

MOLECULAR EVOLUTION AND PHYLOGENETICS  
OF CIRCULAR SINGLE-STRANDED DNA VIRUSES

By

YEE MEY SEAH

A dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

And

The Graduate School of Biomedical Sciences

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Microbiology and Molecular Genetics

Written under the direction of

Siobain Duffy

And approved by

---

---

---

---

New Brunswick, New Jersey

October, 2015

## ABSTRACT OF THE DISSERTATION

Molecular Evolution and Phylogenetics of Circular Single-stranded DNA Viruses

by YEE MEY SEAH

Dissertation Director:

Siobain Duffy, Ph.D.

Viruses infect a wide variety of hosts across all domains of life. Despite their ubiquity, and a long history of virus research, fundamental questions such as what constitutes a virus species, and how viruses evolve and are modeled, have yet to be adequately answered. We compared viral sequences across five genomic architectures (single- and double-stranded DNA and RNA) and demonstrate the presence of substitution bias, especially in single-stranded viruses. Most striking is a consistent pattern of over-represented cytosine-to-thymine substitutions in single-stranded DNA (ssDNA) viruses. This led us to question the validity of using time-reversible nucleotide substitution models in viral phylogenetic inference, as these models assume equal rates of forward and reverse substitutions. We found that an unrestricted substitution model fit the data better for most single- and double-stranded viral datasets, as measured by corrected Akaike Information Criterion, and hierarchical likelihood ratio test scores. We also approached the question of virus species identification by examining members of the most species-rich viral genus *Begomovirus* (Family

*Geminiviridae*), which are circular single-stranded DNA (ssDNA) viral crop pathogens transmitted by whitefly vectors. We used novel sweet potato-infecting begomoviruses (sweepoviruses) collected during a recent vector-enabled metagenomic survey to evaluate the concept of pairwise percent nucleotide identity threshold as a criterion for species demarcation. We demonstrate that species demarcation based on pairwise percent nucleotide identities group divergent sweepovirus clusters together, and is highly influenced by when, and how much sampling occurs.

## **ACKNOWLEDGEMENTS**

NATIONAL SCIENCE FOUNDATION

RUTGERS ARESTY RESEARCH CENTER

RUTGERS DIVISION OF LIFE SCIENCES OFFICE OF UNDERGRADUATE INSTRUCTION

Jana Curry, Gregg Transue

DUFFY LAB

Kendra Avinger, Cassandra Burdziak, Daniel Cardinale, Brian Cully, Shamyla Din, Preshita

Gadkari, Allison Hicks, Natasia Jacko, Dylan McClung, Jennifer Mcconnell, Lisa Nigro,

Yuliya Olifer, Amy G. Patel, Aubrey Watson, LaShanda Williams, Lele Zhao

COLLABORATORS & OTHER LABS

Mya Breitbart, Karyna Rosario | Alison Lima, F. Murilo Zerbini | Thomas Leustek, Rita Pasini |

Hung-kuang Chen, Gerben Zylstra | Sarah Doore, Bentley Fane | Bryan Gelfand,

Dibyendu Kumar, Nicole Wagner, Udi Zelzion

MICROBIOLOGY AND MOLECULAR GENETICS

Carolyn Ambrose, Diane Murano, Andrew Vershon

ECOLOGY, EVOLUTION AND NATURAL RESOURCES

Alison Cariveau, Marsha Morin, Christine Tizzano

FAMILY & FRIENDS



## TABLE OF CONTENTS

Abstract	ii
Acknowledgements	iv
List of tables	vi
List of illustrations	vii
Preface	ix
Introduction	1
Chapter 1 – 98% identical, 100% wrong: Percent nucleotide identity can lead plant virus epidemiology astray	5
Chapter 2 – Cause for UNREST: Unrestricted nucleotide substitution model more likely than general time-reversible in virus phylogenetics	27
Chapter 3 – No clean sweep for the sweepoviruses: Problems with pairwise percent nucleotide identity as a definitive classification tool	52
Chapter 4 – Exploring cytosine→thymine bias in bacteriophage phiX174	82
Appendix	98
Bibliography	117

## LIST OF TABLES

### CHAPTER 1

1. The number of short-form reports in three plant pathology journals that involved sequencing of plant viruses from 2007 to 2009 (2/07 – 1/10 for <i>New Disease Reports</i> ).	10
--	----

### CHAPTER 2

1. Substitution bias analyses	38
S1. Substitution bias analyses on selected codon positions.	101
S2. Virus species in analyses	103
S3. Nucleotide substitution models used for inferring maximum likelihood trees in PAUP*	107

### CHAPTER 3

1. Recombination events and best-matching pairwise nucleotide identities of novel VEM sweepovirus isolates.	66
2. Recombination events and top two best-matching pairwise nucleotide identities for SPLCESV isolates.	68
S1. Sweepovirus sequences downloaded from GenBank for analyses.	111

### CHAPTER 4

1. Primers used in site-directed mutagenesis and sequence confirmation.	87
2. Master mix used in site-directed mutagenesis.	88
3. PCR programs for site-directed mutagenesis and regular PCR amplification.	90
S1. Sequence of wildtype phiX174 obtained from Dr. Bentley A. Fane.	115

## LIST OF ILLUSTRATIONS

### CHAPTER 1

- |   |    |
|---|----|
| 1. Maximum-likelihood tree of a 488-base alignment of partial coat protein sequences of New World isolates of tomato yellow leaf curl viruses with three non-North American isolates, constructed with a Tamura-Nei (TrN) model of nucleotide substitution.                               | 18 |
| 2. Maximum-likelihood tree of a 513-base alignment of coat protein sequences (the 513-base ORF on DNA-3) of <i>Banana bunchy top virus</i> , constructed with a TrN substitution model including the fraction of invariant sites (I) and substitution rate heterogeneity among sites (G). | 21 |
| 3. Maximum-likelihood tree of a 309-base alignment of partial heat shock protein 70 sequences of tomato chlorosis virus, constructed with a transversional (TVM) model of nucleotide substitution.  | 24 |

### CHAPTER 2

- |  |     |
|--|-----|
| 1. Difference in AICc scores between UNREST and GTR substitution models for 40 virus species.  | 42  |
| 2. Relative skew of observed substitutions   | 44  |
| S1. Difference in AICc scores between UNREST and GTR (with ML-estimated nucleotide frequency parameters) substitution models for 40 virus species.                               | 98  |
| S2. Difference in AICc scores, and difference in BIC scores between UNREST and GTR (with ML-estimated nucleotide frequency parameters) substitution models for 40 virus species. | 100 |

### CHAPTER 3

- |  |    |
|--|----|
| 1. Pairwise percent nucleotide identities between all isolates currently identified as <i>Sweet potato leaf curl</i> and non-SPLCV isolates (top), and between all SPLCV isolates only (bottom). | 60 |
| 2a. Maximum likelihood phylogeny of 168 sweepovirus isolates   | 65 |

2b. Pairwise percent nucleotide identity matrix corresponding to sweepovirus clades containing VEM, SPLCV, SPLCUV, and SPLCSCV isolates.	65
3a. Sweepovirus phylogeny as in 2a, with the top clades expanded, and the bottom portion collapsed.	67
3b. Pairwise percent nucleotide identity matrix corresponding to sweepovirus clades containing SPLCV, SPLCHnV, SPLCCaV, SPLCGoV, SPLCCNV, SPLCSiV-1 and 2, SPLCShV, and SPLCGxV isolates.	67
4. Cladogram of sweepovirus complete genomes (n=168).	70
5. Pairwise percent nucleotide identities between the ICTV reference sequences for SPLCV and SPLCESV, and all available sequences of SPLCV.	71
6. SPLCV and SPLCESV sequence clusters merging due to sampling of three additional isolates.	79
S1. Cladogram of sweepovirus genomes without the long intergenic region.	109
S2. Cladogram of sweepovirus coat protein amino acid sequences.	110
CHAPTER 4	
1a. PCR-based site-directed mutagenesis of phiX174.	92
1b. Joining of the two mutagenized DNA fragments.	92
2. Plaque pick assay for confirmation of successful recovery of mutant bacteriophage.	94
3. Titers of virions stored in 4°C and 45°C.	95

## PREFACE

Chapter 1 has been published as “98% identical, 100% wrong: per cent nucleotide identity can lead plant virus epidemiology astray,” (Duffy, S., and Seah, Y.M. *Philosophical Transactions of the Royal Society B*. 2010. 365(1548):1891-7).

Chapter 2 is in preparation for publication under the title “Cause for UNREST: Unrestricted nucleotide substitution model more likely than general time-reversible in virus phylogenetics,” by Seah, Y.M., Burdziak C., McClung, D., and Duffy, S., to be submitted to *Molecular Biology and Evolution*.

Chapter 3 has been submitted to *Virus Evolution*, as “No clean sweep for the sweepoviruses: Problems with pairwise percent nucleotide identity as a definitive classification tool,” by Seah, Y.M., Rosario K., Breitbart, M., and Duffy, S.

## INTRODUCTION

Despite having an arguably basic parasitic lifestyle, viruses have a wide range of characteristics, from their hosts, which span the entire Tree of Life, to their genome type, structure, and size (the smallest known genome is that of the Porcine circovirus-2 at 1.8 kb (Hamel et al. 1998), and the largest Pandoravirus is 1,000 times larger at 2.5 Mb (Philippe et al. 2013)). Viruses are the most numerous biological entities on Earth (Edwards and Rohwer 2005), and can be found in harsh environmental conditions where temperature and acidity are extremely high (Ortmann and Suttle 2005; Iverson and Stedman 2012). Although viruses are ubiquitous, many fundamental questions about how they evolve, and what constitutes a virus species remain unanswered.

The most straightforward way to distinguish between viruses on a rough scale is by the Baltimore classification that groups them by genome structure: classes I through V are respectively, double-stranded DNA (dsDNA), single-stranded DNA (ssDNA), double-stranded RNA (dsRNA), positive-sense and negative-sense single-stranded RNA (+ssRNA, -ssRNA) (Baltimore 1971). Reverse-transcribing viruses fall into classes VI and VII (Baltimore 1971). Taxonomic classification on the other hand, is trickier since viruses do not share a single common feature across all types, unlike for example, bacteria that all have 16S ribosomal RNA sequence that can be used to delineate different groups. Nevertheless, the International Committee on Taxonomy of Viruses (ICTV) was created in 1966

under the auspices of the Virology Division of the International Union of Microbiological Societies and tasked with developing a “single, universal taxonomic scheme for all the viruses infecting animals (vertebrates, invertebrates, and protozoa), plants (higher plants and algae), fungi, bacteria, and archaea (King AMQ, Adams MJ, Carstens EB, Lefkowitz 2012).”

One of the many fundamental questions dealt with by the ICTV includes the nature and definition of a virus species. After much debate, ICTV adopted the proposal that a virus species is “a polythetic class of viruses that constitute a replicating lineage and occupies a particular ecological niche (King AMQ, Adams MJ, Carstens EB, Lefkowitz 2012).” Experts in each virus family are responsible for determining the exact details of classification based on this definition. While determination of virus biological properties is often recommended for confirmation of identity, improved sequencing technology and increased metagenomic sampling have led to the availability of massive amounts of virus sequence data (Bao et al. 2004). This means that taxonomic level classification of many viral isolates depends increasingly on the relative similarity of their genomes to those of other known isolates. Genome comparisons can be done by basic percent identity calculations, or phylogenetic methods that incorporate evolutionary information; both have benefits and disadvantages that are explored in Chapters 1 and 3.

Besides for classification purposes, analyses of virus genomes also provide keys to viral evolutionary patterns and processes. Mutations, substitutions (mutations that become fixed in the population), and recombination are all processes that generate detectable variation in viruses. Viral genome sequences are also used to elucidate evolutionary relationships between isolates through phylogenetic inference methods.

In this thesis, I focus on the molecular evolution and phylogenetics of Class II ssDNA viruses. These viruses can have either circular or linear genomes; all of the research presented here focuses on circular ssDNA viruses. A major group of circular ssDNA viruses are the *Begomovirus* species, which are crop pathogens transmitted by whitefly vectors (Brown et al. 2015). I use novel virus sequences collected via vector-enabled metagenomic sampling (Rosario et al. 2015) to determine the phylogenies of sweet potato-infecting begomoviruses in Chapter 3.

Having circular single-stranded genomes with few permanent secondary structures leaves these viruses more vulnerable to DNA damage than viruses with double-stranded, complementary strands. One type of oxidative damage that can occur is the deamination of cytosine to uracil, which, following one round of replication will manifest as a cytosine-to-thymine transition in the genomes. Single-stranded nucleic acids have been demonstrated to experience more cytosine deamination than double-stranded nucleic acids do (Lindahl and Nyberg



1974; Frederico et al. 1990). Increased cytosine-to-thymine substitution may contribute to the high rates of ssDNA virus substitution, which are on the same order of magnitude as the fast-evolving ssRNA viruses (Duffy et al. 2008). I demonstrate a consistent pattern of significant over-representation of cytosine-to-thymine transitions specifically in ssDNA virus genomes in Chapter 2.

This leads to evaluation of the likelihood of an alternative nucleotide substitution model that incorporates unequal rates of forward and reverse substitutions, in describing the viral sequence data. Nucleotide substitution models are used in inference of phylogenetic relationships between organisms, and models that are commonly employed today are derivatives of a model that assumes reversible rates of substitution (Sumner et al. 2012). I determine that the additional parameters introduced by non-reversible forward and reverse substitution rates allow the alternative model to fit the viral data better than the null model. I also explore the cytosine-to-thymine substitution bias in ssDNA genomes with a laboratory model system comprising the ssDNA bacteriophage phiX174 and its *E. coli* host in my final chapter.

## CHAPTER 1

### **98% IDENTICAL, 100% WRONG: PERCENT NUCLEOTIDE IDENTITY CAN LEAD PLANT VIRUS EPIDEMIOLOGY ASTRAY**

#### **Abstract**

Short-form publications such as Plant Disease reports serve essential functions: the rapid dissemination of information on the geography of established plant pathogens, incidence and symptomology of pathogens in new hosts, and the discovery of novel pathogens. Many of these sentinel publications include viral sequence data, but most use that information only to confirm the virus' species. When researchers use the standard technique of percent nucleotide identity to determine that the new sequence is closely related to another sequence, potentially erroneous conclusions can be drawn from the results. Multiple introductions of the same pathogen into a country are being ignored because researchers know fast-evolving plant viruses can accumulate substantial sequence divergence over time, even from a single introduction. An increased use of phylogenetic methods in short-form publications could speed our understanding of these cryptic second introductions and aid in control of epidemics.

## **Introduction**

In the mid-1990s, the emerging and damaging tomato yellow leaf curl virus (TYLCV) was found in tomato plants in the Caribbean. This Old World virus had never before been seen in the New World and quickly spread to other Caribbean islands, to eastern Mexico and to Florida. From Florida it spread North and West, reaching Alabama in 2005 (Akad et al. 2007) and Texas in 2006 (Isakeit et al. 2007). At the same time, tomato plants in western Mexico were succumbing to TYLCV (Brown and Idris 2006). As the western Mexican sequences were 98 per cent identical to those from the eastern Caribbean, the infections in western Mexico were thought to be an extension of the initial introduction (Idris et al. 2007). Despite this substantial percent nucleotide identity (PNI), subsequent phylogenetic analysis revealed that these western Mexican sequences were more closely related (greater than 99%) to TYLCV in Asia than in other North American isolates and represented a second introduction of this exotic virus into North America (Duffy and Holmes 2007).

If plant viruses were typically introduced into new locations once and only once, then the different, more complete perspective that phylogenetics provided for TYLCV infections in the New World would be interesting, but of ultimately limited value. We know, however, that multiple introductions are a frequent occurrence. In the case of TYLCV in the New World, a third introduction, of a mild TYLCV isolate into Venezuela, has also been documented (Zambrano et al. 2007).

The possibility of overlooking a second or third introduction of the virus into a country or area stems from the temptation to compare the newly obtained sequence to those of viruses previously sequenced from the same country or from nearby countries. However, infected plant material and disease vectors are inadvertently traded around the world and phylogenies of genes or genomes of individual species of plant viruses often reveal that viruses from distant geographical areas are closely related to one another. For example, outbreaks of iris yellow spot virus in onions from the American state of Georgia are very closely related to strains circulating in Peru (Nischwitz et al. 2007). Some pathogens move frequently and are repeatedly re-introduced to certain geographical areas. For instance, cassava mosaic begomoviruses have migrated from eastern Africa to western Africa at least twice (Ndunguru et al. 2005), while the maize-adapted maize streak virus A has frequently moved around Africa (Varsani et al. 2009). Although incomplete sampling of the diversity of plant viruses hampers definitive source tracking (Moury et al. 2006), attempting to find the origin of novel viral sequences can become a useful standard in the field. Placing novel viral sequences into their appropriate phylogeographical context can identify infection source countries, help trace back how the pathogens breached agricultural security measures and give the phytopathology community the most complete picture of each novel viral sequence.

## Materials and Methods

For each analysis, sequences were obtained from GenBank and aligned and trimmed manually with Se-Al v2.0a11 (<http://tree.bio.ed.ac.uk/software>). No statistically significant recombination breakpoints were detected by more than two of the following algorithms as implemented in RDP 3.15 (Martin et al. 2005): RDP, GENECONV, Bootscan, MaxChi, Chimaera, SiScan and 3Seq. Therefore, recombination was not considered in further phylogenetic analyses. Nucleotide substitution models were selected by Akaike's information criterion (AIC) with MODELTEST 3.7 (Posada and Crandall 1998). Maximum-likelihood phylogenetic analyses were performed with PAUP\* 4.0 beta (Swofford 2003) and bootstrapped with 1000 replicates. Trees were manipulated with FIGTREE v.1.2.3 (<http://tree.bio.ed.ac.uk/software/figtree/>), midpoint rooted for clarity and presented with branch lengths scaled to the numbers of substitutions/site.

## Results and Discussion

A survey of several journals that publish short-form reports on new plant diseases or established diseases in new plants or locations revealed that phylogenetic methods are rarely used when describing a novel viral sequence (Table 1). Only 3.6 per cent of the more than 200 viral reports published in three journals (BSPP's *New Disease Reports*, APS' *Plant Disease Reports* and *Plant Health Progress*) contained a phylogenetic tree that a reader could look at and evaluate. Far more popular was reporting the PNI to other strains of the virus.

Sometimes, the PNI was explicitly aided by NCBI's BLAST (Basic Local Alignment Search Tool), which was used to identify closely related isolates in GenBank for comparison. Often, however, PNI was calculated relative to isolates without any rationale for why the specific isolates were selected. A third category was needed for reports that communicated that they had created a phylogenetic tree, but did not provide the tree to the reader (though presumably they would, upon individual request). Most of these reports aimed solely to communicate that a virus had been found in a new plant or place, not to say anything about its biogeography, nor assert where the infection had come from. For mere pathogen identification, PNI is adequate.

**a. Percent nucleotide identity: good, and sometimes good enough**

If the only goal is determining what virus is present in a diseased plant, then obtaining a sequence with species-specific primers and confirming that it is very similar to known isolates of a virus is sufficient, and is more sensitive than serological techniques (Schneider et al. 2004). The vast majority of short-form plant virus reports use sequence data in this way. In fact, some reports do not mention exact PNI values because the authors felt it was sufficient to mention that the sequences were highly similar to one another.

**Table 1.** The number of short-form reports in three plant pathology journals that involved sequencing of plant viruses from 2007 to 2009 (2/07 – 1/10 for *New Disease Reports*).

	Short-form reports of plant viruses with		
	PNI <sup>a</sup>	An unseen phylogenetic tree <sup>b</sup>	A phylogenetic tree <sup>c</sup>
<i>New Disease Reports</i>	47	0	5
<i>Plant Disease</i>	154	7	2
<i>Plant Health Progress</i>	9	0	1

<sup>a</sup>Those reports that used percent nucleotide identity (PNI) for any reason, including viral species identification, but no phylogenetic methods.

<sup>b</sup>Those reports that mentioned a phylogenetic analysis that was not presented (e.g. “a maximum-parsimony analysis showed that these sequences group closely together” (Raj *et al.* 2008).

<sup>c</sup>Those reports that provided a phylogenetic tree. The format of *New Disease Reports* and *Plant Health Briefs*, which allow figures as part of the report instead of supplementary ‘e-Xtra’ information, might make these journals a more welcoming home for reports with phylogenetic trees.

Importantly, many virus families use a threshold PNI to determine if a novel viral sequence represents a new species (determined and revised by the International Committee for Taxonomy of Viruses (Fauquet et al. 2005)). For instance, begomoviruses are of the same species if they are more than 89 per cent identical over the full-length DNA-A segments from previously characterized species, while their sister group, the mastreviruses, use a cut-off of 75 per cent (Fauquet et al. 2008). The single-stranded RNA potyviruses use a threshold value of 85 per cent (Fauquet et al. 2005), but there is discussion of reducing this to 76 per cent (Adams et al. 2005). It is necessary to include PNI when characterizing a novel viral sequence from a family that has a threshold PNI in order to assess whether or not it represents a novel species.

#### **b. BLAST can be better**

PNI is a better measure than simply the presence or absence of a PCR band, since it confirms that what lies between those sequence-specific sites is the expected sequence, and does not ignore insertions and deletions. If the researchers do not wish to undertake a more complete phylogenetic analysis, using BLAST can be an intermediate step (Altschul et al. 1990). BLAST compares a query sequence to the entire GenBank non-redundant nucleotide sequence collection, looks for high identity matches, and selects sequences that closely match the entire query sequence. The matching sequences are ranked by expect scores (E-values) that correspond to the relative likelihood of the match



being identified by chance alone. If one uses BLAST to query a novel viral sequence, and it is a member of a viral species or genus that has many sequences in GenBank, the results will show the publicly available sequences of that group to which the novel sequence has the highest PNI. These sequences increasingly have their country and time of isolation in their GenBank files or in an accompanying publication. However, it is important to note that these details must be explicitly specified; the year of submission to GenBank and the country of the submitting scientists are not reliable indicators of when and where a virus was isolated. By using BLAST, one can find the most similar viruses for PNI comparisons without preconceived notions about sequences from particular geographical areas to which the novel sequence should be most closely related.

### **c. Some situations are perfect for phylogenetics**

If PNI is good, and BLAST is better, then the most thorough placement a novel viral sequence can initially receive is through phylogenetic analysis. Not every new report of a plant disease requires a phylogeny, but reporting the first incidence of a pathogen in a new location, without attempting to determine where it could have come from, shortchanges the scientific community. In order to increase plant biosecurity, each country or region needs to know how and from where pathogens previously entered the region (Rodoni 2009). If novel sequences come with biogeographical information in the initial report, it will speed the process of highlighting frequent sources of infection and consistently

leaky ports of entry in interstate and international commerce. These analyses could lead to increased vigilance when screening imports from a subset of countries that are consistent sources of phytopathogens. This is especially critical, as trade agreements have weakened the ability of many countries to routinely quarantine plant material (Jones 2009; Rodoni 2009).

PNI analyses make some of the same implicit assumptions as phylogenetic analyses: that the alignment of the sequences is correct and reflects homology. By choosing one or a few sequences to calculate PNI against, the author is making the *a priori* decision that these are the most informative sequences with which to compare the new sequence. By contrast, phylogenetic analyses that involve more sequences from a wider geographical range than is typically employed in PNI analyses allow unexpected relationships between sequences to emerge.

Recombination can obfuscate patterns of common descent because it means different sequences with different evolutionary histories are physically joined together in the genome. Recombination is a frequent occurrence in many plant viruses (Chare and Holmes 2006; Lefeuve et al. 2007), and software programs exist to detect recombination breakpoints (many are collated into RDP3, (Martin et al. 2010)) so that the evolutionary relationships of different portions of the genome can be analysed separately. As PNI makes fewer assumptions about

ancestry, it is less affected by the presence of recombination than are phylogenetic analyses. Alignments destined for phylogenies should be screened for recombination, and researchers must be very cautious about further analysis with the entire alignment. One approach is to break up the dataset into smaller alignments at recombination breakpoints and analyse them separately. Another is to aim not for a single, bifurcating phylogenetic tree, but a network (Huson and Bryant 2006). The blended evolutionary history of several plant viruses has been traced using split network methods in SPLITSTREES4 (Hu et al. 2007; Codoñer and Elena 2008; Martín et al. 2009). While networks can better reflect the true ancestry of many recombinant plant viruses, the analysis of migration and hypothesis testing is more complicated on a network than bifurcating trees (Huson and Bryant 2006).

BLAST analysis shares many assumptions with phylogenetic analyses, and has an understandable bias towards finding the most closely related sequence over the longest stretch of nucleotides. BLAST will score a longer but lower similarity match higher than a shorter, more similar match. As many viral sequences in GenBank are from relatively short species-specific PCR amplicons, this means that using BLAST on a longer viral sequence, such as the whole genome, will not necessarily return the highest PNI match over the highly sequenced, species-specific amplicon region. Rather, it will return the highest PNI match for the entire length of the query sequence. Alignments for phylogenetic analyses can be

trimmed such that all sequences have the same length and the algorithms compare across an even amount of information. There is an obvious disadvantage to eliminating part of the new viral sequence from analysis and consideration, but perhaps the solution is to present both a phylogenetic analysis based on a good alignment and the PNI from a BLAST analysis with the full sequence.

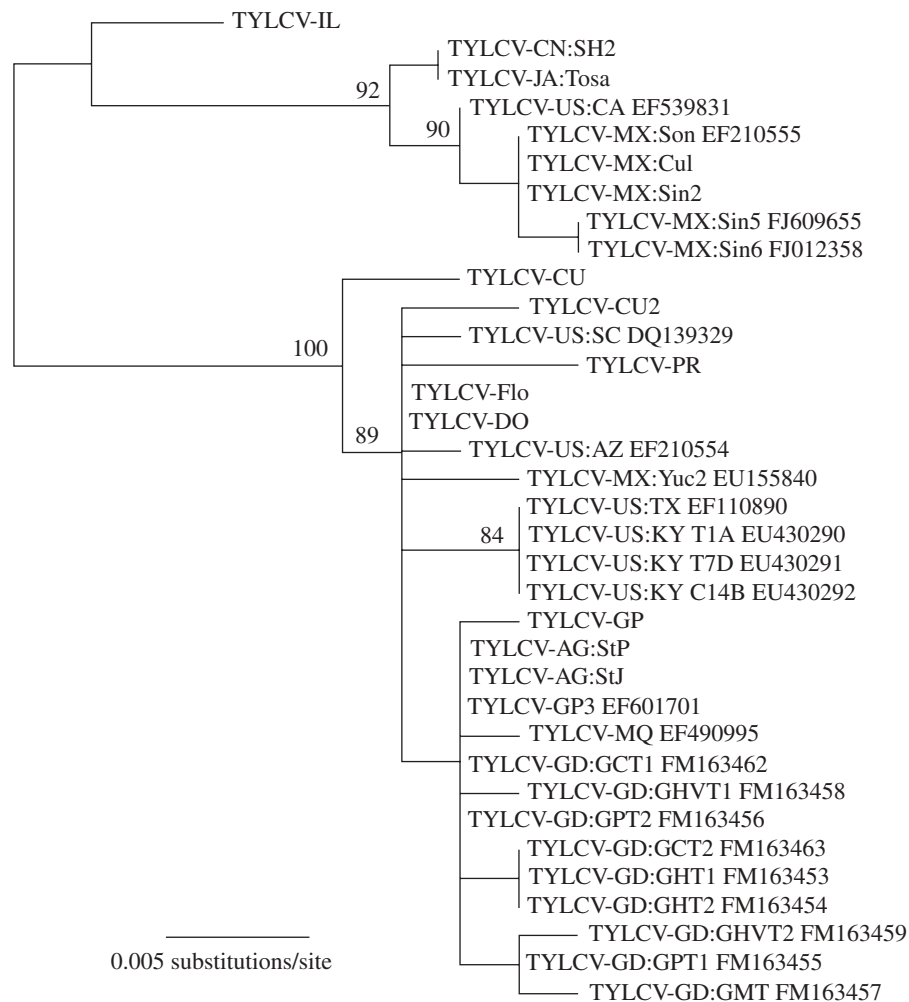
Phylogeographical approaches not only allow epidemiologists to trace the source of infections, they also can provide a measure of how confident researchers should be in their conclusions. Through estimates of support for clades, from bootstrap analyses or from Bayesian posterior probabilities, authors and readers alike can assess how probable it is that the alignment underlying the phylogenetic tree is representative of a real relationship between sequences that are closely grouped on the tree. This allows authors to move beyond suggestion and give a relative measure of the support for their assertions. For instance, when zucchini yellow mosaic virus (ZYMV) was first discovered in Poland, pairwise PNI comparisons indicated it was more closely related to sequences from Asia than to other European strains (Pospieszny et al. 2007). However, these authors were unable to give a measure of how confident they were in these relationships. The phylogenetic relationship between Polish (and now two French) ZYMV isolates and a Chinese isolate from 1999 was published in 2009, and revealed moderate 77 per cent bootstrap support, lending increased

credibility to the authors' conclusions (Lecoq et al. 2009). Several additional examples of the utility of phylogenetic analysis inspired by the recent plant virus report literature are given below.

#### **d. *Tomato yellow leaf curl virus***

As TYLCV has continued to spread in North America and the Caribbean, sequences have recently been added to GenBank from viruses isolated in Arizona (Idris et al. 2007), Guadeloupe, Grenada, Kentucky (de Sa et al. 2008), Martinique (Urbino and Dalmon 2007) and Mexico (Idris et al. 2007; Gamez-Jimenez et al. 2009). An updated tree created from an alignment of partial coat protein genes, which recapitulates the two geographically distinct New World clades (Duffy and Holmes 2007; Zhang et al. 2009), is given in Figure 1. The more recently isolated viruses were published with PNI to a range of closely related TYLCV isolates, but our phylogenetic analysis offers better resolution of the ancestry of some of these strains. The publication describing the partial genome sequences of TYLCV isolated in Kentucky reported them to be 98–99% identical to TYLCV-US:TX, TYLCV-US:AZ and TYLCV-US:SC (de Sa et al. 2008), but our analysis provides support for the Kentucky strains being more closely related to the Texan strain in particular. Similarly, the first Californian TYLCV isolate had a very high PNI to the strains from western Mexico (Rojas et al. 2007), but the tree in Figure 1 places a quantitative measure of the level of

support in the alignment for the grouping of TYLCV-US:CA with the Sonoran and Sinaloan strains.



**Figure 1.** Maximum-likelihood tree of a 488-base alignment of partial coat protein sequences of New World isolates of tomato yellow leaf curl viruses with three non-North American isolates, constructed with a Tamura-Nei (TrN) model of nucleotide substitution. Branches with greater than or equal to 75% bootstrap support are labeled. Taxon labels are as previously published (Duffy and Holmes 2007) or labeled according to convention (Fauquet et al. 2008) and given with its GenBank accession number.

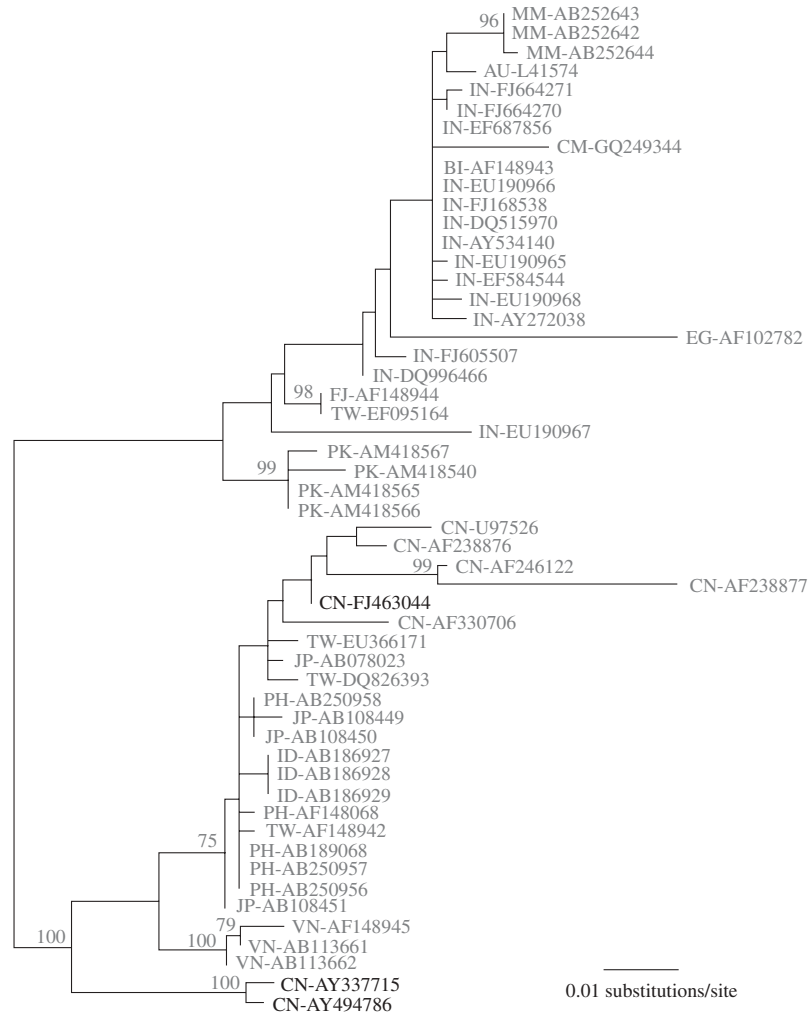
### **e. *Banana bunchy top virus***

The economically important banana bunchy top virus (BBTV) is a threat to banana crops in Asia, on Pacific Islands, in Australia and in the Middle East (Amin et al. 2008). The spread of BBTV into several novel geographical areas has been documented, especially in Hawaii, where the introduction of the virus to each island can be traced and dated using molecular phylogenetic methods (Almeida et al. 2009). That analysis revealed evidence of two introductions of BBTV onto Kauai island, despite 99.6 per cent or more nucleotide identity among the Kauaian sequences. Thus, the use of phylogenetic methods revealed continued exchange of infected plant material and/or infected banana aphids among the Hawaiian islands, which would have almost certainly been overlooked if the researchers only used PNI (Almeida et al. 2009).

We constructed a gene genealogy for all available coat protein genes of BBTV to see if we could detect any other cryptic second introductions of this virus. In figure 2, we highlight sequences from the Hainan province in China and show that two coat protein alleles are circulating within the region. Hainan is an island in the South China Sea and its first BBTV isolates grouped with those of nearby Vietnam (Jun and Zhi-Xin 2005). The more recently sequenced isolate from June 2008 is more closely related to viruses from the Chinese mainland (direct submission to GenBank, accession number FJ463044). The older Hainan BBTV isolates are 99.69 and 99.57 per cent identical to FJ463044—again very high



PNIs that make an incorrect, direct, ancestral relationship between the older and newer Hainan isolates seem plausible.



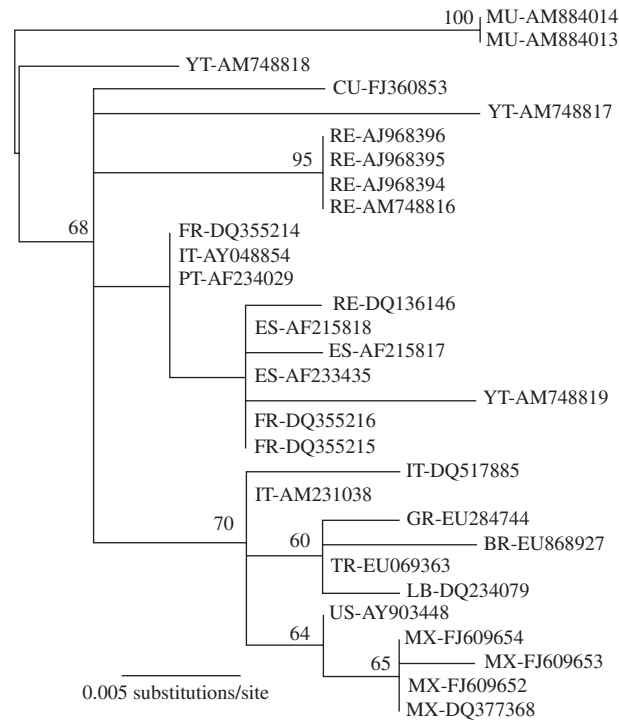
**Figure 2.** Maximum-likelihood tree of a 513-base alignment of coat protein sequences (the 513-base ORF on DNA-3) of *Banana bunchy top virus*, constructed with a TrN substitution model including the fraction of invariant sites (I) and substitution rate heterogeneity among sites (G). Branches with greater than or equal to 75% bootstrap support are labeled. Taxon labels are GenBank accession numbers preceded by two-letter country codes: AU, Australia; BI, Burundi; CM, Cameroon; CN, China; EG, Egypt; FJ, Fiji; ID, Indonesia; IN, India; JP, Japan; MM, Myanmar; PH, Philippines; PK, Pakistan; TW, Taiwan; VN, Vietnam. Isolates from Hainan, China are shown in black, other locations are shown in grey.

#### **f. *Tomato chlorosis virus***

In 2006 – 2007 tomato chlorosis virus (ToCV) was identified for the first time in South America, in Sumaré, Brazil (Barbosa et al. 2008). The researchers who sequenced portions of this Brazilian ToCV found that the strain was 99 per cent identical to ToCV from the USA. The phylogenetic analysis in figure 3 suggests, albeit with lower bootstrap support than that observed in our other analyses, that ToCV in Brazil is more closely related to ToCV from the Mediterranean (Greece, Turkey and Lebanon) than to strains from North and Central America. When the 463 base partial genome sequence was trimmed to make it align with other ToCV sequences in GenBank, the 309 base region of the Brazilian isolate was more than 99 per cent identical to Turkish and Lebanese isolates, and 98.7 per cent or less identical to the other New World isolates. In addition to the lower PNI in this region than what the Brazilian strain shares with the Mediterranean strains, the Brazilian strain does not have two common synapomorphies (mutations) shared among the other New World isolates. This analysis suggests that ToCV might have been introduced twice to the New World, but the relatively weak bootstrap values on the tree make any definitive statement inadvisable. As future isolates are collected and analysed, the hypothesis of multiple introductions can be more thoroughly examined.

These examples show that a phylogenetic approach can provide a geographical context to novel viral sequences, and either provide support for intuitive

relationships, or introduce the notion that viruses have migrated multiple times into a region.



**Figure 3.** Maximum-likelihood tree of a 309-base alignment of partial heat shock protein 70 sequences of tomato chlorosis virus, constructed with a transversional (TVM) model of nucleotide substitution. Branches with greater than or equal to 50% bootstrap support are labeled. Taxon labels are GenBank accession numbers preceded by two-letter country codes: BR, Brazil; CU, Cuba; ES, Spain; FR, France; GR, Greece; IT, Italy; LB, Lebanon; MX, Mexico; MU, Mauritius; PT, Portugal; RE, Reunion Island; TR, Turkey; YT, Mayotte.

**g. No additional experiments**

For researchers who are already using sequencing to identify and confirm the causative agent of a disease, no additional wet-lab work is needed to conduct a phylogenetic analysis. While there is a learning curve for using phylogenetic programs, many of the relevant programs are free or low-cost, and tutorials exist online and in book form. One approachable volume that now assists readers using MEGA (Tamura et al. 2007) is Barry Hall's third edition of *Phylogenetic Trees Made Easy* (Hall 2007). Beginning to use phylogenetic methods opens the door to more advanced hypothesis testing. One directly relevant application would be comparing the likelihood of two hypothetical evolutionary histories: one where a virus is allowed to have multiple introductions to a geographical region, and one where all isolates from the geographical region must be descended from a single introduction (e.g., Duffy and Holmes 2007).

In addition to noting multiple introductions of a virus and identifying weak points in plant biosecurity, it is important for disease management to know that multiple strains of a virus are circulating in the same region. Co-infection of the same plant by multiple strains of a virus can lead to more severe symptoms owing to synergistic action or the rare creation of a more virulent genotype, both illustrated in the cassava mosaic disease outbreak in Uganda in the late 1990s (Legg et al. 2006). With foreknowledge of multiple strains in an area, researchers could begin monitoring to see if recombinant strains emerge and are associated with more

severe symptoms. However, plant pathologists need to be aware of the potential problem before they put the time and resources into increased vigilance.

## CHAPTER 2

### CAUSE FOR UNREST:

#### UNRESTRICTED NUCLEOTIDE SUBSTITUTION MODEL MORE LIKELY THAN GENERAL TIME-REVERSIBLE IN VIRUS PHYLOGENETICS

##### Abstract

Nucleotide substitution models that are used in phylogenetic inference assume substitution rate reversibility, which allows for mathematically convenient calculations, but are not necessarily biologically realistic. This assumption may be particularly inaccurate for viruses, which have genomic architectures, and replication strategies that are prone to introducing substitution biases in their genomes. Alignments and phylogenies of full-length genes and whole genome sequences of 40 virus species were used to determine the strength of substitution bias across different genomic architectures (single- and double-stranded DNA, positive and negative single-stranded RNA, and double-stranded RNA viruses), and to compare the fit of general time-reversible (GTR) and the unrestricted non-time-reversible (UNREST) nucleotide substitution models. Corrected Akaike Information Criterion scores and the likelihood ratio test were applied to determine if UNREST or GTR was more likely to fit alignments to maximum likelihood phylogenies inferred with GTR+G. While single-stranded viruses exhibit more asymmetrical nucleotide substitutions, UNREST fit at least half of the data sets to the hypotheses better than GTR for both single-stranded



and double-stranded viruses. C→T substitutions were most often significantly over-represented ( $X^2$  test,  $p < 0.01$ ) while its reverse T→C substitutions were just as often under-represented; this bias is especially prominent in single-stranded DNA viruses. Analyses on a subset of the species indicate that while third codon position substitutions may be driving most of the overall bias, first and second codon positions also independently exhibit substitution bias. Our results suggest that inference of virus phylogenies may benefit from the extra parameters introduced by incorporating non-time-reversibility in nucleotide substitution models.

## **Introduction**

Substitution models used in phylogenetic inferences make assumptions of the evolutionary processes underlying the observed relationships. One of the earliest models of nucleotide substitution assumed equal probabilities of each base changing to another base as well as equal frequencies of all nucleotides (Jukes and Cantor 1969). Nucleotide substitution models that followed introduced more parameters to approximate empirical observations such as the higher probabilities of transitions occurring over transversions (Kimura 1980) and unequal base frequencies (Felsenstein 1981; Hasegawa et al. 1985). More detailed reviews of various nucleotide, codon-based, and amino acid substitution models, in addition to model corrections for rate heterogeneity over sites have been discussed elsewhere (Liò and Goldman 1998; Sullivan and Joyce 2005).

The general time-reversible (GTR) model, which allows different substitution rates for all nucleotides while still assuming reversibility of rates (Tavare 1986), is the most parameter-rich model commonly available for phylogenetic tree reconstructions, and is among the most frequently used (Sumner et al. 2012). Rate reversibility is a convenient mathematical assumption for Markov models (Felsenstein 1981; Yang 1994; Liò and Goldman 1998), but has been acknowledged to have little biological justification (Yang 1994; Liò and Goldman 1998). While nucleotide substitution models that incorporate non-reversibility have been evaluated (Takahata and Kimura 1981; Gojobori et al. 1982; Blaisdell 1985; Rzhetsky and Nei 1995), these studies are largely at least 20 years old (Boussau and Gouy 2006), and they have yet to be routinely employed in phylogenetic inference methods.

Further, nucleotide substitution models have also mostly been validated against cellular, double-stranded genomes (e.g., Kimura 1980; Felsenstein 1981; Hasegawa et al. 1985; Tamura and Nei 1993), in which accumulation of biased substitutions are constrained by the complementary strands. Lacking a stabilizing complementary strand however, single-stranded nucleic acids are more prone to strand-specific mutational biases than are double-stranded nucleic acids (Lindahl and Nyberg 1974; Frederico et al. 1990). Since viral genomic architecture can range from single- to double-stranded forms of both DNA and RNA, we

attempted to determine if substitution bias patterns were different between five types of viral genomes: single-stranded DNA (ssDNA), positive and negative single-stranded RNA (+ssRNA and -ssRNA), and double-stranded DNA and RNA (dsDNA and dsRNA).

We were also interested in evaluating whether incorporation of non-reversible rates in nucleotide substitution models could lead to significant improvement in fitting inferred viral phylogenies to sequence data. Enzymatic processes contributing to biased substitution patterns have been identified in retro-transcribing viruses (Zhang et al. 2003; Vartanian et al. 2010), RNA viruses (Fehrholz et al. 2012), and at least one dsDNA virus (Vartanian et al. 2008). Substitution bias has also been identified in ssDNA viruses (Duffy and Holmes 2008; Duffy and Holmes 2009). Since amino acid substitution models that take into account specific biological parameters have proven useful in phylogenetic inference of a retrovirus (Dimmic et al. 2002), and influenza virus (Dang et al. 2010), we explore the possibility that a non-time-reversible nucleotide substitution model that more accurately reflects substitution bias might fit viral sequence data better than a general time-reversible model.

## Materials and Methods

### *Viral sequence data collection and alignment*

Full-length protein-coding gene sequences or whole genome sequences of ssDNA, dsDNA, +ssRNA, -ssRNA, and dsRNA viruses were downloaded from GenBank (<http://www.ncbi.nlm.nih.gov/genbank/>) between 2010 and 2013.

Downloaded sequences were then viewed using Se-AL v2.0a11

(<http://tree.bio.ed.ac.uk/software/seal/>); alignments and further analyses for all datasets were done based on the coding strand sequence; sequences that were not already in that sense were reverse-complemented in Se-AL. Any patented sequences, modified microbial nucleic acids, or vaccine strains were removed. Sequence alignment was done manually, or with either Clustal W2 or Clustal Omega (Larkin et al. 2007; Sievers et al. 2011), both alignment tools hosted at the European Bioinformatics Institute (EMBL-EBI) website (Goujon et al. 2010). Software-aided alignments were followed by manual inspection and adjustment when necessary. Identical sequences were removed using RAxML v7.0.3 (Stamatakis 2006), which automatically creates a subset of unique sequences from the original dataset before any analysis.

Since recombination invalidates the assumption of shared evolutionary history between taxa in a bifurcating phylogenetic tree, the RDP3 package of recombination detection algorithms was used to identify sequences detected as

recombinant by three or more methods (Martin et al. 2010). All detected recombinant sequences were then removed from the alignments.

It has been reported that while the correct number of taxa or genes for phylogenetic analyses cannot be generalized, greater than 95% bootstrap support values are often obtained when taxon sampling was greater than 27 (Hedtke et al. 2006). Therefore for our analyses, we rounded up to a minimum of 30 taxa per dataset. While our datasets comprised 40 different virus species (eight species per genomic architecture), some species are represented by analyses of two genes, bringing our total number of datasets to 46.

Once sufficiently large, recombination-free alignments were obtained, they were each manually aligned to a sister taxon outgroup. The sister taxa were chosen based on the recommendation of the Ninth Report of the International Committee on Taxonomy of Viruses (King AMQ, Adams MJ, Carstens EB, Lefkowitz 2012). When such recommendations were not available in the report, sister taxa were selected based on published phylogenies (Supplementary Table 2).

Since transitions in the third codon “wobble” position are likely to be synonymous and not affect amino acid sequence, we also investigated the possibility that the overall observed substitution bias is driven by changes in that position.

Alignments of full-length genes, as well as alignments comprising only the first

and second codon positions, and only the third codon positions of the same genes, were created for two representative virus species per genomic architecture for comparative substitution bias analyses; columns containing gaps that altered the reading frames were removed from the alignments, while gaps that did not affect the reading frames were maintained.

### *Maximum Likelihood Phylogenetic Tree Inferences*

Maximum likelihood (ML) trees were constructed with both PAUP\* v4.0beta (Swofford 2003), and RAxML (Stamatakis 2006) (over the course of the analyses, different versions of RAxML were used for tree inferences: command line v7.0.3 and v7.6.3, and v8.1.11 accessed through the CIPRES portal (Miller et al. 2010)). Trees built in RAxML simply assumed the GTR+G (G=gamma distribution of among-site rate variation) model with average empirical nucleotide frequencies, and were used for testing the hypothesis of UNREST vs. GTR. Meanwhile, nucleotide substitution models for inferring ML trees with PAUP\* were selected by jModelTest v0.1.1 (Posada 2008) or ModelTest v3.7 (Posada and Crandall 1998) based on Akaike Information Criterion scores (models are listed in Supplementary Table 3); for all datasets, with the exception of JC polyomavirus, trees inferred by PAUP\* were then used to conduct substitution bias analyses, also using PAUP\*. The RAxML tree was used for the JC polyomavirus dataset because the PAUP\* tree search did not complete in over

12,000 hours of analysis. All of the models selected by jModelTest and ModelTest were derivatives of the GTR (Sumner et al. 2012).

### *Analyses of Substitution Biases*

Once an outgroup-rooted tree was obtained for each viral species of interest, PAUP\* was used to calculate the most parsimonious number of substitutions observed between the root of the tree to the tips (extant sequences used in the alignment), excluding all substitutions on the branch leading to the outgroup. PAUP\* was also used to infer an ancestral sequence and its nucleotide frequencies, to produce a frequency-adjusted matrix of expected substitutions. The  $\chi^2$  test was performed to determine if each of the observed nucleotide substitutions were significantly over- or under-represented compared to the expected substitutions. In addition, the size of the substitution bias in one direction relative to its opposite direction (substitution skew) was estimated based on a frequency-adjusted matrix of the observed substitutions.

Substitution skew was calculated based on the difference between frequency-adjusted number of observed substitutions from nucleotide  $i \rightarrow j$ , and frequency-adjusted number of observed substitutions in the opposite direction  $j \rightarrow i$ . This difference was scaled to a ratio between 1 and -1 by dividing by the sum of the numerator.

### *Hypothesis Testing*

The likelihood of either the reversible nucleotide substitution model (GTR, or REV), or the unrestricted, i.e., non-reversible model (UNREST) fitting the alignment data with the given ML tree was obtained using PAML v.4.4 (Yang 2007). Hierarchical likelihood ratio test (hLRT) was performed to obtain significance values based on a  $\chi^2$  distribution. Corrected Akaike Information Criterion (AICc) and Bayesian Information Criterion (BIC) scores were also calculated with the number of parameters of each substitution model,  $K = 8$  (GTR) or  $K = 11$  (UNREST) (Hurvich and Tsai 1989; Yang 1994; Burnham and Anderson 2004). The differences between scores of the two substitution models for each species dataset were then obtained by subtracting the AICc or BIC scores of UNREST from those of GTR (Burnham and Anderson 2004; Posada and Buckley 2004).

Since average empirical nucleotide frequencies were used by RAxML to infer the ML trees, the nucleotide frequency parameters for testing the likelihood of GTR in PAML was also set to the average observed frequencies. A similar comparison between GTR with ML-estimated nucleotide frequencies and UNREST was also performed for the same datasets.



## Results

### *Substitution Biases in Viruses*

Comparisons between observed and expected numbers of parsimonious nucleotide changes along maximum likelihood (ML) phylogenies revealed significant bias for specific substitutions in all viral genomic architectures (Table 1). Substitution biases were more often observed in transitions than in transversions. C→T and G→A transitions were most often significantly over-represented ( $\chi^2$  test,  $p < 0.01$ ), with the former being biased in almost all ssDNA viruses. Of the 49 transitions that were significantly over-, or under-represented across all datasets (Table 1, first four substitution columns), there were 42 transitions where the substitution from one nucleotide to another was significantly over-represented (e.g., C→T), while its reverse substitution (e.g., T→C) was significantly under-represented. In other words, there were 21 instances of asymmetrically biased forward and reverse substitution pairs. The remaining seven transitions had reverse substitutions that were of more marginal statistical significance ( $0.01 < p\text{-value} < 0.06$ ). Fifteen of the 21 asymmetric transition pairs were in the single-stranded viruses. Similarly, 22 of the 25 biased transversions had significantly asymmetric forward and reverse substitutions, i.e. 11 transversion pairs were asymmetrically biased, again mostly in the single-stranded viruses (8 out of 10 pairs).












The relative skew of the observed substitutions in one direction with respect to substitutions in the opposite direction were plotted for all datasets (Figure 2).

Points at the boundaries of 1 and -1 result from zero observed substitutions in either one of the directions, and are ignored. Across all genomic architectures, the most highly skewed substitutions are  $A \rightarrow C$ ,  $A \rightarrow T$ , and  $A \rightarrow G$ .

As the majority of the substitutions in our 40 species datasets were in the third codon position (wobble position) in protein-coding genes, we investigated a subset of the data (2 datasets/genomic architecture) to see if this position dominated the substitution biases observed. Each of these datasets were split into two: one with all the aligned first and second codon positions, and another with solely the aligned third codon positions. Out of the 10 representative datasets (Supplementary Table 1), the substitution biases in four of the complete coding gene sequence datasets matched to subsets consisting of only third codon positions, and not to the subsets comprising first and second codon positions (+ssRNA: HAV, both -ssRNA, and dsRNA: RBSDV). On the other hand, two of the 10 datasets (ssDNA: phiX174, and dsRNA: RotA.G9) also had matching, or partially matching biases in the first and second position subsets in addition to matching third codon position biases; a third dataset (+ssRNA: ACLSV) had different substitution biases in the complete, and first and second codon position subsets, but no detected bias in the third codon position subset. The remaining three datasets either had no detectable substitution biases in the full, first and second codon position, and third codon position analyses (both dsDNA datasets), or none detected in either codon position subsets (ssDNA: EACMV).

**Table 1: Substitution bias analyses.** Black boxes indicate that the observed nucleotide substitution (denoted in the top row) is significantly over-represented relative to the expected amount, based on a  $\chi^2$  test ( $p < 0.01$ ); grey boxes indicate significant under-representation. The species abbreviations represent: B19: Human erythrovirus B19; BBTV: Banana bunchy top virus; BFDV: Beak feather disease virus; EACMV: East African cassava mosaic virus; MSV: Maize streak virus; PCV2: Porcine circovirus 2; phiX174: Enterobacteriophage phiX174; WDV: Wheat dwarf virus; ACLSV: Apple chlorotic leaf spot virus; BNYVV: Beet necrotic yellow vein virus; CuMV: Cucumber mosaic virus; GLRaV3: Grapevine leafroll-associated virus 3; HAV: Hepatitis A virus; JEV: Japanese encephalitis virus; PLRV: Potato leafroll virus; TStV: Tobacco streak virus; AKAV: Akabane virus (nucleoprotein); BDV: Borna disease virus; CDV: Canine distemper virus; GBNV: Groundnut bud necrosis virus; HTNV: Hantaan virus; IHN: Infectious hematopoietic necrosis virus; RSV: Rice stripe virus; VHSV: Viral hemorrhagic septicemia virus; BKPyV: BK polyomavirus; GaHV1: Gallid herpesvirus 1; HAdB: Human adenovirus B; HPV: Human papillomavirus; HSV1: Herpes simplex virus 1; JCPyV: JC polyomavirus; VACV: Vaccinia virus; AHSV: African horsesickness virus; ARV: Avian orthoreovirus; EHDV2: Epizootic hemorrhagic disease virus 2; IBDV: Infectious bursal disease virus; IPNV: Infectious pancreatic necrosis virus; RBSDV: Rice black streaked dwarf virus; RotA.G9: Rotavirus A subtype G9; RotC: Rotavirus C. Whole genome sequences were analyzed unless otherwise noted by the protein or segment names in parentheses.

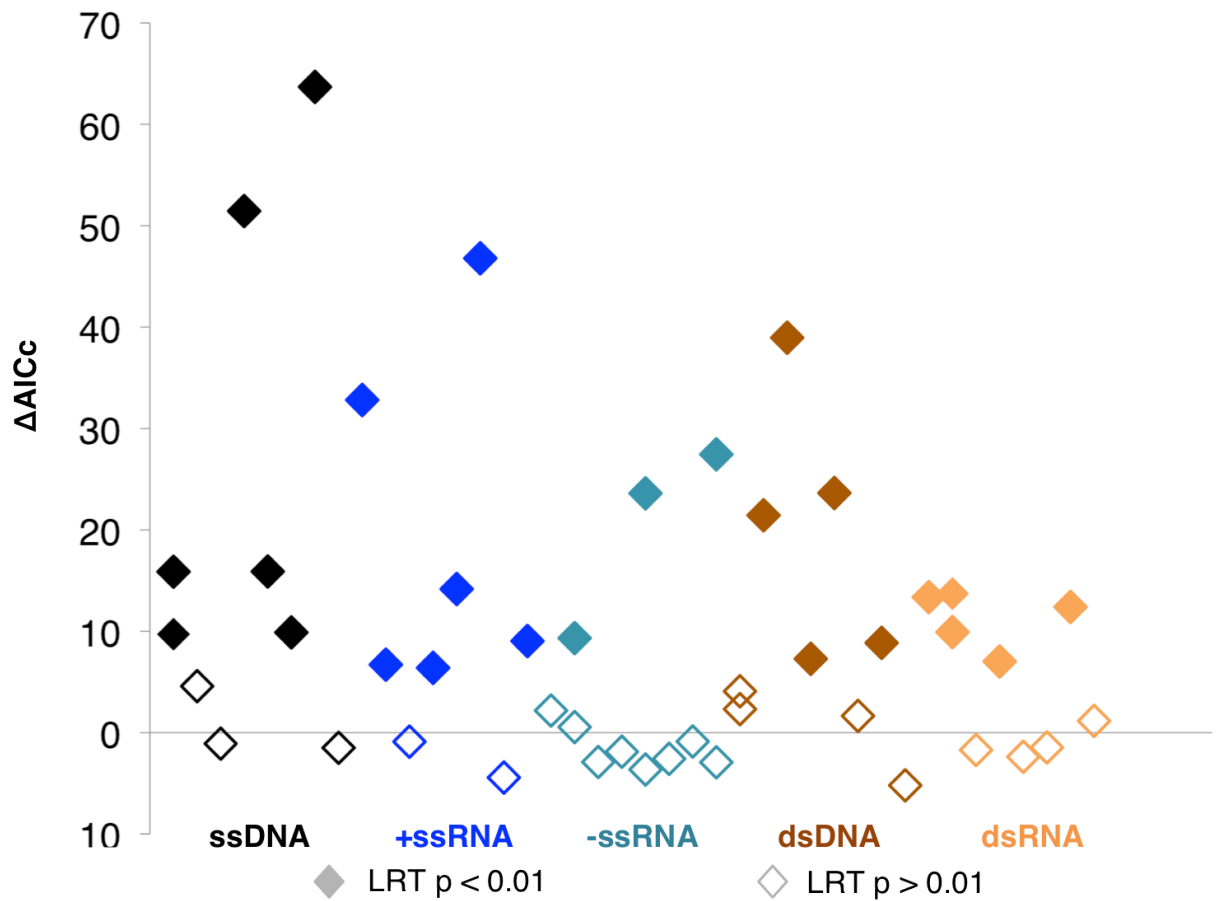


<b>dsDNA</b>				
BKPyV	234	5665		
BKPyV (VP1)	167	1095		
GaHV1 (UL23)	35	1109		
HAdB (L3)	30	2851		
HPV6 (L1)	38	1515		
HPV16	70	8139		
HSV1 (UL23)	417	1139		
JCPyV	434	5501		
VACV (B5R)	39	954		
<b>dsRNA</b>				
AHSV (VP7)	49	1062		
ARV ( $\sigma$ NS)	35	1113		
ARV ( $\sigma$ C)	50	984		
EHDV2 (VP7)	37	1050		
IBDV (RdRP)	94	2678		
IPNV (poly)	33	2920		
RBSDV (CP)	82	1677		
RotA.G9 (VP7)	163	1023		
RotC (VP7)	58	999		

*Hypothesis testing: GTR vs. UNREST*

Based on  $\Delta\text{AICc}$  and hLRT scores, UNREST performed significantly better than GTR at fitting the alignment data to the inferred ML phylogeny for more than half the datasets ( $p < 0.01$ , 25 out of 46 datasets) (Figure 1). The significant improvement of UNREST over GTR was seen in similar proportions of single-stranded virus datasets (15 out of 28 total, 54%) as in the double-stranded viruses (10 out of 18 total, 56%). When UNREST was compared to GTR with ML-estimated nucleotide frequency parameters,  $\Delta\text{AICc}$  scores still favored UNREST in more than half the datasets, although the number of significant results by hLRT was lower (Supplementary Figure 1).

Bayesian Information Criterion (BIC) scores were also calculated (Supplementary Figure 2), and UNREST was still significantly better than GTR for two ssDNA datasets. BIC selects simpler models than AIC does when  $n > 8$  (where  $n$  is the total number of characters per alignment) (Posada and Buckley 2004), and  $n$  in these datasets is  $\geq 1000$ .



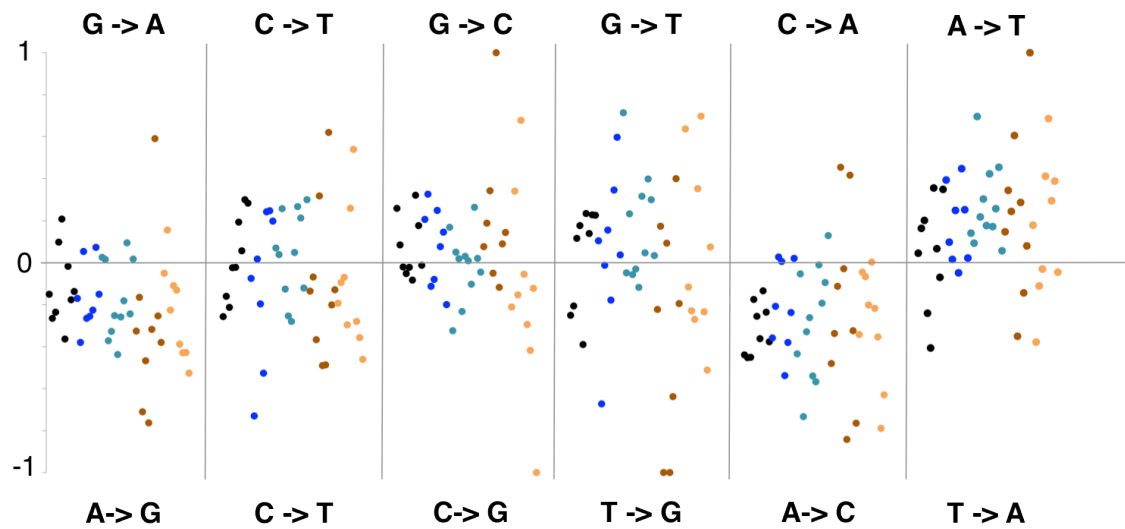
**Figure 1: Difference in AICc scores between UNREST and GTR substitution**

**models for 40 virus species.**  $\Delta AICc = AICc_{GTR} - AICc_{UNREST}$ , therefore  $\Delta AICc > 0$

indicates UNREST being more likely than GTR at fitting the data; filled diamonds show that the difference is significant by the likelihood ratio test ( $p < 0.01$ ). Diamonds that are vertically stacked in the same column represent different data points from the same virus species (e.g., alignments of different gene sequences from one virus species). The 40 virus species, from left to right, are: Human erythrovirus B19 (NS1, top, VP, bottom); Banana bunchy top virus (DNA1 segment); Beak feather disease virus; East African cassava mosaic virus (DNA-A segment); Maize streak virus; Porcine circovirus 2; Enterobacteriophage phiX174; Wheat dwarf virus; [Apple chlorotic leaf spot virus \(CP\)](#); [Beet necrotic yellow vein virus \(CP\)](#); [Cucumber mosaic virus \(CP\)](#); Grapevine leafroll-

associated virus 3 (CP); Hepatitis A virus (polyprotein); Japanese encephalitis virus;  
 Potato leafroll virus (CP); Tobacco streak virus (CP); Akabane virus (nucleoprotein);  
 Borna disease virus (P, top, N bottom); Canine distemper virus (H); Groundnut bud  
 necrosis virus (N); Hantaan virus (Gc, top, N, bottom); Infectious hematopoietic necrosis  
 virus (G); Rice stripe virus (CP); Viral hemorrhagic septicemia virus (G, top, N, bottom);  
 BK polyomavirus (whole genome, top, VP1, bottom); Gallid herpesvirus 1 (UL23);  
 Human adenovirus B (L3); Human papillomavirus 6 (L1); Human papillomavirus 16;  
 Herpes simplex virus 1 (UL23); JC polyomavirus; Vaccinia virus (B5R); African  
 horsesickness virus (VP7); Avian orthoreovirus ( $\sigma$ C, top,  $\sigma$ NS, bottom); Epizootic  
 hemorrhagic disease virus 2 (VP7); Infectious bursal disease virus (RdRP); Infectious  
 pancreatic necrosis virus (polyprotein); Rice black streaked dwarf virus (CP); Rotavirus A  
 subtype G9 (VP7); Rotavirus C (VP7).





**Figure 2: Relative skew of observed substitutions.** Each point represents one dataset and the point colors correspond to the previously used scheme; black: ssDNA, blue: +ssRNA, teal: -ssRNA, brown: dsDNA, and orange: dsRNA viruses.

## Discussion

In situations when nucleotide strands exist in single-stranded form, asymmetrical substitution patterns have occurred and in many cases led to subsequent strand bias in the genomes of, amongst others, metazoan animals (Van Den Bussche et al. 1998; Green et al. 2003; Chen et al. 2011), mitochondria (Tanaka and Ozawa 1994; Reyes et al. 1998), archaea (Lopez and Philippe 2001), bacteria (Lobry 1996), and viruses (Berkhout and van Hemert 1994; Mrázek and Karlin 1998). Bias can be introduced during transcription when the non-transcribed strand is temporarily single-stranded and vulnerable to DNA damage via deamination of cytosine to uracil, or 5'-methylcytosine to thymine (Francino and Ochman 1997). Cytosine deamination does occur at a higher rate in single-stranded DNA than in double-stranded DNA (Lindahl and Nyberg 1974; Frederico et al. 1990), and significantly increased cytosine to thymine transitions have been observed in the non-transcribed strand of bacteria (Beletskii et al. 2000; Lind and Andersson 2008). Single-strandedness during chromosomal replication has also been associated with the generation of strand asymmetry, when the unwound template lagging strand is exposed to higher possibility of damage or primer-template misalignment (Francino and Ochman 1997; Frank and Lobry 1999). Similarly, during mammalian mitochondrial replication the parental H strand can be single-stranded for an extended period of time until the synthesis of a new L strand (Reyes et al. 1998; Frank and Lobry 1999).

As for viruses, specific deamination enzymes that target single-stranded nucleotide strands also affect viral genomic nucleotide composition. The APOBEC cytosine deaminases include members such as APO3G that deaminate cytosines to uracil in viruses that depend on single-stranded DNA intermediates as part of their replication cycle, e.g., in HIV-1 (Senavirathne et al. 2012), as well as in Hepatitis B virus, consequently causing G→A hypermutations in the virus genomes (Zhang et al. 2003; Vartanian et al. 2010). APOBEC3 proteins also appear to hypermutate the genomes of viruses that do not replicate via reverse transcription such as the -ssRNA measles virus (Fehrholz et al. 2012), the ssDNA TT virus (Tsuge et al. 2010), and the dsDNA human papillomavirus (Vartanian et al. 2008). Meanwhile, evidence of biased substitution patterns have also been noted in the genomes of ssDNA East African cassava mosaic virus, Tomato yellow leaf curl virus, Tomato yellow leaf curl China virus, and Maize streak virus, although the causes are posited to be non-enzymatic oxidative deamination (Ge et al. 2007; Duffy and Holmes 2008; van der Walt et al. 2008; Duffy and Holmes 2009).

#### *More biased substitution patterns in single-stranded viruses*

Here we show that substitution biases were observed in at least four datasets per genomic architecture, although the double-stranded viruses tended towards fewer instances of bias. As with other organisms with double-stranded genomes, dsDNA and dsRNA viruses appear to have some of the same physical

constraints preventing excessive mutations. The substitution rates of dsDNA viruses, which approach  $10^{-9}$  substitutions/site/year, are orders of magnitude lower than the substitution rates of ssDNA and ssRNA viruses, which range from approximately  $10^{-2}$  to  $10^{-5}$  substitutions/site/year (Duffy et al. 2008). Viruses with dsRNA genomes also have substitution rates that are lower than, or comparable to those of single-stranded viruses (Carpi et al. 2010). Although higher substitution rates may introduce or increase the possibility of bias, it is not the only explanation for it. Compositional bias in viral genomes may also result from transcriptional mechanisms; the RNA polymerase of dsDNA bacteriophage T7 appears to promote cytosine to thymine mutations in the non-transcribed strand that is temporarily single-stranded, skewing the cytosine to thymine ratio in the T7 genome (Beletskii et al. 2000). Host adaptation too, may be a contributing factor in generating mutational bias, as seen in the -ssRNA genomes of influenza A viruses that evolve from avian to human hosts (Rabadan et al. 2006).

Biased genomic composition naturally leads to the consideration of the effect on codon usage. Genomic nucleotide content is suspected to exert a strong influence on the codon usage biases of both large and small DNA viruses that infect vertebrates (Shackelton et al. 2006), as well as human RNA viruses (Jenkins and Holmes 2003). The extent to which either mutational pressure or selection generates viral codon usage bias still appears uncertain. It has previously been shown that ssDNA bacteriophages exhibit a bias for using

synonymous codons that end with thymine, even when such a codon usage profile does not match that of their bacterial hosts, thus indicating a major role for mutational pressure in ssDNA viral genome evolution; dsDNA bacteriophages on the other hand, do not demonstrate the same bias (Cardinale and Duffy 2011).

While this may support the use of codon-based substitution models, it does not negate a role for a non-time-reversible nucleotide substitution model.

Overlapping reading frames, ambisense genomes, and non-coding regions in viruses are some aspects that may not be appropriately described with codon-based substitution models.

The most frequently observed biased transitions in this analysis are G→A and C→T substitutions. There is no pattern of any particular transition being preferred in any of the genomic architectures examined, except in ssDNA viruses, where C→T is consistently significantly over-represented (Table 1). This bias corresponds with earlier observations in ssDNA plant viruses (Duffy and Holmes 2008; Duffy and Holmes 2009) and is thought to be due to non-enzymatic mechanisms not associated with replication since ssDNA viruses do not possess their own polymerases, replicating instead with their hosts' processive polymerases. The error rates for bacterial and eukaryotic polymerases are estimated at  $10^{-6}$  mutations/site/round of replication (Roberts and Kunkel 1988; Schaaper 1993), whereas the mutation rate of ssDNA viruses are higher by approximately an order of magnitude (Duffy et al. 2008). ssDNA bacteriophage

packaged in viral capsids also appear to exist in its single-stranded state, rather than base-paired in secondary structures, inviting the probability of spontaneous oxidative reactions (Benevides et al. 1991); however, recent computational predictions based on minimum free energy calculations suggest that eukaryotic ssDNA viruses may possess more conserved genomic secondary structures than initially suspected (Muhire et al. 2014).

#### *Non-time reversible substitution model in virus phylogenetics*

The unrestricted (non-time-reversible) nucleotide substitution model UNREST was initially introduced by Yang, as a special case of an earlier model that applied different rate matrices to different branches of a tree (1994). UNREST was estimated using six primate  $\psi\eta$ -globin pseudogenes and part of the mitochondrial DNA genome from nine primates, but its usefulness was dismissed as it offered marginal improvement over the reversible GTR model on these limited, cellular datasets (Yang 1994). Nevertheless, as far as we are aware, UNREST has never been validated against viral sequences, and since asymmetrical substitution patterns are evident in these viral datasets, the additional parameters may well be justified. Since UNREST allows forward and reverse substitution rates to vary, we expected it to better fit the ML tree to the data more often for single-stranded viruses with significant substitution biases; UNREST surprisingly however, also fit significantly better than GTR for just over half of the double-stranded virus datasets.

We are currently unable to apply UNREST to phylogenetic reconstructions of viral datasets because current phylogenetic programs do not adequately implement this model. PAML was not intended for tree-searching (Yang 2007).

Implementing a non-time-reversible substitution model also entails searching through the exponentially larger rooted tree space, with a model that performs poorly at identifying root position (Yang 1994; Huelsenbeck et al. 2002); nevertheless, Huelsenbeck et al. also note that the non-time reversible model is able to identify the tree root in the presence of highly nonreversible substitution processes (2002). The nhPhyML program, which implements another non-reversible evolutionary model (Galtier and Gouy 1998; Boussau and Gouy 2006), requires a starting tree that can only be first inferred using another (reversible) model, while HyPhy offers a batch file method that uses the non-reversible model to maximize likelihood estimates of branch lengths on a user-inputted tree (Kosakovsky Pond et al. 2005; Harkins et al. 2009). However, it is hoped that the updated versions of phylogenetic reconstruction programs will allow for this most flexible of nucleotide substitution models (e.g., RevBayes). Such programs are needed to assure the most accurate reconstruction of viral phylogenies, including for molecular epidemiology during disease outbreaks. As ssDNA viral phylogenies were particularly better fit by UNREST models (by AICc, BIC and hLRT), it could be most important to use this more complicated model to describe the evolution of viruses with ssDNA genomic architectures.

In trying to capture biological reality, the introduction of too many parameters is undesirable, as it reduces the power to discriminate between trees (Liò and Goldman 1998); however, the accuracy of nucleotide substitution models currently used in viral phylogenetic inference has been found to be limited by decaying transition/transversion ratios due to mutational saturation, especially in rapidly-evolving viruses (Duchene et al. 2015). New phylogenetic methods that were developed specifically for inference of viral evolutionary relationships have tended to outperform other types of models (Dimmic et al. 2002; Dang et al. 2010), as has the experimentally-informed substitution model for influenza virus that does not rely on extant sequences of interest – this latter parameter-free model had likelihoods that surpassed other highly-parameterized codon substitution models (Bloom 2014).



## CHAPTER 3

### NO CLEAN SWEEP FOR THE SWEEPOVIRUSES: PROBLEMS WITH PAIRWISE PERCENT NUCLEOTIDE IDENTITY AS A DEFINITIVE CLASSIFICATION TOOL

#### Abstract

Viral classification relies on genomic sequence similarity in addition to other criteria such as genome architecture, serology, and host or vector range. Since viruses lack universally conserved sequences and evolve at different rates, the specific criteria delineating different virus species vary between taxonomic groups. Members of the genus *Begomovirus* (Family: *Geminiviridae*) comprise single-stranded DNA (ssDNA) viral crop pathogens transmitted by the whitefly *Bemisia tabaci*, and infect a large variety of dicotyledonous plants. Based on whole genome sequence comparisons, a *Begomovirus* isolate is considered a novel species if it has <91% pairwise nucleotide identity to all other previously identified begomoviruses. Here we use a subclade of *Begomovirus* that exclusively infects plants of the family *Convolvulaceae* (sweepoviruses: sweet potato-infecting begomoviruses) to explore how uncorrected pairwise nucleotide distance can negatively impact virus species demarcation. We demonstrate two issues by following the currently proposed recommendations: first, the newly revised number of sweepovirus species needs to be further adjusted. Secondly, the identification of a given species based on pairwise identities to a single isolate lead to species groups with highly divergent members and suggests that

the current criteria may be dependent on the order of species discovery. The issues identified here regarding species demarcation based on pairwise identities potentially invalidate the biological significance of classification by species within some viral groups. We determine that sweepovirus species are not monophyletic, and suggest a consensus sequence-based identification of species.

## Introduction

The genus *Begomovirus* (Family: *Geminiviridae*) consists of whitefly-vectored, single-stranded DNA (ssDNA) plant viruses that infect various crops all over the world (King AMQ, Adams MJ, Carstens EB, Lefkowitz 2012), often causing substantial economic losses (Patil and Fauquet 2009; Bernardo et al. 2013). Within the genus, begomoviruses fall into three distinct phylogenetic groupings: begomoviruses that infect a wide variety of dicotyledonous plants, those that primarily infect leguminous plants, and those that primarily infect convolvulaceous plants (Prasanna et al. 2010; King AMQ, Adams MJ, Carstens EB, Lefkowitz 2012). The virus genomes are either monopartite (a single segment homologous to DNA-A) or bipartite (DNA-A and DNA-B), with each segment approximately 2.5-2.6 kb in size (King AMQ, Adams MJ, Carstens EB, Lefkowitz 2012). Begomoviruses found in the New World (NW) are usually bipartite, while most begomoviruses found in the Old World (OW) are monopartite; however, examples of both monopartite NW and bipartite OW begomoviruses have been reported (Melgarejo et al. 2013). *Begomovirus* is the most species-rich viral genus, with species differentiated based on genomic

content (<91 percent DNA-A nucleotide identity (Brown et al. 2015), presence/absence of DNA-B, presence/absence of the AV2 open reading frame (which codes for a pre-coat protein (Ho et al. 2014)), <90% coat protein amino acid sequence identity, and other characteristics, such as natural host range and symptoms, viability of pseudorecombinants, and ability of the replication-associated (Rep) protein to replicate genomic components *in trans* (King AMQ, Adams MJ, Carstens EB, Lefkowitz 2012). With the increasing ease and affordability of sequencing technologies, however, many novel virus species are being recovered through metagenomic surveys, and identified primarily based on genomic similarity (Brown et al. 2015).

Recently, several new programs for calculating pairwise nucleotide percent identities have been introduced (Muhire et al. 2013; Bao et al. 2014; Simmonds 2015), and recommended for use in species demarcation (Muhire et al. 2013; Brown et al. 2015). One of these programs, the Species Demarcation Tool (SDT) was developed and implemented to improve and standardize the classification criteria of ssDNA viral groups (Muhire et al. 2013). In this study, we used SDT to determine the identities of monopartite sweet potato (*Ipomoea batatas*)-infecting begomoviruses (sweepoviruses) identified through vector-enabled metagenomics (VEM) of the whitefly vector (Rosario et al. 2015). Despite the publication of a taxonomic revision of all begomoviruses in 2015 implementing the new <91% nucleotide identity species threshold (Brown et al. 2015), the present analysis of all publicly available sweepovirus sequences suggests further taxonomic

changes that need to be adopted to comply with this threshold. We also explore the pitfalls of relying on pairwise percent nucleotide identity for assignment of novel virus species, and the “lumping” effect resulting from even the seemingly stringent >91% sequence identity threshold for species demarcation.

## **Materials and Methods**

### ***Sweepovirus sequence collection***

Novel sweepovirus sequences were recovered through metagenomic sampling of whiteflies collected in Spain and Puerto Rico. Sample collection and processing methods were reported elsewhere (Rosario et al. 2015). All other available genome-length sweepovirus DNA-A sequences as identified through TaxBrowser were downloaded from the GenBank nucleotide database in August 2014 (Acland et al. 2014).

In total, 168 sweepovirus whole genome sequences were analyzed (Supplementary Table 1), comprising isolates of *Sweet potato leaf curl virus* (SPLCV), *Sweet potato leaf curl Canary virus* (SPLCCaV), *Sweet potato leaf curl China virus* (SPLCCNV), *Sweet potato leaf curl Georgia virus* (SPLCGoV), *Sweet potato leaf curl Sao Paulo virus* (SPLCSPV), *Sweet potato leaf curl South Carolina virus* (SPLCSCV), *Sweet potato leaf curl Uganda virus* (SPLCUV), and *Sweet potato mosaic virus* (SPMV). *Merremia leaf curl virus* (MerLCV) *Sweet potato leaf curl Henan virus* (SPLCHnV), and *Sweet potato leaf curl Sichuan virus 1* and *2* (SPLCSiV-1, SPLCSiV-2), which are recognized as begomovirus

species as of January 2015 (Brown et al. 2015), are also included in this analysis. Isolates of six other sweepoviruses formerly known as Ipomoea yellow vein virus (IYVV), Sweet potato golden vein-associated virus (SPGVV), Sweet potato leaf curl Bengal virus (SPLCBeV), Sweet potato leaf curl Japan virus (SPLCJV), Sweet potato leaf curl Lanzarote virus (SPLCLaV) and Sweet potato leaf curl Spain virus (SPLCESV) have been reclassified as SPLCV (Brown et al. 2015); these names are retained in the following analyses only to identify isolates that may be divergent from the SPLCV cluster, and thus are not italicized. Furthermore, Sweet potato leaf curl Guangxi virus (SPLCGxV), and Sweet potato leaf curl Shanghai virus (SPLCShV) are proposed species that have not yet been recognized by the International Committee on Taxonomy of Viruses (ICTV).

### ***Pairwise percent nucleotide identity***

The initially unaligned sweepovirus sequences were inputted to SDT v1.0 (for Mac OSX) and percent nucleotide identity was calculated using a MUSCLE pairwise-alignment (Muhire et al. 2013). SDT determines sequence identity by calculating the proportion of matching nucleotides out of the total non-gapped columns of the pairwise alignment (Muhire et al. 2013). SDT (v1.2 for Windows) was used to generate a color-coded pairwise identity matrix, and species and strain cutoff values of <91% and <94% respectively were applied.

In addition to the visual identity matrix, we used the exact pairwise identities to further examine isolates within the SPLCV species, and between SPLCV and

other sweepovirus species. Isolates included in the SPLCV species definition comprise SPLCV, and other isolates reclassified as SPLCV (IYVV, SPGVV, SPLCBeV, SPLCJV, SPLCLaV, SPLCESV); three isolates misclassified as SPLCV (Results, I.c.) were omitted. Unrecognized species SPLCGxV and SPLCShV were not included in either SPLCV or non-SPLCV groups. The percent nucleotide identities were also used to visualize within-species clustering around SPLCV and SPLCESV reference sequences selected based on ICTV recommendations (King AMQ, Adams MJ, Carstens EB, Lefkowitz 2012).

### ***Phylogenetic inference***

The sweepovirus complete genome sequences were also aligned with Clustal Omega (Sievers et al. 2011), and manually adjusted in Se-Al v2.0a11 (<http://tree.bio.ed.ac.uk/software/seal/>) for phylogenetic inference. A maximum likelihood tree was constructed based on the resulting alignment with RAxML v7.6.3 (Stamatakis 2006) under the GTRGAMMA model, with 1000 bootstrap replicates. The same methods were also used to infer a tree of the sweepovirus genomes after removal of the long intergenic region.

An amino acid alignment of the sweepovirus coat proteins (CP) was also created in Se-Al. Identical sequences were removed from the final alignment (n=113), and a maximum likelihood tree inferred by RAxML v8.1.11 on the CIPRES Science Gateway (Miller et al. 2010), implementing the amino acid model JTT

selected by ProtTest v3.3 (Abascal et al. 2005). All alignments are available from Dryad (accession numbers pending).

Recombination was identified using RDP 3.44a (Martin et al. 2010), which offers an array of different detection algorithms; the specific programs used here were RDP (Martin and Rybicki 2000), GENECONV (Sawyer 1989; Padidam et al. 1999), Chimaera (Posada and Crandall 2001), MaxChi (Smith 1992), BootScan (Salminen MO, Carr JK, Burke DS 1995), SiScan (Gibbs et al. 2000), and 3SEQ (Boni et al. 2007). Events detected by more than two of the methods with average p-values <0.001 were accepted as significant.

## **Results**

### **I. Application of pairwise percent nucleotide identity**

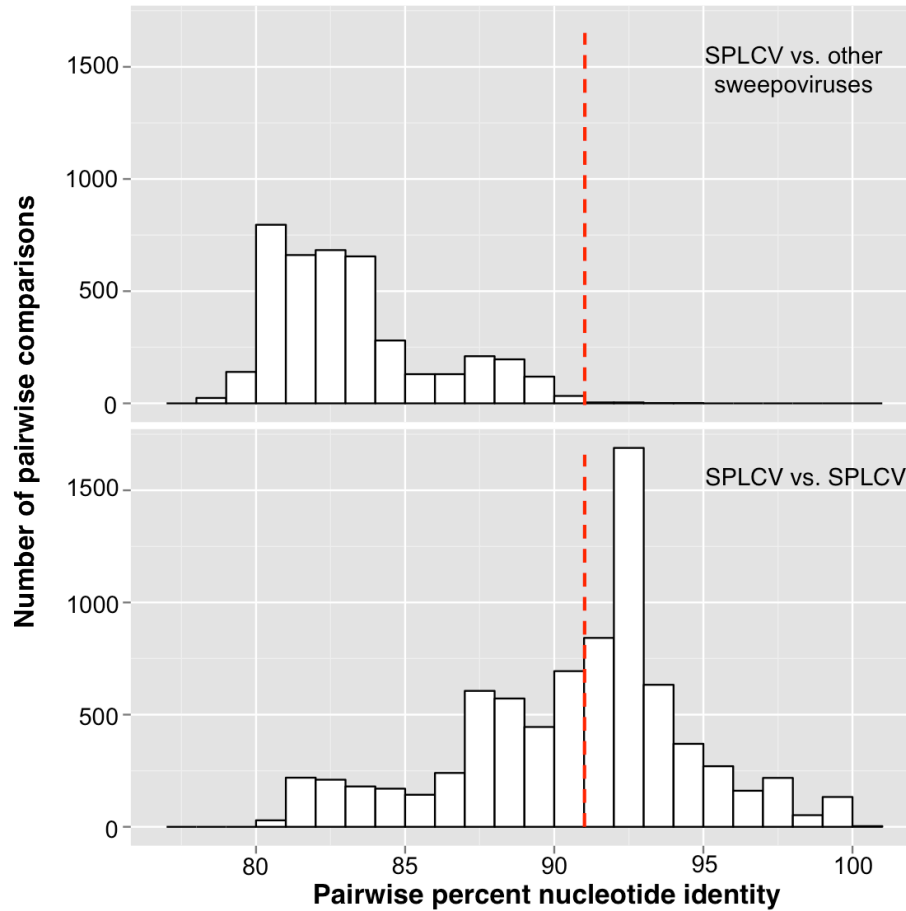
#### **a. Sweepovirus identities**

Pairwise percent nucleotide identities between SPLCV and non-SPLCV sweepovirus isolates indicate that the 91% cutoff effectively distinguished isolates that are not SPLCV (Figure 1, top) from those that are. This was not unexpected, since this group was recently revised by the ICTV *Geminiviridae* study group to be in accordance with the 91% threshold (Brown et al. 2015). Within the SPLCV species however, the pairwise identities ranged between 80.5%-100%, with almost half (45%) of the identities being less than 91% (Figure 1, bottom).

**b. VEM-sampled sweepoviruses are *Sweet potato leaf curl virus***

All five novel sweepovirus genomes have been submitted to GenBank with the following accession numbers: KT099133 (PR3\_1), KT099140 (PR10\_3), KT099143 (PR11\_2), KT099144 (Sp2\_1) and KT099145 (Sp3\_1). The three sweepovirus genomes identified from Puerto Rico by VEM on whiteflies (PR3\_1, PR10\_3, PR11\_2) belong to the SPLCV species, since they are most similar to each other (Table 1), and are 92-95% identical to other SPLCV isolates (Figure 2b). Based on the 94% nucleotide identity cutoff to distinguish strains, PR3\_1 and PR11\_2 are variants of the same strain (Table 1); PR10\_3 could also be a variant based on identity to PR3\_1 (95%), although it is <94% identical to PR11\_2, leading to a discrepancy in classification. Recombination detection analyses shed some light on this strain ambiguity: although a single recombination event was detected in all three isolates, an additional event was detected only in PR3\_1 and PR11\_2, but not in PR10\_3. The identity between PR3\_1 and PR10\_3 might be attributable to a different recombination event in which PR3\_1 was identified as the minor parent contributing to PR10\_3 (Table 1).





**Figure 1.** Pairwise percent nucleotide identities between all isolates currently identified as *Sweet potato leaf curl virus* (SPLCV, including IYVV, SPGVV, SPLCBeV, SPLCJV, SPLCESV, and SPLCLaV) and non-SPLCV isolates (top), and between all SPLCV isolates only (bottom). The red dashed line indicates 91% identity, the cutoff value for a novel begomovirus species.

Each of the sweepovirus genomes from Spain (Sp2\_1 and Sp3\_1) is >94% identical to other known SPLCV isolates (Figure 2b, Table 1); therefore, they would be classified as SPLCV based on current species demarcation guidelines. However, these two sweepovirus genomes are only 84% identical to each other, which would have led to their classification as distinct species, had they been discovered before the SPLCV isolates. Sp3\_1, which is most identical to an isolate formerly known as SPLCESV, is also detectably recombinant, with SPLCV and SPLCESV parental sequences (Table 1).

### **c. Misclassified and novel sweepoviruses**

Three sweepovirus isolates from South Korea (JX961671, JX961973, and JX961674) are annotated as SPLCV in the GenBank database. Under the 91% nucleotide identity threshold all three are instead SPLCCNV variants (Figure 3b). Two isolates identified as novel SPLCGxV share 96% nucleotide identity with each other, and are <91% identical to all other sweepoviruses (Figure 3b). Therefore SPLCGxV should be considered as a novel species by ICTV in the future.

## **II. Ambiguous sweepovirus species**

### **a. *Sweet potato leaf curl Sichuan virus 2* or *Sweet potato leaf curl China virus***

Several sweepovirus isolates from China may belong to either SPLCSiV-2, or SPLCCNV. Four isolates identified as SPLCCNV (KJ013572-KJ013575) and one

isolate identified as SPLCHnV (KJ476509) are 98% identical to SPLCSiV-2 type isolate KF156759 (Brown et al. 2015) (Figure 3b). While KJ476509 is <91% identical to the other two SPLCHnV isolates, the other four SPLCCNV isolates are 91% identical to one other SPLCCNV isolate, KJ013576. In turn, KJ013576 is >91% identical to several other SPLCCNV isolates in addition to SPLCHnV isolate KJ476509, but not to the SPLCSiV-2 type isolate (90% identity). This suggests that these species should be considered one species, likely with the name SPLCCNV, since that designation was recognized before SPLCSiV-2.

**b. *Sweet potato leaf curl virus* or novel *Sweet potato leaf curl Shanghai virus***

All seven SPLCShV isolates are also <91% identical to all other sweepoviruses, with the exception of an isolate currently identified as SPLCV (EU309693), with which they share 94-96% nucleotide identities (Figure 3b). Since this isolate (EU309693) is 91% identical to one other SPLCV isolate (HQ333143), it has been classified as SPLCV (Figure 3b). Therefore, while the SPLCShV isolates are generally divergent from other sweepoviruses, their >91% identity to a single SPLCV sequence requires that these isolates are included in SPLCV.

**c. Divergent Spanish sweepovirus cluster within *Sweet potato leaf curl virus***

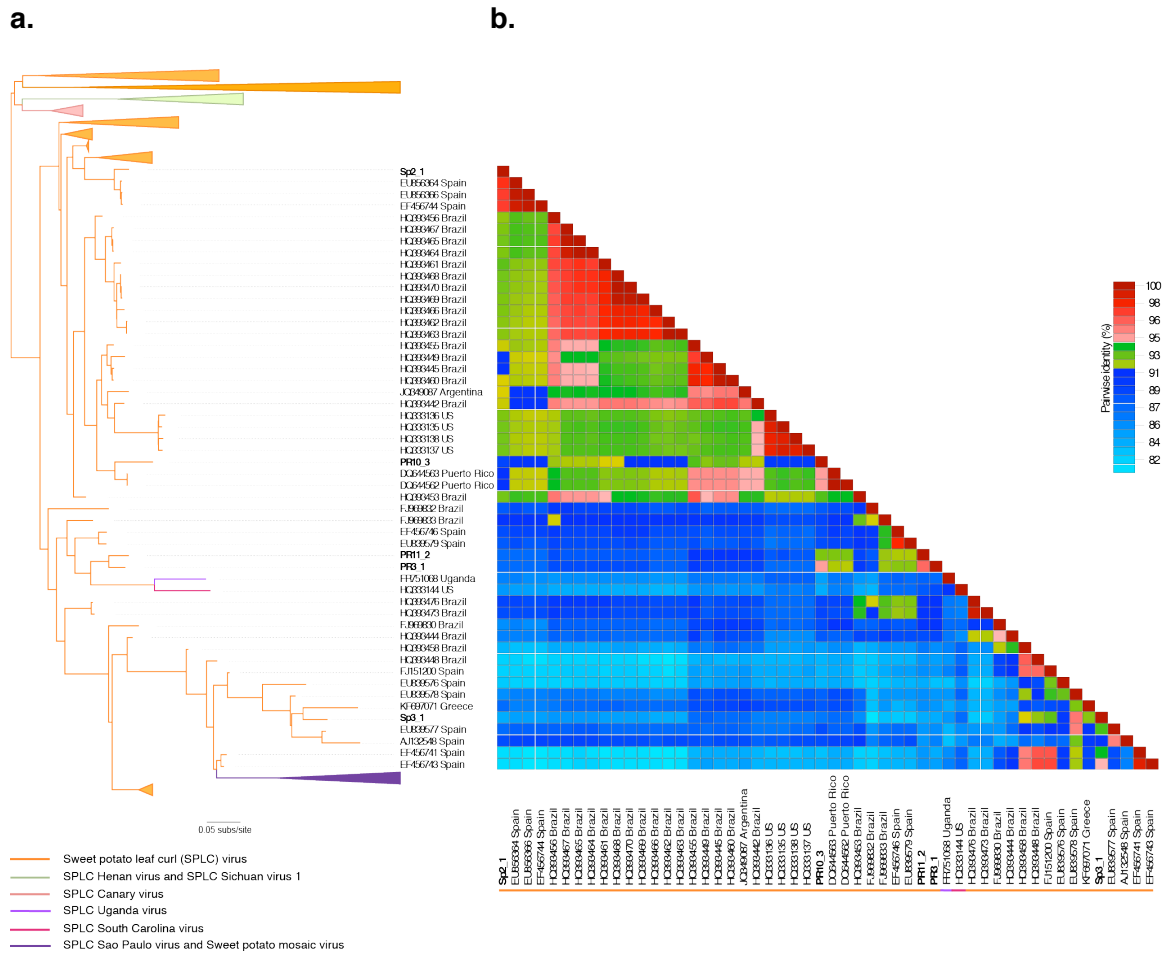
Five SPLCV isolates (EF456741, EF456743, FJ151200, HQ393448, HQ393458), which share the highest nucleotide identities (>94%) with each other, were

previously proposed to be the distinct species SPLCESV (Lozano et al. 2009). One isolate (HQ393458) is >91% identical to three former SPGVV isolates, while all five are 91%-93% identical to two SPLCV isolates (EU839576 and EU839578) formerly known as IYVV (Figure 3b, Table 2). These IYVV isolates are 92% identical to each other and EU839576 only matches one other sequence at >91% (SPLCESV isolate FJ151200). Meanwhile, EU839578 is >91% identical to KF697071, a divergent SPLCV isolate that has <91% identity to all other SPLCV sequences in this analysis, and two other SPLCV isolates formerly known as IYVV (EU839577, AJ132548). Similar to the situation demonstrated for SPLCShV above, these connections through only a single “SPLCV” isolate have brought these two former species into SPLCV (Brown et al. 2015).

Recombination events detected in the SPLCESV isolates, and SPLCV KF697071, revealed that they have IYVV, SPLCESV, and SPLCV parental sequences (Table 2). The pairwise nucleotide identities between all SPLCV isolates and the ICTV-suggested reference genomic sequences for SPLCV (EU253456) and the formerly recognized species SPLCESV (EF456741) (Figure 5), demonstrate that the wide range among SPLCV species’ pairwise nucleotide identities seen in Figure 1, is at least in part due to the inclusion of SPLCESV (and VEM Sp3\_1) isolates.

**d. *Sweet potato leaf curl Canary virus* and *Merremia leaf curl virus***

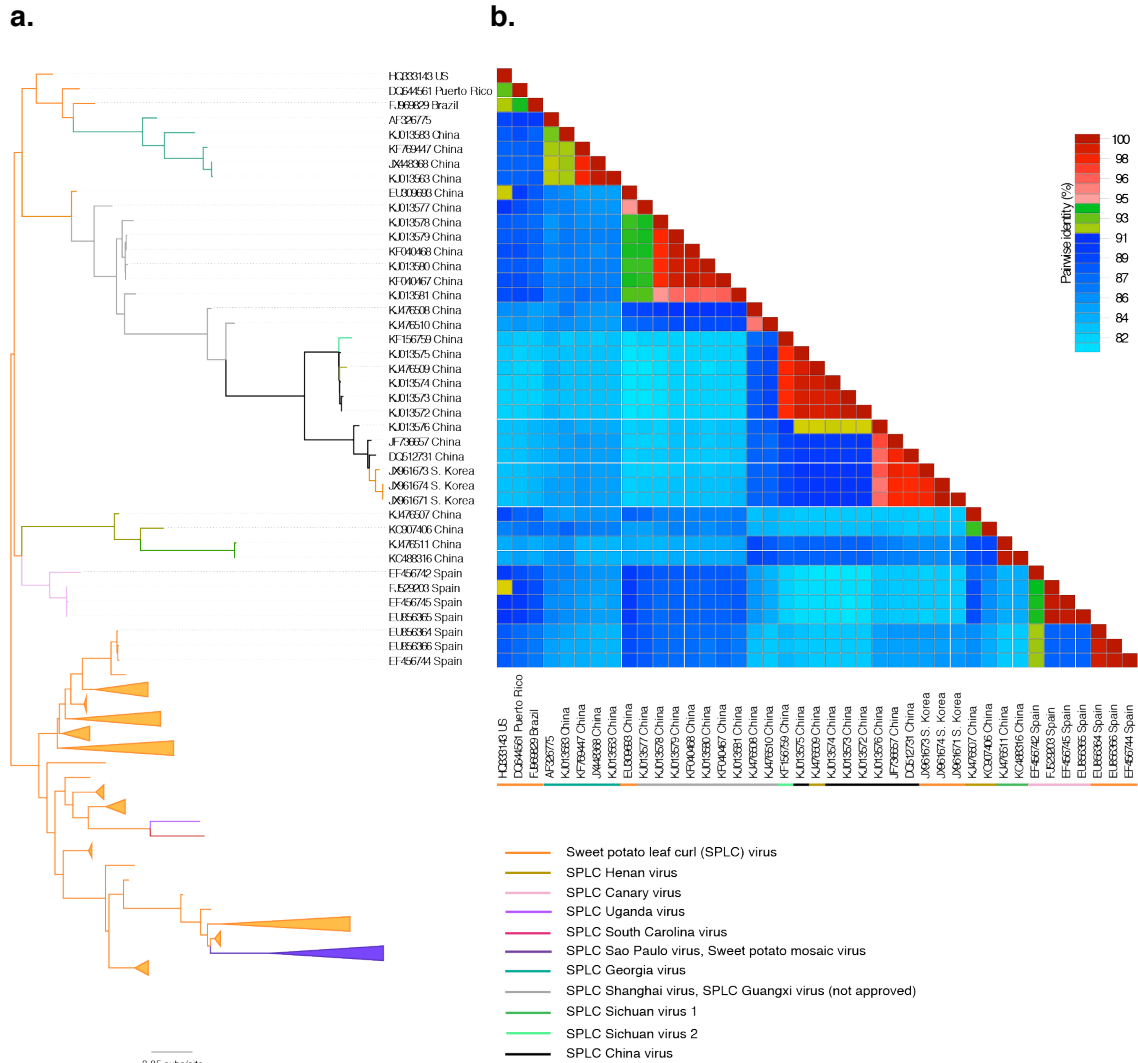
SPLCCaV and MerLCV are both recognized begomovirus species as of January 2015 (Brown et al. 2015). All four of the SPLCCaV isolates (Lozano et al. 2009) in this analysis share  $\geq 94\%$  nucleotide identity with each other, but one of the isolates, EF456742 is 92% identical to SPLCV isolates EU856366, EF456744, and EU856364 (Figure 3b). Similarly, MerLCV isolate DQ644561 is 91%-94% similar to four other SPLCV isolates: FJ969829, HQ333143 (Figure 3b), and DQ644562-3, as noted by Brown et al. (Brown et al. 2015). Therefore, both species would be considered part of the sprawling SPLCV under the current taxonomic scheme.



**Figure 2a.** Maximum likelihood phylogeny comprising all 168 sweepovirus isolates in this study; collapsed orange clades contain multiple sweepovirus species, including SPLCV (see Fig. 3a). At right, **Figure 2b** is a pairwise percent nucleotide identity matrix corresponding to sweepovirus clades containing VEM isolates (in bold) and nearly exclusively SPLCV isolates (signaled by the orange line), but also containing a single isolate of each SPLCUV (purple) and SPLCSCV (red). The matrix cells are colored blue indicating <91% identity (below species threshold), green indicating 91%-94% identity (same species, different strains), and red indicating >94% identity.

**Table 1.** Recombination events and best-matching pairwise nucleotide identities of novel VEM sweepovirus isolates. Significant recombination events detected by more than two methods in RDP v3.44 ( $p < 0.001$ ) were accepted; methods used are indicated as R: RDP, G: GeneConv, B: BootScan, M: MaxChi, C: Chimaera, S: SiScan, 3S: 3Seq.

VEM	Best match (Identity)	Parental sequences		Breakpoints	Methods
		Major	Minor		
PR3_1	PR11_2 (96%)	Unknown	HQ333138 (SPLCV)	1930-2901	G, B, M, C, S, 3S
		HM754639 (SPLCV)	KC907406 (SPLCHnV)	?	G, B, M, C, S, 3S
PR10_3	PR3_1 (95%)	Unknown	HQ333138 (SPLCV)	1686-?	G, B, M, C, S, 3S
		HQ393467 (SPLCV)	PR3_1	50-2229	G, B, M, C, S, 3S
PR11_2	PR3_1 (96%)	Unknown	HQ333138 (SPLCV)	1601-2793	G, B, M, C, S, 3S
		HM754639 (SPLCV)	KC907406 (SPLCHnV)	?	G, B, M, C, S, 3S
		HQ393453 (SPLCV)	HQ393458 (SPLCESV)	616-983	R, G, B, M, C, S, 3S
Sp2_1	EU856364 (SPLCV) (97%)	HQ393444 (SPLCV)	KF697070 (SPLCV)	?-1250	G, S, 3S
Sp3_1	EF456741 (SPLCESV) (94%)	EF456741 (SPLCESV)	KJ013557 (SPLCV)	2360-2943	R, G, B, M, C, S, 3S
		EF456741 (SPLCESV)	KF040465 (SPLCV)	2360-2942	G, B, M, C, S, 3S
		EU839577 (IYVV)	Unknown	?-1869	G, B, M, C, S, 3S



**Figure 3a.** The same phylogeny as in 2a, with top clades expanded, and bottom collapsed. At right, **Figure 3b** is a pairwise percent nucleotide identity matrix corresponding to sweepovirus clades containing SPLCV (orange), SPLCHnV (olive), SPLCCaV (light pink), SPLCGoV (turquoise), SPLCCNV (black), SPLCSiV-1 and 2 (green and light green), and the unofficial SPLCShV and SPLCGxV (both grey) isolates. Species identifications correspond to information in GenBank and/or ICTV, and do not reflect the changes recommended in this paper. The blue matrix cells indicate <91% identity (below species threshold), green indicates 91%-94% identity (same species, different strains), and red indicates >94% identity.



**Table 2.** Recombination events and top two best-matching pairwise nucleotide identities for SPLCESV isolates. Significant recombination events detected by more than two methods in RDP v3.44 ( $p < 0.001$ ) were accepted; methods used are indicated as R: RDP, G: GeneConv, B: BootScan, M: MaxChi, C: Chimaera, S: SiScan, 3S: 3Seq.

SPLCESV	Best match (Identity)	2nd match (Identity)	Parental sequences		Breakpoints	Methods
			Major	Minor		
EF456741	EF456743 (98%)	EU839578 (IYVV) (91%)	Unknown	HQ393470 (SPLCV)	2943-2064	G, B, M, C, S, 3S
EF456743	EF456741 (98%)	EU839578 (IYVV) (91%)	Unknown	HQ393470 (SPLCV)	2943-2112	G, B, M, C, S, 3S
FJ151200	EF456743 (96%)	EU839578 (IYVV) (93%)	EU839577 (IYVV)	Unknown	?	G, B, M, C, S, 3S
			HQ393465 (SPLCV)	EF456743 (SPLCESV)	?-2781	B, S, 3S
			HQ393452 (SPLCV)	FR751068 (SPLCUV)	580-?	B, C, S
			EF456746 (SPLCV)	EF456741 (SPLCESV)	2944-168	M, S, 3S
HQ393448	EF456743 (96%)	EU839578 (IYVV) (90%)	EU839577 (IYVV)	Unknown	?	G, B, M, C, S, 3S
			Unknown	HQ393470 (SPLCV)	392-2097	G, B, M, C, S, 3S
HQ393458	HQ393448 (96%)	EU839578 (IYVV) (92%)	EU839577 (IYVV)	Unknown	?	G, B, M, C, S, 3S
			HQ393465 (SPLCV)	EF456743 (SPLCESV)	?-2943	B, S, 3S
			HQ393452 (SPLCV)	FR751068 (SPLCUV)	533-?	B, C, S
KF697071 (SPLCV)	EU839578 (IYVV) (92%)		EF456741 (SPLCESV)	KJ013557 (SPLCV)	1999-326	R, G, B, M, C, S, 3S
			EU839577 (IYVV)	Unknown	?-1992	G, B, M, C, S, 3S

### **III. Phylogenetic approaches to classification**

As monophyly is considered an essential or, at least, desirable property of a species, we also considered phylogenetic approaches to sweepovirus species demarcation. A bootstrap-supported cladogram representing the phylogenies shown in Figures 2a and 3a shows that bootstrap support for the basal branches in the sweepovirus phylogeny is remarkably weak (Figure 4). The large SPLCV group is clearly paraphyletic, giving rise to smaller, <91% identical groups, but even when these are monophyletic groups, they are not supported with high confidence. Attempts at removing the long intergenic region (which is considerably less well-conserved than the ORFs and harder to align) from the whole genome alignment did not significantly improve node support (Supplementary Figure 1). Finally, as begomovirus classification previously relied on percent amino acid identity in the coat protein, we also looked at a genealogy based solely on CP protein sequences (Supplementary Figure 2); however, this approach had even poorer phylogenetic resolution.





**Figure 5.** Pairwise percent nucleotide identities between the ICTV reference sequences for *Sweet potato leaf curl virus* (SPLCV) and *Sweet potato leaf curl Spain virus* (SPLCESV) (King AMQ, Adams MJ, Carstens EB, Lefkowitz 2012), and all available sequences of SPLCV, including IYVV, SPGVV, SPLCBeV, SPLCJV, SPLCESV, and SPLCLaV (orange squares), and SPLCESV (teal squares). Orange square with teal fill: divergent SPLCV isolate KF697071; teal square with orange fill: VEM isolate SP3\_1.

## DISCUSSION

The definition of what constitutes a species for any biological entity has always been fertile ground for discussion, and continually shifts as cases and exceptions are discovered and debated. As Darwin eloquently wrote in *On The Origin of Species*, “ No one definition has satisfied all naturalists; yet every naturalist knows vaguely what he means when he speaks of a species.” One of the fundamental questions surrounding this problem is whether species are actual biological objects, or mental constructs for the purpose of categorization. From the prokaryotic perspective where horizontal gene transfer can occur across taxonomic boundaries, species has been considered a practical, abstract concept for classifying organisms in some logical manner that incorporates biological reality (Rossello-Mora and Amann 2001; Van Regenmortel 2003). While there was much opposition to the idea of recognizing species as a taxonomic rank for viruses, the ICTV nevertheless saw the need for specific classification, and adopted the definition put forth by van Regenmortel (Van Regenmortel 1989) that a species is “a polythetic class of viruses that constitutes a replicating lineage and occupies a particular ecological niche (King AMQ, Adams MJ, Carstens EB, Lefkowitz 2012).”

The actual task of classifying viruses by that definition however, remains difficult (Van Regenmortel 2003). Biological, ecological, phylogenetic, and phenetic species concepts have all been found lacking for viruses, especially in light of rampant horizontal gene transfer (Morgan and Pitts 2008). The ability to

interbreed within a population, and reproductive isolation from other populations as required in the biological species concept, do not apply to asexual and recombining/reassorting viruses. Further, the ecological and phylogenetic species concepts are respectively confounded by difficulties in defining viruses' ecological niche, and determining lineage and speciation events within reticulate evolutionary histories, while the phenetic species concept has to rely on often-arbitrary selection of categories and thresholds to determine relatedness (Morgan and Pitts 2008). With the variety of properties that can be considered for polythetic classification, species demarcation guidelines as determined by the ICTV Study Groups vary widely between groups of viruses. For example, the species demarcation criteria for the genera *Begomovirus* and *Mastrevirus* (family *Geminiviridae*) both recommend that serological differences between coat proteins, and inability of the replication-associated protein to *trans*-replicate genomic components may be used to indicate different species, while nucleotide sequence identity-based thresholds for novel species are the very different <91% and <78% respectively (King AMQ, Adams MJ, Carstens EB, Lefkowitz 2012; Muhire et al. 2013; Brown et al. 2015). Meanwhile in the *Totiviridae* family of dsRNA viruses, some of the criteria for identifying novel species include infection of distinct host species, and amino acid sequence identity of <50% (or <60% for the genus *Victorivirus*) in the coat protein, or RNA-dependent RNA polymerase (King AMQ, Adams MJ, Carstens EB, Lefkowitz 2012). The range of species demarcation criteria across different families and genera reflects the diversity of virus properties.

Despite being controversial and lacking universality, the concept and definitions for virus species are important mainly because the taxonomic level of genus is still broad enough to encompass groups of viruses that can be distinct phylogenetically, or divided by biologically relevant criteria such as host range (e.g., *Totivirus*), and genome organization (e.g., *Alpha-*, *Beta-*, *Gammacoronavirus*). Better resolution of virus groups at the species level will allow for more precise discussion of the biology and epidemiology of particular virus groups. Another point in favor of the taxonomic ranking of species is that it indicates common evolutionary origins amongst virus isolates, as opposed to a non-hierarchical grouping of isolates purely on the basis of shared properties (Van Regenmortel 2007).

While the ICTV virus species demarcation guidelines consist of multiple criteria, genome sequence information is usually the easiest to obtain and interpret for initial identification of novel isolates. Percent nucleotide identity is not the only species demarcation criteria for begomoviruses, but it is heavily relied on for novel species assignment. With the increasing number of isolates recovered by metagenomic sequencing and availability of next-generation sequencing technologies, virus identification will increasingly rely primarily on genomic sequence information (Bao et al. 2014; Brown et al. 2015; Simmonds 2015). In anticipation of this, the *Geminiviridae* Study Group has updated the guidelines for *Begomovirus* classification; pairwise alignments of whole genome sequences

excluding gaps are used to calculate pairwise nucleotide identities, leading to a revision of the previous species demarcation criterion from 89% to 91%, while strain demarcation criterion was maintained at 94% (Brown et al. 2015).

This pairwise method of calculating percent nucleotide identity is a departure from the one used to establish the former 89% threshold, which depended on multiple sequence alignments, and treated gaps as fifth-state characters (Muhire et al. 2013). It has been noted that identities calculated using different algorithms are not directly comparable (Bao et al. 2014), which supports the effort by the *Geminiviridae* Study Group to standardize percent identity calculations for all geminiviruses. Nevertheless, percent nucleotide identity itself as a measure to determine species membership suffers from other weaknesses in addition to different values resulting from different alignment algorithms and gap treatments. Using percent identity to compare genomes obscures evolutionary history (Simmonds 2015), especially those of recombinant viruses since percent identity is essentially an average distance measure across the whole genome (i.e., two genomes can be 90% identical, and be 100% identical in half the genome, and only 80% identical in the other half). Substitution saturation in divergent sequences can also lead to meaningless percent identity values that in extreme cases only reflects nucleotide frequencies (Xia XH 2009); it has been noted that global alignment of two distantly related random sequences of the same size can result in identities of up to 50% (Bao et al. 2014).



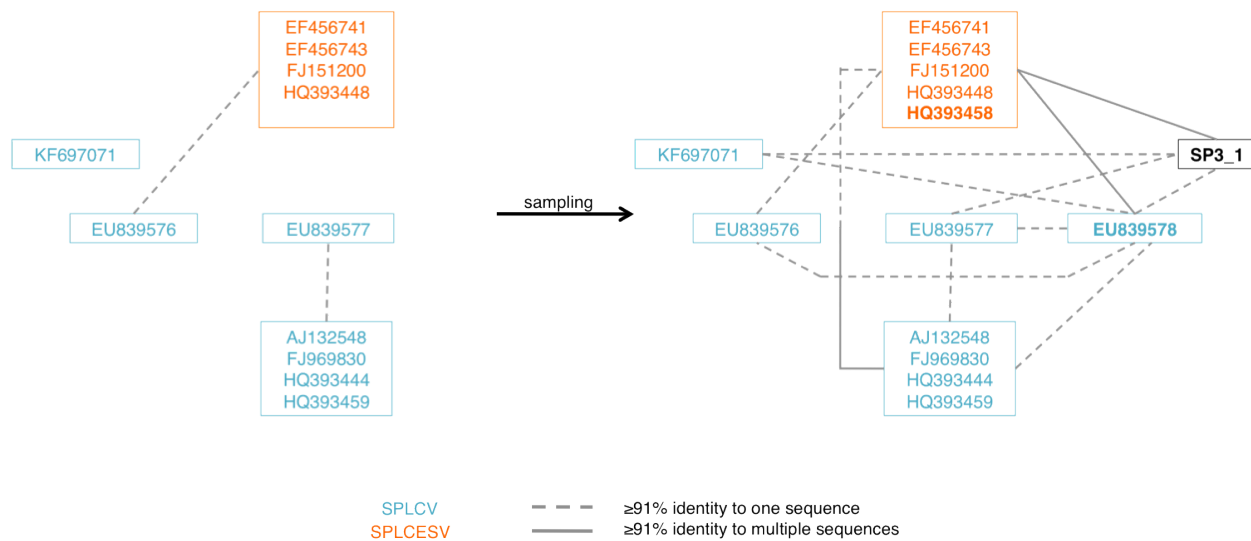
Monopartite *Begomovirus* members that infect convolvulaceous plants such as sweet potato and related wild varieties form a small divergent clade, which afforded us an opportunity to evaluate whole genome pairwise percent nucleotide identity as a means to classify a subset of this most speciose genus. The recent change in begomovirus species demarcation criteria resulted in a decrease in the number of previously identified sweepovirus species from 17 to 12 named species (Brown et al. 2015). By the current guidelines, as long as a sequence shares  $\geq 91\%$  identity to any previously-identified isolate, it should belong to the same species. Therefore, the order of discovery exerts an undue influence in virus identification and nomenclature.

While the new species cutoff value was selected to minimize the number of sequences that might share identities above the cutoff values to more than one recognized species, the Study Group also recommended that in such cases, the isolate should be assigned to the species with which it has the highest match. Following these recommendations means that virus isolates can have low percent nucleotide identities to others within the same species. Therefore, divergent clusters of virus isolates can be classified as one species based on a continuous chain of pairwise identities to single isolates, forming a highly diverse group. This is reminiscent of ring species in macroorganisms, in which two separate (reproductively isolated) species are connected by a chain of interbreeding populations (Irwin et al. 2005). We observed this diversity in our dataset: a large proportion of SPLCV sequences are below the 91% threshold

when compared to other members of the same species (Figure 1). One example from our dataset is SPLCV isolate EU309693, which is >91% to only one other SPLCV isolate, but 93%-95% identical to the novel (but as yet unrecognized) SPLCShV species. Therefore, EU309693 is the sole link between the divergent SPLCShV and SPLCV, even though all seven of the SPLCShV isolates are <91% to other SPLCV isolates. On the other hand, such a tenuous situation based on a single sequence can clearly disappear with increased sampling and recovery of viruses that link two or more sequence clusters, as illustrated by the case of SPLCV and former SPLCESV isolates (Figure 6).

Our larger analysis of all sweepoviruses prompts some changes to the very recently revisited begomovirus species list (Brown et al. 2015). By following current species demarcation criteria, we find no support for separate species named SPLCCaV or MerLCV and suggest that these sequences belong to SPLCV. In addition, since the divergent SPLCShV isolates share >91% nucleotide identity with a single previously identified SPLCV isolate, these genomes should also be lumped into SPLCV. We further propose merging SPLCSiV-2 into SPLCCNV, reducing the number of recognized sweepoviruses by three, but also support the recognition of SPLCGxV as a species. If our recommendations are adopted by ICTV, then there would be 10 current sweepovirus species (Supplementary Table 1).

While we still adhere to the current guidelines, we also show that it is highly problematic when only a single isolate needs to match at the identity threshold. It is easy to imagine how sweepvirus sequences sampled in the future, especially recombinant isolates, will continue to have >91% nucleotide identity with at least one SPLCV. The two most divergent sequences within the current SPLCV species are only 80% identical, and almost half of the within-SPLCV pairwise comparisons are below the 91% threshold. This classification scheme is likely to cause significant lumping of previously separate groups into SPLCV as more viruses are sequenced (e.g., Figure 6) and obscure biologically distinct isolates. Our analysis of the divergent SPLCESV cluster within SPLCV suggests that a consensus sequence-based delineation of species should be considered. Novel isolates can be compared to an ICTV-curated type isolate, and any percent nucleotide identity threshold should be met by all isolates.



**Figure 6.** SPLCV and SPLCESV sequence clusters merging due to sampling of three additional isolates.

These classification considerations are not unique to the begomoviruses, and the adequacy of pairwise distances in virus identification has been investigated.

Lauber and Gorbalenya demonstrated with picornaviruses that since uncorrected pairwise distances do not correct for multiple substitutions, the distances tend to underestimate the genetic divergence between sequences, unlike distance calculations that incorporate an evolutionary model (Lauber and Gorbalenya 2012). Similarly motivated by the idea that sequence information should be sufficient for virus taxonomic classification, Lauber and Gorbalenya devised a method of species demarcation based on pairwise evolutionary distance clustering of picornaviruses. Although they do not address novel species assignment and this method has not been optimized for high throughput use (Bao et al. 2014), this method was found to correspond closely with existing ICTV recommendations that rely on other criteria in addition to genomic sequences (Lauber and Gorbalenya 2012).

Phylogenetic support for species classification has also been explored, specifically with noroviruses, in which sequences that fall within well-supported clades are proposed to belong to the same species (Kroneman et al. 2013). Our dataset proved to have poor phylogenetic support for the species that were supported by overall genetic distance. SPLCV is clearly paraphyletic, and while many paraphyletic eukaryotic species are recognized and accepted (Crisp MD 1996), this clearly conflicts with the overall goal of viral species reflecting monophyly. In addition, there have been efforts to draw clear boundaries

between norovirus species clusters, with the requirement that the average distance between all sequences within a cluster be at least 2 standard deviations from another cluster (Kroneman et al. 2013). Since recombination is a frequent occurrence in noroviruses (as it also is in begomoviruses), a naming convention has been proposed that accounts for the distinct origins of the two relevant ORFs used in norovirus genotyping (Kroneman et al. 2013). Our results show that there are much broader ranges of genetic distance among isolates of sweepoviruses and this approach might not be as applicable to the highly recombinogenic begomoviruses (Lefeuvre and Moriones 2015).

Though percent nucleotide identity may be intuitive for species identification and classification it is not an ideal method, especially if the rank of species is to imply shared evolutionary origins. An average identity-based measure not only obscures reticulate evolution, but is also highly dependent on adequate sampling of virus sequence space.

## CHAPTER 4

### EXPLORING CYTOSINE → THYMINE BIAS IN BACTERIOPHAGE phiX174

#### Abstract

Cytosine-to-thymine substitution bias has been observed in single-stranded (ssDNA) viruses, and is particularly pronounced in the bacteriophage phiX174. Since phiX174 uses its host *E. coli* replication machinery, which is not known to introduce such mutations, we investigate the possibility that this over-represented substitution is a result of spontaneous cytosine deamination. We replaced the cytosine in the start codon of the phiX174 G gene to measure the rate of C→T reversion at different incubation temperatures (4°C and 45°C). However, despite successful start codon mutagenesis, we were unable to recover mutant bacteriophage, due to either the high rate of cytosine transition, or recombination between the mutant bacteriophage and the *trans*-complementing wildtype gene carried by the permissive *E. coli* host on a plasmid. An additional assay to determine the types of mutations that accumulate in wildtype phiX174 genomes incubated at 4°C and 45°C was inconclusive due to *E. coli* contamination.

#### Introduction

The high substitution rates seen in ssDNA viruses, which are on the same order of magnitude as fast-evolving RNA viruses (Duffy et al. 2008), do not result from mutations introduced by error-prone RNA-dependent RNA polymerases as in the

case of RNA viruses since ssDNA viruses co-opt their hosts' processive DNA polymerases for replication. Since positive selection also does not appear to be increasing their substitution rate, it is likely that the ssDNA genomes are undergoing spontaneous oxidative damage. Cytosine is easily deaminated forming uracil, leading to the introduction of thymine in the virus genome in the next round of replication. Single-stranded DNA is also more vulnerable to cytosine deamination than double-stranded DNA (Lindahl and Nyberg 1974; Frederico et al. 1990), and cytosine-to-thymine substitution bias has been observed in ssDNA viruses (Duffy and Holmes 2008; Duffy and Holmes 2009). If spontaneous deamination is responsible for the cytosine-to-thymine substitution bias, which may be driving the high ssDNA virus substitution rates, the rate of deamination can be increased by heat (Lindahl and Nyberg 1974), and it is expected that spontaneous deamination rates will double with every 10°C increase in temperature, as per the Arrhenius equation (Laidler 1984).

The bacteriophage phiX174 (Family *Microviridae*) has a small, circular ssDNA genome that is 5,386 bases in length. PhiX174 and its host bacteria *E. coli* have long been used in the laboratory for molecular biology and experimental evolution studies, and are therefore a well-studied host-virus system. PhiX174 genome sequences also display significant cytosine-to-thymine substitution bias (Chapter 2), and appear to frequently sample C→T mutations during experimental evolution (Rokyta et al. 2005). Therefore, it is chosen as the model organism for



exploring the role of spontaneous cytosine deamination in ssDNA viruses. We employ two assays for investigating cytosine deamination in phiX174: a phenotypic reversion assay utilizing a start codon mutant of phiX174 (ATG→ACG), and a more general mutation accumulation assay. For the former assay, the start codon mutation is introduced in the G gene, which does not have any overlapping genes. The rate of start codon mutant reversion (ACG→ATG) in phiX174 virions incubated at different temperatures could then be measured and compared. Meanwhile, next generation sequencing will be used to probe the accumulation of mutations in wildtype phiX174 virions also incubated at different temperatures. The minimum and maximum incubation temperatures in both assays were 4°C and 45°C, and the 41°C temperature difference will allow us to observe a 16-fold difference in reaction rates.

## Materials and Methods

### *Model system*

The wildtype *E. coli* host BTCC 122 (Fane and Hayashi 1991), and bacteriophage phiX174 were obtained from Dr. Bentley A. Fane of the University of Arizona. The wildtype phiX174 sequence of this strain had been previously confirmed by Duffy lab members (Supplementary Table 1). The permissive *E. coli* host used for complementation of mutant phiX174 is BAF30(pφXG), a BTCC122 derivative (Fane et al. 1992) bearing the wildtype phiX174 G gene on plasmid pSE420 (S. Doore, personal communication), also obtained from the Fane

laboratory. Both *E. coli* hosts were cultured in TK (1.0% tryptone, 0.5% KCl) broth or agar media (Fane et al. 1992); TK media used for culturing BAF30(p $\phi$ XG) were supplemented with 50  $\mu$ g/mL of ampicillin.

Confirmation of p $\phi$ XG and the gene it carries was done by plasmid extraction using the AxyPrep plasmid miniprep kit (Axygen Biosciences, Union City, CA), followed by Sanger sequencing (Genewiz, South Plainfield, NJ). Sequencing primers were based on the G gene sequence (Table 1).

#### *Site-directed mutagenesis*

The protocol for PCR-based site-directed mutagenesis was slightly adapted from the one obtained from the laboratory of Dr. Holly A. Wichman of the University of Idaho. First, PCR was used to simultaneously introduce the desired mutation and amplify two overlapping fragments that cover the entire phiX174 genome.

Primers with the desired point mutations were designed to pair with phiX174 primers that will allow at least a 700-base overlap between the two amplified fragments of half-genomes. PCR was done using the Platinum® *Pfx* high fidelity DNA polymerase kit (Life Technologies, Carlsbad, CA) (Table 2). Presence of mutations was confirmed by Sanger sequencing.

Next, another PCR reaction is used to join the two half-genome products to make a complete genome; the overlapping half-genome products act as both template and primers (Tables 2 and 3).

**Table 1.** Primers used in site-directed mutagenesis and sequence confirmation.

F=forward, R=reverse primers; numbers in primer names refer to starting nucleotide position in phiX174 genome. All primer sequences are listed in 5' to 3' direction. Point mutations are underlined in the sequence.

Purpose	Primer pairs	Sequence
Mutagenesis (fragment 1)	F2384 R5199	GGAGTTTAATC <u>ACG</u> TTTCAGACTTT GGATTAAGCACTCCGTGGA
Mutagenesis (fragment 2)	F3917 R2408	ACCGTCAGGATTGACACC AAAGTCTGAAAC <u>CG</u> TGATTAAACTCC
Amplification for sequence confirmation	F2354 R2953	CCAAGCGAAGCGCGGTAGGT CCGCCAGCAATAGCACC

**Table 2.** Master mix used in site-directed mutagenesis.

Reagent	Amount per	Amount per
	mutagenesis reaction	joining reaction
10X <i>Pfx</i> amplification buffer	10 $\mu$ L	10 $\mu$ L
Millipore purified water	30.1 $\mu$ L	17.1 $\mu$ L
MgSO <sub>4</sub> , 50 mM	1 $\mu$ L	1 $\mu$ L
Mutagenesis primer, 10 $\mu$ M	1.5 $\mu$ L	-
Forward/Reverse primer, 10 $\mu$ M	1.5 $\mu$ L	-
dNTP blend, 10 mM (GeneAmp®, Life Technologies)	1.5 $\mu$ L	1.5 $\mu$ L
<i>Pfx</i> polymerase	0.4 $\mu$ L	0.4 $\mu$ L
Template	4 $\mu$ L	-
Half-genome product	-	10 $\mu$ L x2
<i>Total Reaction Volume</i>	<i>50 <math>\mu</math>L</i>	<i>50 <math>\mu</math>L</i>

### *E. coli transformation via electroporation*

A culture of BAF30(p $\phi$ XG) was started with a single isolated colony inoculated into 20 mL of TK media and incubated overnight at 37°C in a shaking incubator. The overnight culture was then used to start a fresh culture of BAF30(p $\phi$ XG) cells (1:100 dilution) that were harvested at exponential phase, as determined by doubling of optical density measurements at 600 nm. Preparation of electrocompetent cells, and electroporation were according to the standard protocol (Bio-Rad Laboratories MicroPulser<sup>TM</sup>, Hercules, CA).

After electroporation, 30  $\mu$ L of the transformed cells were added to 3 mL of TK soft agar (0.6% agar) for plating. Isopropyl  $\beta$ -D-1-thiogalactopyranoside (IPTG) was also added to the soft agar to induce p $\phi$ XG expression. The plates were then incubated overnight at 37°C till plaques formed. Plaque picks were done to confirm mutagenesis success, and potential mutants were then PCR-amplified with Kapa Taq PCR kit reagents (Kapa Biosystems, Wilmington, MA) and primers designed to cover the G gene region (Tables 1-3). PCR products were purified with Exo-SAP-IT<sup>®</sup> (Affymetrix, Santa Clara, CA) before Sanger sequencing for confirmation (Genewiz, South Plainfield, NJ).

**Table 3.** PCR programs for site-directed mutagenesis and regular PCR amplification.

Protocol	Stage	Temperature (°C)	Time (min:sec)	# of cycles
Site-directed mutagenesis	Initial denaturation	94	4:45	30, 10*
	Denaturation	94	0:15	
	Annealing	42	0:30	
	Extension	68	6:00	
	Final extension			
Amplification	Initial denaturation	94	2:00	25
	Denaturation	94	0:30	
	Annealing	51	0:30	
	Extension	71	1:30	
	Final extension	71	8:30	

\*30 cycles: mutagenesis reaction, 10 cycles: half-genome joining reaction

### *Mutation accumulation in wildtype phiX174*

High-titer lysates of wildtype phiX174 were collected by first overlaying  $10^5$  plaque-forming units/mL (pfu/mL) of bacteriophage stock with *E. coli* in soft agar, onto TK agar plates and incubating at 37°C overnight. The resulting lacey lawns were scraped and centrifuged at 3,000 rpm for 10 minutes to harvest free phage in the supernatant, which was then filtered through 0.22 µm syringe filters (Fisher Scientific, Waltham, MA). The filter-purified lysates were sampled to determine initial virus titer, and then split equally and incubated at 4°C and 45°C for two weeks. Destructive sampling every two days was done to determine virus titers over the course of incubation.

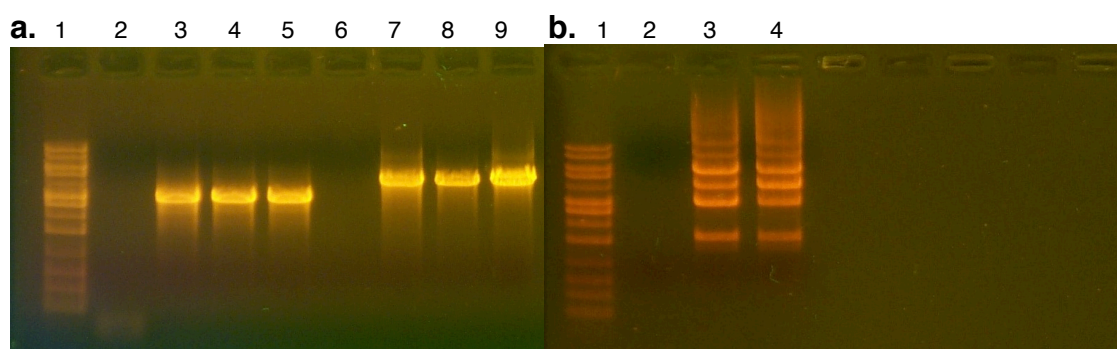
After two weeks, viral nucleic acids were purified using the QIAamp MinElute Virus Spin kit (QIAGEN, Hilden, Germany) and sent for Illumina sequencing on the MiSeq V3 platform at the Waksman Genomics Center (New Brunswick, NJ).

## **Results**

### *Site-directed mutagenesis of phiX174 G gene*

Amplification of phiX174 with both the forward and reverse mutagenesis primers produced bands of the expected lengths between 2.5-3 kb (Figure 1a), and the subsequent joining reaction produced multiple bands, one of which was of the expected length between 5-6 kb (Figure 1b). Sanger sequencing confirmed the presence of the start codon mutations in the first mutagenesis step.





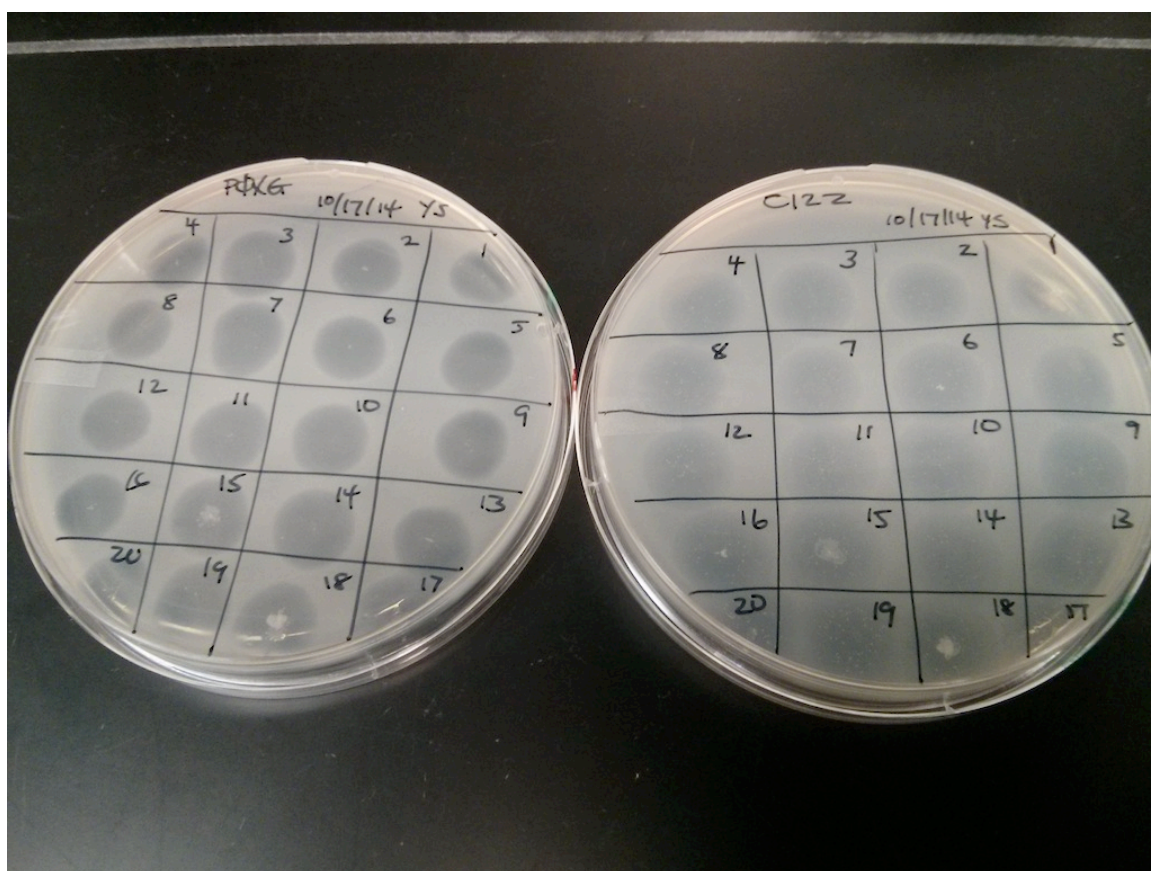
**Figure 1a.** PCR-based site-directed mutagenesis of phiX174, with the *Pfx* polymerase kit (Life Technologies). Lanes 2 and 6: negative control; Lanes 3 and 7: positive control; Lanes 4-5: F2384 (mutagenesis primer) and R5199; Lanes 8-9: F3917 and R2408 (mutagenesis primer). **1b.** Joining of the two mutagenized DNA fragments from Fig. 1a. Lane 2: negative control; Lanes 3-4: replicate joining reactions. Lane 1 in both figures are the 1-kb DNA ladder (Axygen Biosciences).

*Transformation of permissive E. coli host with start codon mutants*

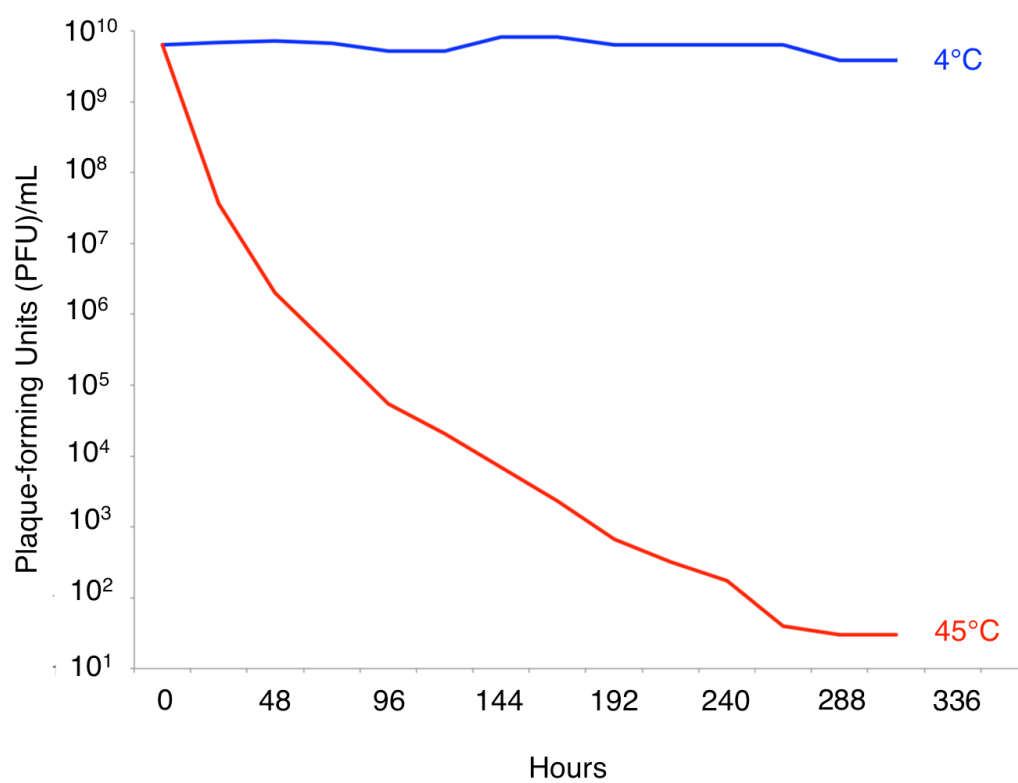
Plaques recovered from transformation were assayed on both the permissive, G gene-bearing *E. coli*, and the non-permissive wildtype *E. coli* hosts. No mutant bacteriophages were recovered, as evidenced by the plaques on the non-permissive host (Figure 2).

*Incubation titers*

Initial virus titers were  $6 \times 10^9$  pfu/mL. At the end of two weeks, the titer of virions incubated at 45°C had dropped below the detection limit ( $<10$  pfu/mL), while those incubated at 4°C remained at the same order of magnitude as the initial titer (Figure 3).



**Figure 2.** Plaque pick assay for confirmation of succesful recovery of mutant bacteriophage.



**Figure 3.** Titers of virions stored in 4°C (blue line) and 45°C (red line).

### *Illumina sequencing*

Preliminary sequencing of 2% of submitted samples indicated that the majority of nucleic acids recovered were of *E. coli* origin; thus, further sequencing was not pursued.

### **Discussion**

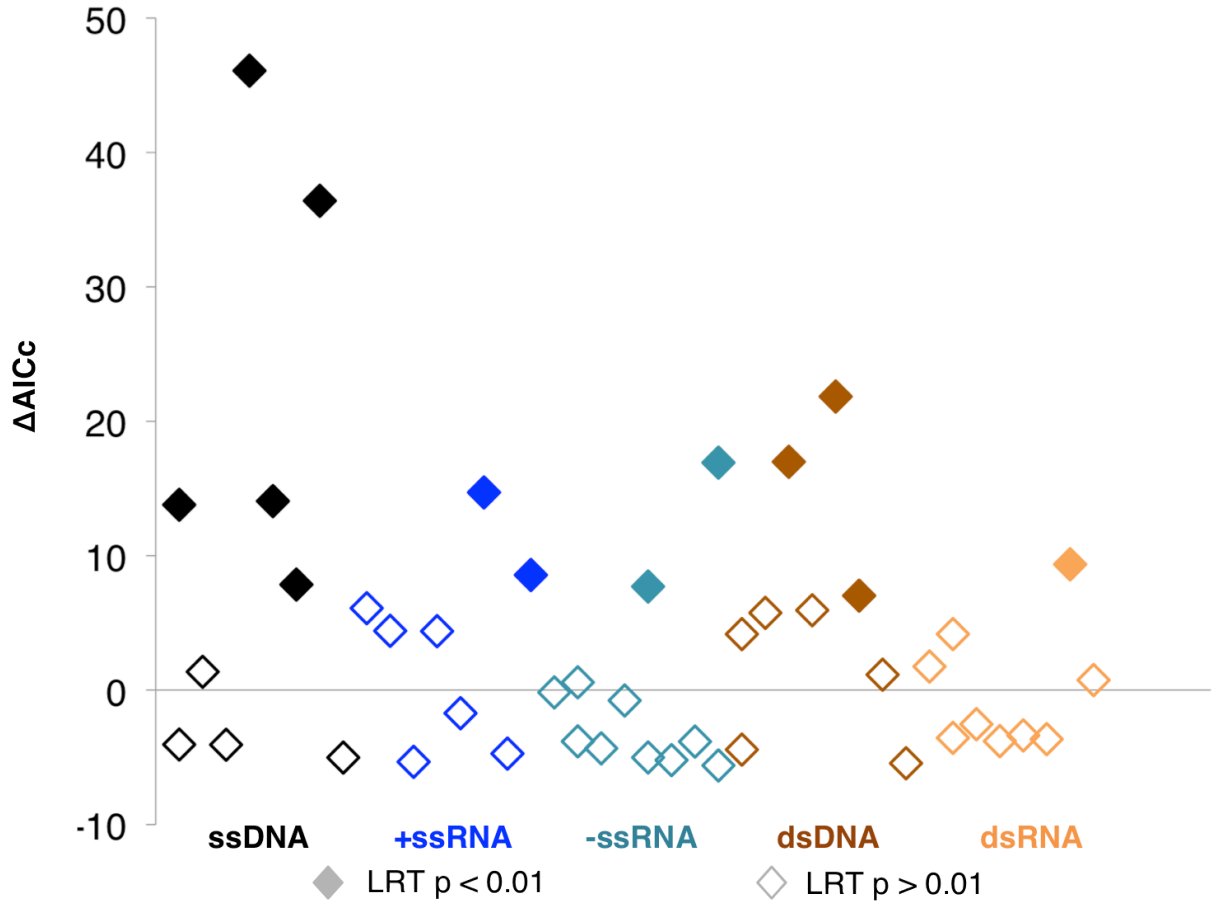
The mutation rate of phiX174 was measured at  $1.1 \times 10^{-6}$  substitutions/nucleotide/cell infection (Sanjuán et al. 2010). Considering the relatively high ratio of *E. coli* to bacteriophage during transformation, and the short time it takes for phiX174 to replicate its genome and lyse its original host to infect a new host cell (average burst size is 180 phage/*E. coli* cell (Gillam et al. 1985) and average time to burst is 25 minutes (Pereira-Gómez and Sanjuán 2014)), it is possible that after transformation, the cytosine mutation introduced into the start codon (ACG) was lost by replacement of cytosine to thymine during bacteriophage replication, resulting in a reversion to the wildtype start codon sequence. Since plaques on bacterial lawns begin to be visible to the naked eye only after approximately four hours post-inoculation, the phiX174 start codon mutant would have undergone multiple rounds of replication before sampling was done to confirm successful transformation into the bacterial host.

Recombination between the plasmid bearing the wildtype G gene and the mutant bacteriophage is another possibility to be considered, and may be detected by

introduction of synonymous “marker” mutations located close to the G gene start codon on the plasmid. The current virus purification method with 0.22  $\mu\text{m}$  syringe filters, while adequate for bacteriophage stock production, appears to be insufficiently thorough for next generation sequencing sample preparation. Prior sequencing of similarly purified phiX174 samples recovered 85% *E. coli* sequences (U. Zelzion, personal communication). Future experiments will require additional purification methods.

## APPENDIX

## Supplementary Information



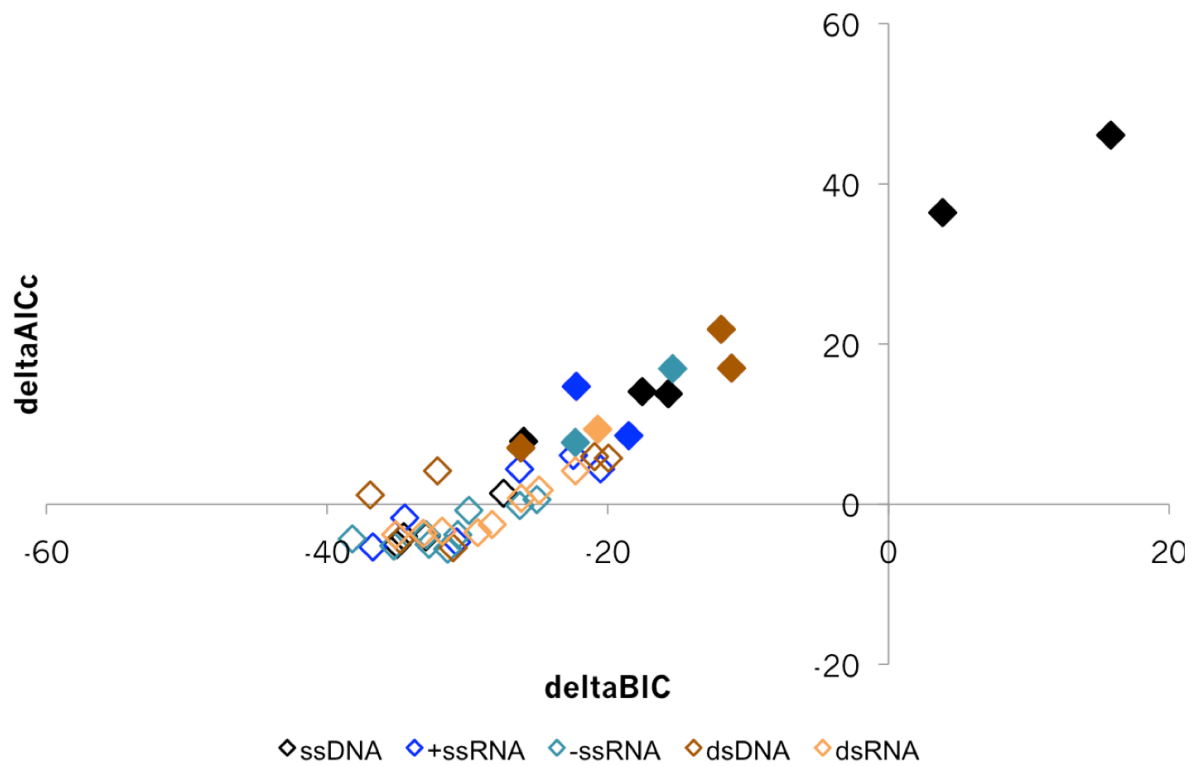
**Chapter 2, Supplementary Figure 1: Difference in AICc scores between UNREST and GTR (with ML-estimated nucleotide frequency parameters) substitution**

**models for 40 virus species.**  $\Delta AICc = AICc_{GTR} - AICc_{UNREST}$ , therefore  $\Delta AICc > 0$










indicates UNREST being more likely than GTR at fitting the data; filled diamonds show that the difference is significant by the likelihood ratio test ( $p < 0.01$ ). Diamonds that are vertically stacked in the same column represent different data points from the same virus species (e.g., alignments of different gene sequences from one virus species). The 40 virus species, from left to right, are: Human erythrovirus B19 (NS1, top, VP, bottom); Banana bunchy top virus (DNA1 segment); Beak feather disease virus; East African








cassava mosaic virus (DNA-A segment); Maize streak virus; Porcine circovirus 2; Enterobacteriophage phiX174; Wheat dwarf virus; Apple chlorotic leaf spot virus (CP); Beet necrotic yellow vein virus (CP); Cucumber mosaic virus (CP); Grapevine leafroll-associated virus 3 (CP); Hepatitis A virus (polyprotein); Japanese encephalitis virus; Potato leafroll virus (CP); Tobacco streak virus (CP); Akabane virus (nucleoprotein); Borna disease virus (P, top, N bottom); Canine distemper virus (H); Groundnut bud necrosis virus (N); Hantaan virus (Gc, top, N, bottom); Infectious hematopoietic necrosis virus (G); Rice stripe virus (CP); Viral hemorrhagic septicemia virus (G, top, N, bottom); BK polyomavirus (whole genome, top, VP1, bottom); Gallid herpesvirus 1 (UL23); Human adenovirus B (L3); Human papillomavirus 6 (L1); Human papillomavirus 16; Herpes simplex virus 1 (UL23); JC polyomavirus; Vaccinia virus (B5R); African horsesickness virus (VP7); Avian orthoreovirus ( $\sigma$ C, top,  $\sigma$ NS, bottom); Epizootic hemorrhagic disease virus 2 (VP7); Infectious bursal disease virus (RdRP); Infectious pancreatic necrosis virus (polyprotein); Rice black streaked dwarf virus (CP); Rotavirus A subtype G9 (VP7); Rotavirus C (VP7).





**Chapter 2, Supplementary Figure 2: Difference in AICc scores, and difference in BIC scores between UNREST and GTR (with ML-estimated nucleotide frequency parameters) substitution models for 40 virus species.**  $\Delta AICc$  scores are calculated as in Supplementary Figure 1, while  $\Delta BIC = BIC_{GTR} - BIC_{UNREST}$  therefore  $\Delta BIC > 0$  indicates UNREST being more likely than GTR at fitting the data.

	#taxa	#nt	C ↓ T	T ↓ C	A ↓ G	G ↓ A	A ↓ C	C ↓ A	A ↓ T	T ↓ A	C ↓ G	G ↓ C	G ↓ T	T ↓ G		
<b>ssDNA</b>																
<i>phiX174</i>																
F, G, H	67	2817														
F, G, H (3)	67	939														
F, G, H (1+2)	67	1878														
<i>EACMV</i>																
CP	63	774														
CP (3)	63	258														
CP (1+2)	63	516														
<b>+ssRNA</b>																
<i>HAV</i>																
Polyprotein	60	6699														
Polyprotein (3)	60	2233														
Polyprotein (1+2)	60	4466														
<i>ACLSV</i>																
CP	172	582														
CP (3)	172	194														
CP (1+2)	172	388														

<b>-ssRNA</b>					
<i>GBNV</i>					
N	145	831			
N (3)	145	277			
N (1+2)	145	554			
<i>RSV</i>					
CP	124	969			
CP (3)	124	646			
CP (1+2)	124	323			
<b>dsDNA</b>					
<i>VACV</i>					
B5R	39	954			
B5R (3)	39	318			
B5R (1+2)	39	636			
<i>HPV 16</i>					
L1	70	1605			
L1 (3)	70	535			
L1 (1+2)	70	1070			
<b>dsRNA</b>					
<i>RotA.G9</i>					
VP7	163	984			
VP7 (3)	163	328			
VP7 (1+2)	163	656			
<i>RBSDV</i>					
CP	82	1677			
CP (3)	82	559			
CP (1+2)	82	1118			

**Chapter 2, Supplementary Table 2: Virus species in analyses.** All sister taxa were selected based on ICTV-9 recommendations (King 2012), unless otherwise denoted by asterisks.

Family	Genus	Species	Sister Taxon
<i>Circoviridae</i>	<i>Circovirus</i>	Beak feather disease virus	Gull circovirus (NC_008521)
		Porcine circovirus-2	Porcine circovirus-1 (NC_001792)
<i>Geminiviridae</i>	<i>Begomovirus</i>	East African cassava mosaic virus	South African cassava mosaic virus (NC_003803)
	<i>Mastrevirus</i>	Maize streak virus	Digitaria streak virus (NC_001478)
		Wheat dwarf virus	Oat dwarf virus (NC_010799)
<i>Microviridae</i>	<i>Microvirus</i>	Enterobacteria phage phiX174	Enterobacteria phage G4
<i>Nanoviridae</i>	<i>Babuvirus</i>	Banana bunchy top virus	Cardamom bushy dwarf virus (JX867551)
<i>Parvoviridae</i>	<i>Erythrovirus</i>	Human parvovirus B19	Pig-tailed macaque parvovirus (AF221123)
<i>Bromoviridae</i>	<i>Cucumovirus</i>	Cucumber mosaic virus	*Peanut stunt virus (NC_002040)
	<i>Ilarvirus</i>	Tobacco streak virus	**Parietaria mottle virus (NC_005854)
<i>Flexiviridae</i>	<i>Trichovirus</i>	Apple chlorotic leaf spot virus	Apricot pseudo-chlorotic leaf spot virus (NC_006946)

<i>Closteroviridae</i>	<i>Ampelovirus</i>	Grapevine leafroll-associated virus 3	Pineapple mealybug wilt-associated virus 2 (JX645775)
<i>Flaviviridae</i>	<i>Flavivirus</i>	Japanese encephalitis virus	Usutu virus (NC_006551)
<i>Picornaviridae</i>	<i>Hepatovirus</i>	Hepatitis A virus	Avian encephalomyelitis virus (NC_003990)
<i>Luteoviridae</i>	<i>Polerovirus</i>	Potato leafroll virus	Sweet potato leaf speckling virus (DQ655700)
<i>Unassigned</i>	<i>Benyvirus</i>	Beet necrotic yellow vein virus	Beet soil-borne mosaic virus (NC_003503)
<i>Bornaviridae</i>	<i>Bornavirus</i>	Borna disease virus	Avian bornavirus (GU249595)
<i>Bunyaviridae</i>	<i>Hantavirus</i>	Hantaan virus	Seoul virus (NC_005236)
	<i>Orthobunyavirus</i>	Akabane virus	Oropouche virus (NC_005777)
	<i>Tospovirus</i>	Groundnut bud necrosis virus	Watermelon silver mottle virus (NC_003843)
<i>Paramyxoviridae</i>	<i>Morbillivirus</i>	Canine distemper virus	Phocine distemper virus (AF479277)
<i>Rhabdoviridae</i>	<i>Novirhabdovirus</i>	Infectious hematopoietic necrosis virus	Hirame rhabdovirus (NC_005093)
		Viral hemorrhagic septicemia virus	Snakehead rhabdovirus (NC_000903)

<i>Unassigned</i>	<i>Tenuivirus</i>	Rice stripe virus	Maize stripe virus (AJ969410)
<i>Adenoviridae</i>	<i>Mastadenovirus</i>	Human adenovirus B	Human adenovirus E (NC_003266)
<i>Herpesviridae</i>	<i>Iltovirus</i>	Gallid herpesvirus 1	Psittacid herpesvirus 1 (NC_005264)
	<i>Simplexvirus</i>	Herpes simplex virus 1	Herpes simplex virus 2 (NC_001798)
<i>Papillomaviridae</i>	<i>Alphapapillomavirus</i>	Human papillomavirus 16	Human papillomavirus 34 (NC_001587)
		Human papillomavirus 6	Human papillomavirus 7 (NC_001595)
<i>Polyomaviridae</i>	<i>Polyomavirus</i>	BK polyomavirus	Simian virus 12 (NC_001538)
		JC polyomavirus	BK polyomavirus (NC_001538)
<i>Poxviridae</i>	<i>Orthopoxvirus</i>	Vaccinia virus	Cowpox virus (NC_003663)
<i>Birnaviridae</i>	<i>Aquabirnavirus</i>	Infectious pancreatic necrosis virus	Yellowtail ascites virus (NC_004168)
	<i>Avibirnavirus</i>	Infectious bursal disease virus	Yellowtail ascites virus (NC_004176)
<i>Reoviridae</i>	<i>Fijivirus</i>	Rice black streaked dwarf virus	Maize rough dwarf virus (L76560)
	<i>Orbivirus</i>	African horsesickness virus	Chuzan virus (NC_005988)
		Epizootic hemorrhagic disease virus 2	Bluetongue virus (NC_006022)

<i>Orthoreovirus</i>	Avian orthoreovirus	Nelson Bay reovirus (AF218360)
<i>Rotavirus</i>	Human rotavirus A Group 9	Rotavirus Group C (X77257)
	Human rotavirus Group C	Porcine Group C rotavirus (M61101)

---

\*Boulila, M. Virus Genes. 2009. 38:435-44

\*Codoner, FM, Cuevas, JM, Sanchez-Navarro, JA, Pallas, V, Elena SF. J Mol Evol. 2005. 61:697-705

\*\*Boulila, M. Virus Genes. 2009. 38:435-44

\*\*Tzanetakis, IE, Martin, RR, Scott, SW. Arch Virol. 2010. 155:557-61.

**Chapter 2, Supplementary Table 3: Nucleotide substitution models used for inferring maximum likelihood trees in PAUP\*.** Models were selected based on AIC scores as determined by jModelTest v0.1.1. Analyses prior to the release of jModelTest v0.1.1 were done based on models selected by ModelTest v3.7. I=Invariant, G=gamma.

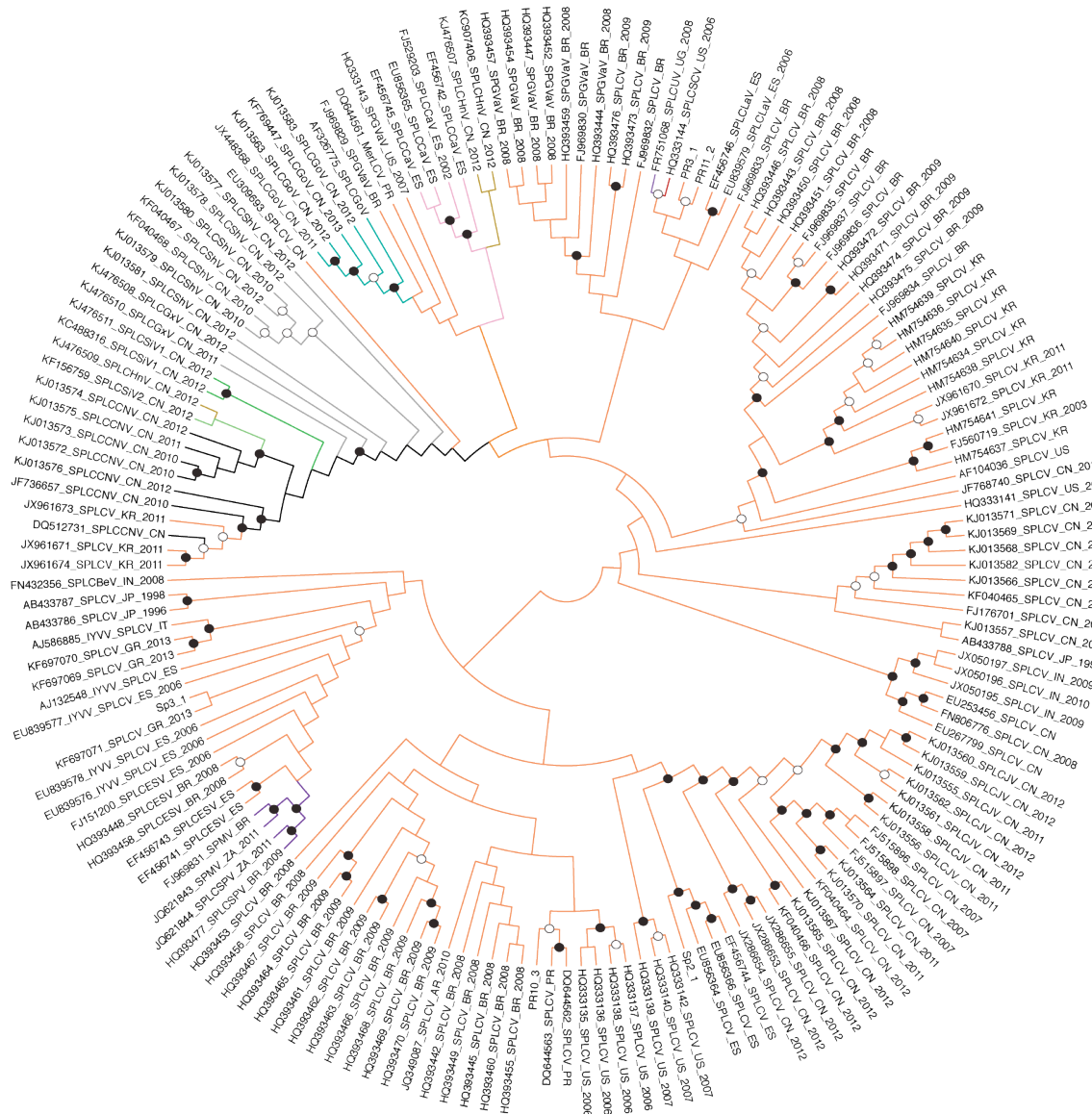
<b>Taxa</b>	<b>Selected Nucleotide Substitution Model</b>
B19 (NS1)	TIM3+I+G
B19 (VP)	TIM3+I+G
BBTV (DNA1)	TIM3+G
BFDV	GTR+G
EACMV (DNA-A)	GTR+I+G
MSV	TIM3+G
PCV2	TIM2+I+G
phiX174	TrN+I+G
WDV	TrN+I+G
ACLSV (CP)	TPM2+I+G
BNYVV (CP)	HKY+G
CuMV (CP)	GTR+G
GLRaV3 (CP)	TIM1+G
HAV (poly)	GTR+I+G
JEV	GTR+I+G
PLRV (CP)	TIM2+G
TStV (CP)	TVMef+G
AKAV (NP)	TVM+G
BDV (N)	TIM2+I+G
BDV (P)	TIM3+I+G
CDV (H)	TVM+I+G
GBNV (N)	GTR+I+G
HTNV (Gc)	GTR+I+G
HTNV (N)	GTR+G
IHNV (G)	GTR+G
RSV (CP)	TIM3+G
VHSV (G)	TVM+I+G
VHSV (N)	GTR+I+G
BKPyV	GTR+I+G



---

GaHV1 (UL23)	TPM3+G
HAdB (L3)	GTR+G
HPV6 (L1)	TIM2+G
HPV16	TVM+I+G
HSV1 (UL23)	GTR+I+G
JCPyV	GTR+I+G
VACV (B5R)	TVM+I+G
AHSV (VP7)	TVM+G
ARV ( $\sigma$ NS)	TVM+G
ARV ( $\sigma$ C)	GTR+I+G
EHDV2 (VP7)	GTR+G
IBDV (RdRP)	GTR+G
IPNV (poly)	GTR+I+G
RBSDV (CP)	TIM1+G
RotA.G9 (VP7)	TVM+G
RotC (VP7)	TIM3+I+G

---



**Chapter 3, Supplementary Figure 1.** Cladogram of sweepovirus genomes (n=168) with the long intergenic region removed. Maximum likelihood tree was inferred by RAxML v7.6.3, implementing the nucleotide model GTR+G. Branches are colored by species: SPLCV (orange), SPLCHnV (olive), SPLCCaV (light pink), SPLCGoV (turquoise), SPLCCNV (black), SPLCSiV-1 and 2 (green and light green), and the unofficial SPLCShV and SPLCGxV (both grey). Solid circles at the nodes of the tree indicate  $\geq 85\%$  bootstrap support (1,000 replicates), while open circles indicate 60%-84% support.

**Chapter 3, Supplementary Figure 2.** Cladogram of sweepovirus coat protein amino acid sequences; identical sequences were removed for this analysis, reducing the number of sequences to 113. Maximum likelihood tree was inferred by RAxML v8.1.11 implementing the JTT amino acid model. Solid circles at the nodes of the tree indicate  $\geq 85\%$  bootstrap support (1,000 replicates), while open circles indicate 60%-84% support.

**Chapter 3, Supplementary Table 1.** Sweepovirus sequences downloaded from GenBank for analyses. Accession number is listed, followed by current species abbreviation as determined by ICTV, two-letter country code, and year of isolation (if available in GenBank).

Proposed Species Membership		Isolate
1.	SPLCCNV	DQ512731_SPLCCNV_CN JF736657_SPLCCNV_CN_2010 JX961671_SPLCV_KR_2011 JX961673_SPLCV_KR_2011 JX961674_SPLCV_KR_2011 KF156759_SPLCSiV2_CN_2012** KJ013572_SPLCCNV_CN_2010 KJ013573_SPLCCNV_CN_2010 KJ013574_SPLCCNV_CN_2012 KJ013575_SPLCCNV_CN_2011 KJ013576_SPLCCNV_CN_2012 KJ476509_SPLCHnV_CN_2012
2.	SPLCGoV	AF326775_SPLCGoV JX448368_SPLCGoV_CN_2011 KF769447_SPLCGoV_CN_2013 KJ013563_SPLCGoV_CN_2012 KJ013583_SPLCGoV_CN_2012
3.	SPLCGxV*	KJ476508_SPLCGxV_CN_2012 KJ476510_SPLCGxV_CN_2011
4.	SPLCHnV	KC907406_SPLCHnV_CN_2012 KJ476507_SPLCHnV_CN_2012
5.	SPLCSCV	HQ333144_SPLCSCV_US_2006
6.	SPLCSiV-1	KC488316_SPLCSiV1_CN_2012 KJ476511_SPLCSiV1_CN_2012
7.	SPLCSPV	HQ393477_SPLCSPV_BR_2009 JQ621844_SPLCSPV_ZA_2011
8.	SPLCUV	FR751068_SPLCUV_UG_2008
9.	SPMV	FJ969831_SPMV_BR JQ621843_SPMV_ZA_2011
10.	SPLCV	AB433786_SPLCV_JP_1996 AB433787_SPLCV_JP_1998 AB433788_SPLCV_JP_1998 AF104036_SPLCV_US AJ132548_SPLCV_ES AJ586885_SPLCV_IT

DQ644561\_MerLCV\_PR\*\*  
DQ644562\_SPLCV\_PR  
DQ644563\_SPLCV\_PR  
EF456741\_SPLCV\_ES  
EF456742\_SPLCCaV\_ES\*\*  
EF456743\_SPLCV\_ES  
EF456744\_SPLCV\_ES  
EF456745\_SPLCCaV\_ES\*\*  
EF456746\_SPLCV\_ES  
EU253456\_SPLCV\_CN  
EU267799\_SPLCV\_CN  
EU309693\_SPLCV\_CN  
EU839576\_SPLCV\_ES\_2006  
EU839577\_SPLCV\_ES\_2006  
EU839578\_SPLCV\_ES\_2006  
EU839579\_SPLCV\_ES\_2006  
EU856364\_SPLCV\_ES  
EU856365\_SPLCCaV\_ES\*\*  
EU856366\_SPLCV\_ES  
FJ151200\_SPLCV\_ES\_2006  
FJ176701\_SPLCV\_CN\_2008  
FJ515896\_SPLCV\_CN\_2007  
FJ515897\_SPLCV\_CN\_2007  
FJ515898\_SPLCV\_CN\_2007  
FJ529203\_SPLCCaV\_ES\_2002\*\*  
FJ560719\_SPLCV\_KR\_2003  
FJ969829\_SPLCV\_BR  
FJ969830\_SPLCV\_BR  
FJ969832\_SPLCV\_BR  
FJ969833\_SPLCV\_BR  
FJ969834\_SPLCV\_BR  
FJ969835\_SPLCV\_BR  
FJ969836\_SPLCV\_BR  
FJ969837\_SPLCV\_BR  
FN432356\_SPLCV\_IN\_2008  
FN806776\_SPLCV\_CN\_2008  
HM754634\_SPLCV\_KR  
HM754635\_SPLCV\_KR  
HM754636\_SPLCV\_KR  
HM754637\_SPLCV\_KR  
HM754638\_SPLCV\_KR  
HM754639\_SPLCV\_KR  
HM754640\_SPLCV\_KR  
HM754641\_SPLCV\_KR  
HQ333135\_SPLCV\_US\_2006  
HQ333136\_SPLCV\_US\_2006  
HQ333137\_SPLCV\_US\_2006  
HQ333138\_SPLCV\_US\_2006  
HQ333139\_SPLCV\_US\_2007  
HQ333140\_SPLCV\_US\_2007  
HQ333141\_SPLCV\_US\_2007  
HQ333142\_SPLCV\_US\_2007  
HQ333143\_SPLCV\_US\_2007  
HQ393442\_SPLCV\_BR\_2008  
HQ393443\_SPLCV\_BR\_2008  
HQ393444\_SPLCV\_BR\_2008

---

HQ393445\_SPLCV\_BR\_2008  
HQ393446\_SPLCV\_BR\_2008  
HQ393447\_SPLCV\_BR\_2008  
HQ393448\_SPLCV\_BR\_2008  
HQ393449\_SPLCV\_BR\_2008  
HQ393450\_SPLCV\_BR\_2008  
HQ393451\_SPLCV\_BR\_2008  
HQ393452\_SPLCV\_BR\_2008  
HQ393453\_SPLCV\_BR\_2008  
HQ393454\_SPLCV\_BR\_2008  
HQ393455\_SPLCV\_BR\_2008  
HQ393456\_SPLCV\_BR\_2008  
HQ393457\_SPLCV\_BR\_2008  
HQ393458\_SPLCV\_BR\_2008  
HQ393459\_SPLCV\_BR\_2008  
HQ393460\_SPLCV\_BR\_2008  
HQ393461\_SPLCV\_BR\_2009  
HQ393462\_SPLCV\_BR\_2009  
HQ393463\_SPLCV\_BR\_2009  
HQ393464\_SPLCV\_BR\_2009  
HQ393465\_SPLCV\_BR\_2009  
HQ393466\_SPLCV\_BR\_2009  
HQ393467\_SPLCV\_BR\_2009  
HQ393468\_SPLCV\_BR\_2009  
HQ393469\_SPLCV\_BR\_2009  
HQ393470\_SPLCV\_BR\_2009  
HQ393471\_SPLCV\_BR\_2009  
HQ393472\_SPLCV\_BR\_2009  
HQ393473\_SPLCV\_BR\_2009  
HQ393474\_SPLCV\_BR\_2009  
HQ393475\_SPLCV\_BR\_2009  
HQ393476\_SPLCV\_BR\_2009  
JF768740\_SPLCV\_CN\_2010  
JQ349087\_SPLCV\_AR\_2010  
JX050195\_SPLCV\_IN\_2009  
JX050196\_SPLCV\_IN\_2010  
JX050197\_SPLCV\_IN\_2009  
JX286653\_SPLCV\_CN\_2012  
JX286654\_SPLCV\_CN\_2012  
JX286655\_SPLCV\_CN\_2012  
JX961670\_SPLCV\_KR\_2011  
JX961672\_SPLCV\_KR\_2011  
KF040464\_SPLCV\_CN\_2012  
KF040465\_SPLCV\_CN\_2012  
KF040466\_SPLCV\_CN\_2012  
KF040467\_SPLCShV\_CN\_2012\*\*\*  
KF040468\_SPLCShV\_CN\_2010\*\*\*  
KF697069\_SPLCV\_GR\_2013  
KF697070\_SPLCV\_GR\_2013  
KF697071\_SPLCV\_GR\_2013  
KJ013555\_SPLCV\_CN\_2011  
KJ013556\_SPLCV\_CN\_2011  
KJ013557\_SPLCV\_CN\_2012  
KJ013558\_SPLCV\_CN\_2011  
KJ013559\_SPLCV\_CN\_2012  
KJ013560\_SPLCV\_CN\_2012

---

KJ013561\_SPLCV\_CN\_2012  
KJ013562\_SPLCV\_CN\_2012  
KJ013564\_SPLCV\_CN\_2011  
KJ013565\_SPLCV\_CN\_2012  
KJ013566\_SPLCV\_CN\_2012  
KJ013567\_SPLCV\_CN\_2012  
KJ013568\_SPLCV\_CN\_2012  
KJ013569\_SPLCV\_CN\_2012  
KJ013570\_SPLCV\_CN\_2011  
KJ013571\_SPLCV\_CN\_2012  
KJ013577\_SPLCShV\_CN\_2012\*\*\*  
KJ013578\_SPLCShV\_CN\_2012\*\*\*  
KJ013579\_SPLCShV\_CN\_2010\*\*\*  
KJ013580\_SPLCShV\_CN\_2010\*\*\*  
KJ013581\_SPLCShV\_CN\_2012\*\*\*  
KJ013582\_SPLCV\_CN\_2012

---

\*Proposed novel species.

\*\*Currently accepted as species (Brown et al. 2015).

\*\*\*Species name in GenBank, never accepted by ICTV (King AMQ, Adams MJ, Carstens EB, Lefkowitz 2012).

# Chapter 4, Supplementary Table 1. Sequence of wildtype phiX174 obtained from Dr.

Bentley A. Fane. The G gene ORF is highlighted.

>KF0\_phiX174Fane

```
GAGTTTATCGCTTCCATGACGCAGAAAGTTAACACTTTCGGATATTTCTGATGAGTCGAAAAATTATCTTGATAAAGCA
GGAATTACTACTGCTTGTTTACGAATTAAATCGAAGTGGACTGCTGGCGGAAAAATGAGAAAATTCGACCTATCCTTGCG
CAGCTCGAGAAGCTCTTACTTTGCGACCTTTTCGCCATCAACTAACGATTCTGTCAAAAATGACGCGTTGGATGAGGAG
AAGTGGCTTAATATGCTTGGCAGCTTCGTCAAGGACTGGTTTAGATATGAGTCACATTTTGTTCATGGTAGAGATTCTC
TTGTTGACATTTTAAAAAGCGTGGATTACTATCTGAGTCCGATGCTGTTTAAACCACTAATAGGTAAGAAATCATGAGT
CAAGTTACTGAACAATCCGTACGTTTCCAGACCGCTTTGGCCCTTATTAAGCTCATTACGAGCTTTCGCCGTTTTGGATT
TAACCGAAGATGATTTTCGATTTTCTGACGAGTAACAAGTTTGGATTGCTACTGACCGCTCTCGTGCTCGTCGCTGCGT
TGAGGCTTGGCTTTATGGTACGCTGGACTTTGTGGGATACCCTCGCTTTCTGCTCCTGTTGAGTTTATTGCTGCCGTC
ATTGCTTATTATGTTTCATCCCGTCAACATTCAAACGGCCTGTCTCATCATGGAAGGCGCTGAATTTACGAAAAACATTA
TTAATGGCGTCGAGCGTCCGTTAAAGCCGCTGAATTTGTCGCGTTTACCTTGCCTGTACGCGCAGGAAACACTGACGT
TCTTACTGACGCAGAAAGAAACGTGCGTCAAAAAATACGTGCAGAAAGGAGTGATGTAATGTCTAAAGGTAAAAACGTT
CTGGCGCTCGCCCTGGTCTGTCGCGACCGCTTGGCAGGACTATAAGGCAAGCGTAAAGGCGCTCTTTGGTATGTAGG
TGGTCAACAATTTTAAATGTCAGGGGCTTCGGCCCTTACTTGAGGATAAAATTATGTCTAATATTCAAACTGGCGCCGAG
CGTATGCCGATGACCTTTCCCATCTTGGCTTCTTGTCTGAGTTCGCTGCTTATTACCATTTCAACTACTCCGG
TTATCGCTGGCGACTCCTTCGAGATGGACGCCGTTGGCGCTCTCGCTCTTTCTCCATTGCGCTCGTGGCCTTGCTATTGA
CTCTACTGTAGACATTTTTACTTTTTATGTCCCTCATCGTCACGTTTATGGTGAACAGTGGATTAAAGTTCATGAAGGAT
GGTGTTAATGCCACTCCTCTCCGACTGTTAACACTACTGGTTATATTGACCATGCCGCTTTTCTTGGCAGCATTAACC
TGATACCAATAAAAAATCCCTAAGCATTTGTTTTCAGGGTTATTTGAATATCTATAACAACATATTTTAAAGCGCCGTGGAT
GCCTGACCGTACCGAGGCTAACCCCTAATGAGCTTAATGAAGATGATGCTCGTTATGGTTTCCGTTGCTGCCATCTCAAA
AACATTTGGACTGCTCCGCTTCTCCTGAGACTGAGCTTTCTCGCCAAATGACGACTTCTACCACATCTATTGACATTA
TGGGTCTGCAAGCTGCTTATGCTAATTTGCATACGACCAAGAACGTGATTACTTCATGCAGCGTTACCGTGATGTTAT
TTCTTCATTTGGAGGTAAACCTCTTATGACGCTGACAACCGTCCCTTACTTGTCTATGCGCTCAATTTCTGGGCATCT
GGCTATGATGTTGATGGAACGTGACCAACGTCGTTAGGCCAGTTTTCTGGTCTGTTTCAACAGACCTATAAACATTTCTG
TGCCGCTTTCTTTGTTCTGAGCATGGCACTATGTTTACTCTTTCGCTTGTTCGTTTTCCGCTACTGCGACTAAAGA
GATTCAGTACCTTAACGCTAAAGGTGCTTTGACTTATACCGATATTGCTGGCGACCTGTTTTGTATGGCAACTTGCCG
CCGCTGAAATTTCTATGAAGGATGTTTTCCGTTTCTGGTGATTCTGCTAAGAAGTTTAAAGATTGCTGAGGGTCACTGGT
ATCGTTATGCGCTTCGTATGTTTCTCCTGTTTTATCACCTTCTTGAAGGCTTCCCATTCATTCAGGAACCGCTTCTGG
TGATTTGCAAGAACGCGTACTTATTCGCCACCATGATTATGACCAAGTGTTCAGTCCGTTTCAAGTGTGTCAGTGGAA
AGTCAGGTTAAATTTAATGTGACCGTTTATCGCAATCTGCGCACCTTCTGCGATTCAATCATGACTTCGATGATAAAAGA
TTGAGTGTGAGGTTATAACGCCGAAGCGGTAAAAATTTTAAATTTTTCGCGCTGAGGGGTTGACCAAGCGAAGCGCGGTA
GGTTTTCTGCTTAGGAGTTTAAATCATGTTTTCAGACTTTTATTTCTCGCCATAATTCAAACCTTTTTTTCTGATAAGCTGG
TTCTCACTTCTGTTACTCCAGCTTCTTCGGCACCTGTTTTACAGACACCTAAAGCTACATCGTCAACGTTATATTTTGA
TAGTTTGACGGTTAATGCTGGTAATGGTGGTTTTCTTCATTGCATTTCAGATGGATACATCTGTCAACGCCGCTAATCAG
GTTGTTTCTGTTGGTGCTGATATTGCTTTTGATGCCGACCCATAAATTTTTTGCCGTGTTTGGTTCGCTTTGAGTCTTCTT
CGGTTCCGACTCCCTCCGACTGCCATGATGTTTATCCTTTGGATGGTTCGCCATGATGGTGGTTATTATACCGCTCAA
GGACTGTGTGACTATTGACGTCCTTCCCCGTACGCCGGGCAATAATGTTTATGTTGGTTTTCATGGTTTGGTCTAACTTT
ACCGCTACTAAATGCCGCGGATTGGTTTTCGCTGAATCAGGTTATTAAAGAGATTATTTGTCTCCAGCCACTTAAGTGAG
GTGATTTATGTTTGGTGCTATTGCTGGCGGATTTGCTTCTGCTCTTGGTGGTGGCCATGTCTAAATGTTTGGAGGC
GGTCAAAAAGCCGCTCCGGTGGCATTCAAGGTGATGTGCTTGCTACCGATAACAATACTGTAGGCATGGGTGATGCTG
GTATTAATCTGCCATTCAAGGCTCTAATGTTCCTAACCATGATGAGGCCGCCCTAATTTTGTCTGGTGCTATGGC
TAAAGCTGGTAAAGGACTTCTTGAAGGTACGTTGCAGGCTGCGACCTTCTGCCGTTTCTGATAAGTTGCTTGATTTGGTT
GGACTTGGTGGCAAGTCTGCCGCTGATAAAGGAAAGGATACTCGTGATTATCTTGTCTGCTGCATTTCCTGAGCTTAATG
CTTGGGAGCGTGCTGGTGCTGATGCTTCTCTGCTGGTATGGTTGACGCCGATTGAGAATCAAAAAGAGCTTACTAA
AATGCAACTGGACAATCAGAAAGAGATTGCCGAGATGCAAAATGAGACTCAAAAAGAGATTGCTGGCATTACGTCGGCG
ACTTCACGCCAGAATACGAAAGACCAGGTATATGCACAAAATGAGATGCTTGCTTATCAACAGAAAGGAGTCTACTGCTC
CGTTTGGCTCTATTATGGAACACCAATCTTTCCAAGCAACAGCAGGTTTCCGAGATTATGCGCCAAATGCTTACTCA
AGCTCAAACCGCTGGTCAAGTATTTTACCAATGACCAAAATCAAAGAAATGACTCGCAAGGTTAGTGCTGAGGTTGACTTA
GTTTCATCAGCAACCGCAGAAATCAGCGGTATGGCTCTTCTCATATTGGCGCTACTGCAAGGATATTTCTAATGTCGTCA
CTGATGCTGCTTCTGGTGTTGATATTTTTTTCATGGTATTGATAAAGCTGTTGCCGATACTTGAACAATTTCTGGAA
AGACGGTAAAGCTGATGGTATTGGCTCTAATTTGCTAGGAAATAACCGTCCGATGACACCTCCCAATTGTATGTT
TTCATGCCTCCAAATCTTGGAGGCTTTTTTATGGTTTCGTTCTTATTACCCTTCTGAATGTCACGCTGATTATTTGACT
TTGAGCGTATCGAGGCTCTTAAACCTGCTATTGAGGCTGTGGCATTTCTACTCTTTCTCAATCCCAATGCTTGGCTT
CCATAAGCAGATGGATAACCGCATCAAGCTCTTGAAGAGATTCTGTCTTTTCTGATGCAGGGCGTTGAGTTCGATAAT
GGTGATATGATGTTGACGGCCATAAGGCTGCTTCTGACGTTTCGTGATGAGTTTGTATCTGTTACTGAGAAGTTAATGG
ATGAATTGGCACAATGCTACAATGTGCTCCCCAACCTTGATATTAATAACACTATAGACCACCGCCCGAAGGGGACGA
AAAATGGTTTTTTAGAGAACGAGAAGACGGTTACGCGAGTTTTGCCGCAAGCTGGCTGCTGAACGCCCTCTTAAGGATATT
CGCGATGAGTATAATTACCCCAAAAAGAAAGGTATTAAGGATGAGTGTTCAAGATTGCTGGAGGCTCCACTATGAAAT
CGCGTAGAGGCTTTACTATTACGCGTTTGATGAATGCAATGCGACAGGCTCATGCTGATGGTTGGTTTATCGTTTTTGA
```



CACTCTCACGTTGGCTGACGACCGATTAGAGGCGTTTTATGATAATCCCAATGCTTTGCGTGACTATTTTCGTGATATT  
GGTCGTATGGTTCTTGCTGCCGAGGGTCGCAAGGCTAATGATTCACACGCCGACTGCTATCAGTATTTTTGTGTGCCTG  
AGTATGGTACAGCTAATGGCCGTCTTCATTTCCATGCGGTGCACCTTTATGCGGACACTTCCTACAGGTAGCGTTGACCC  
TAATTTTGGTCGTCGGGTACGCAATCGCCGCCAGTTAAATAGCTTGCAAAATACGTGGCCTTATGGTTACAGTATGCCC  
ATCGCAGTTCGCTACACGCAGGACGCTTTTTTCACGTTCTGGTTGGTTGTGGCCTGTTGATGCTAAAGGTGAGCCGCTTA  
AAGCTACCAGTTATATGGCTGTTGGTTTCTATGTGGCTAAATACGTTAACAAAAAGTCAGATATGGACCTTGCTGCTAA  
AGGTCTAGGAGCTAAAGAATGGAACAACCTCACTAAAAACCAAGCTGTCGCTACTTCCCAAGAAGCTGTTCAGAATCAGA  
ATGAGCCGCAACTTCGGGATGAAAATGCTCACAATGACAAATCTGTCCACGGAGTGCTTAATCCAACCTACCAAGCTGG  
GTTACGACGCGACGCCGTTCAACCAGATATTGAAGCAGAACGCAAAAAGAGAGATGAGATTGAGGCTGGGAAAAGTTAC  
TGTAGCCGACGTTTTTGGCGGCGCAACCTGTGACGACAAATCTGCTCAAATTTATGCGCGCTTCGATAAAAATGATTGGC  
GTATCCAACCTGCA

## BIBLIOGRAPHY

- Abascal F, Zardoya R, Posada D. 2005. ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics* 21:2104–2105.
- Acland A, Agarwala R, Barrett T, Beck J, Benson D a., Bollin C, Bolton E, Bryant SH, Canese K, Church DM, et al. 2014. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 42:7–17.
- Adams MJ, Antoniw JF, Fauquet CM. 2005. Molecular criteria for genus and species discrimination within the family Potyviridae. *Arch. Virol.* 150:459–479.
- Akad F, Jacobi J, Polston J. 2007. Identification of Tomato yellow leaf curl virus and Tomato mottle virus in two counties in Alabama. *Plant Dis.* 91:906.
- Almeida RPP, Bennett GM, Anhalt MD, Tsai CW, O’Grady P. 2009. Spread of an introduced vector-borne banana virus in Hawaii. *Mol. Ecol.* 18:136–146.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.
- Amin I, Qazi J, Mansoor S, Ilyas M, Briddon RW. 2008. Molecular characterisation of Banana bunchy top virus (BBTV) from Pakistan. *Virus Genes* 36:191–198.
- Baltimore D. 1971. Expression of animal virus genomes. *Bacteriol. Rev.* 35:235–241.
- Bao Y, Chetvernin V, Tatusova T. 2014. Improvements to pairwise sequence comparison (PASC): a genome-based web tool for virus classification. *Arch. Virol.*:3293–3304.
- Bao Y, Federhen S, Leipe D, Pham V, Resenchuk S, Rozanov M, Tatusov R, Tatusova T. 2004. National center for biotechnology information viral genomes project. *J. Virol.* [Internet] 78:7291–7298. Available from: <http://jvi.asm.org/content/78/14/7291.short>
- Barbosa J, Teixeira A, Moreira A, Camargo L, Bergamin Filho A, Kitajima E, Rezende J. 2008. First report of Tomato chlorosis virus infecting tomato crops in Brazil. *Plant Dis.* 92:1709.

- Beletskii A, Grigoriev A, Joyce S, Bhagwat AS. 2000. Mutations induced by bacteriophage T7 RNA polymerase and their effects on the composition of the T7 genome. *J. Mol. Biol.* 300:1057–1065.
- Benevides JM, Stow PL, Ilag LL, Incardona NL, Thomas GJ. 1991. Differences in secondary structure between packaged and unpackaged single-stranded DNA of bacteriophage phi X174 determined by Raman spectroscopy: a model for phi X174 DNA packaging. *Biochemistry* 30:4855–4863.
- Berkhout B, van Hemert FJ. 1994. The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins. *Nucleic Acids Res.* 22:1705–1711.
- Bernardo P, Golden M, Akram M, Naimuddin, Nadarajan N, Fernandez E, Granier M, Rebelo AG, Peterschmitt M, Martin DP, et al. 2013. Identification and characterisation of a highly divergent geminivirus: Evolutionary and taxonomic implications. *Virus Res.* 177:35–45.
- Blaisdell B. 1985. A method of estimating from two aligned present-day DNA sequences their ancestral composition and subsequent rates of substitution, possibly different in the two lineages, corrected for multiple and parallel substitutions at the same site. *J. Mol. Evol.* 22:69–81.
- Bloom JD. 2014. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Mol. Biol. Evol.* 31:1956–1978.
- Boni MF, Posada D, Feldman MW. 2007. An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* 176:1035–1047.
- Boussau B, Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. *Syst. Biol.* 55:756–768.
- Brown J, Idris A. 2006. Introduction of the exotic monopartite Tomato yellow leaf curl virus into west coast Mexico. *Plant Dis.* 90:1360.
- Brown JK, Zerbini FM, Navas-Castillo J, Moriones E, Ramos-Sobrinho R, Silva JCF, Fiallo-Olivé E, Briddon RW, Hernández-Zepeda C, Idris A, et al. 2015. Revision of Begomovirus taxonomy based on pairwise sequence comparisons. *Arch. Virol.* [Internet]. Available from: <http://link.springer.com/10.1007/s00705-015-2398-y>
- Burnham KP, Anderson DR. 2004. Multimodel inference - understanding AIC and BIC in Model Selection. *Sociol. Methods Res.* 33:261–304.

- Van Den Bussche RA, Baker RJ, Huelsenbeck JP, Hillis DM. 1998. Base compositional bias and phylogenetic analyses: a test of the “flying DNA” hypothesis. *Mol. Phylogenet. Evol.* 10:408–416.
- Cardinale DJ, Duffy S. 2011. Single-stranded genomic architecture constrains optimal codon usage. *Bacteriophage* 1:219–224.
- Carpi G, Holmes EC, Kitchen A. 2010. The evolutionary dynamics of bluetongue virus. *J. Mol. Evol.* 70:583–592.
- Chare E, Holmes E. 2006. A phylogenetic survey of recombination frequency in plant RNA viruses. *Arch. Virol.* 151:933–946.
- Chen CL, Duquenne L, Audit B, Guilbaud G, Rappailles A, Baker A, Huvet M, D’Aubenton-Carafa Y, Hyrien O, Arneodo A, et al. 2011. Replication-associated mutational asymmetry in the human genome. *Mol. Biol. Evol.* 28:2327–2337.
- Codoñer FM, Elena SF. 2008. The promiscuous evolutionary history of the family Bromoviridae. *J. Gen. Virol.* 89:1739–1747.
- Crisp MD CG. 1996. Paraphyletic species. In: *Telopea*. p. 813–844. Available from: [http://biology-assets.anu.edu.au/hosted\\_sites/Crisp/pdfs/Crisp1996\\_paraspecies.pdf](http://biology-assets.anu.edu.au/hosted_sites/Crisp/pdfs/Crisp1996_paraspecies.pdf)
- Dang CC, Le QS, Gascuel O, Le VS. 2010. FLU, an amino acid substitution model for influenza proteins. *BMC Evol. Biol.* 10:99.
- Dimmic MW, Rest JS, Mindell DP, Goldstein RA. 2002. rtREV: An amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *J. Mol. Evol.* 55:65–73.
- Duchene S, Ho S, Holmes E. 2015. Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models. *BMC Evol. Biol.* 15:36.
- Duffy S, Holmes EC. 2007. Multiple introductions of the old world begomovirus Tomato yellow leaf curl virus into the new world. *Appl. Environ. Microbiol.* 73:7114–7117.
- Duffy S, Holmes EC. 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 EC. 2009. Validation of high rates of nucleotide substitution in geminiviruses: Phylogenetic evidence from East African cassava mosaic viruses. *J. Gen. Virol.* 90:1539–1547.
- Duffy S, Shackelton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* 9:267–276.
- Edwards RA, Rohwer F. 2005. Opinion: Viral metagenomics. *Nat. Rev. Microbiol.* 3:504–510.
- Fane BA, Hayashi M. 1991. Second-site suppressors of a cold-sensitive prohead accessory protein of bacteriophage  $\phi$ X174. *Genetics* 128:663–671.
- Fane BA, Head S, Hayashi M. 1992. Functional relationship between the J proteins of bacteriophages  $\phi$ X174 and G4 during phage morphogenesis. *J. Bacteriol.* 174:2717–2719.
- Fauquet CM, Briddon RW, Brown JK, Moriones E, Stanley J, Zerbini M, Zhou X. 2008. Geminivirus strain demarcation and nomenclature.
- Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA. 2005. Virus Taxonomy: VIIIth Report of the International Committee on Taxonomy of Viruses. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0168170201003525>
- Fehrholz M, Kendl S, Prifert C, Weissbrich B, Lemon K, Rennick L, Duprex PW, Rima BK, Koning FA, Holmes RK, et al. 2012. The innate antiviral factor APOBEC3G targets replication of measles, mumps and respiratory syncytial viruses. *J. Gen. Virol.* 93:565–576.
- Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17:368–376.
- Francino MP, Ochman H. 1997. Strand asymmetries in DNA evolution. *Trends Genet.* 13:240–245.
- Frank AC, Lobry JR. 1999. Asymmetric substitution patterns: A review of possible underlying mutational or selective mechanisms. *Gene* 238:65–77.
- Frederico LA, Kunkel TA, Shaw BR. 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry* 29:2532–2537.

- Galtier N, Gouy M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* 15:871–879.
- Gamez-Jimenez C, Romero-Romero J, Santos-Cervantes M, Leyva-Lopez N, Mendez-Lozano J. 2009. Tomatillo (*Physalis ixocarpa*) as a natural new host for Tomato yellow leaf curl virus in Sinaloa, Mexico. *Plant Dis.* 93:545.
- 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.
- Gibbs MJ, Armstrong JS, Gibbs AJ. 2000. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* 16:573–582.
- Gillam S, Atkinson T, Markham a, Smith M. 1985. Gene K of bacteriophage phi X174 codes for a protein which affects the burst size of phage production. *J. Virol.* 53:708–709.
- Gojobori T, Ishii K, Nei M. 1982. Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. *J. Mol. Evol.* 18:414–423.
- Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, Lopez R. 2010. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.* 38.
- Green P, Ewing B, Miller W, Thomas PJ, Green ED. 2003. Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.* 33:514–517.
- Hall B. 2007. *Phylogenetic trees made easy: a how-to manual*. Sunderland, MA: Sinauer Associates
- Hamel AL, Lin LL, Nayar GP. 1998. Nucleotide sequence of porcine circovirus associated with postweaning multisystemic wasting syndrome in pigs. *J. Virol.* 72:5262–5267.
- Harkins GW, Delport W, Duffy S, Wood N, Monjane AL, Owor BE, Donaldson L, Saumtally S, Triton G, Briddon RW, et al. 2009. Experimental evidence indicating that mastreviruses probably did not co-diverge with their hosts. *Virol. J.* 6:104.
- Hasegawa M, Kishino H, Yano T. 1985. Dating the human-ape split by a molecular clock of mitochondrial DNA. *Evolution (N. Y.)*. 22:160–174.

- Hedtke SM, Townsend TM, Hillis DM. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst. Biol.* 55:522–529.
- Ho ES, 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* [Internet] 9:e111957. Available from: <http://dx.plos.org/10.1371/journal.pone.0111957>
- Hu J-M, Fu H-C, Lin C-H, Su H-J, Yeh H-H. 2007. Reassortment and concerted evolution in banana bunchy top virus genomes. *J. Virol.* 81:1746–1761.
- Huelsenbeck JP, Bollback JP, Levine AM. 2002. Inferring the root of a phylogenetic tree. *Syst. Biol.* 51:32–43.
- Hurvich C, Tsai C. 1989. Regression and time series model selection in small samples. *Biometrika* [Internet] 76:297–307. Available from: <http://biomet.oxfordjournals.org/content/76/2/297.short>
- Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23:254–267.
- Idris A, Guerrero J, Brown J. 2007. Two distinct isolates of Tomato yellow leaf curl virus threaten tomato production in Arizona and Sonora, Mexico. *Plant Dis.* 91:910.
- Irwin DE, Bensch S, Irwin JH, Price TD. 2005. Speciation by distance in a ring species. *Science* 307:414–416.
- Isakeit T, Idris A, Sunter G, Black M, Brown J. 2007. Tomato yellow leaf curl virus in tomato in Texas, originating from transplant facilities. *Plant Dis.* 91:466.
- Iverson E, Stedman K. 2012. A genetic study of SSV1, the prototypical fusellovirus. *Front. Microbiol.* 3.
- Jenkins GM, Holmes EC. 2003. The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Res.* 92:1–7.
- Jones RAC. 2009. Plant virus emergence and evolution: Origins, new encounter scenarios, factors driving emergence, effects of changing world conditions, and prospects for control. *Virus Res.* 141:113–130.
- Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: *In Mammalian protein metabolism*, Vol. III (1969), pp. 21-132. Vol. III. p. 21–132. Available from: <http://www.citeulike.org/group/1390/article/768582>

- Jun T, Zhi-Xin L. 2005. Cloning and sequencing of DNA components of Banana bunchy top virus Hainan isolate. *China J. Agric. Biotechnol.* 2:91–97.
- Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16:111–120.
- King AMQ, Adams MJ, Carstens EB, Lefkowitz E ed. 2012. *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*. 9th ed. Waltham, MA: Elsevier Academic Press
- Kosakovsky Pond SL, Frost SDW, Muse S V. 2005. HyPhy: Hypothesis testing using phylogenies. *Bioinformatics* 21:676–679.
- Kroneman A, Vega E, Vennema H, Vinjé J, White P a., Hansman G, Green K, Martella V, Katayama K, Koopmans M. 2013. Proposal for a unified norovirus nomenclature and genotyping. *Arch. Virol.* 158:2059–2068.
- Laidler KJ. 1984. The development of the Arrhenius equation. *J. Chem. Educ.* [Internet] 61:494–498. Available from: <http://pubs.acs.org/doi/abs/10.1021/ed061p494>
- Larkin MA, Blackshields G, Brown NP, Chenna R, Mcgettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947–2948.
- Lauber C, Gorbalenya a. E. 2012. Toward Genetics-Based Virus Taxonomy: Comparative Analysis of a Genetics-Based Classification and the Taxonomy of Picornaviruses. *J. Virol.* 86:3905–3915.
- Lecoq H, Wipf-Scheibel C, Chandeysson C, Lê Van A, Fabre F, Desbiez C. 2009. Molecular epidemiology of Zucchini yellow mosaic virus in France: An historical overview. *Virus Res.* 141:190–200.
- Lefeuve P, Martin DP, Hoareau M, Naze F, Delatte H, Thierry M, Varsani a., Becker N, Reynaud B, Lett JM. 2007. Begomovirus “melting pot” in the south-west Indian Ocean islands: Molecular diversity and evolution through recombination. *J. Gen. Virol.* 88:3458–3468.
- Lefeuve P, Moriones E. 2015. Recombination as a motor of host switches and virus emergence: geminiviruses as case studies. *Curr. Opin. Virol.* [Internet] 10:14–19. Available from: <http://www.sciencedirect.com/science/article/pii/S1879625714002375>



- Legg JP, Owor B, Sseruwagi P, Ndunguru J. 2006. Cassava Mosaic Virus Disease in East and Central Africa: Epidemiology and Management of A Regional Pandemic. *Adv. Virus Res.* 67:355–418.
- Lind PA, Andersson DI. 2008. Whole-genome mutational biases in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* 105:17878–17883.
- Lindahl T, Nyberg B. 1974. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry* 13:3405–3410.
- Liò P, Goldman N. 1998. Models of molecular evolution and phylogeny. *Genome Res.* [Internet] 8:1233–1244. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/9872979>
- Lobry JR. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* 13:660–665.
- Lopez P, Philippe H. 2001. Composition strand asymmetries in prokaryotic genomes: Mutational bias and biased gene orientation. *Comptes Rendus l'Academie des Sci. - Ser. III* 324:201–208.
- Lozano G, Trenado HP, Valverde R a., Navas-Castillo J. 2009. Novel begomovirus species of recombinant nature in sweet potato (*Ipomoea batatas*) and *Ipomoea indica*: Taxonomic and phylogenetic implications. *J. Gen. Virol.* 90:2550–2562.
- Martin D, Rybicki E. 2000. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* 16:562–563.
- Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefevre P. 2010. RDP3: A flexible and fast computer program for analyzing recombination. *Bioinformatics* 26:2462–2463.
- Martin DP, Williamson C, Posada D. 2005. RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* 21:260–262.
- Martín S, Sambade A, Rubio L, Vives MC, Moya P, Guerri J, Elena SF, Moreno P. 2009. Contribution of recombination and selection to molecular evolution of Citrus tristeza virus. *J. Gen. Virol.* 90:1527–1538.
- Melgarejo T a, Kon T, Rojas MR, Paz-Carrasco L, Zerbini FM, Gilbertson RL. 2013. Characterization of a new world monopartite begomovirus causing leaf curl disease of tomato in ecuador and peru reveals a new direction in geminivirus evolution. *J. Virol.* [Internet] 87:5397–5413. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23468482>

- Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: 2010 Gateway Computing Environments Workshop, GCE 2010.
- Morgan GJ, Pitts WB. 2008. Evolution without species: The case of mosaic bacteriophages. *Br. J. Philos. Sci.* 59:745–765.
- Moury B, Desbiez C, Jacquemond M, Lecoq H. 2006. Genetic Diversity of Plant Virus Populations: Towards Hypothesis Testing in Molecular Epidemiology. *Adv. Virus Res.* 67:49–87.
- Mrázek J, Karlin S. 1998. Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl. Acad. Sci. U. S. A.* 95:3720–3725.
- Muhire B, Martin DP, Brown JK, Navas-Castillo J, Moriones E, Zerbini FM, Rivera-Bustamante R, Malathi VG, Briddon RW, 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.
- Muhire BM, Golden M, Murrell B, Lefeuvre P, Lett J-M, Gray A, Poon AYW, Ngandu NK, Semegni Y, Tanov EP, et al. 2014. Evidence of pervasive biologically functional secondary structures within the genomes of eukaryotic single-stranded DNA viruses. *J. Virol.* [Internet] 88:1972–1989. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24284329>
- Ndunguru J, Legg JP, Aveling TAS, Thompson G, Fauquet CM. 2005. Molecular biodiversity of cassava begomoviruses in Tanzania: evolution of cassava geminiviruses in Africa and evidence for East Africa being a center of diversity of cassava geminiviruses. *Virol. J.* 2:21.
- Nischwitz C, Pappu HR, Mullis SW, Sparks AN, Langston DR, Csinos AS, Gitaitis RD. 2007. Phylogenetic analysis of Iris yellow spot virus isolates from onion (*Allium cepa*) in Georgia (USA) and Peru. *J. Phytopathol.* 155:531–535.
- Ortmann AC, Suttle CA. 2005. High abundances of viruses in a deep-sea hydrothermal vent system indicates viral mediated microbial mortality. *Deep. Res. Part I Oceanogr. Res. Pap.* 52:1515–1527.
- Padidam M, Sawyer S, Fauquet CM. 1999. Possible emergence of new geminiviruses by frequent recombination. *Virology* 265:218–225.
- Patil BL, Fauquet CM. 2009. Cassava mosaic geminiviruses: Actual knowledge and perspectives. *Mol. Plant Pathol.* 10:685–701.

- Pereira-Gómez M, Sanjuán R. 2014. Delayed lysis confers resistance to the nucleoside analogue 5-Fluorouracil and alleviates mutation accumulation in the single-stranded DNA bacteriophage  $\phi$ 174. *J. Virol.* [Internet] 88:5042–5049. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24554658>
- Philippe N, Legendre M, Doutre G, Couté Y, Poirot O, Lescot M, Arslan D, Seltzer V, Bertaux L, Bruley C, et al. 2013. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* [Internet] 341:281–286. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23869018>
- Posada D, Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst. Biol.* 53:793–808.
- Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14:817–818.
- Posada D, Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl. Acad. Sci. U. S. A.* 98:13757–13762.
- Posada D. 2008. jModelTest: Phylogenetic model averaging. *Mol. Biol. Evol.* 25:1253–1256.
- Pospieszny H, Hasiow B, Borodynko N. 2007. A Polish isolate of Zucchini yellow mosaic virus from zucchini is distinct from other European isolates. *Plant Dis.* 91:639.
- Prasanna HC, Sinha DP, Verma A, Singh M, Singh B, Rai M, Martin DP. 2010. The population genomics of begomoviruses: global scale population structure and gene flow. *Virol. J.* 7:220.
- Rabadan R, Levine AJ, Robins H. 2006. Comparison of avian and human influenza A viruses reveals a mutational bias on the viral genomes. *J. Virol.* 80:11887–11891.
- Van Regenmortel M. 1989. Applying the species concept to plant viruses. *Arch. Virol.* 104:1–17.
- Van Regenmortel MH V. 2003. Viruses are real, virus species are man-made, taxonomic constructions. *Arch. Virol.* 148:2481–2488.
- Van Regenmortel MH V. 2007. Virus species and virus identification: Past and current controversies. *Infect. Genet. Evol.* 7:133–144.

- Reyes A, Gissi C, Pesole G, Saccone C. 1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol. Biol. Evol.* 15:957–966.
- Roberts JD, Kunkel TA. 1988. Fidelity of a human cell DNA replication complex. *Proc. Natl. Acad. Sci. U. S. A.* 85:7064–7068.
- Rodoni B. 2009. The role of plant biosecurity in preventing and controlling emerging plant virus disease epidemics. *Virus Res.* 141:150–157.
- Rojas M, Kon T, Natwick E, Polston J, Akad F, Gilbertson R. 2007. First report of Tomato yellow leaf curl virus associated with tomato yellow leaf curl disease in California. *Plant Dis.* 91:1056.
- Rokyta DR, Joyce P, Caudle SB, Wichman HA. 2005. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. *Nat. Genet.* 37:441–444.
- Rosario K, Seah Y, Marr C, Varsani A, Moriones E, Polston J, Duffy S, Breitbart M. 2015. Begomovirus and associated satellite DNA molecule diversity captured through vector-enabled metagenomic (VEM) surveys using whiteflies (Aleyrodidae). *J. Virol.* Submitted.
- Rossello-Mora R, Amann R. 2001. The species concept for procaryotes. *FEMS Microbiol. Ecol.* 25:39–67.
- Rzhetsky A, Nei M. 1995. Tests of applicability of several substitution models for DNA sequence data. *Mol. Biol. Evol.* 12:131–151.
- De Sa P, Seebold K, Vincelli P. 2008. First report of Tomato yellow leaf curl virus in greenhouse tomatoes in Kentucky. *Plant Heal. Prog.*
- Salminen MO, Carr JK, Burke DS MF. 1995. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS Res. Hum. Retroviruses* 11:1423–1425.
- Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. 2010. Viral mutation rates. *J. Virol.* 84:9733–9748.
- Sawyer S. 1989. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* 6:526–538.
- Schaaper R. 1993. Base selection, proofreading, and mismatch repair during DNA replication in *Escherichia coli*. *J. Biol. Chem.* [Internet] 268:23762–23765. Available from: <http://www.jbc.org/content/268/32/23762.short>

- Schneider WL, Sherman DJ, Stone AL, Damsteegt VD, Frederick RD. 2004. Specific detection and quantification of Plum pox virus by real-time fluorescent reverse transcription-PCR. *J. Virol. Methods* 120:97–105.
- Senavirathne G, Jaszczur M, Auerbach PA, Upton TG, Chelicos L, Goodman MF, Rueda D. 2012. Single-stranded DNA scanning and deamination by APOBEC3G cytidine deaminase at single molecule resolution. *J. Biol. Chem.* 287:15826–15835.
- Shackelton LA, Parrish CR, Holmes EC. 2006. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. *J. Mol. Evol.* 62:551–563.
- Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7.
- Simmonds P. 2015. Methods for virus classification and the challenge of incorporating metagenomic sequence data. *J. Gen. Virol.* [Internet] 96:1193–1206. Available from: <http://jgv.sgmjournals.org/content/journal/jgv/10.1099/vir.0.000016>
- Smith JM. 1992. Analyzing the mosaic structure of genes. *J. Mol. Evol.* 34:126–129.
- Stamatakis A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
- Sullivan J, Joyce P. 2005. MODEL SELECTION IN PHYLOGENETICS. *Annu. Rev. Ecol. Evol. Syst.* 36:445–466.
- Sumner JG, Jarvis PD, Fernández-Sánchez J, Kaine BT, Woodhams MD, Holland BR. 2012. Is the general time-reversible model bad for molecular phylogenetics? *Syst. Biol.* 61:1069–1074.
- Swofford D. 2003. PAUP\* Phylogenetic analysis using parsimony (\*and other methods).
- Takahata N, Kimura M. 1981. A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. *Genetics* 98:641–657.

- Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* 24:1596–1599.
- Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10:512–526.
- Tanaka M, Ozawa T. 1994. Strand asymmetry in human mitochondrial DNA mutations. *Genomics* 22:327–335.
- Tavare S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. In: American Mathematical Society: Lectures on Mathematics in the Life Sciences. Vol. 17. p. 57–86. Available from: citeulike-article-id:4801403\nhttp://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&\npath=ASIN/0821811673
- Tsuge M, Noguchi C, Akiyama R, Matsushita M, Kunihiro K, Tanaka S, Abe H, Mitsui F, Kitamura S, Hatakeyama T, et al. 2010. G to A hypermutation of TT virus. *Virus Res.* 149:211–216.
- Urbino C, Dalmon A. 2007. Occurrence of Tomato yellow leaf curl virus in tomato in Martinique, Lesser Antilles. *Plant Dis.* 91:1058.
- Varsani A, Monjane AL, Donaldson L, Oluwafemi S, Zinga I, Komba EK, Plakoutene D, Mandakombo N, Mboukoulida J, Semballa S, et al. 2009. Comparative analysis of Panicum streak virus and Maize streak virus diversity, recombination patterns and phylogeography. *Viol. J.* 6:194.
- Vartanian J-P, Guétard D, Henry M, Wain-Hobson S. 2008. Evidence for editing of human papillomavirus DNA by APOBEC3 in benign and precancerous lesions. *Science* 320:230–233.
- Vartanian JP, Henry M, Marchio A, Suspène R, Aynaud MM, Guétard D, Cervantes-Gonzalez M, Battiston C, Mazzaferro V, Pineau P, et al. 2010. Massive APOBEC3 editing of hepatitis B viral DNA in cirrhosis. *PLoS Pathog.* 6:1–9.
- Van der Walt E, Martin DP, Varsani A, Polston JE, Rybicki EP. 2008. Experimental observations of rapid Maize streak virus evolution reveal a strand-specific nucleotide substitution bias. *Viol. J.* 5:104.

- Xia XH LP. 2009. Assessing substitution saturation with DAMBE. In: Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. 2nd ed. p. 615–630.
- Yang Z. 1994. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* 39:105–111.
- Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24:1586–1591.
- Zambrano K, Carbaloo O, Geraud F, Chirinos D, Fernandez C, Marys E. 2007. First report of Tomato yellow leaf curl virus in Venezuela. *Plant Dis.* 91:768.
- Zhang H, Gong H, Zhou X. 2009. Molecular characterization and pathogenicity of tomato yellow leaf curl virus in China. *Virus Genes* 39:249–255.
- Zhang H, Yang B, Pomerantz RJ, Zhang C, Arunachalam SC, Gao L. 2003. The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA. *Nature* 424:94–98.