viruses evolve quickly

Evolutionary study of eukaryotic CRESS DNA viruses is not restricted to paleovirology and untangling the deep phylogenetic relationships among families. Some eukaryotic CRESS DNA viruses are emergent pathogens, and the year-to-year evolution of these viruses impacts food security. For instance, novel begomoviruses have been a persistent emerging problem in crops including tomato (Ribeiro et al., 2003). Like emergent RNA viruses, eukaryotic CRESS DNA viruses have been shown to evolve quickly, a signal that is even evident in datasets where statistically detectable recombination has been removed (the effects of recombination are discussed below). Representative lineages have been measured evolving quickly over a period of years: *Geminiviridae*: Tomato yellow leaf curl virus (TYLCV, Duffy and Holmes, 2009), Maize streak virus (MSV, van der Walt et al., 2008) and Sugarcane streak Reunion virus (Harkins et al., 2009a), *Nanoviridae*: Faba bean yellow necrotic virus (Grigoras et al., 2010b). The rates of evolution have been estimated by computational methods for many more species, including members of *Circoviridae* (Firth et al., 2009, Kundu et al., 2012) and *Geminiviridae* (Duffy and Holmes, 2008, Harkins et al., 2009b). High substitution rates in CRESS DNA viruses do not cause commensurate change in protein-coding genes over long periods of evolutionary time. This is expected as part of the generalizable time-dependence of substitution rates – selection to retain function and saturation, especially of third codon positions, affect the measurable substitution rate over longer timespans (Aiewsakun and Katzourakis, 2016). This time dependence explains how mastreviruses may have diverged with their hosts, despite calculated substitution rates that are orders of magnitude slower than observable mastrevirus evolution (Wu et al., 2008).

What is less understandable is how eukaryotic CRESS DNA viruses achieve these high, RNA virus-like substitution rates. RNA virus polymerases mutation rates drive their high substitution rates, but ssDNA viruses replicate with their host polymerases that are not thought to have high error rates (Duffy et al., 2008). As it is easier to measure the mutation rate of CRESS DNA viruses of bacteria than of eukaryotes (due to difficulty knowing and controlling generation times), no rigorous mutation rates have been measured for eukaryotic CRESS DNA viruses. Phages phiX174 and M13 have both been shown to mutate more rapidly than their host *Escherichia coli*, despite the fact that they all use the same DNA polymerase for genome replication (Cuevas et al., 2009, Sanjuán et al., 2010). PhiX174 avoids host DNA repair (Cuevas et al., 2011), but their ssDNA virus mutation rates are still 10 to 100-fold lower than the RNA virus mutation rates that explain their high substitution rates. Eukaryotic CRESS DNA viruses may have similar mutation rates to their phage counterparts, but the lack of mechanistic insight with phage offers nothing to assist in understanding how eukaryotic CRESS DNA viruses achieve the mutation rates necessary for their fast rates of evolution.

Circovirus researchers had noted fifteen years ago that unpaired single-stranded DNA bases could be oxidatively damaged, leading to high rates of transition for cytosines and adenines, which could bolster the baseline mutation rate of BFDV (Ritchie et al., 2003). Phylogenetic and experimental studies of geminiviruses confirmed that cytosine to thymine transitions were elevated compared to the other kinds of transitions (Tomato yellow leaf curl China virus, Ge et al., 2007), TYLCV (Duffy and Holmes, 2008), East African cassava mosaic virus (EACMV, Duffy and Holmes, 2009), Sugarcane streak Reunion virus (Harkins et al., 2009a), but other substitution biases with the potential to be due to oxidative damage (i.e., guanine to thymine transversions) were more frequently observed than the predicted adenine transitions (TYLCV

(Duffy and Holmes, 2008), EACMV (Duffy and Holmes, 2009), MSV (van der Walt et al., 2008)). In begomoviruses, these substitution biases are strand-specific, indicating that the oxidative damage might occur while eukaryotic CRESS DNA viruses are encapsidated. That the cytosine transitions occur on the packaged, virion strand was cleanly demonstrated by an examination of begomovirus codon usage, which is biased towards thymine-ending codons in the mRNA for virion sense CP, but adenine-ending codons for antisense Rep, which correspond to thymines in the virion complement of the Rep gene (Cardinale et al., 2013). While it seems likely that oxidative damage contributes to higher CRESS DNA virus mutation rates, it does not appear that it explains all of the difference between expected mutation rates and measured substitution rates. This open question would benefit greatly from accurate measurement of mutation rate within eukaryotic hosts. Data indicate that geminivirus mutation rates might be very high in plants (Arguello-Astorga et al., 2007), and eukaryotic CRESS DNA viruses have highly diverse populations within single hosts (Sánchez-Campos et al., 2018, Sarker et al., 2014). Technological advances are needed to help the field move from high mutation frequencies to accurately measured mutation rates.

Recombination

Long before high mutation rates were implied in eukaryotic CRESS DNA viruses, these viruses were known to be capable of frequent recombination, and successful recombinant viruses were often isolated (Lefeuvre and Moriones, 2015). Even in the small, 1 kb segments of nanoviruses, statistically detectable recombination is found (Grigoras et al., 2014, Hyder et al., 2011). Recombination can bring more genetic change into a genome at once than a point mutation, potentially causing large, swift phenotypic changes. As mentioned earlier, the genus *Curtovirus* was the result of a successful recombination between ancestors with a begomovirus-like Rep and a mastrevirus-like CP (Rybicki, 1994) – this caused a change in vector from whitefly to leafhopper, a phenotype that could not be conferred with a point mutation. A similar whole-gene recombination birthed the recently emerged porcine circovirus 3, which has a PCV Rep and an avian circovirus CP (Franzo et al., 2018). In chunks of whole genes (or more than one gene) or smaller portions of the genome, statistically detectable recombination is prevalent in eukaryotic CRESS DNA viruses. Undetectable recombination between nearly identical viruses also likely occurs as well, making the frequent recombination observed a conservative estimate of its occurrence.

Both mutation and recombination are known to occur more often in hot spots, and beyond their non-random occurrence, natural selection purges most mutations (and products of recombination) such that surviving sequences show clear signals of regions where recombination is well-tolerated (Lefeuvre et al., 2007). For instance, across eukaryotic CRESS DNA viruses, recombination in intergenic regions is favored compared to recombination within a protein-coding gene. The canalization of recombination hot spots is more pronounced within family and genus. Despite nucleotide dissimilarity, the same regions are recombination hot spots within *Geminiviridae* (Lefeuvre et al., 2009), and separately, within *Circovirus* (Stenzel et al., 2014). The ambisense nature of many eukaryotic CRESS DNA viruses may be one reason for frequent recombination. When a gene is simultaneously replicated and transcribed, the interaction of the enzymes causes a pausing, and paused polymerases are associated with template switching recombination events (Martin et al., 2011). The best data that supports this idea is that there are more recombination breakpoints detected in the antisense genes in ambisense genomes – the orientation where it would be easiest to have replication and transcription approach each other from opposite directions (Martin et al., 2011). While eukaryotic CRESS DNA viruses are able to recombine quite often, mutation still accounts for most of the diversity seen in begomovirus populations worldwide (Lima et al., 2017a).

The multipartite plant CRESS DNA viruses also can experience reassortment (previously called pseudorecombination), wherein a segment from one virus can be used in a successful infection of another. The sequence similarity of the origin of replication in the Rep-encoding segment and the reassorted segment determines whether the segments can productively infect the host – the Rep protein must still be able to recognize the new segment’s intergenic region including its stem-loop origin and nonanucleotide (Martin et al., 2011).

Migration

Researchers have studied the epidemiology of eukaryotic CRESS DNA virus spread around the globe to understand the current patterns of infection and to help prevent pathogens from moving into key agricultural areas. The Old World-New World biogeography of begomoviruses means that it has been easy to see when a virus from the Old World (Tomato yellow leaf curl virus, TYLCV) has migrated into the Americas, as it did to Hispaniola the early 1990s (Mabvakure et al., 2016). From there, TYLCV spread throughout the Caribbean, into Central and North America and at the same time was re-introduced to the west coast of Mexico from Asia (Duffy and Holmes, 2007). The only New World begomovirus that has successfully migrated out of the Americas is Squash leaf curl virus, which is now a problem in Cucurbit production throughout Asia Minor (Lapidot et al., 2014).

Modern phylogenetic software packages have made it possible to examine the likely pathways viruses take as they spread worldwide. Phylogeographic analyses have been conducted for members of the three most well-characterized families with a focus on the geminiviruses MSV and Panicum streak virus (Varsani et al., 2009), TYLCV (Lefeuvre et al., 2010, Mabvakure et al., 2016), East African cassava mosaic virus (De Bruyn et al., 2012), and Sweet potato leaf curl virus (Kim et al., 2018). Livestock trading among countries has been shown to be significant in PCV2 phylogeography (Firth et al., 2009, Vidigal et al., 2012), and the pet trade undoubtedly helped the spread of BFDV out of Australia (Harkins et al., 2014), but banana bunchy top virus (BBTV) has appeared to rarely move long distances, suggesting the modern banana trade is not responsible for the current distribution of BBTV (Stainton et al., 2015). As increased sampling reveals more about the location of related eukaryotic CRESS DNA viruses that are not associated with diseases, some of these same techniques might be applied to viruses that are not necessarily pathogens, but for now the largest datasets that cover the longest periods of sampling are all from pathogens in *Geminiviridae, Circoviridae* and *Nanoviridae*.

Conclusions

Eukaryotic CRESS DNA viruses have been on the vanguard of the transition from detailed, molecular characterization of novel viruses to taxonomy by sequence similarity alone. Their small genome size, prevalence and affinity for rolling-circle replication allow easy molecular surveillance and high return on sampling effort. Virologists can now appreciate their incredible sequence diversity: varied genomic organization, divergence of the homologous Rep protein and novel viral proteins, which are sometimes unlike anything else previously sequenced (part of viral "dark matter," Krishnamurthy and Wang, 2017). There is also no sign that the diversity of eukaryotic CRESS DNA viruses has been thoroughly explored, and there is the strong potential for even more novel species, genera and families to be discovered with increased sampling.

The burgeoning number of families of eukaryotic CRESS DNA viruses reflects some of the extant diversity of this widespread group of viruses. While each of these families (and genera) have important idiosyncrasies, the shared evolutionary history of their Rep protein provides important unity within this group that is deserving of recognition by ICTV as a higher level of taxonomy.

The sharp, recent expansion of our knowledge of eukaryotic CRESS DNA viral diversity was partially a function of how little attention was paid to this genomic architecture for several decades. The diversity of all viral groups will likely expand as more affordable sequencing facilitates larger surveys. However, CRESS DNA viruses have served as the canaries in the metagenomics mine for how cheap sequencing and expensive molecular virology skews our understanding of viruses towards bioinformatics and away from host-virus interactions and ecology. Without dedicated effort to cultivate and study representatives of these families in the laboratory, virologists will still know comparatively less about this widespread, diverse and intriguing viral group.

Acknowledgements

All authors acknowledge funding from the NSF Assembling the Tree of Life program (DEB-1239976 and DEB-1240049) and SD acknowledges funding from NSF OIA 1545553.

The nomenclature for this group of viruses has evolved over the last decade, and the authors would like to acknowledge the input of several ICTV members at the 2018 American Society for Virology meeting for the suggestions, including Poliane Alfenas-Zerbini, Mart Krupovic, Arvind Varsani and Murilo Zerbini.

References

Abraham, A.D., Bencharki, B., Torok, V., Katul, L., Varrelmann, M., Josef Vetten, H., 2010. Two distinct nanovirus species infecting faba bean in Morocco. Arch. Virol. 155, 37-46.

Abraham, A.D., Varrelmann, M., Josef Vetten, H., 2012. Three Distinct Nanoviruses, One of Which Represents a New Species, Infect Faba Bean in Ethiopia. Plant Dis. 96, 1045-1053.

Aiewsakun, P., Katzourakis, A., 2015. Endogenous viruses: Connecting recent and ancient viral evolution. Virology 479-480, 26-37.

Aiewsakun, P., Katzourakis, A., 2016. Time-Dependent Rate Phenomenon in Viruses. J. Virol. 90, 7184-7195.

Al Rwahnih, M., Alabi, O.J., Westrick, N.M., Golino, D., Rowhani, A., 2017. Description of a Novel Monopartite Geminivirus and Its Defective Subviral Genome in Grapevine. Phytopathology 107, 240-251.

Alabi, O., Rayapati, N., Kumar, L., 2011. Cassava mosaic disease: A curse to food security in Sub-Saharan Africa. Online. APS*net* Features. doi:10.1094/APSnetFeature-2011-0701.

Alarcon, P., Rushton, J., Wieland, B., 2013. Cost of post-weaning multi-systemic wasting syndrome and porcine circovirus type-2 subclinical infection in England - an economic disease model. Prev. Vet. Med. 110, 88-102.

Allan, G., Krakowka, S., Ellis, J., Charreyre, C., 2012. Discovery and evolving history of two genetically related but phenotypically different viruses, porcine circoviruses 1 and 2. Virus Res. 164, 4-9.

Allan, G.M., Mc Neilly, F., Meehan, B.M., Kennedy, S., Mackie, D.P., Ellis, J.A., Clark, E.G., Espuna, E., Saubi, N., Riera, P., Bøtner, A., Charreyre, C.E., 1999. Isolation and characterisation of circoviruses from pigs with wasting syndromes in Spain, Denmark and Northern Ireland. Vet. Microbiol. 66, 115-123.

Allen, R.N., 1978. Spread of bunchy top disease in established banana plantations. Australian J. Ag. Res. 29, 1223-1233.

Anderson, P.K., Cunningham, A.A., Patel, N.G., Morales, F.J., Epstein, P.R., Daszak, P., 2004. Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. Trends Ecol. Evol. 19, 535-544.

Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C., Chan, A.M., Haynes, M., Kelley, S., Liu, H., Mahaffy, J.M., Mueller, J.E., Nulton, J., Olson, R., Parsons, R., Rayhawk, S., Suttle, C.A., Rohwer, F., 2006. The Marine Viromes of Four Oceanic Regions. PLOS Biol. 4, e368.

Arguello-Astorga, G., Ascencio-Ibáñez, J.T., Dallas, M.B., Orozco, B.M., Hanley-Bowdoin, L., 2007. High-Frequency Reversion of Geminivirus Replication Protein Mutants during Infection. J. Virol. 81, 11005-11015.

Ashby, E., 1921. Notes on Psephotus haematonotus, the Red-rumped Grass Parrakeet. The Avicultural Magazine 12, 131-133.

Ashby, M.K., Warry, A., Bejarano, E.R., Khashoggi, A., Burrell, M., Lichtenstein, C.P., 1997. Analysis of multiple copies of geminiviral DNA in the genome of four closely related Nicotiana species suggest a unique integration event. Plant Mol. Biol. 35, 313-321.

Babin, M., Ortíz, V., Castro, S., Romero, J., 2000. First Detection of Faba bean necrotic yellow virus in Spain. Plant Dis. 84, 707-707.

Bassami, M.R., Berryman, D., Wilcox, G.E., Raidal, S.R., 1998. Psittacine Beak and Feather Disease Virus Nucleotide Sequence Analysis and Its Relationship to Porcine Circovirus, Plant Circoviruses, and Chicken Anaemia Virus. Virology 249, 453-459.

Basso, M.F., da Silva, J.C.F., Fajardo, T.V.M., Fontes, E.P.B., Zerbini, F.M., 2015. A novel, highly divergent ssDNA virus identified in Brazil infecting apple, pear and grapevine. Virus Res. 210, 27-33.

Bejarano, E.R., Khashoggi, A., Witty, M., Lichtenstein, C., 1996. Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. Proc. Natl. Acad. Sci. U.S.A. 93, 759-764.

Belabess, Z., Dallot, S., El-Montaser, S., Granier, M., Majde, M., Tahiri, A., Blenzar, A., Urbino, C., Peterschmitt, M., 2015. Monitoring the dynamics of emergence of a non-canonical recombinant of Tomato yellow leaf curl virus and displacement of its parental viruses in tomato. Virology 486, 291-306.

Belfort, M., Perlman, P.S., 1995. Mechanisms of Intron Mobility. J. Biol. Chem. 270, 30237-30240.

Belyi, V.A., Levine, A.J., Skalka, A.M., 2010. Sequences from ancestral single-stranded DNA viruses in vertebrate genomes: the *Parvoviridae* and *Circoviridae* are more than 40 to 50 million years old. J. Virol. 84, 12458-12462.

Bernardo, P., Muhire, B., Francois, S., Deshoux, M., Hartnady, P., Farkas, K., Kraberger, S., Filloux, D., Fernandez, E., Galzi, S., Ferdinand, R., Granier, M., Marais, A., Monge Blasco, P., Candresse, T., Escriu, F., Varsani, A., Harkins, G.W., Martin, D.P., Roumagnac, P., 2016. Molecular characterization and prevalence of two capulaviruses: Alfalfa leaf curl virus from France and Euphorbia caput-medusae latent virus from South Africa. Virology 493, 142-153.

Bistolas, K.S.I., Besemer, R.M., Rudstam, L.G., Hewson, I., 2017a. Distribution and Inferred Evolutionary Characteristics of a Chimeric ssDNA Virus Associated with Intertidal Marine Isopods. Viruses 9, 361.

Bistolas, K.S.I., Jackson, E.W., Watkins, J.M., Rudstam, L.G., Hewson, I., 2017b. Distribution of circular single-stranded DNA viruses associated with benthic amphipods of genus Diporeia in the Laurentian Great Lakes. Freshwater Biol. 62, 1220-1231.

Bistolas, K.S.I., Rudstam, L.G., Hewson, I., 2017c. Gene expression of benthic amphipods (genus: Diporeia) in relation to a circular ssDNA virus across two Laurentian Great Lakes. PeerJ 5.

Blanc, S., Drucker, M., Uzest, M., 2014. Localizing viruses in their insect vectors. Annu. Rev. Phytopathol. 52, 403-425.

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

Borzak, R., Sellyei, B., Szekely, C., Doszpoly, A., 2017. Molecular Detection and Genome Analysis of Circoviruses of European Eel (*Anguilla anguilla*) and Sichel (Pelecus cultratus). Acta Vet. Hung. 65, 262-277.

Breitbart, M., Benner, B.E., Jernigan, P.E., Rosario, K., Birsa, L.M., Harbeitner, R.C., Fulford, S., Graham, C., Walters, A., Goldsmith, D.B., Berger, S.A., Nejstgaard, J.C., 2015. Discovery, Prevalence, and Persistence of Novel Circular Single-Stranded DNA Viruses in the Ctenophores Mnemiopsis leidyi and Beroe ovata. Front. Microbiol. 6, 1427.

Breitbart, M., Delwart, E., Rosario, K., Segales, J., Varsani, A., Consortium, I.R., 2017. ICTV Virus Taxonomy Profile: *Circoviridae*. J. Gen. Virol. 98, 1997-1998.

Breitbart, M., Rohwer, F., 2005. Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. BioTechniques 39, 729-736.

Briddon, R.W., Martin, D.P., Roumagnac, P., Navas-Castillo, J., Fiallo-Olivé, E., Moriones, E., Lett, J.-M., Zerbini, F.M., Varsani, A., 2018. *Alphasatellitidae*: a new family with two subfamilies for the classification of geminivirus- and nanovirus-associated alphasatellites. Arch. Virol. 163, 2587-2600.

Brown, J.K., Zerbini, F.M., Navas-Castillo, J., Moriones, E., Ramos-Sobrinho, R., Silva, J.C., Fiallo-Olive, E., Briddon, R.W., Hernandez-Zepeda, C., Idris, A., Malathi, V.G., Martin, D.P., Rivera-Bustamante, R., Ueda, S., Varsani, A., 2015. Revision of *Begomovirus* taxonomy based on pairwise sequence comparisons. Arch. Virol. 160, 1593-1619.

Campos-Olivas, R., Louis, J.M., Clerot, D., Gronenborn, B., Gronenborn, A.M., 2002. The structure of a replication initiator unites diverse aspects of nucleic acid metabolism. Proc. Natl. Acad. Sci. U.S.A. 99, 10310-10315.

Capella-Gutierrez, S., Silla-Martinez, J.M., Gabaldon, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972-1973.

Cardinale, D., DeRosa, K., Duffy, S., 2013. Base Composition and Translational Selection are Insufficient to Explain Codon Usage Bias in Plant Viruses. Viruses 5, 162.

Castrignano, S.B., Nagasse-Sugahara, T.K., Garrafa, P., Monezi, T.A., Barrella, K.M., Mehnert, D.U., 2017. Identification of circo-like virus-Brazil genomic sequences in raw sewage from the metropolitan area of Sao Paulo: evidence of circulation two and three years after the first detection. Mem. Inst. Oswaldo Cruz 112, 175-181.

Castrignano, S.B., Nagasse-Sugahara, T.K., Kisielius, J.J., Ueda-Ito, M., Brandao, P.E., Curti, S.P., 2013. Two novel circo-like viruses detected in human feces: complete genome sequencing and electron microscopy analysis. Virus Res. 178, 364-373.

Cheung, A.K., Ng, T.F., Lager, K.M., Alt, D.P., Delwart, E.L., Pogranichniy, R.M., 2014. Unique circovirus-like genome detected in pig feces. Genome Announc. 2, e00251-14.

Cheung, A.K., Ng, T.F., Lager, K.M., Bayles, D.O., Alt, D.P., Delwart, E.L., Pogranichniy, R.M., Kehrli, M.E., 2013. A divergent clade of circular single-stranded DNA viruses from pig feces. Arch. Virol. 158, 2157-2162.

Cheung, A.K., Ng, T.F.F., Lager, K.M., Alt, D.P., Delwart, E., Pogranichniy, R.M., 2015. Identification of several clades of novel single-stranded circular DNA viruses with conserved stem-loop structures in pig feces. Arch. Virol. 160, 353-358.

Choi, C., Chae, C., 1999. In-situ hybridization for the detection of porcine circovirus in pigs with postweaning multisystemic wasting syndrome. J. Comp. Pathol. 121, 265-270.

Claverie, S., Bernardo, P., Kraberger, S., Hartnady, P., Lefeuvre, P., Lett, J.M., Galzi, S., Filloux, D., Harkins, G.W., Varsani, A., Martin, D.P., Roumagnac, P., 2018. From Spatial Metagenomics to Molecular Characterization of Plant Viruses: A *Geminivirus* Case Study. Adv. Virus. Res. 101, 55-83.

Conceicao-Neto, N., Zeller, M., Heylen, E., Lefrere, H., Mesquita, J.R., Matthijnssens, J., 2015. Fecal virome analysis of three carnivores reveals a novel nodavirus and multiple gemycircularviruses. Virol. J. 12, 6.

Crawford, L.V., 1966. A minute virus of mice. Virology 29, 605-612.

Cuevas, J.M., Duffy, S., Sanjuán, R., 2009. Point Mutation Rate of Bacteriophage ΦX174. Genetics 183, 747-749.

Cuevas, J.M., Pereira-Gomez, M., Sanjuan, R., 2011. Mutation rate of bacteriophage PhiX174 modified through changes in GATC sequence context. Infect. Genet. Evol. 11, 1820-1822.

Cui, L.B., Wu, B.Y., Zhu, X.J., Guo, X.L., Ge, Y.Y., Zhao, K.C., Qi, X., Shi, Z.Y., Zhu, F.C., Sun, L.X., Zhou, M.H., 2017. Identification and genetic characterization of a novel circular single-stranded DNA virus in a human upper respiratory tract sample. Arch. Virol. 162, 3305-3312.

Czosnek, H., Hariton-Shalev, A., Sobol, I., Gorovits, R., Ghanim, M., 2017. The Incredible Journey of Begomoviruses in Their Whitefly Vector. Viruses 9, 273.

Dale, J.L., 1987. Banana Bunchy Top: An Economically Important Tropical Plant Virus Disease, in: Maramorosch, K., Murphy, F.A., Shatkin, A.J. (Eds.), Advances in Virus Research. Academic Press, pp. 301-325.

Dale, J.L., Harding, R.M., 1998. Banana bunchy top disease: current and future stratified for control., in: Hadidi, E.A., Khetarpal, R.K., Koganezawa, H. (Eds.), Plant Virus Disease Control. APS Press, St. Paul, MN, USA, pp. pp. 659-669.

Dayaram, A., Galatowitsch, M., Harding, J.S., Arguello-Astorga, G.R., Varsani, A., 2014. Novel circular DNA viruses identified in Procordulia grayi and Xanthocnemis zealandica larvae using metagenomic approaches. Infect. Genet. Evol. 22, 134-141.

Dayaram, A., Galatowitsch, M.L., Arguello-Astorga, G.R., van Bysterveldt, K., Kraberger, S., Stainton, D., Harding, J.S., Roumagnac, P., Martin, D.P., Lefeuvre, P., Varsani, A., 2016. Diverse circular replication-associated protein encoding viruses circulating in invertebrates within a lake ecosystem. Infect. Genet. Evol. 39, 304-316.

Dayaram, A., Goldstien, S., Arguello-Astorga, G.R., Zawar-Reza, P., Gomez, C., Harding, J.S., Varsani, A., 2015a. Diverse small circular DNA viruses circulating amongst estuarine molluscs. Infect. Genet. Evol. 31, 284-295.

Dayaram, A., Goldstien, S., Zawar-Reza, P., Gomez, C., Harding, J.S., Varsani, A., 2013a. Novel ssDNA virus recovered from estuarine Mollusc (Amphibola crenata) whose replication associated protein (Rep) shares similarities with Rep-like sequences of bacterial origin. J. Gen. Virol. 94, 1104-1110.

Dayaram, A., Opong, A., Jaschke, A., Hadfield, J., Baschiera, M., Dobson, R.C.J., Offei, S.K., Shepherd, D.N., Martin, D.P., Varsani, A., 2012. Molecular characterisation of a novel cassava associated circular ssDNA virus. Virus Res. 166, 130-135.

Dayaram, A., Potter, K.A., Moline, A.B., Rosenstein, D.D., Marinov, M., Thomas, J.E., Breitbart, M., Rosario, K., Arguello-Astorga, G.R., Varsani, A., 2013b. High global diversity of cycloviruses amongst dragonflies. J. Gen. Virol. 94, 1827-1840.

Dayaram, A., Potter, K.A., Pailes, R., Marinov, M., Rosenstein, D.D., Varsani, A., 2015b. Identification of diverse circular single-stranded DNA viruses in adult dragonflies and damselflies (Insecta: Odonata) of Arizona and Oklahoma, USA. Infect. Genet. Evol. 30, 278-287.

De Barro, P.J., Liu, S.S., Boykin, L.M., Dinsdale, A.B., 2011. Bemisia tabaci: a statement of species status. Annu. Rev. Entomol. 56, 1-19.

De Bruyn, A., Villemot, J., Lefeuvre, P., Villar, E., Hoareau, M., Harimalala, M., Abdoul-Karime, A.L., Abdou-Chakour, C., Reynaud, B., Harkins, G.W., Varsani, A., Martin, D.P., Lett, J.-M., 2012. East African cassava mosaic-like viruses from Africa to Indian ocean islands: molecular diversity, evolutionary history and geographical dissemination of a bipartite begomovirus. BMC Evol. Biol. 12, 228.

Dean, F.B., Nelson, J.R., Giesler, T.L., Lasken, R.S., 2001. Rapid Amplification of Plasmid and Phage DNA Using Phi29 DNA Polymerase and Multiply-Primed Rolling Circle Amplification. Genome Res. 11, 1095-1099.

Decaro, N., Martella, V., Desario, C., Lanave, G., Circella, E., Cavalli, A., Elia, G., Camero, M., Buonavoglia, C., 2014. Genomic characterization of a circovirus associated with fatal hemorrhagic enteritis in dog, Italy. PLoS One 9, e105909.

Dennis, T.P.W., de Souza, W.M., Marsile-Medun, S., Singer, J.B., Wilson, S.J., Gifford, R.J., 2018a. The evolution, distribution and diversity of endogenous circoviral elements in vertebrate genomes. Virus Res, in press.

Dennis, T.P.W., Flynn, P.J., Marciel de Souza, W., Singer, J.B., Moreau, C.S., Wilson, S.J., Gifford, R.J., 2018b. Insights into circovirus host range from the genomic fossil record. J. Virol. 92, e00145-18.

Diemer, G.S., Stedman, K.M., 2012. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. Biol. Direct 7, 13.

Doneley, R.J., 2003. Acute beak and feather disease in juvenile African Grey parrots--an uncommon presentation of a common disease. Aust. Vet. J. 81, 206-207.

Du, Z.G., Tang, Y.F., Zhang, S.B., She, X.M., Lan, G.B., Varsani, A., He, Z.F., 2014. Identification and molecular characterization of a single-stranded circular DNA virus with similarities to Sclerotinia sclerotiorum hypovirulence-associated DNA virus 1. Arch. Virol. 159, 1527-1531.

Duffy, S., Holmes, E.C., 2007. Multiple Introductions of the Old World Begomovirus Tomato yellow leaf curl virus into the New World. Appl. Environ. Microbiol. 73, 7114.

Duffy, S., Holmes, E.C., 2008. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. J. Virol. 82, 957-965.

Duffy, S., Holmes, E.C., 2009. Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. J. Gen. Virol. 90, 1539-1547.

Duffy, S., Shackelton, L.A., Holmes, E.C., 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. 9, 267.

Dunlap, D.S., Ng, T.F.F., Rosario, K., Barbosa, J.G., Greco, A.M., Breitbart, M., Hewson, I., 2013. Molecular and microscopic evidence of viruses in marine copepods. Proc. Natl. Acad. Sci. U.S.A. 110, 1375-1380.

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

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res. 32.

Fahsbender, E., Hewson, I., Rosario, K., Tuttle, A.D., Varsani, A., Breitbart, M., 2015. Discovery of a novel circular DNA virus in the Forbes sea star, Asterias forbesi. Arch. Virol. 160, 2349-2351.

Faurez, F., Dory, D., Grasland, B., Jestin, A., 2009. Replication of porcine circoviruses. Virol J. 6, 60.

Feschotte, C., Gilbert, C., 2012. Endogenous viruses: insights into viral evolution and impact on host biology. Nat. Rev. Genet. 13, 283-296.

Fiallo-Olivé, E., Martínez-Zubiaur, Y., Moriones, E., Navas-Castillo, J., 2012. A novel class of DNA satellites associated with New World begomoviruses. Virology 426, 1-6.

Fiallo-Olivé, E., Tovar, R., Navas-Castillo, J., 2016. Deciphering the biology of deltasatellites from the New World: maintenance by New World begomoviruses and whitefly transmission. New Phytol. 212, 680-692.

Filloux, D., Murrell, S., Koohapitagtam, M., Golden, M., Julian, C., Galzi, S., Uzest, M., Rodier-Goud, M., D'Hont, A., Vernerey, M.S., Wilkin, P., Peterschmitt, M., Winter, S., Murrell, B., Martin, D.P., Roumagnac, P., 2015. The genomes of many yam species contain transcriptionally active endogenous geminiviral sequences that may be functionally expressed. Virus Evol. 1, vev002.

Firth, C., Charleston, M.A., Duffy, S., Shapiro, B., Holmes, E.C., 2009. Insights into the Evolutionary History of an Emerging Livestock Pathogen: Porcine Circovirus 2. J. Virol. 83, 12813.

Fogell, D.J., Martin, R.O., Groombridge, J.J., 2016. Beak and feather disease virus in wild and captive parrots: an analysis of geographic and taxonomic distribution and methodological trends. Arch. Virol. 161, 2059-2074.

Fondong, V.N., 2013. Geminivirus protein structure and function. Mol Plant Pathol 14, 635-649.

Fontenele, R.S., Abreu, R.A., Lamas, N.S., Alves-Freitas, D.M.T., Vidal, A.H., Poppiel, R.R., Melo, F.L., Lacorte, C., Martin, D.P., Campos, M.A., Varsani, A., Ribeiro, S.G., 2018. Passion Fruit Chlorotic Mottle Virus: Molecular Characterization of a New Divergent Geminivirus in Brazil. Viruses 10, e169.

Franz, A., Makkouk, K.M., Vetten, H.J., 1998. Acquisition, Retention and Transmission of Faba Bean Necrotic Yellows Virus by Two of its Aphid Vectors, Aphis craccivora (Koch) and Acyrthosiphon pisum (Harris). J. Phytopathol. 146, 347-355.

Franzo, G., Segales, J., Tucciarone, C.M., Cecchinato, M., Drigo, M., 2018. The analysis of genome composition and codon bias reveals distinctive patterns between avian and mammalian circoviruses which suggest a potential recombinant origin for Porcine circovirus 3. PLoS One 13, e0199950.

Gaafar, Y., Cordsen Nielsen, G., and Ziebell, H. 2018. Molecular characterisation of the first occurrence of Pea necrotic yellow dwarf virus in Denmark. New Disease Reports 37, 16.

Gaafar, Y., Timchenko, T., Ziebell, H., 2017. First report of Pea necrotic yellow dwarf virus in The Netherlands. New Dis. Rep. 35, 23.

Gallet, R., Fabre, F., Thébaud, G., Sofonea, M.T., Sicard, A., Blanc, S., Michalakis, Y., 2018. Small bottleneck size in a highly multipartite virus during a complete infectious cycle. J. Virol. 92, e00139-18.

Garigliany, M.M., Hagen, R.M., Frickmann, H., May, J., Schwarz, N.G., Perse, A., Jost, H., Borstler, J., Shahhosseini, N., Desmecht, D., Mbunkah, H.A., Daniel, A.M., Kingsley, M.T., Campos Rde, M., de Paula, V.S., Randriamampionona, N., Poppert, S., Tannich, E., Rakotozandrindrainy, R., Cadar, D., Schmidt-Chanasit, J., 2014. Cyclovirus CyCV-VN species distribution is not limited to Vietnam and extends to Africa. Sci. Rep. 4, 7552.

Ge, L., Zhang, J., Zhou, X., Li, H., 2007. Genetic Structure and Population Variability of Tomato Yellow Leaf Curl China Virus. J. Virol. 81, 5902-5907.

Ge, X.Y., Li, J.L., Peng, C., Wu, L.J., Yang, X.L., Wu, Y.Q., Zhang, Y.Z., Shi, Z.L., 2011. Genetic diversity of novel circular ssDNA viruses in bats in China. J. Gen. Virol. 92, 2646-2653.

Gerlach, H., 1994. *Circoviridae*-psittacine beak and feather disease virus, in: Ritchie, B.W., Harrison, G.J., Harrison, L.R. (Eds.), Avian medicine: principles and application. Wingers Publishing Incorporation, Lake Worth, FL, pp. 894-903.

Gibbs, M.J., Smeianov, V.V., Steele, J.L., Upcroft, P., Efimov, B.A., 2006. Two families of rep-like genes that probably originated by interspecies recombination are represented in viral, plasmid, bacterial, and parasitic protozoan genomes. Mol. Biol. Evol. 23, 1097-1100.

Gibbs, M.J., Weiller, G.F., 1999. Evidence that a plant virus switched hosts to infect a vertebrate and then recombined with a vertebrate-infecting virus. Proc. Natl. Acad. Sci. U.S.A. 96, 8022-8027.

Gillespie, J., Opriessnig, T., Meng, X.J., Pelzer, K., Buechner-Maxwell, V., 2009. Porcine Circovirus Type 2 and Porcine Circovirus-Associated Disease. J. Vet. Intern. Med. 23, 1151-1163.

Goodman, R.M., 1977a. Infectious DNA from a whitefly-transmitted virus of Phaseolus vulgaris. Nature 266, 54.

Goodman, R.M., 1977b. Single-stranded DNA genome in a whitefly-transmitted plant virus. Virology 83, 171-179.

Gorbalenya, A.E., Koonin, E.V., 1993. Helicases: amino acid sequence comparisons and structure-function relationships. Curr. Opin. Struct. Biol. 3, 419-429.

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

Grigoras, I., Ginzo, A.I.d.C., Martin, D.P., Varsani, A., Romero, J., Mammadov, A.C., Huseynova, I.M., Aliyev, J.A., Kheyr-Pour, A., Huss, H., Ziebell, H., Timchenko, T., Vetten, H.-J., Gronenborn, B., 2014. Genome diversity and evidence of recombination and reassortment in nanoviruses from Europe. J. Gen. Virol. 95, 1178-1191.

Grigoras, I., Gronenborn, B., Vetten, H.J., 2010a. First Report of a Nanovirus Disease of Pea in Germany. Plant Dis. 94, 642-642.

Grigoras, I., Timchenko, T., Grande-Perez, A., Katul, L., Vetten, H.J., Gronenborn, B., 2010b. High Variability and Rapid Evolution of a Nanovirus. J. Virol. 84, 9105-9117.

Guindon, S., Gascuel, O., 2003. A Simple, Fast and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Syst. Biol. 52.

Haible, D., Kober, S., Jeske, H., 2006. Rolling circle amplification revolutionizes diagnosis and genomics of geminiviruses. J. Virol. Meth. 135, 9-16.

Hamel, A.L., Lin, L.L., Nayar, G.P.S., 1998. Nucleotide Sequence of Porcine Circovirus Associated with Postweaning Multisystemic Wasting Syndrome in Pigs. J. Virol. 72, 5262-5267.

Hanley-Bowdoin, L., Bejarano, E.R., Robertson, D., Mansoor, S., 2013. Geminiviruses: masters at redirecting and reprogramming plant processes. Nat. Rev. Microbiol. 11, 777-788.

Hanna, Z.R., Runckel, C., Fuchs, J., DeRisi, J.L., Mindell, D.P., Van Hemert, C., Handel, C.M., Dumbacher, J.P., 2015. Isolation of a Complete Circular Virus Genome Sequence from an Alaskan Black-Capped Chickadee (*Poecile atricapillus*) Gastrointestinal Tract Sample. Genome Announc 3, e01081-15.

Harding, J.C.S., Clark, E.G., 1997. Recognizing and diagnosing postweaning multisystemic wasting syndrome (PMWS). Swine Health Prod. 5, 201-203.

Harkins, G.W., Delport, W., Duffy, S., Wood, N., Monjane, A.L., Owor, B.E., Donaldson, L., Saumtally, S., Triton, G., Briddon, R.W., Shepherd, D.N., Rybicki, E.P., Martin, D.P., Varsani, A., 2009a. Experimental evidence indicating that mastreviruses probably did not co-diverge with their hosts. Virol. J. 6, 104.

Harkins, G.W., Martin, D.P., Christoffels, A., Varsani, A., 2014. Towards inferring the global movement of beak and feather disease virus. Virology 450-451, 24-33.

Harkins, G.W., Martin, D.P., Duffy, S., Monjane, A.L., Shepherd, D.N., Windram, O.P., Owor, B.E., Donaldson, L., van Antwerpen, T., Sayed, R.A., Flett, B., Ramusi, M., Rybicki, E.P., Peterschmitt, M., Varsani, A., 2009b. Dating the origins of the maize-adapted strain of maize streak virus, MSV-A. J. Gen. Virol. 90, 3066-3074.

Harrison, B.D., Barker, H., Bock, K.R., Guthrie, E.J., Meredith, G., Atkinson, M., 1977. Plant viruses with circular single-stranded DNA. Nature 270, 760.

Hayward, A., Katzourakis, A., 2015. Endogenous retroviruses. Curr Biol 25, R644-646.

Hefferon, K.L., Moon, Y.S., Fan, Y., 2006. Multi-tasking of nonstructural gene products is required for bean yellow dwarf geminivirus transcriptional regulation. FEBS J. 273, 4482-4494.

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

Heydarnejad, J., Kamali, M., Massumi, H., Kvarnheden, A., Male, M.F., Kraberger, S., Stainton, D., Martin, D.P., Varsani, A., 2017. Identification of a Nanovirus-Alphasatellite Complex in Sophora alopecuroides. Virus Res. 235, 24-32.

Ho, E.S., Kuchie, J., Duffy, S., 2014. Bioinformatic analysis reveals genome size reduction and the emergence of tyrosine phosphorylation site in the movement protein of New World bipartite begomoviruses. PLoS One 9, e111957.

Hooks, C.R.R., Wright, M.G., Kabasawa, D.S., Manandhar, R., Almeida, R.P.P., 2008. Effect of banana bunchy top virus infection on morphology and growth characteristics of banana. Ann. Appl. Biol. 153, 1-9.

Hyder, M.Z., Shah, S.H., Hameed, S., Naqvi, S.M., 2011. Evidence of recombination in the Banana bunchy top virus genome. Infect. Genet. Evol. 11, 1293-1300.

Idris, A.M., Shahid, M.S., Briddon, R.W., Khan, A.J., Zhu, J.-K., Brown, J.K., 2011. An unusual alphasatellite associated with monopartite begomoviruses attenuates symptoms and reduces betasatellite accumulation. J. Gen. Virol. 92, 706-717.

Ilyina, T.V., Koonin, E.V., 1992. Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. Nucleic Acids Res. 20, 3279-3285.

Inoue-Nagata, A.K., Albuquerque, L.C., Rocha, W.B., Nagata, T., 2004. A simple method for cloning the complete begomovirus genome using the bacteriophage φ29 DNA polymerase. J. Virol. Meth. 116, 209-211.

Inoue-Nagata, A.K., Lima, M.F., Gilbertson, R.L., 2016. A review of geminivirus diseases in vegetables and other crops in Brazil: current status and approaches for management. Hort. Bras. 34, 8-18.

Jackson, E.W., Bistolas, K.S.I., Button, J.B., Hewson, I., 2016. Novel Circular Single-Stranded DNA Viruses among an Asteroid, Echinoid and Holothurian (Phylum: Echinodermata). PLoS One 11, 9.

Jeske, H., Lütgemeier, M., Preiß, W., 2001. DNA forms indicate rolling circle and recombination-dependent replication of Abutilon mosaic virus. EMBO J. 20, 6158-6167.

Jones, D.R., 2003. Plant Viruses Transmitted by Whiteflies. Eur. J. Plant Pathol. 109, 195-219.

Kapoor, A., Dubovi, E.J., Henriquez-Rivera, J.A., Lipkin, W.I., 2012. Complete genome sequence of the first canine circovirus. J. Virol. 86, 7018.

Kaszab, E., Marton, S., Forro, B., Bali, K., Lengyel, G., Banyai, K., Feher, E., 2018. Characterization of the genomic sequence of a novel CRESS DNA virus identified in Eurasian jay (Garrulus glandarius). Arch. Virol. 163, 285-289.

Katzourakis, A., Gifford, R.J., 2010. Endogenous viral elements in animal genomes. PLoS Genet. 6, e1001191.

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

Kazlauskas, D., Varsani, A., Krupovic, M., 2018. Pervasive Chimerism in the Replication-Associated Proteins of Uncultured Single-Stranded DNA Viruses. Viruses 10.

Kemenesi, G., Kurucz, K., Zana, B., Foldes, F., Urban, P., Vlaschenko, A., Kravchenko, K., Budinski, I., Szodoray-Paradi, F., Bucs, S., Jere, C., Csosz, I., Szodoray-Paradi, A., Estok, P., Gorfol, T., Boldogh, S., Jakab, F., 2018. Diverse replication-associated protein encoding circular DNA viruses in guano samples of Central-Eastern European bats. Arch. Virol. 163, 671-678.

Kim, A.R., Chung, H.C., Kim, H.K., Kim, E.O., Nguyen, V.G., Choi, M.G., Yang, H.J., Kim, J.A., Park, B.K., 2014. Characterization of a complete genome of a circular single-stranded DNA virus from porcine stools in Korea. Virus Genes 48, 81-88.

Kim, H.K., Park, S.J., Nguyen, V.G., Song, D.S., Moon, H.J., Kang, B.K., Park, B.K., 2012. Identification of a novel single-stranded, circular DNA virus from bovine stool. J. Gen. Virol. 93, 635-639.

Kim, J., Kwak, H.-R., Kim, M., Seo, J.-K., Yang, J.W., Chung, M.-N., Kil, E.-J., Choi, H.-S., Lee, S., 2018. Phylogeographic analysis of the full genome of Sweepovirus to trace virus dispersal and introduction to Korea. PLoS One 13, e0202174.

Kim, K.-H., Bae, J.-W., 2011. Amplification Methods Bias Metagenomic Libraries of Uncultured Single-Stranded and Double-Stranded DNA Viruses. Appl. Environ. Microbiol. 77, 7663.

Kimura, K., Tomaru, Y., 2013. Isolation and characterization of a single-stranded DNA virus infecting the marine diatom Chaetoceros sp. strain SS628-11 isolated from western Japan. PLoS One 8, e82013.

Kimura, K., Tomaru, Y., 2015. Discovery of two novel viruses expands the diversity of single-stranded DNA and single-stranded RNA viruses infecting a cosmopolitan marine diatom. Appl. Environ. Microbiol. 81, 1120-1131.

King, A.M., Adams, M.J., Carstens, E.B., Lefkowitz, E.J., 2012. Virus taxonomy: classification and nomenclature of viruses: Ninth Report of the International Committee on Taxonomy of Viruses. Elsevier, San Diego.

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

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

Koonin, E.V., Ilyina, T.V., 1993. Computer-assisted dissection of rolling circle DNA replication. BioSystems 30, 241-268.

Kraberger, S., Arguello-Astorga, G.R., Greenfield, L.G., Galilee, C., Law, D., Martin, D.P., Varsani, A., 2015a. Characterisation of a diverse range of circular replication-associated protein encoding DNA viruses recovered from a sewage treatment oxidation pond. Infect. Genet. Evol. 31, 73-86.

Kraberger, S., Farkas, K., Bernardo, P., Booker, C., Arguello-Astorga, G.R., Mesleard, F., Martin, D.P., Roumagnac, P., Varsani, A., 2015b. Identification of novel Bromus- and Trifolium-associated circular DNA viruses. Arch. Virol. 160, 1303-1311.

Kraberger, S., Geering, A.D.W., Walters, M., Martin, D.P., Varsani, A., 2017. Novel mastreviruses identified in Australian wild rice. Virus Res. 238, 193-197.

Kraberger, S., Kumari, S.G., Najar, A., Stainton, D., Martin, D.P., Varsani, A., 2018. Molecular characterization of faba bean necrotic yellows viruses in Tunisia. Arch. Virol. 163, 687-694.

Kraberger, S., Stainton, D., Dayaram, A., Zawar-Reza, P., Gomez, C., Harding, J.S., Varsani, A., 2013. Discovery of Sclerotinia sclerotiorum Hypovirulence-Associated Virus-1 in Urban River Sediments of Heathcote and Styx Rivers in Christchurch City, New Zealand. Genome Announc. 1, e00559-13.

Krenz, B., Thompson, J.R., Fuchs, M., Perry, K.L., 2012. Complete genome sequence of a new circular DNA virus from grapevine. J Virol. 86, 7715.

Krishnamurthy, S.R., Wang, D., 2017. Origins and challenges of viral dark matter. Virus Res. 239, 136-142.

Krupovic, M., 2013. Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. Curr. Opin. Virol. 3, 578-586.

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

Krupovic, M., Ravantti, J.J., Bamford, D.H., 2009. Geminiviruses: a tale of a plasmid becoming a virus. BMC Evol. Biol. 9, 112.

Krupovic, M., Zhi, N., Li, J., Hu, G., Koonin, E.V., Wong, S., Shevchenko, S., Zhao, K., Young, N.S., 2015. Multiple layers of chimerism in a single-stranded DNA virus discovered by deep sequencing. Genome Biol. Evol. 7, 993-1001.

Kryukov, K., Ueda, M.T., Imanishi, T., Nakagawa, S., 2018. Systematic survey of non-retroviral virus-like elements in eukaryotic genomes. Virus Res.

Kumar, J., Kumar, J., Singh, S.P., Tuli, R., 2014. betaC1 is a pathogenicity determinant: not only for begomoviruses but also for a mastrevirus. Arch. Virol. 159, 3071-3076.

Kumari, S.G., Attar, N., Mustafayev, E., Akparov, Z., 2009. First Report of Faba bean necrotic yellows virus Affecting Legume Crops in Azerbaijan. Plant Dis. 93, 1220-1220.

Kundu, S., Faulkes, C.G., Greenwood, A.G., Jones, C.G., Kaiser, P., Lyne, O.D., Black, S.A., Chowrimootoo, A., Groombridge, J.J., 2012. Tracking Viral Evolution during a Disease Outbreak: the Rapid and Complete Selective Sweep of a Circovirus in the Endangered Echo Parakeet. J. Virol. 86, 5221-5229.

Lamberto, I., Gunst, K., Muller, H., Zur Hausen, H., de Villiers, E.M., 2014. Mycovirus-like DNA virus sequences from cattle serum and human brain and serum samples from multiple sclerosis patients. Genome Announc. 2, e00848-14.

Lapidot, M., Gelbart, D., Gal-On, A., Sela, N., Anfoka, G., Haj Ahmed, F., Abou-Jawada, Y., Sobh, H., Mazyad, H., Aboul-Ata, A.-A.E., Kamal El-Attar, A., Ali-Shtayeh, M.S., Jamous, R.M., Polston, J.E., Duffy, S., 2014. Frequent migration of introduced cucurbit-infecting begomoviruses among Middle Eastern countries. Virol. J. 11, 181.

Lasken, R.S., Egholm, M., 2003. Whole genome amplification: abundant supplies of DNA from precious samples or clinical specimens. Trends Biotechnol. 21, 531-535.

Laufs, J., Traut, W., Heyraud, F., Matzeit, V., Rogers, S.G., Schell, J., Gronenborn, B., 1995. In vitro cleavage and joining at the viral origin of replication by the replication initiator protein of tomato yellow leaf curl virus. Proc. Natl. Acad. Sci. U.S.A. 92, 3879-3883.

Le, V.T., de Jong, M.D., Nguyen, V.K., Nguyen, V.T., Taylor, W., Wertheim, H.F., van der Ende, A., van der Hoek, L., Canuti, M., Crusat, M., Sona, S., Nguyen, H.U., Giri, A., Nguyen, T.T., Ho, D.T., Farrar, J., Bryant, J.E., Tran, T.H., Nguyen, V.V., van Doorn, H.R., 2014. Limited geographic distribution of the novel cyclovirus CyCV-VN. Sci. Rep. 4, 3967.

Lefeuvre, P., Harkins, G.W., Lett, J.-M., Briddon, R.W., Chase, M.W., Moury, B., Martin, D.P., 2011. Evolutionary Time-Scale of the Begomoviruses: Evidence from Integrated Sequences in the Nicotiana Genome. PLoS One 6, e19193.

Lefeuvre, P., Lett, J.-M., Reynaud, B., Martin, D.P., 2007. Avoidance of Protein Fold Disruption in Natural Virus Recombinants. PLoS Pathog. 3, e181.

Lefeuvre, P., Lett, J.M., Varsani, A., Martin, D.P., 2009. Widely conserved recombination patterns among single-stranded DNA viruses. J. Virol. 83, 2697-2707.

Lefeuvre, P., Martin, D.P., Harkins, G., Lemey, P., Gray, A.J., Meredith, S., Lakay, F., Monjane, A., Lett, J.M., Varsani, A., Heydarnejad, J., 2010. The spread of tomato yellow leaf curl virus from the Middle East to the world. PLoS Pathog. 6, e1001164.

Lefeuvre, P., Moriones, E., 2015. Recombination as a motor of host switches and virus emergence: geminiviruses as case studies. Curr. Opin. Virol. 10, 14-19.

Legg, J.P., Lava Kumar, P., Makeshkumar, T., Tripathi, L., Ferguson, M., Kanju, E., Ntawuruhunga, P., Cuellar, W., 2015. Cassava virus diseases: biology, epidemiology, and management. Adv Virus Res. 91, 85-142.

Li, L., Giannitti, F., Low, J., Keyes, C., Ullmann, L.S., Deng, X., Aleman, M., Pesavento, P.A., Pusterla, N., Delwart, E., 2015a. Exploring the virome of diseased horses. J. Gen. Virol. 96, 2721-2733.

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

Li, L., Shan, T., Soji, O.B., Alam, M.M., Kunz, T.H., Zaidi, S.Z., Delwart, E., 2011. Possible cross-species transmission of circoviruses and cycloviruses among farm animals. J. Gen. Virol. 92, 768-772.

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

Li, W., Gu, Y., Shen, Q., Yang, S., Wang, X., Wan, Y., Zhang, W., 2015b. A novel gemycircularvirus from experimental rats. Virus Genes 51, 302-305.

Li, Y., Khalafalla, A.I., Paden, C.R., Yusof, M.F., Eltahir, Y.M., Al Hammadi, Z.M., Tao, Y., Queen, K., Hosani, F.A., Gerber, S.I., Hall, A.J., Al Muhairi, S., Tong, S., 2017. Identification of diverse viruses in upper respiratory samples in dromedary camels from United Arab Emirates. PLoS One 12, e0184718.

Lian, H., Liu, Y., Li, N., Wang, Y., Zhang, S., Hu, R., 2014. Novel circovirus from mink, China. Emerg. Infect. Dis. 20, 1548-1550.

Lima, A.T.M., Silva, J.C.F., Silva, F.N., Castillo-Urquiza, G.P., Silva, F.F., Seah, Y.M., Mizubuti, E.S.G., Duffy, S., Zerbini, F.M., 2017a. The diversification of begomovirus populations is predominantly driven by mutational dynamics. Virus Evol. 3, vex005.

Lima, D.A., Cibulski, S.P., Finkler, F., Teixeira, T.F., Varela, A.P.M., Cerva, C., Loiko, M.R., Scheffer, C.M., dos Santos, H.F., Mayer, F.Q., Roehe, P.M., 2017b. Faecal virome of healthy chickens reveals a large diversity of the eukaryote viral community, including novel circular ssDNA viruses. J. Gen. Virol. 98, 690-703.

Lima, F.E.D., Cibulski, S.P., dos Santos, H.F., Teixeira, T.F., Varela, A.P.M., Roehe, P.M., Delwart, E., Franco, A.C., 2015. Genomic Characterization of Novel Circular ssDNA Viruses from Insectivorous Bats in Southern Brazil. PLoS One 10, 11.

Liu, H.Q., Fu, Y.P., Li, B., Yu, X., Xie, J.T., Cheng, J.S., Ghabrial, S.A., Li, G.Q., Yi, X.H., Jiang, D.H., 2011. Widespread Horizontal Gene Transfer from Circular Single-stranded DNA Viruses to Eukaryotic Genomes. BMC Evol. Biol. 11, 15.

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

Lozano, G., Trenado, H.P., Fiallo-Olive, E., Chirinos, D., Geraud-Pouey, F., Briddon, R.W., Navas-Castillo, J., 2016. Characterization of Non-coding DNA Satellites Associated with Sweepoviruses (Genus *Begomovirus*, *Geminiviridae*) - Definition of a Distinct Class of Begomovirus-Associated Satellites. Front. Microbiol. 7, 162.

Lu, Q.Y., Wu, Z.J., Xia, Z.S., Xie, L.H., 2015. Complete genome sequence of a novel monopartite geminivirus identified in mulberry (Morus alba L.). Arch. Virol. 160, 2135-2138.

Ma, Y., Navarro, B., Zhang, Z., Lu, M., Zhou, X., Chi, S., Di Serio, F., Li, S., 2015. Identification and molecular characterization of a novel monopartite geminivirus associated with mulberry mosaic dwarf disease. J. Gen. Virol. 96, 2421-2434.

Mabvakure, B., Martin, D.P., Kraberger, S., Cloete, L., van Brunschot, S., Geering, A.D.W., Thomas, J.E., Bananej, K., Lett, J.M., Lefeuvre, P., Varsani, A., Harkins, G.W., 2016. Ongoing geographical spread of Tomato yellow leaf curl virus. Virology 498, 257-264.

Macera, L., Focosi, D., Vatteroni, M.L., Manzin, A., Antonelli, G., Pistello, M., Maggi, F., 2016. Cyclovirus Vietnam DNA in immunodeficient patients. J. Clin. Virol. 81, 12-15.

Magar, R., Larochelle, R., Thibault, S., Lamontagne, L., 2000. Experimental transmission of porcine circovirus type 2 (PCV2) in weaned pigs: a sequential study. J. Comp. Pathol. 123, 258-269.

Magee, C.J., 1927. Investigation on the bunchy top disease of the banana. Bull. Counc. Sci. Ind. Res. Aust. 30, 1-64.

Makkouk, K.M., Kumari, S.G., 2009. Epidemiology and integrated management of persistently transmitted aphid-borne viruses of legume and cereal crops in West Asia and North Africa. Virus Res. 141, 209-218.

Maldonado, J., Segales, J., Martinez-Puig, D., Calsamiglia, M., Riera, P., Domingo, M., Artigas, C., 2005. Identification of viral pathogens in aborted fetuses and stillborn piglets from cases of swine reproductive failure in Spain. Vet. J. 169, 454-456.

Male, M.F., Kami, V., Kraberger, S., Varsani, A., 2015. Genome Sequences of Poaceae-Associated Gemycircularviruses from the Pacific Ocean Island of Tonga. Genome Announc. 3.

Male, M.F., Kraberger, S., Stainton, D., Kami, V., Varsani, A., 2016. Cycloviruses, gemycircularviruses and other novel replication-associated protein encoding circular viruses in Pacific flying fox (Pteropus tonganus) faeces. Infect. Genet. Evol. 39, 279-292.

Martin, D.P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P., Varsani, A., 2011. Recombination in Eukaryotic Single Stranded DNA Viruses. Viruses 3, 1699-1738.

Marzano, S.L., Domier, L.L., 2016. Novel mycoviruses discovered from metatranscriptomics survey of soybean phyllosphere phytobiomes. Virus Res. 213, 332-342.

McDaniel, L.D., Rosario, K., Breitbart, M., Paul, J.H., 2014. Comparative metagenomics: natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. Environ. Microbiol. 16, 570-585.

Medina, C.G.V., Teppa, E., Bornancini, V.A., Flores, C.R., Marino-Buslje, C., Lambertini, P.M.L., 2018. Tomato Apical Leaf Curl Virus: A Novel, Monopartite Geminivirus Detected in Tomatoes in Argentina. Front. Microbiol. 8.

Meng, X.J., 2013. Porcine circovirus type 2 (PCV2): pathogenesis and interaction with the immune system. Annu. Rev. Anim. Biosci. 1, 43-64.

Mi, S., Lee, X., Li, X.-p., Veldman, G.M., Finnerty, H., Racie, L., LaVallie, E., Tang, X.-Y., Edouard, P., Howes, S., Keith Jr, J.C., McCoy, J.M., 2000. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403, 785.

Mu, F., Xie, J., Cheng, S., You, M.P., Barbetti, M.J., Jia, J., Wang, Q., Cheng, J., Fu, Y., Chen, T., Jiang, D., 2017. Virome Characterization of a Collection of S. sclerotiorum from Australia. Front. Microbiol. 8, 2540.

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

Muñoz-Martín, A., Collin, S., Herreros, E., Mullineaux, P.M., Fernández-Lobato, M., Fenoll, C., 2003. Regulation of MSV and WDV virion-sense promoters by WDV nonstructural proteins: a role for their retinoblastoma protein-binding motifs. Virology 306, 313-323.

Murphy, F.A., Fauquet, C.M., Bishop, D.H.L., Ghabrial, S.A., Jarvis, A.W., Martelli, G., P. Mayo, M.A., Summers, M.D., 1995. Virus Taxonomy. Sixth report of the International Committee on Taxonomy of Viruses. Arch. Virol. Suppl. 10, 590.

Naccache, S.N., Greninger, A.L., Lee, D., Coffey, L.L., Phan, T., Rein-Weston, A., Aronsohn, A., Hackett, J., Delwart, E.L., Chiu, C.Y., 2013. The Perils of Pathogen Discovery: Origin of a Novel Parvovirus-Like Hybrid Genome Traced to Nucleic Acid Extraction Spin Columns. J. Virol. 87, 11966.

Nagasaki, K., Tomaru, Y., Takao, Y., Nishida, K., Shirai, Y., Suzuki, H., Nagumo, T., 2005. Previously unknown virus infects marine diatom. Appl. Environ. Microbiol. 71, 3528-3535.

Nagy, G.N., Suardiaz, R., Lopata, A., Ozohanics, O., Vekey, K., Brooks, B.R., Leveles, I., Toth, J., Vertessy, B.G., Rosta, E., 2016. Structural Characterization of Arginine Fingers: Identification of an Arginine Finger for the Pyrophosphatase dUTPases. J. Am. Chem. Soc. 138, 15035-15045.

Nakasu, E.Y.T., Melo, F.L., Michereff, M., Nagata, T., Ribeiro, B.M., Ribeiro, S.G., Lacorte, C., Inoue-Nagata, A.K., 2017. Discovery of two small circular ssDNA viruses associated with the whitefly Bemisia tabaci. Arch. Virol. 162, 2835-2838.

Nawaz-ul-Rehman, M.S., Fauquet, C.M., 2009. Evolution of geminiviruses and their satellites. FEBS Lett. 583, 1825-1832.

Nawaz-ul-Rehman, M.S., Nahid, N., Mansoor, S., Briddon, R.W., Fauquet, C.M., 2010. Post-transcriptional gene silencing suppressor activity of two non-pathogenic alphasatellites associated with a begomovirus. Virology 405, 300-308.

Nayar, G.P., Hamel, A., Lin, L., 1997. Detection and characterization of porcine circovirus associated with postweaning multisystemic wasting syndrome in pigs. Can. Vet. J. 38, 385-386.

Ng, T.F., Chen, L.F., Zhou, Y., Shapiro, B., Stiller, M., Heintzman, P.D., Varsani, A., Kondov, N.O., Wong, W., Deng, X., Andrews, T.D., Moorman, B.J., Meulendyk, T., MacKay, G., Gilbertson, R.L., Delwart, E., 2014. Preservation of viral genomes in 700-y-old caribou feces from a subarctic ice patch. Proc. Natl. Acad. Sci. USA 111, 16842-16847.

Ng, T.F., Duffy, S., Polston, J.E., Bixby, E., Vallad, G.E., Breitbart, M., 2011a. Exploring the diversity of plant DNA viruses and their satellites using vector-enabled metagenomics on whiteflies. PLoS One 6, e19050.

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

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

Ng, T.F.F., Zhang, W., Sachsenröder, J., Kondov, N.O., da Costa, A.C., Vega, E., Holtz, L.R., Wu, G., Wang, D., Stine, C.O., Antonio, M., Mulvaney, U.S., Muench, M.O., Deng, X., Ambert-Balay, K., Pothier, P., Vinjé, J., Delwart, E., 2015. A diverse group of small circular ssDNA viral genomes in human and non-human primate stools. Virus Evol. 1, vev017.

Nguyen, T.T., Robertsen, E.M., Landfald, B., 2017. Viral assemblage variation in an Arctic shelf seafloor. Aquat. Microb. Ecol. 78, 135-145.

Oba, M., Katayama, Y., Naoi, Y., Tsuchiaka, S., Omatsu, T., Okumura, A., Nagai, M., Mizutani, T., 2017. Discovery of fur seal feces-associated circular DNA virus in swine feces in Japan. J. Vet. Med. Sci. 79, 1664-1666.

Opriessnig, T., Meng, X.-J., Halbur, P.G., 2007. Porcine Circovirus Type 2–Associated Disease: Update on Current Terminology, Clinical Manifestations, Pathogenesis, Diagnosis, and Intervention Strategies. J. Vet. Diagn. Invest. 19, 591-615.

Opriessnig, T., Xiao, C.-T., Halbur, P.G., Gerber, P.F., Matzinger, S.R., Meng, X.-J., 2017. A commercial porcine circovirus (PCV) type 2a-based vaccine reduces PCV2d viremia and shedding and prevents PCV2d transmission to naïve pigs under experimental conditions. Vaccine 35, 248-254.

Padilla-Rodriguez, M., Rosario, K., Breitbart, M., 2013. Novel cyclovirus discovered in the Florida woods cockroach Eurycotis floridana (Walker). Arch. Virol. 158, 1389-1392.

Pass, D.A., Perry, R.A., 1984. The pathology of psittacine beak and feather disease. Aust. Vet. J. 61, 69-74.

Pearson, V.M., Caudle, S.B., Rokyta, D.R., 2016. Viral recombination blurs taxonomic lines: examination of single-stranded DNA viruses in a wastewater treatment plant. PeerJ 4, 18.

Phan, T.G., da Costa, A.C., Mendoza, J.D., Bucardo-Rivera, F., Nordgren, J., O'Ryan, M., Deng, X.T., Delwart, E., 2016. The fecal virome of South and Central American children with diarrhea includes small circular DNA viral genomes of unknown origin. Arch. Virol. 161, 959-966.

Phan, T.G., Kapusinszky, B., Wang, C., Rose, R.K., Lipton, H.L., Delwart, E.L., 2011. The Fecal Viral Flora of Wild Rodents. PLOS Pathog. 7, e1002218.

Phan, T.G., Luchsinger, V., Avendano, L.F., Deng, X.T., Delwart, E., 2014. Cyclovirus in nasopharyngeal aspirates of Chilean children with respiratory infections. J. Gen. Virol. 95, 922-927.

Phan, T.G., Mori, D., Deng, X.T., Rajindrajith, S., Ranawaka, U., Ng, T.F.F., Bucardo-Rivera, F., Orlandi, P., Ahmed, K., Delwart, E., 2015. Small circular single stranded DNA viral genomes in unexplained cases of human encephalitis, diarrhea, and in untreated sewage. Virology 482, 98-104.

Piewbang, C., Jo, W.K., Puff, C., van der Vries, E., Kesdangsakonwut, S., Rungsipipat, A., Kruppa, J., Jung, K., Baumgartner, W., Techangamsuwan, S., Ludlow, M., Osterhaus, A., 2018. Novel canine circovirus strains from Thailand: Evidence for genetic recombination. Sci. Rep. 8, 7524.

Quaiser, A., Krupovic, M., Dufresne, A., Francez, A.J., Roux, S., 2016. Diversity and comparative genomics of chimeric viruses in Sphagnum-dominated peatlands. Virus Evol. 2, vew025.

Rahaus, M., Desloges, N., Probst, S., Loebbert, B., Lantermann, W., Wolff, M.H., 2008. Detection of beak and feather disease virus DNA in embryonated eggs of psittacine birds. Vet. Med. (Praha) 53, 53-58.

Raidal, S.R., Sarker, S., Peters, A., 2015. Review of psittacine beak and feather disease and its effect on Australian endangered species. Aust. Vet. J. 93, 466-470.

Reavy, B., Swanson, M.M., Cock, P.J.A., Dawson, L., Freitag, T.E., Singh, B.K., Torrance, L., Mushegian, A.R., Taliansky, M., 2015. Distinct Circular Single-Stranded DNA Viruses Exist in Different Soil Types. Appl. Environ. Microbiol. 81, 3934-3945.

Regenmortel, M.H.V.v., Fauquet, C.M., Bishop, D.H.L., Carstens, E.B., Estes, M.K., Lemon, S.M., Maniloff, J., Mayo, M.A., McGeoch, D.J., Pringle, C.R., Wickner, R.B., 2000. Virus taxonomy: classification and nomenclature of viruses. Seventh report of the International Committee on Taxonomy of Viruses. Academic Press, San Diego.

Regnard, G.L., Rybicki, E.P., Hitzeroth, II, 2017. Recombinant expression of beak and feather disease virus capsid protein and assembly of virus-like particles in Nicotiana benthamiana. Virol. J. 14, 174.

Reuter, G., Boros, A., Delwart, E., Pankovics, P., 2014. Novel circular single-stranded DNA virus from turkey faeces. Arch. Virol. 159, 2161-2164.

Rey, M.E.C., Ndunguru, J., Berrie, L.C., Paximadis, M., Berry, S., Cossa, N., Nuaila, V.N., Mabasa, K.G., Abraham, N., Rybicki, E.P., Martin, D., Pietersen, G., Esterhuizen, L.L., 2012. Diversity of Dicotyledenous-Infecting Geminiviruses and Their Associated DNA Molecules in Southern Africa, Including the South-West Indian Ocean Islands. Viruses 4, 1753-1791.

Ribeiro, S.G., Ambrozevicius, L.P., Avila, A.C., Bezerra, I.C., Calegario, R.F., Fernandes, J.J., Lima, M.F., de Mello, R.N., Rocha, H., Zerbini, F.M., 2003. Distribution and genetic diversity of tomato-infecting begomoviruses in Brazil. Arch. Virol. 148, 281-295.

Ritchie, B.W., Niagro, F.D., Lukert, P.D., Steffens, W.L., Latimer, K.S., 1989. Characterization of a new virus from cockatoos with psittacine beak and feather disease. Virology 171, 83-88.

Ritchie, P.A., Anderson, I.L., Lambert, D.M., 2003. Evidence for specificity of psittacine beak and feather disease viruses among avian hosts. Virology 306, 109-115.

Robino, P., Grego, E., Rossi, G., Bert, E., Tramuta, C., Stella, M.C., Bertoni, P., Nebbia, P., 2014. Molecular analysis and associated pathology of beak and feather disease virus isolated in Italy from young Congo African grey parrots (Psittacus erithacus) with an "atypical peracute form" of the disease. Avian Pathol. 43, 333-344.

Roossinck, M.J., 2011. The good viruses: viral mutualistic symbioses. Nat. Rev. Microbiol. 9, 99-108.

Roossinck, M.J., Saha, P., Wiley, G.B., Quan, J., White, J.D., Lai, H., Chavarria, F., Shen, G., Roe, B.A., 2010. Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol. Ecol. 19 Suppl 1, 81-88.

Rosario, K., Breitbart, M., Harrach, B., Segales, J., Delwart, E., Biagini, P., Varsani, A., 2017. Revisiting the taxonomy of the family *Circoviridae*: establishment of the genus Cyclovirus and removal of the genus Gyrovirus. Arch. Virol. 162, 1447-1463.

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

Rosario, K., Duffy, S., Breitbart, M., 2009. Diverse circovirus-like genome architectures revealed by environmental metagenomics. J. Gen. Virol. 90, 2418-2424.

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

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

Rosario, K., Marr, C., Varsani, A., Kraberger, S., Stainton, D., Moriones, E., Polston, J.E., Breitbart, M., 2016. Begomovirus-Associated Satellite DNA Diversity Captured Through Vector-Enabled Metagenomic (VEM) Surveys Using Whiteflies (Aleyrodidae). Viruses 8, 36.

Rosario, K., Mettel, K.A., Benner, B.E., Johnson, R., Scott, C., Yusseff-Vanegas, S.Z., Baker, C.C.M., Cassill, D.L., Storer, C., Varsani, A., Breitbart, M., 2018. Virus discovery in all three major lineages of terrestrial arthropods highlights the diversity of single-stranded DNA viruses associated with invertebrates. PeerJ 6, e5761.

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

Rosario, K., Schenck, R.O., Harbeitner, R.C., Lawler, S.N., Breitbart, M., 2015a. Novel circular single-stranded DNA viruses identified in marine invertebrates reveal high sequence diversity and consistent predicted intrinsic disorder patterns within putative structural proteins. Front. Microbiol. 6, 13.

Rosario, K., Seah, Y., Marr, C., Varsani, A., Kraberger, S., Stainton, D., Moriones, E., Polston, J., Duffy, S., Breitbart, M., 2015b. Vector-Enabled Metagenomic (VEM) Surveys Using Whiteflies (Aleyrodidae) Reveal Novel Begomovirus Species in the New and Old Worlds. Viruses 7, 2895.

Rose, N., Opriessnig, T., Grasland, B., Jestin, A., 2012. Epidemiology and transmission of porcine circovirus type 2 (PCV2). Virus Res. 164, 78-89.

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

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., Colombet, J., Sime-Ngando, T., Debroas, D., 2012. Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. PLoS One 7.

Roux, S., Solonenko, N.E., Dang, V.T., Poulos, B.T., Schwenk, S.M., Goldsmith, D.B., Coleman, M.L., Breitbare, M., Sullivan, M.B., 2016. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. PeerJ 4, 17.

Rwahnih, M.A., Dave, A., Anderson, M.M., Rowhani, A., Uyemoto, J.K., Sudarshana, M.R., 2013. Association of a DNA Virus with Grapevines Affected by Red Blotch Disease in California. Phytopathology 103, 1069-1076.

Rybicki, E.P., 1994. A phylogenetic and evolutionary justification for three genera of *Geminiviridae*. Arch. Virol. 139, 49-77.

Sachsenroder, J., Braun, A., Machnowska, P., Ng, T.F.F., Deng, X.T., Guenther, S., Bernstein, S., Ulrich, R.G., Delwart, E., Johne, R., 2014. Metagenomic identification of novel enteric viruses in urban wild rats and genome characterization of a group A rotavirus. J. Gen. Virol. 95, 2734-2747.

Sachsenroder, J., Twardziok, S., Hammerl, J.A., Janczyk, P., Wrede, P., Hertwig, S., Johne, R., 2012. Simultaneous Identification of DNA and RNA Viruses Present in Pig Faeces Using Process-Controlled Deep Sequencing. PLoS One 7, 11.

Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J., Walker, A.W., 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biology 12, 87.

Sanchez, R., Nauwynch, G., Pensaert, M., 2001. Serological survey of PCV-2 antibodies in domestic and feral pig populations in Belgium. Proc. ssDNA viruses of plants, birds, pigs and primates. St. Malo, France, p. 122.

Sánchez-Campos, S., Domínguez-Huerta, G., Díaz-Martínez, L., Tomás, D.M., Navas-Castillo, J., Moriones, E., Grande-Pérez, A., 2018. Differential Shape of Geminivirus Mutant Spectra Across Cultivated and Wild Hosts With Invariant Viral Consensus Sequences. Front. Plant Sci. 9, 932.

Sanjuán, R., Nebot, M.R., Chirico, N., Mansky, L.M., Belshaw, R., 2010. Viral Mutation Rates. J. Virol. 84, 9733.

Sarker, S., Forwood, J.K., Ghorashi, S.A., Peters, A., Raidal, S.R., 2015a. Beak and feather disease virus genotypes in Australian parrots reveal flexible host-switching. Aust. Vet. J. 93, 471-475.

Sarker, S., Moylan, K.G., Ghorashi, S.A., Forwood, J.K., Peters, A., Raidal, S.R., 2015b. Evidence of a deep viral host switch event with beak and feather disease virus infection in rainbow bee-eaters (Merops ornatus). Sci. Rep. 5, 14511.

Sarker, S., Patterson, E.I., Peters, A., Baker, G.B., Forwood, J.K., Ghorashi, S.A., Holdsworth, M., Baker, R., Murray, N., Raidal, S.R., 2014. Mutability Dynamics of an Emergent Single Stranded DNA Virus in a Naïve Host. PLoS One 9, e85370.

Sasaki, M., Orba, Y., Ueno, K., Ishii, A., Moonga, L., Hang'ombe, B.M., Mweene, A.S., Ito, K., Sawa, H., 2015. Metagenomic analysis of the shrew enteric virome reveals novel viruses related to human stool-associated viruses. J. Gen. Virol. 96, 440-452.

Sato, G., Kawashima, T., Kiuchi, M., Tohya, Y., 2015. Novel cyclovirus detected in the intestinal contents of Taiwan squirrels (Callosciurus erythraeus thaiwanensis). Virus Genes 51, 148-151.

Sattar, M.N., Kvarnheden, A., Saeed, M., Briddon, R.W., 2013. Cotton leaf curl disease - an emerging threat to cotton production worldwide. J. Gen. Virol. 94, 695-710.

Saunders, K., Bedford, I.D., Yahara, T., Stanley, J., 2003. The earliest recorded plant virus disease. Nature 422, 831.

Schoemaker, N.J., Dorrestein, G.M., Latimer, K.S., Lumeij, J.T., Kik, M.J., van der Hage, M.H., Campagnoli, R.P., 2000. Severe leukopenia and liver necrosis in young African grey parrots (Psittacus erithacus erithacus) infected with psittacine circovirus. Avian Dis. 44, 470-478.

Seal, S.E., Jeger, M.J., Van den Bosch, F., 2006a. Begomovirus Evolution and Disease Management. Ad. Vir. Res. 67, 297-316.

Seal, S.E., vandenBosch, F., Jeger, M.J., 2006b. Factors Influencing Begomovirus Evolution and Their Increasing Global Significance: Implications for Sustainable Control. Crit. Rev. Plant Sci. 25, 23-46.

Sertic, V., Bulgakov, N., 1935. Classification et identification des typhiphage. C R Soc. Biol. Paris 119, 1270-1272.

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

Sharman, M., Thomas, J.E., Skabo, S., Holton, T.A., 2008. Abaca bunchy top virus, a new member of the genus Babuvirus (family *Nanoviridae*). Arch. Virol. 153, 135-147.

Shen, H., Wang, C., Madson, D.M., Opriessnig, T., 2010. High prevalence of porcine circovirus viremia in newborn piglets in five clinically normal swine breeding herds in North America. Prev. Vet. Med. 97, 228-236.

Sicard, A., Yvon, M., Timchenko, T., Gronenborn, B., Michalakis, Y., Gutierrez, S., Blanc, S., 2013. Gene copy number is differentially regulated in a multipartite virus. Nat. Commun. 4, 2248.

Sikorski, A., Arguello-Astorga, G.R., Dayaram, A., Dobson, R.C.J., Varsani, A., 2013a. Discovery of a novel circular single-stranded DNA virus from porcine faeces. Arch. Virol. 158, 283-289.

Sikorski, A., Kearvell, J., Elkington, S., Dayaram, A., Arguello-Astorga, G.R., Varsani, A., 2013b. Novel ssDNA viruses discovered in yellow-crowned parakeet (Cyanoramphus auriceps) nesting material. Arch. Virol. 158, 1603-1607.

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

Simmonds, P., Adams, M.J., Benko, M., Breitbart, M., Brister, J.R., Carstens, E.B., Davison, A.J., Delwart, E., Gorbalenya, A.E., Harrach, B., Hull, R., King, A.M., Koonin, E.V., Krupovic, M., Kuhn, J.H., Lefkowitz, E.J., Nibert, M.L., Orton, R., Roossinck, M.J., Sabanadzovic, S., Sullivan, M.B., Suttle, C.A., Tesh, R.B., van der Vlugt, R.A., Varsani, A., Zerbini, F.M., 2017. Consensus statement: Virus taxonomy in the age of metagenomics. Nat. Rev. Microbiol. 15, 161-168.

Smits, S.L., Zijlstra, E.E., van Hellemond, J.J., Schapendonk, C.M., Bodewes, R., Schurch, A.C., Haagmans, B.L., Osterhaus, A.D., 2013. Novel cyclovirus in human cerebrospinal fluid, Malawi, 2010-2011. Emerg. Infect. Dis. 19, 1511-1513.

Soffer, N., Brandt, M.E., Correa, A.M.S., Smith, T.B., Thurber, R.V., 2014. Potential role of viruses in white plague coral disease. ISME J. 8, 271-283.

Stainton, D., Martin, D.P., Muhire, B.M., Lolohea, S., Halafihi, M.i., Lepoint, P., Blomme, G., Crew, K.S., Sharman, M., Kraberger, S., Dayaram, A., Walters, M., Collings, D.A., Mabvakure, B., Lemey, P., Harkins, G.W., Thomas, J.E., Varsani, A., 2015. The global distribution of Banana bunchy top virus reveals little evidence for frequent recent, human-mediated long distance dispersal events. Virus Evol. 1, vev009-vev009.

Stedman, K.M., 2015. Deep Recombination: RNA and ssDNA Virus Genes in DNA Virus and Host Genomes. Ann. Rev. Virol. 2, 203-217.

Steel, O., Kraberger, S., Sikorski, A., Young, L.M., Catchpole, R.J., Stevens, A.J., Ladley, J.J., Coray, D.S., Stainton, D., Dayarama, A., Julian, L., van Bysterveldt, K., Varsani, A., 2016. Circular replication-associated protein encoding DNA viruses identified in the faecal matter of various animals in New Zealand. Infect. Genet. Evol. 43, 151-164.

Steinfeldt, T., Finsterbusch, T., Mankertz, A., 2006. Demonstration of nicking/joining activity at the origin of DNA replication associated with the rep and rep' proteins of porcine circovirus type 1. J. Virol. 80, 6225-6234.

Stenzel, T., Piasecki, T., Chrzastek, K., Julian, L., Muhire, B.M., Golden, M., Martin, D.P., Varsani, A., 2014. Pigeon circoviruses display patterns of recombination, genomic secondary structure and selection similar to those of beak and feather disease viruses. J. Gen. Virol. 95, 1338-1351.

Stover, R.H., 1972. Banana, plantain, and abaca diseases. Commonwealth Mycological Institute, Kew, England.

Tan, L.V., van Doorn, H.R., Nghia, H.D.T., Chau, T.T.H., Tu, L.T.P., de Vries, M., Canuti, M., Deijs, M., Jebbink, M.F., Baker, S., Bryant, J.E., Tham, N.T., Bkrong, N.T.T.C., Boni, M.F., Loi, T.Q., Phuong, L.T., Verhoeven, J.T.P., Crusat, M., Jeeninga, R.E., Schultsz, C., Chau, N.V.V., Hien, T.T., van der Hoek, L., Farrar, J., de Jong, M.D., 2013. Identification of a New Cyclovirus in Cerebrospinal Fluid of Patients with Acute Central Nervous System Infections. mBio 4, e00231-00213.

Thresh, J.M., Cooter, R.J., 2005. Strategies for controlling cassava mosaic virus disease in Africa. Plant Pathology 54, 587-614.

Tijssen, P., Penzes, J.J., Yu, Q., Pham, H.T., Bergoin, M., 2016. Diversity of small, single-stranded DNA viruses of invertebrates and their chaotic evolutionary past. J. Invertebr. Pathol. 140, 83-96.

Timchenko, T., de Kouchkovsky, F., Katul, L., David, C., Vetten, H.J., Gronenborn, B., 1999. A Single Rep Protein Initiates Replication of Multiple Genome Components of Faba Bean Necrotic Yellows Virus, a Single-Stranded DNA Virus of Plants. Journal of Virology 73, 10173-10182.

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

Tischer, I., Gelderblom, H., Vettermann, W., Koch, M.A., 1982. A very small porcine virus with circular single-stranded DNA. Nature 295, 64-66.

Tischer, I., Rasch, R., Tochtermann, G., 1974. Characterization of papovavirus-and picornavirus-like particles in permanent pig kidney cell lines. Zentralblatt fur Bakteriologie, Parasitenkunde, Infektionskrankheiten und Hygiene. Erste Abteilung Originale. Reihe A: Medizinische Mikrobiologie und Parasitologie 226, 153-167.

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

Tomaru, Y., Shirai, Y., Suzuki, H., Nagumo, T., Nagasaki, K., 2008. Isolation and characterization of a new single-stranded DNA virus infecting the cosmopolitan marine diatom *Chaetoceros dehilis*. Aquat. Microb. Ecol. 50, 103-112.

Tomaru, Y., Shirai, Y., Toyoda, K., Nagasaki, K., 2011a. Isolation and characterisation of a single-stranded DNA virus infecting the marine planktonic diatom *Chaetoceros tenuissimus*. Aquat. Microb. Ecol. 64, 175-184.

Tomaru, Y., Takao, Y., Suzuki, H., Nagumo, T., Koike, K., Nagasaki, K., 2011b. Isolation and Characterization of a Single-Stranded DNA Virus Infecting *Chaetoceros lorenzianus Grunow*. Appl. Environ. Microbiol. 77, 5285-5293.

Tomaru, Y., Toyoda, K., Kimura, K., Hata, N., Yoshida, M., Nagasaki, K., 2012. First evidence for the existence of pennate diatom viruses. ISME J. 6, 1445-1448.

Tomaru, Y., Toyoda, K., Suzuki, H., Nagumo, T., Kimura, K., Takao, Y., 2013. New single-stranded DNA virus with a unique genomic structure that infects marine diatom *Chaetoceros setoensi*s. Sci. Rep. 3, 8.

Toyoda, K., Kimura, K., Hata, N., Nakayama, N., Nagasaki, K., Tomaru, Y., 2012. Isolation and characterization of a single-stranded DNA virus infecting the marine planktonic diatom *Chaetoceros* sp. (strain TG07-C28). Plank. Benthos Res. 7, 20-28.

Tu, J., Guo, J., Li, J., Gao, S., Yao, B., Lu, Z., 2015. Systematic Characteristic Exploration of the Chimeras Generated in Multiple Displacement Amplification through Next Generation Sequencing Data Reanalysis. PLoS One 10, e0139857.

Uch, R., Fournier, P.E., Robert, C., Blanc-Tailleur, C., Galicher, V., Barre, R., Jordier, F., de Micco, P., Raoult, D., Biagini, P., 2015. Divergent Gemycircularvirus in HIV-Positive Blood, France. Emerg. Infect. Dis. 21, 2096-2098.

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

van der Walt, E., Martin, D.P., Varsani, A., Polston, J.E., Rybicki, E.P., 2008. Experimental observations of rapid Maize streak virus evolution reveal a strand-specific nucleotide substitution bias. Virol. J. 5, 104.

Varma, A., Mandal, B., Singh, M.K., 2011. Global Emergence and Spread of Whitefly (Bemisia tabaci) Transmitted Geminiviruses, in: Thompson, W.M.O. (Ed.), The Whitefly, Bemisia tabaci (Homoptera: Aleyrodidae) Interaction with Geminivirus-Infected Host Plants: Bemisia tabaci, Host Plants and Geminiviruses. Springer Netherlands, Dordrecht, pp. 205-292.

Varsani, A., Krupovic, M., 2017. Sequence-based taxonomic framework for the classification of uncultured single-stranded DNA viruses of the family *Genomoviridae*. Virus Evol. 3, vew037.

Varsani, A., Krupovic, M., 2018. *Smacoviridae*: a new family of animal-associated single-stranded DNA viruses. Arch. Virol. 163, 2005-2015.

Varsani, A., Martin, D.P., Navas-Castillo, J., Moriones, E., Hernandez-Zepeda, C., Idris, A., Murilo Zerbini, F., Brown, J.K., 2014a. Revisiting the classification of curtoviruses based on genome-wide pairwise identity. Arch. Virol. 159, 1873-1882.

Varsani, A., Monjane, A.L., Donaldson, L., Oluwafemi, S., Zinga, I., Komba, E.K., Plakoutene, D., Mandakombo, N., Mboukoulida, J., Semballa, S., Briddon, R.W., Markham, P.G., Lett, J.-M., Lefeuvre, P., Rybicki, E.P., Martin, D.P., 2009. Comparative analysis of Panicum streak virus and Maize streak virus diversity, recombination patterns and phylogeography. Virol. J. 6, 194.

Varsani, A., Navas-Castillo, J., Moriones, E., Hernández-Zepeda, C., Idris, A., Brown, J.K., Murilo Zerbini, F., Martin, D.P., 2014b. Establishment of three new genera in the family *Geminiviridae*: Becurtovirus, Eragrovirus and Turncurtovirus. Arch. Virol. 159, 2193-2203.

Varsani, A., Roumagnac, P., Fuchs, M., Navas-Castillo, J., Moriones, E., Idris, A., Briddon, R.W., Rivera-Bustamante, R., Murilo Zerbini, F., Martin, D.P., 2017. Capulavirus and Grablovirus: two new genera in the family *Geminiviridae*. Arch. Virol. 162, 1819-1831.

Vega-Rocha, S., Byeon, I.J., Gronenborn, B., Gronenborn, A.M., Campos-Olivas, R., 2007a. Solution structure, divalent metal and DNA binding of the endonuclease domain from the replication initiation protein from porcine circovirus 2. J. Mol. Biol. 367, 473-487.

Vega-Rocha, S., Gronenborn, B., Gronenborn, A.M., Campos-Olivas, R., 2007b. Solution Structure of the Endonuclease Domain from the Master Replication Initiator Protein of the Nanovirus Faba Bean Necrotic Yellows Virus and Comparison with the Corresponding Geminivirus and Circovirus Structures. Biochemistry 46, 6201-6212.

Vetten, H.J., 2008. Nanoviruses, in: Mahy, B.W.J., Regenmortel, M.V. (Eds.), Encyclopedia of Virology, 3rd edn ed. Elsevier, Oxford, pp. 385-391.

Vidigal, P.M.P., Mafra, C.L., Silva, F.M.F., Fietto, J.L.R., Silva Júnior, A., Almeida, M.R., 2012. Tripping over emerging pathogens around the world: A phylogeographical approach for determining the epidemiology of Porcine circovirus-2 (PCV-2), considering global trading. Virus Res. 163, 320-327.

Vincent, I.E., Carrasco, C.P., Herrmann, B., Meehan, B.M., Allan, G.M., Summerfield, A., McCullough, K.C., 2003. Dendritic cells harbor infectious porcine circovirus type 2 in the absence of apparent cell modulation or replication of the virus. J. Virol. 77, 13288-13300.

Waits, K., Edwards, M.J., Cobb, I.N., Fontenele, R.S., Varsani, A., 2018. Identification of an anellovirus and genomoviruses in ixodid ticks. Virus Genes 54, 155-159.

Walker, I.W., Konoby, C.A., Jewhurst, V.A., McNair, I., McNeilly, F., Meehan, B.M., Cottrell, T.S., Ellis, J.A., Allan, G.M., 2000. Development and Application of a Competitive Enzyme-

Linked Immunosorbent Assay for the Detection of Serum Antibodies to Porcine Circovirus Type 2. J. Vet. Diagn. Invest. 12, 400-405.

Wang, H., Li, S.X., Mahmood, A., Yang, S.X., Wang, X.C., Shen, Q., Shan, T.L., Deng, X.T., Li, J.J., Hua, X.G., Cui, L., Delwart, E., Zhang, W., 2018. Plasma virome of cattle from forest region revealed diverse small circular ssDNA viral genomes. Virol. J. 15, 11.

Wege, C., Gotthardt, R.-D., Frischmuth, T., Jeske, H., 2000. Fulfilling Koch's postulates for Abutilon mosaic virus. Arch. Virol. 145, 2217-2225.

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

Woo, P.C., Lau, S.K., Teng, J.L., Tsang, A.K., Joseph, M., Wong, E.Y., Tang, Y., Sivakumar, S., Bai, R., Wernery, R., Wernery, U., Yuen, K.Y., 2014. Metagenomic analysis of viromes of dromedary camel fecal samples reveals large number and high diversity of circoviruses and picobirnaviruses. Virology 471-473, 117-125.

Wu, B., Melcher, U., Guo, X., Wang, X., Fan, L., Zhou, G., 2008. Assessment of codivergence of Mastreviruses with their plant hosts. BMC Evol. Biol. 8, 335.

Wu, Z., Yang, L., Ren, X., He, G., Zhang, J., Yang, J., Qian, Z., Dong, J., Sun, L., Zhu, Y., Du, J., Yang, F., Zhang, S., Jin, Q., 2016. Deciphering the bat virome catalog to better understand the ecological diversity of bat viruses and the bat origin of emerging infectious diseases. ISME J. 10, 609-620.

Wyant, P.S., Strohmeier, S., Schäfer, B., Krenz, B., Assunção, I.P., Lima, G.S.d.A., Jeske, H., 2012. Circular DNA genomics (circomics) exemplified for geminiviruses in bean crops and weeds of northeastern Brazil. Virology 427, 151-157.

Xia, H., Wang, Y., Shi, C., Atoni, E., Zhao, L., Yuan, Z., 2018. Comparative Metagenomic Profiling of Viromes Associated with Four Common Mosquito Species in China. Virol Sin 33, 59-66.

Yang, J. G., Wang, S. P., Liu, W., Li, Y., Shen, L. L., Qian, Y. M., Wang, F. L., and Du, Z. G. 2015. First Report of Milk vetch dwarf virus Associated With a Disease of Nicotiana tabacum in China. Plant Dis. 100, 1255.

Yoon, H.S., Price, D.C., Stepanauskas, R., Rajah, V.D., Sieracki, M.E., Wilson, W.H., Yang, E.C., Duffy, S., Bhattacharya, D., 2011. Single-Cell Genomics Reveals Organismal Interactions in Uncultivated Marine Protists. Science 332, 714.

Yoshida, M., Takaki, Y., Eitoku, M., Nunoura, T., Takai, K., 2013. Metagenomic analysis of viral communities in (hado)pelagic sediments. PLoS One 8, e57271.

Yu, X., Li, B., Fu, Y., Xie, J., Cheng, J., Ghabrial, S.A., Li, G., Yi, X., Jiang, D., 2013. Extracellular transmission of a DNA mycovirus and its use as a natural fungicide. Proc. Natl. Acad. Sci. USA 110, 1452-1457.

Yu, X., Li, B., Fu, Y.P., Jiang, D.H., Ghabrial, S.A., Li, G.Q., Peng, Y.L., Xie, J.T., Cheng, J.S., Huang, J.B., Yi, X.H., 2010. A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. Proc. Natl. Acad. Sci. USA 107, 8387-8392.

Zawar-Reza, P., Arguello-Astorga, G.R., Kraberger, S., Julian, L., Stainton, D., Broady, P.A., Varsani, A., 2014. Diverse small circular single-stranded DNA viruses identified in a freshwater pond on the McMurdo Ice Shelf (Antarctica). Infect. Genet. Evol. 26, 132-138.

Zerbini, F.M., Briddon, R.W., Idris, A., Martin, D.P., Moriones, E., Navas-Castillo, J., Rivera-Bustamante, R., Roumagnac, P., Varsani, A., Consortium, I.R., 2017. ICTV Virus Taxonomy Profile: *Geminiviridae*. J. Gen. Virol. 98, 131-133.

Zhang, W., Li, L., Deng, X., Kapusinszky, B., Pesavento, P.A., Delwart, E., 2014. Faecal virome of cats in an animal shelter. J. Gen. Virol. 95, 2553-2564.

Zhang, W., Yang, S., Shan, T., Hou, R., Liu, Z., Li, W., Guo, L., Wang, Y., Chen, P., Wang, X., Feng, F., Wang, H., Chen, C., Shen, Q., Zhou, C., Hua, X., Cui, L., Deng, X., Zhang, Z., Qi, D., Delwart, E., 2017. Virome comparisons in wild-diseased and healthy captive giant pandas. Microbiome 5, 90.

Zhao, G., Vatanen, T., Droit, L., Park, A., Kostic, A.D., Poon, T.W., Vlamakis, H., Siljander, H., Härkönen, T., Hämäläinen, A.-M., Peet, A., Tillmann, V. Ilonen, J., Wang, D., Knip, M., Xavier, R.J., Virgin, H.W., 2017. Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. Proc. Natl. Acad. Sci. USA 114, E6166.

Zhou, X., 2013. Advances in understanding begomovirus satellites. Annu. Rev. Phytopathol. 51, 357-381.

Zhou, X., Liu, Y., Calvert, L., Munoz, C., Otim-Nape, G.W., Robinson, D.J., Harrison, B.D., 1997. Evidence that DNA-A of a geminivirus associated with severe cassava mosaic disease in Uganda has arisen by interspecific recombination. J. Gen. Virol. 78, 2101-2111.

Chapter 4

The genealogy of the replication associated protein of circular rep-encoding single-stranded DNA viruses (CRESS DNA viruses)

Abstract

Large numbers of novel circular Rep-encoding ssDNA viruses (CRESS DNA viruses) have been discovered in the past decade, prompting a new appreciation for the ubiquity and genomic diversity of this group of viruses. Although highly divergent in the eukaryotic hosts they infect or are associated with, CRESS DNA viruses are united by the homologous replication-associated protein (Rep). An accurate genealogy of Rep can therefore provide insights into how these pathogens are related to each other. We worked with a dataset of CRESS DNA RefSeq genomes (n=926), which included representatives from all six established families and unclassified species. To assure an optimal Rep genealogy, we derived and tested a bespoke amino acid substitution model (named CRESS), which outperformed existing protein matrices in describing the evolution of Rep. The CRESS matrix was also selected as the best fitting to describe the evolution of several CRESS DNA families' capsid protein sequences and *Parvoviridae* NS1/Rep sequences, suggesting that the CRESS matrix has captured substitution patterns universal to CRESS DNA viruses (and perhaps ssDNA viral proteins in general). The CRESS model-estimated Rep genealogy revealed several intriguing relationships that had not been previously observed: most significantly a potential single origin of intron-containing Reps, which causes several geminivirus genera to group with *Genomoviridae* (bootstrap support 55%, aLRT SH-like support 0.997). While there are numerous unclassified CRESS DNA viruses throughout the tree, there is

significant intermingling with sequences assigned to *Circoviridae*, suggesting a needed expansion of *Circoviridae* or creation of further new CRESS DNA virus families.

Introduction

Our understanding of eukaryotic circular Rep-encoding ssDNA (CRESS DNA) viruses is changing, as this group is no longer restricted to plant and livestock infecting pathogens, but instead are now considered ubiquitous. With the application of the highly processive phi29 polymerase to enrich for circular DNA through rolling circle amplification, numerous publications have found CRESS DNA viruses (Haible et al., 2006; Inoue-Nagata et al., 2004; Li et al., 2010a; Rosario et al., 2009, 2012; Wyant et al., 2012). CRESS DNA viruses are found all around the world: in North America (Bistolas et al., 2017a; Bistolas et al., 2017b; Rosario et al., 2015), South America (Castrignano et al., 2017), Europe (Belabess et al., 2015; Reavy et al., 2015), Africa (Garigliany et al., 2014; Smits et al., 2013), Antarctica (Zawar-Reza et al., 2014), Arctic (Nguyen et al., 2017), Australia (Dayaram et al., 2016), Asia (Yu et al., 2010). Many CRESS DNA viruses have been associated with animals such as insects (Nakasu et al., 2017; Ng et al., 2011; Rosario et al., 2018; Waits et al., 2018; Xia et al., 2018), birds (Hanna et al., 2015; Kaszab et al., 2018; Sikorski et al., 2013), rodents (Phan et al., 2011; Sachsenroder et al., 2014; Sasaki et al., 2015), bats (Ge et al., 2011; Kemenesi et al., 2018; Li et al., 2010b; Lima et al., 2015; Male et al., 2016; Wu et al., 2016), chimps (Blinkova et al., 2010; Li et al., 2010a) and humans (Cui et al., 2017; Lamberto et al., 2014; Phan et al., 2014) and in myriad samples: water (Kimura and Tomaru, 2013; Kraberger et al., 2015a; McDaniel et al., 2014; Roux et al., 2012; Roux et al., 2016; Tomaru et al., 2013), feces (Castrignano et al., 2013; Conceicao-Neto et al., 2015; Garigliany et al., 2014; Kim et al., 2012; Li et al., 2010a; Ng et al., 2014; Ng et al., 2015; Phan et al., 2016; Reuter et al., 2014; Sikorski et al., 2013; Woo et al., 2014) and plant samples (Du et al., 2014; Gaafar et al., 2018; Kraberger et al., 2018; Male et al., 2015; Marzano and Domier, 2016; Yang et al., 2015), sometimes blood (Lamberto et al., 2014; Li et al., 2015; Uch et al., 2015), tissue (Decaro et al., 2014; Ge et al., 2011; Li et al., 2010a; Piewbang et al., 2018; Stenzel et al., 2014), and cerebrospinal fluids (Macera et al., 2016; Phan et al., 2015; Smits et al., 2013; Tan et al., 2013). Although we understood more of the ecological features of this group than before, continuous phylogenetic analyses and classifications are still necessary and ongoing. Alongside the discovery driven accelerated accumulation of new viral sequences deposited to Genbank, we are seeing commensurate taxonomical revisions and proposals to change the groupings of CRESS DNA viruses (Kazlauskas et al., 2017; Krupovic et al., 2016; Rosario et al., 2017; Varsani and Krupovic, 2017; Varsani and Krupovic, 2018; Varsani et al., 2014a; Varsani et al., 2014b; Varsani et al., 2017).

CRESS DNA viruses are known for their high mutation frequencies (Ge et al., 2007; Grigoras et al., 2014), fast evolution (Duffy and Holmes, 2008), wide host range and high recombination rates (Pearson et al., 2016), all of which contribute to the diversity and complexity of the CRESS DNA viral genealogy. Although some new sequences are highly similar to well defined families and are easily classified by percent similarity in genome composition and structure, there are still a great number of novel eukaryotic-associated CRESS viral sequences that do not cluster well with any characterized group (Ng et al., 2015). The relationships among CRESS DNA viruses are further complicated by recombination within this group (Kazlauskas et al., 2018). Several new CRESS DNA viral families remain to be proposed (Kazlauskas et al., 2018) and additional unclassified sequences await the discovery of similar sequences before classification may be attempted.

Building a phylogenetic tree is an established way to put new viral sequences into the context of well characterized viruses. However, unlike eukaryotic cytochrome c oxidase I and prokaryotic 16S phylogenetic trees, viruses do not share a universal gene that can be used to reconstruct their evolutionary history. Instead, the deep phylogeny of viruses is typically restricted to groups that share at least one homologous protein. RNA dependent RNA polymerase is the shared gene used to study the relationships among RNA viruses (Koonin, 1991; Koonin and Dolja, 2012; Payne, 2017). Glycoprotein B and DNA polymerase protein sequences have been used for the dsDNA herpesvirus phylogenetic analyses and taxonomic classifications (Chmielewicz et al., 2003; Ehlers et al., 1999; McGeoch and Gatherer, 2005). Despite varying genomic structure and size, all CRESS DNA viruses, by definition, share a replication associated protein (Rep) sequence. Therefore, Rep sequences are always used to study the phylogeny of CRESS DNA viruses. While Rep gene sequences can be useful for studying the relationships within an individual family of CRESS DNA viruses (Simmonds et al., 2017), ssDNA viruses are known to evolve as quickly as RNA viruses (Duffy and Holmes, 2008; Duffy et al., 2008; Firth et al., 2009; Harkins et al., 2009; Shackelton et al., 2005), quickly saturating the information in their nucleotide sequences (Melcher 2010). Therefore, it is necessary to use protein sequences to determine the evolutionary relationships of divergent members of CRESS DNA viral group.

Unfortunately, methods available for describing CRESS viral protein evolution are not ideal. Previous attempts to describe CRESS DNA viral Rep protein's evolutionary history have used a variety of amino acid substitution matrices, but there is reason to believe these matrices are poorly parameterized for CRESS DNA viruses. General matrices, such as LG, WAG, BLOSUM62, VT and PAM, are either estimated from aged and short protein sequence alignments, or from datasets containing protein sequences of multiple biological sources (mostly cellular). Organismal biologists in several areas have noticed the non-specific performance of these general matrices, and compensated for this inadequacy by estimating substitution matrices from highly specific protein sequences and then constructing phylogenies. Thus, the amino acid substitution matrices RtREV, cpREV, mtREV, HIVb, HIVw and FLU were developed from retro-transcribing elements, chloroplast, mitochondrial, HIV and influenza sequences respectively (Adachi and Hasegawa, 1996; Adachi et al., 2000; Dang et al., 2010; Dimmic et al., 2002; Nickle et al., 2007). Unsurprisingly, these specific matrices repeatedly outperform the general matrices in describing their designated sequences. The rtREV matrix has even been selected as the best-fitting matrix by ProtTest for CRESS DNA viral sequences on occasions (Dayaram et al., 2016;

Dayaram et al., 2015a; Dayaram et al., 2015b; Kraberger et al., 2015b; Rosario et al., 2015) – a result that highlights the inadequacies of the general substitution matrices to describe CRESS DNA viral evolution more than suggests that CRESS DNA viruses evolve identically to retro-transcribing viruses.

To best understand the evolution of CRESS DNA viruses, we need to optimize the methods used in describing the evolutionary patterns and constructing the genealogy of all CRESS DNA viruses. We attempted to create an amino acid matrix specific to CRESS DNA viruses, as has been done for other viral groups. The homologous protein Rep was used since it is present in all CRESS DNA viruses and divergence of Rep proteins defines some taxonomic classifications (Kazlauskas et al., 2017). Therefore, a CRESS Rep derived amino acid substitution matrix was estimated and validated with sequence data from GenBank. As expected, this CRESS matrix outperformed all other established matrices in describing CRESS Rep phylogeny. In addition, this matrix was the best fitting to describe datasets of other ssDNA viruses, including CRESS DNA viral capsid proteins (CP) and linear ssDNA virus *Parvoviridae* NS1 protein sequences. The CRESS matrix has captured the protein evolutionary pattern of ssDNA virus sequences, including for proteins unrelated to the Rep. The resulting CRESS DNA Rep genealogy also strongly supports a common origin for the intron-containing form of the Rep among some members of *Geminiviridae* and all *Genomoviridae*.

Material and Methods

Dataset generation

Collection: Rep sequences were downloaded from NCBI RefSeq December, 2017. RefSeq sequences were chosen to include unclassified viral sequences but exclude repeating sequences from the same species. Manual editing of concatenated *Mastrevirus* sequences was done after MUSCLE v3.8.31 (Edgar, 2004) alignment of all *Geminiviridae* Reps (details of edits in Appendix 4.1).

Alignment and trimming: All Rep sequences were pooled into one FASTA file and aligned using MUSCLE v3.8.31 (default setting: max 16 iterations) (Edgar, 2004). The multiple sequence alignments were trimmed using trimAl v1.2 command line (Capella-Gutierrez et al., 2009), resulting in alignments with the length of 329 aa, the average length of the Rep data set. This cut-off was chosen to represent the approximate size of an ideal Rep dataset. This was achieved by removing all columns with gaps in more than 40% (-gt 0.6) of the sequences while respecting the conservation of 17.12% of the columns (trimAl v1.2 adds columns in decreasing order of score when necessary). The trimmed MUSCLE alignment was used for matrix estimation.

Matrix estimation

ProtTest 3.4 (GUI) (Darriba et al., 2011) determined the best model for building the initial maximum likelihood (ML) tree of the trimmed MUSCLE alignment. A total of 80 models' NNI ML trees were compared: JTT, LG, Dayhoff, WAG, Blosum62, VT, rtREV, DCMut, MtREV, MtArt with combinations of +I, +G, +F. VT+G+F was the overall best performing model according to AIC, AICc, BIC scores. A full data set ML tree was constructed with VT+G+F using PhyML 3.1(Guindon et al., 2010) for subsequent computations.

Then, the 926 sequence alignment were randomly jackknifed into two halves as the training and test datasets using a python script (https://github.com/lzhao-virevol/matrix). Ten pairs of jackknifed datasets were generated as replicates. The initial maximum likelihood tree was also divided accordingly into ten training trees and ten test trees manually using Figtree (http://tree.bio.ed.ac.uk/software/figtree/). The training set was used to estimate an amino acid substitution matrix with a modified HyPhy batch file from Nickle et al. (Nickle et al., 2007) (https://github.com/lzhao-virevol/matrix) and FastMG (Dang et al., 2014). We used two matrices to seed our matrix estimation: VT, the best-fitting matrix to the full dataset and LG, the most recent general amino acid substitution matrix (Le and Gascuel, 2008), which is frequently used to describe CRESS DNA virus evolution (Bistolas et al., 2017a; Castrignano et al., 2017; Kaszab et al., 2018; Male et al., 2016). The HyPhy estimated matrices initiated with LG and VT are named fitLG and fitVT matrices. The FastMG estimated matrices initiated with LG and VT are named fmgLG and fmgVT matrices. Given the ten training datasets, 40 matrices were estimated.

Matrix evaluation

Likelihood scores were calculated for each matrix of interest (LG, RtREV, VT, fitLG1-10, fitVT1-10, fmgLG1-10, and fmgVT1-10) describing the ten test datasets (half alignment and tree) in PAML: codeml using model 3, Empirical+F (Yang, 1997). We included the LG and VT matrices that the estimated matrices were based on, and included rtREV because it is a specific matrix that has been previously used to describe CRESS DNA virus Rep evolution (Dayaram et al., 2016; Dayaram et al., 2015a; Dayaram et al., 2015b; Kraberger et al., 2015b; Rosario et al., 2015). To examine how similar the estimated matrices are to each other (and to the three established matrices), we calculated Pearson correlations in Excel (Redmond, WA) for the matrices.

The best-fitting matrix for of the ten test sets according to codeml results was used to build a maximum likelihood tree (PhyML3 with 4 discrete gamma rate categories, and empirical aa frequency options), producing ten test sets trees each. The maximum likelihood scores of these 100 trees were rank ordered by -In score within each test set. The matrix with the highest overall ranking was chosen as the best performing matrix, named the CRESS matrix. We compared the CRESS matrix to the established matrices VT, LG and rtREV using $\log_{10}$ ratios (Excel).

Dataset size validation

A series of datasets with identical sequence number and length as the training datasets were simulated to test if the training sets are sufficient to recover the amino acid substitution patterns. These pseudo datasets were simulated under the WAG model (Whelan and Goldman, 2001), with different site rate variation parameters (alpha: 0.1, 0.5, 1, and 2) in Seq-Gen v1.3.3 (Rambaut and

Grassly, 1997). We simulated alignments of similar size (463 sequences with the length of 329aa) to the training set Rep alignment. Four alignments (one for each parameter for the gamma distribution) were combined with the tree from training set 10 to produce eight substitution matrices by the same matrix estimation methods used to generate the CRESS Rep-derived matrices: fmg and fit (Dang et al., 2014; Nickle et al., 2007), seed matrix VT. The sequences-specific amino acid substitution matrix of each pseudo-dataset was estimated using the algorithms described above, and then compared to WAG through Pearson correlation tests. The $\log_{10}$ ratio matrices of WAG/Pseudo set matrix were also compared with Pearson correlation to screen for systematic patterns.

CRESS DNA viral Rep tree construction

Four aLRT SH-like supported maximum likelihood Rep trees were also built in PhyML3 using CRESS, VT, LG and rtREV matrices with +G+F options. 1000 bootstrapped trees were also estimated using RAxML v8.2.10 (Stamatakis, 2014) on CIPRES (RAxML-HPC2 on XSEDE) with +G+F options using CRESS, VT, LG and rtREV matrices. The bootstrapped trees were mapped onto corresponding aLRT SH-like supported ML trees with local non-MPI RAxML-HPC v8.1.17 (Stamatakis, 2014).

As there is a hotspot of recombination between the two domains of Rep (Kazlauskas et al., 2018), we separated the endonuclease and helicase domain alignments by splitting the trimmed MUSCLE aligned full-length Rep alignment, according to previously published descriptions (Kazlauskas et al., 2018). The endonuclease and helicase domain maximum likelihood trees were built in PhyML3 (CRESS+G+F).

We also removed sequences from our alignment that were determined to be recombinant by another study (Kazlauskas et al., 2018), and reduced the number of sequences used from the most overrepresented genus, *Begomovirus*. A total of 123 recombinants and 319 begomoviruses were separately removed from the full dataset's MUSCLE alignment. The two trees were built in PhyML3 using CRESS+G+F.

We aligned the 926 Rep sequences using MAFFT v7. 271 (options L-INS-I –ep 0.123) (Katoh and Standley, 2013) and also extracted the taxa from the MAFFT (and from the analogous MUSCLE) alignment that were members of *Geminiviridae* and *Genomoviridae* for additional analysis. Trees based on these three datasets were built using PhyML3 with CRESS matrix and +G+F options.

All trees were visualized with Figtree (http://tree.bio.ed.ac.uk/software/figtree/) and edited in Adobe Illustrator.

Model comparison using modified ProtTest

Capsid protein (CP) sequences from several families of CRESS DNA viruses and *Parvoviridae* RefSeq NS1 sequences were procured from NCBI in May 2018. Lists of CRESS DNA viral genome accession numbers were downloaded from the relevant family description from the International Committee on the Taxonomy of Viruses (ICTV, talk.ictvonline.org/taxonomy/) and if a CP was not identified in the sequence then the CP ORF was predicted using the NCBI ORF caller and validated through BLAST (accession numbers in Appendix 4.2). The 102 RefSeq *Parvoviridae* NS1/Rep sequences were either confirmed through database label or through BLAST (Appendix 4.2). Each CP dataset and the parvovirus NS1/Rep dataset were aligned using MUSCLE (default setting: max 16 iterations) and left untrimmed. The phage major capsid protein sequence alignment was from published dataset (Creasy et al., 2018), generously provided by K. Rosario (University of South Florida), and trimmed using -gappyout option in TrimAl 1.2. These alignments were provided to a modified version of ProtTest that includes the CRESS model (https://github.com/lzhao-virevol/matrix) to evaluate various model performance. All viral specific models (CRESS, rtREV, HIVb, HIVw, FLU) and two relevant general models (LG and

VT) with different combinations of +G, +F, +I (a total of 56 models) were tested using strategymode: NNI maximum likelihood tree.

Results

Matrix generation

926 Rep sequences from eukaryotic-infecting or eukaryotic-associated CRESS DNA viruses were aligned, trimmed, and an initial maximum likelihood tree was built with the best-fitting model chosen in ProtTest3 (uniformly chosen by AIC, AICc and BIC scores) which was VT+G+F. The trimmed alignment was jackknifed ten times and the tree split accordingly. Each training set half was used for amino acid substitution matrix estimation in four ways: with HyPhy or FastMG, and starting with the seed matrix of VT or LG. The HyPhy-estimated matrices for each of the ten training datasets are named fitVT1-10 and fitLG1-10, the FastMG estimated matrices are named fmgVT1-10 and fmgLG1-10 (a total of 40 estimated matrices). The four estimated matrices for each training dataset along with VT, LG and rtREV were fitted to the appropriate test set half and tree in PAML. The goodness of fit of each matrix were compared through maximum likelihood scores (FIGURE 4.1). In six out of ten jackknifed sets, the fmgVT matrix outperformed all other matrices and fmgLG was the best-fitting in three of the remaining four cases. The FastMG estimated matrices always outperformed the established matrices, but the HyPhy estimated matrices did not, further demonstrating that the FastMG alogorithm produced superior results. Interestingly, 9/10 times rtREV fit the test half of the dataset better than LG and VT, which had described the full dataset best.

We evaluated the consistency of jackknifed matrices by comparing their Pearson correlation coefficients (FIGURE 4.2). Matrices estimated by FastMG are very similar to each other, regardless of starting with the VT or LG matrix, with correlation coefficients ranging from 0.989 to 0.999. The HyPhy fit algorithm produced matrices with correlation coefficients ranging from 0.824 to 1. The lowest correlations (as low as 0.79) were seen when comparing the three established matrices (VT, LG and rtREV) to the estimated matrices, substantiating that these matrices may not best model with evolution of the Rep protein. Additionally, these correlations show that the FastMG approach finds the same signal from different jackknifed datasets, from two different starting points.

					Test Sets					
	1	2	3	4	5	6	7	8	9	10
fmgVT	-164332.68	-166702.6	-170493.13	-163893.1	-162400.5	-164353.67	-164055.57	-163685	-166218	-160477.89
fmgLG	-164333.16	-166723.23	-170488.5	-163926.06	-162399.85	-164364.54	-164059.56	-163708.96	-166207.96	-160485.4
fitVT	-164365	-166805	-173241	-163870	-165070	-164871.27	-166668.49	-166267	-168903	-160486
fitLG	-164480	-166782	-170548	-164061	-163354.91	-164486	-164167	-163933	-166264	-160718
rtREV	-165204.17	-167660.35	-171312.9	-164696.69	-163349	-165230	-164932.72	-164514	-167039	-161288
LG	-165232.14	-167637.29	-171371.36	-164760.32	-163354.91	-165268.77	-164929.03	-164516.53	-167081	-161365.11
VT	-166963.2	-169318.62	-173240.67	-166528.33	-165070.19	-166948.23	-166668.49	-166267	-168903	-163120

FIGURE 4.1. Comparison of likelihood scores of matrices fitting the test sets. The better the performance the darker the color. Numbers shown are -log likelihood score.

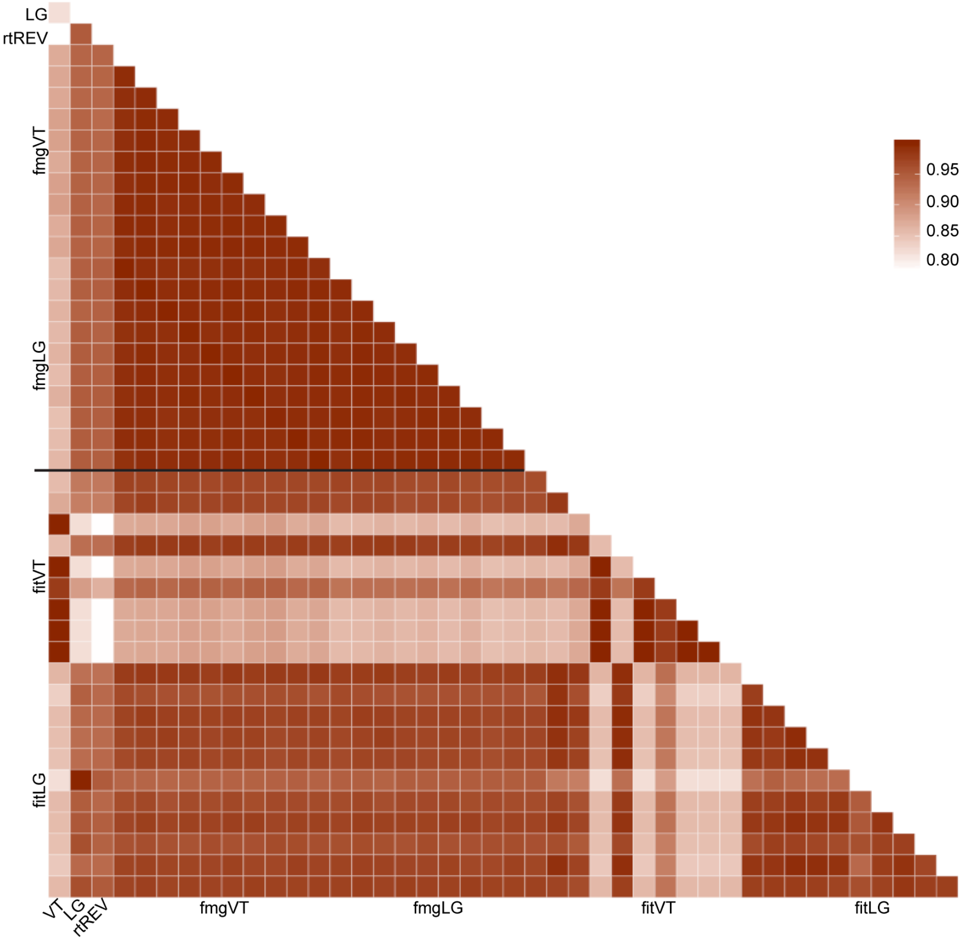


FIGURE 4.2. Pairwise Pearson correlation of amino acid substitution rates of matrices.

Dataset size validation

To validate the size of our training dataset, pseudo alignments were generated to evaluate the sufficiency of information within a dataset similar to our dataset. Using Seq-Gen v1.3.3, four alignments of similar size (463 sequences with the length of 329aa) to the training set Rep alignment were simulated with WAG matrix, each alignment with a different parameter in the gamma distribution of rate variation across sites (alpha=0.1, 0.5, 1, and 2). Identical methods as above for estimating substitution matrices were used, producing four fmg and four fit matrices. Pearson correlation coefficients of fmg matrices' rates to WAG rates range from 0.93 to 0.97, fit matrices' rates to WAG rates range from 0.87 to 0.98, indicate high similarity of the estimated matrices to the true matrix, WAG. This shows that the dataset size is sufficient for extracting protein substitution patterns. Ratios of simulated data derived substitution rates over WAG model substitution rates on a log scale were compared. Scattered over-/under-representations of rates inside the matrices are randomly distributed when compared across all simulated matrices over WAG (fmg/WAG log-ratio matrices have Pearson correlation coefficients ranging -0.5~0.33, fit/WAG log-ratio matrices have Pearson correlation coefficients ranging -0.14~0.27) (Appendix 4.3 & 4.4). This is expected because of the stochastic presence of certain low frequency amino acids with simulation, which might result in the estimated substitution rates to vary greatly from WAG. The variations among these ratio matrices show low similarity ruled out any extant systematic patterns within the two matrix estimation algorithms. Altogether, the results from these simulations validated our dataset size and the estimation methods we used.

Matrix selection

To select one jackknife matrix out of the ten to build a CRESS DNA viral rep tree, we built maximum likelihood trees from ten test set alignments using the best performing matrix of every jackknifed training sets. Then, we rank ordered the best matrices by the likelihood scores of the trees they generated (Appendix 4.5 & 4.6). Each matrix was responsible for estimating the tree with the lowest likelihood score when it comes to its corresponding test set, this is because the corresponding test set contained no overlapping information with the training set used to estimate the matrix, while other test sets may share sequence information with the training set from another jackknife set. The best matrix (fmgVT-10) with the lowest sum in ranking was named the CRESS matrix and used to construct the CRESS DNA viral rep tree.

Model comparison

We compared the default amino acid frequencies among models: CRESS, rtREV, LG and VT, which are used if users do not specify that they wish to use empirical amino acid frequencies from their alignments (FIGURE 4.3). There are many discordances among the amino acid frequencies within these models, showing that the amino acid sequences used to estimate the matrices are very different. Three amino acid frequencies were noticeably higher in CRESS than the other matrices. They are arginine (R), aspartic acid (D), and glutamic acid (E), all negatively charged, polar and hydrophilic. Hydrophobic leucine (L) is lower in frequency in CRESS compared to the other three models. The two matrices that are derived from viral sequences (rtREV, CRESS) differ from the general matrices in several amino acid frequencies as well: they have less alanine (A) and methionine (M), more proline (P) and tryptophan (W) . These differences in default amino acid frequencies further support the idea that specific models are necessary for describing viral protein evolution – the empirical amino acid frequencies of viral proteins are far from the default for proteins in general.

We also compared the substitution rates of these models, shown as heat maps are the $\log_{10}$ ratio of amino acid substitution rates of CRESS over rtREV, LG and VT (FIGURE 4.4). The CRESS matrix has generally lower rates of substitution compared to rtREV, has similar rates compared to LG and higher rates compared to VT. The CRESS matrix seems to have a consistently low proline (P) - Isoleucine (I) interchange rate compared to all other three matrices. More prominent than the general pattern, the CRESS matrix has higher rates of substitution when involving cysteine (C), methionine (M), histidine (H) and tryptophan (W) compared to VT.

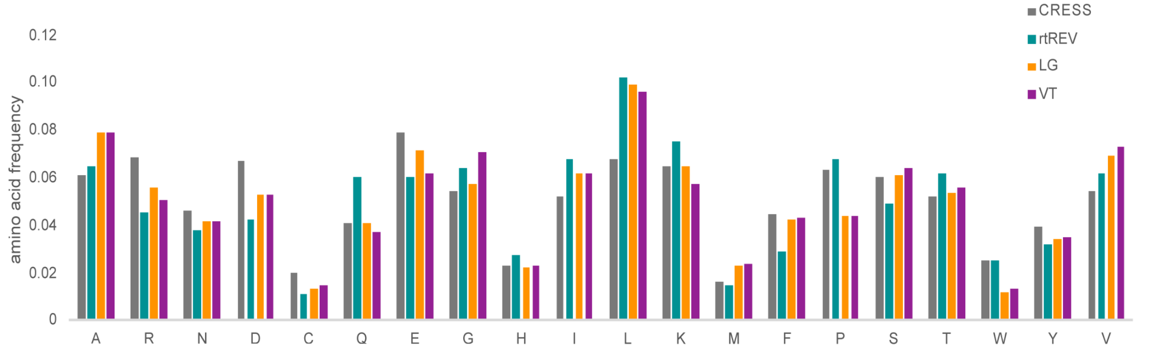


FIGURE 4.3. Comparison of amino acid frequencies among the matrices.

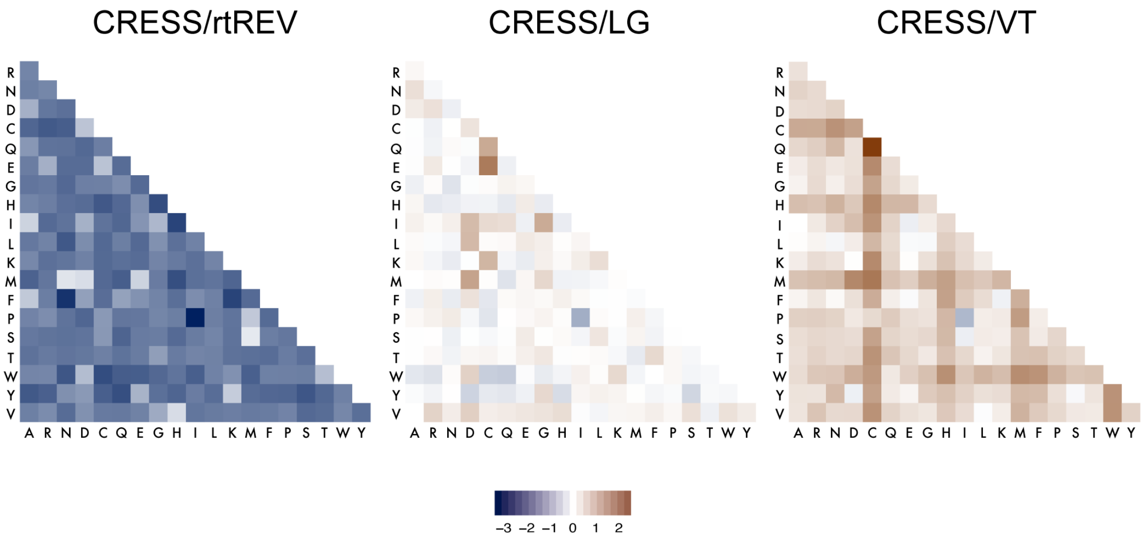


FIGURE 4.4. Comparison of four matrices. Log-ratio rate matrices of CRESS rates over rtREV, LG and VT are shown. The depth of blue color indicates substitutions where CRESS has lower rates, and the depth of brown color indicates substitutions where CRESS has higher rates.

Genealogy construction and comparison

Four aLRT SH-like support maximum likelihood tree were built using CRESS, rtREV, LG and VT in PhyML3. 1000 bootstrap trees were also generated for each matrix in RaxML, and the bootstrap support were mapped to the PhyML trees. Tree topologies of the four trees were similar with minor differences. All trees showed a single origin for the intron-containing form Rep. The CRESS matrix generated tree has 55% bootstrap support and 0.997 aLRT SH-like support for the *Genomoviridae* and intron-containing geminiviruses clade, while the three other matrices produced 45% bootstrap support, and 0.915-0.997 aLRT SH-like support for the clade (FIGURE 4.5 and Appendix 4.7- 4.10). With further inspection, most of the unclassified virus Rep (7/11) inside this clade also contain annotated introns in GenBank. And blastp results show the remaining four Reps to be highly similar to mastrevirus Rep, genomovirus Reps and other unclassified intron-containing Reps. To further confirm the single ancestral origin of intron-containing Reps in these two families, we divided the MUSCLE aligned full dataset into helicase and endonuclease domains as described in previous literature (Kazlauskas et al., 2018). The two resulting trees independently show support for the intron-containing form of Reps (endonuclease tree aLRT SH-like support 0.842, helicase tree aLRT SH-like support 0.909, (FIGURE 4.6 and FIGURE 4.7; Appendix 4.11 and 4.12)).

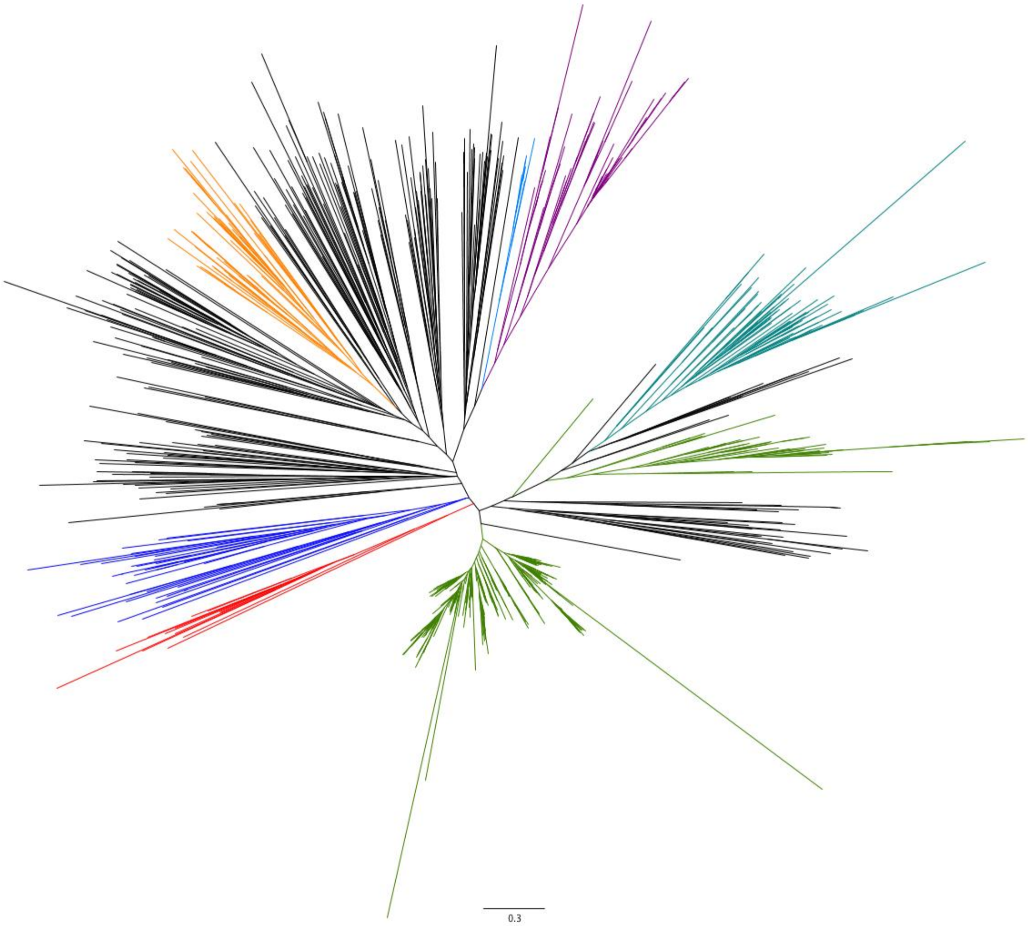


FIGURE 4.5 Maximum likelihood tree built using CRESS matrix. Green for *Geminiviridae*, teal for *Genomoviridae*, dark blue for *Smacoviridae*, red for *Bacilladnaviridae*, orange for *Circoviridae*, light blue for *Nanoviridae*, purple for the alphasatellites (*Alphasatellitidae*).

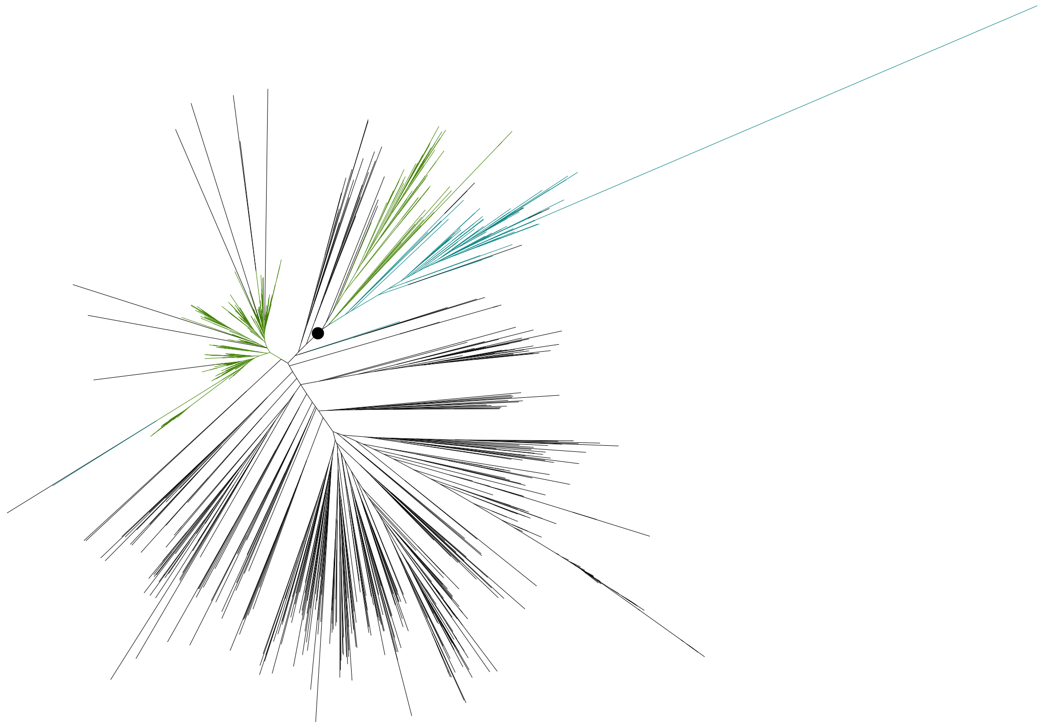


FIGURE 4.6 Unrooted maximum likelihood tree of CRESS Rep endonuclease domain. Genomovirus Reps colored in teal, geminivirus Reps colored in green. The black circle indicates the node for the common ancestor of intro-containing Reps (aLRT SH-like support 0.842).

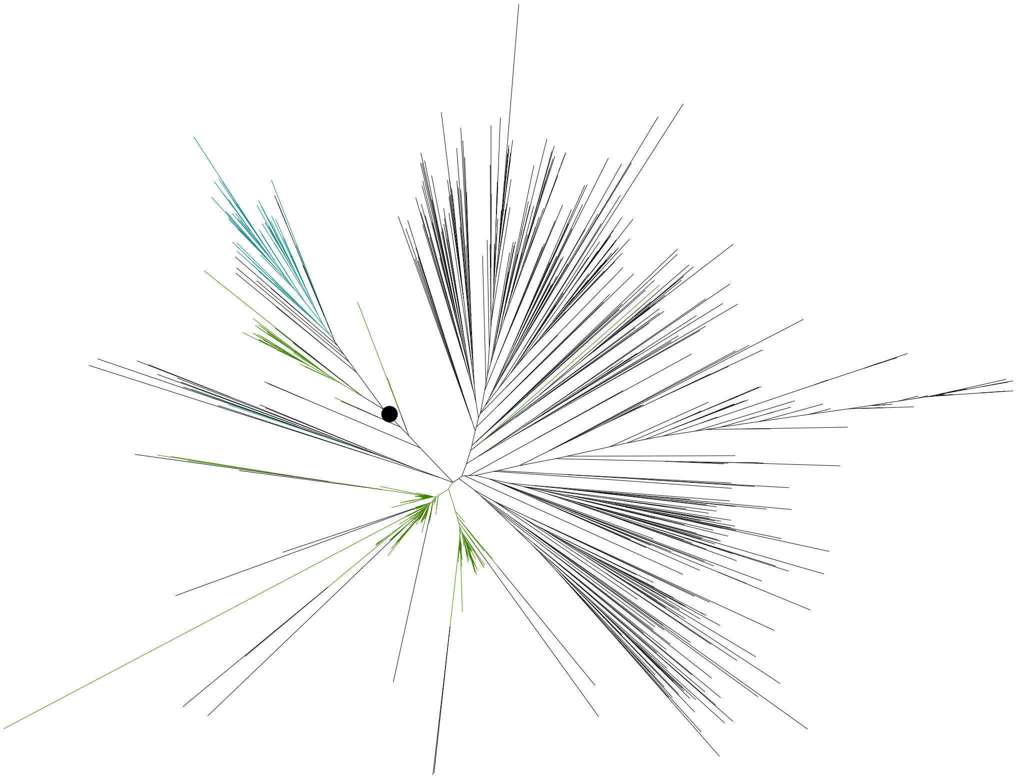


FIGURE 4.7 Unrooted maximum likelihood tree of CRESS Rep helicase domain. Genomovirus Reps colored in teal, geminivirus Reps colored in green. The black circle indicates the node for the common ancestor of intro-containing Reps (aLRT SH-like support 0.909).

.

The viral sequence-derived matrices (rtREV, CRESS) construct similar trees (FIGURE 4.5 and Appendix 4.8) which are more topologically different from those resulting from the two general matrices, which construct similar trees (Appendix 4.9 and Appendix 4.10). Both specific matrices placed *Nanoviridae* as a sister group to Alphasatellites, while the general matrices placed *Nanoviridae* inside the Alphasatellite clade. The general matrices place *Smacoviridae* within the *Geminiviridae* and *Genomoviridae* clade, but the specific matrices placed *Smacoviridae* outside of the said clade. CRESS matrix tree is the only tree that grouped all currently classified *Bacilladnaviridae* in one clade, while all other matrices failed to place *Thalassionema nitzschioides* DNA virus Rep (BAN59850) inside the *Bacilladnaviridae* clade. And finally, unclassified Reps mostly scatter themselves around classified *Circoviridae*, as have seen in previous Rep trees.

Some unclassified Reps were placed in unexpected places in the CRESS matrix-built tree, such as the two protruding tips inside the Geminivirus clade. This suggests that the dataset does not have too many sequences that are similar to these unclassified Reps, so the alignment algorithm was not able to accurately align this sequence to other sequences present in the dataset. Thus, after alignment and trimming, the positions that are conserved of these unclassified orphan CRESS DNA virus Reps contained information which might not be representative of its true phylogenetic status or place in the tree. In addition, updates are still pending for many annotations in Genbank. Inconsistency in classification updates will also lead to Rep sequences showing up mislabeled in the tree.

Model performance on other sequences

Capsid protein(CP) sequences from CRESS DNA viral families *Bacilladnaviridae*, *Circoviridae*, *Genomoviridae*, *Nanoviridae* and *Smacoviridae* were downloaded and aligned using MUSCLE (details of the sequences used are found in Appendix 4.2). Model performance was ranked in Prottest3, comparing 56 different models (CRESS, LG, VT, rtREV, HIV-B, HIV-W, FLU in combination with +I, +G, +F parameters). The CRESS models (CRESS+G+F, CRESS+I+G+F) outperformed all other models in building ML trees with CRESS DNA viral CP alignments (TABLE 4.1). The *Parvoviridae* NS1/Rep sequence (Appendix 4.2) alignment was also tested, CRESS+I+G+F was the best model chosen based on three different information criteria (TABLE 4.1). A circular ssDNA phage major capsid protein multiple sequence alignment was provided by collaborators (K. Rosario and M. Breitbart, details in (Creasy et al., 2018)) and LG+I+G+F outperformed all other matrices tested, though the CRESS matrix was ranked second. These results indicate that the CRESS model is good at describing both Rep and CP protein evolution of eukaryote-associated ssDNA viruses, and potentially would be useful for some datasets of ssDNA phage protein sequences in the future.

TABLE 4.1. Best model chosen by ProtTest3 for each tested ssDNA protein alignment. The table shows the models chosen by these information criteria and their corresponding weights. The best performing models were ranked by Akaike information criterion (AIC), corrected Akaike information criterion (AICc) and Bayesian information criterion (BIC) scores and the weights of these models are shown in parenthesis. +G assumes gamma-distributed rate variation across sites. +I estimates the proportion of invariant sites, +F uses empirical amino acid frequency. * indicates that the top two models' weights do not sum to 1.

Alignment	AIC	AICc	BIC
Bacilladnaviridae CP	CRESS+I+G+F (1)	CRESS+I+G+F (1)	CRESS+I+G+F (1)
Circoviridae CP	CRESS+I+G+F (0.72) CRESS+G+F (0.28)	CRESS+I+G+F (0.59) CRESS+G+F (0.41)	CRESS+I+G+F (0.76) CRESS+G+F (0.24)
Genomoviridae CP	CRESS+G+F (1)	CRESS+G+F (1)	CRESS+G+F (1)
Smacoviridae CP	CRESS+I+G+F (0.96) CRESS+G+F (0.04)	CRESS+I+G+F (0.96) CRESS+G+F (0.04)	CRESS+I+G+F (0.76) CRESS+G+F (0.24)
Nanoviridae CP	LG+G+F (0.83) LG+I+G+F (0.17)	LG+G+F (0.87) LG+I+G+F (0.09)*	LG+G (0.92) LG+I+G (0.08)
Parvoviridae NS1/Rep	CRESS+I+G+F (0.99) CRESS+G+F (0.01)	CRESS+I+G+F (0.99) CRESS+G+F (0.01)	CRESS+I+G+F (0.90) CRESS+G+F (0.10)
ssDNA phage MCP	LG+I+G+F (1)	LG+I+G+F (1)	LG+I+G+F (1)

Discussion

In this study, we estimated the CRESS matrix and built a CRESS DNA viral Rep genealogy with the matrix. The CRESS matrix was generated and selected from ten jackknifed Rep datasets. It outperformed other existing models in describing CRESS DNA viral Reps and several multiple sequence alignments of ssDNA viral sequences. We also observed an unprecedented single evolutionary origin for intron-containing Rep shared by some members of the *Geminiviridae* and all in *Genomoviridae*. While some members of *Circoviridae* are also predicted to have an intron in their Rep gene, this is not in the same location as in the other two families, and likely represents a unique integration event (Mankertz and Hillenbrand, 2001).

The utility of the CRESS matrix.

Prior to phylogenetic reconstruction, tools such as JModelTest and ProtTest select for the most appropriate model for a given alignment with which to estimate phylogenetic relationships. General models such as Dayhoff, BLOSUM, VT, WAG, LG, are all estimated from unconstrained, non-specific aligned protein sequences. The most recent general model, LG, published in 2008, used the entire Pfam database (Finn et al., 2016) at the time, but the database is heavily weighted towards non-viral sequences (Skewes-Cox et al., 2014). Unfortunately, the majority of CRESS DNA viruses were discovered and listed in public available database (NCBI nucleotide database) after 2009 (Chapter 3), which marked the start of extensive application of phi29 processive polymerase to CRESS virus research. Thus, the LG model could not have incorporated the bulk of CRESS DNA protein sequence information we are now trying to understand, and since many of these viruses contain only two ORFs, their sequence diversity contributes little to the overall Pfam database. While general models do increasingly well at describing the patterns dominated by genetic code constraints and the physiochemical properties of the amino acids (Murrell et al., 2011), there will always be opportunities for specific matrices to resolve protein evolution in biological entities that have more unique lifestyles and constraints, such as fast-evolving viruses with single-stranded mutational biases (Cardinale et al., 2013; Frederico et al., 1990; Xia and Yuen, 2005).

Seeing a need to resolve amino acid substitution specific problems, researchers have developed specific protein substitution matrices for specific proteins, which account for different proteins experiencing different selective pressures and mutational biases. Among these matrices there are mtREV for mitochondrial sequences (Adachi and Hasegawa, 1996), rtREV for retro-transcribing sequences (Dimmic et al., 2002), cpREV for chloroplastic sequences (Adachi et al., 2000), HIV-Wm, HIV-Bm (Nickle et al., 2007) for within and between host HIV sequences, and FLU (Dang et al., 2010) for influenza sequences. There is also an attempt to provide algorithms for sequence specific estimation with limited data (Murrell et al., 2011). All of the specific matrices outperformed the general matrices in describing the sequences they were designated to model, both because they model different amino acid substitution patterns and because they account for differences in amino acid composition.

Published CRESS DNA phylogenetic trees based on Rep sequences select a variety of different protein substitution matrices – there has never been a dominant, preferred model. Sometimes general models (predominantly LG and VT) were chosen (Bistolas et al., 2017a; Castrignano et al., 2017; Kaszab et al., 2018; Male et al., 2016), but other alignments preferred rtREV, a specific model that is very distinct from the general matrices (Dayaram et al., 2016; Dayaram et al., 2015a; Dayaram et al., 2015b; Kraberger et al., 2015b; Rosario et al., 2015). Known for their similarly high rates of substitution to RNA viruses (Duffy and Holmes, 2008; Sarker et al., 2014), it is reasonable that ssDNA viruses might prefer a specific matrix partially based on retro-transcribing viral sequences. However, ssDNA viruses use their host's replication machinery rather than an RdRp, which may contribute to different substitution patterns (Duffy et al., 2008). The variety of selected matrices, and the evidence for mutational bias in ssDNA viruses (Cardinale et al., 2013; Frederico et al., 1990; Xia and Yuen, 2005) prompted us to evaluate a

Rep-specific matrix, which appears to be universally useful across the CRESS DNA viral proteins, and was a decent fit to a ssDNA phage capsid alignment as well. As the annotation of ssDNA phages known by sequences alone improves, we welcome the opportunity to test the fit of the CRESS matrix to their Rep homologs – and are curious if CRESS would outcompete established matrices to fit phage Rep evolution. Regardless, the CRESS matrix will likely become the consistently chosen matrix for eukaryotic CRESS DNA virus phylogenies, and hopefully the substitution rates estimated from CRESS DNA viral sequences helps accurately model their protein evolution, create more accurate alignments and be useful in searching databases for similar sequences (Thorne, 2000).

We found the fmg algorithm to perform better than fit. Most impressively, fmg successfully extracted a consistent signal of protein evolution that overcame the stochastic effects of different randomly split datasets. This stable performance occurred using both LG and VT separately as seed matrices. For the creation of future specific amino acid models, when reliable, known phylogenies are not available (cf. rtREV (Dimmic et al., 2002) and HIV (Nickle et al., 2007)), we would prefer FastMG (Dang et al., 2014).

Comparing the four matrices used in this study, it seems the CRESS matrix falls in the large gulf between the fast-changing rtREV and the slower general matrices (FIGURE 4.4). While we have not compared CRESS to all described matrices, it may fill a useful niche for other proteins' evolution. Just as the rtREV model has been useful beyond studies on retro-transcribing elements, the inclusion of the CRESS matrix in ProtTest may result in the model being used to develop phylogenies for proteins from distantly related viruses. Perhaps its moderate substitution rate would be preferred by other viruses that evolve faster than cells but more slowly than many RNA viruses (Duffy et al., 2008)

Single origin intron-containing Rep

The Reps from *Genomoviridae* and four genera within *Geminiviridae* grouped together within a supported clade (55% bootstrap, aLRT SH-llike support 0.997) in the tree with 926 sequences (FIGURE 4.5). This suggest a common ancestry for the intron-containing form of Rep from an non-spliced form of Rep. This branching pattern has not been previously published, as all previous trees showed Reps from *Genomoviridae* and *Geminiviridae* as reciprocally monophyletic (e.g., (Dayaram et al., 2015a; Simmonds et al., 2017; Zawar-Reza et al., 2014)). However, the grouping of intron containing Reps is appealing from the perspective of Occam's razor: it is more plausible that an intron was only inserted once in the same location in evolutionary history of the CRESS DNA viruses instead of being inserted at the same location multiple times in two diverged groups. The fact that others have not observed this in previous trees could be due to the common practice of not including all geminivirus sequences (or even all genera) in trees. As *Geminiviridae* is the most speciose viral family, most researchers use only a small number of representative geminiviruses in a dataset, and these representatives most often come from *Begomovirus*, which comprises 75% of the annotated geminivirus species (https://talk.ictvonline.org/taxonomy/) and the vast majority of geminivirus sequences in GenBank. Begomoviruses Reps do not contain an intron, so often the sequence diversity of geminivirus Reps was not being adequately represented (Castrignano et al., 2017; Simmonds et al., 2017; Varsani and Krupovic, 2018). Our complete dataset contained the Reps of all species in *Geminiviridae*, including the intron-containing Rep sequences of all members of the genera *Mastrevirus, Becurtovirus, Grablovirus,* and *Capulavirus*. Interestingly, when we reduced the number of *Begomovirus* Reps in our dataset to the same size as *Mastrevirus* (n=37), the single origin of intron-containing geminivirus and genomovirus Reps loses support (FIGURE 4.8 and Appendix 4.13). Conversely, trimming the multiple sequence alignment did not affect this single intron integration event; trimmed and untrimmed alignments of our 926 sequences did not change major tree branching patterns (data not shown). Using all CRESS DNA virus species' Rep sequences, regardless of alignment trimming, contributed to our novel, supported clade connecting the intron-containing Reps of *Genomoviridae* and four genera of *Geminiviridae*.

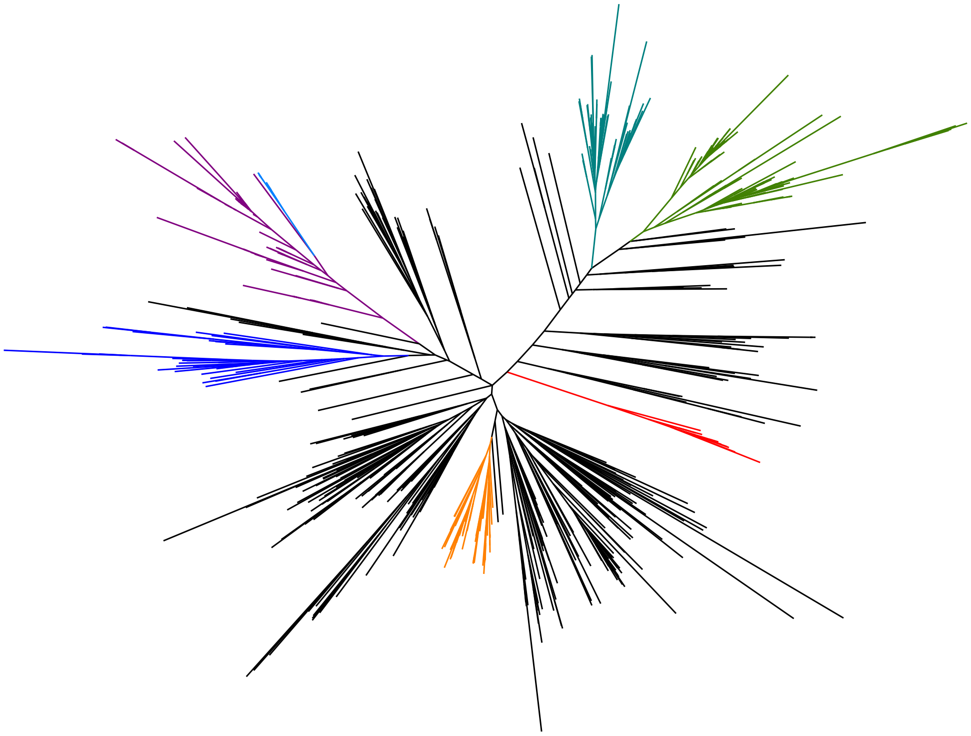


FIGURE 4.8 Unrooted maximum likelihood trees with equal number of *Begomovirus* and *Mastrevirus* Reps built with CRESS matrix. Alignment produced by MUSCLE. Green for *Geminiviridae*, teal for *Genomoviridae*, dark blue for *Smacoviridae*, red for *Bacilladnaviridae*, orange for *Circoviridae*, light blue for *Nanoviridae*, purple for the alphasatellites (*Alphasatellitidae*).

To further probe the difference between our tree and a recently published Rep tree that showed members of *Geminiviridae* and *Genomoviridae* as separate clades (Kazlauskas et al., 2018), we examined the methods used to make the two Rep trees. The previously published study used a smaller dataset of CRESS DNA virus Rep sequences (n=647), which was further pruned by removing detectable recombinants (final n = 380). When we removed the recombinant sequences detected in that study from our dataset, we still supported the single origin of intron-containing Reps in *Genomoviridae* and the appropriate genera from *Geminiviridae* (FIGURE 4.9). Additionally, there were differences in alignment algorithms, as the other study used MAFFT instead of MUSCLE. We aligned our original dataset with MAFFT and compared its ML tree (FIGURE 4.10 and Appendix 4.15) with our original MUSCLE alignment derived tree (FIGURE 4.5). The MAFFT tree of our entire dataset did not show single intron origin for members of *Geminiviridae* and *Genomoviridae*. However, when we extracted the geminivirus and genomovirus sequences from the MAFFT alignment and the original MUSCLE alignment and built smaller trees (FIGURE 4.11 &4.12 and Appendix 4.16 & 4.17), both showed the intron-containing clade of geminiviruses and genomoviruses separate from the non-intron containing clade of geminiviruses. This suggests that the inclusion of Reps from other families obscured the potential single origin of intron-containing Rep in the MAFFT alignment tree. While both algorithms have been used to analyze CRESS DNA virus evolution, there isn’t an a priori or a posteriori reason to prefer one algorithm over the other. Regardless, some datasets aligned with both algorithms provided support for a single origin of intron-containing Reps.

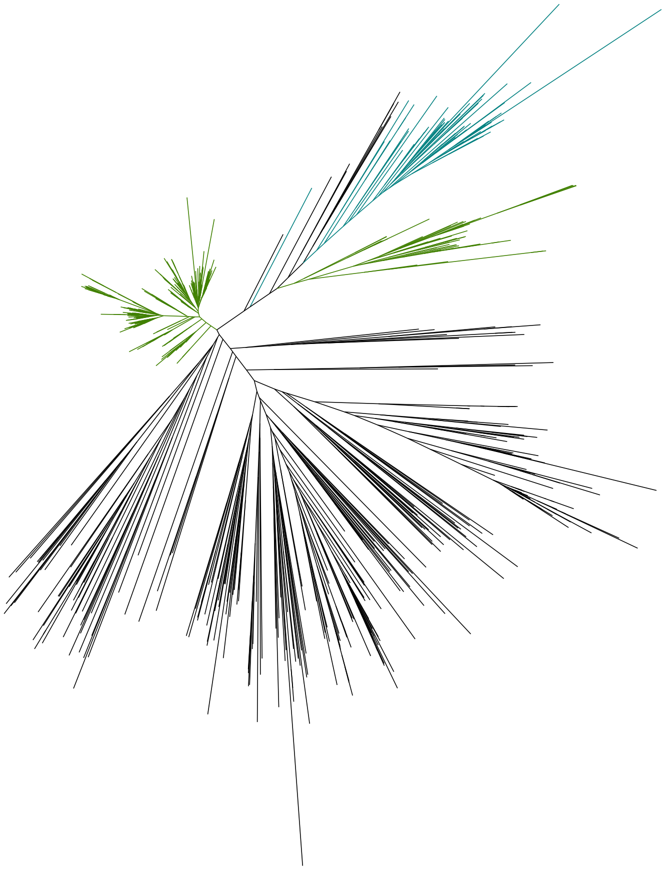


FIGURE 4.9 Unrooted maximum likelihood tree built with 123 recombinant sequences removed from the 926 CRESS Rep dataset. Green for *Geminiviridae*, teal for *Genomoviridae*.

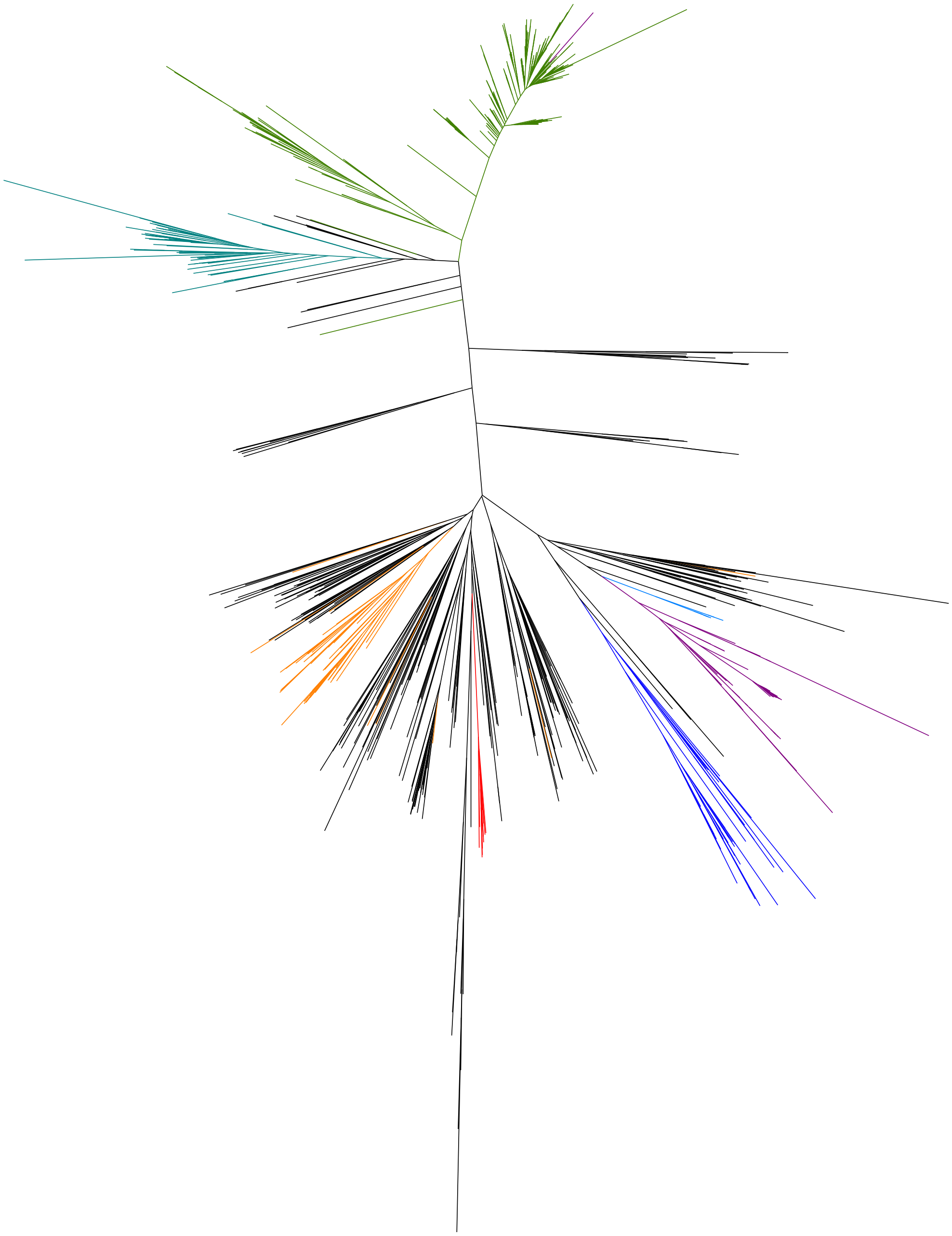


FIGURE 4.10 Unrooted maximum likelihood tree built with the full 926 CRESS Rep dataset. Alignment produced by MAFFT. Green for *Geminiviridae*, teal for *Genomoviridae*, dark blue for *Smacoviridae*, red for *Bacilladnaviridae*, orange for *Circoviridae*, light blue for *Nanoviridae*, purple for the alphasatellites (*Alphasatellitidae*).

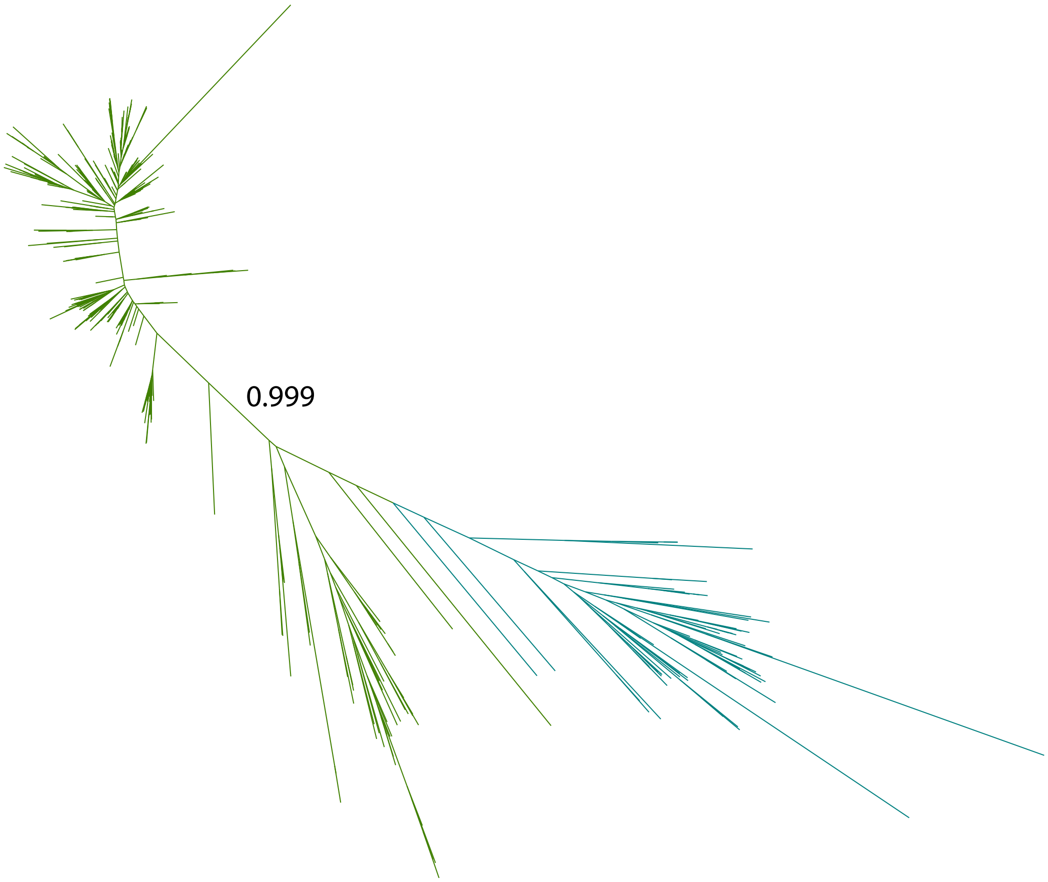


FIGURE 4.11 Unrooted maximum likelihood tree with Geminivirus and Genomovirus Reps built with CRESS matrix. Alignment produced by MAFFT. Green for *Geminiviridae*, teal for *Genomoviridae*. SH-like support labeled on the branch between the no intron Reps (left) and intron-containing Reps (right).

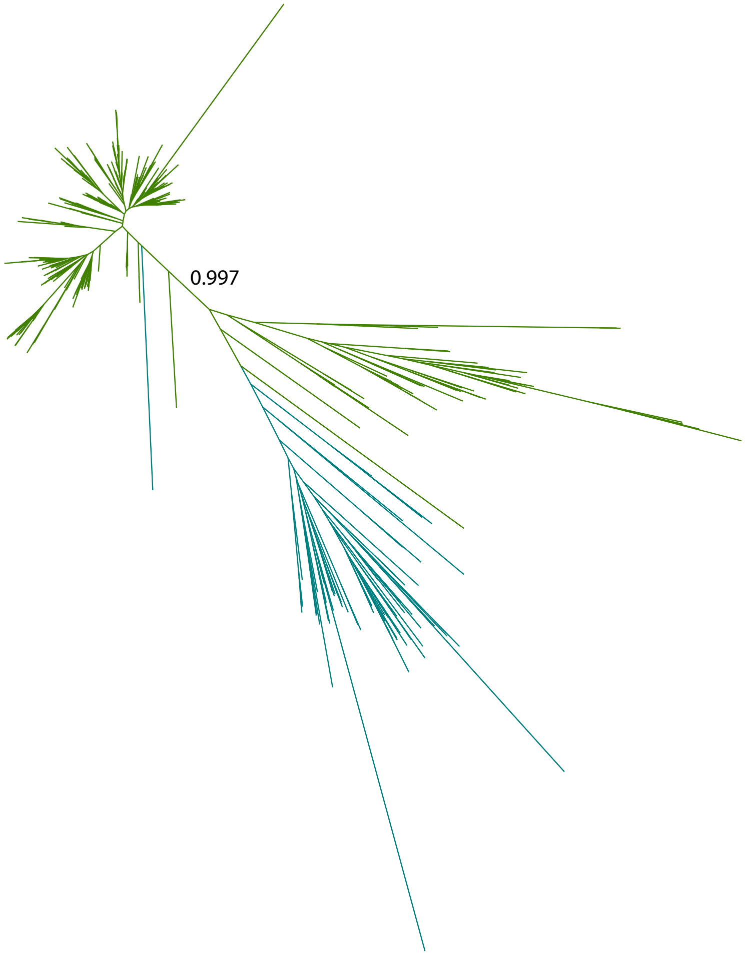


FIGURE 4.12 Unrooted maximum likelihood trees with Geminivirus and Genomovirus Reps built with CRESS matrix. Alignment produced by MUSCLE. Green for *Geminiviridae*, teal for *Genomoviridae*. SH-like support labeled on the branch between the no intron Reps and one intron-containing genomovirus (left) and the majority of the intron-containing Reps (right).

Relationships among other CRESS DNA viral families

Some of the large branching patterns observed on our full dataset tree are consistent with previous analyses. Notably, the close association of members of *Nanoviridae* and alphasatellites, and the close placement of members of *Genomoviridae* and *Geminiviridae*. In contrast, without strong support, bacilladnavirus and smacovirus Reps were close to each other in our tree compared to previous trees (Dayaram et al., 2015b; Varsani and Krupovic, 2018). The relatively few members of *Bacilladnaviridae* are known to infect diatoms (Kazlauskas et al., 2017; Nagasaki, 2008), which are phylogenetically closer to plants than animals. Therefore, we might expect their Reps to be more similar to other plant-infecting CRESS DNA viruses (*Nanoviridae, Geminiviridae*) instead of smacoviruses, which are associated with animals. In previous trees, Reps from *Smacoviridae* closely grouped with those of nanoviruses and alphasatellites, with strong support (Kazlauskas et al., 2018).

A majority of the Reps from unclassified CRESS DNA viruses are sister taxa to circovirus Reps. *Circoviridae* used to be the catch-all group for all circular eukaryotic infecting viruses, and species assigned to *Circoviridae* but not to a genus may represent additional genera or families of CRESS DNA viruses (Rosario et al., 2017). As previously proposed, many of the unclassified CRESS DNA viruses should be organized into novel families (Kazlauskas et al., 2018).

While we do not know what the true phylogeny of CRESS DNA viruses looks like, especially as discovery of this group is still ongoing, we believe that our comprehensive genealogy is the best representation of the relationship among the diverse Rep sequences that have been sampled to date. Our bespoke CRESS matrix will help expand phylogenetic analyses on CRESS DNA viruses as others discover additional novel representatives of this group.

Acknowledgements

This work was supported by the National Science Foundation’s Assembling the Tree of Life program, DEB1240049. We thank David Swofford for advice on data sufficiency validation methods, Li Li for developing the initial matrix estimation scheme, Steen Hoyer for technical help and support, and LaShanda Williams and Natasia Jacko for useful comments and support.

References

Adachi, J., Hasegawa, M., 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. Journal of molecular evolution 42, 459-468.

Adachi, J., Waddell, P.J., Martin, W., Hasegawa, M., 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. Journal of molecular evolution 50, 348-358.

Belabess, Z., Dallot, S., El-Montaser, S., Granier, M., Majde, M., Tahiri, A., Blenzar, A., Urbino, C., Peterschmitt, M., 2015. Monitoring the dynamics of emergence of a non-canonical recombinant of Tomato yellow leaf curl virus and displacement of its parental viruses in tomato. Virology 486, 291-306.

Bistolas, K.S.I., Jackson, E.W., Watkins, J.M., Rudstam, L.G., Hewson, I., 2017a. Distribution of circular single-stranded DNA viruses associated with benthic amphipods of genus Diporeia in the Laurentian Great Lakes. Freshwater Biology 62, 1220-1231.

Bistolas, K.S.I., Rudstam, L.G., Hewson, I., 2017b. Gene expression of benthic amphipods (genus: Diporeia) in relation to a circular ssDNA virus across two Laurentian Great Lakes. PeerJ 5, e3810.

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

Capella-Gutierrez, S., Silla-Martinez, J.M., Gabaldon, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics (Oxford, England) 25, 1972-1973.

Cardinale, J.D., DeRosa, K., Duffy, S., 2013. Base Composition and Translational Selection are Insufficient to Explain Codon Usage Bias in Plant Viruses. Viruses 5, 162-181.

Castrignano, S.B., Nagasse-Sugahara, T.K., Garrafa, P., Monezi, T.A., Barrella, K.M., Mehnert, D.U., 2017. Identification of circo-like virus-Brazil genomic sequences in raw sewage from the metropolitan area of Sao Paulo: evidence of circulation two and three years after the first detection. Mem. Inst. Oswaldo Cruz 112, 175-181.

Castrignano, S.B., Nagasse-Sugahara, T.K., Kisielius, J.J., Ueda-Ito, M., Brandao, P.E., Curti, S.P., 2013. Two novel circo-like viruses detected in human feces: complete genome sequencing and electron microscopy analysis. Virus Res. 178, 364-373.

Chmielewicz, B., Goltz, M., Franz, T., Bauer, C., Brema, S., Ellerbrok, H., Beckmann, S., Rziha, H.J., Lahrmann, K.H., Romero, C., Ehlers, B., 2003. A novel porcine gammaherpesvirus. Virology 308, 317-329.

Conceicao-Neto, N., Zeller, M., Heylen, E., Lefrere, H., Mesquita, J.R., Matthijnssens, J., 2015. Fecal virome analysis of three carnivores reveals a novel nodavirus and multiple gemycircularviruses. Virol. J. 12, 79.

Creasy, A., Rosario, K., Leigh, B.A., Dishaw, L.J., Breitbart, M., 2018. Unprecedented Diversity of ssDNA Phages from the Family Microviridae Detected within the Gut of a Protochordate Model Organism (Ciona robusta). Viruses 10, E404.

Cui, L.B., Wu, B.Y., Zhu, X.J., Guo, X.L., Ge, Y.Y., Zhao, K.C., Qi, X., Shi, Z.Y., Zhu, F.C., Sun, L.X., Zhou, M.H., 2017. Identification and genetic characterization of a novel circular single-stranded DNA virus in a human upper respiratory tract sample. Archives of Virology 162, 3305-3312.

Dang, C.C., Le, Q.S., Gascuel, O., Le, V.S., 2010. FLU, an amino acid substitution model for influenza proteins. BMC Evolutionary Biology 10, 99.

Dang, C.C., Le, V.S., Gascuel, O., Hazes, B., Le, Q.S., 2014. FastMG: a simple, fast, and accurate maximum likelihood procedure to estimate amino acid replacement rate matrices from large data sets. BMC Bioinformatics 15, 341.

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

Dayaram, A., Galatowitsch, M.L., Arguello-Astorga, G.R., van Bysterveldt, K., Kraberger, S., Stainton, D., Harding, J.S., Roumagnac, P., Martin, D.P., Lefeuvre, P., Varsani, A., 2016. Diverse circular replication-associated protein encoding viruses circulating in invertebrates within a lake ecosystem. Infect. Genet. Evol. 39, 304-316.

Dayaram, A., Goldstien, S., Arguello-Astorga, G.R., Zawar-Reza, P., Gomez, C., Harding, J.S., Varsani, A., 2015a. Diverse small circular DNA viruses circulating amongst estuarine molluscs. Infect. Genet. Evol. 31, 284-295.

Dayaram, A., Potter, K.A., Pailes, R., Marinov, M., Rosenstein, D.D., Varsani, A., 2015b. Identification of diverse circular single-stranded DNA viruses in adult dragonflies and damselflies (Insecta: Odonata) of Arizona and Oklahoma, USA. Infect. Genet. Evol. 30, 278-287.

Decaro, N., Martella, V., Desario, C., Lanave, G., Circella, E., Cavalli, A., Elia, G., Camero, M., Buonavoglia, C., 2014. Genomic characterization of a circovirus associated with fatal hemorrhagic enteritis in dog, Italy. PLoS One 9, e105909.

Dimmic, M.W., Rest, J.S., Mindell, D.P., Goldstein, R.A., 2002. rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. Journal of molecular evolution 55, 65-73.

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

Duffy, S., Holmes, E.C., 2008. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. J Virol 82, 957-965.

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

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

Ehlers, B., Borchers, K., Grund, C., Frolich, K., Ludwig, H., Buhk, H.J., 1999. Detection of new DNA polymerase genes of known and potentially novel herpesviruses by PCR with degenerate and deoxyinosine-substituted primers. Virus genes 18, 211-220.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A., 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research 44, D279-D285.

Firth, C., Charleston, M.A., Duffy, S., Shapiro, B., Holmes, E.C., 2009. Insights into the Evolutionary History of an Emerging Livestock Pathogen: Porcine Circovirus 2. Journal of Virology 83, 12813-12821.

Frederico, L.A., Kunkel, T.A., Shaw, B.R., 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry 29, 2532-2537.

Gaafar, Y., G., C.N., Ziebell, H., 2018. Molecular characterisation of the first occurrence of Pea necrotic yellow dwarf virus in Denmark. New Disease Reports 37, 16-16.

Garigliany, M.M., Hagen, R.M., Frickmann, H., May, J., Schwarz, N.G., Perse, A., Jost, H., Borstler, J., Shahhosseini, N., Desmecht, D., Mbunkah, H.A., Daniel, A.M., Kingsley, M.T., Campos Rde, M., de Paula, V.S., Randriamampionona, N., Poppert, S., Tannich, E., Rakotozandrindrainy, R., Cadar, D., Schmidt-Chanasit, J., 2014. Cyclovirus CyCV-VN species distribution is not limited to Vietnam and extends to Africa. Sci Rep 4, 7552.

Ge, L., Zhang, J., Zhou, X., Li, H., 2007. Genetic Structure and Population Variability of Tomato Yellow Leaf Curl China Virus. Journal of Virology 81, 5902-5907.

Ge, X.Y., Li, J.L., Peng, C., Wu, L.J., Yang, X.L., Wu, Y.Q., Zhang, Y.Z., Shi, Z.L., 2011. Genetic diversity of novel circular ssDNA viruses in bats in China. J. Gen. Virol. 92, 2646-2653.

Grigoras, I., Ginzo, A.I.d.C., Martin, D.P., Varsani, A., Romero, J., Mammadov, A.C., Huseynova, I.M., Aliyev, J.A., Kheyr-Pour, A., Huss, H., Ziebell, H., Timchenko, T., Vetten, H.-J., Gronenborn, B., 2014. Genome diversity and evidence of recombination and reassortment in nanoviruses from Europe. J. Gen. Virol. 95, 1178-1191.

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

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

Hanna, Z.R., Runckel, C., Fuchs, J., DeRisi, J.L., Mindell, D.P., Van Hemert, C., Handel, C.M., Dumbacher, J.P., 2015. Isolation of a Complete Circular Virus Genome Sequence from an Alaskan Black-Capped Chickadee (Poecile atricapillus) Gastrointestinal Tract Sample. Genome Announc 3, e01081-15.

Harkins, G.W., Martin, D.P., Duffy, S., Monjane, A.L., Shepherd, D.N., Windram, O.P., Owor, B.E., Donaldson, L., van Antwerpen, T., Sayed, R.A., Flett, B., Ramusi, M., Rybicki, E.P., Peterschmitt, M., Varsani, A., 2009. Dating the origins of the maize-adapted strain of maize streak virus, MSV-A. The Journal of General Virology 90, 3066-3074.

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

Kaszab, E., Marton, S., Forro, B., Bali, K., Lengyel, G., Banyai, K., Feher, E., 2018. Characterization of the genomic sequence of a novel CRESS DNA virus identified in Eurasian jay (Garrulus glandarius). Archives of Virology 163, 285-289.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution 30, 772-780.

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

Kazlauskas, D., Varsani, A., Krupovic, M., 2018. Pervasive Chimerism in the Replication-Associated Proteins of Uncultured Single-Stranded DNA Viruses. Viruses 10, 187.

Kemenesi, G., Kurucz, K., Zana, B., Foldes, F., Urban, P., Vlaschenko, A., Kravchenko, K., Budinski, I., Szodoray-Paradi, F., Bucs, S., Jere, C., Csosz, I., Szodoray-Paradi, A., Estok, P., Gorfol, T., Boldogh, S., Jakab, F., 2018. Diverse replication-associated protein encoding circular DNA viruses in guano samples of Central-Eastern European bats. Arch Virol 163, 671-678.

Kim, H.K., Park, S.J., Nguyen, V.G., Song, D.S., Moon, H.J., Kang, B.K., Park, B.K., 2012. Identification of a novel single-stranded, circular DNA virus from bovine stool. J. Gen. Virol. 93, 635-639.

Kimura, K., Tomaru, Y., 2013. Isolation and characterization of a single-stranded DNA virus infecting the marine diatom Chaetoceros sp. strain SS628-11 isolated from western Japan. PLoS One 8, e82013.

Koonin, E.V., 1991. The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. J. Gen. Virol. 72, 2197-2206.

Koonin, E.V., Dolja, V.V., 2012. Expanding networks of RNA virus evolution. BMC Biology 10, 54.

Kraberger, S., Arguello-Astorga, G.R., Greenfield, L.G., Galilee, C., Law, D., Martin, D.P., Varsani, A., 2015a. Characterisation of a diverse range of circular replication-associated protein encoding DNA viruses recovered from a sewage treatment oxidation pond. Infect. Genet. Evol. 31, 73-86.

Kraberger, S., Farkas, K., Bernardo, P., Booker, C., Arguello-Astorga, G.R., Mesleard, F., Martin, D.P., Roumagnac, P., Varsani, A., 2015b. Identification of novel Bromus- and Trifolium-associated circular DNA viruses. Archives of Virology 160, 1303-1311.

Kraberger, S., Kumari, S.G., Najar, A., Stainton, D., Martin, D.P., Varsani, A., 2018. Molecular characterization of faba bean necrotic yellows viruses in Tunisia. Arch Virol 163, 687-694.

Krupovic, M., Ghabrial, S.A., Jiang, D., Varsani, A., 2016. Genomoviridae: a new family of widespread single-stranded DNA viruses. Arch Virol 161, 2633-2643.

Lamberto, I., Gunst, K., Muller, H., Zur Hausen, H., de Villiers, E.M., 2014. Mycovirus-like DNA virus sequences from cattle serum and human brain and serum samples from multiple sclerosis patients. Genome Announc 2, e00848-14.

Le, S.Q., Gascuel, O., 2008. An improved general amino acid replacement matrix. Molecular biology and evolution 25, 1307-1320.

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

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

Li, W., Gu, Y., Shen, Q., Yang, S., Wang, X., Wan, Y., Zhang, W., 2015. A novel gemycircularvirus from experimental rats. Virus genes 51, 302-305.

Lima, F.E.D., Cibulski, S.P., dos Santos, H.F., Teixeira, T.F., Varela, A.P.M., Roehe, P.M., Delwart, E., Franco, A.C., 2015. Genomic Characterization of Novel Circular ssDNA Viruses from Insectivorous Bats in Southern Brazil. Plos One 10, e0118070.

Macera, L., Focosi, D., Vatteroni, M.L., Manzin, A., Antonelli, G., Pistello, M., Maggi, F., 2016. Cyclovirus Vietnam DNA in immunodeficient patients. J. Clin. Virol. 81, 12-15.

Male, M.F., Kami, V., Kraberger, S., Varsani, A., 2015. Genome Sequences of Poaceae-Associated Gemycircularviruses from the Pacific Ocean Island of Tonga. Genome Announc 3, e01144-15.

Male, M.F., Kraberger, S., Stainton, D., Kami, V., Varsani, A., 2016. Cycloviruses, gemycircularviruses and other novel replication-associated protein encoding circular viruses in Pacific flying fox (Pteropus tonganus) faeces. Infect. Genet. Evol. 39, 279-292.

Mankertz, A., Hillenbrand, B., 2001. Replication of porcine circovirus type 1 requires two proteins encoded by the viral rep gene. Virology 279, 429-438.

Marzano, S.L., Domier, L.L., 2016. Novel mycoviruses discovered from metatranscriptomics survey of soybean phyllosphere phytobiomes. Virus Res 213, 332-342.

McDaniel, L.D., Rosario, K., Breitbart, M., Paul, J.H., 2014. Comparative metagenomics: natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. Environ Microbiol 16, 570-585.

McGeoch, D.J., Gatherer, D., 2005. Integrating Reptilian Herpesviruses into the Family Herpesviridae. Journal of Virology 79, 725-731.

Murrell, B., Weighill, T., Buys, J., Ketteringham, R., Moola, S., Benade, G., du Buisson, L., Kaliski, D., Hands, T., Scheffler, K., 2011. Non-Negative Matrix Factorization for Learning Alignment-Specific Models of Protein Evolution. PLOS ONE 6, e28898.

Nagasaki, K., 2008. Dinoflagellates, diatoms, and their viruses. The Journal of Microbiology 46, 235-243.

Nakasu, E.Y.T., Melo, F.L., Michereff, M., Nagata, T., Ribeiro, B.M., Ribeiro, S.G., Lacorte, C., Inoue-Nagata, A.K., 2017. Discovery of two small circular ssDNA viruses associated with the whitefly Bemisia tabaci. Archives of Virology 162, 2835-2838.

Ng, T.F., Chen, L.F., Zhou, Y., Shapiro, B., Stiller, M., Heintzman, P.D., Varsani, A., Kondov, N.O., Wong, W., Deng, X., Andrews, T.D., Moorman, B.J., Meulendyk, T., MacKay, G., Gilbertson, R.L., Delwart, E., 2014. Preservation of viral genomes in 700-y-old caribou feces from a subarctic ice patch. Proceedings of the National Academy of Sciences of the United States of America 111, 16842-16847.

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

Ng, T.F.F., Zhang, W., Sachsenröder, J., Kondov, N.O., da Costa, A.C., Vega, E., Holtz, L.R., Wu, G., Wang, D., Stine, C.O., Antonio, M., Mulvaney, U.S., Muench, M.O., Deng, X., Ambert-Balay, K., Pothier, P., Vinjé, J., Delwart, E., 2015. A diverse group of small circular ssDNA viral genomes in human and non-human primate stools. Virus Evolution 1, vev017.

Nguyen, T.T., Robertsen, E.M., Landfald, B., 2017. Viral assemblage variation in an Arctic shelf seafloor. Aquat. Microb. Ecol. 78, 135-145.

Nickle, D.C., Heath, L., Jensen, M.A., Gilbert, P.B., Mullins, J.I., Pond, S.K., 2007. HIV-Specific Probabilistic Models of Protein Evolution. PLoS ONE 2, e503.

Payne, S., 2017. Chapter 10 - Introduction to RNA Viruses, in: Payne, S. (Ed.), Viruses. Academic Press, pp. 97-105.

Pearson, V.M., Caudle, S.B., Rokyta, D.R., 2016. Viral recombination blurs taxonomic lines: examination of single-stranded DNA viruses in a wastewater treatment plant. PeerJ 4, e2585.

Phan, T.G., da Costa, A.C., Mendoza, J.D., Bucardo-Rivera, F., Nordgren, J., O'Ryan, M., Deng, X.T., Delwart, E., 2016. The fecal virome of South and Central American children with diarrhea includes small circular DNA viral genomes of unknown origin. Archives of Virology 161, 959-966.

Phan, T.G., Kapusinszky, B., Wang, C., Rose, R.K., Lipton, H.L., Delwart, E.L., 2011. The Fecal Viral Flora of Wild Rodents. PLOS Pathogens 7, e1002218.

Phan, T.G., Luchsinger, V., Avendano, L.F., Deng, X.T., Delwart, E., 2014. Cyclovirus in nasopharyngeal aspirates of Chilean children with respiratory infections. J. Gen. Virol. 95, 922-927.

Phan, T.G., Mori, D., Deng, X.T., Rajindrajith, S., Ranawaka, U., Ng, T.F.F., Bucardo-Rivera, F., Orlandi, P., Ahmed, K., Delwart, E., 2015. Small circular single stranded DNA viral genomes in unexplained cases of human encephalitis, diarrhea, and in untreated sewage. Virology 482, 98-104.

Piewbang, C., Jo, W.K., Puff, C., van der Vries, E., Kesdangsakonwut, S., Rungsipipat, A., Kruppa, J., Jung, K., Baumgartner, W., Techangamsuwan, S., Ludlow, M., Osterhaus, A., 2018. Novel canine circovirus strains from Thailand: Evidence for genetic recombination. Sci Rep 8, 7524.

Rambaut, A., Grassly, N.C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Computer applications in the biosciences : CABIOS 13, 235-238.

Reavy, B., Swanson, M.M., Cock, P.J.A., Dawson, L., Freitag, T.E., Singh, B.K., Torrance, L., Mushegian, A.R., Taliansky, M., 2015. Distinct Circular Single-Stranded DNA Viruses Exist in Different Soil Types. Appl. Environ. Microbiol. 81, 3934-3945.

Reuter, G., Boros, A., Delwart, E., Pankovics, P., 2014. Novel circular single-stranded DNA virus from turkey faeces. Archives of Virology 159, 2161-2164.

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

Rosario, K., Duffy, S., Breitbart, M., 2009. Diverse circovirus-like genome architectures revealed by environmental metagenomics. J. Gen. Virol. 90, 2418-2424.

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

Rosario, K., Mettel, K.A., Benner, B.E., Johnson, R., Scott, C., Yusseff-Vanegas, S.Z., Baker, C.C.M., Cassill, D.L., Storer, C., Varsani, A., Breitbart, M., 2018. Virus discovery in all three major lineages of terrestrial arthropods highlights the diversity of single-stranded DNA viruses associated with invertebrates. PeerJ 6, e5761.

Rosario, K., Schenck, R.O., Harbeitner, R.C., Lawler, S.N., Breitbart, M., 2015. Novel circular single-stranded DNA viruses identified in marine invertebrates reveal high sequence diversity and consistent predicted intrinsic disorder patterns within putative structural proteins. Front. Microbiol. 6, 696.

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., Colombet, J., Sime-Ngando, T., Debroas, D., 2012. Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. PLoS ONE 7, e33641.

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

Sachsenroder, J., Braun, A., Machnowska, P., Ng, T.F.F., Deng, X.T., Guenther, S., Bernstein, S., Ulrich, R.G., Delwart, E., Johne, R., 2014. Metagenomic identification of novel enteric viruses in urban wild rats and genome characterization of a group A rotavirus. J. Gen. Virol. 95, 2734-2747.

Sarker, S., Patterson, E.I., Peters, A., Baker, G.B., Forwood, J.K., Ghorashi, S.A., Holdsworth, M., Baker, R., Murray, N., Raidal, S.R., 2014. Mutability Dynamics of an Emergent Single Stranded DNA Virus in a Naïve Host. PLOS ONE 9, e85370.

Sasaki, M., Orba, Y., Ueno, K., Ishii, A., Moonga, L., Hang'ombe, B.M., Mweene, A.S., Ito, K., Sawa, H., 2015. Metagenomic analysis of the shrew enteric virome reveals novel viruses related to human stool-associated viruses. J Gen Virol 96, 440-452.

Shackelton, L.A., Parrish, C.R., Truyen, U., Holmes, E.C., 2005. High rate of viral evolution associated with the emergence of carnivore parvovirus. Proceedings of the National Academy of Sciences of the United States of America 102, 379-384.

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

Simmonds, P., Adams, M.J., Benko, M., Breitbart, M., Brister, J.R., Carstens, E.B., Davison, A.J., Delwart, E., Gorbalenya, A.E., Harrach, B., Hull, R., King, A.M., Koonin, E.V., Krupovic, M., Kuhn, J.H., Lefkowitz, E.J., Nibert, M.L., Orton, R., Roossinck, M.J., Sabanadzovic, S., Sullivan, M.B., Suttle, C.A., Tesh, R.B., van der Vlugt, R.A., Varsani, A., Zerbini, F.M., 2017. Consensus statement: Virus taxonomy in the age of metagenomics. Nat Rev Microbiol 15, 161-168.

Skewes-Cox, P., Sharpton, T.J., Pollard, K.S., DeRisi, J.L., 2014. Profile hidden Markov models for the detection of viruses within metagenomic sequence data. PLoS One 9, e105067.

Smits, S.L., Zijlstra, E.E., van Hellemond, J.J., Schapendonk, C.M., Bodewes, R., Schurch, A.C., Haagmans, B.L., Osterhaus, A.D., 2013. Novel cyclovirus in human cerebrospinal fluid, Malawi, 2010-2011. Emerg Infect Dis 19, 1511.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics (Oxford, England) 30, 1312-1313.

Stenzel, T., Piasecki, T., Chrzastek, K., Julian, L., Muhire, B.M., Golden, M., Martin, D.P., Varsani, A., 2014. Pigeon circoviruses display patterns of recombination, genomic secondary structure and selection similar to those of beak and feather disease viruses. J. Gen. Virol. 95, 1338-1351.

Tan, L.V., van Doorn, H.R., Nghia, H.D.T., Chau, T.T.H., Tu, L.T.P., de Vries, M., Canuti, M., Deijs, M., Jebbink, M.F., Baker, S., Bryant, J.E., Tham, N.T., Bkrong, N.T.T.C., Boni, M.F., Loi, T.Q., Phuong, L.T., Verhoeven, J.T.P., Crusat, M., Jeeninga, R.E., Schultsz, C., Chau, N.V.V., Hien, T.T., van der Hoek, L., Farrar, J., de Jong, M.D., 2013. Identification of a New Cyclovirus in Cerebrospinal Fluid of Patients with Acute Central Nervous System Infections. mBio 4, e00231-13.

Thorne, J.L., 2000. Models of protein sequence evolution and their applications. Current opinion in genetics & development 10, 602-605.

Tomaru, Y., Toyoda, K., Suzuki, H., Nagumo, T., Kimura, K., Takao, Y., 2013. New single-stranded DNA virus with a unique genomic structure that infects marine diatom Chaetoceros setoensis. Scientific Reports 3, 3337.

Uch, R., Fournier, P.E., Robert, C., Blanc-Tailleur, C., Galicher, V., Barre, R., Jordier, F., de Micco, P., Raoult, D., Biagini, P., 2015. Divergent Gemycircularvirus in HIV-Positive Blood, France. Emerg Infect Dis 21, 2096-2098.

Varsani, A., Krupovic, M., 2017. Sequence-based taxonomic framework for the classification of uncultured single-stranded DNA viruses of the family Genomoviridae. Virus Evolution 3, vew037.

Varsani, A., Krupovic, M., 2018. Smacoviridae: a new family of animal-associated single-stranded DNA viruses. Arch Virol, 163, 2005-2015.

Varsani, A., Martin, D.P., Navas-Castillo, J., Moriones, E., Hernandez-Zepeda, C., Idris, A., Murilo Zerbini, F., Brown, J.K., 2014a. Revisiting the classification of curtoviruses based on genome-wide pairwise identity. Arch Virol 159, 1873-1882.

Varsani, A., Navas-Castillo, J., Moriones, E., Hernández-Zepeda, C., Idris, A., Brown, J.K., Murilo Zerbini, F., Martin, D.P., 2014b. Establishment of three new genera in the family Geminiviridae: Becurtovirus, Eragrovirus and Turncurtovirus. Archives of Virology 159, 2193-2203.

Varsani, A., Roumagnac, P., Fuchs, M., Navas-Castillo, J., Moriones, E., Idris, A., Briddon, R.W., Rivera-Bustamante, R., Murilo Zerbini, F., Martin, D.P., 2017. Capulavirus and Grablovirus: two new genera in the family Geminiviridae. Arch Virol 162, 1819-1831.

Waits, K., Edwards, M.J., Cobb, I.N., Fontenele, R.S., Varsani, A., 2018. Identification of an anellovirus and genomoviruses in ixodid ticks. Virus genes 54, 155-159.

Whelan, S., Goldman, N., 2001. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. Molecular biology and evolution 18, 691-699.

Woo, P.C., Lau, S.K., Teng, J.L., Tsang, A.K., Joseph, M., Wong, E.Y., Tang, Y., Sivakumar, S., Bai, R., Wernery, R., Wernery, U., Yuen, K.Y., 2014. Metagenomic analysis of viromes of dromedary camel fecal samples reveals large number and high diversity of circoviruses and picobirnaviruses. Virology 471-473, 117-125.

Wu, Z., Yang, L., Ren, X., He, G., Zhang, J., Yang, J., Qian, Z., Dong, J., Sun, L., Zhu, Y., Du, J., Yang, F., Zhang, S., Jin, Q., 2016. Deciphering the bat virome catalog to better understand the ecological diversity of bat viruses and the bat origin of emerging infectious diseases. ISME J 10, 609-620.

Wyant, P.S., Strohmeier, S., Schäfer, B., Krenz, B., Assunção, I.P., Lima, G.S.d.A., Jeske, H., 2012. Circular DNA genomics (circomics) exemplified for geminiviruses in bean crops and weeds of northeastern Brazil. Virology 427, 151-157.

Xia, H., Wang, Y., Shi, C., Atoni, E., Zhao, L., Yuan, Z., 2018. Comparative Metagenomic Profiling of Viromes Associated with Four Common Mosquito Species in China. Virol Sin 33, 59-66.

Xia, X., Yuen, K.Y., 2005. Differential selection and mutation between dsDNA and ssDNA phages shape the evolution of their genomic AT percentage. BMC genetics 6, 20.

Yang, J.G., Wang, S.P., Liu, W., Li, Y., Shen, L.L., Qian, Y.M., Wang, F.L., Du, Z.G., 2015. First Report of Milk vetch dwarf virus Associated With a Disease of Nicotiana tabacum in China. Plant Disease 100, 1255-1255.

Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer applications in the biosciences : CABIOS 13, 555-556.

Yu, X., Li, B., Fu, Y.P., Jiang, D.H., Ghabrial, S.A., Li, G.Q., Peng, Y.L., Xie, J.T., Cheng, J.S., Huang, J.B., Yi, X.H., 2010. A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. Proceedings of the National Academy of Sciences of the United States of America 107, 8387-8392.

Zawar-Reza, P., Arguello-Astorga, G.R., Kraberger, S., Julian, L., Stainton, D., Broady, P.A., Varsani, A., 2014. Diverse small circular single-stranded DNA viruses identified in a freshwater pond on the McMurdo Ice Shelf (Antarctica). Infect. Genet. Evol. 26, 132-138.

Chapter 5

Truly ubiquitous, CRESS DNA viruses scattered across the eukaryotic tree of life

Abstract

Until recently, most viruses detected and well-characterized were of economic significance, associated with agricultural and medical diseases. This was true for the circular Rep-encoding (replication-associated protein) single-stranded DNA (CRESS DNA) viruses, which were thought to be a relatively small group of viruses. With the explosion of metagenomic sequencing over the past decade and increasing use of the highly processive phi29 polymerase for sequence amplification, scientists have identified and annotated copious numbers of novel CRESS DNA viruses – many without known hosts, but which have been found in association with eukaryotes. While harnessing host machinery for replication, DNA viruses have been known to integrate into host genomes. Detecting endogenous sequences homologous to viral Reps will not only provide us with "fossil records" for protein evolution studies but also reveal potential host species for these viruses. The Rep protein is a conserved protein the CRESS DNA viruses all use to assist rolling circle replication, and which is known to be endogenized in a few plant lineages (some tobacco, yam). A systematic search for endogenous Rep sequences in GenBank non-redundant eukaryotic database was performed using tblastn. We utilized relaxed search criteria for the capture of integrated Rep sequence within eukaryotic genomes, identifying 163 unique species with an endogenized fragment of Rep in their nuclear, plasmid (2 species), mitochondrial (10 species) or chloroplast (10 species) genomes. These species come from 24 different phyla, scattered across the eukaryotic tree of life. Exogenous and endogenous CRESS DNA viral Rep tree topology suggested potential hosts for two families of uncharacterized viruses.

Introduction

Recent metagenomics advances have widened our knowledge of a previously understudied group of viruses, circular Rep-encoding ssDNA (CRESS DNA) viruses (see Chapter 3). The homologous Rep protein these viruses all share is a replication-associated protein that facilitates rolling-circle replication (Rosario et al., 2012b). The ubiquitous presence of these viruses in different environments has been confirmed by numerous sequencing efforts, but seldom have cellular hosts been identified in these studies. Many of these viruses may not be very virulent in their hosts (Roossinck and Bazán, 2017) and isolating these hundreds of unclassified viruses to screen against thousands to millions of potential hosts is a daunting task with low probability of success. Another way to narrow the potential host range for viruses known by sequence alone could be through finding "fossil" records inside host genomes, which would indicate that a related virus infected that host some time ago (Dennis et al., 2018b; Patel et al., 2011).

Many viruses integrate themselves inside host genomes during an infection, including retro-transcribing viruses replicating through an integrated DNA intermediate (Nisole and Saïb, 2004). Eight percent of the human genome consists of retroviral elements because they happened to insert themselves inside germline cells (Hayward and Katzourakis, 2015); the same process is currently ongoing in the koala genome (Stoye, 2006). Phages, with both dsDNA and ssDNA genomes, can be equipped with intergrases and tranposases to facilitate endogenization (Krupovic and Forterre, 2015). In eukaryotes, viruses that replicate in the host's nucleus will have a better chance of endogenization than others (Gilbert and Feschotte, 2010). While the mechanisms for retrovirus and dsDNA viral integration are well-studied, how other viral sequences become endogenized is an active area of research (Tu et al., 2017). The mechanism by which the CRESS DNA viruses integrate into their eukaryotic host genomes is still not clear (Krupovic and Forterre, 2015). The conventional wisdom is that viral use of host replication machinery inside the nucleus facilitates illegitimate recombination between viral and host genome (Belyi et al., 2010; Gilbert and Feschotte, 2010). This must sometimes occur in germline cell infection to allow CRESS DNA viruses' integration into the host's lineage.

Many previous studies have found endogenous CRESS DNA viruses in some eukaryotic genomes. One of the earliest studies identified 35 species with Rep-like sequences after a blast search using circovirus (animal-infecting CRESS DNA viruses), geminivirus and nanovirus (two families of plant-infecting CRESS DNA viruses) Rep proteins as queries (Liu et al., 2011). Circovirus-like Rep sequences were found in many vertebrates such cat, dog, panda, frog, opossum and sloth, and assuming these came from a single genomic integration event, it was dated to ~55 million years ago (Belyi et al., 2010). Geminiviruses have been found in a number of plant species. Multiple studies have reported *Begomovirus* (a genus within the plant-infecting *Geminiviridae*)-derived geminivirus-related DNA sequence inside several tobacco *Nicotiana* species: *Nicotiana tabacum, N. tomentosiformis, N. tomentosa* and *N. kawakamii*, (Ashby et al., 1997; Bejarano et al., 1996; Kenton et al., 1995; Murad et al., 2004). Evidence showed geminivirus might have integrated more than once into the ancestors of *Nicotiana* species (Murad et al., 2004). Two endogenized Rep fragments similar to geminiviruses have been found in the genome of the water yam (*Dioscorea alata*) and 22 other *Dioscorea* species. These Rep fragments appear to be actively expressed: they are under purifying selection, small RNAs and the expressed proteins have been detected in *Dioscorea* (Filloux et al., 2015). Recently, large magnitude surveys were conducted searching for traces of all non-retrotranscribing viral sequences in over four thousand eukaryotic genomes (Kryukov et al., 2018). While CRESS DNA viruses were not the only focus of that study, it primarily showed the distribution of endogenous sequences among diverse eukaryotic taxa. A more detailed search was also done for circovirus endogenous elements in vertebrate genome assemblies, suggesting that circovirus-like sequences had been introduced nineteen times into vertebrate germlines (Dennis et al., 2018a).

Paleovirology is the emergent field of studying ancient extinct viruses through endogenized sequences or viral “fossil records”. These sequences are not only useful in studying the origin and evolution of viruses, but also has many implications for how viruses have shaped the evolution of their hosts (Feschotte and Gilbert, 2012). Endogenized viral sequences are used to answer questions concerning evolutionary time-scale of the exogenous viruses, such as ancient host-shifting events and determining long-term substitution rates (Gilbert and Feschotte, 2010). Endogenized lentiviruses were used to estimate endogenization events about 4.2 million years ago, and suggested that endogenous sequences are very useful in studying ancestral host-pathogen dynamics and reconstructing ancient viruses (Gilbert et al., 2009). Paleovirology has already yielded important insights for CRESS DNA viruses. While studies of extant crop virus nucleotide sequences often coalesce around the time of the dawn of agriculture (~10,000 years ago), with the evidence of an endogenized Rep sequence from *Nicotiana* spp., we know that geminiviruses originated more than ten million years ago (Gibbs et al., 2006; Lefeuvre et al., 2011).

In this study, we performed a relaxed tblastn search of the non-redundant eukaryotic nucleotide database for endogenized CRESS DNA viral Reps, including a family of Rep-encoding alphasatellites that are known to be related to CRESS DNA viral Reps. We detected endogenous CRESS DNA Rep sequences in 543 unique accession entries from 163 unique eukaryotic species. The endogenous Rep fragments came from species of 24 different phyla, scattered across the eukaryotic tree of life. All viral families displayed intriguing findings, for example, genomovirus Reps were closely related to endogenous sequences from fungal genomes, showing that fungal species might indeed serve as hosts for all uncharacterized genomoviruses. The circovirus tree showed strong intermingling of exogenous and endogenous sequences. Geminivirus Reps were surrounded by endogenous sequences from *Nicotiana* and *Dioscorea* spp. as previously discovered. Endogenous sequences found by searching with nanoviruses and alphasatellites (associated with CRESS DNA viral infection of plants) Reps were from a wide range of hosts, not just their current plant host range. The few endogenous sequences found by searching with bacilladnavirus Reps were distantly related to exogenous viral sequences, which did not help explicate the evolutionary history of these unusual, diatom-infecting viruses. Finally, smacovirus Reps, which are from the only CRESS DNA virus family without a single cultured member, were unexpectedly found to be similar to sequences from diatoms, not from the animal species with which smacovirus sequences are often found in association.

Material and Methods

BLAST searches

Local tblastn runs were carried out with the replication-associated protein (Rep) sequences of CRESS DNA viruses from the RefSeq database (downloaded December 2017) as queries, against the non-redundant (nr) eukaryote nucleotide database (taxid: 2759; downloaded March 2018) from NCBI. Queries dataset include 8 Reps from *Bacilladnaviridae*, 154 Reps from *Circoviridae*, 416 Reps from *Geminiviridae*, 67 Reps from *Genomoviridae*, 8 Reps from *Nanoviridae*, 26 Reps from *Smacoviridae*, 66 Reps from Alphasatellites, and 160 Reps from unclassified ssDNA viruses (Appendix 4.1). The tblastn ran with the following relaxed criteria: BLOSUM50 matrix, word size 6, e-value threshold 0.001, gap penalty 15 and extension penalty 1 (Altschul et al., 1990).

BLAST results processing

Repetitive and overlapping hits of the same accession entry resulted from queries from the same family were merged into one consensus sequence manually using Seaview (Gouy et al., 2010). Consensus sequences were omitted if they are less than 50 amino acids in length. One thousand nucleotides up and down stream of the consensus sequence were extracted from Genbank and scanned for repetitive elements using WSCensor (http://www.girinst.org/censor/).

Endogenous and viral sequence alignment and tree generation

All endogenous consensus sequences and the viral Reps used to search for them were aligned in MUSCLE (default maximum 16 iteration) (Edgar, 2004) and trimmed using TrimAl (-gappyout) (Capella-Gutierrez et al., 2009). 256 begomoviruses were taken out of the dataset, leaving 100 representative members of *Begomovirus* within the geminivirus and endogenous sequences dataset. This is to save computation time and avoid overrepresentation of *Begomovirus*, the most speciose genus of all classified viruses, within the alignment. All trimmed endogenous and viral sequence alignments were inputs to PhyML 3.0 (Guindon and Gascuel, 2003) to estimate maximum likelihood trees using CRESS+G+F model (see Chapter 4 for details of CRESS protein matrix). Trees were visualized and colored using Figtree (http://tree.bio.ed.ac.uk/software/figtree/).

Results

tblastn results

With our relaxed search criteria and using 905 CRESS DNA viral Rep amino acid sequences as queries, we were able to obtain 111,344 raw hits after the search, which collapsed to 543 unique accession entries, 163 unique species and 24 eukaryotic phyla (Table 5.1 and Figure 5.1). Bacilladnavirus Reps found the smallest number of similar sequences in eukaryotic genomes and geminivirus Reps found the most number of hits per viral Rep. However, circovirus Reps were the most wide-spread sequences, as they were found in 89 unique species genomes. These endogenous Rep sequences spread across Plantae, Chromalveolates, Unikonts, and Excavates across the eukaryotic tree of life. The phyla with most representing species are Magnoliophyta, Ascomycota, Arthropoda and Chordata. The identified phyla had varying representation of endogenous sequences from the CRESS viral families.

TABLE 5.1 Summary of results from tblastn using viral Rep queries.

Virus family	Queries	Raw hits	Unique species	Number of species hit more than once	Consensus sequences (cutoff 50 amino acids)
Alphasatellitidae	66	2396	57	28	214
Bacilladnaviridae	8	55	4	2	13
Circoviridae	154	15680	89	47	366
Geminiviridae	416	71708	58	32	292
Genomoviridae	67	7290	41	25	175
Nanoviridae	8	359	22	9	64
Smacoviridae	26	233	11	2	15
Unclassified	160	13623	138	66	N/A

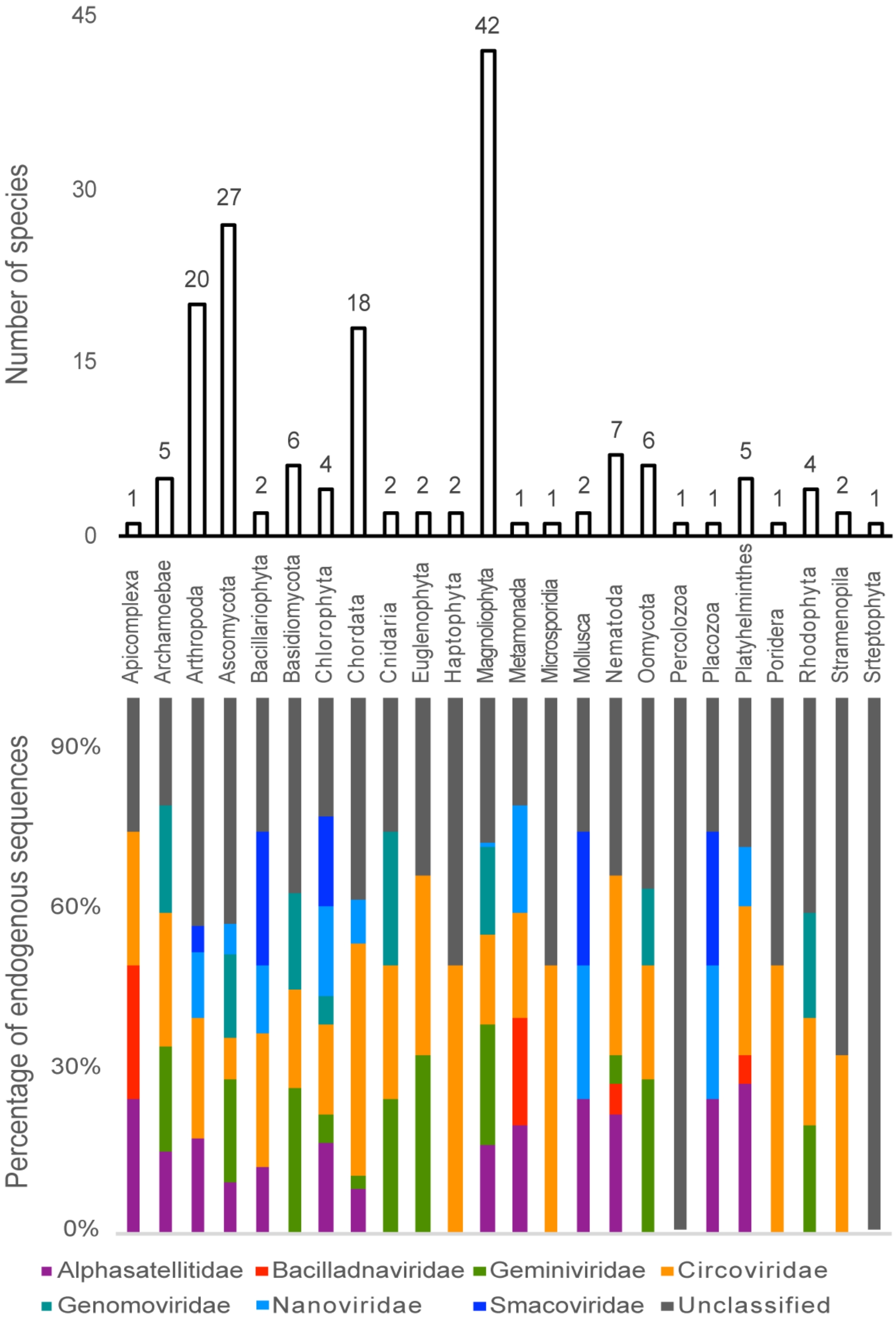


FIGURE 5.1 Distribution of endogenous Reps across eukaryotic phyla. The numbers represent number of species containing endogenous viral sequences. The percentages represent relative ratio of identified endogenous sequence by each viral family.

Maximum likelihood trees

Seven maximum likelihood trees were built with PhyML3 using the CRESS+G+F model, one for each family of Reps used to query the database. The geminivirus and similar endogenous sequences ML tree is shown in FIGURE 5.2. While the majority of eukaryotic genomes identified were plant species, there are some very surprising results showing geminivirus-like sequences integrated into protists, animals, oomycetes and fungi. Several sequences from Brown spotted pitviper (*Protobothrops mucrosquamatus*) grouped within the Reps from *Mastrevirus*, *Becurtovirus*, *Grablovirus*, *Capulavirus*, two unclassified intron-containing Reps and several other unclassified geminivirus (wild vitis virus, Baminivirus, etc.) with strong 0.983 aLRT SH-like support. The unclassified species Niminivirus is in a branch containing species from the fungal groups Ascomycota and Basidiomycota (0.916 sh-like support). And a sequence from *Eragrovirus* is closest to sequences from Nematoda and Basidiomycota (albeit weakly, SH-like support 0.539). One large clade (SH-like support 0.975) containing all Reps from *Begomovirus*, *Curtovirus*, *Topocurtovirus*, and *Turncurtovirus* also includes a mix of endogenous Rep sequences from all three groups of crown eukaryotes, specifically phyla Magnoliophyta, Cnidaria, and Ascomycota. Specifically, these species are the common sunflower (*Helianthus annuus*), narrow-leafed ash (*Fraxinus angustifolia*), billy goat weed (*Ageratum conyzoides*), the mitochondrion of *Amborella trichopoda* (understory shrubs), wild olive (*Olea europaea* var. *sylvestris*), acroporid coral (*Acropora digitifera*), and the beetle fungus (*Metarhizium majus* ARSEF297). The sister group to this well-supported clade is composed entirely of *Nicotiana* species (tobacco). And the sister group to these two clades exclusively contains more endogenized sequences, mostly *Dioscorea* species (yam) and two plant fungal pathogens. These two clades of endogenized sequences and the clade dominated by extant begomovirus sequences formed a well-supported group (SH-like 0.982). Deeper in the tree, the intron-containing and non-intron-containing geminivirus Rep sequences were reciprocally monophyletic, as have seen in the analysis of extant Rep sequences from the previous chapter. There is a large, weakly supported clade (SH-like support 0.742) that contains only endogenous sequences, dominated by species of Entamoeba, plasmids of red algae and the mitochondria of oomycetes (*Phytophthora nicotianae, Phytophthora infestans, Peronospora tabacina* (KT893455-56) and *Phytophthora sojae* (DQ832717)), though there are some sequences from plant species as well (e.g., banana, quinoa). Geminivirus-like Rep sequences were also found in the chloroplast genomes of two *Euglena garcilis* species (X70810, KP686076) and *Paradoxia multiseta* (KM462879).

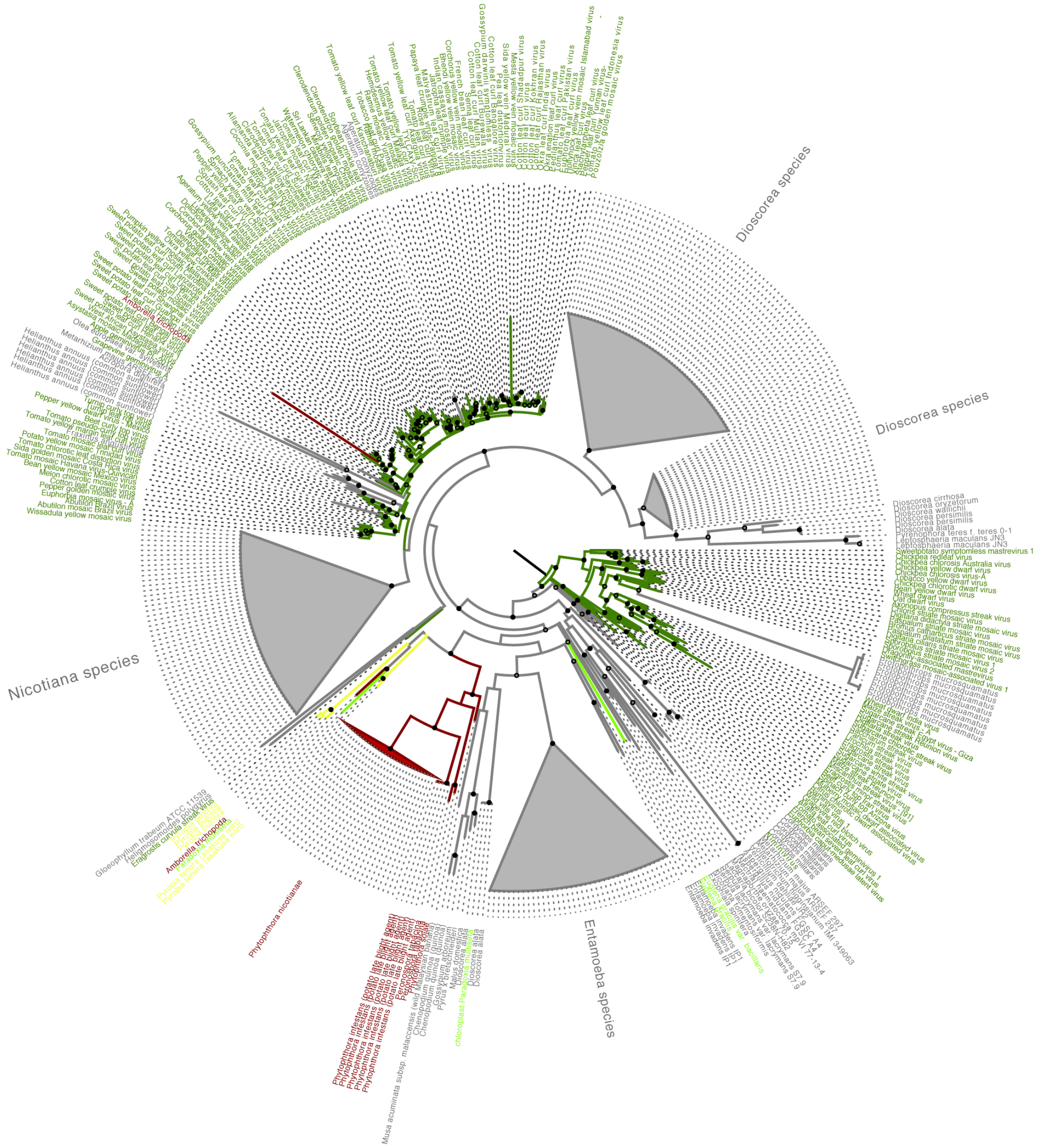

Dioscorea species
Dioscorea species
Nicotiana species
Entamoeba species
Dioscorea cirrhosa
Dioscorea oryzetorum
Dioscorea wallichii
Dioscorea persimilis
Dioscorea persimilis
Dioscorea alata
Pyrenophora teres f. teres 0-1
Leptosphaeria maculans JN3
Leptosphaeria maculans JN3
Sweetpotato symptomless mastrevirus 1
Chickpea redleaf virus
Chickpea chlorosis Australia virus
Chickpea yellow dwarf virus
Chickpea chlorosis virus-A
Tobacco yellow dwarf virus
Chickpea chlorotic dwarf virus
Bean yellow dwarf virus
Wheat dwarf virus
Oat dwarf virus
Axonopus compressus streak virus
Chloris striate mosaic virus
Digitaria didactyla striate mosaic virus
Paspalum striate mosaic virus
Bromus catharticus striate mosaic virus
Paspalum dilatatum striate mosaic virus
Digitaria ciliaris striate mosaic virus
Sporobolus striate mosaic virus 1
Sporobolus striate mosaic virus 2
Dragonfly-associated mastrevirus
Switchgrass mosaic-associated virus 1
Protobothrops mucrosquamatus
Protobothrops mucrosquamatus
Protobothrops mucrosquamatus
Protobothrops mucrosquamatus
Protobothrops mucrosquamatus
Protobothrops mucrosquamatus
Protobothrops mucrosquamatus
Wheat dwarf India virus
Maize streak virus - A
Sugarcane streak Egypt virus - Giza
Sugarcane streak Reunion virus
Digitaria streak virus
Sugarcane chlorotic streak virus
Saccharum streak virus
Panicum streak virus
Sugarcane white streak virus
Sugarcane striate virus
Streak Reunion virus
Panicum streak virus - [91]
Amborella trichopoda
Olea europaea var. sylvestris
Metarhizium majus ARSEF 297
Acropora digitifera
Helianthus annuus (common sunflower)
Helianthus annuus (common sunflower)
Helianthus annuus (common sunflower)
Helianthus annuus (common sunflower)
Helianthus annuus (common sunflower)
Apple geminivirus PL-2015
Grapevine geminivirus A
Asystasia mosaic Madagascar virus
West African Asystasia virus
Turnip curly top virus
Turnip leaf roll virus
Pepper yellow dwarf virus - Mexico
Beet curly top virus
Tomato pseudo-curly top virus
Tomato yellow margin leaf curl virus
Fraxinus angustifolia
Tomato mosaic leaf curl virus
Potato yellow mosaic Trinidad virus
Tomato chlorotic leaf distortion virus
Sida golden mosaic Costa Rica virus
Tomato mosaic Havana virus-Quivican
Bean yellow mosaic Mexico virus
Melon chlorotic mosaic virus
Cotton leaf crumple virus
Pepper golden mosaic virus
Euphorbia mosaic virus - A
Abutilon Brazil virus
Abutilon mosaic Brazil virus
Wissadula yellow mosaic virus
Gloeophyllum trabeum ATCC 11539
Heligmosomoides polygyrus
Eragrostis curvula streak virus
Pyropia pulchra
Pyropia pulchra
Pyropia pulchra
Amborella trichopoda
Paradoxia multiseta
Pyropia tenera (asakusa nori)
Pyropia tenera (asakusa nori)
Phytophthora nicotianae
Phytophthora infestans (potato late blight agent)
Phytophthora infestans (potato late blight agent)
Phytophthora infestans (potato late blight agent)
Phytophthora infestans (potato late blight agent)
Peronospora tabacina
Peronospora tabacina
Phytophthora sojae
Musa acuminata subsp. malaccensis (wild Malaysian banana)
Chenopodium quinoa (quinoa)
Chenopodium quinoa (quinoa)
Gossypium arboreum
Pyrus x bretschneideri
Malus domestica
Dioscorea alata
chloroplast Paradoxia multiseta
Dioscorea alata
Dioscorea alata

FIGURE 5.2 Geminivirus Reps and endogenous sequences midpoint-rooted maximum likelihood tree. Eukaryote sequences are colored in grey, geminivirus Reps are colored green, mitochondrial sequences are colored dark red, chloroplast sequences are colored bright green, plasmids from red algae are colored dark yellow. Open circles indicate SH-like support between 0.75 and 0.90, while filled circles indicate SH-like support ≥0.90.

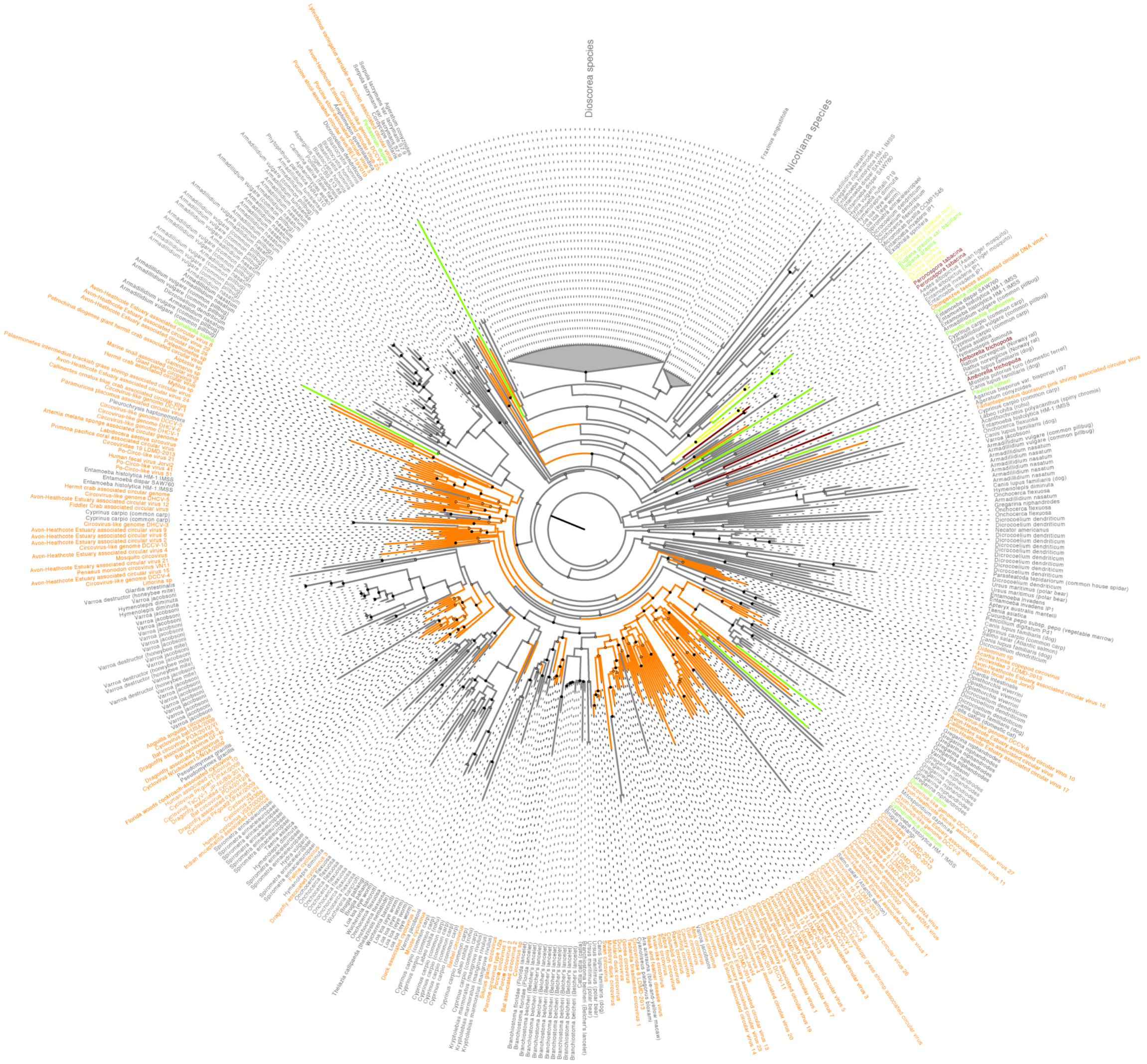


FIGURE 5.3 Circovirus Reps and endogenous sequences midpoint-rooted maximum likelihood tree. Eukaryote sequences are colored in grey, circovirus Reps are colored orange, mitochondrial sequences are colored dark red, chloroplast sequences are colored bright green, plasmids from red algae are colored dark yellow. Open circles indicate SH-like support between 0.75 and 0.90, while filled circles indicate SH-like support ≥0.90.

The circovirus Reps and similar endogenous sequences are shown in a tree in FIGURE 5.3. The intermingled complexity of the tree strongly supported the amount of endogenous element research done on this viral family. The tree can be roughly divided into two groups, one containing the majority of circovirus Reps with their closely related endogenous sequences, the other containing six unclassified circovirus sequences and their integrated counterparts (this clade supported by SH-like support 0.806). The well-mixed group of circovirus exogenous and endogenous rep sequence tree encompassed sequences from many eukaryotic groups, including Apicomplexa, Archamoebae, Ascomycota, Rhodaphyta, Basidiomycota, Stramenopila, Chlorophyta, Metamonada, Euglenophyta, Haptophyta, Chordata, Cnidaria, Arthropoda, Nematoda, Trematode, Cestoda, Porifera, Platyhelminthes, Microsporidia. All sequences officially assigned to the genus *Cyclovirus* (Rosario et al., 2017) are within one monophyletic branch with the SH-like support of 0.858, however, this clade included integrated sequences from an ant species (*Pseudomyrmex gracilis*). The official members of genus *Circovirus* were not similarly united on the tree, with classified members forming clades with both unclassified circoviruses and endogenized eukaryotic sequences. Some clades reinforced similar host ranges for extant circoviruses and their related integrated sequences, for instance the clade containing Barbel circovirus contains only sequences from fish genomes (SH-like support 0.95). However, this was not the only part of the tree that included fish – salmon appears twice in the tree, including once as the sole endogenous sequence in a different well-supported clade of exogenous, unclassified members of *Circoviridae*. Carp also makes appearances in other parts of the tree, including as the sole two endogenous sequences nested within unclassified circoviruses. Another clade with notably consistent host range is of the bird circoviruses, where the two endogenous sequences come from members of the parrot family (SH-like support 0.822). The distribution of terrestrial vertebrate and invertebrate species is less defined throughout the tree. There is a substantial overlap with endogenous sequences identified by querying the database with geminivirus sequences – the *Discorea* and *Nicotiana* clades likely due to endogenization of geminiviruses, not circoviruses, but this reflects the homology between the Reps of these two distantly related families (see FIGURE 4.5 for a tree of all CRESS DNA viruses of eukaryotes). Several endogenous sequences were found in organelles: in the chloroplast of Pediastrum *duplex* (KY114064), *Dunaliella salina* (GQ250046), *Euglena gracilis* (X70810, KP686076, same as in the geminivirus tree), *Pavlova lutheri* (KC573041), *Pseudo-nitzschia multiseries* (KR709240), and *Cylindrotheca closterium* (KC509522) and from the mitochondrion of *Amborella trichopoda* (KF754803, same as in the geminivirus tree), *Peronospora tabacina* (KT893455-56).

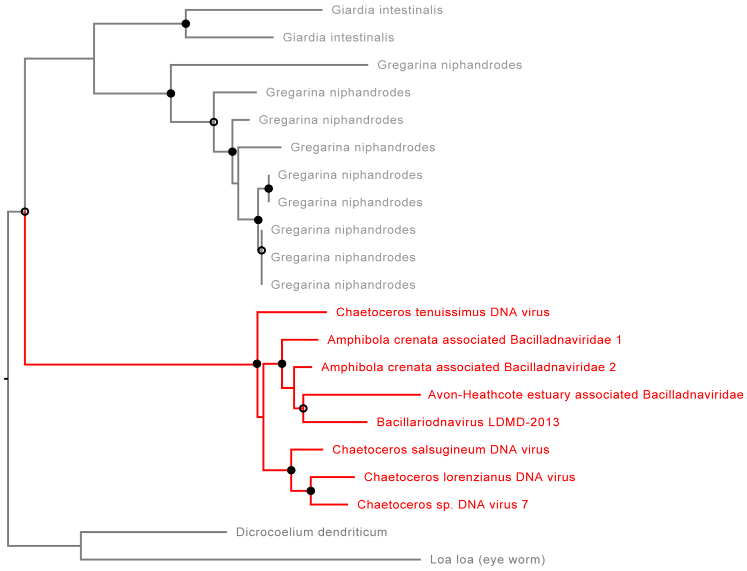

Giardia intestinalis
Giardia intestinalis
Gregarina niphandrodes
Gregarina niphandrodes
Gregarina niphandrodes
Gregarina niphandrodes
Gregarina niphandrodes
Gregarina niphandrodes
Gregarina niphandrodes
Gregarina niphandrodes
Gregarina niphandrodes
Chaetoceros tenuissimus DNA virus
Amphibola crenata associated Bacilladnaviridae 1
Amphibola crenata associated Bacilladnaviridae 2
Avon-Heathcote estuary associated Bacilladnaviridae
Bacillariodnavirus LDMD-2013
Chaetoceros salsugineum DNA virus
Chaetoceros lorenzianus DNA virus
Chaetoceros sp. DNA virus 7
Dicrocoelium dendriticum
Loa loa (eye worm)

FIGURE 5.4 Bacilladnavirus Reps and endogenous sequences midpoint-rooted maximum likelihood tree. Eukaryote sequences are colored in grey, bacilladnavirus Reps are red. Open circles indicate SH-like support between 0.75 and 0.90, while filled circles indicate SH-like support ≥0.90.

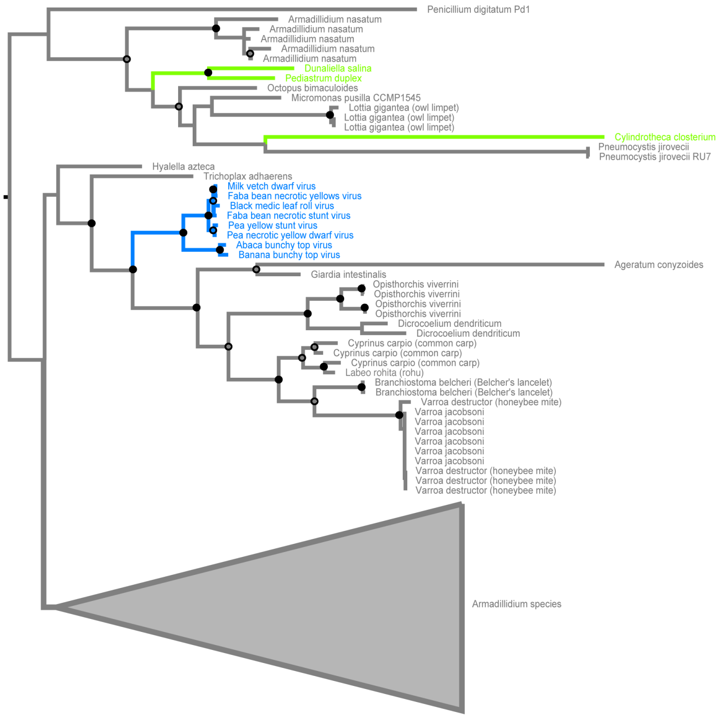


FIGURE 5.5 Nanovirus Reps and endogenous sequences midpoint-rooted maximum likelihood tree. Eukaryote sequences are colored in grey, nanovirus Reps are light blue, chloroplast sequences are colored bright green. Open circles indicate SH-like support between 0.75 and 0.90, while filled circles indicate SH-like support ≥0.90.

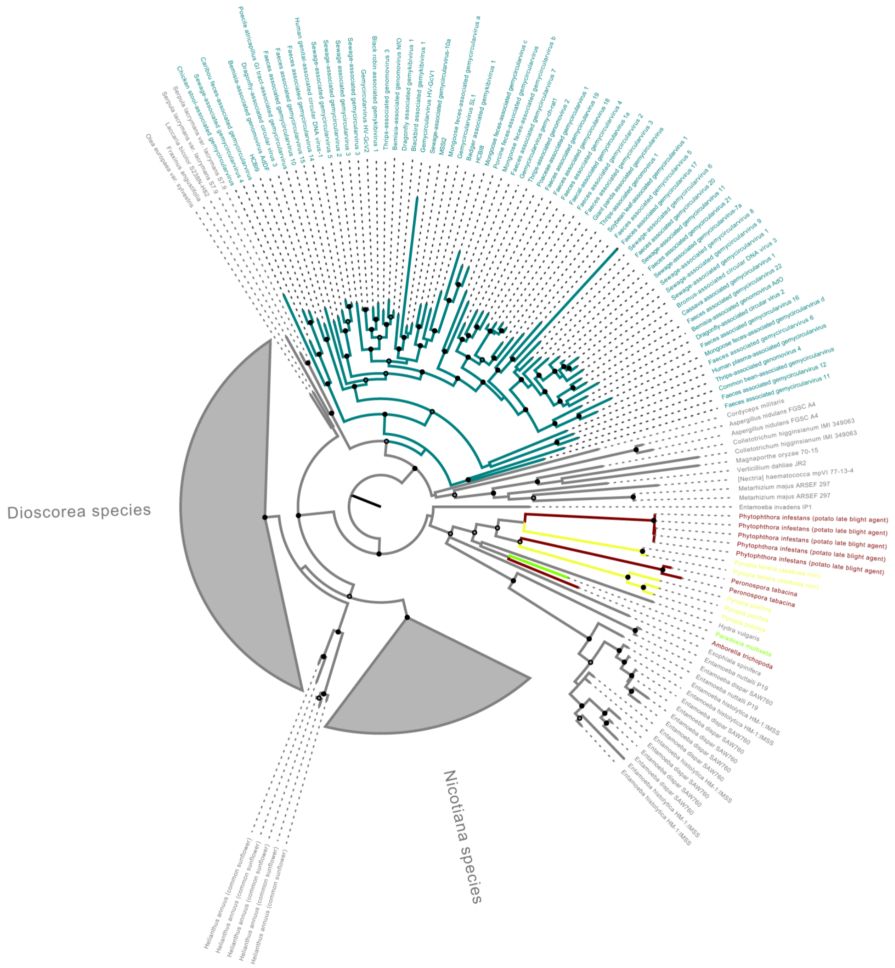

Olea europaea var. sylvestris
Fraxinus angustifolia
Laccaria bicolor S238N-H82
Serpula lacrymans var. lacrymans S7.9
Serpula lacrymans var. lacrymans S7.9
Chicken stool-associated gemycircularvirus
Sewage-associated gemycircularvirus 4
Caribou feces-associated gemycircularvirus
HCBI9
Bemisia-associated genomovirus AdDF
Dragonfly-associated circular virus 3
Poecile atricapillus GI tract-associated gemycircularvirus
Faeces associated gemycircularvirus 10
Faeces associated gemycircularvirus 15
Faeces associated gemycircularvirus 14
Human genital-associated circular DNA virus-1
Sewage-associated gemycircularvirus 5
Sewage associated gemycircularvirus 2
Sewage associated gemycircularvirus 3
Sewage-associated gemycircularvirus 3
Gemycircularvirus HV-GcV2
Black robin associated gemykibivirus 1
Thrips-associated genomovirus 3
Bemisia-associated genomovirus NfO
Dragonfly associated gemykibivirus 1
Blackbird associated gemykibivirus 1
Gemycircularvirus HV-GcV1
Sewage-associated gemycircularvirus-10a
MSSI2
Mongoose feces-associated gemycircularvirus a
Gemycircularvirus SL1
Badger associated gemykibivirus 1
HCBI8
Mongoose feces-associated gemycircularvirus c
Porcine feces-associated gemycircularvirus
Mongoose feces-associated gemycircularvirus b
Faeces associated gemycircularvirus 7
Gemycircularvirus gemy-ch-rat1
Thrips-associated genomovirus 2
Poaceae-associated gemycircularvirus 1
Faeces associated gemycircularvirus 19
Faeces associated gemycircularvirus 18
Faecal-associated gemycircularvirus 4
Faeces associated gemycircularvirus 1a
Faeces associated gemycircularvirus 2
Faeces associated gemycircularvirus 3
Giant panda associated gemycircularvirus
Thrips-associated genomovirus 1
Soybean leaf-associated gemycircularvirus 1
Faeces associated gemycircularvirus 5
Faeces associated gemycircularvirus 17
Sewage-associated gemycircularvirus 6
Faeces associated gemycircularvirus 20
Sewage-associated gemycircularvirus 11
Faeces associated gemycircularvirus 21
Sewage-associated gemycircularvirus-7a
Sewage-associated gemycircularvirus 8
Sewage-associated gemycircularvirus 9
Sewage-associated gemycircularvirus 1
Sewage-associated circular DNA virus 3
Bromus-associated gemycircularvirus 1
Cassava associated gemycircularvirus 1
Faeces associated gemycircularvirus 22
Bemisia-associated genomovirus AdO
Dragonfly-associated circular virus 2
Faeces associated gemycircularvirus 16
Mongoose feces-associated gemycircularvirus d
Faeces associated gemycircularvirus 6
Human plasma-associated gemycircularvirus
Thrips-associated genomovirus 4
Common bean-associated gemycircularvirus
Faeces associated gemycircularvirus 12
Faeces associated gemycircularvirus 11
Cordyceps militaris
Aspergillus nidulans FGSC A4
Aspergillus nidulans FGSC A4
Colletotrichum higginsianum IMI 349063
Colletotrichum higginsianum IMI 349063
Magnaporthe oryzae 70-15
Verticillium dahliae JR2
[Nectria] haematococca mpVI 77-13-4
Metarhizium majus ARSEF 297
Metarhizium majus ARSEF 297
Entamoeba invadens IP1
Phytophthora infestans (potato late blight agent)
Phytophthora infestans (potato late blight agent)
Phytophthora infestans (potato late blight agent)
Phytophthora infestans (potato late blight agent)
Phytophthora infestans (potato late blight agent)
Pyropia tenera (asakusa nori)
Pyropia tenera (asakusa nori)
Peronospora tabacina
Peronospora tabacina
Pyropia pulchra
Pyropia pulchra
Pyropia pulchra
Hydra vulgaris
Paradoxia multiseta
Amborella trichopoda
Exophiala spinifera
Entamoeba nuttalli P19
Entamoeba dispar SAW760
Entamoeba nuttalli P19
Entamoeba histolytica HM-1:IMSS
Entamoeba histolytica HM-1:IMSS
Entamoeba dispar SAW760
Entamoeba dispar SAW760
Entamoeba dispar SAW760
Entamoeba dispar SAW760
Entamoeba dispar SAW760
Entamoeba histolytica HM-1:IMSS
Entamoeba dispar SAW760
Entamoeba dispar SAW760
Entamoeba dispar SAW760
Entamoeba histolytica HM-1:IMSS
Entamoeba histolytica HM-1:IMSS
Nicotiana species
Helianthus annuus (common sunflower)
Helianthus annuus (common sunflower)
Helianthus annuus (common sunflower)
Helianthus annuus (common sunflower)
Helianthus annuus (common sunflower)
Dioscorea species

FIGURE 5.6 Genomovirus Reps and endogenous sequences midpoint-rooted maximum likelihood tree. Eukaryote sequences are colored in grey, genomovirus Reps are teal, mitochondrial sequences are colored dark red, chloroplast sequences are colored bright green, plasmids from red algae are colored dark yellow. Open circles indicate SH-like support between 0.75 and 0.90, while filled circles indicate SH-like support ≥0.90.

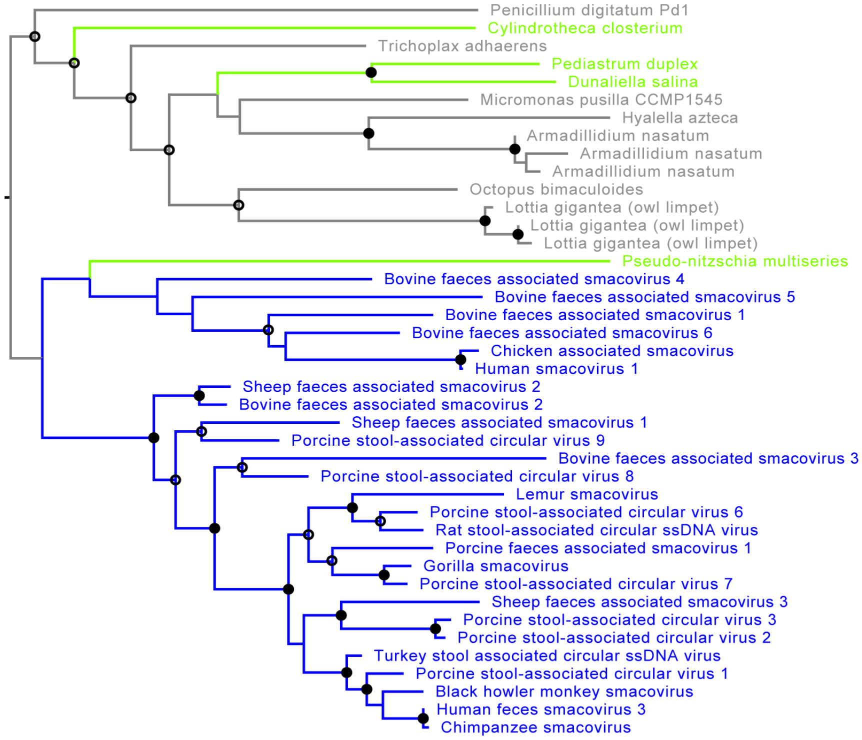

Penicillium digitatum Pd1
Cylindrotheca closterium
Trichoplax adhaerens
Pediastrum duplex
Dunaliella salina
Micromonas pusilla CCMP1545
Hyalella azteca
Armadillidium nasatum
Armadillidium nasatum
Armadillidium nasatum
Octopus bimaculoides
Lottia gigantea (owl limpet)
Lottia gigantea (owl limpet)
Lottia gigantea (owl limpet)
Pseudo-nitzschia multiseries
Bovine faeces associated smacovirus 4
Bovine faeces associated smacovirus 5
Bovine faeces associated smacovirus 1
Bovine faeces associated smacovirus 6
Chicken associated smacovirus
Human smacovirus 1
Sheep faeces associated smacovirus 2
Bovine faeces associated smacovirus 2
Sheep faeces associated smacovirus 1
Porcine stool-associated circular virus 9
Bovine faeces associated smacovirus 3
Porcine stool-associated circular virus 8
Lemur smacovirus
Porcine stool-associated circular virus 6
Rat stool-associated circular ssDNA virus
Porcine faeces associated smacovirus 1
Gorilla smacovirus
Porcine stool-associated circular virus 7
Sheep faeces associated smacovirus 3
Porcine stool-associated circular virus 3
Porcine stool-associated circular virus 2
Turkey stool associated circular ssDNA virus
Porcine stool-associated circular virus 1
Black howler monkey smacovirus
Human feces smacovirus 3
Chimpanzee smacovirus

FIGURE 5.7 Smacovirus Reps and endogenous sequences unrooted maximum likelihood tree. Eukaryote sequences are colored in grey, smacovirus Reps are blue, chloroplast sequences are colored bright green. Open circles indicate SH-like support between 0.75 and 0.90, while filled circles indicate SH-like support ≥0.90.

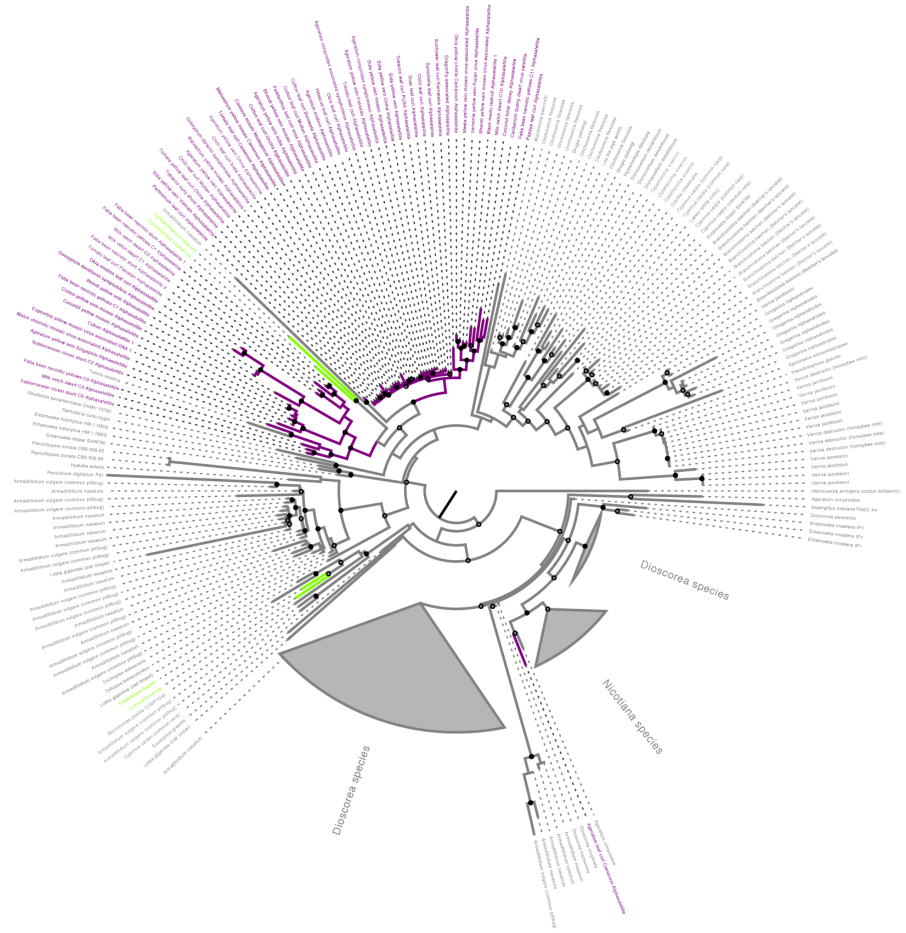

Dioscorea species
Nicotiana species
Dioscorea species
Faba bean necrotic yellows C7 Alphasatellite
Bhindi yellow vein Alphasatellite
Croton yellow vein mosaic Alphasatellite
Cucurbit yellow mosaic Alphasatellite
Cuban Alphasatellite 1
Euphorbia yellow mosaic virus associated DNA 1
Melon chlorotic mosaic virus-associated Alphasatellite
Ageratum yellow vein Singapore Alphasatellite
Subterranean clover stunt C2 Alphasatellite
Taenia asiatica
Faba bean necrotic yellows C9 Alphasatellite
Milk vetch dwarf C3 Alphasatellite
Subterranean clover stunt C6 Alphasatellite
Baudoinia panamericana UAMH 10762
Ramularia collo-cygni
Entamoeba histolytica HM-1:IMSS
Entamoeba histolytica HM-1:IMSS
Entamoeba dispar SAW760
Penicilliopsis zonata CBS 506.65
Penicilliopsis zonata CBS 506.65
Hyalella azteca
Penicillium digitatum Pd1
Armadillidium vulgare (common pillbug)
Armadillidium nasatum
Armadillidium vulgare (common pillbug)
Armadillidium vulgare (common pillbug)
Armadillidium nasatum
Armadillidium nasatum
Armadillidium nasatum
Armadillidium nasatum
Armadillidium vulgare (common pillbug)
Armadillidium vulgare (common pillbug)
Lottia gigantea (owl limpet)
Armadillidium nasatum
Armadillidium nasatum
Armadillidium vulgare (common pillbug)
Armadillidium vulgare (common pillbug)
Armadillidium vulgare (common pillbug)
Armadillidium nasatum
Armadillidium vulgare (common pillbug)
Armadillidium nasatum
Armadillidium vulgare (common pillbug)
Armadillidium vulgare (common pillbug)
Armadillidium nasatum
Branchiostoma belcheri (Belcher's lancelet)
Varroa jacobsoni
Gregarina niphandrodes
Gregarina niphandrodes
Gregarina niphandrodes
Gregarina niphandrodes
Gregarina niphandrodes
Gregarina niphandrodes
Gregarina niphandrodes
Spirometra erinaceieuropaei
Pseudomyrmex gracilis
Varroa destructor (honeybee mite)
Varroa jacobsoni
Varroa jacobsoni
Varroa jacobsoni
Varroa jacobsoni
Varroa jacobsoni
Varroa jacobsoni
Varroa destructor (honeybee mite)
Varroa destructor (honeybee mite)
Varroa destructor (honeybee mite)
Varroa jacobsoni
Varroa jacobsoni
Varroa jacobsoni
Varroa jacobsoni
Helicoverpa armigera (cotton bollworm)
Ageratum conyzoides
Aspergillus nidulans FGSC A4
Dioscorea persimilis
Entamoeba invadens IP1
Entamoeba invadens IP1
Entamoeba invadens IP1

FIGURE 5.8 Alphasatellites Reps and endogenous sequences midpoint-rooted maximum likelihood tree. Eukaryote sequences are colored in grey, alphasatellite Reps are purple, chloroplast sequences are colored bright green. Open circles indicate SH-like support between 0.75 and 0.90, while filled circles indicate SH-like support ≥0.90.

The *Geminiviridae* and *Circoviridae* are the oldest, most established groups of CRESS DNA viruses. The maximum likelihood trees of Reps from exogenous and endogenous sequences similar to other families had more straightforward branching patterns, with less intermingling. A tightly clustered clade of bacilladnavirus Reps is well-supported (SH-like support 0.843), and nested within diverged eukaryotic sequences from protists (phyla Metamonada, Apicomplexa) to worms (phyla Nematoda and Platyhelminthes) (FIGURE 5.4). These endogenized sequences are from hosts very distantly related to the heterokont diatoms that bacilladnaviruses are known to infect.

In FIGURE 5.5, the small number of nanovirus Reps formed their own clade (SH-like support 0.965). This was nested within a supported clade that contains animals (including fish, arthropod mites and the smallest marine animal, *Trichoplax adhaerens*), a sequence from the parasitic microorganism (*Giardia intestinalis*) and a plant (*Ageratum conyzoides*). More distantly related to nanoviruses are a two species from an arthropod genus (*Armadilidum nasatum* and *Aramdilidum vulgare*), fungi, an alga, the California two-spot octopus and another fish species. Endogenous sequences from the chloroplast of *Pediastrum duplex* (KY114064), *Dunaliella salina* (GQ250046), and *Cylindrotheca closterium* (KC509522), which were also BLAST hits to circovirus Rep queries. Although currently identified nanoviruses infect plants, we did not find nanoviruses endogenized in the phylum Magnoliaphyta like the other plant infecting family *Geminiviridae*, but only in one Streptophyta species (*Ageratum conyzoides*), one Chlorophyta species and inside chloroplasts derived from Chlorophyta.

Genomovirus Reps formed a weakly supported clade (SH-like support 0.539) in the viral and eukaryotic sequence tree in FIGURE 5.6. Together with a clade of sequences from Basidiomycota and a clade from Ascomycota, this larger clade containing the exogenous genomoviruses is well-supported (SH-like support 0.983). The Basidiomycota clade contains sequences from the mushroom bicoloured deceiver (*Laccaria bicolor*) and dry rot fungus

(*Serpula lacrymans* var. lacrymans S7.9). The Ascomycota clade contains sequences from fungal plant pathogens (*Verticillium dahliae* JR2, *Nectria haematococca* mpVI 77-13-4, *Colletotrichum higginsianum* IMI 349063, *Pyricularia oryzae* (syn. *Magnaporthe oryzae*) 70-15), a fungal beetle pathogen (*Metarhizium majus* ARSEF 297), a filamentous fungus used as a model organism (*Aspergillus nidulans* FGSC A4), and an insect pathogenic fungus (*Cordyceps militaris*). The only cultivated genomovirus infects a fungus, and this tree provides support for the current and extinct genomovirus host range being restricted to fungi. More distantly genomovirus Reps are related to sequences that appeared in the geminivirus tree: the endogenized sequences in *Dioscorea*, *Nicotiana*, sunflower, and those in the clade formed by solely endogenous geminivirus Rep-like sequences, including a cnidarian, *Entamoeba* species and in the plasmid of red algae. Similarly, the organelle-integrated elements found with genomovirus Rep queries are also found in the geminivirus tree: the chloroplast of *Paradoxia multiseta* (KM462879), the mitochondria of *Amborella trichopoda* (KF754803), *Peronospora tabacina* (KT893455-56) and *Phytophthora infestans*.

Smacovirus Reps did not produce many hits, however, many of the identified endogenous sequences were also found when querying with Reps from other families. Sequences from the chloroplast of a diatom (*Pseudo-nitzschia multiseries*), which had been identified in the circovirus search, were most closely related to smacovirus Reps (FIGURE 5.7). Forming a separate clade from the smacovirus Reps are almost all of the species in the nanovirus tree that were more distantly related to the nanovirus Reps (SH-like support 0.852). Only two taxa appear in the nanovirus tree outside of the clade containing the nanoviruses and do not appear in the smacovirus tree: *Pneumocystis jirovecii* and *Armadillidium nasatum*. Therefore, the smacovirus-like endogenous sequences originated from a basal metazoan, an arthropod, a fungus, an alga, the California two-spot octopus and a fish. However, it should be noted that the smacovirus search yielded fewer BLAST hits even within these taxa than the nanovirus search.

The alphasatellite Reps, which group with nanovirus Reps in phylogenies of CRESS DNA viruses (see FIGURE 4.5) found sequences similar to BLAST hits to both the nanoviruses and geminiviruses with which they associate (FIGURE 5.8). All but one of the alphasatellites formed two distinct clades with strong support – one exclusively containing geminivirus-associated alphasatellites and one endogenous sequence from *Armadillidium nasatum* (SH-like support 0.902) and the other containing a mix of nanovirus- and geminivirus-associated alphasatellites containing a single endogenous sequence (from tapeworm, *Taenia asiatica*, SH-like support 0.921). Between these two clades, there are two sequences very closely related to these exogenous alphasatellites from chloroplast of a diatom (*Cyinrotheca closterium*). More distantly related to these alphasatellite clades are sequences from taxa that grouped with nanoviruses (FIGURE 5.5): *Opisthorchis viverrini*, *Dicrocoelium dendriticum, Branchiostoma belcheri, Varroa* species and two species of fish. However, the endogenous sequences in a supported clade with the bulk of the alphasatellites (SH-like support 0.807) also include taxa less closely related to nanoviruses (*Trichoplax adhaerans*, *Hyalella azteca*, *Armadillidium* species, *Penicillum digitatum*), those identified by querying with bacilladnavirus Rep sequences (*Loa loa, Gregarina niphandrodes*), those found in both trees (*Giardia intestinalis*), and unique fungal species (e.g., *Ramularia collo-cygni*), nematodes and platyhelminths not found in other searches.

The remaining taxa in the tree formed a supported clade, and included only one alphasatellite sequence. Ageratum leaf curl Cameroon alphasatellite grouped with endogenous sequences from *Nicotiana* and other plant species. The close relationship of this alphasatellite to geminivirus Reps is well-known, as it groups with *Geminiviridae* in the previous chapter, so it was not surprising that it was found near endogenous sequences from the geminivirus Rep tree (FIGURE 5.2). The remaining endogenous taxa in the larger clade included those found in the geminivirus and genomovirus trees (*Entamoeba* species, *Dioscorea* species) and nanovirus (and smacovirus) trees (*Armadillidium* species, *Lottia gigantea*, *Micromonas pusilla*, *Octopus bimaculoides, Penicillum* *digitatum*), one found in both geminivirus and nanovirus trees (*Ageratum conyzoides*) and one taxon unique to this clade (*Eucalyptus grandis*). Endogenous sequences were also found in the chloroplasts of *Pediastrum duplex* (KY114064) and *Dunaliella salina* (GQ250046) – along with *Cyinrotheca closterium* from the other part of the tree, these three species' organelle genomes were also identified in the circovirus and nanovirus queries.

Discussion

Relaxed search criteria

Despite the proliferation of papers discussing endogenized CRESS DNA viral sequences, we were able to find more endogenous circovirus Rep sequences and fragments than any previous study (Liu et al., 2011; Metegnier et al., 2015; Theze et al., 2014). This is because we used relaxed search criteria, allowing for distantly related sequences to be included, which also accounts for the rapid evolution of proteins in ssDNA viruses (Duffy and Holmes, 2008; Firth et al., 2009) and diversification over the potential millions of years since integration (Gibbs et al., 2006; Lefeuvre et al., 2011).

Many of the sequences identified here have been previously observed to be related to CRESS DNA viruses of eukaryotes. Rep-like sequences in the mitochondrial sequences of oomycetes (*Phytophthora sojae* and *Phytophthora infestans*) (Liu et al., 2011)*, Giardia intestinalis* and *Entamoeba histolytica* HM-1:IMSS (Gibbs et al., 2006) were found by multiple different viral family searches in our study. Similarly, multiple viral searches found the endogenous elements of yam (*Dioscorea*) and tobacco (*Nicotiana*) species that have been well-described (Filloux et al., 2015; Gibbs et al., 2006; Lefeuvre et al., 2011). Some overlap between the different family trees was expected, especially with our relaxed search criteria. The *Genomoviridae* were named the Gemini-like No Movement protein viruses, and they are more similar to geminiviruses than other CRESS DNA viral families. Therefore, it is understandable that the distantly related taxa in the genomovirus tree are all from the geminivirus tree, but it is clear that certain fungal taxa are truly more closely related to genomoviruses. Our relaxed search criteria mean that each query pulls in more distantly related Rep-like sequences, making representation in multiple trees understandable. In addition to potential confusing about which exogenous family is most closely related to some shared endogenous sequences, there is also the potential for false positive results. While we have taken a census of the BLAST hits, we have conservatively discussed our results that are based on strongly supported relationships with exogenous Rep sequences.

Geminivirus tree

Upon further examination of our results, we found that the reason Brown spotted pitviper (*Protobothrops mucrosquamatus*) predicted mRNA sequences ATPase plasma membrane Ca2+ transporting 3 was identified in our search was because the Rep of Paspalum striate mosaic virus (YP_006659974) blasted to it with a high e-value of 0.0005. Although multiple geminivirus accessions form the clade with the pitviper, the result should still be viewed with some skepticism – and it is hard to imagine how a plant-infecting geminivirus would find its way into the pitviper germline. As relatives of the Brown spotted pitviper have their genomes sequenced, we may find additional evidence for the Rep-like hit, reflecting an ancient endogenization event. However, we are more confident in our identified new endogenous elements from additional plant species. Apart from the mentioned *Nicotiana* and *Dioscorea* spp., previous searches with geminiviruses only found endogenous sequences inside black cottonwood (*Populus trichocarpa*, (Liu et al., 2011)), but our results widened the group that contained endogenized Rep sequences to other plants: common sunflower (*Helianthus annuus*), narrow-leafed ash (*Fraxinus angustifolia*), billy goat weed (*Ageratum conyzoides*), the mitochondrion of *Amborella trichopoda* (understory shrubs), wild olive (*Olea europaea* var. *sylvestris*), and acroporid coral (*Acropora digitifera*). Our analyses do not permit the resolution of which exogenous species is most closely related to most endogenized sequences, but the better-resolved integration into the mitochondrion of *Amborella trichopoda* is clearly related to sweepoviruses – sweet potato-infecting begomoviruses. Even though this mitochondrial sequence was detected by searches with other CRESS DNA virus family queries, no other phylogenies show the close relationship seen in the geminivirus tree; we have reason to think that this Rep-like sequence originated from a sweet potato begomovirus-like virus. Both the sequence similarity of the mitochondrial Rep-like sequence, as evidenced by the high support for the mixed clade, and the geography of the species -- *Amborella* are endemic to New Caledonia and sweet potatoes originate in the India (Poncet et al., 2012; Srivastava et al., 2018) – support the endogenization of the viral Rep sequence. Most of the other sequences that BLASTed in our relaxed search are more distantly related to extant viral sequences. Significantly, there was a clade of entirely endogenous sequences that appeared in the largest form in the geminivirus tree but appeared in other viral family trees as well (including Entamoebae species, mitochondria of oomycetes and others). Sequences in this clade show homology to CRESS DNA viral Reps, but it is unclear which family might be most closely related (Liu et al., 2011).

Circovirus tree

Endogenous sequences similar to circovirus Rep proteins were found in 89 eukaryotic species from 20 phyla, encompassing a huge diversity of eukaryotes. The endogenous sequences identified in dog, cat, bird, fish, rat, ant, spider, mite, worms and plants only partially agreed with circovirus extant host range (mammals, birds, fish and likely invertebrates, Chapter 3). As endogenous sequences could serve as evidence of past or current host use, we can deduce likely hosts for uncharacterized circoviruses (Aiewsakun and Katzourakis, 2015). All official members from the genus Cyclovirus shared a monophyletic clade with two ant sequences, suggesting a constrained host use for this genus, and validating the invertebrate host range inferred for many of its species (Rosario et al., 2012a; Rosario et al., 2018). Official members of the genus circovirus, however, intermingled with a more diverse selection of metazoan sequences, though many endogenized animal sequences grouped near their exogenous virus of similar species (e.g. fish sequences by Barbel circovirus). The unclassified circovirus species had sister groups from a much wider range than either of the two genera, suggesting a wider host range than just metazoans, commensurate with the high sequence diversity among unclassified circoviruses (Dayaram et al., 2014; Li et al., 2010; Steel et al., 2016). One of the potential hosts for unclassified circoviruses comes from one of the endogenous Salmon sequences, it suggests a salmoid host for circoviruses 14, 16 and the first sequence isolated from the New Zealand Avon-Heathcote estuary, FIGURE 5.3. A cocolithophore species (*Pleurochyrsis haptonemofera*) suggests phytoplankton may be the hosts for several sister taxa (circovirus-like genome DHCV-1, -6, DCCV-1, -5, and several species associated with corals). Nonetheless, some of the close associations of endogenized sequences with exogenous viruses were surprising, for instance, a sequence from a honeybee mite genome is closely related to mammalian circoviruses (mink, dog). The most puzzling association might be how circovirus-like sequences, ones closely related to currently circulating genomes, have related sequences in chloroplast genomes. While the chloroplasts of *Pediastrum duplex*, *Dunaliella salina* and *Cyinrotheca closterium* appeared in three family trees, they are each most closely grouped with exogenous viruses in the circovirus tree. This suggests that the circulating circoviruses may be most representative of the CRESS DNA viruses that integrated into these chloroplasts, especially for *P. duplex* and *D. salina*, since *C. closterium* was also found to be closely related to alphasatellite sequences (FIGURE 5.8).

Bacilladnavirus tree

Bacilladnavirus Rep sequences are very distinct from their endogenous BLAST hits. This could be due, in part, to the small number of representative Reps from this newly codified family (Kazlauskas et al., 2017), compared to the diversity within the geminivirus and circovirus query sets. Alternatively, this could reflect the true relationship between the bacilladnavirus Reps and endogenous sequences, and that no close relatives to bacilladnaviruses have integrated into eukaryotes, or that such events happened so long ago as to erase any close sequence relationship.

The endogenous sequences in this tree appear in other CRESS DNA viral family trees (notably the alphasatellite tree), and the distance between them and the extant bacilladnavirus sequences suggest bacilladnaviruses are unlikely to be the true sister taxa to the endogenized sequences.

Genomovirus tree

This was previously observed that all fungal virus-like sequences cluster closely with *Sclerotinia sclerotiorum* hypovirulence associated DNA virus 1 (SsHADV-1) and Rep sequences isolated from viral metagenomics samples (Liu et al., 2011). Being the first and only characterized species of the *Genomoviridae* family, SsHADV-1 was isolated from the plant fungal pathogen *Sclerotinia sclerotiorum* (Yu et al., 2010). The endogenous sequences closely related to genomovirus Reps are from two fungi phyla: Ascomycota and Basidiomycota. Sequences from the mushroom bicoloured deceiver (*Laccaria bicolor* S238N-H82), plant fungal pathogens (*Nectria haematococca* mpVI 77-13-4, *Magnaporthe oryzae* 70-15 (rice blast fungus)) and model organism (*Aspergillus nidulans* FGSC A4) that clustered closely to SsHADV-1 in Liu et al.'s work were also detected in this study. We have expanded the number of fungi with genomovirus Rep-like endogenized sequences to include dry rot fugus (*Serpula lacrymans* var. lacrymans S7.9), fungal plant pathogen (*Verticillium dahliae* JR2, *Colletotrichum higginsianum* IMI 349063,), fungal beetle pathogen (*Metarhizium majus* ARSEF 297), and herbal fungus (*Cordyceps militaris*). Almost all the fungal genomic sequences found by genomovirus Reps belong to plant fungal pathogens, just like the host of SsHADV-1. Recently, virus-like particles with genome sequences most similar to genomoviruses has been found in fungus-farming termites (Kerr et al., 2018). All of these findings suggest a fungal host range – possibly exclusively so – for all members of *Genomoviridae*. In addition, it has been proposed that SsHADV-1 could be a biological fungal pathogen control agent (Yu et al., 2010). Our findings would undermine that application, since they indicate genomoviruses likely have infected fungi for a long period of time, and the integrated genomovirus-like sequences inside fungal hosts suggest some fungi may be able to resistant viral infection through the production of small antiviral RNAs (Campo et al., 2016).

Smacovirus tree

Smacovirus is a recently proposed family, primarily found in fecal associated samples. None of these viruses have been studied as isolated virions or characterized in ways other than sequencing. This is the only CRESS DNA viral family with no confirmed hosts, so our mixed viral and endogenous sequence tree is potentially very important for our understanding of this family. Chloroplast sequences found in a diatom are very closely related to smacovirus Reps, suggesting that diatoms might be a potential past or current host for some of these viruses. The only identified CRESS DNA viruses to infect diatoms are bacilladnaviruses, which we have recently shown are more closely related to smacoviruses than previously thought (Chapter 4, FIGURE 4.5), so this is an independent piece of evidence suggesting that closer relationship may be accurate. What is unclear, however, is why smacoviruses, which have been so widely associated with vertebrate and invertebrate samples (Varsani and Krupovic, 2018), might have an endogenous fossil in diatoms.

Nanovirus and alphasatellite trees

With a relative diverse set of sequences, the alphasatellite queries were able to search through a large amount of the eukaryotic genome sequence space close to Rep proteins, because these Reps still found the *Nicotiana* and *Dioscorea* endogenous elements that are unambiguously due to geminivirus integration events. Geminivirus Reps are not closely related to alphasatellite Reps (see Chapter 4), so it was surprising that even under a relaxed search alphasatellites were able to find the same sequences. However, nanoviruses had far fewer search results, even though they are very closely related to alphasatellites phylogenetically (Chapter 4). This is likely due to the small number of queries used in the nanovirus endogenous search compared to the alphasatellite queried search – very few nanovirus species have been identified and sequenced. Both the nanovirus and alphasatellite trees shared many of the same eukaryotic sequences, evincing their close ancestry. These included terrestrial and marine animals, which was unexpected from exclusively plant-associated viruses and satellites (Briddon et al., 2018). This was exacerbated in the alphasatellite tree as the only taxon an otherwise exclusively satellite clade was from a tapeworm, which we would not think would be in close contact with alphasatellites.

Endogenous elements in organelles

There are ten mitochondria and ten chloroplasts from eukaryotes that appear in these seven trees, often in more than one tree. The mechanism of how eukaryotic CRESS DNA viruses would integrate into these erstwhile prokaryotes is an open area of inquiry, though it is thought that some replication genes in mitochondria trace back to T-odd-like phages (Shutt and Gray, 2006). Other researchers that found CRESS DNA virus-like sequences in mitochondria did not attempt a mechanistic explanation (Liu et al., 2011), but there are some theoretical ways Rep proteins might be useful in an organelle. Mitochondrial genomes and their plasmids have been detected in single-stranded states, undergoing rolling-circle replication in higher plants (Backert et al., 1996). This suggests the requirement of a protein with a similar function as Rep, which would not have been ancestral to mitochondrial genomes. Similarly, strand displacement mode of mtDNA synthesis was also suggested as a mode of human mitochondrial DNA replication (Miralles Fusté et al., 2014). That these Rep-like sequences would be functional in the organelles is not surprising, since they are so genome-reduced that it is hard to imagine useless integrated sequence lasting over long evolutionary times (Smith and Keeling, 2015). Although no direct evidence of a Rep from any CRESS DNA virus has been shown to be harnessed by mitochondria, we identified several potential candidates for such investigations. The mitochondria of *Amborella trichopoda*, *Peronospora tabacina*, *Phytophthora sojae*, *Phytophthora nicotianae* and *Phytophthora infestans* and the chloroplasts of *Pediastrum duplex*, *Dunaliella salina*, *Euglena* *gracilis*, *Pavlova lutheri*, *Pseudo-nitzschia multiseries*, and *Cylindrotheca closterium* might hold intriguing evidence showing tangents of the evolution of the CRESS DNA viral Rep protein.

The four species of oomycete with Rep-like sequences are closely related to each other, in a supported clade in the geminivirus tree, and three are closely related members of the same *Phytophthora* genus (Yuan et al., 2017)). Yet these similar sequences likely do not reflect a single integration event. In *P. infestans*, it is clear the Rep-like protein is encoded in an insertion present in only some isolates (typically type II, which have ~2kb larger genomes (Yang et al., 2013)), but the other species do not share the same insertion, nor organization of the insertion's hypothetical protein ORFs. The weak phylogenetic relationship between CRESS DNA viruses and the endogenous sequences in the oomycetes (and the sequences grouping with them) could indicate a very ancient endogenization, where divergence and extinction of exogenous viruses could obscure the signal, or it could indicate that these Rep-like proteins are more similar to replication-associated proteins present in ssDNA plasmids. The Reps of eukaryotic CRESS DNA viruses form a clade with the Reps of ssDNA plasmids, including those of red algae, and it has been proposed that the viruses evolved from the plasmid (Krupovic et al., 2009; Saccardo et al., 2011). When a Rep-like sequence was first observed in *P. infestans*, it grouped with a Rep from ssDNA algal plasmids (Liu et al., 2011), and the plasmids of *Pyropia pulchra* are in the same supported clade with the *Phytophthora* mitochondrial sequences. These eukaryotic sequences set apart from the more geminivirus-like clades could well represent Rep elements from plasmids, not viruses. While plasmids are more expected than eukaryotic viruses inside organelles, this cannot explain the Rep-like sequences that are very closely related to exogenous viruses, such as one of the integration events into the *Amborella trichopoda* mitochondrion or the *Cylindrotheca closterium* chloroplast.

The preponderance of CRESS DNA viruses known by sequence alone has stymied our understanding of the ecological impact of this group. Although all genomic sequences inferred from metagenomes that we used in this study were verified with PCR and Sanger sequencing to obtain the complete genome sequences, we lack the biological isolates to conduct any host range screening for nearly all of these species (Male et al., 2016; Rosario et al., 2015; Steel et al., 2016). We hope the host taxa showing evidence of a historical host range for these viruses will help researchers target the hosts of a given CRESS DNA virus family, and increase the odds of isolating viral particles for in depth characterization. We find evidence to support the host range assertions of several CRESS DNA virus groups, including arthropod-infecting circoviruses (Dayaram et al., 2014; Dayaram et al., 2013; Rosario et al., 2012a; Rosario et al., 2011; Rosario et al., 2018), the fungal host range of genomoviruses (Yu et al., 2010), and the closely related species that likely expand the host range of vertebrate-infecting circoviruses (Dennis et al., 2018a; Dennis et al., 2018b).

Acknowledgements

This work was supported by the National Science Foundation’s Assembling the Tree of Life program, DEB1240049.

References

Aiewsakun, P., Katzourakis, A., 2015. Endogenous viruses: Connecting recent and ancient viral evolution. Virology 479-480, 26-37.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. Journal of molecular biology 215, 403-410.

Ashby, M.K., Warry, A., Bejarano*, E.R., Khashoggi, A., Burrell, M., Lichtenstein*, C.P., 1997. Analysis of multiple copies of geminiviral DNA in the genome of four closely related Nicotiana species suggest a unique integration event. Plant Molecular Biology 35, 313-321.

Backert, S., Dörfel, P., Lurz, R., Börner, T., 1996. Rolling-circle replication of mitochondrial DNA in the higher plant Chenopodium album (L.). Molecular and Cellular Biology 16, 6285-6294.

Bejarano, E.R., Khashoggi, A., Witty, M., Lichtenstein, C., 1996. Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. Proceedings of the National Academy of Sciences of the United States of America 93, 759-764.

Belyi, V.A., Levine, A.J., Skalka, A.M., 2010. Sequences from ancestral single-stranded DNA viruses in vertebrate genomes: the parvoviridae and circoviridae are more than 40 to 50 million years old. J Virol 84, 12458-12462.

Briddon, R.W., Martin, D.P., Roumagnac, P., Navas-Castillo, J., Fiallo-Olive, E., Moriones, E., Lett, J.M., Zerbini, F.M., Varsani, A., 2018. Alphasatellitidae: a new family with two subfamilies for the classification of geminivirus- and nanovirus-associated alphasatellites. Arch Virol, 1-14.

Campo, S., Gilbert, K.B., Carrington, J.C., 2016. Small RNA-Based Antiviral Defense in the Phytopathogenic Fungus Colletotrichum higginsianum. PLoS pathogens 12, e1005640.

Capella-Gutierrez, S., Silla-Martinez, J.M., Gabaldon, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics (Oxford, England) 25, 1972-1973.

Dayaram, A., Galatowitsch, M., Harding, J.S., Arguello-Astorga, G.R., Varsani, A., 2014. Novel circular DNA viruses identified in Procordulia grayi and Xanthocnemis zealandica larvae using metagenomic approaches. Infect. Genet. Evol. 22, 134-141.

Dayaram, A., Potter, K.A., Moline, A.B., Rosenstein, D.D., Marinov, M., Thomas, J.E., Breitbart, M., Rosario, K., Arguello-Astorga, G.R., Varsani, A., 2013. High global diversity of cycloviruses amongst dragonflies. J. Gen. Virol. 94, 1827-1840.

Dennis, T.P.W., de Souza, W.M., Marsile-Medun, S., Singer, J.B., Wilson, S.J., Gifford, R.J., 2018a. The evolution, distribution and diversity of endogenous circoviral elements in vertebrate genomes. Virus Res, In press.

Dennis, T.P.W., Flynn, P.J., Marciel de Souza, W., Singer, J.B., Moreau, C.S., Wilson, S.J., Gifford, R.J., 2018b. Insights into circovirus host range from the genomic fossil record. J Virol 92, e00145-18.

Duffy, S., Holmes, E.C., 2008. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. J Virol 82, 957-965.

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

Feschotte, C., Gilbert, C., 2012. Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet 13, 283-296.

Filloux, D., Murrell, S., Koohapitagtam, M., Golden, M., Julian, C., Galzi, S., Uzest, M., Rodier-Goud, M., D'Hont, A., Vernerey, M.S., Wilkin, P., Peterschmitt, M., Winter, S., Murrell, B., Martin, D.P., Roumagnac, P., 2015. The genomes of many yam species contain transcriptionally active endogenous geminiviral sequences that may be functionally expressed. Virus evolution 1, vev002.

Firth, C., Charleston, M.A., Duffy, S., Shapiro, B., Holmes, E.C., 2009. Insights into the Evolutionary History of an Emerging Livestock Pathogen: Porcine Circovirus 2. Journal of Virology 83, 12813.

Gibbs, M.J., Smeianov, V.V., Steele, J.L., Upcroft, P., Efimov, B.A., 2006. Two families of rep-like genes that probably originated by interspecies recombination are represented in viral, plasmid, bacterial, and parasitic protozoan genomes. Molecular biology and evolution 23, 1097-1100.

Gilbert, C., Feschotte, C., 2010. Genomic Fossils Calibrate the Long-Term Evolution of Hepadnaviruses. PLOS Biology 8, e1000495.

Gilbert, C., Maxfield, D.G., Goodman, S.M., Feschotte, C., 2009. Parallel germline infiltration of a lentivirus in two Malagasy lemurs. PLoS Genet 5, e1000425.

Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Molecular Biology and Evolution 27, 221-224.

Guindon, S., Gascuel, O., 2003. A Simple, Fast and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Syst Biol 52, 696-704.

Hayward, A., Katzourakis, A., 2015. Endogenous retroviruses. Curr Biol 25, R644-646.

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

Kenton, A., Khashoggi, A., Parokonny, A., Bennett, M.D., Lichtenstein, C., 1995. Chromosomal location of endogenous geminivirus-related DNA sequences in Nicotiana tabacum L. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology 3, 346-350.

Kerr, M., Rosario, K., Baker, C.C.M., Breitbart, M., 2018. Discovery of Four Novel Circular Single-Stranded DNA Viruses in Fungus-Farming Termites. Genome announcements 6, e00318-00318.

Krupovic, M., Forterre, P., 2015. Single-stranded DNA viruses employ a variety of mechanisms for integration into host genomes. Annals of the New York Academy of Sciences 1341, 41-53.

Krupovic, M., Ravantti, J.J., Bamford, D.H., 2009. Geminiviruses: a tale of a plasmid becoming a virus. BMC Evol. Biol. 9, 112.

Kryukov, K., Ueda, M.T., Imanishi, T., Nakagawa, S., 2018. Systematic survey of non-retroviral virus-like elements in eukaryotic genomes. Virus Research, In press.

Lefeuvre, P., Harkins, G.W., Lett, J.-M., Briddon, R.W., Chase, M.W., Moury, B., Martin, D.P., 2011. Evolutionary Time-Scale of the Begomoviruses: Evidence from Integrated Sequences in the Nicotiana Genome. PLOS ONE 6, e19193.

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

Liu, H.Q., Fu, Y.P., Li, B., Yu, X., Xie, J.T., Cheng, J.S., Ghabrial, S.A., Li, G.Q., Yi, X.H., Jiang, D.H., 2011. Widespread Horizontal Gene Transfer from Circular Single-stranded DNA Viruses to Eukaryotic Genomes. BMC Evol. Biol. 11, 15.

Male, M.F., Kraberger, S., Stainton, D., Kami, V., Varsani, A., 2016. Cycloviruses, gemycircularviruses and other novel replication-associated protein encoding circular viruses in Pacific flying fox (Pteropus tonganus) faeces. Infect. Genet. Evol. 39, 279-292.

Metegnier, G., Becking, T., Chebbi, M.A., Giraud, I., Moumen, B., Schaack, S., Cordaux, R., Gilbert, C., 2015. Comparative paleovirological analysis of crustaceans identifies multiple widespread viral groups. Mobile DNA 6, 16.

Miralles Fusté, J., Shi, Y., Wanrooij, S., Zhu, X., Jemt, E., Persson, Ö., Sabouri, N., Gustafsson, C.M., Falkenberg, M., 2014. In Vivo Occupancy of Mitochondrial Single-Stranded DNA Binding Protein Supports the Strand Displacement Mode of DNA Replication. PLOS Genetics 10, e1004832.

Murad, L., Bielawski, J.P., Matyasek, R., Kovarik, A., Nichols, R.A., Leitch, A.R., Lichtenstein, C.P., 2004. The origin and evolution of geminivirus-related DNA sequences in Nicotiana. Heredity 92, 352-358.

Nisole, S., Saïb, A., 2004. Early steps of retrovirus replicative cycle. Retrovirology 1, 9-9.

Patel, M.R., Emerman, M., Malik, H.S., 2011. Paleovirology—ghosts and gifts of viruses past. Current Opinion in Virology 1, 304-309.

Poncet, V., Couderc, M., Tranchant-Dubreuil, C., Gomez, C., Hamon, P., Hamon, S., Pillon, Y., Munzinger, J., de Kochko, A., 2012. Microsatellite markers for Amborella (Amborellaceae), a monotypic genus endemic to New Caledonia. American journal of botany 99, e411-414.

Roossinck, M.J., Bazán, E.R., 2017. Symbiosis: Viruses as Intimate Partners. Annual Review of Virology 4, 123-139.

Rosario, K., Breitbart, M., Harrach, B., Segales, J., Delwart, E., Biagini, P., Varsani, A., 2017. Revisiting the taxonomy of the family Circoviridae: establishment of the genus Cyclovirus and removal of the genus Gyrovirus. Arch. Virol. 162, 1447-1463.

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

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

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

Rosario, K., Mettel, K.A., Benner, B.E., Johnson, R., Scott, C., Yusseff-Vanegas, S.Z., Baker, C.C.M., Cassill, D.L., Storer, C., Varsani, A., Breitbart, M., 2018. Virus discovery in all three major lineages of terrestrial arthropods highlights the diversity of single-stranded DNA viruses associated with invertebrates. PeerJ 6, e5761.

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

Saccardo, F., Cettul, E., Palmano, S., Noris, E., Firrao, G., 2011. On the alleged origin of geminiviruses from extrachromosomal DNAs of phytoplasmas. BMC Evol. Biol. 11, 185.

Shutt, T.E., Gray, M.W., 2006. Bacteriophage origins of mitochondrial replication and transcription proteins. Trends in Genetics 22, 90-95.

Smith, D.R., Keeling, P.J., 2015. Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. Proceedings of the National Academy of Sciences of the United States of America 112, 10177-10184.

Srivastava, G., Mehrotra, R.C., Dilcher, D.L., 2018. Paleocene Ipomoea (Convolvulaceae) from India with implications for an East Gondwana origin of Convolvulaceae. Proceedings of the National Academy of Sciences 115, 6028-6033.

Steel, O., Kraberger, S., Sikorski, A., Young, L.M., Catchpole, R.J., Stevens, A.J., Ladley, J.J., Coray, D.S., Stainton, D., Dayarama, A., Julian, L., van Bysterveldt, K., Varsani, A., 2016. Circular replication-associated protein encoding DNA viruses identified in the faecal matter of various animals in New Zealand. Infect. Genet. Evol. 43, 151-164.

Stoye, J.P., 2006. Koala retrovirus: a genome invasion in real time. Genome biology 7, 241-241.

Theze, J., Leclercq, S., Moumen, B., Cordaux, R., Gilbert, C., 2014. Remarkable diversity of endogenous viruses in a crustacean genome. Genome biology and evolution 6, 2129-2140.

Tu, T., Budzinska, M.A., Shackel, N.A., Urban, S., 2017. HBV DNA Integration: Molecular Mechanisms and Clinical Implications. Viruses 9, 75.

Varsani, A., Krupovic, M., 2018. Smacoviridae: a new family of animal-associated single-stranded DNA viruses. Arch Virol 163, 2005-2015.

Yang, Z.-H., Qi, M.-X., Qin, Y.-X., Zhu, J.-H., Gui, X.-M., Tao, B., Xu, X.-H., Zhang, F.-G., 2013. Mitochondrial DNA polymorphisms in Phytophthora infestans: New haplotypes are identified and re-defined by PCR. Journal of Microbiological Methods 95, 117-121.

Yu, X., Li, B., Fu, Y.P., Jiang, D.H., Ghabrial, S.A., Li, G.Q., Peng, Y.L., Xie, J.T., Cheng, J.S., Huang, J.B., Yi, X.H., 2010. A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. Proceedings of the National Academy of Sciences of the United States of America 107, 8387-8392.

Yuan, X., Feng, C., Zhang, Z., Zhang, C., 2017. Complete Mitochondrial Genome of Phytophthora nicotianae and Identification of Molecular Markers for the Oomycetes. Frontiers in Microbiology 8, 1484.

Appendices

Appendix 1.1 SNPs above 0.1% called by VarScan for sample phi6-WT on PP. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP, p-value: p-value of SNP detection. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cur_aa	freq	P-value
119			C			T			0.22%	2.50E-02
120			C			T			0.25%	4.71E-03
229			G			A			0.26%	1.21E-04
259			C			T			0.24%	1.09E-03
338	p8	12	C	CCT	P	T	TCT	S	0.20%	3.53E-02
516	p8	71	T	CTG	L	C	CCG	P	0.23%	1.21E-02
582	p8	93	C	CCG	P	T	CTG	L	0.19%	3.64E-02
598	p8	98	C	AAC	N	T	AAT	N	0.26%	5.92E-04
599	p8	99	C	CTC	L	T	TTC	F	0.24%	1.89E-03
709	p8	135	C	TAC	Y	T	TAT	Y	0.20%	1.06E-02
738	p8	145	T	ATG	M	C	ACG	T	0.33%	1.84E-06
899	p12	49	T	CTG	L	C	CCG	P	0.19%	9.87E-03
910	p12	53	C	CCT	P	T	TCT	S	0.58%	4.70E-18
1224	p12	157	C	ACC	T	T	ACT	T	0.27%	8.02E-03
1230	p12	159	T	CGT	R	C	CGC	R	0.23%	2.60E-02
1273	p12	174	T	TCG	S	C	CCG	P	0.24%	1.20E-02
1320	p12	189	C	GTC	V	T	GTT	V	0.47%	1.64E-06
1321	p12	190	C	CAC	H	T	TAC	Y	0.43%	7.43E-06
1353	p9	5	C	CTG	L	T	TTG	L	0.19%	4.93E-02
1368	p9	10	C	CCA	P	T	TCA	S	0.33%	4.44E-05
1369	p9	10	C	CCA	P	T	CTA	L	0.19%	4.47E-02
1423	p9	28	T	CTG	L	C	CCG	P	0.31%	1.85E-05
1517	p9	59	G	TGG	W	T	TGT	C	0.18%	1.64E-02
1618			T			C			0.21%	4.67E-03
1673	p5a	18	A	CAA	Q	C	CAC	H	0.18%	3.64E-02
1696	p5a	26	T	GTG	V	C	GCG	A	0.33%	2.56E-05
1788	p5a	57	T	TCG	S	C	CCG	P	0.34%	4.84E-05
1913	p5a	98	G	GTG	V	A	GTA	V	0.36%	1.76E-05
1914	p5a	99	G	GTG	V	A	ATG	M	0.19%	4.93E-02
1915	p5a	99	T	GTG	V	C	GCG	A	0.73%	4.73E-14

1975	p5a	119	G	CGG	R	A	CAG	Q	0.20%	4.35E-02
1976	p5a	119	G	CGG	R	A	CGA	R	0.25%	6.73E-03
1983	p5a	122	T	TGG	W	C	CGG	R	0.32%	9.29E-04
2004	p5a	129	G	GGC	G	T	TGC	C	0.23%	2.60E-02
2065	p5a	149	T	ATC	I	C	ACC	T	0.19%	4.92E-02
2128	p5a	170	C	ACT	T	T	ATT	I	0.18%	2.43E-02
2246	p5a	209	T	GCT	A	C	GCC	A	0.26%	2.09E-04
2482			C			T			0.21%	9.27E-03
2504			T			C			0.28%	8.13E-04
2750			C			T			0.36%	6.86E-06
2751			T			C			0.43%	3.66E-07
2754			C			T			0.26%	3.19E-03
2764			T			C			0.78%	1.07E-14
2781			C			T			0.34%	1.65E-03
3121			T			C			0.31%	8.20E-04
3295			C			T			0.24%	6.89E-03
3296			C			T			0.64%	2.56E-12
3297			C			T			0.21%	2.50E-02
3330	p10	5	T	CTC	L	C	CCC	P	0.53%	8.60E-09
3354	p10	13	C	TCT	S	T	TTT	F	0.30%	1.51E-04
3423	p10	36	G	GGC	G	T	GTC	V	0.23%	7.75E-04
3488			C			T			0.19%	2.43E-02
3493			G			A			0.39%	9.58E-07
3539			C			T			0.24%	6.89E-03
3767			T			C			0.30%	1.64E-05
3796			G			A			0.33%	7.70E-06
3830			C			T			0.31%	2.32E-04
3903			T			C			0.71%	1.14E-15
3931	p6	6	C	TCG	S	T	TTG	L	0.23%	8.29E-03
3978	p6	22	C	CTT	L	T	TTT	F	0.18%	2.21E-02
4005	p6	31	C	CCG	P	T	TCG	S	0.44%	1.09E-11
4006	p6	31	C	CCG	P	T	CTG	L	0.34%	3.97E-07
4016	p6	34	T	CTT	L	C	CTC	L	0.17%	4.59E-02
4102	p6	63	C	TCT	S	T	TTT	F	0.20%	1.06E-02
4168	p6	85	C	TCT	S	T	TTT	F	0.18%	2.00E-02
4190	p6	92	T	TCT	S	C	TCC	S	0.18%	2.00E-02
4259	p6	115	T	CCT	P	C	CCC	P	0.30%	1.44E-05
4453	p3	10	T	CTG	L	C	CCG	P	0.25%	1.53E-03
4654	p3	77	T	ATC	I	C	ACC	T	0.37%	2.50E-06
4894	p3	157	T	CTG	L	C	CCG	P	0.18%	2.00E-02
4906	p3	161	C	TCG	S	T	TTG	L	0.19%	8.14E-03
4992	p3	190	T	TGG	W	C	CGG	R	0.27%	3.39E-04
5001	p3	193	T	TCG	S	C	CCG	P	0.19%	3.64E-02
5248	p3	275	C	CCT	P	T	CTT	L	0.33%	3.97E-07
5253	p3	277	C	CTG	L	T	TTG	L	0.32%	9.64E-07
5360	p3	312	T	GTT	V	C	GTC	V	0.19%	3.25E-02
5671	p3	416	C	ACT	T	T	ATT	I	0.48%	1.63E-11

5805	p3	461	C	CTG	L	T	TTG	L	0.24%	2.58E-04
5904	p3	494	T	TGG	W	C	CGG	R	0.23%	6.89E-03
6001	p3	526	C	ACT	T	T	ATT	I	0.16%	4.26E-02
6025	p3	534	C	GCT	A	T	GTT	V	0.18%	1.50E-02
6081	p3	553	T	TCC	S	C	CCC	P	0.33%	1.49E-06
6218	p3	598	C	AAC	N	T	AAT	N	0.36%	1.74E-07
6219	p3	599	C	CCA	P	T	TCA	S	0.26%	1.95E-04
6229	p3	602	T	CTC	L	C	CCC	P	0.19%	1.06E-02
6232	p3	603	T	GTC	V	C	GCC	A	0.20%	6.30E-03
6236	p3	604	A	GAA	E	G	GAG	E	0.20%	8.13E-03
6241	p3	606	T	GTC	V	C	GCC	A	0.36%	3.79E-08
6251	p3	609	C	GCC	A	T	GCT	A	0.30%	4.79E-06
6257	p3	611	C	TTC	F	T	TTT	F	0.21%	3.09E-03
6258	p3	612	C	CTG	L	T	TTG	L	0.24%	3.74E-04
6272	p3	616	T	GCT	A	C	GCC	A	0.32%	8.65E-07
6318	p3	632	A	ATG	M	G	GTG	V	0.19%	1.11E-02
6380			T			C			0.55%	2.87E-12
6443			C			T			0.39%	4.25E-06
6454			G			A			0.21%	1.75E-02
6620	p13	54	T	ATC	I	C	ACC	T	0.18%	3.64E-02
6747			C			T			0.23%	1.59E-03
6873			T			C			0.87%	5.57E-16
6914			T			C			0.45%	3.50E-02
6917			T			C			0.56%	1.94E-02
7272			T			C			0.45%	6.31E-03
7485	p7	5	T	CTG	L	C	CCG	P	0.40%	6.42E-03
7489	p7	6	C	GTC	V	T	GTT	V	0.53%	2.06E-03
7490	p7	7	C	CCT	P	T	TCT	S	0.53%	2.06E-03
7503	p7	11	C	TCG	S	T	TTG	L	0.36%	1.06E-02
7555	p7	28	C	CTC	L	T	CTT	L	0.34%	8.39E-03
7568	p7	33	C	CTT	L	T	TTT	F	0.24%	3.83E-02
7635	p7	55	T	CTG	L	C	CCG	P	0.41%	3.56E-03
7701	p7	77	C	CCC	P	T	CTC	L	0.36%	9.54E-03
8080	p2	42	G	GAA	E	T	TAA	X	0.24%	3.17E-02
8351	p2	132	T	GTC	V	C	GCC	A	0.31%	1.75E-02
8429	p2	158	C	TCT	S	T	TTT	F	0.81%	6.68E-07
8490	p2	178	T	GCT	A	C	GCC	A	0.34%	1.54E-02
9103	p2	383	C	CCG	P	T	TCG	S	0.35%	9.54E-03
9112	p2	386	C	CCT	P	T	TCT	S	0.34%	1.54E-02
9113	p2	386	C	CCT	P	T	CTT	L	0.47%	7.61E-04
9136	p2	394	T	TCC	S	C	CCC	P	0.28%	2.86E-02
9402	p2	482	C	AAC	N	T	AAT	N	0.46%	3.73E-03
9403	p2	483	C	CCC	P	T	TCC	S	0.34%	2.86E-02
9405	p2	483	C	CCC	P	T	CCT	P	0.31%	4.60E-02
9498	p2	514	C	ATC	I	T	ATT	I	0.28%	3.83E-02
9642	p2	562	T	TGT	C	C	TGC	C	0.26%	4.60E-02
9673	p2	573	G	GAG	E	T	TAG	X	0.43%	2.15E-03

9799	p2	615	C	CTC	L	T	TTC	F	0.33%	3.26E-02
9843	p2	629	C	CTC	L	T	CTT	L	0.34%	1.92E-02
9974	p4	6	T	ACT	T	C	ACC	T	0.59%	2.06E-03
10069	p4	38	C	ACT	T	T	ATT	I	0.57%	3.58E-04
10177	p4	74	C	GCT	A	T	GTT	V	0.27%	1.73E-02
10219	p4	88	C	TCT	S	T	TTT	F	0.39%	7.48E-04
10495	p4	180	T	CTC	L	C	CCC	P	0.37%	5.86E-03
10636	p4	227	T	GTC	V	C	GCC	A	0.33%	4.60E-02
10729	p4	258	C	GCT	A	T	GTT	V	0.26%	4.60E-02
10804	p4	283	C	GCT	A	T	GTT	V	0.26%	3.83E-02
11045	p1	27	C	AAC	N	T	AAT	N	0.31%	2.86E-02
11270	p1	102	C	ATC	I	T	ATT	I	0.38%	3.27E-03
11271	p1	103	T	TGG	W	C	CGG	R	0.70%	1.60E-07
11359	p1	132	T	CTG	L	C	CCG	P	0.50%	5.53E-05
11885	p1	307	C	ATC	I	T	ATT	I	0.27%	4.60E-02
12175	p1	404	T	ATC	I	C	ACC	T	0.50%	3.73E-03
12592	p1	543	C	CCT	P	T	CTT	L	0.30%	2.86E-02
12593	p1	543	T	CCT	P	C	CCC	P	0.55%	2.39E-04
12611	p1	549	C	ATC	I	T	ATT	I	0.37%	1.75E-02
12616	p1	551	A	TAC	Y	C	TCC	S	0.45%	3.73E-03
12751	p1	596	C	ACC	T	T	ATC	I	0.40%	1.92E-02
12752	p1	596	C	ACC	T	T	ACT	T	0.48%	6.41E-03
13007	p1	681	T	ATT	I	C	ATC	I	0.85%	5.86E-05

Appendix 1.2 SNPs above 0.1% called by VarScan for sample phi6-WT on PA. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP, p-value: p-value of SNP detection. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq	P-value
146			T			C			0.30%	5.44E-06
149			T			C			0.17%	3.51E-02
229			G			A			0.32%	5.59E-07
259			C			T			0.20%	1.29E-02
600	p8	99	T	CTC	L	C	CCC	P	0.24%	5.40E-04
738	p8	145	T	ATG	M	C	ACG	T	0.20%	5.20E-03
779	p12	9	T	CTC	L	C	CCC	P	0.29%	2.19E-05
910	p12	53	C	CCT	P	T	TCT	S	0.46%	1.45E-12
1097	p12	115	A	AAG	K	C	ACG	T	0.31%	9.76E-05
1104	p12	117	A	TCA	S	G	TCG	S	1%	4.99E-25
1106	p12	118	A	AAC	N	G	AGC	S	0.19%	4.47E-02
1122	p12	123	C	GGC	G	G	GGG	G	0.80%	3.75E-22
1128	p12	125	T	GGT	G	G	GGG	G	0.19%	1.64E-02
1212	p12	153	T	GCT	A	G	GCG	A	0.67%	2.20E-22
1224	p12	157	C	ACC	T	T	ACT	T	0.22%	1.57E-03
1230	p12	159	T	CGT	R	C	CGC	R	0.29%	1.66E-05
1272	p12	173	T	ACT	T	G	ACG	T	0.31%	1.49E-06
1273	p12	174	T	TCG	S	C	CCG	P	0.46%	6.42E-12
1277	p12	175	C	ACC	T	A	AAC	N	0.60%	3.97E-18
1278	p12	175	C	ACC	T	T	ACT	T	0.19%	1.46E-02
1279	p12	176	T	TTT	F	C	CTT	L	1.90%	1.16E-74
1288	p12	179	A	AAA	K	C	CAA	Q	1.43%	3.29E-62
1308	p12	185	T	TTT	F	C	TTC	F	0.91%	3.16E-33
1309	p12	186	G	GTG	V	A	ATG	M	0.32%	8.65E-07
1316	p12	188	T	ATG	M	C	ACG	T	0.40%	5.52E-10
1317	p12	188	G	ATG	M	A	ATA	I	0.59%	2.73E-18
1318	p12	189	G	GTC	V	A	ATC	I	0.85%	1.18E-30
1320	p12	189	C	GTC	V	T	GTT	V	1.33%	6.76E-56
1321	p12	190	C	CAC	H	T	TAC	Y	0.20%	4.30E-03
1328	p12	192	A	AAG	K	G	AGG	R	1.65%	3.65E-74
1331	p12	193	A	GAT	D	G	GGT	G	2.07%	1.60E-96
1334	p12	194	T	GTT	V	G	GGT	G	0.19%	1.11E-02
1335	p12	194	T	GTT	V	G	GTG	V	0.45%	3.99E-12

1341	p12	196	A	TAA	X	T	TAT	Y	0.38%	3.72E-09
1349	p9	3	T	TTT	F	G	TTG	L	0.35%	3.80E-08
1353	p9	5	C	CTG	L	T	TTG	L	0.19%	1.11E-02
1357	p9	6	T	GTA	V	C	GCA	A	0.28%	1.66E-05
1359	p9	7	A	AAG	K	G	GAG	E	0.39%	8.92E-10
1363	p9	8	A	CAA	Q	G	CGA	R	5.03%	7.07E-270
1364	p9	8	A	CAA	Q	G	CAG	Q	0.42%	1.29E-10
1381	p9	14	C	GCT	A	G	GGT	G	0.24%	3.74E-04
1384	p9	15	T	TTC	F	G	TGC	C	0.23%	1.10E-03
1423	p9	28	T	CTG	L	C	CCG	P	0.63%	1.93E-20
1470	p9	44	C	CAC	H	T	TAC	Y	0.18%	1.50E-02
1541	p9	67	C	GGC	G	T	GGT	G	0.20%	4.30E-03
1559	p9	73	T	CGT	R	C	CGC	R	0.21%	4.30E-03
1758	p5a	47	G	GAG	E	C	CAG	Q	0.73%	4.04E-23
1788	p5a	57	T	TCG	S	C	CCG	P	0.29%	1.10E-05
1913	p5a	98	G	GTG	V	A	GTA	V	0.19%	9.87E-03
1914	p5a	99	G	GTG	V	A	ATG	M	0.24%	7.67E-04
1915	p5a	99	T	GTG	V	C	GCG	A	0.60%	5.13E-17
1983	p5a	122	T	TGG	W	C	CGG	R	0.21%	5.20E-03
2065	p5a	149	T	ATC	I	C	ACC	T	0.45%	6.59E-12
2128	p5a	170	C	ACT	T	T	ATT	I	0.17%	3.51E-02
2351			T			C			0.18%	2.66E-02
2380			T			C			0.21%	3.73E-03
2504			T			C			0.18%	2.00E-02
2712			G			A			0.17%	2.66E-02
2750			C			T			0.18%	2.44E-02
2754			C			T			0.18%	1.64E-02
2764			T			C			0.19%	1.64E-02
2825			C			T			0.77%	8.17E-16
2851			C			T			0.62%	2.39E-05
2891			A			G			0.68%	3.49E-02
2901			T			G			1.89%	1.15E-04
3121			T			C			0.25%	4.27E-04
3159			T			C			0.17%	4.93E-02
3292			C			T			0.23%	1.32E-03
3296			C			T			0.23%	1.90E-03
3330	p10	5	T	CTC	L	C	CCC	P	0.31%	2.84E-05
3354	p10	13	C	TCT	S	T	TTT	F	0.24%	3.74E-04
3381	p10	22	C	GCT	A	T	GTT	V	0.32%	1.33E-06
3411	p10	32	T	GTC	V	C	GCC	A	0.27%	5.53E-05
3549			G			A			0.17%	2.44E-02
3644			T			C			0.24%	3.74E-04
3702			G			A			0.52%	1.41E-14
3796			G			A			0.31%	2.05E-06
3818			T			C			0.28%	2.49E-05
3869			T			C			0.31%	3.14E-06
3903			T			C			0.56%	2.39E-17

3931	p6	6	C	TCG	S	T	TTG	L	0.20%	8.13E-03
4005	p6	31	C	CCG	P	T	TCG	S	0.27%	8.19E-05
4006	p6	31	C	CCG	P	T	CTG	L	0.27%	7.50E-05
4016	p6	34	T	CTT	L	C	CTC	L	0.83%	1.59E-28
4102	p6	63	C	TCT	S	T	TTT	F	0.26%	1.31E-04
4172	p6	86	T	CTT	L	C	CTC	L	0.53%	2.91E-15
4188	p6	92	T	TCT	S	C	CCT	P	0.17%	4.26E-02
4190	p6	92	T	TCT	S	C	TCC	S	0.36%	1.51E-08
4259	p6	115	T	CCT	P	C	CCC	P	0.20%	8.13E-03
4515	p3	31	G	GCT	A	A	ACT	T	0.35%	5.99E-08
4527	p3	35	G	GAC	D	T	TAC	Y	1.14%	9.72E-44
4528	p3	35	A	GAC	D	C	GCC	A	4.03%	8.01E-196
4555	p3	44	T	GTG	V	C	GCG	A	0.17%	4.26E-02
4608	p3	62	C	CGT	R	T	TGT	C	0.17%	2.66E-02
4654	p3	77	T	ATC	I	C	ACC	T	0.50%	3.99E-14
4741	p3	106	T	CTC	L	C	CCC	P	0.25%	1.77E-04
4822	p3	133	C	GCG	A	T	GTG	V	18.29%	0.00E+00
4843	p3	140	A	CAG	Q	G	CGG	R	3.33%	3.32E-170
4855	p3	144	A	AAG	K	G	AGG	R	18.78%	0.00E+00
4861	p3	146	A	AAT	N	G	AGT	S	1.53%	2.92E-65
4894	p3	157	T	CTG	L	G	CGG	R	0.27%	3.72E-05
5060	p3	212	A	GCA	A	G	GCG	A	0.55%	7.02E-17
5099	p3	225	T	GCT	A	C	GCC	A	0.23%	9.26E-04
5163	p3	247	G	GGC	G	C	CGC	R	0.85%	1.92E-28
5251	p3	276	T	CTC	L	C	CCC	P	0.19%	9.87E-03
5264	p3	280	C	GTC	V	T	GTT	V	0.27%	3.72E-05
5320	p3	299	C	TCG	S	G	TGG	W	0.64%	6.17E-19
5404	p3	327	C	GCT	A	T	GTT	V	0.17%	3.24E-02
5471	p3	349	T	CCT	P	C	CCC	P	0.34%	1.48E-07
5671	p3	416	C	ACT	T	T	ATT	I	0.35%	9.42E-08
6081	p3	553	T	TCC	S	C	CCC	P	0.24%	3.74E-04
6112	p3	563	A	TAC	Y	T	TTC	F	1.06%	3.86E-42
6114	p3	564	T	TTG	L	C	CTG	L	0.60%	5.30E-19
6154	p3	577	C	TCC	S	T	TTC	F	0.17%	4.58E-02
6218	p3	598	C	AAC	N	T	AAT	N	0.25%	3.61E-04
6241	p3	606	T	GTC	V	C	GCC	A	0.18%	2.00E-02
6272	p3	616	T	GCT	A	C	GCC	A	0.43%	7.87E-11
6380			T			C			0.29%	5.77E-05
6443			C			T			0.23%	1.32E-03
6454			G			A			1.15%	7.23E-42
6541	p13	28	C	CTC	L	T	TTC	F	0.35%	1.02E-07
6620	p13	54	T	ATC	I	C	ACC	T	0.23%	1.91E-03
6724			T			C			0.21%	2.21E-03
6873			T			C			0.50%	2.29E-09
6903			T			C			0.37%	1.23E-03
6907			T			C			0.31%	8.40E-03
6914			T			C			0.26%	2.44E-02

6917			T			C			0.44%	4.49E-04
6928			T			A			0.33%	1.06E-02
7272			T			C			0.30%	8.40E-03
7485	p7	5	T	CTG	L	C	CCG	P	0.37%	5.86E-03
7490	p7	7	C	CCT	P	T	TCT	S	0.32%	2.86E-02
7568	p7	33	C	CTT	L	T	TTT	F	0.29%	2.06E-02
7753	p7	94	T	TCT	S	C	TCC	S	0.28%	2.44E-02
8193	p2	79	C	AAC	N	T	AAT	N	0.32%	2.44E-02
8228	p2	91	T	ATG	M	C	ACG	T	0.26%	3.83E-02
8351	p2	132	T	GTC	V	C	GCC	A	0.46%	4.49E-04
8429	p2	158	C	TCT	S	T	TTT	F	0.33%	9.54E-03
8966	p2	337	T	CTG	L	C	CCG	P	0.54%	2.40E-04
9334	p2	460	T	TGG	W	C	CGG	R	0.28%	2.44E-02
9389	p2	478	A	GAG	E	G	GGG	G	0.42%	2.15E-03
9498	p2	514	C	ATC	I	T	ATT	I	0.25%	3.83E-02
9799	p2	615	C	CTC	L	T	TTC	F	0.41%	3.56E-03
9974	p4	6	T	ACT	T	C	ACC	T	0.29%	2.86E-02
10069	p4	38	C	ACT	T	T	ATT	I	0.31%	1.54E-02
10219	p4	88	C	TCT	S	T	TTT	F	0.25%	3.91E-02
10375	p4	140	T	CTC	L	C	CCC	P	0.32%	5.27E-03
10495	p4	180	T	CTC	L	C	CCC	P	0.25%	3.17E-02
10636	p4	227	T	GTC	V	C	GCC	A	0.29%	3.83E-02
10797	p4	281	C	CTT	L	T	TTT	F	0.68%	2.12E-07
10944	p4	330	T	TTC	F	C	CTC	L	0.25%	4.80E-02
11131	p1	56	T	GTG	V	C	GCG	A	0.26%	4.80E-02
11271	p1	103	T	TGG	W	C	CGG	R	0.60%	7.88E-07
11359	p1	132	T	CTG	L	C	CCG	P	0.43%	9.58E-05
12019	p1	352	T	ATG	M	C	ACG	T	0.40%	2.19E-03
12175	p1	404	T	ATC	I	C	ACC	T	0.40%	4.50E-04
12474	p1	504	T	TCG	S	C	CCG	P	0.30%	1.32E-02
12593	p1	543	T	CCT	P	C	CCC	P	1%	1.07E-13
12616	p1	551	A	TAC	Y	C	TCC	S	0.25%	2.06E-02
13007	p1	681	T	ATT	I	C	ATC	I	1.57%	2.89E-19

Appendix 1.3 SNPs above 0.1% called by VarScan for sample phi6-E8G on PT. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP, p-value: p-value of SNP detection. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq	P-value
738	p8	145	T	ATG	M	C	ACG	T	0.21%	2.50E-02
779	p12	9	T	CTC	L	C	CCC	P	0.25%	9.98E-03
910	p12	53	C	CCT	P	T	TCT	S	0.34%	2.31E-05
1067	p12	105	A	AAC	N	G	AGC	S	0.22%	1.47E-02
1273	p12	174	T	TCG	S	C	CCG	P	0.27%	8.01E-03
1423	p9	28	T	CTG	L	C	CCG	P	0.35%	3.03E-05
1594	p9	85	A	AAC	N	G	AGC	S	0.29%	2.26E-03
1618			T			C			0.29%	2.97E-03
1696	p5a	26	T	GTG	V	C	GCG	A	0.24%	1.44E-02
1788	p5a	57	T	TCG	S	C	CCG	P	0.31%	4.03E-03
1913	p5a	98	G	GTG	V	A	GTA	V	0.21%	4.35E-02
1915	p5a	99	T	GTG	V	C	GCG	A	0.54%	3.86E-08
1975	p5a	119	G	CGG	R	A	CAG	Q	0.21%	2.60E-02
1983	p5a	122	T	TGG	W	C	CGG	R	0.34%	6.52E-04
2065	p5a	149	T	ATC	I	C	ACC	T	0.42%	7.51E-06
2754			C			T			0.26%	2.15E-03
2764			T			C			0.49%	7.78E-08
3121			T			C			0.23%	2.13E-02
3296			C			T			0.21%	2.50E-02
3330	p10	5	T	CTC	L	C	CCC	P	0.62%	4.59E-10
3354	p10	13	C	TCT	S	T	TTT	F	0.26%	4.49E-03
3493			G			A			0.25%	3.28E-03
3644			T			C			0.18%	4.02E-02
3767			T			C			0.21%	4.67E-03
3796			G			A			0.29%	8.69E-05
3869			T			C			0.28%	1.01E-04
3903			T			C			0.75%	2.92E-21
4005	p6	31	C	CCG	P	T	TCG	S	0.26%	1.21E-04
4006	p6	31	C	CCG	P	T	CTG	L	0.26%	1.42E-04
4190	p6	92	T	TCT	S	C	TCC	S	0.20%	8.14E-03
4259	p6	115	T	CCT	P	C	CCC	P	0.22%	1.88E-03
4447	p3	8	G	GGG	G	A	GAG	E	0.18%	4.93E-02
4654	p3	77	T	ATC	I	C	ACC	T	0.69%	1.15E-16

4741	p3	106	T	CTC	L	C	CCC	P	0.22%	1.57E-03
4747	p3	108	T	CTT	L	C	CCT	P	0.17%	2.66E-02
4894	p3	157	T	CTG	L	G	CGG	R	0.29%	6.63E-05
5671	p3	416	C	ACT	T	T	ATT	I	0.27%	6.95E-04
5904	p3	494	T	TGG	W	C	CGG	R	0.23%	8.25E-03
6081	p3	553	T	TCC	S	C	CCC	P	0.26%	2.89E-04
6241	p3	606	T	GTC	V	C	GCC	A	0.30%	5.44E-06
6272	p3	616	T	GCT	A	C	GCC	A	0.39%	1.49E-09
6380			T			C			0.62%	7.27E-15
6443			C			T			0.20%	4.35E-02
6620	p13	54	T	ATC	I	C	ACC	T	0.20%	3.53E-02
6873			T			C			1.15%	8.44E-19
6917			T			C			0.71%	1.93E-02
7272			T			C			0.18%	3.64E-02
7485	p7	5	T	CTG	L	C	CCG	P	0.52%	1.25E-09
7489	p7	6	C	GTC	V	T	GTT	V	0.21%	2.13E-02
7568	p7	33	C	CTT	L	T	TTT	F	0.17%	3.94E-02
7635	p7	55	T	CTG	L	C	CCG	P	0.17%	4.41E-02
7753	p7	94	T	TCT	S	C	TCC	S	0.19%	2.97E-02
8230	p2	92	A	AAC	N	G	GAC	D	0.18%	3.25E-02
8351	p2	132	T	GTC	V	C	GCC	A	0.34%	1.48E-05
8429	p2	158	C	TCT	S	T	TTT	F	0.27%	1.81E-03
8490	p2	178	T	GCT	A	C	GCC	A	0.18%	2.97E-02
8639	p2	228	T	GTC	V	C	GCC	A	0.25%	8.29E-03
8652	p2	232	A	GAA	E	C	GAC	D	1%	1.60E-21
8966	p2	337	T	CTG	L	C	CCG	P	0.34%	2.12E-04
9231	p2	425	T	AGT	S	C	AGC	S	0.18%	4.93E-02
9334	p2	460	T	TGG	W	C	CGG	R	0.32%	2.99E-04
9351	p2	465	T	GCT	A	C	GCC	A	0.19%	4.93E-02
9673	p2	573	G	GAG	E	T	TAG	X	0.71%	8.41E-14
9948	p2	664	G	CCG	P	T	CCT	P	0.25%	4.49E-03
9974	p4	6	T	ACT	T	C	ACC	T	0.29%	2.59E-03
10072	p4	39	T	GTC	V	C	GCC	A	0.20%	4.35E-02
10375	p4	140	T	CTC	L	C	CCC	P	0.27%	1.10E-03
10495	p4	180	T	CTC	L	C	CCC	P	0.29%	1.67E-03
10561	p4	202	C	GCG	A	T	GTG	V	0.21%	4.65E-02
11271	p1	103	T	TGG	W	C	CGG	R	0.77%	4.69E-12
11359	p1	132	T	CTG	L	C	CCG	P	0.50%	3.32E-07
12175	p1	404	T	ATC	I	C	ACC	T	0.46%	2.62E-04
12426	p1	488	G	GCG	A	A	ACG	T	0.26%	4.60E-02
12593	p1	543	T	CCT	P	C	CCC	P	0.65%	9.34E-06
12886	p1	641	T	GTC	V	C	GCC	A	0.38%	1.12E-02
12950	p1	662	C	GCC	A	T	GCT	A	0.37%	3.26E-02
13007	p1	681	T	ATT	I	C	ATC	I	0.83%	1.96E-04

Appendix 1.4 SNPs above 0.1% called by VarScan for sample phi6-E8G on PA. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP, p-value: p-value of SNP detection. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq	P-value
159			G			T			0.17%	4.26E-02
165			C			T			0.18%	2.00E-02
229			G			A			0.18%	2.00E-02
259			C			T			0.23%	4.67E-03
260			C			T			0.19%	1.57E-02
329	p8	9	G	GCG	A	T	TCG	S	0.18%	3.29E-02
649	p8	115	C	TTC	F	T	TTT	F	0.18%	2.66E-02
690	p8	129	T	CTG	L	C	CCG	P	0.18%	4.02E-02
705	p8	134	A	AAG	K	G	AGG	R	1.62%	1.25E-50
718	p8	138	G	CTG	L	C	CTC	L	0.17%	4.93E-02
779	p12	9	T	CTC	L	C	CCC	P	0.19%	4.00E-02
811	p12	20	C	CGC	R	A	AGC	S	0.34%	1.62E-05
835	p12	28	G	GAA	E	T	TAA	X	0.21%	9.71E-03
858	p12	35	G	ACG	T	T	ACT	T	0.18%	4.41E-02
910	p12	53	C	CCT	P	T	TCT	S	0.58%	7.81E-16
944	p12	64	C	TCT	S	T	TTT	F	0.17%	3.51E-02
1224	p12	157	C	ACC	T	T	ACT	T	0.30%	4.79E-06
1273	p12	174	T	TCG	S	C	CCG	P	0.38%	9.48E-09
1278	p12	175	C	ACC	T	T	ACT	T	0.19%	1.99E-02
1318	p12	189	G	GTC	V	A	ATC	I	0.19%	1.50E-02
1320	p12	189	C	GTC	V	T	GTT	V	0.32%	8.66E-07
1321	p12	190	C	CAC	H	T	TAC	Y	0.28%	2.49E-05
1351	p9	4	C	CCT	P	T	CTT	L	0.16%	4.59E-02
1352	p9	4	T	CCT	P	C	CCC	P	0.20%	8.13E-03
1353	p9	5	C	CTG	L	T	TTG	L	0.27%	2.49E-05
1376	p9	12	G	TCG	S	C	TCC	S	0.18%	1.50E-02
1381	p9	14	C	GCT	A	G	GGT	G	0.16%	4.59E-02
1423	p9	28	T	CTG	L	C	CCG	P	0.42%	7.89E-11
1473	p9	45	G	GAG	E	C	CAG	Q	0.18%	2.00E-02
1517	p9	59	G	TGG	W	T	TGT	C	0.18%	2.44E-02
1542	p9	68	G	GAG	E	T	TAG	X	0.29%	1.26E-05
1547	p9	69	G	CTG	L	C	CTC	L	0.24%	3.74E-04
1696	p5a	26	T	GTG	V	C	GCG	A	0.19%	7.19E-03

1788	p5a	57	T	TCG	S	C	CCG	P	0.35%	2.75E-07
1868	p5a	83	T	GTT	V	C	GTC	V	0.18%	2.21E-02
1913	p5a	98	G	GTG	V	A	GTA	V	0.30%	3.55E-06
1914	p5a	99	G	GTG	V	A	ATG	M	0.29%	3.32E-05
1915	p5a	99	T	GTG	V	C	GCG	A	0.78%	3.99E-24
1983	p5a	122	T	TGG	W	C	CGG	R	0.24%	5.28E-04
2065	p5a	149	T	ATC	I	C	ACC	T	0.39%	6.24E-09
2128	p5a	170	C	ACT	T	T	ATT	I	0.28%	1.10E-05
2237	p5a	206	G	GCG	A	C	GCC	A	0.24%	4.45E-04
2276	p5a	219	G	TGG	W	T	TGT	C	0.19%	2.65E-02
2422			G			T			0.18%	1.50E-02
2425			G			T			0.29%	7.28E-06
2504			T			C			0.20%	1.06E-02
2624			G			C			0.39%	1.44E-09
2750			C			T			0.40%	1.08E-08
2751			T			C			0.39%	2.94E-08
2754			C			T			0.23%	1.58E-03
2764			T			C			0.21%	8.01E-03
2792			G			C			0.38%	2.62E-06
2972			C			T			0.26%	4.60E-02
3153			T			C			0.22%	1.18E-02
3181			T			C			0.20%	1.57E-02
3211			G			T			0.17%	4.80E-02
3292			C			T			0.23%	1.88E-03
3295			C			T			0.26%	2.46E-04
3296			C			T			0.55%	5.90E-15
3330	p10	5	T	CTC	L	C	CCC	P	0.51%	9.09E-12
3354	p10	13	C	TCT	S	T	TTT	F	0.28%	6.63E-05
3406	p10	30	G	GCG	A	T	GCT	A	0.17%	3.94E-02
3411	p10	32	T	GTC	V	C	GCC	A	0.18%	2.21E-02
3471			G			C			0.21%	3.09E-03
3493			G			A			0.27%	3.97E-04
3517			G			T			0.31%	5.59E-05
3539			C			T			0.19%	1.78E-02
3587			T			G			0.18%	2.97E-02
3644			T			C			0.40%	6.68E-09
3680			G			T			0.18%	2.21E-02
3796			G			A			0.35%	3.80E-08
3802			G			C			0.21%	4.51E-03
3830			C			T			0.27%	1.66E-04
3869			T			C			0.27%	6.45E-05
3903			T			C			0.84%	3.47E-28
3931	p6	6	C	TCG	S	T	TTG	L	0.20%	1.29E-02
4005	p6	31	C	CCG	P	T	TCG	S	0.42%	6.78E-09
4006	p6	31	C	CCG	P	T	CTG	L	0.40%	2.09E-07
4007	p6	31	G	CCG	P	T	CCT	P	0.19%	4.02E-02
4101	p6	63	T	TCT	S	C	CCT	P	0.17%	2.96E-02

4102	p6	63	C	TCT	S	T	TTT	F	0.36%	4.34E-08
4170	p6	86	C	CTT	L	T	TTT	F	0.20%	5.93E-03
4188	p6	92	T	TCT	S	C	CCT	P	0.25%	6.26E-04
4190	p6	92	T	TCT	S	C	TCC	S	0.39%	1.00E-08
4345	p6	144	T	CTC	L	C	CCC	P	0.25%	2.89E-04
4447	p3	8	G	GGG	G	A	GAG	E	1.22%	1.89E-31
4549	p3	42	A	AAG	K	G	AGG	R	0.33%	5.59E-07
4654	p3	77	T	ATC	I	C	ACC	T	0.46%	7.57E-11
4668	p3	82	G	GTT	V	T	TTT	F	0.17%	4.80E-02
4741	p3	106	T	CTC	L	C	CCC	P	0.19%	1.50E-02
4747	p3	108	T	CTT	L	C	CCT	P	0.23%	7.75E-04
4822	p3	133	C	GCG	A	T	GTG	V	94.75%	0.00E+00
4855	p3	144	A	AAG	K	G	AGG	R	0.36%	1.87E-07
4863	p3	147	C	CTC	L	T	TTC	F	0.20%	8.73E-03
4894	p3	157	T	CTG	L	G	CGG	R	0.26%	1.21E-04
5001	p3	193	T	TCG	S	C	CCG	P	0.20%	1.67E-02
5042	p3	206	G	GGG	G	T	GGT	G	0.21%	9.27E-03
5046	p3	208	G	GCC	A	T	TCC	S	0.18%	3.29E-02
5069	p3	215	G	CAG	Q	T	CAT	H	0.17%	4.80E-02
5122	p3	233	G	TGG	W	T	TTG	L	0.17%	2.66E-02
5151	p3	243	C	CAC	H	A	AAC	N	0.17%	4.26E-02
5248	p3	275	C	CCT	P	T	CTT	L	0.25%	3.61E-04
5251	p3	276	T	CTC	L	C	CCC	P	0.18%	2.70E-02
5253	p3	277	C	CTG	L	T	TTG	L	0.18%	2.96E-02
5265	p3	281	G	GAC	D	T	TAC	Y	0.22%	1.57E-03
5294	p3	290	A	GCA	A	C	GCC	A	0.76%	4.25E-26
5320	p3	299	C	TCG	S	G	TGG	W	2.33%	######
5329	p3	302	G	CGG	R	C	CCG	P	0.19%	8.14E-03
5417	p3	331	G	ACG	T	T	ACT	T	0.19%	2.34E-02
5471	p3	349	T	CCT	P	C	CCC	P	0.26%	3.39E-04
5531	p3	369	C	ACC	T	A	ACA	T	0.28%	2.65E-04
5596	p3	391	G	CGT	R	C	CCT	P	0.30%	9.43E-06
5671	p3	416	C	ACT	T	T	ATT	I	0.49%	1.12E-13
5694	p3	424	G	GCC	A	T	TCC	S	0.21%	3.09E-03
5805	p3	461	C	CTG	L	T	TTG	L	0.22%	1.10E-03
6025	p3	534	C	GCT	A	T	GTT	V	0.19%	1.20E-02
6081	p3	553	T	TCC	S	C	CCC	P	0.30%	5.44E-06
6107	p3	561	G	GTG	V	T	GTT	V	0.16%	4.59E-02
6218	p3	598	C	AAC	N	T	AAT	N	0.28%	1.01E-04
6219	p3	599	C	CCA	P	T	TCA	S	0.24%	1.58E-03
6232	p3	603	T	GTC	V	C	GCC	A	0.22%	2.73E-03
6251	p3	609	C	GCC	A	T	GCT	A	0.19%	1.11E-02
6258	p3	612	C	CTG	L	T	TTG	L	0.17%	2.66E-02
6260	p3	612	G	CTG	L	C	CTC	L	0.18%	2.00E-02
6272	p3	616	T	GCT	A	C	GCC	A	0.27%	8.19E-05
6380			T			C			0.54%	4.89E-12
6443			C			T			0.20%	2.50E-02

6620	p13	54	T	ATC	I	C	ACC	T	0.19%	2.65E-02
6657	p13	66	G	ACG	T	T	ACT	T	0.20%	2.17E-02
6673	p13	72	C	CTC	L	T	TTC	F	0.31%	5.59E-05
6747			C			T			0.21%	2.21E-03
6834			G			T			0.21%	6.61E-03
6873			T			C			0.84%	1.39E-16
6914			T			C			0.36%	2.85E-02
6962			A			G			0.68%	3.10E-02
7272			T			C			0.33%	1.75E-02
7485	p7	5	T	CTG	L	C	CCG	P	0.56%	1.16E-03
7490	p7	7	C	CCT	P	T	TCT	S	0.41%	3.26E-02
8067	p2	37	G	GCG	A	C	GCC	A	0.68%	2.39E-05
8105	p2	50	G	CGG	R	C	CCG	P	0.36%	1.75E-02
8121	p2	55	G	AAG	K	T	AAT	N	0.30%	4.60E-02
8429	p2	158	C	TCT	S	T	TTT	F	0.33%	2.86E-02
8637	p2	227	G	ATG	M	T	ATT	I	0.32%	3.26E-02
8652	p2	232	A	GAA	E	C	GAC	D	0.64%	1.97E-04
8793	p2	279	G	GCG	A	T	GCT	A	0.31%	3.26E-02
9260	p2	435	G	CGC	R	T	CTC	L	0.26%	4.60E-02
9368	p2	471	G	CGT	R	T	CTT	L	0.38%	1.12E-02
9498	p2	514	C	ATC	I	T	ATT	I	0.32%	2.86E-02
9548	p2	531	A	TAC	Y	G	TGC	C	0.84%	3.64E-07
9559	p2	535	T	TCG	S	C	CCG	P	0.29%	2.86E-02
9598	p2	548	C	CCC	P	T	TCC	S	0.36%	1.92E-02
9599	p2	548	C	CCC	P	T	CTC	L	0.42%	6.42E-03
9799	p2	615	C	CTC	L	T	TTC	F	0.28%	4.60E-02
10034	p4	26	G	TCG	S	C	TCC	S	0.43%	1.11E-02
10069	p4	38	C	ACT	T	T	ATT	I	0.54%	7.33E-04
10071	p4	39	G	GTC	V	C	CTC	L	0.29%	4.60E-02
10219	p4	88	C	TCT	S	T	TTT	F	0.29%	3.83E-02
10260	p4	102	G	GGA	G	T	TGA	X	0.25%	4.60E-02
10636	p4	227	T	GTC	V	C	GCC	A	0.32%	3.26E-02
10742	p4	262	G	GTG	V	A	GTA	V	0.43%	6.31E-03
10943	p4	329	C	AAC	N	T	AAT	N	0.28%	2.86E-02
11271	p1	103	T	TGG	W	C	CGG	R	0.65%	5.28E-06
11359	p1	132	T	CTG	L	C	CCG	P	0.43%	7.47E-04
11724	p1	254	C	CTG	L	T	TTG	L	0.31%	3.26E-02
12151	p1	396	G	CGT	R	C	CCT	P	0.47%	4.48E-04
12175	p1	404	T	ATC	I	C	ACC	T	0.38%	3.57E-03
12185	p1	407	G	ACG	T	T	ACT	T	0.28%	2.86E-02
12200	p1	412	G	GCG	A	T	GCT	A	0.26%	4.60E-02
12405	p1	481	C	CGG	R	A	AGG	R	0.36%	1.06E-02
12419	p1	485	C	AAC	N	T	AAT	N	0.82%	6.67E-07
12474	p1	504	T	TCG	S	C	CCG	P	0.33%	1.75E-02
12593	p1	543	T	CCT	P	C	CCC	P	0.41%	3.56E-03
12696	p1	578	T	TGG	W	C	CGG	R	0.30%	4.60E-02
12751	p1	596	C	ACC	T	T	ATC	I	0.33%	2.86E-02

12752	p1	596	C	ACC	T	T	ACT	T	0.31%	4.60E-02
12901	p1	646	C	ACT	T	T	ATT	I	0.33%	1.75E-02
12950	p1	662	C	GCC	A	T	GCT	A	1.01%	3.11E-08
12980	p1	672	G	GTG	V	A	GTA	V	0.35%	1.75E-02
13007	p1	681	T	ATT	I	C	ATC	I	1.36%	5.75E-11
13156	p1	731	T	GTT	V	C	GCT	A	0.43%	6.31E-03

Appendix 1.5 SNPs above 0.1% called by VarScan for sample phi6-G515S on PT. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP, p-value: p-value of SNP detection. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq	P-value
229			G			A			0.37%	7.48E-04
1320	p12	189	C	GTC	V	T	GTT	V	0.27%	4.60E-02
1423	p9	28	T	CTG	L	C	CCG	P	0.54%	1.27E-03
1788	p5a	57	T	TCG	S	C	CCG	P	0.73%	6.42E-04
1915	p5a	99	T	GTG	V	C	GCG	A	0.36%	1.75E-02
1976	p5a	119	G	CGG	R	A	CGA	R	0.37%	1.06E-02
2065	p5a	149	T	ATC	I	C	ACC	T	0.32%	2.86E-02
2634			G			A			0.58%	7.88E-07
2712			G			A			0.27%	1.73E-02
2751			T			C			0.25%	4.80E-02
2764			T			C			1.14%	2.60E-12
2781			C			T			0.28%	4.60E-02
3397	p10	27	A	GTA	V	G	GTG	V	0.38%	1.75E-02
3644			T			C			0.26%	4.60E-02
3771			C			T			43.35 %	0.00E+00
3903			T			C			0.72%	2.24E-06
3931	p6	6	C	TCG	S	T	TTG	L	0.26%	4.60E-02
4005	p6	31	C	CCG	P	T	TCG	S	0.31%	7.27E-03
4006	p6	31	C	CCG	P	T	CTG	L	0.24%	3.17E-02
4190	p6	92	T	TCT	S	C	TCC	S	0.35%	7.70E-05
4286	p6	124	A	GAA	E	G	GAG	E	1.18%	3.95E-13
4447	p3	8	A	GAG	E	G	GGG	G	0.44%	3.66E-03
4563	p3	47	A	ACT	T	G	GCT	A	0.62%	4.29E-05
4894	p3	157	T	CTG	L	G	CGG	R	0.37%	5.86E-03
5263	p3	280	T	GTC	V	C	GCC	A	0.41%	2.19E-03
5967	p3	515	A	AGT	S	G	GGT	G	0.34%	1.92E-02
6241	p3	606	T	GTC	V	C	GCC	A	0.33%	2.94E-03
6380			T			C			0.60%	2.60E-07
6724			T			C			0.40%	3.26E-02
6873			T			C			1.11%	1.43E-06
7165			C			T			0.19%	4.00E-02
7272			T			C			0.33%	1.22E-04
7485	p7	5	T	CTG	L	C	CCG	P	0.65%	8.68E-09

7489	p7	6	C	GTC	V	T	GTT	V	0.27%	3.17E-02
7490	p7	7	C	CCT	P	T	TCT	S	0.27%	3.17E-02
8231	p2	92	A	AAC	N	G	AGC	S	0.27%	1.73E-02
8318	p2	121	C	TCT	S	T	TTT	F	0.25%	1.73E-02
8429	p2	158	C	TCT	S	T	TTT	F	0.30%	4.64E-03
8639	p2	228	T	GTC	V	C	GCC	A	0.23%	4.80E-02
8966	p2	337	T	CTG	L	C	CCG	P	0.32%	1.54E-02
9334	p2	460	T	TGG	W	C	CGG	R	0.28%	1.32E-02
9642	p2	562	T	TGT	C	C	TGC	C	0.31%	2.44E-02
9799	p2	615	C	CTC	L	T	TTC	F	0.26%	2.44E-02
10069	p4	38	C	ACT	T	T	ATT	I	0.30%	3.83E-02
10375	p4	140	T	CTC	L	C	CCC	P	0.32%	9.55E-03
10495	p4	180	T	CTC	L	C	CCC	P	0.34%	5.87E-03
10827	p4	291	G	GAT	D	T	TAT	Y	0.29%	4.60E-02
11270	p1	102	C	ATC	I	T	ATT	I	0.32%	4.60E-02
11271	p1	103	T	TGG	W	C	CGG	R	0.42%	1.06E-02
11359	p1	132	T	CTG	L	C	CCG	P	0.41%	1.06E-02
12175	p1	404	T	ATC	I	C	ACC	T	0.64%	1.68E-03
12426	p1	488	G	GCG	A	A	ACG	T	0.44%	1.94E-02
12593	p1	543	T	CCT	P	C	CCC	P	0.70%	3.13E-03
13007	p1	681	T	ATT	I	C	ATC	I	1.15%	9.51E-04

Appendix 1.6 SNPs above 0.1% called by VarScan for sample phi6-G515S on PA. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP, p-value: p-value of SNP detection. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq	P-value
229			G			A			0.25%	2.15E-03
625	p8	107	A	GAA	E	G	GAG	E	1.07%	1.11E-20
728	p8	142	C	CTG	L	T	TTG	L	0.69%	1.93E-11
779	p12	9	T	CTC	L	C	CCC	P	0.44%	9.58E-05
858	p12	35	G	ACG	T	T	ACT	T	0.23%	3.19E-02
1243	p12	164	G	GTT	V	T	TTT	F	0.32%	3.36E-04
1273	p12	174	T	TCG	S	C	CCG	P	0.18%	4.93E-02
1274	p12	174	C	TCG	S	T	TTG	L	0.34%	4.84E-05
1423	p9	28	T	CTG	L	C	CCG	P	0.26%	3.79E-03
1587	p9	83	C	CTG	L	T	TTG	L	0.68%	1.13E-09
1788	p5a	57	T	TCG	S	C	CCG	P	0.39%	7.06E-04
1886	p5a	89	G	CAG	Q	T	CAT	H	0.21%	4.64E-02
1915	p5a	99	T	GTG	V	C	GCG	A	0.23%	3.19E-02
1983	p5a	122	T	TGG	W	C	CGG	R	0.23%	2.61E-02
2351			T			C			0.21%	3.19E-02
2459			G			T			0.17%	4.80E-02
2592			G			T			0.23%	2.61E-02
2634			G			A			0.21%	2.50E-02
2712			G			A			0.20%	3.53E-02
2713			C			A			0.48%	7.78E-08
2769			A			G			0.78%	1.97E-09
3411	p10	32	T	GTC	V	C	GCC	A	1.09%	1.15E-13
3771			C			T			58.26%	0.00E+00
3818			T			C			0.29%	1.32E-02
3903			T			C			0.51%	4.48E-04
3987	p6	25	A	ATC	I	G	GTC	V	0.33%	9.54E-03
4232	p6	106	G	GTG	V	T	GTT	V	0.21%	4.64E-02
4286	p6	124	A	GAA	E	G	GAG	E	1.06%	1.70E-11
4527	p3	35	G	GAC	D	T	TAC	Y	0.56%	2.06E-06
4528	p3	35	A	GAC	D	C	GCC	A	10.15%	1.23E-176
4563	p3	47	A	ACT	T	G	GCT	A	1.78%	1.28E-26
4654	p3	77	T	ATC	I	C	ACC	T	0.24%	4.80E-02
4822	p3	133	C	GCG	A	T	GTG	V	83.05%	0.00E+00

4823	p3	133	G	GCG	A	A	GCA	A	0.74%	2.88E-07
4855	p3	144	A	AAG	K	G	AGG	R	0.70%	1.60E-07
4904	p3	160	T	AGT	S	C	AGC	S	0.49%	3.23E-05
5069	p3	215	G	CAG	Q	T	CAT	H	0.27%	2.44E-02
5141	p3	239	T	CTT	L	C	CTC	L	0.34%	6.52E-04
5147	p3	241	T	TAT	Y	C	TAC	Y	0.45%	1.24E-05
5967	p3	515	A	AGT	S	G	GGT	G	1.10%	3.72E-10
6143	p3	573	T	CCT	P	C	CCC	P	0.64%	2.61E-10
6241	p3	606	T	GTC	V	C	GCC	A	0.23%	2.16E-02
6272	p3	616	T	GCT	A	C	GCC	A	0.72%	6.32E-11
6380			T			C			0.22%	3.76E-02
6733			T			C			0.63%	3.76E-07
6872			A			C			0.69%	1.67E-06
6873			T			C			0.45%	7.61E-04
9673	p2	573	G	GAG	E	T	TAG	X	0.92%	3.10E-02
10918	p4	321	A	AAT	N	G	AGT	S	0.86%	1.55E-02
12593	p1	543	T	CCT	P	C	CCC	P	1%	1.91E-03
12722	p1	586	G	GAG	E	A	GAA	E	1.59%	4.69E-04
13007	p1	681	T	ATT	I	C	ATC	I	0.67%	3.10E-02

Appendix 2.1 SNPs above 1% called by VarScan for sample S/P Day 30 lineage 1. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
5882	p3	486	A	GCA	A	G	GCG	A	39.51%
6537	p13	26	T	GGT	G	C	GGC	G	35.43%
3883			G			A			21.16%
5035	p3	204	T	GTC	V	C	GCC	A	12.98%
4410	p6	166	A	AGT	S	G	GGT	G	12.87%
3779			A			G			11.28%
12546	p1	528	A	ATC	I	C	CTC	L	10.19%
2355			G			A			7.58%
6224	p3	600	T	CGT	R	C	CGC	R	7.02%
4093	p6	60	T	ATC	I	C	ACC	T	6.66%
1874	p5a	85	T	GCT	A	C	GCC	A	6.62%
7976	p2	7	C	GCG	A	T	GTG	V	6.12%
3701			G			A			4.42%
3644			T			C			4.05%
3217			T			G			3.51%
4301	p6	129	A	GTA	V	C	GTC	V	3.42%
10760	p4	268	G	CGG	R	C	CGC	R	3.39%
2768			G			A			3.34%
4128	p6	72	C	CTG	L	T	TTG	L	3.11%
1143	p12	130	A	GCA	A	G	GCG	A	2.88%
5842	p3	473	A	AAC	N	G	AGC	S	2.86%
5161	p3	246	G	AGT	S	C	ACT	T	2.82%
2764			T			C			2.78%
2441			A			G			2.45%
9917	p2	654	A	AAG	K	G	AGG	R	2.45%
1826	p5a	69	A	GAA	E	G	GAG	E	2.36%
10233	p4	93	C	CTG	L	T	TTG	L	2.34%
5788	p3	455	A	AAG	K	G	AGG	R	2.24%

9444	p2	496	C	GCC	A	T	GCT	A	2.18%
2567			G			C			2.04%
13156	p1	731	T	GTT	V	C	GCT	A	2.02%
1230	p12	159	T	CGT	R	C	CGC	R	1.83%
1236	p12	161	T	TGT	C	C	TGC	C	1.82%
10454	p4	166	T	ACT	T	G	ACG	T	1.63%
9999	p4	15	A	ATC	I	G	GTC	V	1.60%
2267	p5a	216	A	AAA	K	G	AAG	K	1.55%
1915	p5a	99	T	GTG	V	C	GCG	A	1.52%
4820	p3	132	G	TCG	S	C	TCC	S	1.51%
13007	p1	681	T	ATT	I	C	ATC	I	1.49%
11756	p1	264	C	TCC	S	T	TCT	S	1.47%
69			T			C			1.40%
5286	p3	288	T	TTG	L	C	CTG	L	1.36%
9480	p2	508	T	CGT	R	C	CGC	R	1.26%
3903			T			C			1.18%
8479	p2	175	G	GCG	A	A	ACG	T	1.06%
3600			T			C			1.04%
6873			T			C			1.01%

Appendix 2.2 SNPs above 1% called by VarScan for sample S/P Day 30 lineage 2. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
510	p8	69	A	CAG	Q	G	CGG	R	54.12%
9879	p2	641	T	AAT	N	C	AAC	N	30.80%
6508	p13	17	G	GTT	V	C	CTT	L	11.95%
4563	p3	47	A	ACT	T	T	TCT	S	5.93%
5501	p3	359	T	GTT	V	C	GTC	V	5.44%
6652	p13	65	A	ACG	T	G	GCG	A	4.64%
6008	p3	528	C	GGC	G	T	GGT	G	4.01%
1393	p9	18	C	GCC	A	T	GTC	V	3.88%
3608			A			T			3.75%
6209	p3	595	T	CGT	R	C	CGC	R	3.63%
4703	p3	93	T	GAT	D	C	GAC	D	3.60%
7453			T			C			3.57%
11047	p1	28	A	CAG	Q	G	CGG	R	3.56%
6879			A			C			3.34%
1173	p12	140	T	GTT	V	C	GTC	V	3.09%
1619			C			T			2.85%
11876	p1	304	C	GCC	A	T	GCT	A	2.66%
6843			C			T			2.41%
1442	p9	34	T	CCT	P	C	CCC	P	2.15%
7927	p7	152	A	CAA	Q	G	CAG	Q	2.08%
406	p8	34	C	TCC	S	T	TCT	S	2.05%
12836	p1	624	C	CTC	L	T	CTT	L	1.85%
8121	p2	55	G	AAG	K	A	AAA	K	1.74%
10225	p4	90	T	GTC	V	C	GCC	A	1.70%
6297	p3	625	T	TTG	L	C	CTG	L	1.68%
9732	p2	592	T	CTT	L	C	CTC	L	1.59%
3903			T			C			1.59%
2764			T			C			1.55%

7351			G			A			1.46%
7204			T			C			1.43%
1895	p5a	92	T	ATT	I	C	ATC	I	1.38%
1929	p5a	104	G	GAA	E	C	CAA	Q	1.38%
4247	p6	111	T	GCT	A	C	GCC	A	1.33%
6730			A			G			1.26%
6877			G			A			1.21%
10019	p4	21	T	GAT	D	C	GAC	D	1.18%
5310	p3	296	G	GGG	G	A	AGG	R	1.17%
6873			T			C			1.13%
6749			T			C			1.12%
8514	p2	186	G	AAG	K	A	AAA	K	1.10%
10607	p4	217	G	TCG	S	A	TCA	S	1.08%
10164	p4	70	C	CTG	L	T	TTG	L	1.04%

Appendix 2.3 SNPs above 1% called by VarScan for sample S/P Day 30 lineage 3. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
4913	p3	163	T	GCT	A	C	GCC	A	16.09%
12128	p1	388	C	GAC	D	T	GAT	D	11.62%
10722	p4	256	T	TTG	L	C	CTG	L	10.49%
1820	p5a	67	A	ACA	T	G	ACG	T	10.33%
3145			A			C			8.01%
4410	p6	166	A	AGT	S	G	GGT	G	7.83%
5471	p3	349	T	CCT	P	C	CCC	P	6.93%
12767	p1	601	C	CGC	R	T	CGT	R	4.54%
1823	p5a	68	T	GCT	A	A	GCA	A	4.42%
2567			G			A			4.12%
6904			A			C			3.54%
6537	p13	26	T	GGT	G	C	GGC	G	3.06%
502	p8	66	C	GCC	A	T	GCT	A	3.01%
5171	p3	249	T	CCT	P	C	CCC	P	2.65%
2236	p5a	206	C	GCG	A	T	GTG	V	2.56%
7213			T			C			2.47%
123			A			G			2.42%
6570	p13	37	T	AGT	S	C	AGC	S	2.41%
1938	p5a	107	G	GTT	V	A	ATT	I	2.33%
2369			A			G			2.25%
3878			A			G			2.13%
4143	p6	77	T	TCG	S	C	CCG	P	2.12%
3189			T			C			2.02%
8382	p2	142	T	GTT	V	C	GTC	V	1.99%
510	p8	69	A	CAG	Q	G	CGG	R	1.88%
1855	p5a	79	A	CAA	Q	G	CGA	R	1.84%
8286	p2	110	T	CGT	R	C	CGC	R	1.73%
3526			C			T			1.72%

1827	p5a	70	G	GTC	V	A	ATC	I	1.72%
6435			C			T			1.64%
2486			T			C			1.63%
715	p8	137	C	GGC	G	T	GGT	G	1.54%
8287	p2	111	G	GCC	A	A	ACC	T	1.47%
2121	p5a	168	G	GTT	V	T	TTT	F	1.42%
7810	p7	113	T	GAT	D	C	GAC	D	1.35%
3767			T			C			1.31%
7062			T			C			1.24%
970	p12	73	G	GGC	G	A	AGC	S	1.18%
10262	p4	102	A	GGA	G	G	GGG	G	1.15%
4985	p3	187	A	AAA	K	G	AAG	K	1.12%
6471	p13	4	A	CTA	L	C	CTC	L	1.10%
8628	p2	224	G	AAG	K	A	AAA	K	1.08%
3997	p6	28	A	AAG	K	G	AGG	R	1.08%
4113	p6	67	A	ACC	T	T	TCC	S	1.08%
6873			T			C			1.08%
2678			C			T			1.02%
9879	p2	641	T	AAT	N	C	AAC	N	1.02%
2390			A			G			1.01%

Appendix 2.4 SNPs above 1% called by VarScan for sample S/P Day 30 lineage 4. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
5936	p3	504	T	GCT	A	C	GCC	A	7.96%
1281	p12	176	T	TTT	F	C	TTC	F	7.65%
5351	p3	309	T	GAT	D	C	GAC	D	7.22%
3187			A			G			7.11%
6658	p13	67	T	TTG	L	C	CTG	L	6.79%
10727	p4	257	T	TGT	C	C	TGC	C	6.57%
4209	p6	99	A	ATG	M	G	GTG	V	5.96%
8589	p2	211	T	GCT	A	G	GCG	A	5.39%
12161	p1	399	T	GCT	A	C	GCC	A	4.28%
6221	p3	599	A	CCA	P	G	CCG	P	4.09%
2871			T			C			3.94%
2063	p5a	148	A	GCA	A	G	GCG	A	3.63%
7654	p7	61	C	CAC	H	T	CAT	H	3.46%
9005	p2	350	A	TAC	Y	G	TGC	C	3.31%
1403	p9	21	C	CGC	R	T	CGT	R	2.77%
5876	p3	484	A	AAA	K	C	AAC	N	2.74%
4547	p3	41	T	GCT	A	C	GCC	A	2.64%
3072			T			C			2.48%
3550			C			T			2.45%
10631	p4	225	C	TGC	C	T	TGT	C	2.34%
7897	p7	142	G	ACG	T	A	ACA	T	2.16%
9281	p2	442	A	CAA	Q	G	CGA	R	2.11%
5615	p3	397	T	GGT	G	C	GGC	G	2.02%
4119	p6	69	A	ACG	T	G	GCG	A	2.01%
8229	p2	91	G	ATG	M	A	ATA	I	1.99%
4697	p3	91	A	ACA	T	G	ACG	T	1.81%
837	p12	28	A	GAA	E	G	GAG	E	1.78%
1014	p12	87	A	GAA	E	G	GAG	E	1.76%

2373			T			C			1.72%
7936	p7	155	T	GAT	D	C	GAC	D	1.51%
6102	p3	560	G	GCT	A	A	ACT	T	1.44%
8550	p2	198	T	GGT	G	C	GGC	G	1.43%
4780	p3	119	C	GCT	A	T	GTT	V	1.41%
3411	p10	32	T	GTC	V	G	GGC	G	1.37%
2232	p5a	205	G	GCG	A	A	ACG	T	1.28%
13004	p1	680	G	ATG	M	A	ATA	I	1.27%
13007	p1	681	T	ATT	I	C	ATC	I	1.23%
8613	p2	219	A	GGA	G	G	GGG	G	1.21%
7591	p7	40	C	CTC	L	T	CTT	L	1.20%
1136	p12	128	A	CAG	Q	T	CTG	L	1.14%
5960	p3	512	A	CAA	Q	G	CAG	Q	1.14%
5273	p3	283	T	CAT	H	C	CAC	H	1.13%
3903			T			C			1.12%
1761	p5a	48	T	TTG	L	C	CTG	L	1.01%
3549			G			A			1%

Appendix 2.5 SNPs above 1% called by VarScan for sample G/P Day 30 lineage 1. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
304			C			T			30.37%
3933	p6	7	T	TTG	L	C	CTG	L	15.48%
4447	p3	8	G	GGG	G	A	GAG	E	12.16%
4181	p6	89	G	ACG	T	A	ACA	T	9.90%
957	p12	68	C	CCC	P	T	CCT	P	9.36%
12950	p1	662	C	GCC	A	T	GCT	A	6.66%
13184	p1	740	C	CGC	R	T	CGT	R	5.94%
6774			T			C			5.09%
13097	p1	711	C	AGC	S	T	AGT	S	3.94%
12134	p1	390	G	GAG	E	A	GAA	E	3.85%
360	p8	19	C	GCG	A	T	GTG	V	3.37%
7454			T			C			3.10%
5393	p3	323	T	AAT	N	C	AAC	N	3.08%
3203			T			C			2.87%
13073	p1	703	C	GCC	A	T	GCT	A	2.75%
6537	p13	26	T	GGT	G	C	GGC	G	2.74%
3827			T			C			2.62%
786	p12	11	T	CCT	P	C	CCC	P	2.48%
1712	p5a	31	T	CGT	R	C	CGC	R	2.36%
2764			T			C			2.35%
7807	p7	112	C	AAC	N	T	AAT	N	2.23%
510	p8	69	A	CAG	Q	G	CGG	R	2.19%
5882	p3	486	A	GCA	A	G	GCG	A	2.18%
10016	p4	20	C	CTC	L	T	CTT	L	2.15%
11054	p1	30	C	TCC	S	T	TCT	S	2.01%
4367	p6	151	C	GCC	A	T	GCT	A	2%
2460			C			T			1.88%
3374	p10	20	A	ACC	T	G	GCC	A	1.85%

3847			A			G			1.68%
12251	p1	429	A	ACA	T	G	ACG	T	1.61%
13007	p1	681	T	ATT	I	C	ATC	I	1.54%
3354	p10	13	C	TCT	S	T	TTT	F	1.52%
5909	p3	495	C	CAC	H	T	CAT	H	1.51%
5462	p3	346	G	GCG	A	A	GCA	A	1.40%
1259	p12	169	A	GAT	D	G	GGT	G	1.35%
8137	p2	61	T	TAT	Y	G	GAT	D	1.30%
3903			T			C			1.24%
13190	p1	742	A	GAA	E	G	GAG	E	1.15%
11658	p1	232	G	GCC	A	A	ACC	T	1.09%
12701	p1	579	G	CCG	P	C	CCC	P	1.06%
12806	p1	614	C	CTC	L	T	CTT	L	1.05%
3059			T			C			1.04%
6873			T			C			1.03%
3570			G			A			1%

Appendix 2.6 SNPs above 1% called by VarScan for sample G/P Day 30 lineage 2. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
5364	p3	314	G	GCT	A	A	ACT	T	12.78%
3334	p10	6	T	GAT	D	G	GAG	E	10.66%
6520	p13	21	G	GCA	A	A	ACA	T	9.72%
9923	p2	656	A	GAG	E	G	GGG	G	7.16%
1997	p5a	126	T	CGT	R	C	CGC	R	7.05%
1788	p5a	57	T	TCG	S	G	GCG	A	5.82%
87			A			G			5.67%
11787	p1	275	T	TTG	L	C	CTG	L	5.47%
5162	p3	246	T	AGT	S	C	AGC	S	5.41%
4271	p6	119	T	GCT	A	C	GCC	A	4.31%
3473			T			C			4.31%
2466			A			G			3.98%
4447	p3	8	G	GGG	G	A	GAG	E	3.44%
10841	p4	295	C	AGC	S	T	AGT	S	3.14%
12950	p1	662	C	GCC	A	T	GCT	A	2.82%
4598	p3	58	T	GCT	A	C	GCC	A	2.77%
7352			C			T			2.54%
4463	p3	13	C	GCC	A	T	GCT	A	2.50%
12011	p1	349	C	GTC	V	T	GTT	V	2.44%
1946	p5a	109	T	GGT	G	C	GGC	G	2.37%
3896			C			T			2.36%
3544			G			A			2.32%
4283	p6	123	C	ACC	T	T	ACT	T	2.28%
510	p8	69	A	CAG	Q	G	CGG	R	2.19%
6350	p3	642	G	CTG	L	C	CTC	L	2.04%
10457	p4	167	G	CTG	L	C	CTC	L	2.01%
3859			A			G			1.94%
2255	p5a	212	C	AAC	N	T	AAT	N	1.76%

2764			T			C			1.70%
5669	p3	415	C	CTC	L	T	CTT	L	1.63%
9190	p2	412	C	CTG	L	T	TTG	L	1.59%
2751			T			C			1.53%
11603	p1	213	G	ACG	T	A	ACA	T	1.44%
10947	p4	331	T	TCT	S	G	GCT	A	1.37%
3134			T			G			1.36%
2663			C			G			1.32%
10646	p4	230	C	GTC	V	T	GTT	V	1.29%
9194	p2	413	T	GTG	V	C	GCG	A	1.23%
12096	p1	378	C	CTG	L	T	TTG	L	1.16%
11495	p1	177	C	ACC	T	T	ACT	T	1.15%
5374	p3	317	T	ATT	I	G	AGT	S	1.12%
9774	p2	606	G	ATG	M	A	ATA	I	1.07%
4343	p6	143	A	GCA	A	G	GCG	A	1.05%
2525			T			C			1.05%
3561			T			C			1.05%
6873			T			C			1.02%
3903			T			C			1.01%
7852	p7	127	G	TCG	S	A	TCA	S	1%

Appendix 2.7 SNPs above 1% called by VarScan for sample G/P Day 30 lineage 3. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
3350	p10	12	T	TTT	F	C	CTT	L	25.12%
2992			A			G			9.35%
12818	p1	618	C	CGC	R	T	CGT	R	9%
110			C			T			6.98%
5358	p3	312	G	GTT	V	A	ATT	I	5.89%
7498	p7	9	G	CTG	L	A	CTA	L	5.40%
1786	p5a	56	A	CAG	Q	G	CGG	R	4.86%
12821	p1	619	A	GAA	E	G	GAG	E	4.84%
3825			A			G			4.10%
10832	p4	292	T	CGT	R	C	CGC	R	4.10%
1272	p12	173	T	ACT	T	C	ACC	T	3.67%
7882	p7	137	T	CGT	R	G	CGG	R	3.36%
3368	p10	18	G	GCG	A	A	ACG	T	3.32%
4344	p6	144	C	CTC	L	T	TTC	F	3.25%
9711	p2	585	T	CGT	R	C	CGC	R	3.20%
2533			T			C			3.20%
791	p12	13	T	GTT	V	C	GCT	A	3.04%
10103	p4	49	G	ATG	M	A	ATA	I	3.02%
6771			T			G			2.95%
4586	p3	54	C	ATC	I	T	ATT	I	2.71%
9999	p4	15	A	ATC	I	G	GTC	V	2.62%
1526	p9	62	T	GCT	A	C	GCC	A	2.61%
8298	p2	114	C	GAC	D	T	GAT	D	2.50%
9306	p2	450	C	ATC	I	T	ATT	I	2.37%
12164	p1	400	G	ACG	T	A	ACA	T	2%
12554	p1	530	T	GTT	V	C	GTC	V	1.96%
1232	p12	160	A	CAG	Q	G	CGG	R	1.94%
4029	p6	39	A	ATC	I	G	GTC	V	1.82%

7288			C			T			1.80%
4571	p3	49	T	GGT	G	G	GGG	G	1.53%
5561	p3	379	G	GCG	A	A	GCA	A	1.40%
3661			C			T			1.37%
6766			A			G			1.34%
10373	p4	139	G	AAG	K	A	AAA	K	1.23%
2745			A			G			1.18%
13168	p1	735	T	ATG	M	C	ACG	T	1.05%

Appendix 2.8 SNPs above 1% called by VarScan for sample G/P Day 30 lineage 4. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
3293			A			T			31.17%
7328			A			G			15.90%
7212			T			C			13.62%
9321	p2	455	T	GAT	D	C	GAC	D	13.61%
3779			A			G			9.99%
5618	p3	398	T	ACT	T	C	ACC	T	8.13%
9864	p2	636	C	GCC	A	T	GCT	A	7.63%
7630	p7	53	C	AAC	N	T	AAT	N	7.60%
4529	p3	35	C	GAC	D	T	GAT	D	4.08%
1692	p5a	25	G	GTA	V	A	ATA	I	3.53%
6844			A			G			3.18%
3141			T			C			2.99%
1556	p9	72	C	TAC	Y	T	TAT	Y	2.95%
4595	p3	57	C	CGC	R	T	CGT	R	2.93%
8652	p2	232	A	GAA	E	C	GAC	D	2.69%
4456	p3	11	G	GGT	G	A	GAT	D	2.61%
3535			T			C			2.42%
4323	p6	137	A	ATC	I	G	GTC	V	2.33%
9099	p2	381	T	GGT	G	C	GGC	G	2.30%
3569			A			G			2.20%
2619			A			G			2.16%
9944	p2	663	T	ATG	M	C	ACG	T	1.78%
6887			G			A			1.69%
4410	p6	166	A	AGT	S	G	GGT	G	1.64%
7329			A			G			1.59%
9315	p2	453	T	TCT	S	C	TCC	S	1.58%
2475			G			A			1.52%
12717	p1	585	A	ACC	T	G	GCC	A	1.48%

6752			T			C			1.48%
8802	p2	282	T	GCT	A	C	GCC	A	1.44%
3879			A			C			1.42%
5180	p3	252	C	GGC	G	G	GGG	G	1.40%
8337	p2	127	C	CTC	L	T	CTT	L	1.39%
3814			A			C			1.38%
3008			A			G			1.31%
244			T			C			1.30%
6803			C			T			1.29%
5990	p3	522	C	CGC	R	T	CGT	R	1.28%
6617	p13	53	C	GCC	A	T	GTC	V	1.27%
7687	p7	72	T	GCT	A	C	GCC	A	1.27%
5742	p3	440	G	GGG	G	A	AGG	R	1.23%
6548	p13	30	C	GCG	A	T	GTG	V	1.20%
6477	p13	6	T	ATT	I	C	ATC	I	1.15%
9673	p2	573	G	GAG	E	T	TAG	X	1.14%
2684			A			G			1.10%
2144	p5a	175	T	TAT	Y	C	TAC	Y	1.07%
370	p8	22	T	CCT	P	C	CCC	P	1.06%
3119			G			A			1.03%
5330	p3	302	G	CGG	R	A	CGA	R	1.03%

Appendix 2.9 SNPs above 1% called by VarScan for sample G/P Day 29 lineage 1. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
304			C			T			32.21%
3933	p6	7	T	TTG	L	C	CTG	L	31.58%
957	p12	68	C	CCC	P	T	CCT	P	10.42%
4181	p6	89	G	ACG	T	A	ACA	T	10.28%
6774			T			C			6.09%
13097	p1	711	C	AGC	S	T	AGT	S	5.15%
13184	p1	740	C	CGC	R	T	CGT	R	4.85%
360	p8	19	C	GCG	A	T	GTG	V	4.80%
5393	p3	323	T	AAT	N	C	AAC	N	3.92%
786	p12	11	T	CCT	P	C	CCC	P	3.35%
3203			T			C			3.34%
7454			T			C			3.09%
3827			T			C			3.06%
7807	p7	112	C	AAC	N	T	AAT	N	3%
12134	p1	390	G	GAG	E	A	GAA	E	2.95%
1712	p5a	31	T	CGT	R	C	CGC	R	2.85%
4367	p6	151	C	GCC	A	T	GCT	A	2.60%
5462	p3	346	G	GCG	A	A	GCA	A	2.56%
12806	p1	614	C	CTC	L	T	CTT	L	2.51%
5909	p3	495	C	CAC	H	T	CAT	H	2.45%
2460			C			T			2.41%
13073	p1	703	C	GCC	A	T	GCT	A	2.25%
10016	p4	20	C	CTC	L	T	CTT	L	2.06%
11054	p1	30	C	TCC	S	T	TCT	S	2.05%
3374	p10	20	A	ACC	T	G	GCC	A	2%
3847			A			G			1.73%
3354	p10	13	C	TCT	S	T	TTT	F	1.70%
12251	p1	429	A	ACA	T	G	ACG	T	1.57%

1259	p12	169	A	GAT	D	G	GGT	G	1.57%
13190	p1	742	A	GAA	E	G	GAG	E	1.55%
13007	p1	681	T	ATT	I	C	ATC	I	1.43%
2244	p5a	209	G	GCT	A	A	ACT	T	1.41%
556	p8	84	G	CTG	L	C	CTC	L	1.40%
2764			T			C			1.33%
6011	p3	529	T	GGT	G	G	GGG	G	1.28%
8137	p2	61	T	TAT	Y	G	GAT	D	1.25%
6873			T			C			1.23%
1393	p9	18	C	GCC	A	T	GTC	V	1.19%
9673	p2	573	G	GAG	E	T	TAG	X	1.16%
3059			T			C			1.14%
8538	p2	194	C	CAC	H	T	CAT	H	1.09%
4547	p3	41	T	GCT	A	G	GCG	A	1.04%
12317	p1	451	C	CTC	L	T	CTT	L	1.01%
11658	p1	232	G	GCC	A	A	ACC	T	1%

Appendix 2.10 SNPs above 1% called by VarScan for sample G/P Day 29 lineage 2. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
5364	p3	314	G	GCT	A	A	ACT	T	17.10%
6520	p13	21	G	GCA	A	A	ACA	T	10.83%
3334	p10	6	T	GAT	D	G	GAG	E	10.68%
1997	p5a	126	T	CGT	R	C	CGC	R	7.46%
87			A			G			6.44%
9923	p2	656	A	GAG	E	G	GGG	G	6.43%
11787	p1	275	T	TTG	L	C	CTG	L	5.06%
3473			T			C			4.56%
5162	p3	246	T	AGT	S	C	AGC	S	4.54%
1788	p5a	57	T	TCG	S	G	GCG	A	4.29%
2466			A			G			3.99%
4271	p6	119	T	GCT	A	C	GCC	A	3.62%
4463	p3	13	C	GCC	A	T	GCT	A	3.01%
4598	p3	58	T	GCT	A	C	GCC	A	2.98%
7352			C			T			2.84%
10841	p4	295	C	AGC	S	T	AGT	S	2.73%
5669	p3	415	C	CTC	L	T	CTT	L	2.33%
4283	p6	123	C	ACC	T	T	ACT	T	2.30%
3544			G			A			2.04%
3896			C			T			2%
12011	p1	349	C	GTC	V	T	GTT	V	1.96%
2525			T			C			1.86%
9673	p2	573	G	GAG	E	T	TAG	X	1.77%
10457	p4	167	G	CTG	L	C	CTC	L	1.69%
1946	p5a	109	T	GGT	G	C	GGC	G	1.68%
11603	p1	213	G	ACG	T	A	ACA	T	1.65%
9194	p2	413	T	GTG	V	C	GCG	A	1.44%
6350	p3	642	G	CTG	L	C	CTC	L	1.36%

2255	p5a	212	C	AAC	N	T	AAT	N	1.30%
9786	p2	610	C	GCC	A	T	GCT	A	1.24%
510	p8	69	A	CAG	Q	G	CGG	R	1.20%
2764			T			C			1.13%
12875	p1	637	T	TAT	Y	C	TAC	Y	1.11%
8370	p2	138	T	GAT	D	C	GAC	D	1.11%
5374	p3	317	T	ATT	I	G	AGT	S	1.10%
6873			T			C			1.09%
9190	p2	412	C	CTG	L	T	TTG	L	1.07%
12096	p1	378	C	CTG	L	T	TTG	L	1.05%
4074	p6	54	T	TTC	F	C	CTC	L	1.04%
3859			A			G			1.03%
2751			T			C			1.03%
9774	p2	606	G	ATG	M	A	ATA	I	1.02%
5700	p3	426	T	TTG	L	C	CTG	L	1.02%
4343	p6	143	A	GCA	A	G	GCG	A	1.01%
3134			T			G			1.01%

Appendix 2.11 SNPs above 1% called by VarScan for sample G/P Day 29 lineage 3. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
3350	p10	12	T	TTT	F	C	CTT	L	27.38%
12818	p1	618	C	CGC	R	T	CGT	R	9.21%
2992			A			G			7.72%
7498	p7	9	G	CTG	L	A	CTA	L	7.08%
110			C			T			6.64%
1272	p12	173	T	ACT	T	C	ACC	T	5.23%
1786	p5a	56	A	CAG	Q	G	CGG	R	4.59%
12821	p1	619	A	GAA	E	G	GAG	E	4.41%
10832	p4	292	T	CGT	R	C	CGC	R	4.05%
2533			T			C			4.04%
9999	p4	15	A	ATC	I	G	GTC	V	3.81%
3825			A			G			3.72%
9306	p2	450	C	ATC	I	T	ATT	I	3.55%
5358	p3	312	G	GTT	V	A	ATT	I	3.21%
6771			T			G			3.16%
8298	p2	114	C	GAC	D	T	GAT	D	3.15%
3368	p10	18	G	GCG	A	A	ACG	T	2.97%
791	p12	13	T	GTT	V	C	GCT	A	2.84%
7882	p7	137	T	CGT	R	G	CGG	R	2.80%
4344	p6	144	C	CTC	L	T	TTC	F	2.73%
7059			A			G			2.55%
10103	p4	49	G	ATG	M	A	ATA	I	2.33%
4571	p3	49	T	GGT	G	G	GGG	G	2.24%
5561	p3	379	G	GCG	A	A	GCA	A	1.98%
4586	p3	54	C	ATC	I	T	ATT	I	1.76%
1526	p9	62	T	GCT	A	C	GCC	A	1.76%
12554	p1	530	T	GTT	V	C	GTC	V	1.61%
9711	p2	585	T	CGT	R	C	CGC	R	1.61%

5375	p3	317	T	ATT	I	C	ATC	I	1.60%
40			G			A			1.47%
7288			C			T			1.43%
4029	p6	39	A	ATC	I	G	GTC	V	1.41%
3661			C			T			1.36%
10904	p4	316	T	GAT	D	C	GAC	D	1.30%
1802	p5a	61	C	AGC	S	T	AGT	S	1.15%
8982	p2	342	T	TGT	C	C	TGC	C	1.14%
4796	p3	124	C	GGC	G	T	GGT	G	1.09%
2497			A			G			1.04%

Appendix 2.12 SNPs above 1% called by VarScan for sample G/P Day 29 lineage 4. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
3293			A			T			31.39%
7328			A			G			15.03%
9321	p2	455	T	GAT	D	C	GAC	D	13.19%
7212			T			C			12.34%
3779			A			G			9.10%
5618	p3	398	T	ACT	T	C	ACC	T	8.67%
9864	p2	636	C	GCC	A	T	GCT	A	8.20%
7630	p7	53	C	AAC	N	T	AAT	N	7.38%
1692	p5a	25	G	GTA	V	A	ATA	I	3.93%
4529	p3	35	C	GAC	D	T	GAT	D	3.59%
6844			A			G			3.28%
4595	p3	57	C	CGC	R	T	CGT	R	2.95%
6887			G			A			2.85%
3569			A			G			2.78%
3141			T			C			2.74%
8652	p2	232	A	GAA	E	C	GAC	D	2.72%
4456	p3	11	G	GGT	G	A	GAT	D	2.70%
7329			A			G			2.45%
9099	p2	381	T	GGT	G	C	GGC	G	2.18%
3879			A			C			2.04%
1556	p9	72	C	TAC	Y	T	TAT	Y	2.02%
2619			A			G			1.93%
4323	p6	137	A	ATC	I	G	GTC	V	1.92%
4410	p6	166	A	AGT	S	G	GGT	G	1.89%
2475			G			A			1.86%
8337	p2	127	C	CTC	L	T	CTT	L	1.74%
3535			T			C			1.70%
3008			A			G			1.69%

5990	p3	522	C	CGC	R	T	CGT	R	1.63%
6617	p13	53	C	GCC	A	T	GTC	V	1.60%
244			T			C			1.55%
9673	p2	573	G	GAG	E	T	TAG	X	1.50%
2684			A			G			1.49%
6477	p13	6	T	ATT	I	C	ATC	I	1.47%
9944	p2	663	T	ATG	M	C	ACG	T	1.46%
3814			A			C			1.45%
6752			T			C			1.45%
7687	p7	72	T	GCT	A	C	GCC	A	1.43%
5180	p3	252	C	GGC	G	G	GGG	G	1.42%
12717	p1	585	A	ACC	T	G	GCC	A	1.40%
5979	p3	519	A	ACT	T	G	GCT	A	1.40%
2144	p5a	175	T	TAT	Y	C	TAC	Y	1.38%
6131	p3	569	G	GCG	A	A	GCA	A	1.35%
3903			T			C			1.29%
6548	p13	30	C	GCG	A	T	GTG	V	1.23%
8802	p2	282	T	GCT	A	C	GCC	A	1.20%
9315	p2	453	T	TCT	S	C	TCC	S	1.19%
4487	p3	21	T	GGT	G	C	GGC	G	1.11%
5330	p3	302	G	CGG	R	A	CGA	R	1.09%
3119			G			A			1.07%
13007	p1	681	T	ATT	I	C	ATC	I	1%
1140	p12	129	T	TAT	Y	C	TAC	Y	1%

Appendix 2.13 SNPs above 1% called by VarScan for sample G/PT Day 30 lineage 1. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
4813	p3	130	A	CAG	Q	G	CGG	R	98.15%
6925			G			A			14.18%
6167	p3	581	A	GAA	E	G	GAG	E	10.55%
1358	p9	6	A	GTA	V	G	GTG	V	8.31%
10920	p4	322	T	TCC	S	C	CCC	P	7.16%
9585	p2	543	C	CGC	R	T	CGT	R	2.86%
8664	p2	236	G	ACG	T	A	ACA	T	2.70%
12582	p1	540	A	ATC	I	G	GTC	V	2.54%
837	p12	28	A	GAA	E	G	GAG	E	2.24%
670	p8	122	G	CTG	L	A	CTA	L	2.15%
3596			T			C			2.09%
10775	p4	273	T	CGT	R	C	CGC	R	1.98%
3803			T			C			1.98%
2764			T			C			1.93%
2781			C			T			1.88%
4806	p3	128	G	GGC	G	A	AGC	S	1.87%
8925	p2	323	T	GCT	A	C	GCC	A	1.85%
9673	p2	573	G	GAG	E	T	TAG	X	1.64%
2862			T			C			1.63%
4985	p3	187	A	AAA	K	G	AAG	K	1.56%
9450	p2	498	T	CTT	L	C	CTC	L	1.40%
1315	p12	188	A	ATG	M	G	GTG	V	1.38%
2489			A			G			1.34%
9828	p2	624	T	GCT	A	C	GCC	A	1.31%
11978	p1	338	G	GAG	E	A	GAA	E	1.21%
6873			T			C			1.17%
2291			T			C			1.16%
13007	p1	681	T	ATT	I	C	ATC	I	1.15%

2377			T			C			1.12%
12995	p1	677	G	AAG	K	A	AAA	K	1.10%
6128	p3	568	T	ACT	T	C	ACC	T	1.10%
8057	p2	34	A	AAA	K	C	ACA	T	1.05%
5726	p3	434	T	CGT	R	C	CGC	R	1.02%

Appendix 2.14 SNPs above 1% called by VarScan for sample G/PT Day 30 lineage 2. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
4813	p3	130	A	CAG	Q	G	CGG	R	79.01%
1378	p9	13	A	AAG	K	G	AGG	R	21.21%
2048	p5a	143	C	AAC	N	T	AAT	N	17.87%
4487	p3	21	T	GGT	G	C	GGC	G	16.06%
2229	p5a	204	G	GTC	V	A	ATC	I	8.20%
2777			C			T			7.69%
10268	p4	104	T	GCT	A	C	GCC	A	7.37%
12900	p1	646	A	ACT	T	G	GCT	A	6.47%
10457	p4	167	G	CTG	L	C	CTC	L	4.84%
5531	p3	369	C	ACC	T	T	ACT	T	4.24%
10134	p4	60	G	GCC	A	A	ACC	T	3.86%
4733	p3	103	C	TGC	C	T	TGT	C	3.38%
10796	p4	280	T	CCT	P	C	CCC	P	3.29%
1272	p12	173	T	ACT	T	G	ACG	T	3.02%
1322	p12	190	A	CAC	H	G	CGC	R	2.81%
6232	p3	603	T	GTC	V	C	GCC	A	2.63%
8046	p2	30	G	AAG	K	A	AAA	K	2.12%
2254	p5a	212	A	AAC	N	G	AGC	S	2.10%
11960	p1	332	A	AAA	K	G	AAG	K	2.01%
13248	p1	762	A	ACA	T	G	GCA	A	1.89%
9944	p2	663	T	ATG	M	C	ACG	T	1.89%
8322	p2	122	G	GAG	E	A	GAA	E	1.69%
2412			C			T			1.67%
3286			C			T			1.57%
8371	p2	139	C	CTA	L	T	TTA	L	1.49%
9673	p2	573	G	GAG	E	T	TAG	X	1.42%
10736	p4	260	C	GGC	G	T	GGT	G	1.28%
3431	p10	39	C	CTG	L	T	TTG	L	1.25%

6873			T			C			1.22%
4205	p6	97	C	CTC	L	T	CTT	L	1.21%
8631	p2	225	T	GAT	D	C	GAC	D	1.19%
7927	p7	152	A	CAA	Q	G	CAG	Q	1.17%
4067	p6	51	C	TCC	S	T	TCT	S	1.10%
5297	p3	291	G	CTG	L	T	CTT	L	1.06%
2764			T			C			1.04%
1190	p12	146	A	AAG	K	C	ACG	T	1.02%

Appendix 2.15 SNPs above 1% called by VarScan for sample G/PT Day 30 lineage 3. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
487	p8	61	C	AGC	S	T	AGT	S	86.36%
3702			G			A			84.41%
9394	p2	480	A	AAG	K	G	GAG	E	83.40%
10811	p4	285	T	ACT	T	C	ACC	T	82.08%
3695			G			A			10.89%
6877			G			A			7.74%
8577	p2	207	A	CAA	Q	G	CAG	Q	4.43%
10226	p4	90	C	GTC	V	T	GTT	V	4.34%
481	p8	59	C	ATC	I	T	ATT	I	3.35%
472	p8	56	C	CTC	L	T	CTT	L	2.32%
11764	p1	267	A	AAG	K	G	AGG	R	2.15%
6879			A			G			2.14%
9072	p2	372	T	GCT	A	C	GCC	A	2.10%
2521			A			G			1.99%
2638			T			C			1.82%
8805	p2	283	T	CCT	P	A	CCA	P	1.74%
6658	p13	67	T	TTG	L	C	CTG	L	1.56%
4301	p6	129	A	GTA	V	C	GTC	V	1.52%
1069	p12	106	A	AGC	S	G	GGC	G	1.42%
6851			T			C			1.41%
10805	p4	283	T	GCT	A	A	GCA	A	1.38%
9673	p2	573	G	GAG	E	T	TAG	X	1.37%
5839	p3	472	A	AAA	K	G	AGA	R	1.26%
1664	p5a	15	C	GCC	A	T	GCT	A	1.24%
7420			A			G			1.20%
13034	p1	690	C	GCC	A	T	GCT	A	1.15%
3988	p6	25	T	ATC	I	C	ACC	T	1.15%
6478	p13	7	A	AGC	S	T	TGC	C	1.14%

6245	p3	607	G	CAG	Q	A	CAA	Q	1.14%
2159	p5a	180	A	CAA	Q	G	CAG	Q	1.13%
2378			T			C			1.11%
2652			A			G			1.07%
2552			A			C			1%

Appendix 2.16 SNPs above 1% called by VarScan for sample G/PT Day 30 lineage 4. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
3664			T			C			91.47%
1353	p9	5	C	CTG	L	T	TTG	L	29.01%
1263	p12	170	A	GTA	V	G	GTG	V	12.79%
2620			A			G			8.94%
12516	p1	518	C	CTG	L	T	TTG	L	6.81%
4496	p3	24	C	GGC	G	T	GGT	G	6.78%
4860	p3	146	A	AAT	N	C	CAT	H	6.64%
6862			A			G			6.36%
10392	p4	146	A	ATT	I	G	GTT	V	5.82%
1268	p12	172	A	CAA	Q	G	CGA	R	5.48%
5894	p3	490	C	TCC	S	T	TCT	S	4.55%
5276	p3	284	C	GTC	V	T	GTT	V	4.07%
12338	p1	458	G	GCG	A	A	GCA	A	4%
1706	p5a	29	G	GAG	E	A	GAA	E	3.80%
6114	p3	564	T	TTG	L	C	CTG	L	3.29%
4088	p6	58	A	GGA	G	G	GGG	G	2.89%
6684			A			C			2.85%
6891			A			G			2.79%
4479	p3	19	G	GCA	A	T	TCA	S	2.67%
129			A			G			2.53%
1631	p5a	4	T	GAT	D	C	GAC	D	2.31%
8736	p2	260	T	GAT	D	C	GAC	D	2.25%
1843	p5a	75	C	GCG	A	T	GTG	V	2.13%
7895	p7	142	A	ACG	T	G	GCG	A	2%
6744			C			T			1.90%
6406			C			T			1.76%
484	p8	60	A	GGA	G	G	GGG	G	1.71%
5087	p3	221	C	GGC	G	T	GGT	G	1.68%

8571	p2	205	T	CGT	R	C	CGC	R	1.47%
2063	p5a	148	A	GCA	A	G	GCG	A	1.44%
3903			T			C			1.38%
2678			C			T			1.36%
2379			T			C			1.24%
2239	p5a	207	T	GTT	V	C	GCT	A	1.22%
9291	p2	445	G	GAG	E	A	GAA	E	1.19%
2311			T			C			1.18%
3690			G			T			1.17%
9673	p2	573	G	GAG	E	T	TAG	X	1.16%
2145	p5a	176	T	TTG	L	C	CTG	L	1.16%
1724	p5a	35	T	GAT	D	C	GAC	D	1.09%
8781	p2	275	T	GGT	G	C	GGC	G	1.04%
286			A			G			1%

Appendix 2.17 SNPs above 1% called by VarScan for sample G/PT Day 29 lineage 1. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
4813	p3	130	A	CAG	Q	G	CGG	R	96.89%
6925			G			A			15.53%
6167	p3	581	A	GAA	E	G	GAG	E	11.93%
10920	p4	322	T	TCC	S	C	CCC	P	8.90%
1358	p9	6	A	GTA	V	G	GTG	V	7.01%
9585	p2	543	C	CGC	R	T	CGT	R	3.90%
670	p8	122	G	CTG	L	A	CTA	L	3.28%
837	p12	28	A	GAA	E	G	GAG	E	3.23%
8664	p2	236	G	ACG	T	A	ACA	T	2.81%
9450	p2	498	T	CTT	L	C	CTC	L	2.54%
2781			C			T			2.49%
3596			T			C			2.46%
9673	p2	573	G	GAG	E	T	TAG	X	2.37%
2524			T			C			1.73%
8925	p2	323	T	GCT	A	C	GCC	A	1.66%
12848	p1	628	A	GTA	V	G	GTG	V	1.56%
12179	p1	405	C	GGC	G	T	GGT	G	1.52%
2862			T			C			1.51%
10775	p4	273	T	CGT	R	C	CGC	R	1.46%
4806	p3	128	G	GGC	G	A	AGC	S	1.35%
3803			T			C			1.35%
8057	p2	34	A	AAA	K	C	ACA	T	1.33%
3903			T			C			1.17%
2764			T			C			1.12%
2489			A			G			1.10%
6552	p13	31	T	TAT	Y	C	TAC	Y	1.10%
6873			T			C			1.06%
11978	p1	338	G	GAG	E	A	GAA	E	1.05%

12582	p1	540	A	ATC	I	G	GTC	V	1.02%
8313	p2	119	T	CCT	P	G	CCG	P	1.02%
13007	p1	681	T	ATT	I	C	ATC	I	1%

Appendix 2.18 SNPs above 1% called by VarScan for sample G/PT Day 29 lineage 2. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
4813	p3	130	A	CAG	Q	G	CGG	R	64.94%
4487	p3	21	T	GGT	G	C	GGC	G	25.12%
1378	p9	13	A	AAG	K	G	AGG	R	15.81%
2048	p5a	143	C	AAC	N	T	AAT	N	13.84%
10268	p4	104	T	GCT	A	C	GCC	A	7.66%
2229	p5a	204	G	GTC	V	A	ATC	I	7.35%
12900	p1	646	A	ACT	T	G	GCT	A	6.86%
6232	p3	603	T	GTC	V	C	GCC	A	6.65%
5531	p3	369	C	ACC	T	T	ACT	T	5.79%
2777			C			T			5.68%
10457	p4	167	G	CTG	L	C	CTC	L	5.09%
6923			T			C			4.88%
4733	p3	103	C	TGC	C	T	TGT	C	4.12%
10796	p4	280	T	CCT	P	C	CCC	P	3.10%
10134	p4	60	G	GCC	A	A	ACC	T	2.95%
10736	p4	260	C	GGC	G	T	GGT	G	2.92%
3286			C			T			2.85%
1322	p12	190	A	CAC	H	G	CGC	R	2.63%
11960	p1	332	A	AAA	K	G	AAG	K	2.11%
13248	p1	762	A	ACA	T	G	GCA	A	1.95%
9673	p2	573	G	GAG	E	T	TAG	X	1.95%
11951	p1	329	C	TCC	S	T	TCT	S	1.94%
2254	p5a	212	A	AAC	N	G	AGC	S	1.94%
8360	p2	135	T	ATG	M	C	ACG	T	1.62%
8322	p2	122	G	GAG	E	A	GAA	E	1.56%
8371	p2	139	C	CTA	L	T	TTA	L	1.53%
2412			C			T			1.40%
4205	p6	97	C	CTC	L	T	CTT	L	1.35%

6920			A			G			1.24%
9944	p2	663	T	ATG	M	C	ACG	T	1.22%
1242	p12	163	T	CAT	H	C	CAC	H	1.19%
1272	p12	173	T	ACT	T	G	ACG	T	1.18%
12821	p1	619	A	GAA	E	G	GAG	E	1.16%
6604	p13	49	T	TCT	S	C	CCT	P	1.15%
5934	p3	504	G	GCT	A	A	ACT	T	1.11%
6873			T			C			1.10%
5228	p3	268	T	GCT	A	C	GCC	A	1.04%
8046	p2	30	G	AAG	K	A	AAA	K	1.01%

Appendix 2.19 SNPs above 1% called by VarScan for sample G/PT Day 29 lineage 3. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
487	p8	61	C	AGC	S	T	AGT	S	79.83%
9394	p2	480	A	AAG	K	G	GAG	E	76.26%
10811	p4	285	T	ACT	T	C	ACC	T	74.75%
3702			G			A			73.20%
3695			G			A			16.97%
6877			G			A			9.55%
3361	p10	15	A	GAA	E	C	GAC	D	4.07%
9075	p2	373	T	CCT	P	C	CCC	P	3.97%
481	p8	59	C	ATC	I	T	ATT	I	3.96%
10226	p4	90	C	GTC	V	T	GTT	V	3.85%
8577	p2	207	A	CAA	Q	G	CAG	Q	3.67%
2521			A			G			3.28%
11102	p1	46	T	GCT	A	C	GCC	A	3.22%
472	p8	56	C	CTC	L	T	CTT	L	3.06%
9673	p2	573	G	GAG	E	T	TAG	X	2.92%
9072	p2	372	T	GCT	A	C	GCC	A	2.52%
6879			A			G			2.14%
7573	p7	34	C	CAC	H	T	CAT	H	2.06%
6851			T			C			1.85%
1664	p5a	15	C	GCC	A	T	GCT	A	1.68%
11764	p1	267	A	AAG	K	G	AGG	R	1.59%
2652			A			G			1.51%
4301	p6	129	A	GTA	V	C	GTC	V	1.43%
2552			A			C			1.33%
2159	p5a	180	A	CAA	Q	G	CAG	Q	1.26%
7744	p7	91	A	GCA	A	G	GCG	A	1.24%
8136	p2	60	G	CGG	R	A	CGA	R	1.23%
5839	p3	472	A	AAA	K	G	AGA	R	1.16%

6245	p3	607	G	CAG	Q	A	CAA	Q	1.10%
2638			T			C			1.10%
9019	p2	355	G	GTT	V	A	ATT	I	1.09%
5033	p3	203	T	CGT	R	C	CGC	R	1.06%
3381	p10	22	C	GCT	A	T	GTT	V	1.03%
730	p8	142	G	CTG	L	A	CTA	L	1.01%
2378			T			C			1.01%

Appendix 2.20 SNPs above 1% called by VarScan for sample G/PT Day 29 lineage 4. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
3664			T			C			90.84%
1353	p9	5	C	CTG	L	T	TTG	L	24.30%
1263	p12	170	A	GTA	V	G	GTG	V	11.15%
2620			A			G			7.81%
10392	p4	146	A	ATT	I	G	GTT	V	5.71%
4496	p3	24	C	GGC	G	T	GGT	G	5.35%
6862			A			G			5.27%
1268	p12	172	A	CAA	Q	G	CGA	R	4.24%
12516	p1	518	C	CTG	L	T	TTG	L	4.22%
4860	p3	146	A	AAT	N	C	CAT	H	4.20%
129			A			G			4.09%
1631	p5a	4	T	GAT	D	C	GAC	D	3.95%
4479	p3	19	G	GCA	A	T	TCA	S	3.33%
6114	p3	564	T	TTG	L	C	CTG	L	3.33%
5894	p3	490	C	TCC	S	T	TCT	S	3.21%
12338	p1	458	G	GCG	A	A	GCA	A	2.90%
7895	p7	142	A	ACG	T	G	GCG	A	2.90%
6684			A			C			2.83%
4088	p6	58	A	GGA	G	G	GGG	G	2.73%
5276	p3	284	C	GTC	V	T	GTT	V	2.62%
9673	p2	573	G	GAG	E	T	TAG	X	2.57%
1706	p5a	29	G	GAG	E	A	GAA	E	2.57%
5087	p3	221	C	GGC	G	T	GGT	G	2.26%
3903			T			C			2.08%
2379			T			C			1.79%
6406			C			T			1.67%
6891			A			G			1.58%
1843	p5a	75	C	GCG	A	T	GTG	V	1.55%

8781	p2	275	T	GGT	G	C	GGC	G	1.47%
2311			T			C			1.34%
8736	p2	260	T	GAT	D	C	GAC	D	1.29%
1724	p5a	35	T	GAT	D	C	GAC	D	1.25%
2612			T			C			1.20%
8571	p2	205	T	CGT	R	C	CGC	R	1.19%
12175	p1	404	T	ATC	I	C	ACC	T	1.18%
13007	p1	681	T	ATT	I	C	ATC	I	1.18%
2294			A			G			1.12%
2063	p5a	148	A	GCA	A	G	GCG	A	1.07%
5630	p3	402	A	GCA	A	G	GCG	A	1.06%
2678			C			T			1.06%

Appendix 2.21 SNPs above 1% called by VarScan for sample G/PE Day 30 lineage 1. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
2072	p5a	151	C	TTC	F	T	TTT	F	99.81%
1920	p5a	101	G	GGT	G	T	TGT	C	99.73%
3513			C			T			91.58%
6892			C			T			24.12%
1777	p5a	53	C	CCG	P	T	CTG	L	11.21%
12458	p1	498	T	AAT	N	C	AAC	N	8.57%
13182	p1	740	C	CGC	R	T	TGC	C	7.12%
6109	p3	562	A	AAC	N	G	AGC	S	4.48%
7629	p7	53	A	AAC	N	G	AGC	S	3.93%
7063			G			A			2.85%
526	p8	74	T	CAT	H	C	CAC	H	2.31%
8381	p2	142	T	GTT	V	C	GCT	A	2.20%
5558	p3	378	T	ACT	T	C	ACC	T	2.06%
13007	p1	681	T	ATT	I	C	ATC	I	1.91%
12113	p1	383	T	AAT	N	C	AAC	N	1.67%
6780			C			T			1.47%
5136	p3	238	C	CTG	L	T	TTG	L	1.46%
9673	p2	573	G	GAG	E	T	TAG	X	1.40%
2788			A			G			1.30%
6831			A			G			1.26%
9239	p2	428	A	AAG	K	C	ACG	T	1.25%
6873			T			C			1.24%
9886	p2	644	G	GAG	E	C	CAG	Q	1.22%
12175	p1	404	T	ATC	I	C	ACC	T	1.14%
2113	p5a	165	A	AAA	K	G	AGA	R	1.09%

Appendix 2.22 SNPs above 1% called by VarScan for sample G/PE Day 30 lineage 2. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
2179	p5a	187	A	CAG	Q	G	CGG	R	98%
9944	p2	663	T	ATG	M	C	ACG	T	72.25%
3818			T			C			50.05%
4439	p3	5	C	GGC	G	T	GGT	G	36.33%
5913	p3	497	A	ATG	M	G	GTG	V	21.76%
2684			A			G			10.42%
8241	p2	95	A	CCA	P	G	CCG	P	6.89%
3811			G			A			3.89%
2144	p5a	175	T	TAT	Y	C	TAC	Y	3.48%
10917	p4	321	A	AAT	N	G	GAT	D	3.15%
379	p8	25	T	GTT	V	C	GTC	V	3.14%
4949	p3	175	C	GCC	A	T	GCT	A	2.92%
3179			T			C			2.51%
11312	p1	116	C	CGC	R	T	CGT	R	2.21%
8097	p2	47	A	GTA	V	G	GTG	V	2.10%
5645	p3	407	T	AAT	N	C	AAC	N	1.72%
9691	p2	579	G	GCG	A	A	ACG	T	1.56%
10769	p4	271	A	AAA	K	G	AAG	K	1.54%
4673	p3	83	A	CCA	P	G	CCG	P	1.53%
2555			T			C			1.51%
1915	p5a	99	T	GTG	V	C	GCG	A	1.22%
9678	p2	574	T	CGT	R	G	CGG	R	1.19%
6811			A			G			1.17%
6917			T			C			1.17%
12851	p1	629	C	GCC	A	T	GCT	A	1.15%
9673	p2	573	G	GAG	E	T	TAG	X	1.14%
6873			T			C			1.06%
9623	p2	556	A	AAA	K	G	AGA	R	1.02%

Appendix 2.23 SNPs above 1% called by VarScan for sample G/PE Day 30 lineage 3. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
1920	p5a	101	G	GGT	G	T	TGT	C	99.61%
3591			C			T			74.28%
9419	p2	488	A	AAG	K	C	ACG	T	29.25%
3815			T			C			25.96%
12665	p1	567	T	GAT	D	C	GAC	D	21.68%
5705	p3	427	T	ACT	T	C	ACC	T	11.21%
9220	p2	422	C	CAC	H	T	TAC	Y	5.91%
3818			T			C			5.83%
11720	p1	252	C	TCC	S	T	TCT	S	4.96%
42			G			A			4.24%
12908	p1	648	A	GCA	A	G	GCG	A	3.10%
7512	p7	14	A	AAA	K	C	ACA	T	2.91%
11894	p1	310	C	GAC	D	T	GAT	D	2.89%
12216	p1	418	A	ACC	T	G	GCC	A	2.66%
12621	p1	553	A	AAG	K	C	CAG	Q	2.24%
257			T			G			2.18%
7950	p7	160	G	AGC	S	A	AAC	N	1.97%
3749			C			T			1.79%
7448			C			T			1.76%
2811			A			G			1.75%
12360	p1	466	A	ATC	I	G	GTC	V	1.63%
5240	p3	272	C	CTC	L	T	CTT	L	1.59%
8615	p2	220	A	AAA	K	G	AGA	R	1.54%
9228	p2	424	C	AAC	N	T	AAT	N	1.48%
8122	p2	56	A	AAC	N	G	GAC	D	1.47%
11348	p1	128	A	CCA	P	G	CCG	P	1.43%
361	p8	19	G	GCG	A	A	GCA	A	1.30%

1391	p9	17	A	GAA	E	G	GAG	E	1.24%
1447	p9	36	G	GGC	G	A	GAC	D	1.15%
8984	p2	343	A	GAT	D	G	GGT	G	1.13%
7077			A			G			1.11%
7438			G			A			1.05%

Appendix 2.24 SNPs above 1% called by VarScan for sample G/PE Day 30 lineage 4. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
701	p8	133	A	ATC	I	C	CTC	L	87.92%
12473	p1	503	T	GTT	V	C	GTC	V	87.83%
2239	p5a	207	T	GTT	V	C	GCT	A	30.56%
2772			C			T			28.97%
3518			T			C			12.59%
4266	p6	118	A	ACG	T	G	GCG	A	12.58%
3365	p10	17	G	GCC	A	A	ACC	T	11.91%
383	p8	27	C	CGA	R	A	AGA	R	9.39%
6652	p13	65	A	ACG	T	G	GCG	A	5.81%
113			A			G			3.21%
6748			T			G			2.77%
1167	p12	138	G	AAG	K	A	AAA	K	2.48%
360	p8	19	C	GCG	A	T	GTG	V	2.45%
5462	p3	346	G	GCG	A	A	GCA	A	2.43%
1197	p12	148	A	GAA	E	G	GAG	E	2.16%
2060	p5a	147	T	TAT	Y	C	TAC	Y	2.11%
6059	p3	545	C	TCC	S	T	TCT	S	2.02%
975	p12	74	G	GCG	A	A	GCA	A	1.91%
2179	p5a	187	A	CAG	Q	G	CGG	R	1.81%
6813			G			A			1.72%
2336			C			T			1.70%
10635	p4	227	G	GTC	V	A	ATC	I	1.56%
9843	p2	629	C	CTC	L	T	CTT	L	1.38%
11587	p1	208	A	AAG	K	C	ACG	T	1.33%
3117			G			A			1.32%
3818			T			C			1.32%
6773			G			A			1.29%
1672	p5a	18	A	CAA	Q	G	CGA	R	1.26%

5553	p3	377	G	GTC	V	A	ATC	I	1.23%
13007	p1	681	T	ATT	I	C	ATC	I	1.19%
3864			T			C			1.19%
9944	p2	663	T	ATG	M	C	ACG	T	1.12%
3903			T			C			1.12%
13196	p1	744	A	GAA	E	G	GAG	E	1.03%
10804	p4	283	C	GCT	A	T	GTT	V	1.01%

Appendix 2.25 SNPs above 1% called by VarScan for sample G/PE Day 29 lineage 1. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
2072	p5a	151	C	TTC	F	T	TTT	F	99.93%
1920	p5a	101	G	GGT	G	T	TGT	C	99.74%
3513			C			T			86.90%
6892			C			T			20.08%
12458	p1	498	T	AAT	N	C	AAC	N	7.67%
1777	p5a	53	C	CCG	P	T	CTG	L	7.27%
13182	p1	740	C	CGC	R	T	TGC	C	5.90%
7063			G			A			5.80%
6109	p3	562	A	AAC	N	G	AGC	S	5.19%
7629	p7	53	A	AAC	N	G	AGC	S	3.05%
6780			C			T			2.41%
526	p8	74	T	CAT	H	C	CAC	H	1.91%
9673	p2	573	G	GAG	E	T	TAG	X	1.82%
5558	p3	378	T	ACT	T	C	ACC	T	1.66%
9239	p2	428	A	AAG	K	C	ACG	T	1.49%
12113	p1	383	T	AAT	N	C	AAC	N	1.21%
2113	p5a	165	A	AAA	K	G	AGA	R	1.20%
2764			T			C			1.15%
8381	p2	142	T	GTT	V	C	GCT	A	1.13%
9563	p2	536	G	GGC	G	C	GCC	A	1.13%
9886	p2	644	G	GAG	E	C	CAG	Q	1.12%
6873			T			C			1.07%

Appendix 2.26 SNPs above 1% called by VarScan for sample G/PE Day 29 lineage 2. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
2179	p5a	187	A	CAG	Q	G	CGG	R	98.75%
4439	p3	5	C	GGC	G	T	GGT	G	60.52%
9944	p2	663	T	ATG	M	C	ACG	T	52.38%
5913	p3	497	A	ATG	M	G	GTG	V	25.26%
3818			T			C			21.50%
2684			A			G			13.03%
8241	p2	95	A	CCA	P	G	CCG	P	7.08%
8097	p2	47	A	GTA	V	G	GTG	V	5.45%
9678	p2	574	T	CGT	R	G	CGG	R	4.56%
3811			G			A			4.39%
4949	p3	175	C	GCC	A	T	GCT	A	3.92%
10917	p4	321	A	AAT	N	G	GAT	D	3.47%
3179			T			C			3.31%
379	p8	25	T	GTT	V	C	GTC	V	3.30%
2144	p5a	175	T	TAT	Y	C	TAC	Y	3.26%
5645	p3	407	T	AAT	N	C	AAC	N	3.08%
2555			T			C			2.87%
7402			C			T			2.34%
11312	p1	116	C	CGC	R	T	CGT	R	2.34%
9673	p2	573	G	GAG	E	T	TAG	X	2.18%
6811			A			G			2.11%
4673	p3	83	A	CCA	P	G	CCG	P	1.75%
3843			G			A			1.49%
11080	p1	39	A	CAG	Q	G	CGG	R	1.39%
2000	p5a	127	T	CGT	R	C	CGC	R	1.17%
10769	p4	271	A	AAA	K	G	AAG	K	1.11%
9880	p2	642	A	ATC	I	G	GTC	V	1.10%
6873			T			C			1.10%

9009	p2	351	T	GCT	A	C	GCC	A	1.08%
9691	p2	579	G	GCG	A	A	ACG	T	1.08%

Appendix 2.27 SNPs above 1% called by VarScan for sample G/PE Day 29 lineage 3. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
1920	p5a	101	G	GGT	G	T	TGT	C	99.72%
3591			C			T			74.39%
9419	p2	488	A	AAG	K	C	ACG	T	25.42%
3815			T			C			16.74%
12665	p1	567	T	GAT	D	C	GAC	D	14.97%
5705	p3	427	T	ACT	T	C	ACC	T	12.16%
42			G			A			3.53%
257			T			G			3.46%
11720	p1	252	C	TCC	S	T	TCT	S	3.12%
12621	p1	553	A	AAG	K	C	CAG	Q	2.86%
8615	p2	220	A	AAA	K	G	AGA	R	2.71%
3749			C			T			2.55%
7448			C			T			2.53%
12360	p1	466	A	ATC	I	G	GTC	V	2.46%
9220	p2	422	C	CAC	H	T	TAC	Y	2.27%
3818			T			C			2.24%
7784	p7	105	C	CTG	L	T	TTG	L	1.93%
9673	p2	573	G	GAG	E	T	TAG	X	1.83%
1826	p5a	69	A	GAA	E	G	GAG	E	1.83%
11894	p1	310	C	GAC	D	T	GAT	D	1.60%
1391	p9	17	A	GAA	E	G	GAG	E	1.59%
7512	p7	14	A	AAA	K	C	ACA	T	1.57%
12908	p1	648	A	GCA	A	G	GCG	A	1.46%
5240	p3	272	C	CTC	L	T	CTT	L	1.33%
7077			A			G			1.25%
361	p8	19	G	GCG	A	A	GCA	A	1.24%
2480			A			C			1.21%
2872			A			G			1.21%

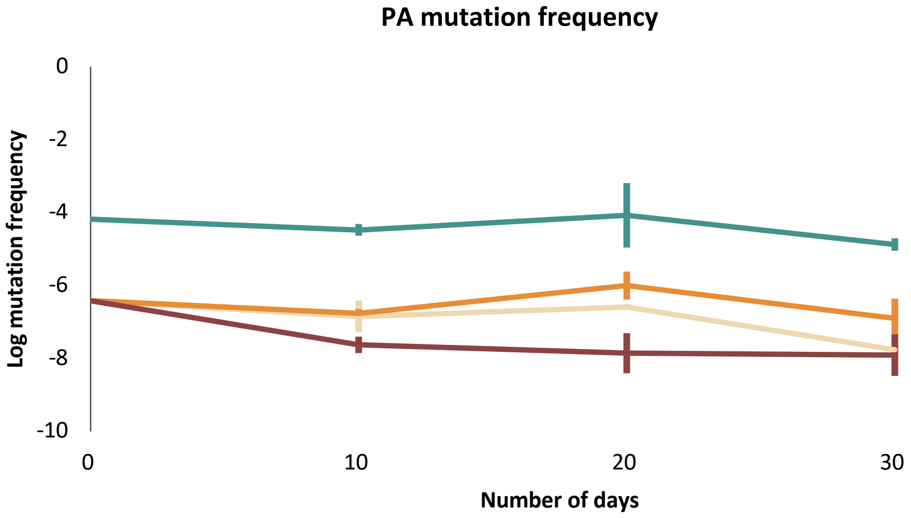
7950	p7	160	G	AGC	S	A	AAC	N	1.21%
8961	p2	335	A	GGA	G	C	GGC	G	1.20%
1447	p9	36	G	GGC	G	A	GAC	D	1.17%
4772	p3	116	T	TAT	Y	C	TAC	Y	1.16%
6109	p3	562	A	AAC	N	G	AGC	S	1.16%
2811			A			G			1.03%
9228	p2	424	C	AAC	N	T	AAT	N	1.03%
7799	p7	110	C	CTG	L	T	TTG	L	1.03%
1604	p9	88	C	TTC	F	T	TTT	F	1.03%

Appendix 2.28 SNPs above 1% called by VarScan for sample G/PE Day 29 lineage 4. Columns abbreviations stand for: ntpos: nucleotide position on concatenated genome (Small segment: 1-2949, Medium Segment: 2950-7014, Large segment: 7015-13388), protname: protein name; aapos: amino acid position of SNP; pre_nt: previous nucleotide; pre_codon: previous codon; pre_aa: previous amino acid; cur_nt: current nucleotide with SNP, cur_codon: current codon with SNP, cur_aa: current amino acid with SNP, freq: frequency of SNP. When protein name not indicated, SNPs were found in non-coding regions.

ntpos	protname	aapos	pre_nt	pre_codon	pre_aa	cur_nt	cur_codon	cut_aa	freq
701	p8	133	A	ATC	I	C	CTC	L	92.90%
12473	p1	503	T	GTT	V	C	GTC	V	90.91%
2239	p5a	207	T	GTT	V	C	GCT	A	20.84%
2772			C			T			20.60%
6652	p13	65	A	ACG	T	G	GCG	A	9.98%
3518			T			C			7.24%
4266	p6	118	A	ACG	T	G	GCG	A	6.77%
5462	p3	346	G	GCG	A	A	GCA	A	6.53%
383	p8	27	C	CGA	R	A	AGA	R	5.71%
3365	p10	17	G	GCC	A	A	ACC	T	4.78%
1197	p12	148	A	GAA	E	G	GAG	E	4.36%
113			A			G			3.36%
6748			T			G			3.35%
6059	p3	545	C	TCC	S	T	TCT	S	2.53%
3864			T			C			2.37%
6813			G			A			1.84%
13196	p1	744	A	GAA	E	G	GAG	E	1.82%
2060	p5a	147	T	TAT	Y	C	TAC	Y	1.71%
9843	p2	629	C	CTC	L	T	CTT	L	1.70%
1672	p5a	18	A	CAA	Q	G	CGA	R	1.68%
3903			T			C			1.52%
379	p8	25	T	GTT	V	C	GTC	V	1.46%
1167	p12	138	G	AAG	K	A	AAA	K	1.40%
11945	p1	327	A	GAA	E	G	GAG	E	1.39%
7336			T			C			1.36%
360	p8	19	C	GCG	A	T	GTG	V	1.35%
6773			G			A			1.32%
2336			C			T			1.25%

3710			T			C			1.04%
376	p8	24	G	CTG	L	A	CTA	L	1.02%

Appendix 2.29 Frequency of PA mutations among lineages. S/P lineages in teal, G/P in beige, G/PT in orange, G/PE in maroon. Data are collected every 10 days during the 30-day passaging, points are average values of four lineages, error bar are standard deviations.



Appendix 4.1 926 CRESS Rep sequence accession numbers. Green and blue highlighted are sequences used to build Geminivirus and Genomovirus Rep trees. Blue highlighted are removed begomoviruses in Appendix 4.13. Grey highlighted are collaborator provided sequences.

YP_009252314	YP_009142778	YP_009109636	YP_009333618	I0831_H8	YP_001004144	YP_009130622	YP_009170680	YP_002224032	NP_579976
YP_009237555	YP_009237588	YP_009109649	YP_009110682	YP_009058942	YP_729220	YP_009428564	YP_009143526	YP_001718634	NP_543117
YP_009047130	YP_009115534	YP_009259720	YP_004152327	YP_009116889	YP_764516	YP_009408646	YP_009140567	YP_001718628	NP_443744
YP_009237524	YP_009252340	YP_009126916	YP_008130363	YP_009163938	YP_729214	YP_002261483	YP_009137892	YP_001718622	NP_148988
YP_009109686	YP_009051960	YP_009237546	YP_009104366	YP_009116891	YP_764447	YP_006459	YP_009129288	YP_006905833	NP_148955
YP_009237602	YP_009021879	YP_009389439	YP_009047065	YP_009237554	YP_459911	YP_009325424	YP_009129278	YP_003915075	NP_148932
YP_009237608	YP_009252389	YP_009126879	YP_009021871	YP_009177700	YP_001661654	YP_001960973	YP_008198573	YP_003587821	NP_066371
YP_009237498	YP_009112559	YP_009237585	YP_009021850	NP_620696	YP_425033	YP_009408633	YP_003622544	YP_003560503	NP_062871
YP_009116892	YP_009237504	YP_009116913	YP_009021843	NP_619760	YP_001285758	YP_009408600	NP_047218	YP_003345516	NP_049918
YP_009117082	BAN59850	YP_009259718	YP_007392931	NP_619565	YP_001285740	YP_009047200	YP_005352908	YP_002124298	NP_047249
YP_009117074	AB781089	YP_009163920	YP_004152331	YP_009021880	YP_001285734	YP_009315916	YP_005352903	YP_001950237	NP_047237
YP_009021243	YP_009345097	YP_009237575	YP_009362252	NP_619759	YP_001285746	YP_009345717	YP_005086957	YP_005352660	NP_047233
YP_009163897	YP_004046698	YP_009259682	YP_004152333	YP_009021872	YP_740297	YP_009216586	YP_004564604	YP_005351246	NP_047223
YP_009163932	YP_009111348	YP_009001737	YP_004152329	NP_619572	YP_001086462	YP_009344823	YP_004429245	YP_004346960	NP_044925
YP_009126901	YP_473359	YP_003084291	YP_009021893	NP_619768	YP_277439	YP_009338003	YP_004429235	YP_004339041	NP_040770
YP_009117057	BAL05205	YP_009109683	YP_009110680	NP_620700	YP_851011	YP_009241413	YP_003084137	YP_004222843	YP_003987454
YP_009237511	YP_009001777	YP_009117061	YP_009021845	NP_619761	YP_001648980	YP_001936689	YP_002791014	YP_004191799	YP_006491266
YP_009337827	YP_004286322	YP_009001739	YP_009158862	YP_009252330	YP_717918	YP_764453	YP_002775436	YP_004123092	
YP_009259708	YP_009109635	YP_009259675	YP_009109670	YP_009117070	YP_009362978	YP_009329832	YP_002753151	YP_004123721	
YP_009259695	YP_009345107	YP_009237565	YP_009237592	YP_009117066	YP_009226627	YP_009325937	YP_002643049	YP_004123085	
YP_009237578	YP_009345086	YP_009237594	YP_009109663	YP_009237509	YP_003778178	YP_009316020	YP_002608335	YP_004046692	
YP_003084285	YP_009252327	YP_009116894	YP_009109643	YP_009116898	YP_009111309	YP_009316183	YP_002576171	YP_004021922	
YP_008828162	YP_009252335	YP_009389456	YP_009109668	YP_009351871	YP_006666535	YP_009310418	YP_002455912	YP_003987461	
YP_009126927	YP_009252333	YP_009259732	YP_009163927	YP_003084297	YP_006659974	YP_009310094	YP_002317398	YP_003934914	
YP_009109644	YP_009237599	YP_009126877	YP_009117058	YP_009116647	YP_006666523	YP_009310080	YP_001974416	YP_002268198	
YP_009001747	YP_009126930	YP_009116902	YP_009237577	YP_009117064	YP_004089627	YP_009310065	YP_001974402	YP_001911132	
YP_003084290	YP_009126895	YP_009117076	YP_009237543	YP_009116789	NP_597785	YP_009310086	YP_001955943	YP_003622559	
YP_009109630	YP_009126889	YP_009237549	YP_009126903	YP_009126941	YP_003915159	YP_009310060	YP_001876452	YP_003288773	
YP_009237570	YP_009252324	YP_009237495	YP_007517186	YP_009021041	NP_569147	YP_009310071	YP_001856215	YP_003254633	
YP_009237538	YP_009252310	YP_009109640	YP_009126919	YP_009259705	YP_004465364	YP_009305128	YP_980246	YP_002875764	
YP_009259692	YP_009054989	YP_007878130	YP_009389445	YP_009237598	YP_009021763	YP_009270641	YP_699993	YP_002640509	
YP_009121932	YP_009252312	YP_009252331	YP_009047144	YP_009116781	NP_542349	YP_009270647	YP_665625	YP_002455918	
YP_009021245	YP_009252306	YP_009237496	YP_003084143	YP_004046687	YP_001941162	YP_005352898	YP_619888	YP_001994911	
YP_009126882	YP_009252308	YP_009259689	YP_009160329	YP_009237591	YP_009026388	YP_009269405	YP_619883	YP_001974397	
YP_009126938	YP_009118278	YP_009237541	YP_001661660	YP_009337838	YP_009389276	NP_795354	YP_232878	YP_002268205	

YP_009259745	YP_009054991	YP_009389527	NP_604483	YP_009237528	NP_042590	NP_620892	YP_006469	YP_005352917
YP_009237530	YP_009252320	YP_009047134	YP_008997794	YP_009237522	YP_006331073	YP_004821545	YP_006452	YP_005352893
YP_009126890	YP_009054987	YP_009047142	YP_008992018	YP_009237563	YP_006273070	YP_009256563	NP_981938	NP_047243
YP_009126896	YP_009118276	YP_009163907	YP_008997797	YP_001144448	NP_620492	YP_009249831	NP_871726	YP_459930
YP_009237516	I1022-I77	YP_009126898	YP_003104737	YP_001285874	YP_001686794	YP_009249837	NP_871720	YP_001718510
YP_009163905	YP_007974230	YP_009226561	NP_619769	YP_001249281	YP_001941155	NP_620884	NP_852654	YP_271822
YP_009259714	YP_007974228	YP_009226565	NP_619567	YP_794171	YP_009154763	NP_050017	NP_808900	YP_213935
YP_003084140	YP_009163761	YP_009126922	YP_008997789	YP_003966137	NP_040965	YP_002117527	NP_808893	YP_213943
YP_009389529	YP_009054993	YP_009237560	NP_579867	NP_066185	YP_003288768	NP_612597	NP_808734	YP_184754
YP_009109615	I1035	YP_009109660	YP_009246456	YP_009305428	NP_840053	NP_808844	NP_803141	YP_133834
YP_009109623	YP_009118272	YP_009237502	YP_003433564	NP_040557	YP_009325925	NP_671446	NP_795340	YP_115511
YP_009109620	YP_009408603	YP_007353980	YP_003966132	YP_009129272	NP_045945	NP_049348	NP_722556	YP_006426
YP_003084299	YP_009118274	YP_009259711	YP_003987456	NP_620735	YP_009362982	YP_009237910	NP_689461	YP_006437
YP_009021890	YP_009022025	YP_009126884	YP_008169853	YP_001294922	YP_004046663	YP_009226417	NP_671461	YP_006431
YP_009237518	YP_009054985	YP_003084282	YP_003828902	YP_794165	YP_009104363	YP_009175085	NP_660168	NP_995306
YP_006281010	YP_009252322	YP_009109626	NP_040944	YP_001294917	YP_008472704	YP_009175078	NP_660081	NP_991336
YP_009237500	YP_009030025	YP_009389534	YP_007004040	YP_579170	YP_004046667	YP_003254640	NP_632006	NP_958320
YP_009259679	YP_009252318	YP_009126925	YP_008854133	YP_001285764	NP_620726	YP_009216263	NP_620883	NP_958314
YP_009226569	YP_006331067	YP_009259700	YP_004123952	YP_001210303	YP_002014712	NP_835275	NP_619682	NP_955735
YP_009226571	YP_009337832	YP_009237586	YP_007250412	YP_001040016	NP_612221	YP_009177713	NP_619557	NP_955741
YP_009163917	YP_009116879	YP_009109659	YP_006666512	YP_001655008	YP_006666527	YP_009175012	NP_077735	NP_955747
YP_009226572	YP_009252326	YP_009001753	YP_004778177	YP_271828	YP_007004038	YP_009174975	NP_077091	NP_817119
YP_009021241	YP_009252316	YP_009389520	YP_009230209	YP_009259496	YP_006666531	YP_006443	YP_009116883	NP_817112
YP_009252337	YP_009237600	YP_009237604	YP_002941920	YP_001661461	YP_009408627	YP_009121940	YP_008691088	NP_817105
YP_009237512	YP_009109653	YP_009022029	YP_003082245	YP_001040012	YP_008400117	YP_009058924	YP_009112876	NP_808911
YP_009237514	YP_009237571	YP_009126892	YP_006666513	YP_803222	YP_006590005	YP_009056858	YP_009109707	NP_808869
YP_009237520	YP_009001745	YP_009115538	YP_003082247	YP_459905	YP_009417307	YP_009042058	YP_009109613	NP_808826
YP_009252390	YP_009001751	YP_009126935	YP_003334471	YP_001285752	YP_009255247	YP_008719948	YP_009109607	NP_808832
YP_009116896	YP_009117079	YP_009126915	YP_004046695	YP_009021233	YP_009162635	YP_008400131	YP_009091993	NP_808850
YP_009163929	YP_009226567	YP_009126905	YP_004123953	YP_001004150	YP_009130624	YP_008400125	YP_009051958	NP_808804
YP_009237533	YP_009163918	YP_009237583	YP_006742179	YP_009252350	YP_009130661	YP_007024783	YP_009051691	NP_808771
YP_009163928	YP_009126881	YP_009163901	YP_001285945	YP_009109715	YP_009181999	YP_006860604	YP_003896044	NP_808777
YP_009001743	YP_009226563	YP_009109675	YP_009154718	YP_009337824	YP_009252362	YP_006860599	YP_003622552	NP_808758
YP_009237564	YP_009237559	YP_009116909	YP_009344825	YP_009115508	YP_009130630	YP_006590064	YP_003288785	NP_808765
YP_009237550	YP_009021888	YP_009315918	NP_878244	YP_009058947	YP_009109727	YP_004958249	YP_003104750	NP_808753
YP_009259723	YP_009163936	YP_009380542	YP_001285943	YP_009051835	YP_009388636	YP_004958242	YP_003084280	NP_808783
YP_009259728	YP_009163904	YP_009116910	NP_878243	YP_009259551	YP_003104796	YP_004958233	YP_002154620	NP_803551
YP_003084293	YP_009001756	YP_009021847	YP_851014	YP_009181996	YP_009115514	YP_004958227	YP_002004579	NP_803557
YP_009237506	YP_009047125	YP_009237526	YP_009073582	YP_009345109	YP_009115523	YP_004958222	YP_009030010	NP_803536
YP_009237534	YP_009047137	YP_004376332	YP_003104752	YP_009273015	YP_009021856	YP_004901696	YP_008411025	NP_803408

YP_009259737	YP_009237567	YP_164517	YP_009254744	YP_009163759	YP_009021043	YP_004429239	YP_009029992	NP_803243
YP_009047132	YP_009163899	YP_271918	YP_007004041	YP_009388634	YP_009252368	YP_004207815	YP_009026403	NP_803249
YP_009117078	YP_009163922	NP_877978	YP_009001905	YP_009021852	YP_009252353	YP_004207827	YP_009026396	NP_803224
YP_009116905	YP_009126936	YP_009091698	YP_009044085	YP_009164036	YP_009252356	YP_004207822	YP_008997814	NP_803150
YP_009117067	YP_009163924	NP_150368	YP_009052483	YP_009109729	YP_009116901	YP_004123935	YP_006390088	NP_795350
YP_009237596	YP_009389442	NP_047275	YP_009134750	YP_009109725	YP_009115515	YP_004064947	YP_009021225	NP_786880
YP_009259699	YP_009163909	YP_803546	YP_425036	YP_009130628	YP_009115519	YP_003856011	YP_008492942	NP_783156
YP_009237542	YP_009163913	NP_059527	YP_004347413	YP_009130626	YP_009116887	YP_003828907	YP_008901146	NP_694933
YP_009001742	YP_009163911	NP_573442	YP_009337826	YP_009408609	YP_009115517	YP_003620408	YP_003864098	NP_690916
YP_009109685	YP_009163915	YP_610960	YP_009174978	YP_009109719	YP_009164033	YP_002941855	YP_001974391	NP_690101
YP_009116906	YP_009109676	YP_009134739	YP_009174977	YP_009252371	YP_009345091	YP_002519381	YP_008888542	NP_671468
I0178	YP_009021875	YP_803549	YP_009002582	YP_009252344	YP_009252365	YP_001960948	YP_008492926	NP_671475
I1020_I75	YP_009001750	YP_764455	YP_009044083	YP_009130657	YP_009252359	YP_001960955	YP_007974235	NP_671452
I1021	YP_003084287	YP_009253909	YP_009337831	YP_009115511	YP_009109731	YP_001960962	YP_007974225	NP_665674
I1393	YP_009126929	YP_009091696	YP_007518505	YP_009115526	YP_009109741	YP_001974407	YP_005087580	NP_658995
I1022_I72	YP_009259555	YP_009170674	YP_003934916	YP_009345088	YP_009109735	YP_001333680	YP_008060402	NP_632018
I1022_G1	YP_009047139	YP_009351873	YP_031730	YP_009388638	YP_009109733	YP_001333687	YP_008003516	NP_632012
I1360	YP_009408607	YP_007974237	YP_007011043	YP_009021860	YP_009389447	YP_717930	YP_007517192	NP_631955
I1338b	YP_009021877	YP_003422530	YP_009337830	YP_009252347	YP_009345062	YP_717094	YP_007438888	NP_620852
I1347	YP_009237581	NP_065678	YP_007011042	YP_009115529	YP_001715621	YP_459916	YP_007438883	NP_620741
SDBVL	YP_009237552	NP_937956	YP_006488615	YP_009259554	YP_004207925	YP_293693	YP_007438878	NP_620748
I1077B	YP_009126932	YP_009021891	YP_003433566	YP_009109723	YP_009021229	NP_957674	YP_007438873	NP_620665
I0171	YP_009237536	YP_009423856	YP_009338004	YP_009051832	YP_009109717	NP_808887	YP_007438868	NP_620424
I0910	YP_009237606	YP_007697652	NP_619574	YP_009109721	YP_009129321	NP_620867	YP_007419083	NP_620297
I0991	YP_009252342	YP_009000900	YP_009058890	YP_009115532	YP_006522422	NP_049355	YP_006905839	NP_620012
I0960	YP_009237540	YP_009052458	YP_009021881	YP_009051838	YP_002875759	NP_040324	YP_007250561	NP_598190

Additional manual edits: YP_009408627 and YP_009408626 were concatenated by removing amino acids 207-261 from YP_009408627, amino acids 1-21 from YP_009408626. YP_008400117 and YP_008400116 were concatenated by removing amino acids 205-264 from YP_009408627, amino acids 1-16 from YP_009408626. NP_040965 was NP_040964 and NP_040965 concatenated with the removal of amino acids 214-266 from NP_040965. NP_045945 was NP_045945 and NP_045944 concatenated with the removal of amino acids 224-276 from NP_045945. NP_597785 was NP_597785 and NP_899201 concatenated with the removal of amino acids 272-336 from NP_587785. NP_840053 was NP_840053 and NP_840052 concatenated with the removal of amino acids 219-312 from NP_840053. YP_006522422 was YP_006522422 and YP_006522421 concatenated with the removal of amino acids 205-269 from YP_006522422. Other concatenated sequences include NP_569147+46, YP_001716521+20, YP_004207925+24, without removal of amino acids from sequence.

Appendix 4.2 Genome accession numbers from which capsid protein sequences were extracted.

Genomovirus		Circovirus	Smacovirus	Nanovirus	Bacilladnavirus	Parvovirus	
KT862253	KF371630	KM382269	KM573772	Y003104738	AB597949	NC_004285	NC_035186
KR912221	KF371631	KJ641740	KT862221	N619570	AB844272	NC_015115	NC_016647
KT732792	LK931484	KT732785	KT862225	Y008997806	AB193315	NC_011317	NC_023673
KT732793	KP974693	KT732786	KT862219	Y008997802	AB553581	NC_012636	NC_016031
KF371640	KJ547634	KM017740	KT862223	Y008992019	AB781089	NC_004290	NC_016032
KT862251	KJ938717	JX185424	KM573769	N619767	KY405008	NC_022748	NC_012042
KF371643	KP263543	KT732787	KT862224	N620699	KF133809	NC_004289	NC_012729
KF371641	LK931483	KC771281	KT862218	Y009508036	KY405006	NC_015718	NC_012564
KF371642	KP263545	KF031466	KT862222	N604477	KY405007	NC_006555	NC_007455
GQ365709	KP263546	GQ404857	JN634851	Y001661657		NC_018450	NC_014358
KF268025	KP987887	HQ738634	KM598409	Y009508220		NC_004287	NC_031695
KF268026	KT862250	KJ641712	KT862228			NC_018399	NC_029133
KF268027	KP133076	JF938079	KM573774			NC_004288	NC_028973
KF268028	KP133077	GQ404846	KP233194			NC_023842	NC_024453
KM598382	KP133078	GQ404845	KT600068			NC_005040	NC_029300
KM598383	KP133079	KJ831064	GQ351272			NC_012685	NC_030873
KM598384	KP133080	HQ738643	GQ351275			NC_001899	NC_022800
KF371639	KP133075	HQ738637	KJ577810			NC_019492	NC_017823
KT253577	KP263547	GQ404849	KJ577817			NC_004286	NC_020499
KT253578	KJ641737	JF938081	KF880727			NC_005341	NC_004442
KT253579	KP263544	KC512919	KJ577819			NC_000936	NC_025825
KF371637	HQ335086	HQ738636	KM573770			NC_030296	NC_031751
JX185429	JX185430	JX185426	KC545226			NC_026943	NC_034445
KM598385	KT862249	JF938082	KC545227			NC_005041	NC_001662
KM598386	KJ641726	HM228874	KP233191			NC_031450	NC_001718
KM598387	KT732790	JX569794	KP233178			NC_014357	NC_001510
KM598388	KT732791	GQ404844	KP233180			NC_007218	NC_001539
KT732801	KJ413144	KC512920	KY086301			NC_011545	NC_029797
KT732802	KJ547635	KJ641715	KP233174			NC_032097	NC_024888
KF371635	KT862254	KJ641720	KY086300			NC_004284	NC_030837
JQ412056	KT862238	KM382270	KP860906			NC_002190	NC_028650
JQ412057	KT862239	GQ404854	KT862229			NC_022564	NC_026815
KT732794	JN704610	JX185422	KP233190			NC_022089	
KF371633	KJ547642	KJ641717	JX274036			NC_037053	
KT862244	KT732813	KF726984	KJ577811			NC_000883	
KT862243	KJ547644	KJ641728	KM573771			NC_004295	

KT862246	KP974694	GQ404855	KM573775			NC_026251	
KT363839	KJ547645	JX185419	KJ577816			NC_016752	
KT732795	LK931485	KC512918	KP233189			NC_006259	
KT732796	KM510192	LC018134	KM598410			NC_031959	
KJ641719	KT732814	GQ404847	KT862220			NC_023860	
KT732800	KJ547639	KC512916	KJ577812			NC_014665	
KT732806	KJ547637	KR902499	GQ351273			NC_023020	
KT732804	KJ547640	KC241982				NC_027429	
KT732805	KJ547638	KJ641711				NC_006148	
JX185428	KM821747	KJ641727				NC_006147	
KT862242	KT862245	AF071878				NC_001701	
KT309029	KT862247	AF252610				NC_014468	
KT732798	KF371634	DQ172906				NC_006263	
KT732799	KJ547643	GQ404851				NC_004828	
KT732808	KT862255	DQ845074				NC_006261	
KT732812		AJ301633				NC_002077	
KT732810		DQ146997				NC_006260	
KT732811		DQ845075				NC_001729	
KT732807		KP793918				NC_001829	
KT732809		GU799606				NC_005889	
KT732797		DQ100076				NC_006152	
KF413620		AJ304456				NC_025891	
KT732803		EU056309				NC_035180	
KF371632		JQ814849				NC_025965	
KT598248		KT783484				NC_016744	
KF371638		KC339249				NC_007018	
KT862241		AF027217				NC_022104	
KT862240		AF071879				NC_031670	
KT862252		JX863737				NC_028136	
KU343137		KJ020099				NC_024452	
KJ547641		JQ011377				NC_024454	
KJ547636		GQ404856				NC_030402	
KF371636		KJ641723				NC_001540	
KT862248		KJ641724				NC_035185	

Appendix 4.3 Pearson correlation of log ratio of WAG rates over fmg method simulated data derived matrix rates. G01 means coefficient alpha for gamma distribution is 0.1 during simulation, G05 means coefficient alpha for gamma distribution is 0.5, G1 means coefficient alpha for gamma distribution is 1, G2 means coefficient alpha for gamma distribution is 2.

WAG/fmg-G05	0.09		
WAG/fmg-G1	0.257	-0.046	
WAG/fmg-G2	0.263	0.33	-0.02
	WAG/fmg-G01	WAG/fmg-G05	WAG/fmg-G1

Appendix 4.4 Pearson correlation of log ratio of WAG rates over fit method simulated data derived matrix rates. G01 means coefficient alpha for gamma distribution is 0.1 during simulation, G05 means coefficient alpha for gamma distribution is 0.5, G1 means coefficient alpha for gamma distribution is 1, G2 means coefficient alpha for gamma distribution is 2.

WAG/fit-G05	0.049		
WAG/fit-G1	0.066	0.038	
WAG/fit-G2	0.011	0.27	-0.165
	WAG/fit-G01	WAG/fit-G05	WAG/fit-G1

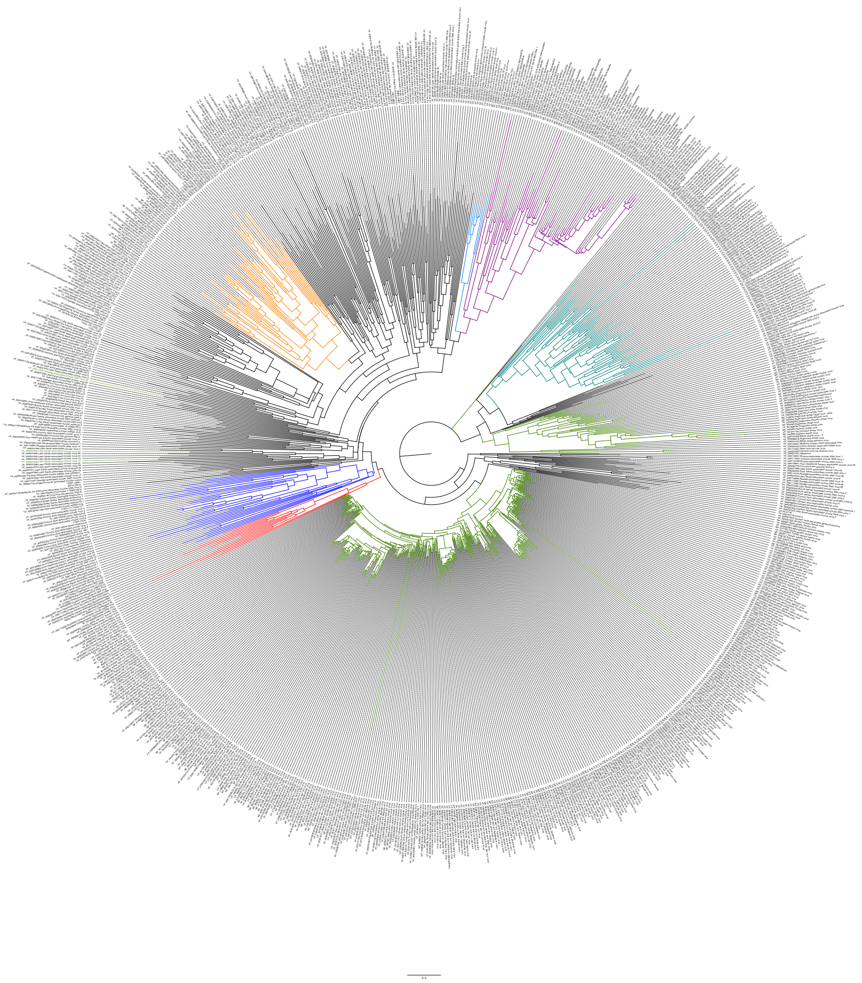
Appendix 4.5 Maximum likelihood scores of trees built by best matrices from 10 test set alignments.

						test set					
		1	2	3	4	5	6	7	8	9	10
Best matrix	fmgVT-1	-163729	-166092	-170039	-163298	-161822	-163795	-163296	-163092	-165413	-159692
	fmgVT-2	-163649	-166082	-170040	-163291	-161776	-163754	-163313	-163030	-165399	-159710
	fmgLG-3	-163660	-166018	-170105	-163276	-161817	-163773	-163346	-163052	-165400	-159725
	fitVT-4	-163667	-166037	-169988	-163324	-161761	-163739	-163325	-163021	-165413	-159685
	fmgLG-5	-163647	-166037	-170026	-163282	-161860	-163775	-163340	-163045	-165361	-159727
	fmgVT-6	-163612	-166027	-169959	-163315	-161793	-163836	-163301	-163058	-165378	-159695
	fmgVT-7	-163631	-166013	-169952	-163263	-161791	-163755	-163376	-163082	-165380	-159657
	fmgVT-8	-163676	-166050	-170006	-163280	-161788	-163754	-163334	-163131	-165412	-159745
	fmgLG-9	-163698	-166062	-170021	-163291	-161832	-163771	-163329	-163038	-165483	-159734
	fmgVT-10	-163600	-166014	-169998	-163254	-161785	-163735	-163299	-163041	-165375	-159761

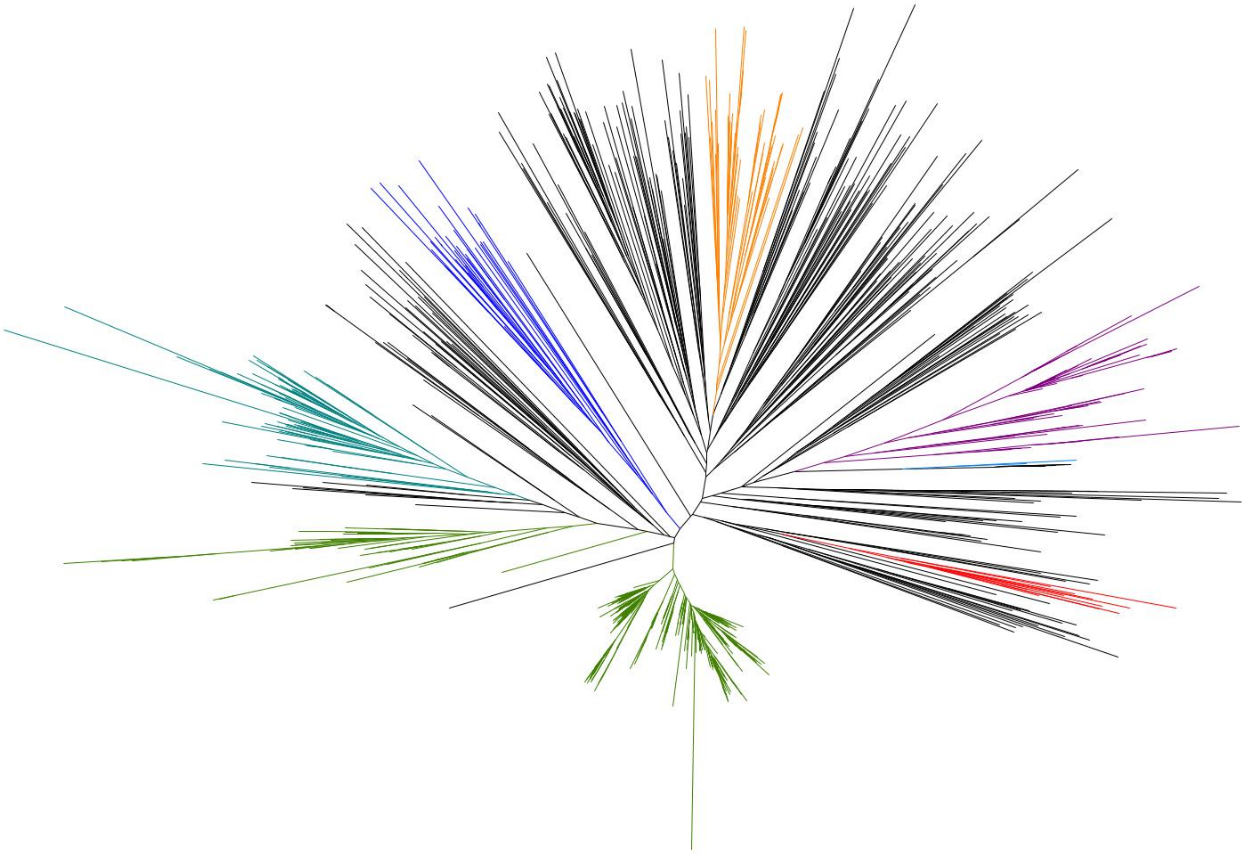
Appendix 4.6 Rank order of maximum likelihood scores of trees built by best matrices from 10 test set alignments. Best matrix with the lowest ranking is highlighted in peach color.

		test set										
		1	2	3	4	5	6	7	8	9	10	sum
Best matrix	fmgVT-1	10	6	6	6	9	9	1	9	9	2	67
	fmgVT-2	5	10	7	4	2	2	5	2	6	5	48
	fmgLG-3	9	8	10	5	4	7	7	5	7	4	66
	fitVT-4	3	9	9	8	8	8	6	1	5	7	64
	fmgLG-5	6	7	8	7	10	5	4	7	4	6	64
	fmgVT-6	2	4	2	9	7	10	3	6	2	3	48
	fmgVT-7	4	1	1	2	6	4	10	8	3	1	40
	fmgVT-8	8	5	4	3	5	3	8	10	8	9	63
	fmgLG-9	7	3	5	10	1	6	9	3	10	8	62
	fmgVT-10	1	2	3	1	3	1	2	4	1	10	28

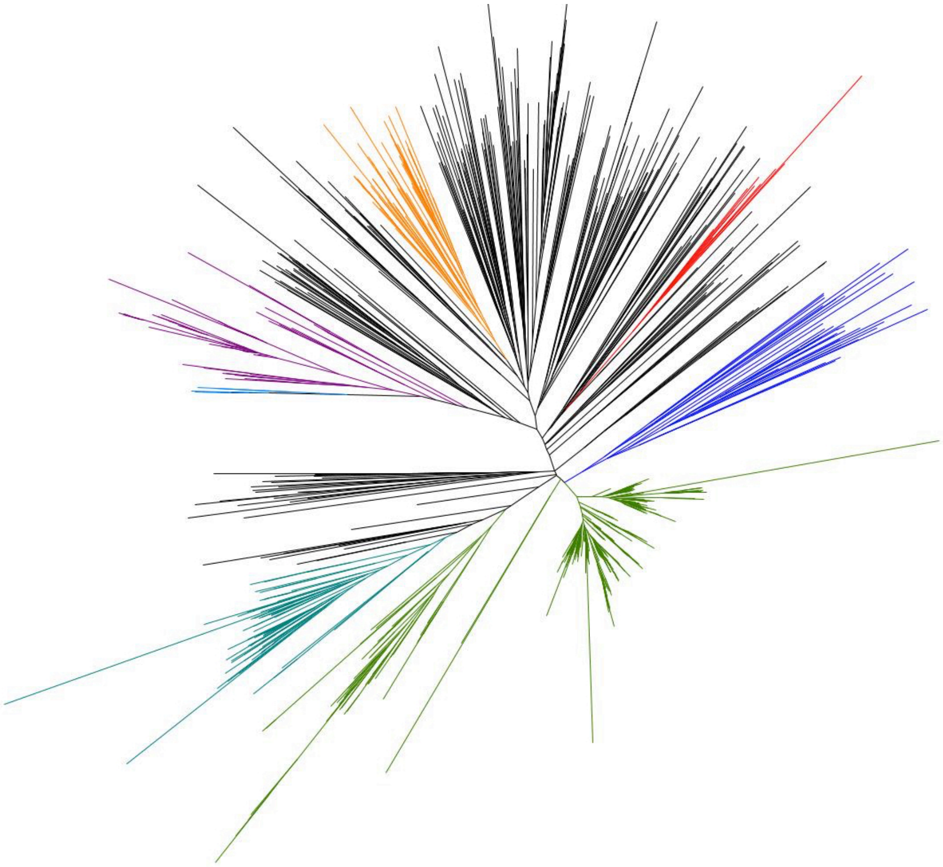
Appendix 4.7 Midpoint rooted Maximum likelihood tree built with CRESS. Alignment product by MUSCLE. Green for *Geminiviridae*, teal for *Genomoviridae*, dark blue for *Smacoviridae*, red for *Bacilladnaviridae*, orange for *Circoviridae*, light blue for *Nanoviridae*, purple for the alphasatellites (*Alphasatellitidae*).



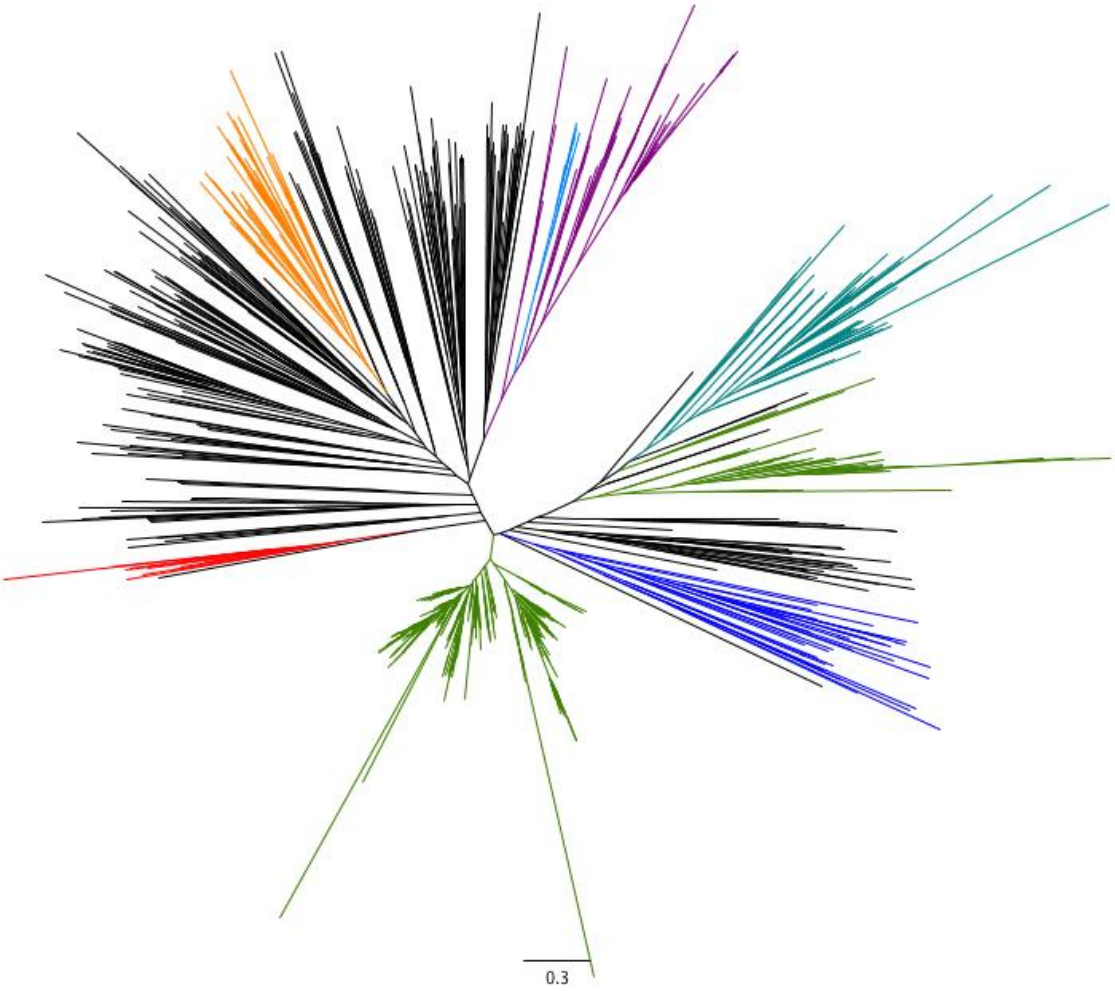
Appendix 4.8 Maximum likelihood tree built using rtREV matrix. Green for *Geminiviridae*, teal for *Genomoviridae*, dark blue for *Smacoviridae*, red for *Bacilladnaviridae*, orange for *Circoviridae*, light blue for *Nanoviridae*, purple for the alphasatellites (*Alphasatellitidae*).



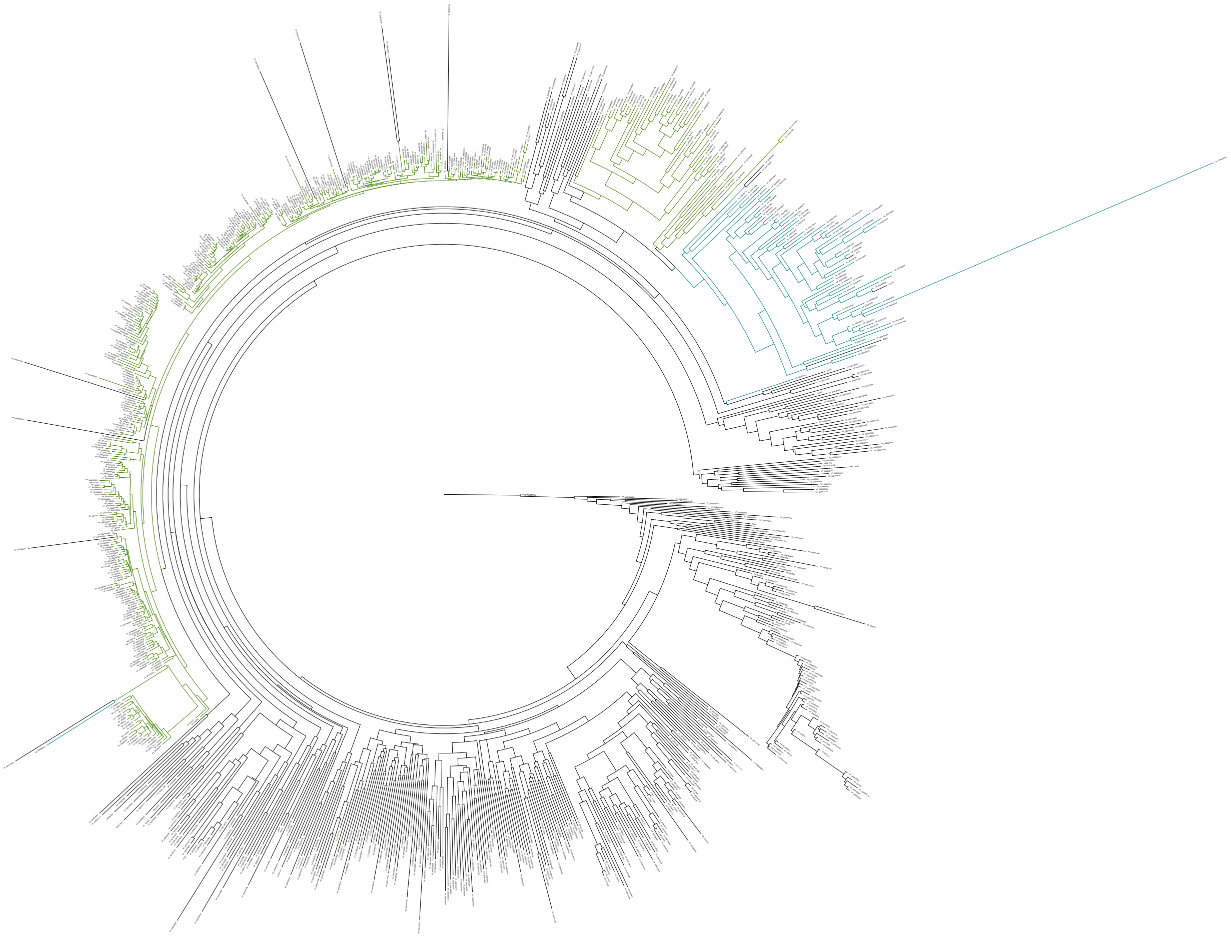
Appendix 4.9 Maximum likelihood tree built using LG matrix. Green for *Geminiviridae*, teal for *Genomoviridae*, dark blue for *Smacoviridae*, red for *Bacilladnaviridae*, orange for *Circoviridae*, light blue for *Nanoviridae*, purple for the alphasatellites (*Alphasatellitidae*).



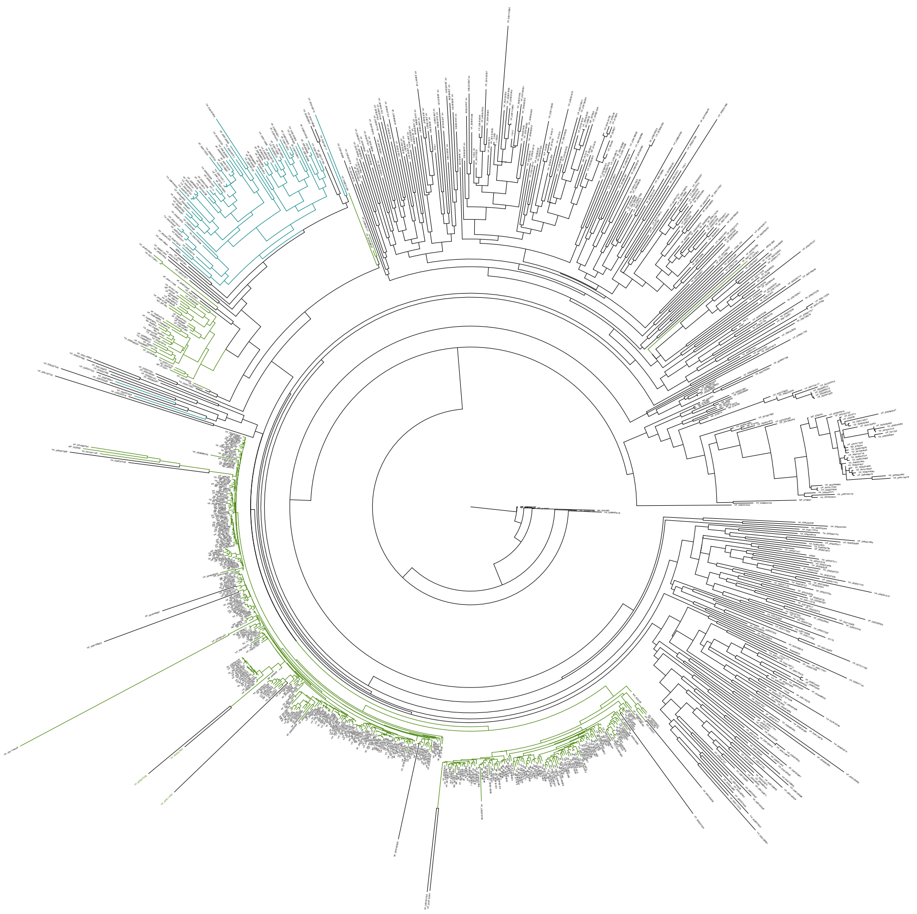
Appendix 4.10 Maximum likelihood tree built using VT matrix. Green for *Geminiviridae*, teal for *Genomoviridae*, dark blue for *Smacoviridae*, red for *Bacilladnaviridae*, orange for *Circoviridae*, light blue for *Nanoviridae*, purple for the alphasatellites (*Alphasatellitidae*).



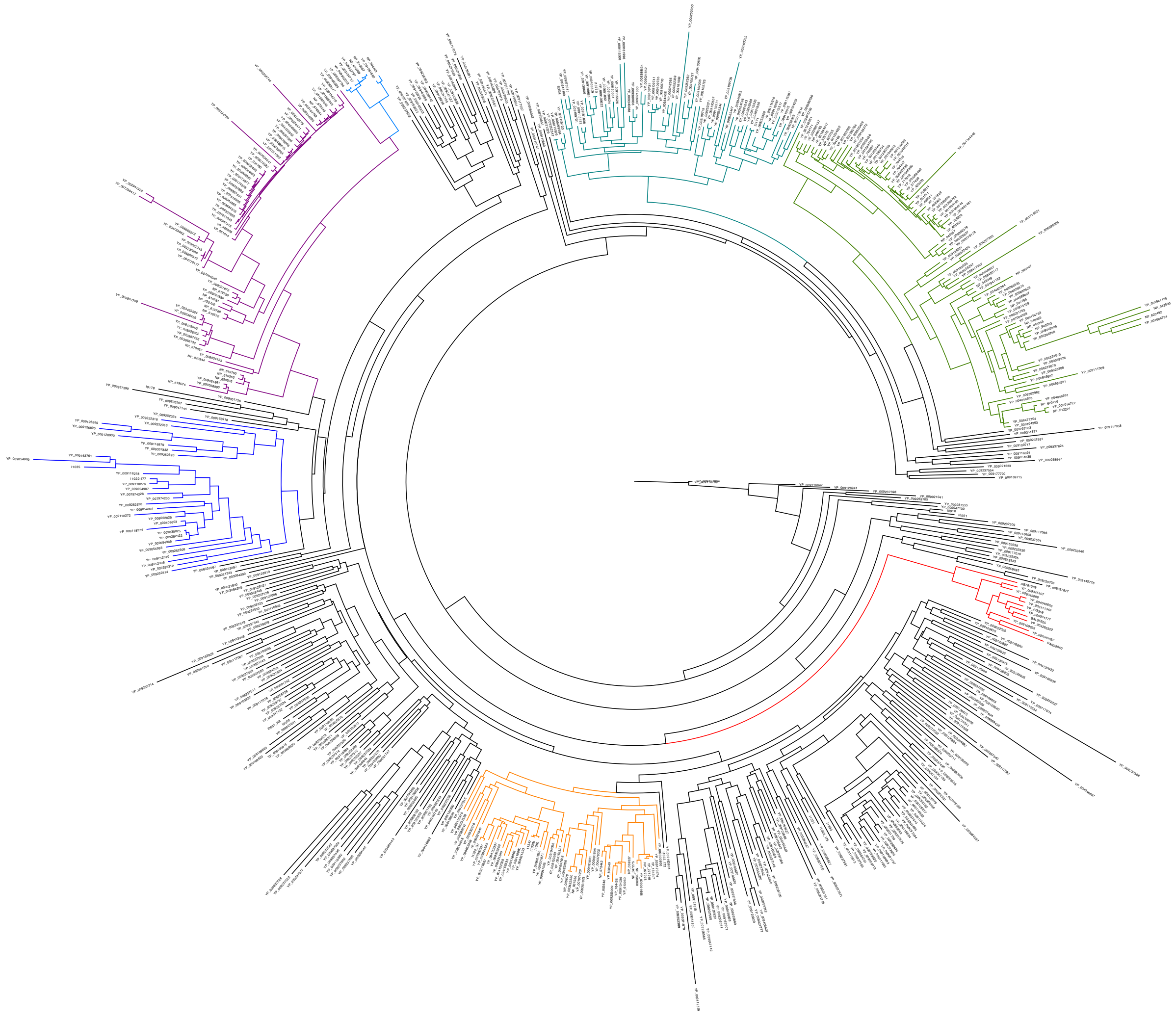
Appendix 4.11 Maximum likelihood tree built with the endonuclease portion of the full CRESS Rep MUSCLE alignment. Green for *Geminiviridae*, teal for *Genomoviridae*, dark blue for *Smacoviridae*, red for *Bacilladnaviridae*, orange for *Circoviridae*, light blue for *Nanoviridae*, purple for the alphasatellites (*Alphasatellitidae*).



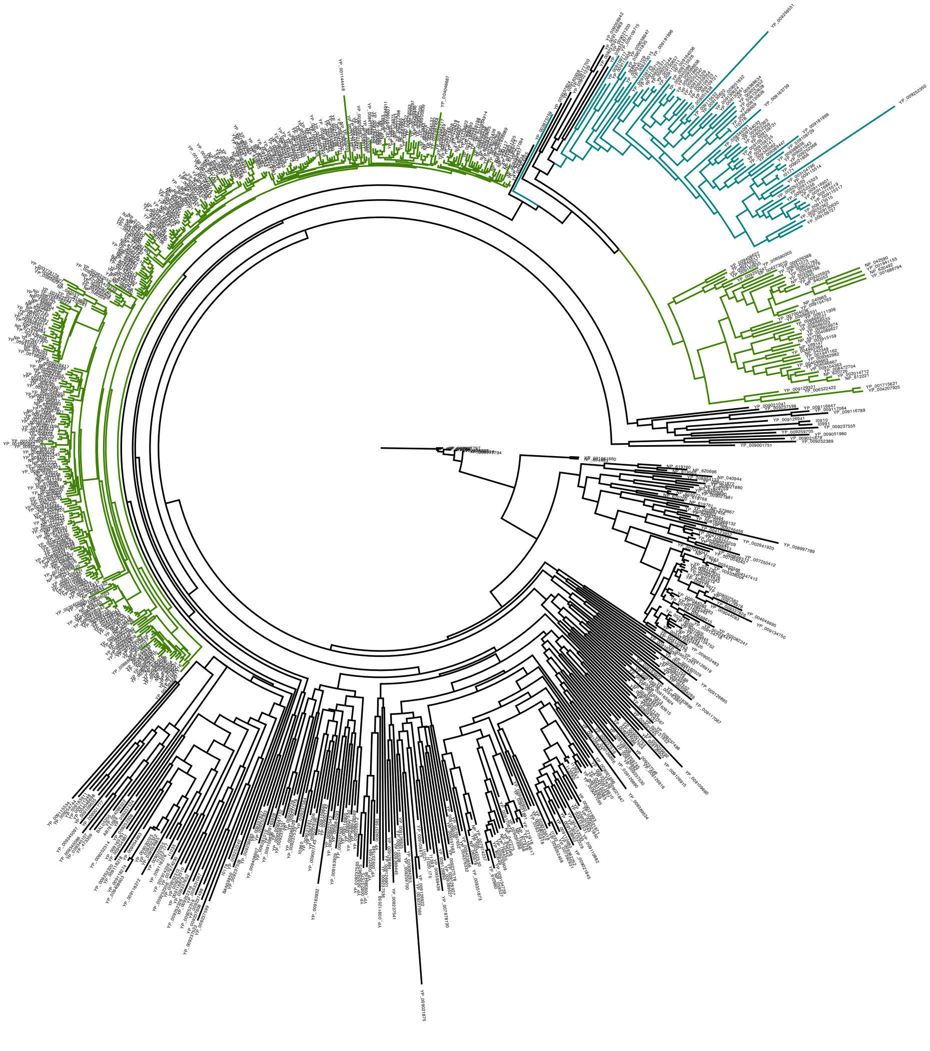
Appendix 4.12 Maximum likelihood tree built with the helicase portion of the full CRESS Rep MUSCLE alignment. Green for *Geminiviridae*, teal for *Genomoviridae*, dark blue for *Smacoviridae*, red for *Bacilladnaviridae*, orange for *Circoviridae*, light blue for *Nanoviridae*, purple for the alphasatellites (*Alphasatellitidae*).



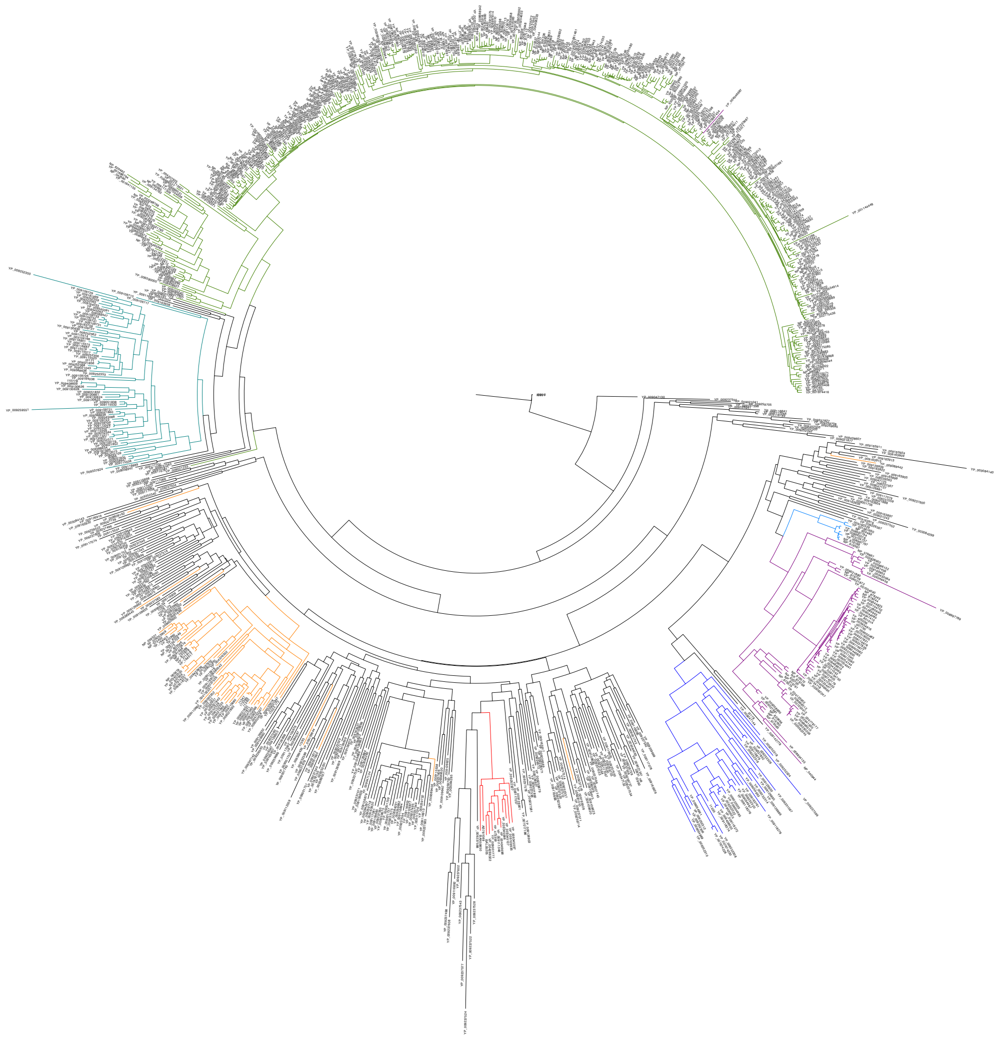
Appendix 4.13 Maximum likelihood tree with equal number of Begomovirus and Mastrevirus Reps built with CRESS. Alignment produced by MUSCLE. Green for *Geminiviridae*, teal for *Genomoviridae*, dark blue for *Smacoviridae*, red for *Bacilladnaviridae*, orange for *Circoviridae*, light blue for *Nanoviridae*, purple for the alphasatellites (*Alphasatellitidae*).



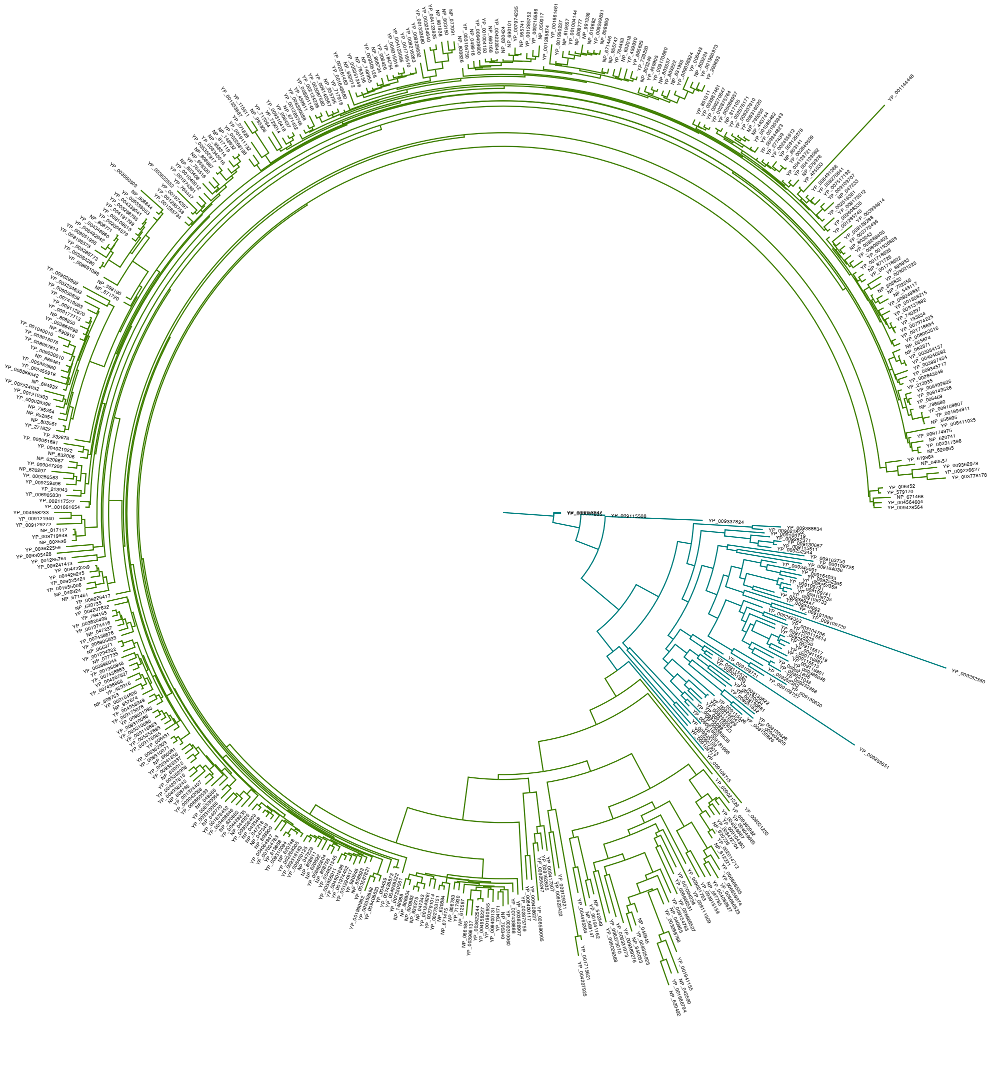
Appendix 4.14 Unrooted maximum likelihood tree built with 123 recombinant sequences removed from the 926 CRESS Rep dataset. Green for *Geminiviridae*, teal for *Genomoviridae*.



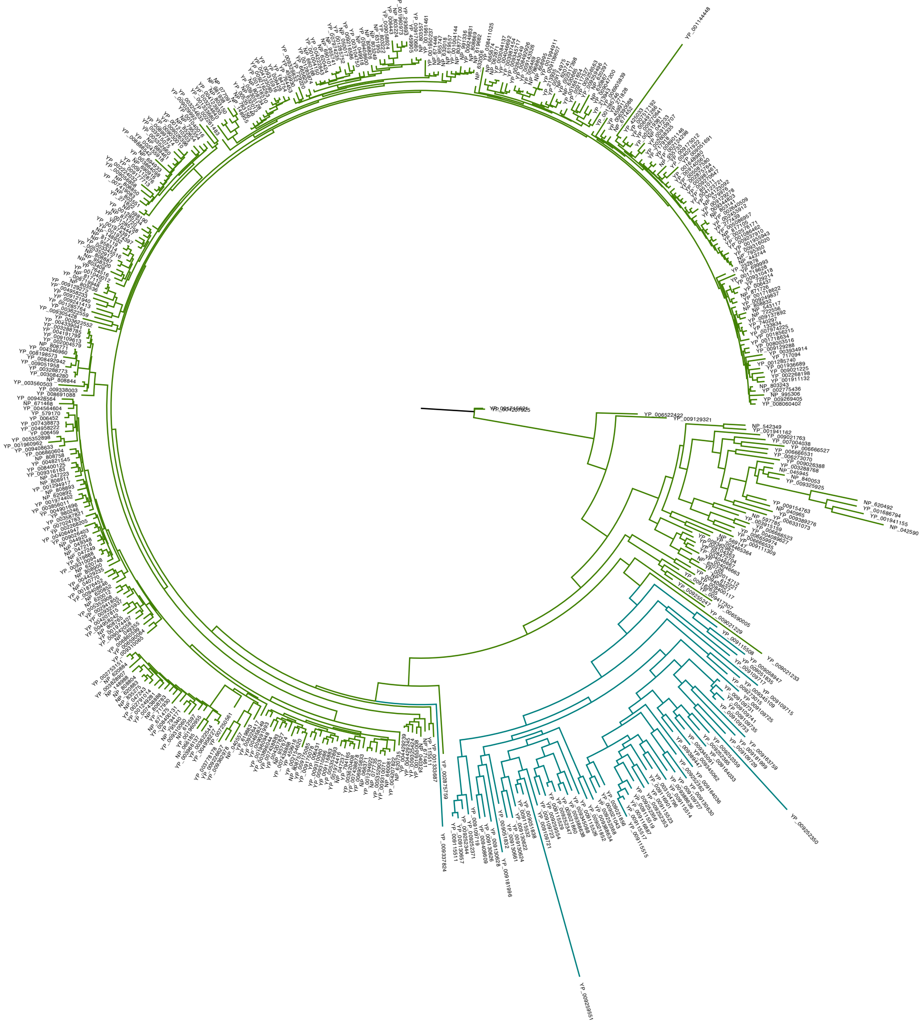
Appendix 4.15 Maximum likelihood tree built with the full 926 CRESS Rep dataset. Alignment produced by MAFFT. Green for *Geminiviridae*, teal for *Genomoviridae*, dark blue for *Smacoviridae*, red for *Bacilladnaviridae*, orange for *Circoviridae*, light blue for *Nanoviridae*, purple for the alphasatellites (*Alphasatellitidae*).



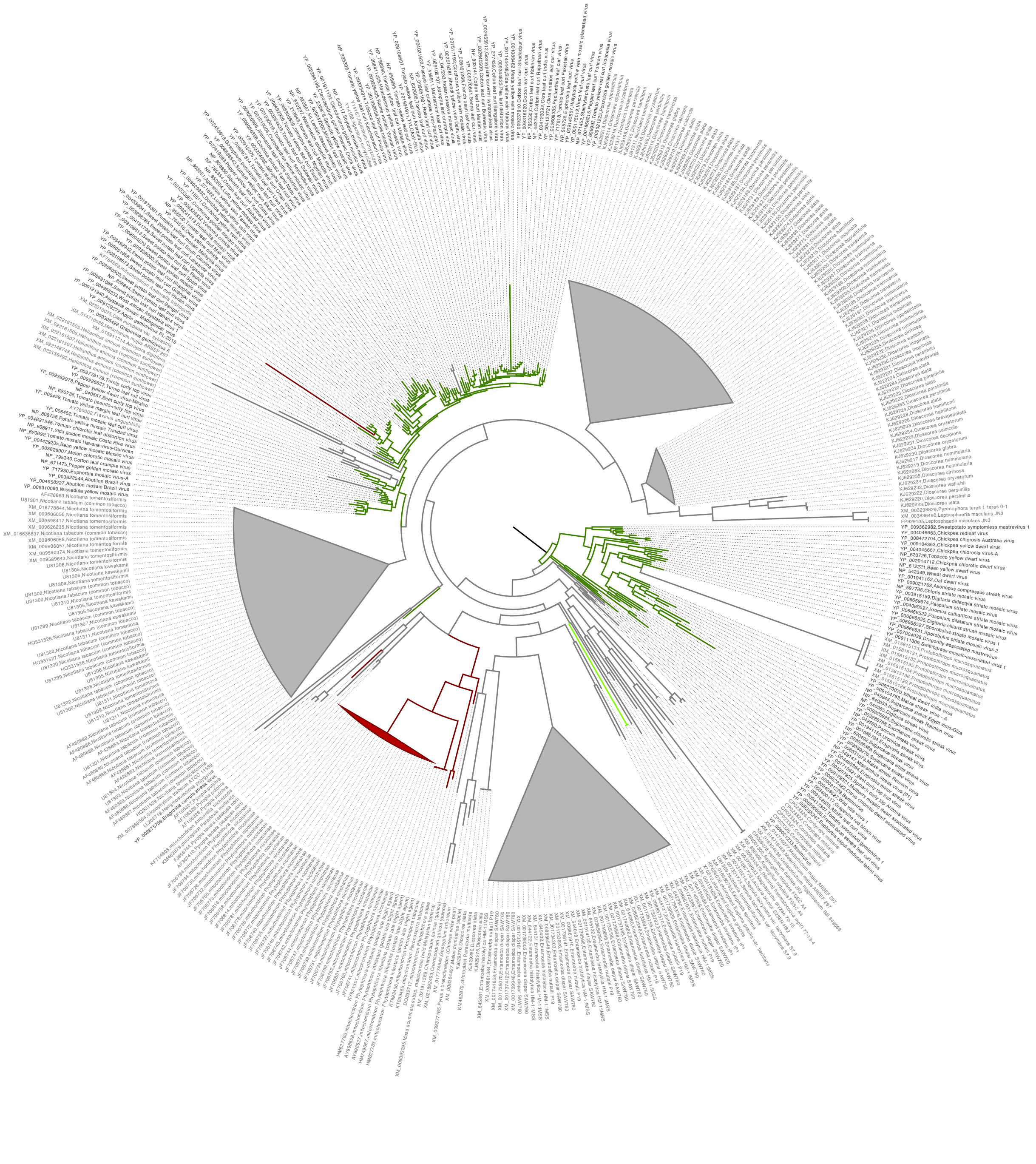
Appendix 4.16 Maximum likelihood tree with Geminivirus and Genomovirus Reps built with CRESS. Alignment product by MAFFT. Green for *Geminiviridae*, teal for *Genomoviridae*.



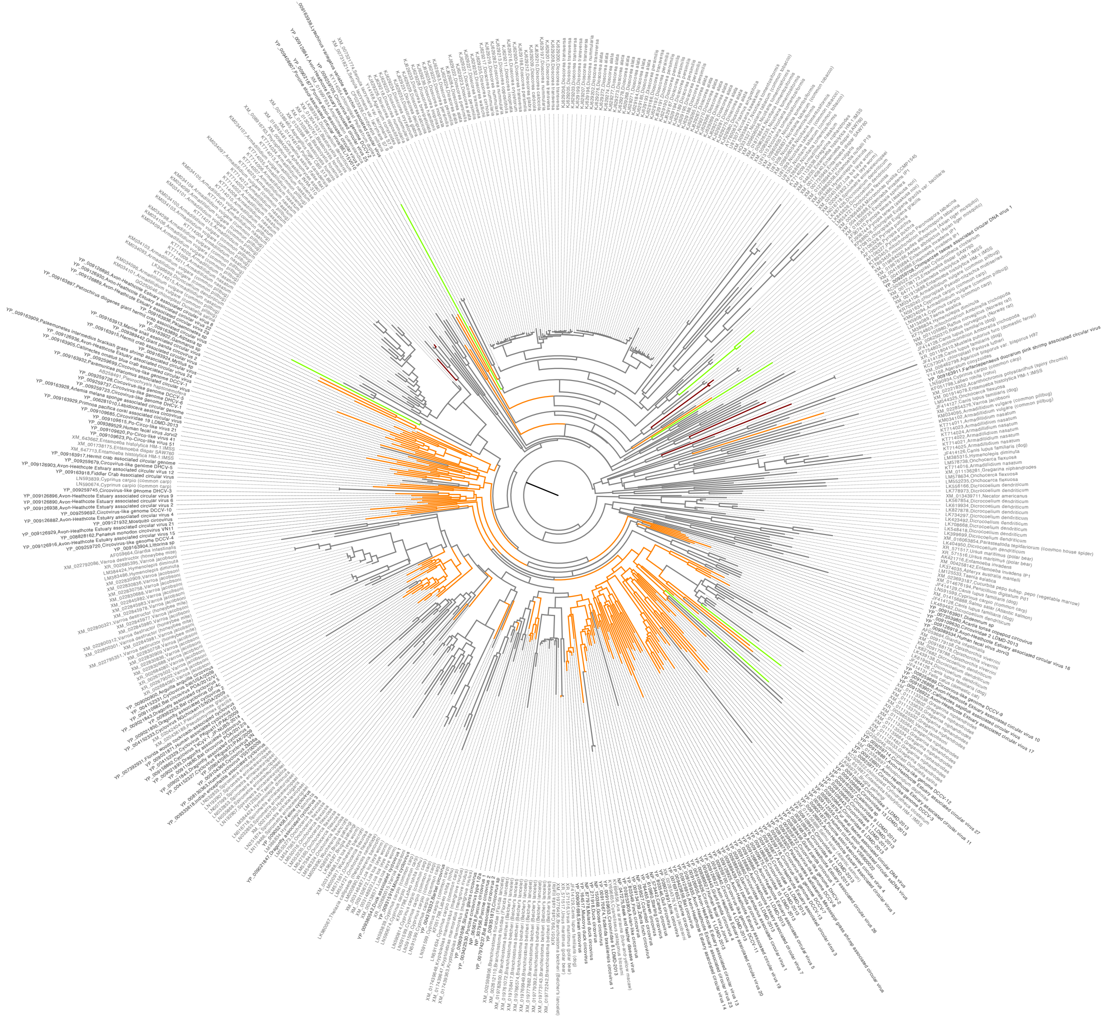
Appendix 4.17 Maximum likelihood trees with Geminivirus and Genomovirus Reps built with CRESS. Alignment produced by MUSCLE. Green for *Geminiviridae*, teal for *Genomoviridae*.



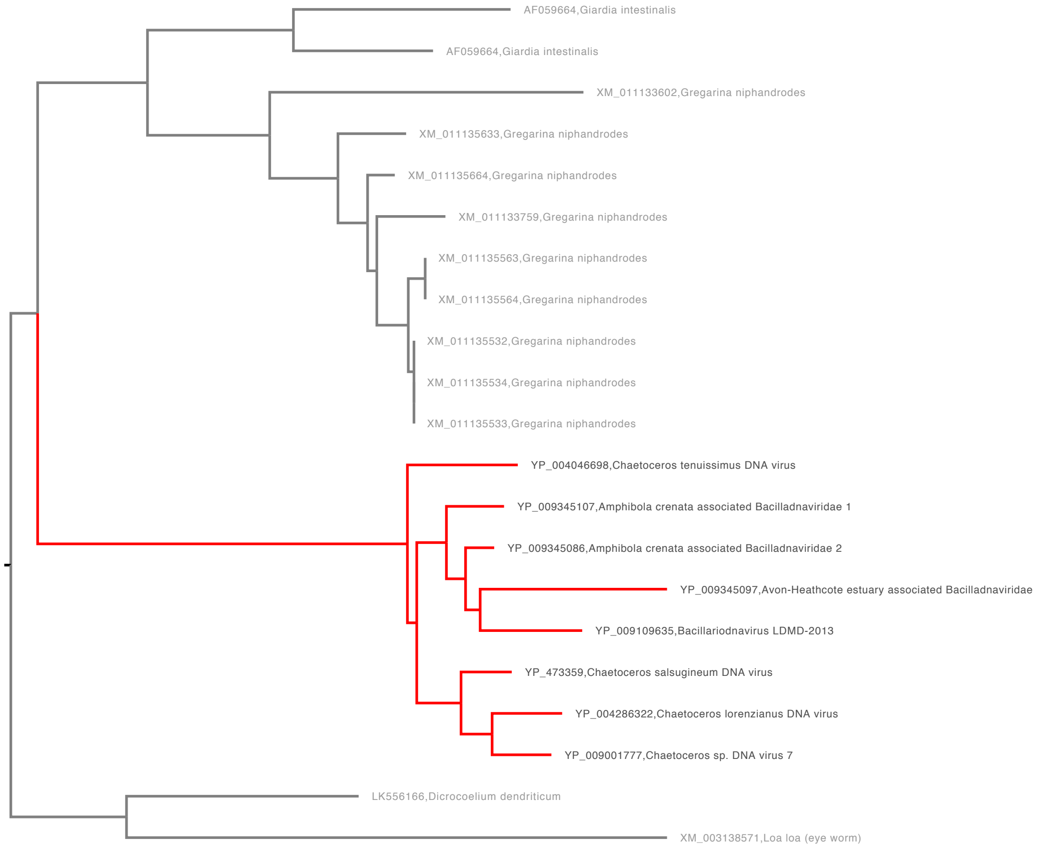
Appendix 5.1 Geminivirus Reps and endogenous sequences midpoint-rooted maximum likelihood tree. Eukaryote sequences are colored in grey, geminivirus Reps are colored green, mitochondrial sequences are colored dark red, chloroplast sequences are colored bright green. Labels show the accession numbers from which the sequences came from.



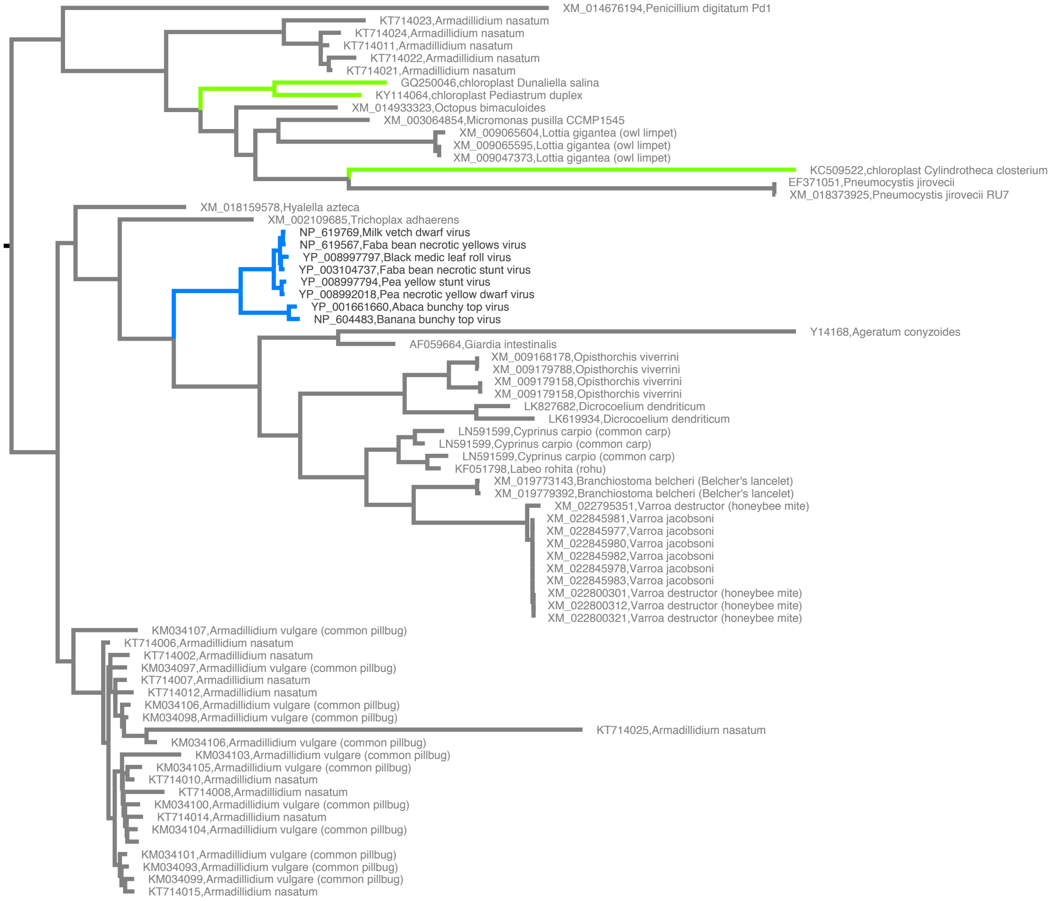
Appendix 5.2 Circovirus Reps and endogenous sequences midpoint-rooted maximum likelihood tree. Eukaryote sequences are colored in grey, circovirus Reps are colored orange, mitochondrial sequences are colored dark red, chloroplast sequences are colored bright green. Labels show the accession numbers from which the sequences came from.



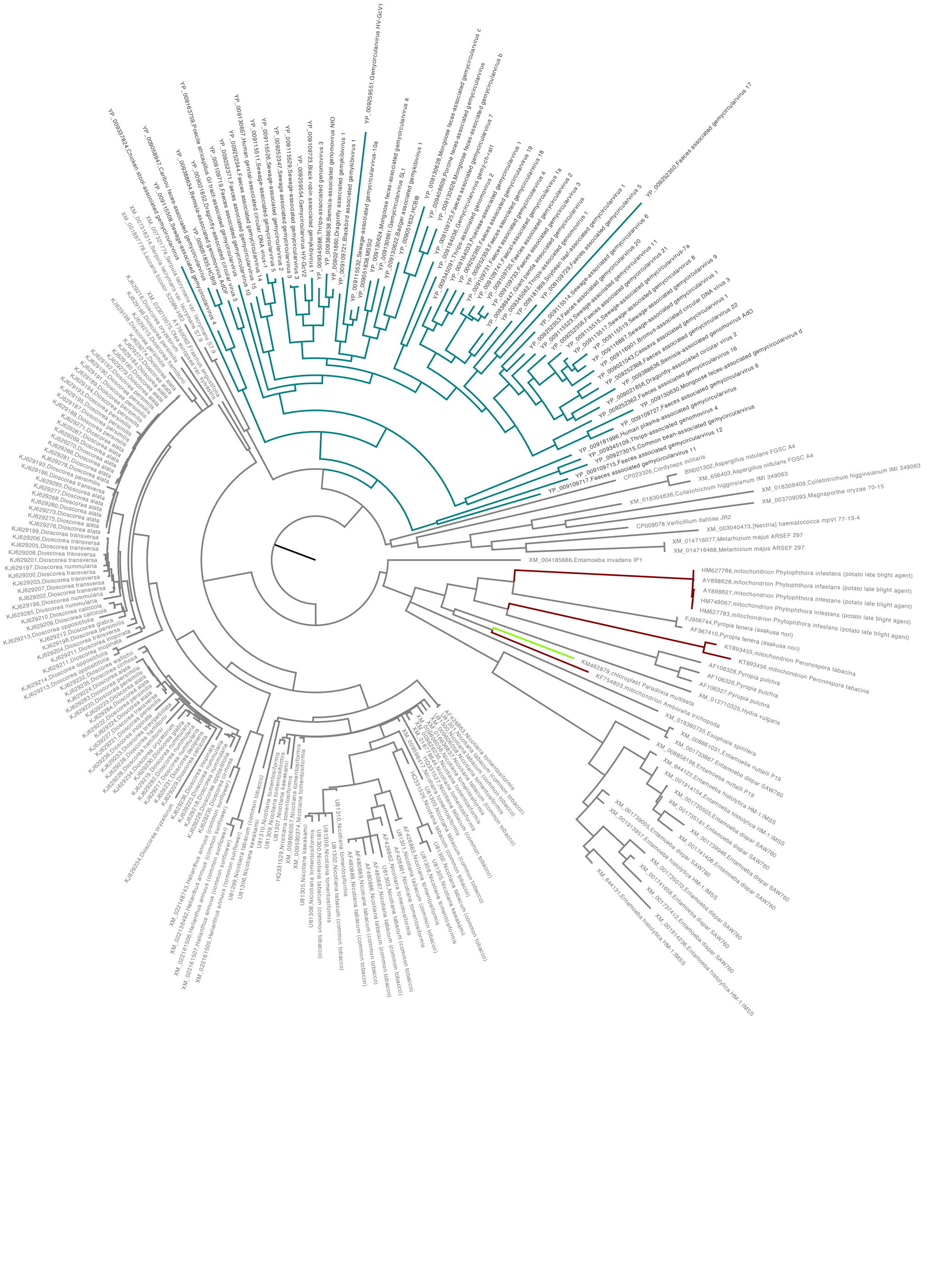
Appendix 5.3 Bacilladnavirus Reps and endogenous sequences midpoint-rooted maximum likelihood tree. Eukaryote sequences are colored in grey, bacilladnavirus Reps are colored red. Labels show the accession numbers from which the sequences came from.


AF059664,Giardia intestinalis
AF059664,Giardia intestinalis
XM_011133602,Gregarina niphandrodes
XM_011135633,Gregarina niphandrodes
XM_011135664,Gregarina niphandrodes
XM_011133759,Gregarina niphandrodes
XM_011135563,Gregarina niphandrodes
XM_011135564,Gregarina niphandrodes
XM_011135532,Gregarina niphandrodes
XM_011135534,Gregarina niphandrodes
XM_011135533,Gregarina niphandrodes
YP_004046698,Chaetoceros tenuissimus DNA virus
YP_009345107,Amphibola crenata associated Bacilladnaviridae 1
YP_009345086,Amphibola crenata associated Bacilladnaviridae 2
YP_009345097,Avon-Heathcote estuary associated Bacilladnaviridae
YP_009109635,Bacillariodnavirus LDMD-2013
YP_473359,Chaetoceros salsugineum DNA virus
YP_004286322,Chaetoceros lorenzianus DNA virus
YP_009001777,Chaetoceros sp. DNA virus 7
LK556166,Dicrocoelium dendriticum
XM_003138571,Loa loa (eye worm)

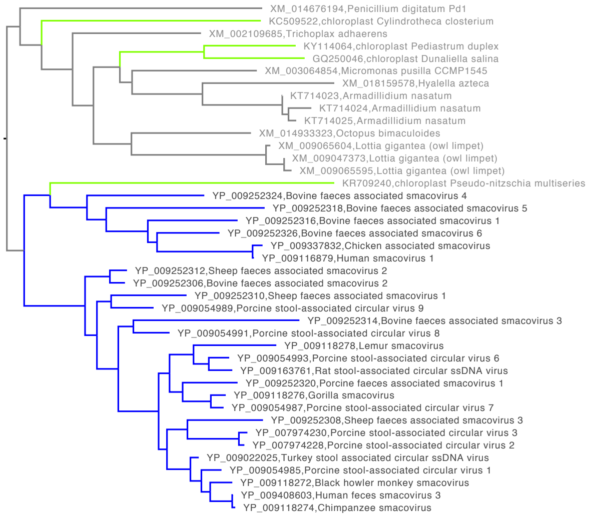
Appendix 5.4 Nanovirus Reps and endogenous sequences midpoint-rooted maximum likelihood tree. Eukaryote sequences are colored in grey, nanovirus Reps are colored blue, chloroplast sequences are colored bright green. Labels show the accession numbers from which the sequences came from.


XM_014676194,Penicillium digitatum Pd1
KT714023,Armadillidium nasatum
KT714024,Armadillidium nasatum
KT714011,Armadillidium nasatum
KT714022,Armadillidium nasatum
KT714021,Armadillidium nasatum
GQ250046,chloroplast Dunaliella salina
KY114064,chloroplast Pediastrum duplex
XM_014933323,Octopus bimaculoides
XM_003064854,Micromonas pusilla CCMP1545
XM_009065604,Lottia gigantea (owl limpet)
XM_009065595,Lottia gigantea (owl limpet)
XM_009047373,Lottia gigantea (owl limpet)
KC509522,chloroplast Cylindrotheca closterium
EF371051,Pneumocystis jirovecii
XM_018373925,Pneumocystis jirovecii RU7
XM_018159578,Hyalella azteca
XM_002109685,Trichoplax adhaerens
NP_619769,Milk vetch dwarf virus
NP_619567,Faba bean necrotic yellows virus
YP_008997797,Black medic leaf roll virus
YP_003104737,Faba bean necrotic stunt virus
YP_008997794,Pea yellow stunt virus
YP_008992018,Pea necrotic yellow dwarf virus
YP_001661660,Abaca bunchy top virus
NP_604483,Banana bunchy top virus
Y14168,Ageratum conyzoides
AF059664,Giardia intestinalis
XM_009168178,Opisthorchis viverrini
XM_009179788,Opisthorchis viverrini
XM_009179158,Opisthorchis viverrini
XM_009179158,Opisthorchis viverrini
LK827682,Dicrocoelium dendriticum
LK619934,Dicrocoelium dendriticum
LN591599,Cyprinus carpio (common carp)
LN591599,Cyprinus carpio (common carp)
LN591599,Cyprinus carpio (common carp)
KF051798,Labeo rohita (rohu)
XM_019773143,Branchiostoma belcheri (Belcher's lancelet)
XM_019779392,Branchiostoma belcheri (Belcher's lancelet)
XM_022795351,Varroa destructor (honeybee mite)
XM_022845981,Varroa jacobsoni
XM_022845977,Varroa jacobsoni
XM_022845980,Varroa jacobsoni
XM_022845982,Varroa jacobsoni
XM_022845978,Varroa jacobsoni
XM_022845983,Varroa jacobsoni
XM_022800301,Varroa destructor (honeybee mite)
XM_022800312,Varroa destructor (honeybee mite)
XM_022800321,Varroa destructor (honeybee mite)
KM034107,Armadillidium vulgare (common pillbug)
KT714006,Armadillidium nasatum
KT714002,Armadillidium nasatum
KM034097,Armadillidium vulgare (common pillbug)
KT714007,Armadillidium nasatum
KT714012,Armadillidium nasatum
KM034106,Armadillidium vulgare (common pillbug)
KM034098,Armadillidium vulgare (common pillbug)
KT714025,Armadillidium nasatum
KM034106,Armadillidium vulgare (common pillbug)
KM034103,Armadillidium vulgare (common pillbug)
KM034105,Armadillidium vulgare (common pillbug)
KT714010,Armadillidium nasatum
KT714008,Armadillidium nasatum
KM034100,Armadillidium vulgare (common pillbug)
KT714014,Armadillidium nasatum
KM034104,Armadillidium vulgare (common pillbug)
KM034101,Armadillidium vulgare (common pillbug)
KM034093,Armadillidium vulgare (common pillbug)
KM034099,Armadillidium vulgare (common pillbug)
KT714015,Armadillidium nasatum

Appendix 5.5 Genomovirus Reps and endogenous sequences midpoint-rooted maximum likelihood tree. Eukaryote sequences are colored in grey, genomovirus Reps are colored teal, mitochondrial sequences are colored dark red, chloroplast sequences are colored bright green. Labels show the accession numbers from which the sequences came from.


YP_009259551,Gemycircularvirus HV-GcV1
YP_009252350,Faeces associated gemycircularvirus 17
YP_009130630,Mongoose feces-associated gemycircularvirus d
YP_009252362,Faeces associated gemycircularvirus 16
YP_009021856,Dragonfly-associated circular virus 2
YP_009109727,Faeces associated gemycircularvirus 6
YP_009181996,Human plasma-associated gemycircularvirus
YP_009345109,Thrips-associated genomovirus 4
YP_009273015,Common bean-associated gemycircularvirus
YP_009109715,Faeces associated gemycircularvirus 12
YP_009109717,Faeces associated gemycircularvirus 11
CP023326,Cordyceps militaris
BN001302,Aspergillus nidulans FGSC A4
XM_656403,Aspergillus nidulans FGSC A4
XM_018304636,Colletotrichum higginsianum IMI 349063
XM_018309409,Colletotrichum higginsianum IMI 349063
XM_003709093,Magnaporthe oryzae 70-15
CP009078,Verticillium dahliae JR2
XM_003040473,[Nectria] haematococca mpVI 77-13-4
XM_014716077,Metarhizium majus ARSEF 297
XM_014716468,Metarhizium majus ARSEF 297
XM_004185686,Entamoeba invadens IP1
HM627786,mitochondrion Phytophthora infestans (potato late blight agent)
AY898628,mitochondrion Phytophthora infestans (potato late blight agent)
AY898627,mitochondrion Phytophthora infestans (potato late blight agent)
HM749067,mitochondrion Phytophthora infestans (potato late blight agent)
HM627783,mitochondrion Phytophthora infestans (potato late blight agent)
FJ956744,Pyropia tenera (asakusa nori)
AF367410,Pyropia tenera (asakusa nori)
KT893455,mitochondrion Peronospora tabacina
KT893456,mitochondrion Peronospora tabacina
AF106328,Pyropia pulchra
AF106326,Pyropia pulchra
AF106327,Pyropia pulchra
XM_012710325,Hydra vulgaris
KM462979,chloroplast Paradoxia multiseta
KF754803,mitochondrion Amborella trichopoda
XM_016380735,Exophiala spinifera

Appendix 5.6 Smacovirus Reps and endogenous sequences midpoint-rooted maximum likelihood tree. Eukaryote sequences are colored in grey, smacovirus Reps are colored blue, , chloroplast sequences are colored bright green. Labels show the accession numbers from which the sequences came from.


XM_014676194,Penicillium digitatum Pd1
KC509522,chloroplast Cylindrotheca closterium
XM_002109685,Trichoplax adhaerens
KY114064,chloroplast Pediastrum duplex
GQ250046,chloroplast Dunaliella salina
XM_003064854,Micromonas pusilla CCMP1545
XM_018159578,Hyalella azteca
KT714023,Armadillidium nasatum
KT714024,Armadillidium nasatum
KT714025,Armadillidium nasatum
XM_014933323,Octopus bimaculoides
XM_009065604,Lottia gigantea (owl limpet)
XM_009047373,Lottia gigantea (owl limpet)
XM_009065595,Lottia gigantea (owl limpet)
KR709240,chloroplast Pseudo-nitzschia multiseries
YP_009252324,Bovine faeces associated smacovirus 4
YP_009252318,Bovine faeces associated smacovirus 5
YP_009252316,Bovine faeces associated smacovirus 1
YP_009252326,Bovine faeces associated smacovirus 6
YP_009337832,Chicken associated smacovirus
YP_009116879,Human smacovirus 1
YP_009252312,Sheep faeces associated smacovirus 2
YP_009252306,Bovine faeces associated smacovirus 2
YP_009252310,Sheep faeces associated smacovirus 1
YP_009054989,Porcine stool-associated circular virus 9
YP_009252314,Bovine faeces associated smacovirus 3
YP_009054991,Porcine stool-associated circular virus 8
YP_009118278,Lemur smacovirus
YP_009054993,Porcine stool-associated circular virus 6
YP_009163761,Rat stool-associated circular ssDNA virus
YP_009252320,Porcine faeces associated smacovirus 1
YP_009118276,Gorilla smacovirus
YP_009054987,Porcine stool-associated circular virus 7
YP_009252308,Sheep faeces associated smacovirus 3
YP_007974230,Porcine stool-associated circular virus 3
YP_007974228,Porcine stool-associated circular virus 2
YP_009022025,Turkey stool associated circular ssDNA virus
YP_009054985,Porcine stool-associated circular virus 1
YP_009118272,Black howler monkey smacovirus
YP_009408603,Human feces smacovirus 3
YP_009118274,Chimpanzee smacovirus

Appendix 5.7 Alphasatellite Reps and endogenous sequences midpoint-rooted maximum likelihood tree. Eukaryote sequences are colored in grey, alphasatellite Reps are colored purple, chloroplast sequences are colored bright green. Labels show the accession numbers from which the sequences came from.