EVOLUTION OF INDIVIDUAL GENES OF SUGARCANE MOSAIC VIRUS

By

CHRIS NJAGI

A thesis submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Master of science

Graduate Program in Cell and Developmental Biology

Written under the direction of

Dr. Siobain Duffy

And approved by

_____

Dr. Siobain Duffy

_____

Dr. Yana Bromberg

_____

Dr. Elizabeth Snyder

New Brunswick, New Jersey

October 2018

ABSTRACT OF THE THESIS

Evolution of Individual Genes of Sugarcane Mosaic Virus

By Chris Njagi

Thesis Director:
Siobain Duffy, Ph.D.

Sugarcane mosaic virus (SCMV) belongs to the *Potyvirus* genus and *Potyviridae* family of single-stranded RNA viruses. It is a disease of great economic importance, especially in sugarcane and maize in many parts of the world. SCMV has been understudied at both the genomic level with a greater emphasis put on the coat protein gene (CP) for molecular epidemiology. This was a computational study using publicly available sequences (NCBI GenBank) that sought to understand the evolutionary processes that shape SCMV molecular evolution such as phylogeny, recombination, selection pressure and nucleotide diversity by focusing on the individual genes of the polypeptide. Recombination was found to be a major driver of evolution with breakpoints found in P3, HC-Pro, C1, NIa-Pro, NIb, and CP. No statistically significant recombination was detected in the P1, 6K1, 6K2, and VPg genes most probably because of their relatively shorter length. After the removal of recombinants from the initial 82 full polyprotein-coding sequences, 57 sequences were used in phylogenetic analysis using a Bayesian MCMC framework implemented in BEAST2 with particular substitution models for each gene. All the ten genes gave similar low mean nucleotide substitution rates of between 3.35×10^{-3} (C1) to 4.29×10^{-3} (CP) and time to the most recent common ancestor (TMRCA) of 219 years (CP) to 264 years (P1). This would mean that the community's

over-reliance on the CP may produce results accurate for the whole genome, provided the researchers controlled for the effects of recombination, but that sufficient data exists for other genes as well, so there is no rational justification for excluding them from similar studies. The SCMV strains clustered into two distinct groups with sub-clustering well defined by their geographical isolation points, with sequences from Argentina closely grouping with Chinese strains. This observation may not be conclusive as most of the sequences studied were from China. Analysis of dN/dS shows significant negative selection in all the genes with P1 and CP registering relatively lower levels (SLAC: 0.192 for P1 and 0.153 for CP). A few sites in P1, HC-Pro, P3, NIa-VP, NIb, and CP appear to be under diversifying selection. P1 and CP also had the highest nucleotide diversity while the overlapping P3N-PIPO region had the least diversity relative to other regions of the polypeptide. The overlapping region is understandably highly conserved and presumably under strong purifying selection due to its double role in translation. This knowledge from this study enhances the understanding of SCMV evolution, highlights residues in several genes that may be affecting SCMV-crop interactions and will help in developing strategies for the control of the diseases in plants.

DEDICATION

I dedicate this work to my family, my wife Florence and my children Trevor and Anita with much love.

ACKNOWLEDGMENTS

I thank my Academic Advisor Dr. Siobain Duffy for her tireless work in guiding me through to the completion of this research. I also wish to extend special recognition to Natasia Jacko and all the members of Duffy lab including Lashada Williams, Lele Zhao, Alvin Crespo, Steen Hoyer, Erik Lavington and Mansha Seth Pasricha for being there whenever I had a question.

I also acknowledge my first academic adviser and former Program director, Cell and Developmental Biology Dr. Richard Padgett, and Biomedical Sciences administrative assistants Carolyn Ambrose and Diane Murano.

I would also like to appreciate my program sponsor, East African Development Bank (EADB) and program facilitator, Africa-America Institute (AAI) for facilitating my studies at Rutgers University.

I also thank my family for their patience and moral support even as they had to cope without me for the entire period of study.

Finally, it has taken the grace of God to get this far, glory to Him.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Sugarcane mosaic virus (SCMV)

Figure 1 Symptoms of sugarcane mosaic virus disease in a) Sugarcane, b) Maize Images in the public domain courtesy of flickr.com

Sugarcane mosaic virus (SCMV) infects poaceous crops such as maize, sugarcane, and sorghum (H Koike & Gillaspie, 1989; Y. Li, Liu, Zhou, & Fan, 2013). SCMV is a systemic virus affecting most parts of the plant. Its most distinguishing symptom of SCMV infection is the development of leaf mosaicism; patterns of contrasting shades of green, often a background of light or yellow-green chlorotic areas interspersed with normal green (Figure 1). There may also be evidence of reddening and necrosis on leaves and stems (Hideo Koike, 1988). SCMV has also been implicated in the synergistic or additive co-infection with other viruses such as maize dwarf mosaic virus (MDMV) in causing Maize lethal necrosis disease in parts of East Africa (Wangai et al., 2012), or ratoon stunting disease (Koike H, 1974). Diagnosis is done symptomatically or more definitively by use of electron microscopy to identify the flexuous virions with

characteristic cytoplasmic inclusions, use of RT-PCR ELISA based methods (Wu, Zu, Wang, & Chen, 2012). The virus is most effectively controlled using resistant varieties, but the disease persists due to synergistic effects of co-infection and the continual evolution of new strains (Wu et al., 2012).

The origin of the virus is not precisely known but is believed to have had a common point of origin with its principal host in New Guinea (Artschwager & Brandes, 1958). The first reported case of sugarcane mosaic was in 1898 by Van Musschenbroek who gave it the name 'gelestrepenziekte' meaning 'yellow stripe disease'(Artschwager, 1948). In the Americas, sugarcane mosaic disease was first reported in Puerto Rico in 1916 and subsequently in Louisiana and several other southern US states, but the identification of the virus was not until 1963 in Ohio (Wu et al., 2012). SCMV is transmitted by aphids in a non-persistent manner and may also spread through infected setts and mechanical inoculation (Adams, Antoniw, & Fauquet, 2005; Holkar, Kumar, Meena, & Lal, 2017). Today, the virus has been reported in over 25 countries (Figure 2) and is known to cause losses of up to 50% in susceptible cultivars in both maize and sugarcane (Chandran & Gajjeraman, 2015; Luo et al., 2016).

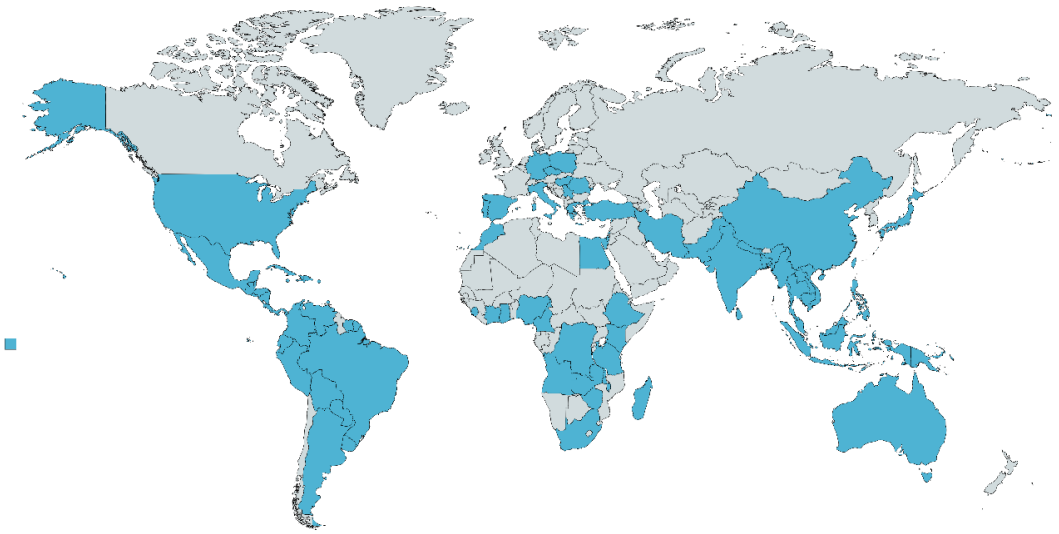

Figure 2 Countries with identified SCMV infection (Created with mapchart.net)

SCMV is a ssRNA positive-strand viruses, in the family *Potyviridae* and genus *Potyvirus* (Wylie et al., 2017). According to the International Committee on Taxonomy of Viruses (ICTV), *Potyviridae* has ten genera with *Potyvirus* being by far the largest with 168 species (King et al., 2018). Species within the *Potyvirus* genus share 50-55% nucleotide identity with a species demarcation criterion of <76% nucleotide identity and <82% amino acid identity (Adams et al., 2005; Wylie et al., 2017). The genus also includes *Maize dwarf mosaic virus*, *Johnson grass mosaic virus*, *Sorghum mosaic virus, Zea mosaic virus*, *Pennisetum mosaic virus*, and *Cocksfoot strike virus*, (Wu et al., 2012). SCMV has a monopartite, 9.6kb linear genome with a poly (A) tract on the 3' terminus and a genome-linked protein (VPg) on the 5' terminus (Wylie et al., 2017). The viral RNA is translated into a single polyprotein which is then cleaved into ten functional proteins; first protein (P1), helper component-protease (HC-Pro), third protein (P3), (6-kDa peptide 1) 6K1, cylindrical inclusion protein (CI), 6-kDa peptide 2 (6K2), small nuclear inclusion protein (NIa) comprised of viral genome-linked protein (VPg) and a

proteinase domain (NIa-Pro), nuclear inclusion protein b (Nib), and Coat protein (CP). An additional protein denoted PIPO (Pretty Interesting Potyviridae ORF), a ~139 amino-acid P3N-PIPO fusion product as a result of ribosomal frameshifting or transcriptional slippage within the N-terminal end of P3 was described in 2008 (Chandran & Gajjeraman, 2015; Chung, Miller, Atkins, & Firth, 2008; D. D. Shukla, Frcnkel, & Ward, 1991); Figure 1).

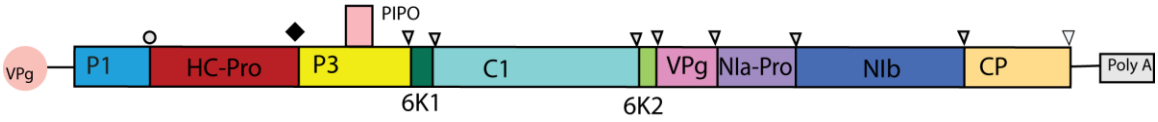

Figure 3 Genome organization of Genus *Potyviridae*.
Cleavage sites of P1-Pro (○), HC-Pro (◆) and NIa-Pro (▼) are indicated. (Wylie et al., 2017)

The virion RNA plays the double role of a genome and viral messenger RNA. As an RNA, it is translated into ten proteins by three self-encoded proteases (P1 protein, HC-Pro, and NIa). P1 and HC-Pro have autocatalytic activity at their C-terminal end that enable their cleavage from the polypeptide while NIa has both cis and trans proteolytic mechanisms that allow the release of other proteins from the peptide at different stages of their lifecycle (Chandrika Ray & Hema, 2016). P1 stimulates genome amplification and enhances virus infection while P3 has been proposed to be involved in viral replication, pathogenicity and symptom development (Revers & Garcia, 2015). CP is involved in RNA encapsidation, viral movement within the host and host specificity while in combination with HC-Pro, it participates in aphid transmissibility (Y. Li et al., 2013; Pirone & Blanc, 1996). HC-Pro is required for the aphid vector transmission of noncirculative viruses such as potyviruses and caulimoviruses. The helper protein mediates the interaction between the viral aspartic acid-alanine-glycine (DAG) motif near

the N-terminal end of CP with the vector stylet (Pirone & Blanc, 1996). HC-Pro also is involved in suppressing gene silencing in vector transmission (Wylie et al., 2017) that may occur through phosphorylation and consequent inactivation of NIa-Pro by the plant cell kinases. VPg is an intrinsically disordered protein with multiple functions due to its diversified protein interactions: VPg interacts with both host and viral proteins in functions such as formation of the viral replication complexes, interaction with the translation eukaryotic initiation factor 4E (eIF4E), modulation of NIa-Pro and suppression of RNA silencing (Jiang & Laliberte, 2011; Wylie et al., 2017). PIPO has been postulated to be involved in a variety of functions including movement, replication, and suppression of systemic silencing (Chung et al., 2008). It has been shown that PIPO functions as a movement protein (MP) in potyviruses (Cheng et al., 2017). CI has both ATPase and RNA helicase activities required for genome replication and together with PIPO aids in virus movement. NIa-Pro is involved in the processing of the viral polyprotein into functional proteins. NIb is the RNA-dependent RNA polymerase (RdRP) responsible for genome replication (Revers & Garcia, 2015).

1.2 Phylogenetic Inference frameworks

Plant viruses exhibit a high potential for genetic variation and diversity that enables them to adapt to different ecological and host internal environments (Y. Li et al., 2013). Genome evolution and phylogenetic studies are primarily purposed to determine genealogical relatedness of organisms, which can be used to estimate the divergence time or Time to the Most Recent Common Ancestor (TMRCA) (S. Y. Li, Pearl, & Doss, 2000). Phylogenetic inference frameworks are broadly grouped into frequentist and Bayesian inferences. The difference between the two approaches is that the former relies

on probability based on actual data or experimental outcomes while the latter relies on both the probability of the data and prior assumptions (Huelsenbeck, Ronquist, Nielsen, & Bollback, 2001). Evolutionary trees centered around frequentist inference employ three approaches; distance-based methods, parsimony, and likelihood approaches (S. Y. Li et al., 2000). They all use an optimality criterion to assess the fitness of given data to a particular hypothesis aiding in the selection of the best tree: maximum likelihood and maximum parsimony are character-based whereas distance-based methods use pairwise distances. Over time, molecular data has come to be primarily analyzed through likelihood analyses (Whelan, Liò, & Goldman, 2001)

Likelihood methods measure the probability of the data given the model assumptions (*e.g.*, how likely one nucleotide is likely to mutate into another) and a phylogenetic tree under consideration. They allow the use of complex evolution models and even allow simultaneous estimation of model parameters (Whelan et al., 2001). They tend to be computationally exerting especially with large sequences because each tree topology is analyzed individually, and not all possible trees can be evaluated once the dataset is even of moderate size (over a few dozen taxa).

In Bayesian statistics, one starts with a model and then derives a probability distribution representing the uncertainty in the parameters of the model (prior) which when combined with the available data gives the posterior distribution from which inferences are made (Paul, 2015; Wang & Yang, 2014). Inference is facilitated by a stochastic process like Markov chain Monte Carlo (MCMC) algorithm which draws samples from the posterior distribution. Bayesian phylogenetics utilizes maximum likelihood frameworks and thus are also a kind of likelihood analysis. “The Bayes theorem is used to combine the

phylogeny (Pr[Tree]) with the likelihood (Pr[Data/Tree]) to produce a posterior probability distribution on trees (Pr[Tree/Data])"(Huelsenbeck et al., 2001).

$$\Pr[\text{Tree} \mid \text{Data}] = \frac{\Pr[\text{Data} \mid \text{Tree}] \times \Pr[\text{Tree}]}{\Pr[\text{Data}]}$$

Bayesian inference has been demonstrated to be more superior than maximum likelihood especially with large sequence datasets (Larget & Simon, 1999; Mar, Harlow, & Ragan, 2005) mainly due to it's ability to accommodate varied sources of data while adjusting for uncertainty (dos Reis, Donoghue, & Yang, 2016).

BEAST, the Bayesian phylogenetic analysis package employed in this study uses the Metropolis-Hastings MCMC algorithm (A. J. Drummond & Rambaut, 2007). MCMC allows inference of posterior distributions under the highly demanding phylogenetic modeling environment provided by Bayesian statistics (Baele et al., 2012). The length of the chain should be sufficiently long for proper mixing allowing for convergence. The initial samples in an MCMC chain referred to as burn-in are discarded to eliminate bias. A Bayesian inference gives a set of topologies having the highest relative frequency, usually within 95% highest probability density (HPD) which are further resolved to provide the tree with the best estimate (Cheon & Liang, 2014).

1.3 Modeling evolution

Evolution models are built on either the empirical or parametric approaches. Empirical approaches involve the use of fixed parameter values calculated from comparisons of observed sequence data while parametric/mechanistic approaches allow the parameter values to be derived from the dataset in each analysis (Whelan et al., 2001). Empirical methods do not consider other factors that may influence the substitution process.

Evolution models used in a BEAST phylogenetic analysis include; site models, clock models and tree prior models. These are discussed in the following pages.

1.3.1 Site model

The site model examines evolutionary rates across sites – either nucleotides or amino acids in an alignment (Uzzell & Corbin, 1971). A site model may incorporate three aspects; the substitution model (how often one character becomes a specific other character), use of gamma rate heterogeneity (Γ, the distribution of how often a site in the dataset experiences mutation, since sites can have multiple substitutions over evolutionary time) and proportion of invariant sites (I, sites that do not experience substitution in the evolutionary time examined) in the analysis (R. R. Bouckaert & Drummond, 2017). The proportion of invariant sites considers sites so constrained by purifying selection that any mutation would be lethal (Whelan et al., 2001; Xia, Xie, Salemi, Chen, & Wang, 2003).

BEAST 2.5.0, which is the version of BEAST used in this analysis, employs 27 nucleotide site models, ranging in complexity. The most basic site model is the Jukes-Cantor 1969 (JC69), which assumes equal base frequencies with all substitutions equally likely (Jukes & Cantor, 1969). It is robust because it has fewer parameters to be estimated and computationally feasible when little change has occurred over the evolutionary times under consideration (Felsenstein, 1981), but it rarely the optimal model for modern datasets. Variations on the JC69 model, such as the Felsenstein 1981 (F81) and Kimura 2-parameter (K80) were developed to incorporate variable base frequencies with all substitutions equally likely (Felsenstein, 1981) or to allow transitions and transversions to occur at different rates, but with equal base frequencies (Kimura, 1980). Parameters were

added to these models creating named variants through the most complex nucleotide substitution model in widespread use, the General Time Reversible (GTR) model. It assumes variable base frequencies and allows every substitution pair to have its own rate (six nucleotide substitution rates in contrast to JC69's one and K80's two, (Lanave, Preparata, Sacone, & Serio, 1984). It is the most widely used substitution model because of its flexible parameterization (Sumner et al., 2012).

1.3.2 Clock model

The molecular clock is used to infer divergence dates in the phylogenetic tree. The strict molecular clock assumes that at the molecular level, the rate of evolution is constant through time and among species (dos Reis et al., 2016). The strict clock is applicable for very closely related sequences or as a default assumption in population statistics. Distantly related sequences are best analyzed using the relaxed molecular clock that assumes different rates of evolution among species (dos Reis et al., 2016). The rate can be correlated or non-correlated between ancestor and descendant branches in relaxed clock models.

1.3.4 Tree prior models

Tree prior models describe the population dynamics within the phylogenetic tree. BEAST implements both birth-death models and coalescent models, which were those used in this study. The coalescent models used include constant population size, exponential growth, logistic growth and Bayesian skyline models, which allow the population size to vary over time and be determined by the analysis.

"The coalescent is the genealogical process of joining lineages when one traces the genealogy of the sample backwards in time" (Hillis, 2015). BEAST uses the sample data

available to run the time machine backward to estimate coalescence time separating two species from their common ancestor. Thus, we aim to use sequentially sampled sequences – sequences from individuals isolated at different points in time from the same population –to generate a coalescent tree that is used to estimate ancestral timelines based on prior information.

1.4 Selection pressure

The high mutation rate in RNA viruses, primarily a result of the error-prone nature of their polymerase, would be expected to result in numerous disadvantageous variants which would impact negatively on their fitness due to negative selection pressure (Y. Li et al., 2013). However, in a large population, some rarer mutations may be beneficial and positive selection may increase their frequency over time, leading to fixation in the population. The ratio of nonsynonymous mutations to the synonymous mutations (dN/dS) is commonly used to determine whether a protein-coding gene in a given population is undergoing positive or diversifying selection (<1) or negative/ purifying selection (>1). It relies on the assumption that nonsynonymous mutations are likely to have an effect on the protein, but synonymous mutations will not – a simplifying assumption that is known not to be entirely correct as they may result in an altered conformation of the protein (Choudhuri, 2014). Regardless, dN/dS remains a frequently used tool in molecular evolution.

1.5 Study Justification

RNA viruses exhibit high mutation rates due to lack of replication fidelity hence tend to evolve fast though this does not mean that viruses with DNA genomes must evolve more slowly or that evolutionary rates are solely a consequence of polymerase fidelity (Duffy,

Shackelton, & Holmes, 2008). Data from CP studies from different ssDNA, RNA viruses and one pararetrovirus that all infect plant hosts carried out in BEAST software show that these viruses evolve at more or less the same high rate (Mbewe, 2017). It has been demonstrated that ssRNA phages exhibit faster evolution rates than ssDNA phages when under high selective pressure but this changes in well-adapted populations where their rates were shown to be comparable to ssDNA viruses (Domingo-Calap & Sanjuan, 2011). The high rate of evolution of ssDNA and RNA viruses enables them to adapt to new environments fast (Duffy et al., 2008). Plant viruses, which have RNA or ssDNA genomes, had long been thought to evolve at a slower rate than animal viruses but emerging evidence shows that they evolve at similar rates as those of animal viruses (A. J. Gibbs, Fargette, Garcia-Arenal, & Gibbs, 2010). Zucchini yellow mosaic virus, a *Potyvirus* has been shown to evolve at a rate of 5.0×10^{-4} substitution per site per year, a rate that is comparable to animal viruses, as do *Tobamoviruses* at 10^{-3} to 10^{-5} subs/site/year and members of family *Luteoviridae* at 10^{-3} to 10^{-4} subs/site/year (Pagan, Firth, & Holmes, 2010; Pagan & Holmes, 2010; Simmons, Holmes, & Stephenson, 2008).

The evolutionary biology of viruses is critical in explaining such phenomena as changes in virulence, disease epidemics, geographic distribution, the emergence of new strains or adaptation to new hosts which ultimately results to better management of economically important viruses such as SCMV (Moradi, Nazifi, & Mehrvar, 2017). Evolutionary changes start with novel genetic variation caused by mutations together with recombination. The fate of this novel variation is also shaped by other factors such as the host/vector genetics and environment. A pathogen is often under positive selection to

increase fitness in a specific host, which usually results in a fitness decrease in other hosts, limiting cross-species jumps (McLeish, Fraile, & Garcia-Arenal, 2018). Viruses overcoming host resistance may not of necessity mean that they are the optimal genotype in all environments. Often, the resistant viral genotypes are less fit in susceptible hosts than the non-resistant viruses (Fraile, Pagan, Anastasio, Saez, & Garcia-Arenal, 2011). Recombination is often a consequence of co-infection and can lead to the emergence of new viral strains which may be more virulent, more infective or more suited to jump hosts. (Faillace, Lorusso, & Duffy, 2017).

SCMV has been understudied at the genomic level with a greater emphasis put on the coat protein gene (CP) (Chen, Chen, & Adams, 2002). CP has for long been studied to show the molecular diversity and the epidemiology of potyviruses (Adams et al., 2005; D. D. Shukla et al., 1992). The heavy focus on the CP could be due to its involvement in virtually all stages in the SCMV lifecycle. Systematic study on SCMV diversity, especially with the use of whole genomes, has been scarce as compared to other potyviruses such as *Turnip mosaic virus, Sweet potato feathery mottle virus, Watermelon mosaic virus,* and *Potato virus Y* (Y. Li et al., 2013). This study was aimed at understanding the evolutionary processes that shape SCMV molecular evolution such as phylogeny, recombination, selection pressure and nucleotide diversity by focusing on the individual genes of the polypeptide. The study also was to investigate whether evolution rates are constant or variable along the entire SCMV polypeptide.

CHAPTER 2: METHODS

2.1 Data set

Eighty-two sugarcane mosaic virus (SCMV) full polyprotein-coding sequences irrespective of their plant host were downloaded from GenBank and saved as a single file in November 2017. Only full genome sequences were considered with the ten viral genes sequences intact. The sequences were from various studies comprising of 70 sequences from China, two from Argentina, one from Estonia, three from Ethiopia, three from Rwanda, two from Iran, one from Ecuador. The sequences were then renamed to include their GenBank accession number, country of origin (ISO two letter country code) and year of isolation (Table 1). We rigorously vetted the backgrounds of each sequence to assure their years of isolation were accurate and did not include extensive passaging while maintained in the lab. Years of isolation ranged from 1998 to 2016 from seven countries.

Table 1: Accession numbers of nucleotide sequences of SCMV polyproteins with reliable dates and country of isolation

1. JX047384_CN_2010	23. JX047398_CN_2010	43. KT895081_IR_2013	63. KR611110_CN_2012
2. JX047399_CN_2010	24. JX047406_CN_2010	44. JX047431_CN_2011	64. KR611111_CN_2012
3. JX047393_CN_2010	25. JX047405_CN_2010	45. JX047420_CN_2011	65. KR611112_CN_2013
4. JX047394_CN_2010	26. JX047403_CN_2010	46. JX047417_CN_2011	66. KR611114_CN_2013
5. JX047395_CN_2010	27. JX047402_CN_2010	47. JX047418_CN_2011	67. KR611113_CN_2013
6. JX047397_CN_2010	28. JX047401_CN_2010	48. JX047419_CN_2011	68. KY006657_EC_2016
7. JX047404_CN_2010	29. JX047400_CN_2010	49. JX047422_CN_2011	69. KP860936_ET_2014
8. JX047408_CN_2010	30. JX047415_CN_2011	50. JX047427_CN_2011	70. KP772216_ET_2014
9. JX047409_CN_2010	31. JX047414_CN_2011	51. JX047428_CN_2011	71. KF744391_RW_2013
10. JX047381_CN_2010	32. JX047413_CN_2010	52. KR108213_CN_2014	72. KP860935_ET_2014
11. JX047410_CN_2010	33. JX047412_CN_2010	53. KT895080_IR_2013	73. AJ297628_CN_2000
12. JX047388_CN_2010	34. JX047411_CN_2010	54. JX237862_AR_2010	74. AJ310105_CN_2000
13. JX047387_CN_2010	35. JX047407_CN_2010	55. KR108212_CN_2014	75. AF494510_CN_2010
14. JX047386_CN_2010	36. JX047429_CN_2011	56. KF744390_RW_2013	76. AM110759_ES_1998
15. JX047385_CN_2010	37. JX047426_CN_2011	57. KF744392_RW_2013	77. JX047421_CN_2011
16. JX047383_CN_2010	38. JX047425_CN_2011	58. KR611105_CN_2012	78. AJ310102_CN_2000
17. JX047382_CN_2010	39. JX047424_CN_2011	59. KR611106_CN_2013	79. JN021933_CN_2009
18. JX047389_CN_2010	40. JX047423_CN_2011	60. KR611107_CN_2012	80. JX237863_AR_2007
19. JX047390_CN_2010	41. JX047416_CN_2011	61. KR611108_CN_2012	81. AJ310104_CN_2000
20. JX047391_CN_2010	42. JX047430_CN_2011	62. KR611109_CN_2013	82. AJ310103_CN_2000
21. JX047392_CN_2010			
22. JX047396_CN_2010			

2.2 Alignment

Pairwise alignment was done computationally using clustalW multiple alignment algorithm in Geneious v11.1 (Kearse et al., 2012) and with some manual editing in SEAL 2.1 (http://compbio.case.edu/seal/), followed by trimming off the untranslated ends from any sequences that had them. The sequences' nucleotide pairwise identity was calculated in Species Demarcation Tool (SDT) v1.2 (Muhire, Varsani, & Martin, 2014) to generate a color-coded pairwise identity matrix. SDT aligns the sequences and calculates the sequence similarity score for each pair and uses a rooted Neighbor-joining phylogenetic

tree (a fast distance-based method) to cluster closely related sequences based on similarity scores. The identity scores were also calculated in Geneious.

2.3 Gene extraction

Gene annotation of the sequences was examined in Geneious and annotations were homologously inferred for the sequences that had not been adequately annotated. This enabled the identification of the polyprotein cleavage sites and the subsequent extraction of the ten genes of the polyprotein in separate alignment files for analysis. The whole polyprotein alignment was not modified other than being cut on the polyprotein sites; the individual gene alignments were not re-aligned. PIPO was not examined because it is an entirely overlapping reading frame with P3.

2.4 Recombination analysis

Presence of recombinant sequences affects the accurate resolution of phylogenetic trees. For phylodynamic analyses, recombination results in an overestimation of the substitution rates, shorter TMRCA, loss of molecular clock and overestimation of nucleotide substitution rates (A. J. Gibbs et al., 2010; Schierup & Hein, 2000).

The BEAST2 software package (R. Bouckaert et al., 2014) used for phylodynamic analyses cannot detect and accommodate recombination within the dataset, and therefore it was necessary to identify and remove recombinants before the commencement of phylogenetic analysis. Recombination was analyzed using seven detection algorithms in RDP 4.95 package (D. P. Martin, Murrell, Golden, Khoosal, & Muhire, 2015). RDP (D. Martin & Rybicki, 2000), GENECONV (Padidam, Sawyer, & Fauquet, 1999), Chimaera (Posada & Crandall, 2001), MaxChi (Maynard Smith, 1992), BootScan (DP Martin, Posada, Crandall, & Williamson, 2005), SiScan (M. J. Gibbs, Armstrong, & Gibbs, 2000)

and 3Seq (Boni, Posada, & Feldman, 2007) recombination detection and analysis methods were used. Recombination events considered significant were those that were detected by at least three of these programs. General settings included a p-value of 0.05 (with a Bonferroni correction for multiple comparisons) and requirement of topological evidence. No reference sequences were used, and the K80 model was applied for the boot scan, DSS, NJ and distance plots. Each gene was analyzed separately. Offspring from the same recombination event sometimes had different major or minor parents called, but the RDP program determined that the phylogenetic relatedness of the parents and constancy of the recombination breakpoints make it highly likely that the recombinants are products of the same event.

2.5 Temporal signal detection

TempEst v1.5.1 (Rambaut, Lam, Max Carvalho, & Pybus, 2016) was used to determine if the sequences show sufficient temporal signal for phylodynamic analysis. Groups of sequences that do not show a clear correlation between genetic divergence and time cannot give reliable estimates on substitution rates, even if they are from the same species and were sampled heterochronously. TempEst software tests for the overall temporal signal through a simple linear regression, which also helps identify incongruent sequences (Rambaut et al., 2016). Geneious was used to generate a nexus file with embedded trees from the aligned sequences as the input file for TempEst.

2.6 Model selection

Statistical selection of best-fit models of nucleotide substitution was done in Jmodeltest2 v2.1.10 (Darriba, Taboada, Doallo, & Posada, 2012; Guindon, Gascuel, & Rannala, 2003) via the CIPRES Science Gateway (Miller, Pfeiffer, & Schwartz, 2010). All the 203

nucleotide substitution schemes were tested, but since BEAST2 only handles 27 models by default, only the best-fitting model implemented in BEAST2 was selected. The best tree topology was chosen via the maximum likelihood search criteria allowing for rate variation among sites (+Γ, +I). Both the clock and tree prior models were then selected through path-sampling in the Path-sampler that is bundled with BEAST2 using the selected substitution model(R. Bouckaert et al., 2014). Clock model selection involved the comparison between a strict clock, the relaxed clock (exponential distribution), and relaxed clock (lognormal distribution) models while the tree prior model selection involved a comparison between the constant population size, the exponentially increasing population size, and the non-parametric Bayesian Skyline models. The *.xml* file was generated in BEAUti2 and then manually edited to run in the Cipres server by replacing the opening run statement with the following statement:

<run spec='beast.inference.PathSampler' chainLength="2000000" alpha='0.3' rootdir='/tmp/' burnInPercentage='50' preBurnin="0" deleteOldLogs='true' nrOfSteps='100'>

cd $(dir)

java -cp $(java.class.path) beast.app.beastapp.BeastMain $(resume/overwrite) -java -seed $(seed) beast.xml

(Adapted from github.com/BEAST2-Dev/BEASTLabs/examples/testPathSampler.xml)

BEAUti2 v2.4.8 was used to generate the *.xml* files for the clock models and v2.5.0 used for the tree prior models, but all ran through the CIPRES Gateway server. The use of the different versions was due to a persistent infinity error with the Coalescent Exponential Population prior in the earlier version. A chain-length of 2,000,000 running for 100 steps was adopted for both clock and tree prior models. The best fitting model was chosen from these results by likelihood score.

2.7 Phylodynamic analysis

BEAST2 analysis was done via the Cipres Gateway server using BEAST2 v2.4.8 with a standard MCMC run of 1 billion and logging after every 100,000 for both the trace and tree files. The *.xml* file was generated in BEAUti2 using the appropriate models as realized in model selection. These settings were sufficient to ensure adequate mixing of the tree for P1-Pro, HC-Pro, CI, VPg, NIa-Pro and CP as evidenced by high effective sample size (ESS) values. Convergence was achieved in the 6K1, 6K2 and NIb genes by increasing the rate of sampling to every 50,000 while in P3, the chain-length was doubled to 2 billion for proper convergence. Tracer was used to examine convergence and ensure the ESS was above 200 for every parameter as recommended (Alexei J. Drummond & Bouckaert, 2015). The consensus trees and Bayesian posterior probability values at nodes were calculated with a 10% burn-in removed from each run. Consistency in the results was ensured by repeating the MCMC runs once using the same parameters. As a further control, the parameters of each analysis were run without the sequence data (sampling from the prior) to ensure that patterns in the data, and not the parameterization, were driving the results. Non-overlapping credibility intervals for the results of interest (TMRCA, substitution rate) were used to assess that sampling from the prior produced different results.

2.8 Selection pressure and diversity analysis

Nonsynonymous (dN) and synonymous (dS) substitution ratios (dN/dS) was calculated using codon-based maximum-likelihood tools; single-likelihood ancestor counting (SLAC), fixed-effects likelihood (FEL), and Mixed Effects Model of Evolution (MEME) in HyPhy package and implemented in the Datamonkey server (http://datamonkey.org/)

(Delport, Poon, Frost, & Kosakovsky Pond, 2010; Pond, Frost, & Muse, 2005). The significance levels were set at P <0.1 for all three tools. MEME identifies sites undergoing episodic positive selection whereas SLAC and FEL infer non-synonymous (dN) and synonymous (dS) substitution rates on a per-site basis assuming that the selection pressure for each site is constant along the entire phylogeny (Delport et al., 2010). The most conservative method, SLAC, employs both maximum likelihood and counting techniques to achieve this while the other two use ML method only. Nucleotide diversity was assessed using DNA Sequence Polymorphism (DnaSP) v6.11.01 (Rozas et al., 2017).

CHAPTER 3: RESULTS

3.1 Alignment

After trimming of the 5' and 3' untranslated regions, full-length polyprotein-coding sequences (nucleotide length of 9189-9234) were aligned, resulting in an overall alignment of 9,265nt. All the gaps were found in CP with multiple insertions and deletions encompassing the first quartile (N-terminal end) of the gene, and sites in the CP region had, by eye, more diversity than other parts of the alignment. The gaps were due to variability in the length (939-984) of CP while the other genes had no variance. The alignment had an overall 41.6% GC content, 85.1% (range of 74-99.9%) pairwise matching of nucleotides, and 54.5% of all sites were invariant. The low full-length nucleotide sequence identity is indicative of high diversity within the species. The pairwise identity scores from SDT ranged from 74% to 99.9% (figure 4). Average pairwise identities for each gene were: P1 78.5 %, HC-Pro 86%, P3 86.7%, 6K1 83.5%, C1 85.7%, 6K2 83.7%, VPg 84.1%, NIa-Pro 85%, NIb 85.4%, and CP 86.2%. After removal of recombinants, the nucleotide pairwise identity scores rose to 86.7% (77-99.9%; figure 5). The CP N-terminal end DAG box (Asp-Ala-Gly), which is important in aphid transmission was found in all isolates (Urcuqui-Inchima, Haenni, & Bernardi, 2001).

The genes had varying lengths ranging from 159nt in 6K2 to 1914nt in C1 (Table 2).

Table 2: Descriptions of SCMV gene position in alignment (5'-3'), with gene function

Gene	Position in alignment	Length	Function
P1	0-699	699	Viral replication
HC-Pro	700-2079	1380	Suppression of gene silencing, vector transmission
P3	2080-3120	1041	Viral replication, host range, and symptom development.
6K1	3121-3321	201	Involved in replication but the specific function is not known
C1	3322-5235	1914	Genome replication- helicase activity
6K2	5236-5394	159	Anchors the replication complex to the ER
VPg	5395-5961	567	Component of viral replication complexes interacts with eIF4E translation initiation factor, suppression of RNA silencing/ modulates NIa-Pro.
NIa-Pro	5962-6687	726	Cleavage of most sites in the polyprotein
NIb	6688-8250	1563	RNA-dependent RNA polymerase
CP	8251-9265	1003	Coat protein, virus movement, genome amplification and vector transmission, host specificity

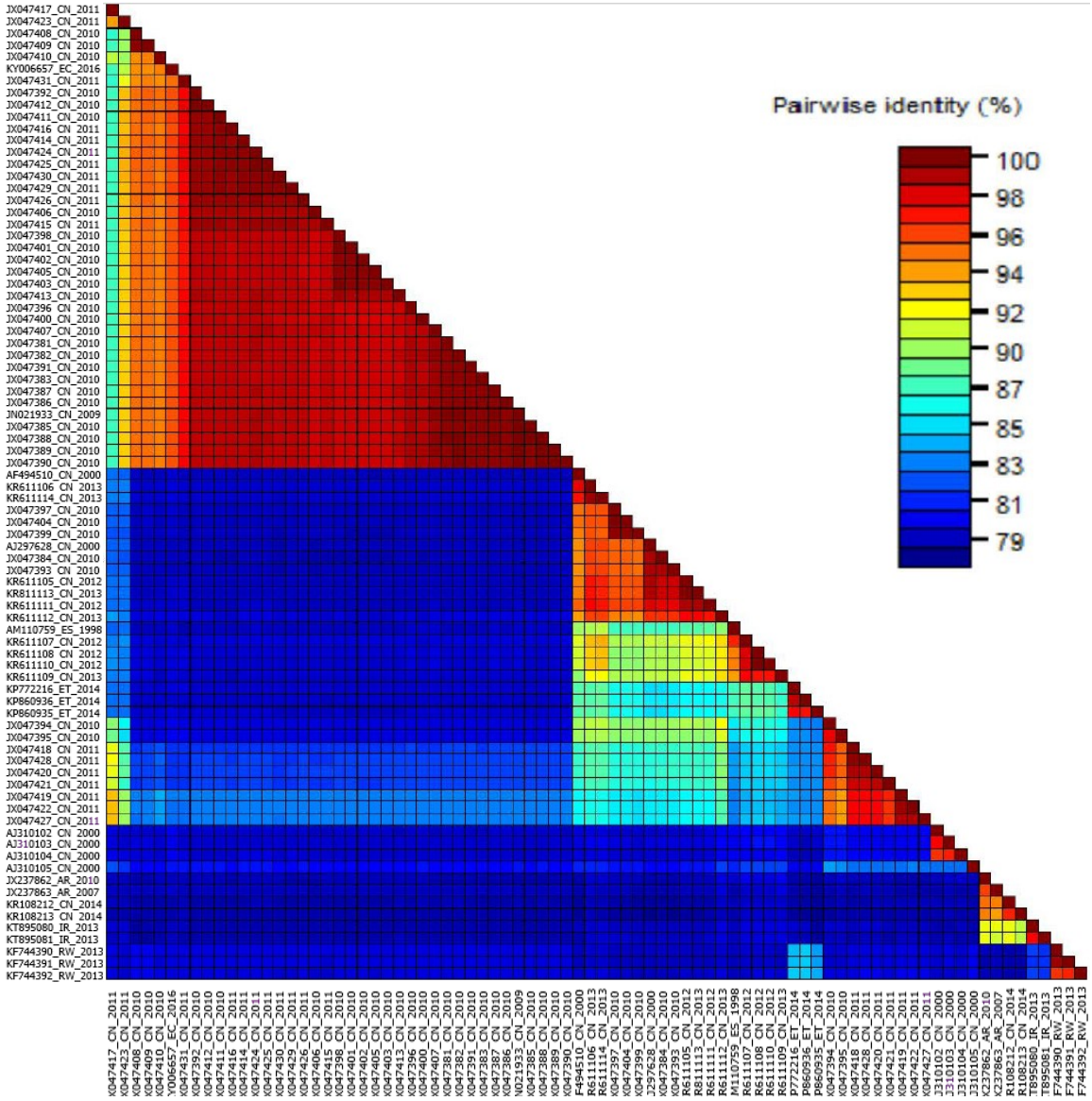

Figure 4 Heat map of percentage nucleotide pairwise identity for 82 sequences. Sequences are clustered based on similarity scores

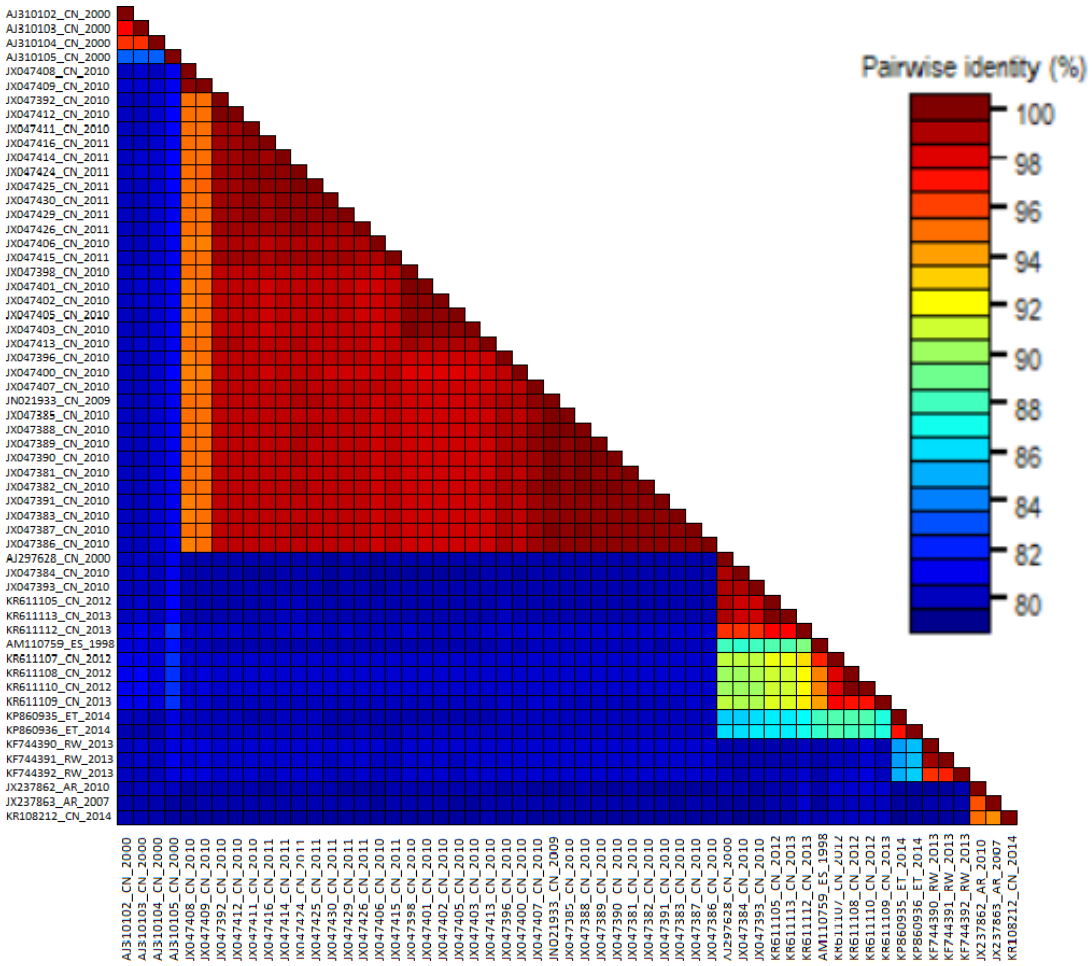

Figure 5 Heat map of percentage nucleotide pairwise identity after removal of recombinants (57 sequences). Sequences are clustered based on similarity scores

3.2 Recombination

Recombination analysis led to the exclusion of 25 sequences from phylogenetic analysis. There was no evidence of recombination found in 6K1, 6K2, VPg and P1. As a rule, this research ignored all recombination involving unknown parents in the RDP program as this may reflect recombination with sequences not in this analysis (what evolutionary biologists would consider migration instead of recombination). Ultimately, most of these recombinant sequences ended up being excluded from the analysis as they were strongly identified in other genes. All recombination breakpoints were mapped relative to AF494510-CN-2000 sequence. Any recombination detected any gene led to the exclusion of the whole polypeptide sequence from further study. Significant recombination results are presented for genes in order from the 5' end of the polyprotein to the 3' end.

P3 had only one event of recombination with the breakpoint occurring at 590 to 1041nt. This led to the removal of JX047423-CN-2011 from further analysis (Table 3).

Table 3: Recombination in P3

The breakpoints are the nucleotide numbers from the start of the gene's alignment, and the values for each algorithm are p-values, NS = not significant.

Recombinant		JX047423-CN-2011
Major parent		JX047421-CN-2011
Minor parent		JX047405-CN-2010
Break	Begin	590
points	End	1041*
Detection methods	RDP	NS
	GENECONV	2.46E-14
	Bootscan	3.72E-16
	Maxchi	1.71E-15
	Chimaera	1.97E-15
	SiSscan	9.02E-17
	3Seq	2.41E-32

* RDP called this an approximate breakpoint position. The actual breakpoint position is undetermined; RDP assumes a subsequent recombination event most

likely overprinted it, but it could also be in the adjoining gene in the polyprotein. Asterisks after breakpoints in subsequent recombination tables mean the same ambiguity applies.

Four significant recombination events were identified in HC-Pro which led to the removal of 13 recombinants (Table 4).

Table 4: Recombination in HC-Pro

RDP Event No		1+	2+	3	4+
Recombinants		KT895080-IR-2013 KT895081-IR-2013	JX047421-CN-2011 JX047394-CN-2010, JX047395-CN-2010, JX047418-CN-2011, JX047419-CN-2011, JX047420-CN-2011, JX047422-CN-2011, JX047423-CN-2011, JX047427-CN-2011, JX047428-CN-2011	KR108213-CN-2014	KR611111-CN-2012 KR611112-CN-2013
Major parent		KR108212-CN-2014	KP860935-ET-2014	KR108212-CN-2014	AJ297628-CN-2000
Minor parent		KF744390-RW-2013	AJ310105-CN-2000	JX237863-AR-2007	KR611107-CN-2012
Break	Begin	572*	4*	1295	1130
points	End	1372	1170	496*	1380*
Detection Methods	RDP	4.35E-09	NS	1.04E-07	NS
	GENECONV	1.87E-05	NS	4.78E-05	NS
	Bootscan	2.06E-08	NS	4.67E-05	1.28E-02
	Maxchi	4.41E-13	2.61E-04	2.87E-05	NS
	Chimaera	NS	3.07E-04	9.75E-03	2.98E-02
	SiSscan	1.28E-21	2.87E-21	NS	1.72E-03
	3Seq	4.76E-20	1.75E-10	1.14E-02	2.08E-02

There were five recombination breakpoints in CI (Table 5).

Table 5: Recombination in C1

RDP Event No		1	2+	3+	4+	5
Recombinant		JX047417-CN-2011	AF494510-CN-2000 KR611106-CN-2013, KR611114-CN-2013	JX047395-CN-2010 JX047394-CN-2010 JX047418-CN-2011 JX047419-CN-2011 JX047420-CN-2011 JX047421-CN-2011 JX047422-CN-2011 JX047427-CN-2011 JX047428-CN-2011	KR611111-CN-2012 JX047399-CN-2010	AF494510-CN-2000
Major parent		JX047410-CN-2010	AJ297628-CN-2000	KP860935-ET-2014	KR611106-CN-2013	Unknown
Minor parent		JX047428-CN-2011	KR611107-CN-2012	AJ310105-CN-2000	JX047384_CN_2010	KR611114-CN-2013
Break points	Begin	982*	1580	1664	266	99*
	End	1914*	1893*	1914*	820	844
Detection Methods	RDP	2.08E-37	2.82E-12	1.68E-09	4.18E-05	NS
	GENECONV	2.13E-28	2.99E-11	2.23E-04	6.91E-07	NS
	Bootscan	4.18E-38	7.00E-13	2.48E-10	1.90E-05	NS
	Maxchi	6.95E-27	8.86E-05	5.85E-06	1.88E-04	1.36E-03
	Chimaera	3.26E-20	7.82E-06	1.33E-06	2.55E-08	2.54E-02
	SiSscan	1.45E-35	2.86E-07	1.92E-04	3.39E-10	6.33E-07
	3Seq	4.33E-74	3.37E-09	1.11E-08	9.28E-09	2.05E-02

NIb had nine recombination events recognized by at least three detection methods resulting in 10 sequences being marked for removal (Table 6).

Table 6: Recombination in NIb

RDP Event No		1+	2+	3+	4	5	6+	7	8+	9
Recombinant		JX047427-CN-2011 JX047419-CN-2011 JX047419-CN-2011	JX047421-CN-2011 JX047418-CN-2011 JX047420-CN-2011 JX047428-CN-2011	KT895080-IR-2013 KT895081-IR-2013	JX047431-CN-2011	JX047395-CN-2010	KR611110-CN-2012 AM110759-ES-1998 KR611108-CN-2012 KR611109-CN-2012	AJ310102-CN-2000	JX047417-CN-2011 JX047410-CN-2010	KP860935-ET-2014
Major parent		JX047417-CN-2011	JX047394-CN-2010	KP772216-ET-2014	Unknown	KR611112-CN-2013	KR611113-CN-2013	AJ310103-CN-2000	JX047411-CN-2010	Unknown
Minor parent		JX047394-CN-2010	JX047417-CN-2011	JX237863-AR-2007	JX047400-CN-2010	JX047394-CN-2010	AJ310103-CN-2000	Unknown	JX047408-CN-2010	KP772216-ET-2014
Break points	Begin	444*	8*	1106	822	868	948	972	166*	940
	End	1563*	1284	1460	1526	1191*	1480*	178*	777*	1271*
Detection Methods	RDP	7.75E-25	9.76E-13	3.37E-20	6.21E-03	NS	NS	NS	NS	NS
	GENECONV	1.27E-17	2.29E-06	2.64E-15	1.86E-04	NS	NS	NS	NS	3.72E-02
	Bootscan	3.14E-25	3.73E-12	5.22E-18	4.28E-03	2.96E-02	4.06E-02	1.29E-04	NS	4.36E-04
	Maxchi	9.49E-19	1.12E-10	7.04E-13	2.88E-07	3.49E-07	6.65E-05	3.48E-03	3.60E-04	4.90E-05
	Chimaera	NS	NS	1.03E-12	6.77E-08	9.44E-05	1.59E-06	5.95E-03	1.94E-02	NS
	SiSscan	1.11E-42	5.95E-45	1.26E-08	1.46E-04	6.96E-05	3.28E-10	3.10E-07	5.50E-08	5.10E-12
	3Seq	3.03E-56	6.89E-42	1.37E-12	1.44E-08	1.40E-04	NS	1.43E-02	5.56E-04	NS

The CP showed extensive recombination, with 45 of the 82 sequences in the dataset implicated as recombinant by the second event identified by RDP. However, 16 of these had an unknown minor parent and 19 of those with known major and minor parents, resulting in only partial evidence of recombination – likely due to recombination with a sequence outside of the dataset (migration rather than intraspecific recombination). The recombinants listed in Table 7 reflect the sequences removed from the dataset because there was the strongest evidence for recombination for these ten sequences. A total of 13 sequences were marked as recombinant in the CP region (Table 7).

Table 7: Recombination in CP

RDP Event No		1+	2+	3
Recombinant		KR108213-CN-2014 KR108212-CN-2014	JX047422-CN-2011 JX047417-CN-2011, JX047418-CN-2011, JX047419-CN-2011, JX047420-CN-2011, JX047421-CN-2011, JX047422-CN-2011, JX047427-CN-2011, JX047428-CN-2011, JX047410-CN-2012, KY006657-EC-2016	KR108213-CN-2014
Major parent		JX237863-AR-2007	Unknown	KR108212-CN-2014
Minor parent		Unknown/	JX047431-CN-2011	AF494510-CN-2000
Break	Begin	299	347	902
points	End	918*	937*	421*
Detection Methods	RDP	NS	NS	1.16E-04
	GENECONV	NS	NS	NS
	Bootscan	NS	NS	NS
	Maxchi	7.25E-14	2.53E-09	1.00E-06
	Chimaera	3.67E-05	2.08E-04	1.93E-03
	SiSscan	4.82E-20	7.47E-07	NS
	3Seq	4.40E-04	1.60E-02	3.16E-02

Table 8: Recombinant sequences removed

Sequence	Recombinant in:					
	P3	**HC-Pro**	**C1**	**NIa-Pro**	**NIb**	**CP**
JX047417-CN-2011		√	√			√
JX047394-CN-2010		√	√			
JX047395-CN-2010		√	√		√	
JX047418-CN-2011		√	√		√	√
JX047419-CN-2011		√	√		√	√
JX047420-CN-2011		√	√		√	√
JX047421-CN-2011		√	√			√
JX047422-CN-2011		√	√		√	√
JX047427-CN-2011		√	√		√	√
JX047428-CN-2011		√	√		√	√
KR611111-CN-2012			√			
AF494510-CN-2000			√			
KR611106-CN-2013			√			
KR611114-CN-2013			√			
JX047423-CN-2011	√	√				
KP772216-ET-2014				√		
JX047410-CN-2012						√
KY006657-EC-2016						√
KT895080-IR-2013		√			√	
KT895081-IR-2013		√			√	
JX047431-CN-2011					√	
JX047404-CN-2010			√			
JX047399-CN-2010			√			
KR108213-CN-2014		√				√
KR108212-CN-2014						√
	1	14	16	1	10	12

3.3 Temporal signal

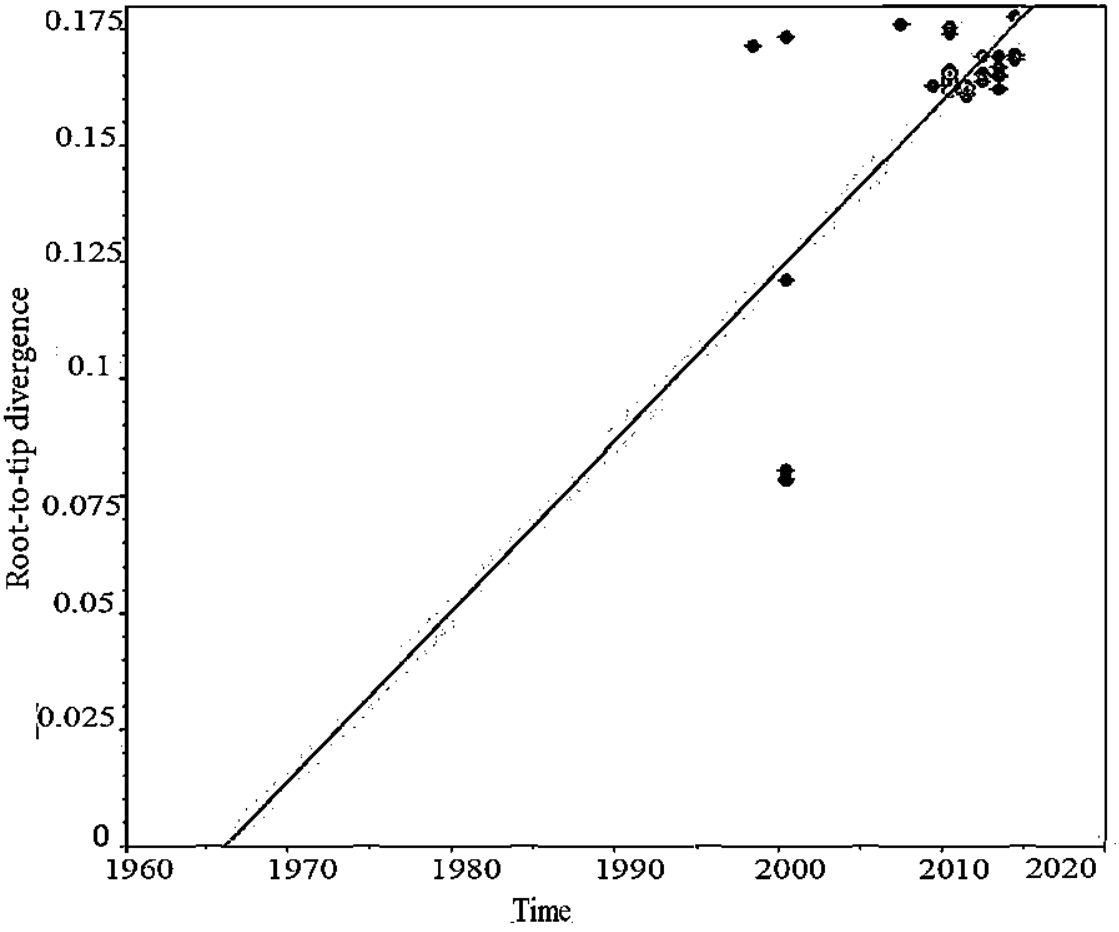

Figure 6 Temporal signal of 57 sequences

The correlation coefficient (r) of root-to-tip distance against sampling date for the whole polypeptide sequence was 0.644 indicating that there was relatively strong temporal signal to allow rate estimation. This was also found to be consistent with the individual genes (P1: r=0.57, HC-Pro: r=0.64, P3: r=0.62, 6K1: r=0.52, C1: r=0.62, 6K2: r=0.31, VPg: r=0.58, NIa-Pro: r=0.62, NIb: r=0.63, CP: r=0.61), justifying analyses on rate of evolution for each gene in SCMV).

3.4 Model selection

Table 9: Models chosen for each gene's analysis

	P1	**HC-Pro**	**P3**	**6K1**	**C1**	**6K2**	**VPg**	**NIa-Pro**	**NIb**	**CP**
Substitution Model	TIM2+I+G	TIM3+I+G	GTR+I+G	TPM2+G	GTR+I+G	K80+G	TrN+I+G	TrN+I+G	GTR+I+G	GTR+G
Maximum likelihood estimates (MLE) from Path sampling (Chain length: 2,000,000, No of steps: 100)										
	6K1	**6K2**	**C1**	**CP**	**HC-Pro**	**NIa-Pro**	**NIa-VP**	**NIb**	**P1**	**P3**
Clock model										
Strict clock	-1532.4	-	-12591.4	-6149.1	-9028.6	-5142.2	-4253.5	-11086.2	-5715.5	-6632.1
Lognormal clock	-1545.7	--	-12546.6	-6185.9	-9030.2	-5139.2	-4258.3	-12191.9	-5688.6	-6599.4
Exponential clock	**-1523.9**	**-1230.4**	**-12530.6**	**-6100.5**	**-8983.6**	**-5125.7**	**-4238.7**	**-11066.0**	**-5681.7**	**-6580.18**
Tree model										
Constant pop	-1520.8	-1226.2	-12520.2	-6095.7	**-8978.1**	-5123.6	-4237.3	-11066.4	-5679.8	-6572.2
Exponential	-1522.1	-1228.8	-12522.9	-6097.4	-8981.9	-5126.2	-4238.4	**-11055.1**	-5679.5	-6579.3
Bayesian	**-1512.5**	**-1222.7**	**-12513.5**	**-6093.1**	-8983.0	**-5115.5**	**-4233.1**	-11056.2	**-5674.7**	**-6565.8**

The best fitting site models were GTR + Γ_4 + I for C1, NIb and C3; TrN+Γ_4 +I for VPg and NIa-Pro; TIM2+ Γ_4 + I for P1; TIM3+ Γ_4 + I for HC-Pro; TPM2+ Γ_4 for 6K1; K80+ Γ_4 for 6K2 andGTR+ Γ_4 for CP (Table 9). The K80 (transitions have one rate, transversions another: two rates) and GTR (all substitution pairs have their own rate: six rates) were discussed in the introduction, the TrN model has three rates (each transition pair has its own rate, all transversions have another), the TPM model has three rates (all transitions have one rate, and transversions are grouped into two sets, each with their own rate) and the TIM model is an expansion of the TPM with four rates (each transition pair has its own rate and transversions are grouped into two sets, each with their own rate). The exponential clock model was chosen as it consistently outperformed the other models tested. The exponential clock model has been recommended for a conservative approach to divergence time estimation (Ho, Phillips, Drummond, & Cooper, 2005). Various tree models were chosen as demographic priors, with the most common model being the flexible Bayesian skyline model.

3.5 Polymorphism

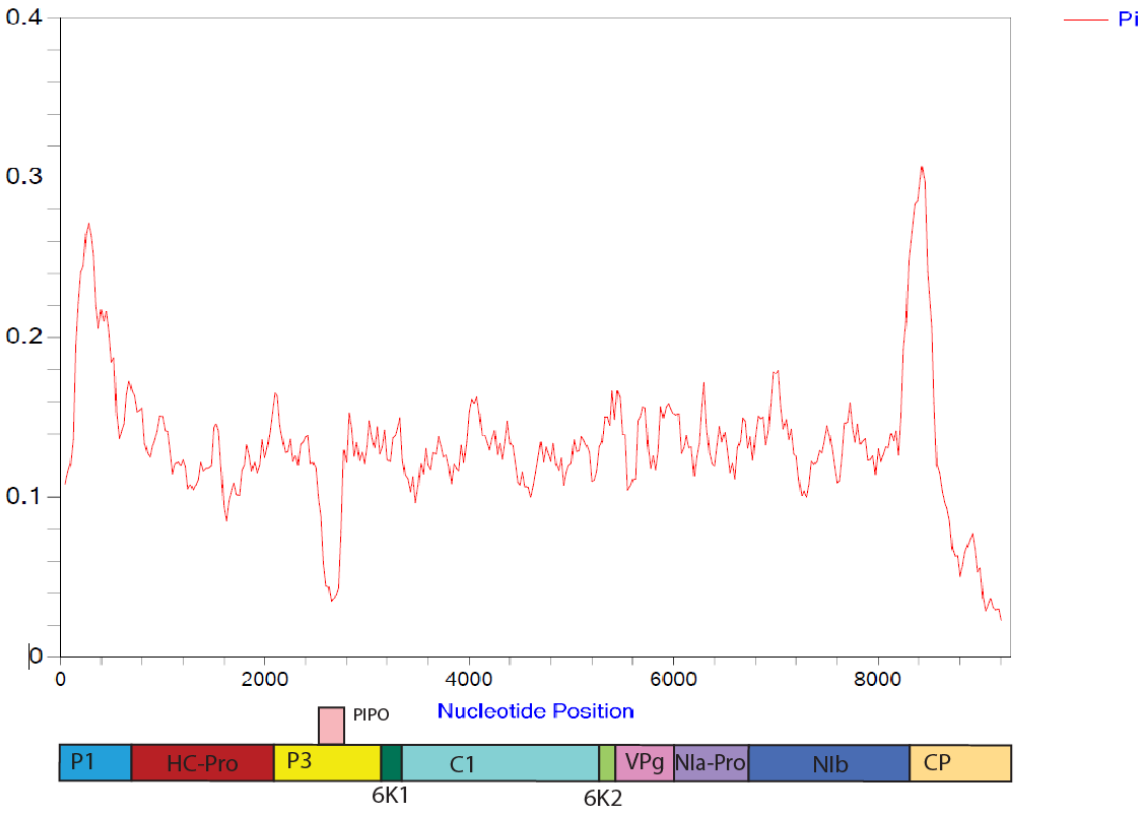

Figure 7 Nucleotide diversity (pi) for the complete polyprotein sequence (sites 1-9153) excluding sites with gaps mapped to a scaled representation of the position of the various genes. The sliding window used for the analysis was 100 nucleotides.

The two regions of the polyprotein with the highest diversity were the 5' and 3' genes: P1 and N-terminus of CP (Figure 7). This matches the diversity seen in CP length during alignment, when it was the only ORF with insertions and deletions, focused on its 5' end. The lowest diversity in the polyprotein was observed in P3, exactly were the overlapping PIPO ORF would constrain the evolution of nucleotides. Low diversity was also identified in the 3' end of CP.

3.6 Selection pressure

Table 10: Non-synonymous/synonymous rate ratios for SLAC, FEL and MEME tests

Gene	Non-synonymous/ synonymous rate ratios		Unique sequences	Total sites	Sites with pervasive positive/diversifying selection			Sites with pervasive negative/ purifying selection	
	SLAC	FEL/MEME			SLAC	FEL	MEME (episodic)	SLAC	FEL
P1	0.192	0.158	48	233	0	1	3	91	121
HC-Pro	0.034	0.0252	53	460	0	0	2	284	386
P3	0.0763	0.0688	50	347	0	4	6	153	213
6K1	0.0466	0.0358	30	67	0	0	0	40	54
C1	0.0241	0.0167	51	638	0	0	6	630	543
6K2	0.06	0.039	28	53	0	0	0	25	36
VPg	0.0647	0.0505	44	189	0	0	3	116	154
NIa-Pro	0.0409	0.031	46	242	0	0	0	161	196
NIb	0.0479	0.0376	54	521	0	0	4	342	420
CP	0.153	0.134	47	333	0	0	11	101	148

Selection analyses confirmed that each ORF in the SCMV genome is under significant selection to retain function, as demonstrated by overall strong purifying selection. The dN/dS ratios for each gene were very low, below 0.1 by multiple methods for all but two genes (P1 and CP, Table 10). Furthermore, more than half of the sites in all of the genes except P1 and CP are under detectable purifying selection (FEL results, Table 10). Some diversifying (or positive selection) was detected in more than half of the genes (by MEME, numbering is codon in the appropriate gene): the variable P1 (30, 68, 174) and CP (16, 70, 149, 172, 198, 230, 246, 258, 277, 295, 325), P3 (14, 93, 184, 197, 200, 221), VPg (2, 25 58), NIb (143, 154, 323, 463), C1(291,421, 427, 430, 492, 592), and HC-Pro (245, 279). FEL detected the signal of positive selection for five sites in P3 (14, 184,

221, 222) and P1 (68), but in no case, was positive selection detected by the conservative SLAC approach.

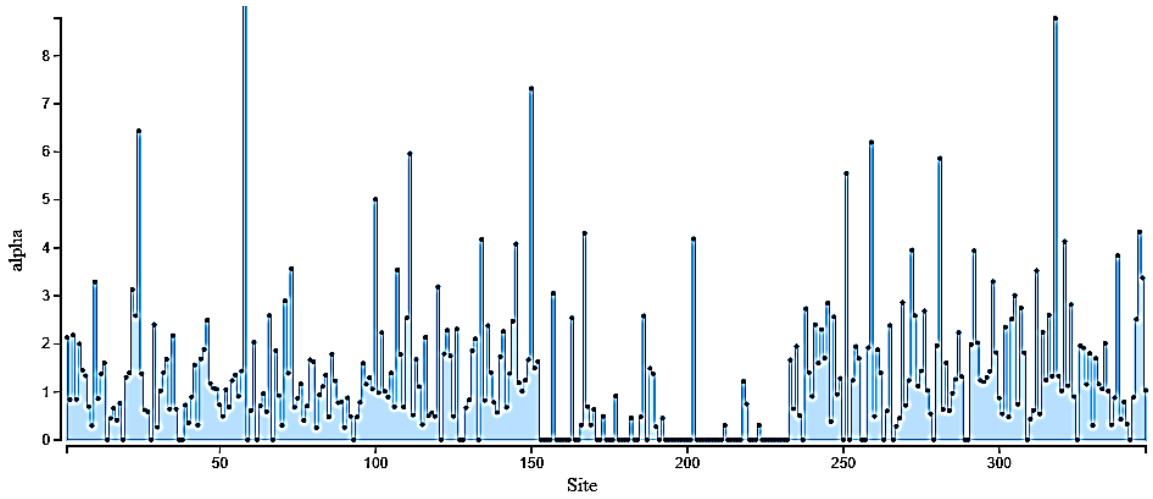

Figure 8 FEL Site Plot for P3

The highly conserved region on P3 was from site 153 to 232 (Figure 8). Site 153 coincided with the start of the highly conserved P3N-PIPO motif of G2A6 that was identified from amino-acids 459-466 (Cheng et al., 2017; Chung et al., 2008).

3.7 Phylogeny

There is little support for any of the deep branches in the phylogenetic trees, but strong support for clades in every gene trees (Figures 9-18). The nucleotide identity matrices (Figure 4) clearly show that the sequences from similar geographical backgrounds have the highest similarity, and cluster together. The phylogenetic analysis emphasizes this with most terminal clustering being from the same countries (Figures 4-5). Since most of the sequences studied were from China, Chinese sequences form the bulk of the tree, but Chinese sequences do not form a monophyletic lineage in any of the trees.

Sequences from Argentina closely associate with the same Chinese sequence (KR106212-CN-2014) in all genes indicative of a close relationship and recent divergence – likely due to recent migration of the disease from one location to the other (the trees offer mixed evidence for the direction and do not resolve this). Rwandan strains form their own distinct sub cluster in 8 genes and are only seen to associate with Ethiopian strains in the VPg, NIb and CP on the 3' end of the polyprotein – providing some evidence that African sequences are more closely associated with each other than with sequences from other continents (Figure 14, 16 and 17). In the 5' proteins, the Ethiopian strains had grouped more closely with the single sequence from Spain, a pattern consistent with recombination occurring between the 6K2 and VPg genes. Since we used recombination methods only within each gene's dataset, we did not detect more global patterns of recombination, but recombination events at or near junctions of genes would not compromise the suitability of any gene's dataset for BEAST analysis. Nonetheless, our data suggest that the Ethiopian 2014 sequences may be descended from European and East African parental sequences.

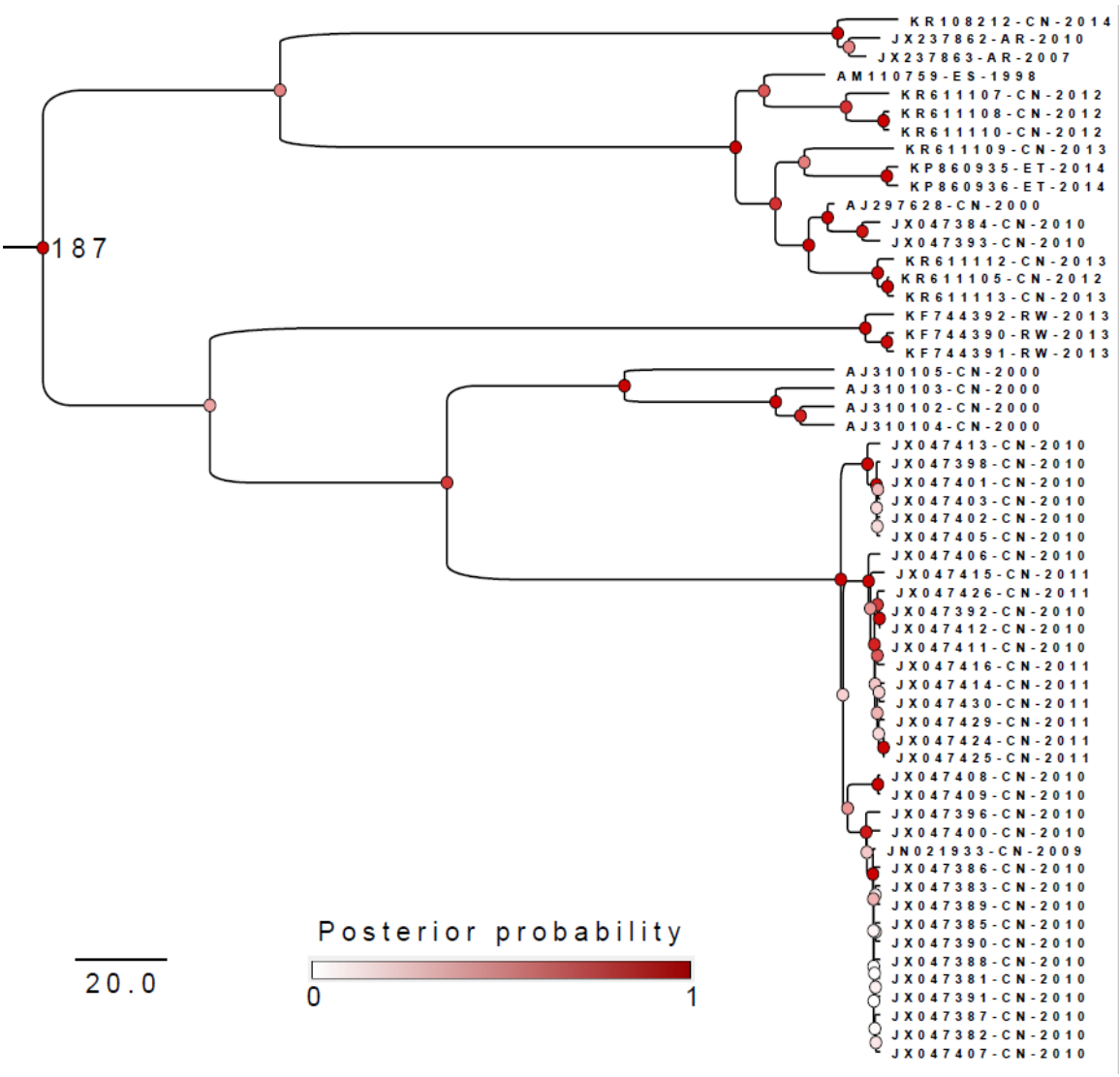

Figure 9 P1 Maximum clade credibility (MCC) phylogenetic tree with time along the x-axis.

The scale bar is in years, the number at the deepest bifurcation in the tree is the mean estimated TMRCA. Node support given by depth of red color, where white is no support and the deepest red is a posterior probability of 1.

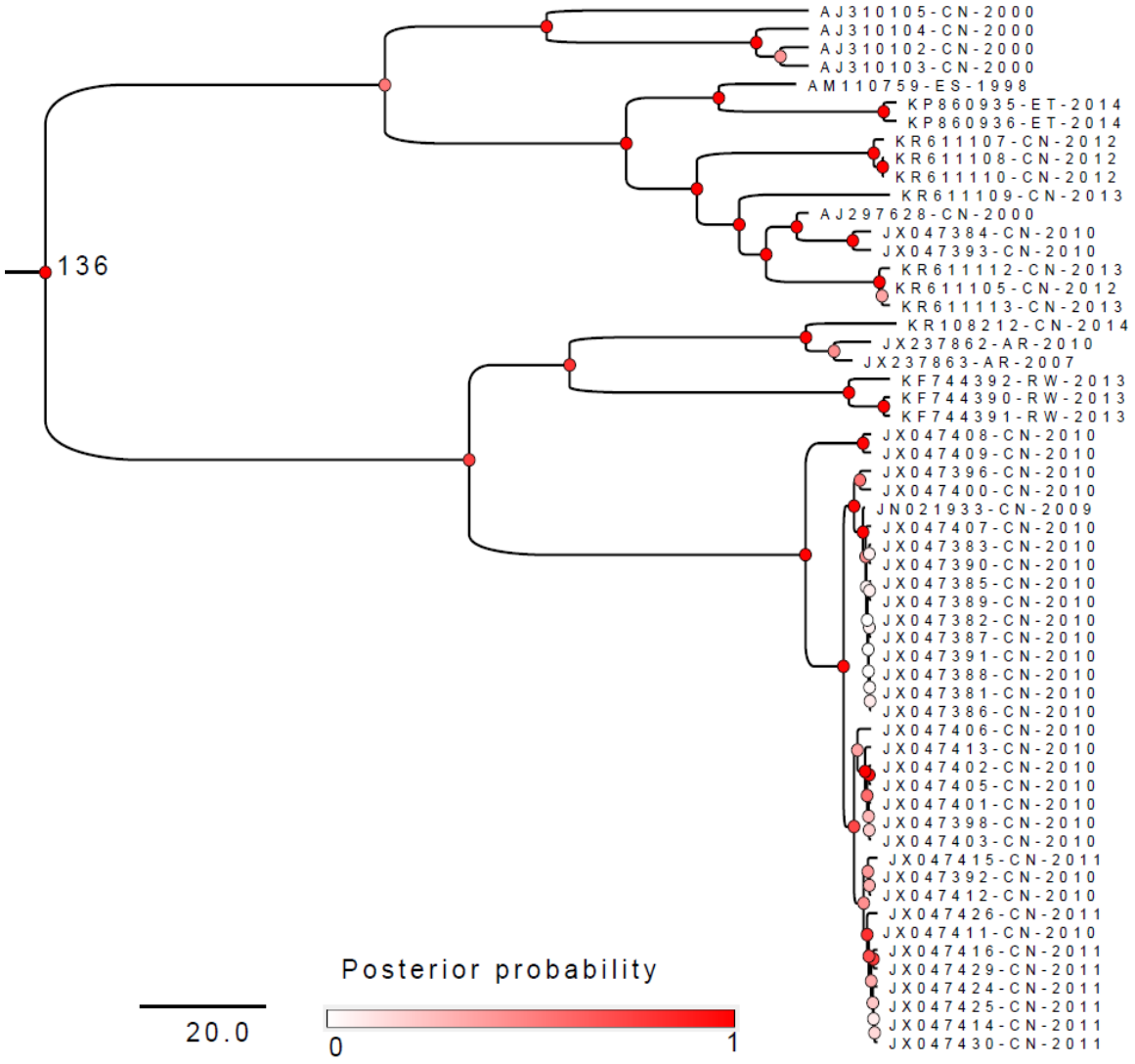

Figure 10 HC-Pro Maximum clade credibility (MCC) phylogenetic tree with time along the x-axis.

The scale bar is in years, the number at the deepest bifurcation in the tree is the mean estimated TMRCA. Node support given by depth of red color, where white is no support and the deepest red is a posterior probability of 1.

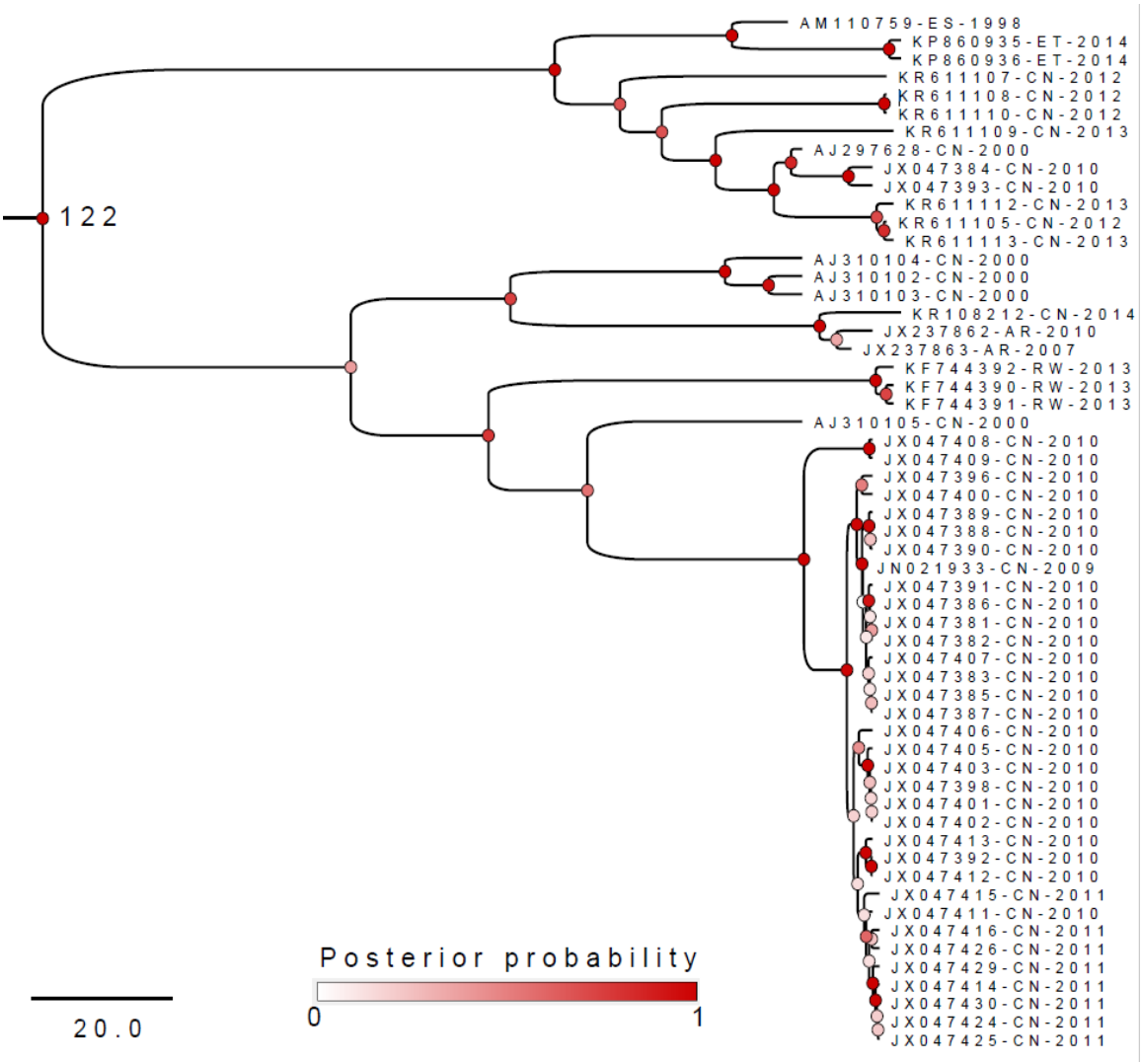

Figure 11 P3 Maximum clade credibility (MCC) phylogenetic tree with time along the x-axis.

The scale bar is in years, the number at the deepest bifurcation in the tree is the mean estimated TMRCA. Node support given by depth of red color, where white is no support and the deepest red is a posterior probability of 1.

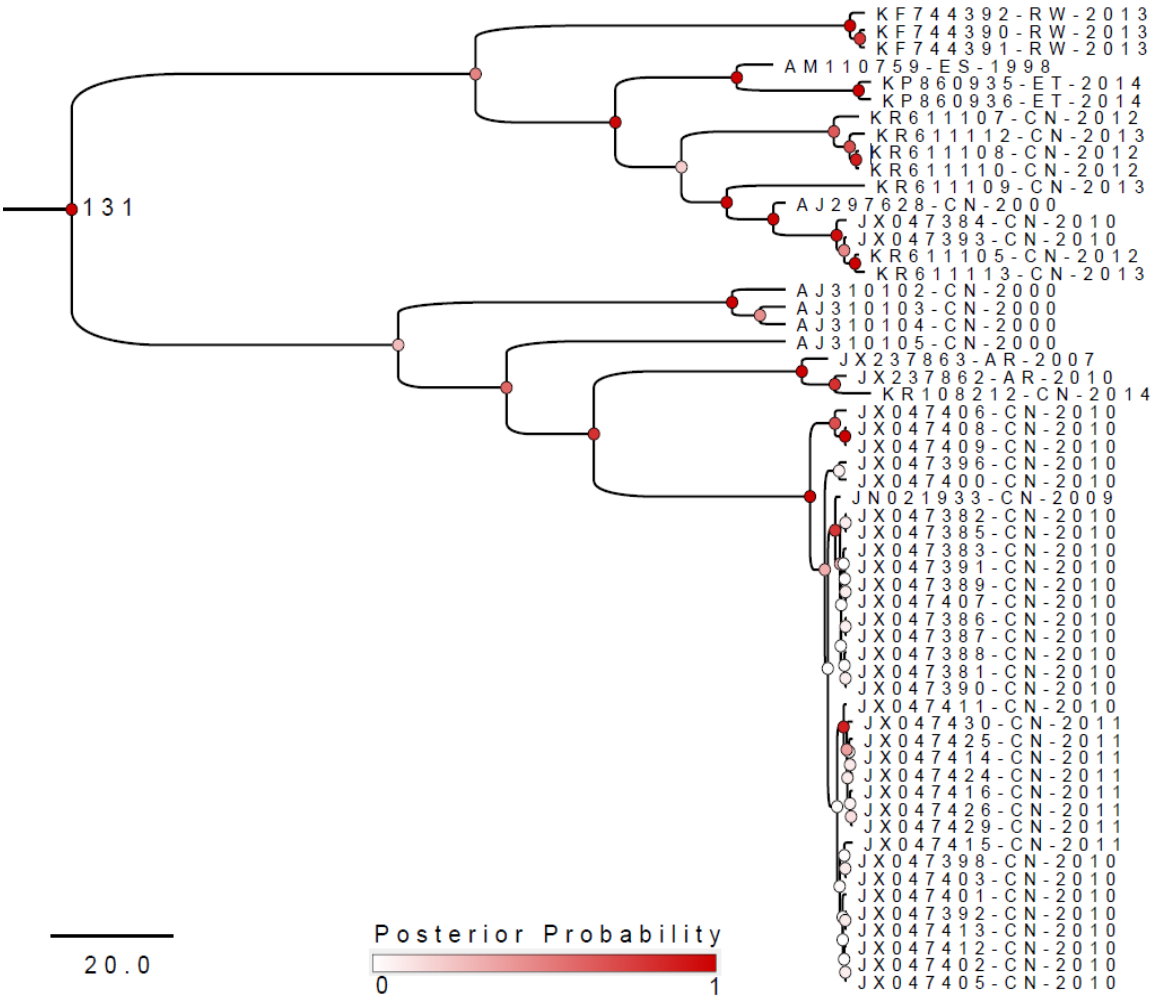

Figure 12 6K1 Maximum clade credibility (MCC) phylogenetic tree with time along the x-axis.

The scale bar is in years, the number at the deepest bifurcation in the tree is the mean estimated TMRCA. Node support given by depth of red color, where white is no support and the deepest red is a posterior probability of 1.

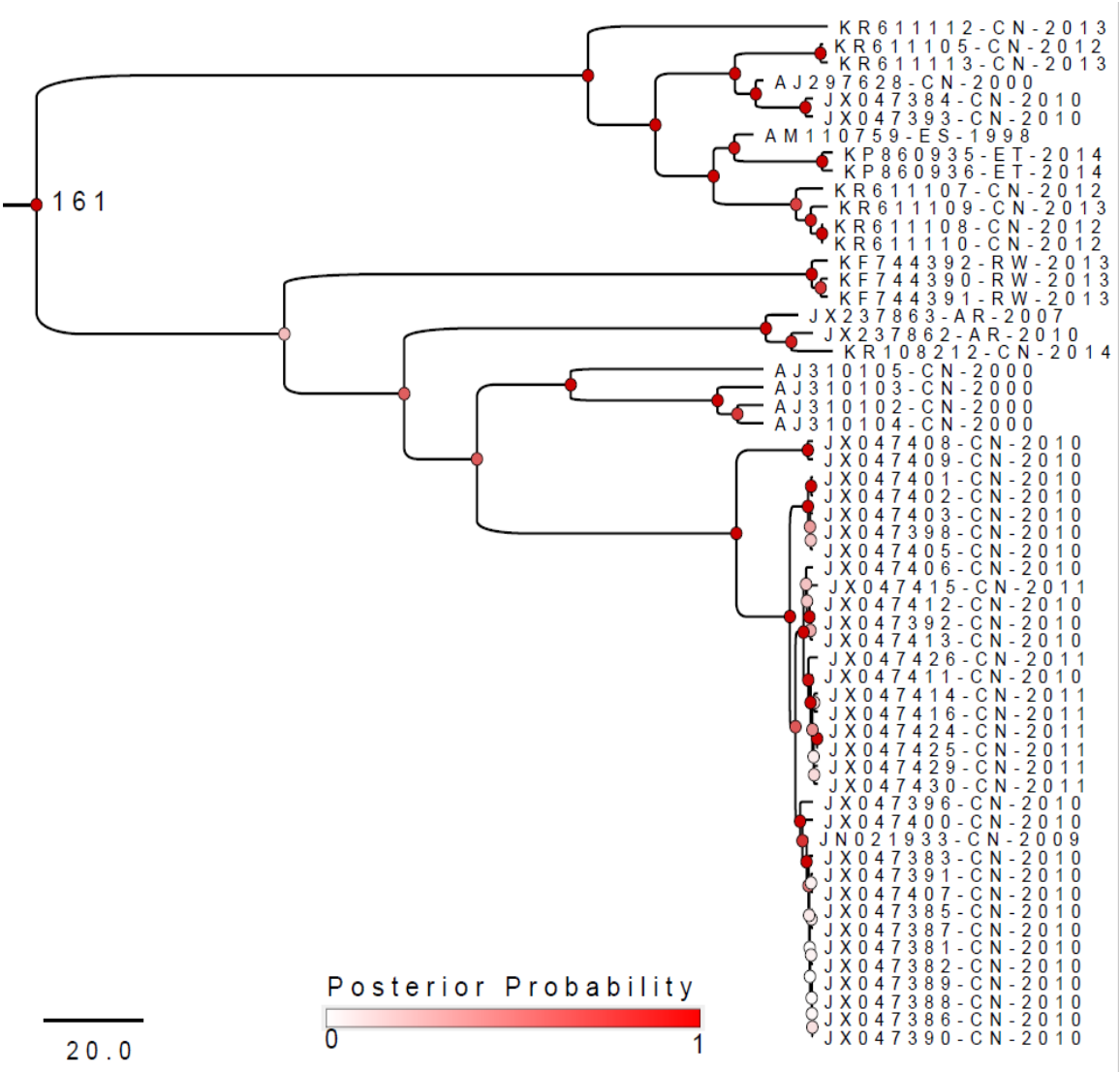

Figure 13 C1 Maximum clade credibility (MCC) phylogenetic tree with time along the x-axis.

The scale bar is in years, the number at the deepest bifurcation in the tree is the mean estimated TMRCA. Node support given by depth of red color, where white is no support and the deepest red is a posterior probability of 1.

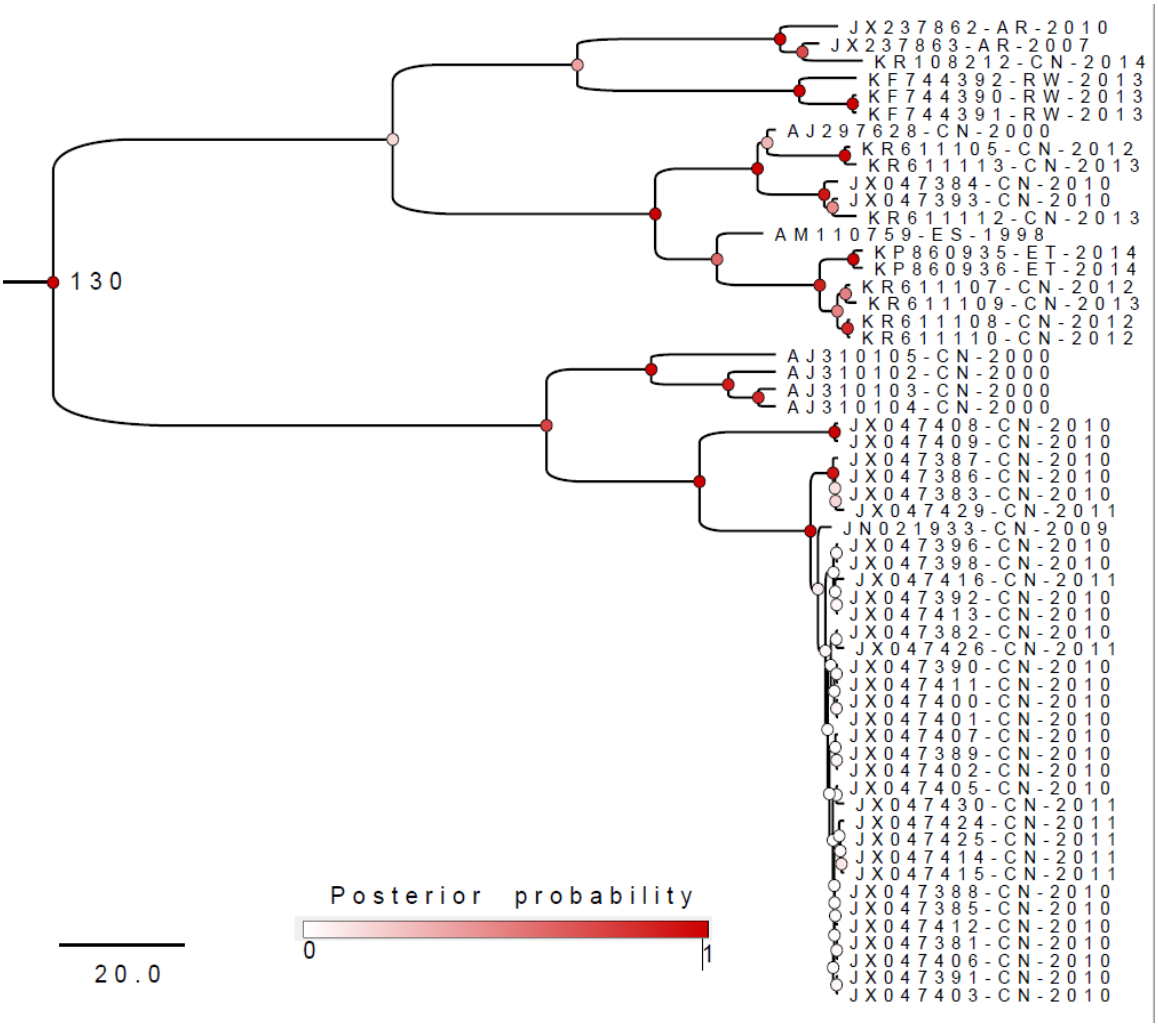

Figure 14 6K2 Maximum clade credibility (MCC) phylogenetic tree with time along the x-axis.

The scale bar is in years, the number at the deepest bifurcation in the tree is the mean estimated TMRCA. Node support given by depth of red color, where white is no support and the deepest red is a posterior probability of 1.

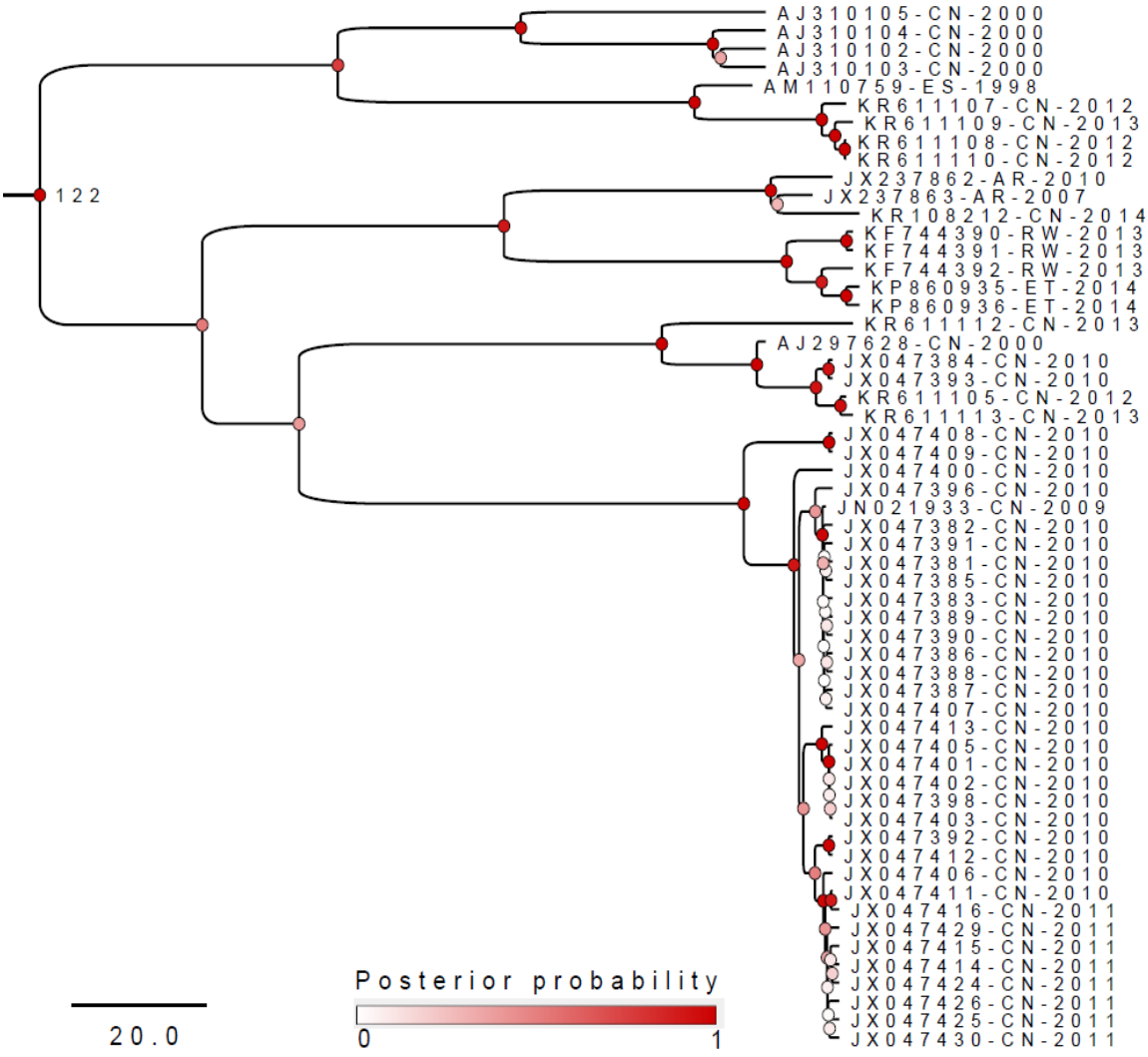

Figure 15 VPg Maximum clade credibility (MCC) phylogenetic tree with time along the x-axis.

The scale bar is in years, the number at the deepest bifurcation in the tree is the mean estimated TMRCA. Node support given by depth of red color, where white is no support and the deepest red is a posterior probability of 1.

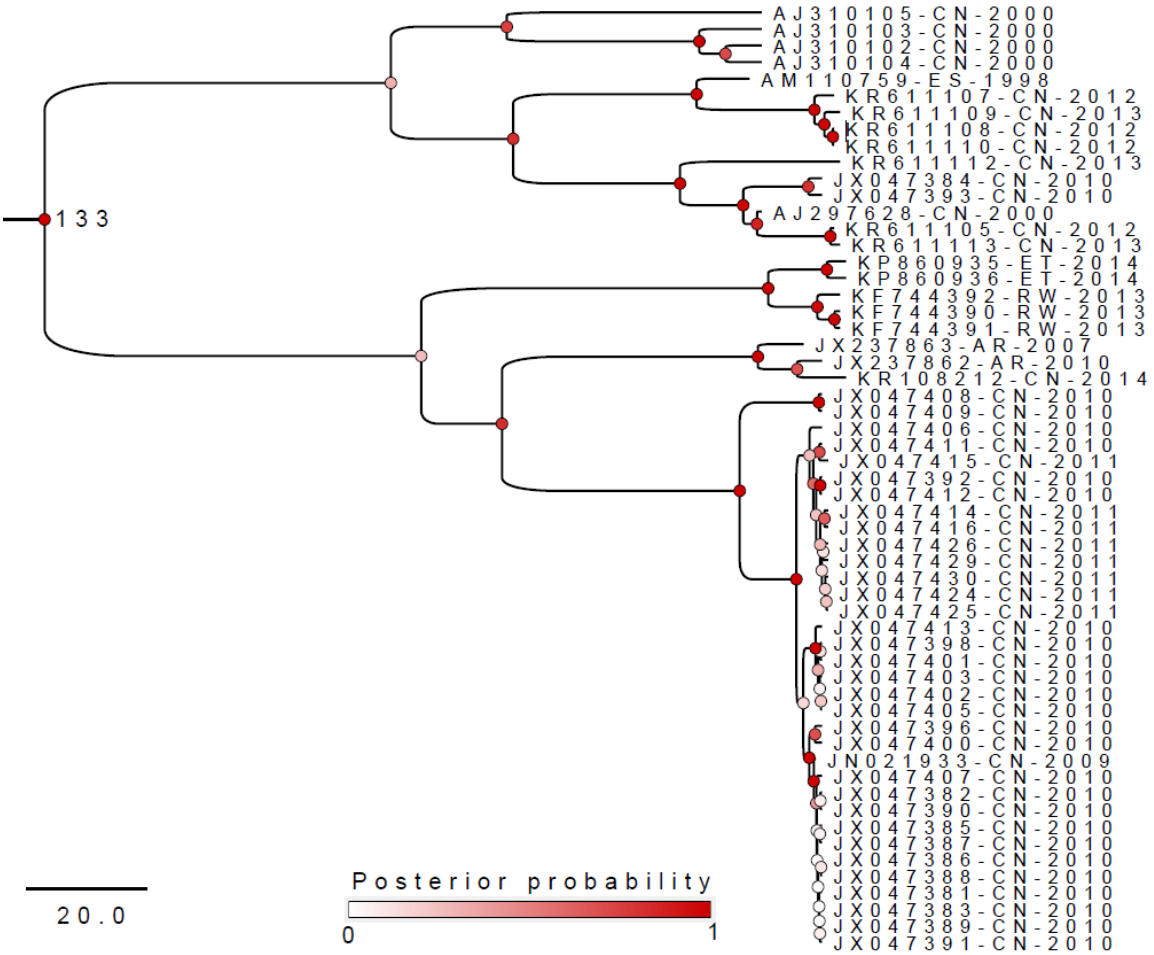

Figure 16 NIa-Pro Maximum clade credibility (MCC) phylogenetic tree with time along the x-axis.

The scale bar is in years, the number at the deepest bifurcation in the tree is the mean estimated TMRCA. Node support given by depth of red color, where white is no support and the deepest red is a posterior probability of 1.

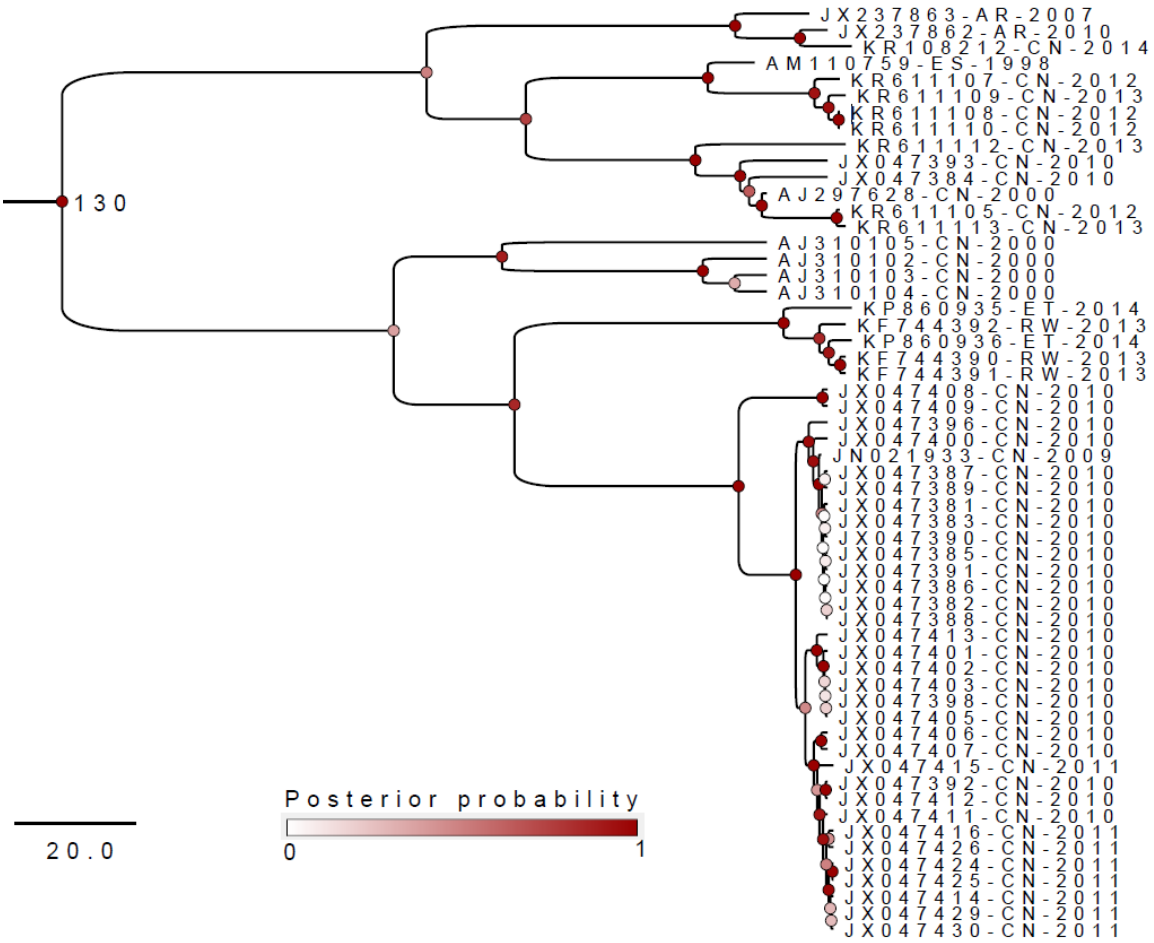

Figure 17 NIb Maximum clade credibility (MCC) phylogenetic tree with time along the x-axis.

The scale bar is in years, the number at the deepest bifurcation in the tree is the mean estimated TMRCA. Node support given by depth of red color, where white is no support and the deepest red is a posterior probability of 1.

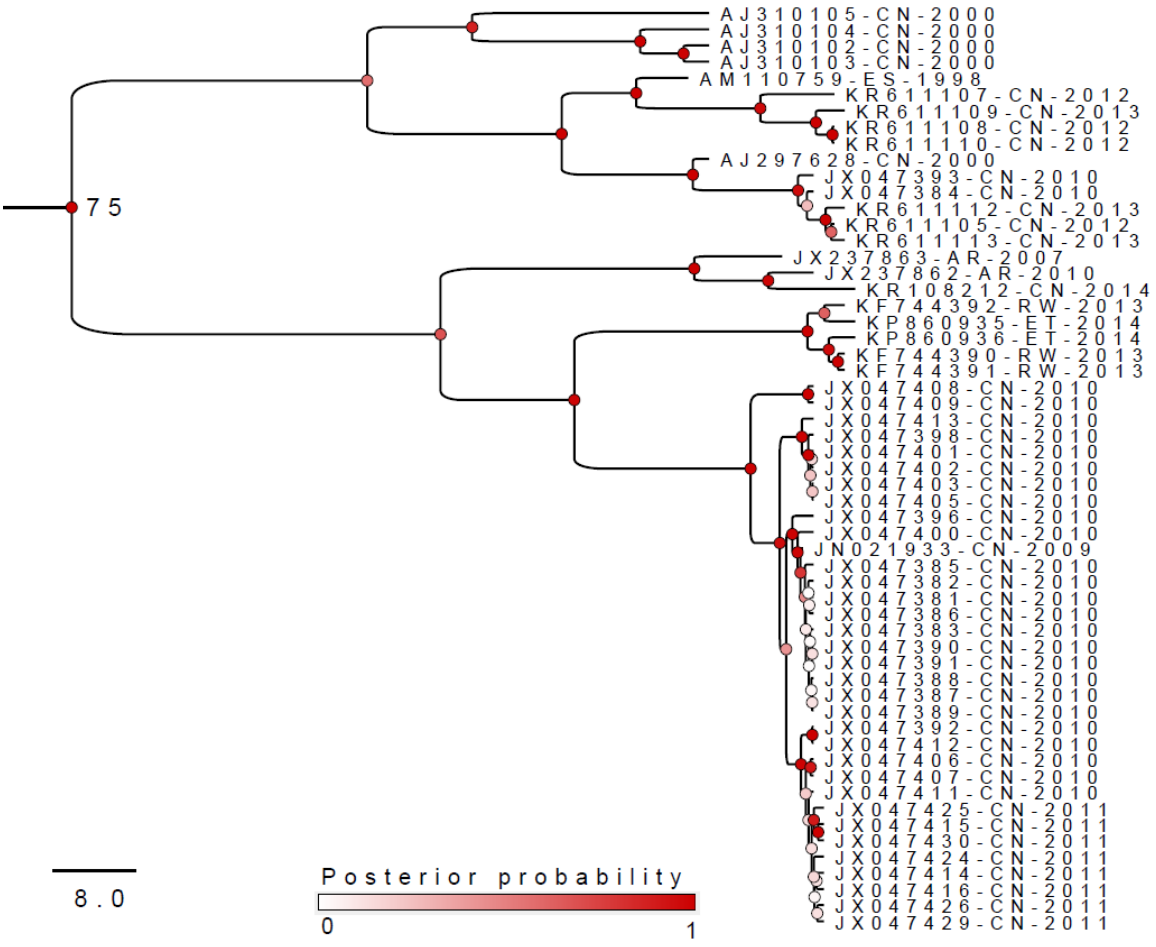

Figure 18 CP Maximum clade credibility (MCC) phylogenetic tree with time along the x-axis.

The scale bar is in years, the number at the deepest bifurcation in the tree is the mean estimated TMRCA. Node support given by depth of red color, where white is no support and the deepest red is a posterior probability of 1.

3.8 Substitution rates and TMRCA

The substitution rates estimated for each gene were fairly similar, with significantly overlapping 95% credibility intervals (the region of highest posterior density, Figure 18). This indicates that the polyprotein is evolving at the same consistent rate of 3.3-4.3 $x10^{-3}$ substitutions/site/year, despite differences in selection pressures (Table 10). Importantly, the substitution rates are different from the rates produced by the prior alone (without the dataset, Figure 19). Each gene's data led to a similar rate of evolution, suggesting that SCMV evolves consistently despite different functional constraints and levels of selection pressure.

With similar rates of evolution comes the expectation the genes will have similar times to most recent common ancestor (TMRCA). Indeed, all ten SCMV genes show overlapping distributions for a range of coalescent age, though CP has a more recent coalescent age than the other genes, due to its faster-estimated rate of evolution. The mean coalescent year for the genes of SCMV range from 1750-1795. While sampling from the prior for TMRCA overlapped entirely with the results of the analysis (Figures 20 and 21), the 95% credibility interval of the prior alone ranged from 700 years ago to 500 years into the future – as in the most recent common ancestor of SCMV circulating today has yet to occur, hundreds of years in the future. Because of the non-overlapping distribution of substitution rates, we disregarded the overlap of these results and considered the SCMV dataset to have produced significant results. The exact values for mean estimated rate and age, with 95% credibility intervals are given in tabular form in Table 11.

Table 11: Estimated gene substitution rates and TMRCA

Substitution rate										
mean	3.46E-03	3.38E-03	3.36E-03	3.81E-03	3.35E-03	4.16E-03	4.02E-03	3.66E-03	3.70E-03	4.29E-03
95% HPD interval	3.0E-3 to 4.41E-3	3E-3 to 4.1282E-3	3E-3 to 4.05E-3	3E-3 to 5.40E-3	3E-3 to 4.03E-3	3E-3 to 6.4799E-3	3.00E-3 to 5.92E-3	3.00E-3 to 4.95E-3	3E-3 to 5.08E-3	3.00E-3 to 6.31E-3
TMRCA										
mean	-264	-255	-260	-240	-259	-238	-246	-253	-260	-219
95% HPD interval	-285 to -241	-273 to -237	-278 to -242	-270 to -210	-277 to -240	-272 to -199	-272 to -217	-275 to -227	-284 to -234	-246 to -191

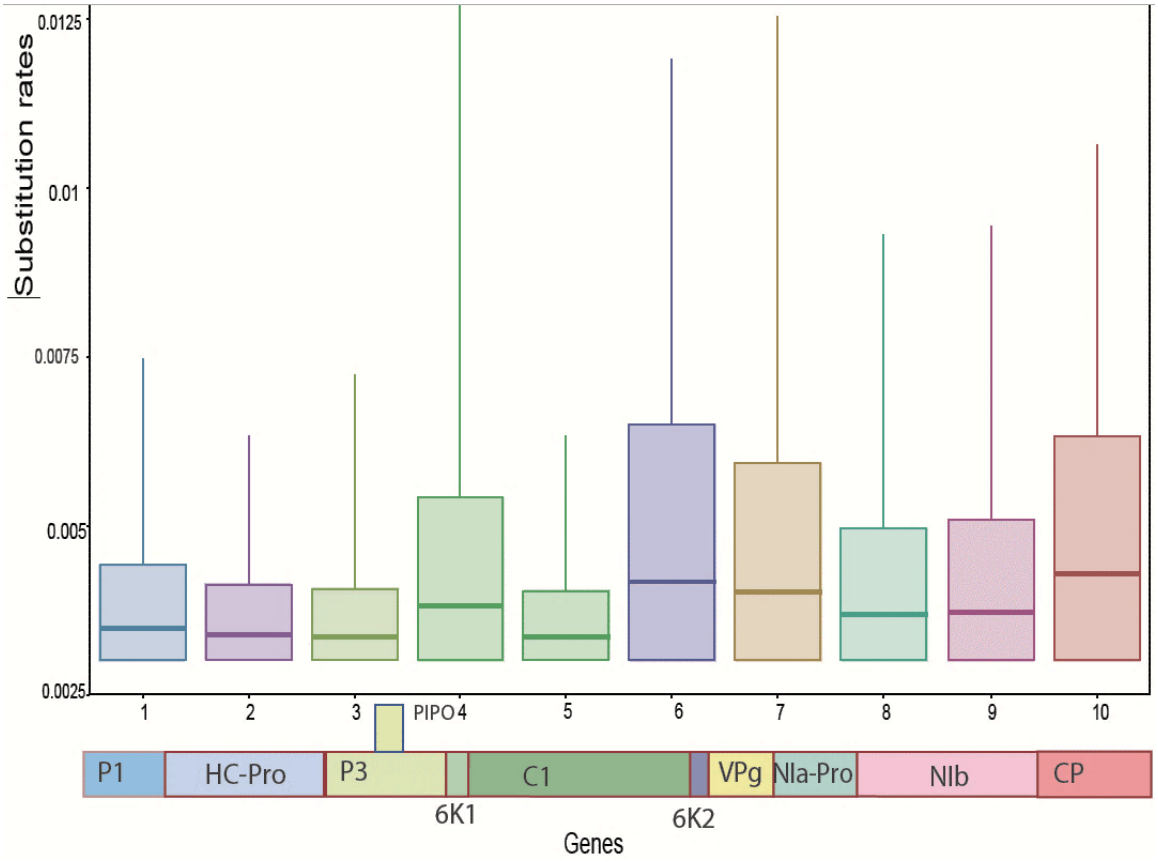

Figure 19 Substitution rates for the 10 SCMVgenes (57 sequences). Mean estimated rates are shown by the bars, the boxes show the 95% highest posterior density and whiskers show the full range of results.

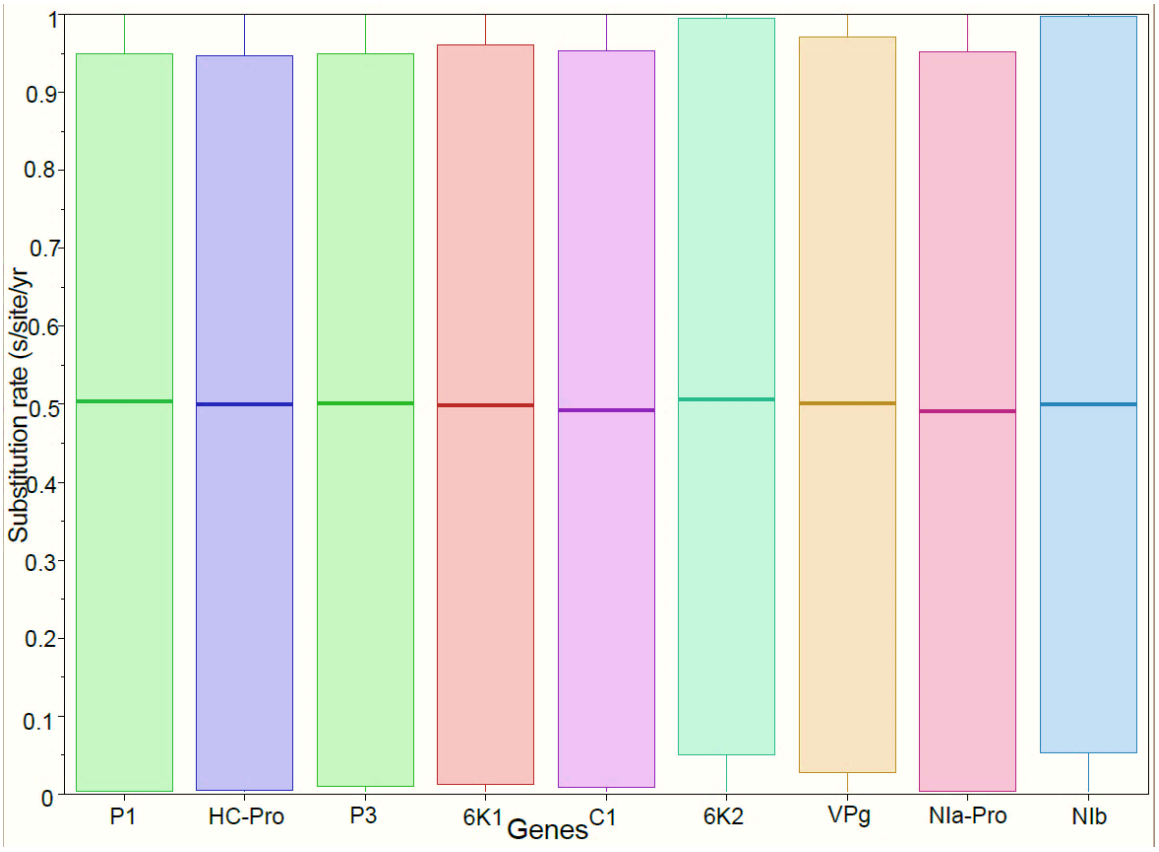

Figure 20 Substitution rates (sampling from the prior)
Shown for 9 genes, 95% HPD for CP was out of the scope of this scale

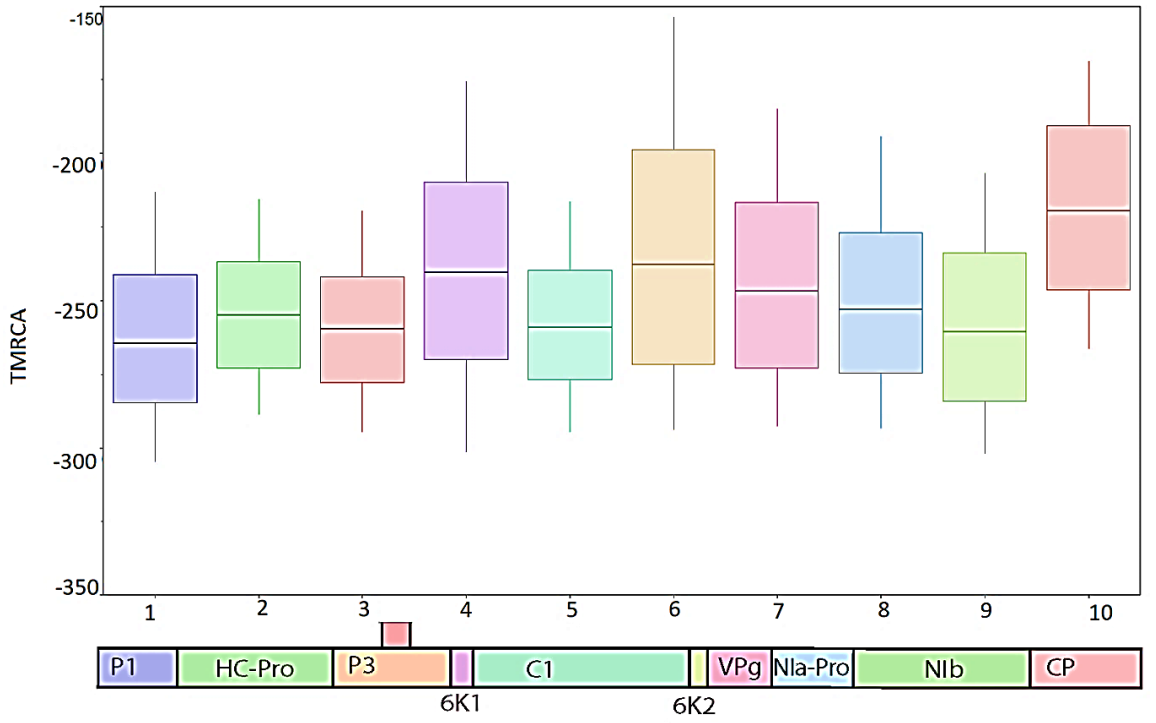

Figure 21 TMRCA (value in years) for each gene from 57 SCMV sequences (years before 2014, the most recently isolated viruses in the dataset).

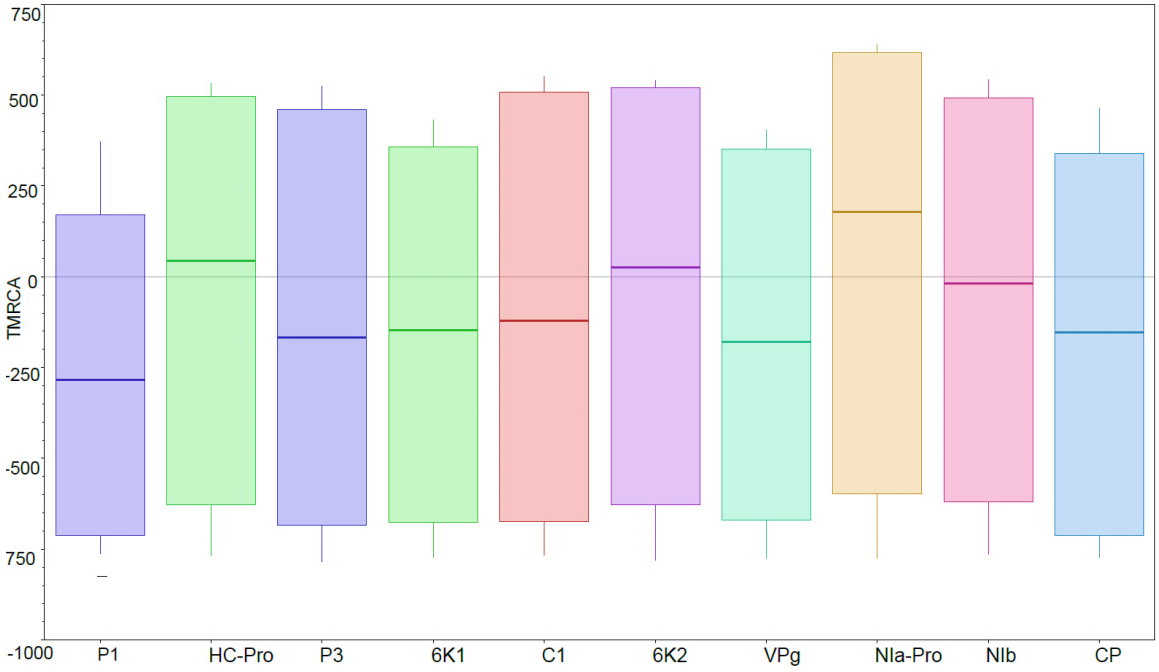

Figure 22 TMRCA (sampling from the prior)

CHAPTER 4: DISCUSSION

Mean estimated substitution rates of between $3.35x10^{-3}$ in CI to $4.29x10^{-3}$ I CP with the 95% HPD significantly overlapping to suggest that all the genes of the polyprotein evolve at more or less the same rate. This is consistent with the observation that most RNA viruses have mutations ranging from 10^{-2} to 10^{-5} nucleotide substitutions per site, per year (Hicks & Duffy, 2014; Mbewe, 2017). TMRCAs ranging from 219 to 264 years before 2014 places the divergence time of the different clades more than 100 years prior to the first recorded sugarcane mosaic sighting in 1898 (Artschwager, 1948). This is a significant age difference with animal-infecting RNA viruses, which tend to be recognized within a decade or two of their coalescent ages (Hicks & Duffy, 2012). On the other hand, this coalescent is similar to many plant viruses that have been studied, in the range of 150-300 years before present (Pagan et al., 2010; Pagan & Holmes, 2010).

The nucleotide identity of 74-99% for the genes in the polyprotein was well within the species demarcation criterion of <76% genomic nucleotide identity set for potyviruses (Adams et al., 2005) even though the 5' and 3' terminals were not considered. Sequences from the same geographic origin were shown to have higher nucleotide identities, often >95%, consistent with the clustering observed in the phylogenetic tree and previous studies (Chen et al., 2002). The association of sequence from Argentina with those from China shows that this plant virus can move long distances in short periods of time, confounding the assumption that closely clustering sequences must be geographically co-located. These more recent migrations prevent us from seeing the history of the sugarcane host moving around the globe, from probable domestication in Pacific islands to the rest of Asia, to Hawaii and then the Caribbean islands and into the Americas. The first

observation of sugarcane mosaic symptoms followed some of this path: Java, Indonesia followed by Puerto Rico in 1916 and later Louisiana in the US (Artschwager & Brandes, 1958; Wu et al., 2012), but we cannot substantiate this path from molecular data, especially with the heavy sampling skew towards Chinese SCMV isolates in GenBank.

Recombination is key driver of evolution in Potyviruses (Y. Li et al., 2013), and it was not surprising that recombination breakpoints were frequently detected. While we saw statistical evidence of recombination in most of the longer genes (P3, HC-Pro, C1, NIa-Pro, NIb, and CP) and not in the shortest genes 6K1, 6K2 or P1 and VPg, in two previous independent studies, recombination had been detected in 6K1, VPg, NIaPro and NIb in SCMV (Y. Li et al., 2013; Padhi & Ramu, 2011). The N-terminus of SCMV CP gene and indeed, for most potyviruses has previously been reported to be highly recombinant (Li et al., 2013) and contains highly variable, virus-specific epitopes (Revers & Garcia, 2015; D. Shukla, M. Strike, L. Tracy, H. Gough, & Ward, 1988). This was quite evident from this study as CP had wide gaps in alignment, especially towards the N-terminus. The variability in CP is likely due to many different ways of accommodating the protein interactions necessary for viral replication and transmission (Revers & Garcia, 2015). N- and C-terminal regions of the coat proteins are exposed on the surfaces of the virus particles and tend to be more variable while the core is relatively conserved (D. Shukla et al., 1988).

Consistent with previous studies, the overlapping section of P3N-PIPO was found to be under strong purifying selection pressure. Proposals put forward to explain the evolution of overlapping reading frames include as a strategy to compress the genome in response to the high mutation rates, coupling gene expression to regulation or simply as a way of

accommodating more genes in a short genome limited by the capsid size (Chirico, Vianelli, & Belshaw, 2010). The highly conserved P3N-PIPO motif of G2A6 was identified from amino-acids 459-466 (Cheng et al., 2017; Chung et al., 2008), in the region of lowest nucleotide diversity in the SCMV genome. The length of PIPO is reported to be highly variable both within and between *Potyviruses* species and was determined to range from 1 to 89 amino-acids (Hillung, Elena, & Cuevas, 2013). Because the G2A6 P3N-PIPO motif coincided with the start of the highly conserved region of P3 and assuming that PIPO encompasses this entire region, we speculate that the size of PIPO is approximately 79 amino-acids long. Three out of four of P3 positive sites detected by FEL and four out of six sites detected by MEME were found in the highly conserved region with least nucleotide diversity, suggesting that these are true positive signals of selection for different amino acids. The positive selection detected by MEME and FEL was concentrated in certain regions of the genome such as P3N-PIPO region in P3, N-terminus in VPg, 3' end of C1 but nonspecific in other genes such as CP. These genes/regions must play a major role in the emergence of new strains capable of adapting to new environments and infecting new hosts as these genes are also implicated in virus-host and virus-vector interactions. Previous studies had looked at positive selection in the CP region only, and we did not find evidence that the sites identified by others were under positive selection in our dataset (Y. Li et al., 2013; Xie et al., 2016)).

Conserved regions on the genome can be exploited in developing control strategies for the disease. The highly conserved PIPO sequence can be a good candidate for a homology-dependent gene silencing in generating resistance against crop viral infections. Genetic engineering can break the plant-virus interaction by disturbing the host gene

expression resulting in the development of resistant varieties (Revers & Garcia, 2015). Since aphid transmission involves the interaction between CP and HC-Pro protein and the aphid stylet, CP and HC-Pro can be disrupted to control spread by the vector (Simmons et al., 2008). In many vectored plant viruses, interactions with both a host and a vector can lead to lower variability, making them good targets for control (Chare & Holmes, 2004). Gan et al. (2010) have demonstrated the use of short hairpin RNA with sequence homology to SCMV CP sequence segments in the control of Sugarcane mosaic disease, showing that regions even in the most variable SCMV gene can be targets of antiviral control.

Viral genomes are usually under great selection pressure due to the density of genes in their compact genomes. P1, P3, and CP are known to be the most variable regions in potyviruses (Mbewe, 2017; D. D. Shukla et al., 1991). Consistent with these previously published results, these were the three genes with the highest dN/dS ratios (SLAC and FEL, Table 10). In fact, these three ORFs also showed the most residues under episodic positive selection (Table 10). Hillung et al. (2013) have noted that there is a significant positive selection in the PIPO overlapping section of P3 in a variety of potyviruses. Except for these known variable genes, purifying selection dominates SCMV evolution, as is common for viruses (Hughes & Hughes, 2007; Xie et al., 2016).

CP has been the gene of choice in viral phylogenetic studies and has been deemed to be representative of the whole genome to the exclusion of other genes. This study shows that if recombination has been accounted for as in this study, all the genes can give consistent rates of evolution and there need not be an over-reliance on CP.

Y. Li et al. (2013) in a CP/ 3′-UTR phylogenetic study, showed that SCMV strains are genetically differentiated based on their host and geographical isolation points. It would be interesting to conduct a phylogeographic analysis on all the genes while factoring the host species and specific isolation points to find out if this holds true along the entire polypeptide. However, because of the higher availability of CP sequences of SCMV in GenBank, more advanced phylodynamic analyses such as these may only be possible using just the CP gene.

This knowledge from this study enhances the understanding of SCMV evolution, highlights residues in several genes that may be affecting SCMV-crop interactions and may help in developing strategies for the control of the diseases in plants.

REFERENCES

Adams, M. J., Antoniw, J. F., & Fauquet, C. M. (2005). Molecular criteria for genus and species discrimination within the family Potyviridae. *Arch Virol, 150*(3), 459-479. doi:10.1007/s00705-004-0440-6

Artschwager, E. (1948). *Vegetative characteristics of some wild forms of saccharum and related grasses*. Washington: Dept. of Agriculture.

Artschwager, E., & Brandes, E. W. (1958). *Sugarcane (Saccharum officinarum L.): origin, classification, characteristics, and descriptions of representative clones*. [Washington, D.C.]: Agricultural Research Service, Crops Research Division.

Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M. A., & Alekseyenko, A. V. (2012). Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty. *Molecular Biology and Evolution, 29*(9), 2157-2167. doi:10.1093/molbev/mss084

Boni, M. F., Posada, D., & Feldman, M. W. (2007). An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics, 176*(2), 1035-1047. doi:10.1534/genetics.106.068874

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., . . . Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLOS Computational Biology, 10*(4), e1003537. doi:10.1371/journal.pcbi.1003537

Bouckaert, R. R., & Drummond, A. J. (2017). bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evol Biol, 17*(1), 42. doi:10.1186/s12862-017-0890-6

Chandran, V., & Gajjeraman, P. (2015). Molecular Diversity Analysis of Pretty Interesting Potyviridae ORF (PIPO) Coding Region in Indian Isolates of Sugarcane streak mosaic virus. *Sugar Tech, 18*(2), 214-221. doi:10.1007/s12355-015-0376-z

Chare, E. R., & Holmes, E. C. (2004). Selection pressures in the capsid genes of plant RNA viruses reflect mode of transmission. *J Gen Virol, 85*(Pt 10), 3149-3157. doi:10.1099/vir.0.80134-0

Chen, J., Chen, J., & Adams, M. J. (2002). Characterisation of potyviruses from sugarcane and maize in China. *Arch Virol, 147*(6), 1237-1246. doi:10.1007/s00705-001-0799-6

Cheng, G., Dong, M., Xu, Q., Peng, L., Yang, Z., Wei, T., & Xu, J. (2017). Dissecting the Molecular Mechanism of the Subcellular Localization and Cell-to-cell Movement of the Sugarcane mosaic virus P3N-PIPO. *Sci Rep, 7*(1), 9868. doi:10.1038/s41598-017-10497-6

Cheon, S., & Liang, F. (2014). Bayesian phylogeny analysis. In M.-H. Chen, L. Kuo, & P. O. Lewis (Eds.), *BAYESIAN PHYLOGENETICS Methods, Algorithms, and Applications* (pp. 129-162). Boca Raton, FL: CRC Press Taylor & Francis Group.

Chirico, N., Vianelli, A., & Belshaw, R. (2010). Why genes overlap in viruses. *Proc Biol Sci, 277*(1701), 3809-3817. doi:10.1098/rspb.2010.1052

Choudhuri, S. (2014). Bioinformatics for beginners : genes, genomes, molecular evolution, databases and analytical tools.

Chung, B. Y., Miller, W. A., Atkins, J. F., & Firth, A. E. (2008). An overlapping essential gene in the Potyviridae. *Proc Natl Acad Sci U S A, 105*(15), 5897-5902. doi:10.1073/pnas.0800468105

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

Delport, W., Poon, A. F. Y., Frost, S. D. W., & Kosakovsky Pond, S. L. (2010). Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics, 26*(19), 2455-2457. doi:10.1093/bioinformatics/btq429

Domingo-Calap, P., & Sanjuan, R. (2011). Experimental evolution of RNA versus DNA viruses. *Evolution, 65*(10), 2987-2994. doi:10.1111/j.1558-5646.2011.01339.x

dos Reis, M., Donoghue, P. C., & Yang, Z. (2016). Bayesian molecular clock dating of species divergences in the genomics era. *Nat Rev Genet, 17*(2), 71-80. doi:10.1038/nrg.2015.8

Drummond, A. J., & Bouckaert, R. R. (2015). *Bayesian Evolutionary Analysis with BEAST*. Cambridge: Cambridge University Press.

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol, 7*, 214. doi:10.1186/1471-2148-7-214

Duffy, S., Shackelton, L. A., & Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics, 9*, 267. doi:10.1038/nrg2323

Faillace, C. A., Lorusso, N. S., & Duffy, S. (2017). Overlooking the smallest matter: viruses impact biological invasions. *Ecol Lett, 20*(4), 524-538. doi:10.1111/ele.12742

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution, 17*(6), 368-376. doi:10.1007/BF01734359

Fraile, A., Pagan, I., Anastasio, G., Saez, E., & Garcia-Arenal, F. (2011). Rapid genetic diversification and high fitness penalties associated with pathogenicity evolution in a plant virus. *Molecular Biology and Evolution, 28*(4), 1425-1437. doi:10.1093/molbev/msq327

Gan, D., Zhang, J., Jiang, H., Jiang, T., Zhu, S., & Cheng, B. (2010). Bacterially expressed dsRNA protects maize against SCMV infection. *Plant Cell Reports, 29*(11), 1261-1268. doi:10.1007/s00299-010-0911-z

Gibbs, A. J., Fargette, D., Garcia-Arenal, F., & Gibbs, M. J. (2010). Time--the emerging dimension of plant virus studies. *J Gen Virol, 91*(Pt 1), 13-22. doi:10.1099/vir.0.015925-0

Gibbs, M. J., Armstrong, J. S., & Gibbs, A. J. (2000). Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics, 16*(7), 573-582.

Guindon, S., Gascuel, O., & Rannala, B. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology, 52*(5), 696-704. doi:10.1080/10635150390235520

Hicks, A. L., & Duffy, S. (2012). One misdated sequence of rabbit hemorrhagic disease virus prevents accurate estimation of its nucleotide substitution rate. *BMC Evol Biol, 12*, 74. doi:10.1186/1471-2148-12-74

Hicks, A. L., & Duffy, S. (2014). Cell tropism predicts long-term nucleotide substitution rates of mammalian RNA viruses. *PLoS Pathog, 10*(1), e1003838. doi:10.1371/journal.ppat.1003838

Hillis, D. M. (2015). Molecular Evolution: A Statistical Approach by Ziheng Yang. *The Quarterly Review of Biology, 90*(1), 89-90. doi:10.1086/679928

Hillung, J., Elena, S. F., & Cuevas, J. M. (2013). Intra-specific variability and biological relevance of P3N-PIPO protein length in potyviruses. *BMC Evolutionary Biology, 13*(1), 249. doi:10.1186/1471-2148-13-249

Ho, S. Y., Phillips, M. J., Drummond, A. J., & Cooper, A. (2005). Accuracy of rate estimation using relaxed-clock models with a critical focus on the early metazoan radiation. *Molecular Biology and Evolution, 22*(5), 1355-1363. doi:10.1093/molbev/msi125

Holkar, S. K., Kumar, A., Meena, M. R., & Lal, R. J. (2017). Detection and partial molecular characterization of sugarcane mosaic virus infecting sugarcane genotypes. *Journal of Environmental Biology, 38*(3), 409-417. doi:10.22438/jeb/38/3/MS-219

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science, 294*(5550), 2310-2314. doi:10.1126/science.1065889

Hughes, A. L., & Hughes, M. A. (2007). More effective purifying selection on RNA viruses than in DNA viruses. *Gene, 404*(1-2), 117-125. doi:10.1016/j.gene.2007.09.013

Jiang, J., & Laliberte, J. F. (2011). The genome-linked protein VPg of plant viruses-a protein with many partners. *Curr Opin Virol, 1*(5), 347-354. doi:10.1016/j.coviro.2011.09.010

Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. In H. N. Munro (Ed.), Mammalian protein metabolism (Vol. 3, pp. 21-123). New York: Academic Press. doi:http://dx.doi.org/10.1016/B978-1-4832-3211-9.50009-7

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., . . . Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics, 28*(12), 1647-1649. doi:10.1093/bioinformatics/bts199

Kimura, M. (1980). A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide-Sequences. *Journal of Molecular Evolution, 16*(2), 111-120. doi:Doi 10.1007/Bf01731581

King, A. M. Q., Lefkowitz, E. J., Mushegian, A. R., Adams, M. J., Dutilh, B. E., Gorbalenya, A. E., . . . Davison, A. J. (2018). Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2018). *Arch Virol*. doi:10.1007/s00705-018-3847-1

Koike, H. (1988). *Sugar-cane Diseases: A Guide for Field Identification*. In (pp. 127). Retrieved from http://hdl.handle.net/10568/45118

Koike H. (1974). *Interaction between diseases on sugarcane: sugarcane mosaic and ratoon stunting disease.* Paper presented at the Proceedings XV Congress, International Society of Sugar Cane Technologists, Durban, South Africa.

Koike, H., & Gillaspie, A. G., Jr. (1989). Mosaic. In C. E. Ricaud, B.T.; Gillaspie, A.G.; Hughes, C.G. (Ed.), Diseases of Sugarcane: Major Diseases. Amsterdam: Elsevier Science.

Lanave, C., Preparata, G., Sacone, C., & Serio, G. (1984). A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution, 20*(1), 86-93. doi:10.1007/bf02101990

Larget, B., & Simon, D. L. (1999). *Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees* (Vol. 16).

Li, S. Y., Pearl, D. K., & Doss, H. (2000). Phylogenetic tree construction using Markov chain Monte Carlo. *Journal of the American Statistical Association, 95*(450), 493-508. doi:Doi 10.2307/2669394

Li, Y., Liu, R., Zhou, T., & Fan, Z. (2013). Genetic diversity and population structure of Sugarcane mosaic virus. *Virus Res, 171*(1), 242-246. doi:10.1016/j.virusres.2012.10.024

Luo, Q., Ahmad, K., Fu, H. Y., Wang, J. D., Chen, R. K., & Gao, S. J. (2016). Genetic diversity and population structure ofSorghum mosaic virusinfectingSaccharumspp. hybrids. *Annals of Applied Biology, 169*(3), 398-407. doi:10.1111/aab.12310

Mar, J. C., Harlow, T. J., & Ragan, M. A. (2005). Bayesian and maximum likelihood phylogenetic analyses of protein sequence data under relative branch-length differences and model violation. *BMC Evol Biol, 5*, 8. doi:10.1186/1471-2148-5-8

Martin, D., Posada, D., Crandall, K. A., & Williamson, C. (2005). A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res Hum Retroviruses, 21*(1), 98-102. doi:10.1089/aid.2005.21.98

Martin, D., & Rybicki, E. (2000). RDP: detection of recombination amongst aligned sequences. *Bioinformatics, 16*(6), 562-563.

Martin, D. P., Murrell, B., Golden, M., Khoosal, A., & Muhire, B. (2015). RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol, 1*(1), vev003. doi:10.1093/ve/vev003

Maynard Smith, J. (1992). Analyzing the mosaic structure of genes. *J Mol Evol 34*, 126-129.

Mbewe, W. D., S. (2017). *Comparative evolvability of plant viruses*. Paper presented at the Microbe New Orleans, LA

McLeish, M. J., Fraile, A., & Garcia-Arenal, F. (2018). Ecological Complexity in Plant Virus Host Range Evolution. *Advances in Virus Research, Vol 91: Control of Plant Virus Diseases Vegetatively-Propagated Crops, 101*, 293-339. doi:10.1016/bs.aivir.2018.02.009

Miller, M. A., Pfeiffer, W., & Schwartz, T. (2010, 14-14 Nov. 2010). *Creating the CIPRES Science Gateway for inference of large phylogenetic trees.* Paper presented at the 2010 Gateway Computing Environments Workshop (GCE).

Moradi, Z., Nazifi, E., & Mehrvar, M. (2017). Occurrence and Evolutionary Analysis of Coat Protein Gene Sequences of Iranian Isolates of Sugarcane mosaic virus. *Plant Pathol J, 33*(3), 296-306. doi:10.5423/PPJ.OA.10.2016.0219

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

Padhi, A., & Ramu, K. (2011). Genomic evidence of intraspecific recombination in sugarcane mosaic virus. *Virus Genes, 42*(2), 282-285. doi:10.1007/s11262-010-0564-6

Padidam, M., Sawyer, S., & Fauquet, C. M. (1999). Possible emergence of new geminiviruses by frequent recombination. *Virology, 265*(2), 218-225. doi:10.1006/viro.1999.0056

Pagan, I., Firth, C., & Holmes, E. C. (2010). Phylogenetic analysis reveals rapid evolutionary dynamics in the plant RNA virus genus tobamovirus. *J Mol Evol, 71*(4), 298-307. doi:10.1007/s00239-010-9385-4

Pagan, I., & Holmes, E. C. (2010). Long-term evolution of the Luteoviridae: time scale and mode of virus speciation. *J Virol, 84*(12), 6177-6187. doi:10.1128/JVI.02160-09

Paul, H. (2015). Bayesian Data Analysis 3rd edn A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin, 2013 Boca Raton, Chapman and Hall–CRC 676 pp., £44.99 ISBN 1-439-84095-4. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 178*(1), 301-301. doi:doi:10.1111/j.1467-985X.2014.12096_1.x

Pirone, T. P., & Blanc, S. (1996). HELPER-DEPENDENT VECTOR TRANSMISSION OF PLANT VIRUSES. *Annual Review of Phytopathology, 34*(1), 227-247. doi:10.1146/annurev.phyto.34.1.227

Pond, S. L. K., Frost, S. D. W., & Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics, 21*(5), 676-679. doi:10.1093/bioinformatics/bti079

Posada, D., & Crandall, K. A. (2001). Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci U S A, 98*(24), 13757-13762. doi:10.1073/pnas.241370698

Rambaut, A., Lam, T. T., Max Carvalho, L., & Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol, 2*(1), vew007. doi:10.1093/ve/vew007

Revers, F., & Garcia, J. A. (2015). Molecular biology of potyviruses. *Advances in Virus Research, Vol 91: Control of Plant Virus Diseases Vegetatively-Propagated Crops, 92*, 101-199. doi:10.1016/bs.aivir.2014.11.006

Rozas, J., Ferrer-Mata, A., Sanchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., & Sanchez-Gracia, A. (2017). DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Molecular Biology and Evolution, 34*(12), 3299-3302. doi:10.1093/molbev/msx248

Schierup, M. H., & Hein, J. (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics, 156*(2), 879-891.

Shukla, D., M. Strike, P., L. Tracy, S., H. Gough, K., & Ward, C. (1988). *The N and C Termini of the Coat Proteins of Potyviruses Are Surface-located and the N Terminus Contains the Major Virus-specific Epitopes* (Vol. 69).

Shukla, D. D., Frcnkel, M. J., & Ward, C. W. (1991). Structure and function of the potyvirus genome with special reference to the coat protein coding region. *Canadian Journal of Plant Pathology, 13*(2), 178-191. doi:10.1080/07060669109500953

Shukla, D. D., Frenkel, M. J., McKern, N. M., Ward, C. W., Jilka, J., Tosic, M., & Ford, R. E. (1992). Present status of the sugarcane mosaic subgroup of potyviruses. In O. W. Barnett (Ed.), *Potyvirus Taxonomy* (pp. 363-373). Vienna: Springer Vienna.

Simmons, H., Holmes, E., & Stephenson, A. (2008). *Rapid evolutionary dynamics of zucchini yellow mosaic virus* (Vol. 89).

Sumner, J. G., Jarvis, P. D., Fernandez-Sanchez, J., Kaine, B. T., Woodhams, M. D., & Holland, B. R. (2012). Is the general time-reversible model bad for molecular phylogenetics? *Syst Biol, 61*(6), 1069-1074. doi:10.1093/sysbio/sys042

Urcuqui-Inchima, S., Haenni, A.-L., & Bernardi, F. (2001). Potyvirus proteins: a wealth of functions. *Virus Research, 74*(1-2), 157-175. doi:10.1016/s0168-1702(01)00220-9

Uzzell, T., & Corbin, K. W. (1971). Fitting Discrete Probability Distributions to Evolutionary Events. *Science, 172*(3988), 1089.

Wang, Y., & Yang, Z. (2014). Priors in Bayesian phylogenetics. In M.-H. Chen, L. Kuo, & P. O. Lewis (Eds.), *BAYESIAN PHYLOGENETICS Methods, Algorithms, and Applications* (pp. 5-23). Boca Raton, FL: CRC Press, Taylor & Francis Group.

Wangai, A., Redinbaugh, M., M. Kinyua, Z., Miano, D., K. Leley, P., Kasina, M., . . . Jeffers, D. (2012). *First Report of Maize chlorotic mottle virus and Maize Lethal Necrosis in Kenya* (Vol. 96).

Whelan, S., Liò, P., & Goldman, N. (2001). Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends in Genetics, 17*(5), 262-272. doi:10.1016/s0168-9525(01)02272-7

Wu, L., Zu, X., Wang, S., & Chen, Y. (2012). Sugarcane mosaic virus – Long history but still a threat to industry. *Crop Protection, 42*, 74-78. doi:10.1016/j.cropro.2012.07.005

Wylie, S. J., Adams, M., Chalam, C., Kreuze, J., Lopez-Moya, J. J., Ohshima, K., . . . Ictv Report, C. (2017). ICTV Virus Taxonomy Profile: Potyviridae. *J Gen Virol, 98*(3), 352-354. doi:10.1099/jgv.0.000740

Xia, X. H., Xie, Z., Salemi, M., Chen, L., & Wang, Y. (2003). An index of substitution saturation and its application. *Molecular Phylogenetics and Evolution, 26*(1), 1-7. doi:Doi 10.1016/S1055-7903(02)00326-3

Xie, X., Chen, W., Fu, Q., Zhang, P., An, T., Cui, A., & An, D. (2016). Molecular Variability and Distribution of Sugarcane Mosaic Virus in Shanxi, China. *PLoS One, 11*(3), e0151549. doi:10.1371/journal.pone.0151549