Identification and ranking of pervasive secondary structures in positive sense single-stranded ribonucleic acid viral genomes

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor Philosophiae (Bioinformatics) at the South African National Bioinformatics Institute, University of the Western Cape.

Author: Emil Pavlov Tanov

Supervisor: Dr. Gordon W. Harkins

Co-Supervisor: Prof. Alan Christoffels

November 2017

Declaration

I declare that *Identification and ranking of pervasive secondary structures in positive sense single-stranded ribonucleic acid viral genomes* is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Full name..... Emil Pavlov Tanov Date..... 10 November 2017

Signed.....

Abstract

The plasticity of single-stranded viral genomes permits the formation of secondary structures through complementary base-pairing of their component nucleotides. Such structures have been shown to regulate a number of biological processes during the viral life-cycle including, replication, translation, transcription, post-transcriptional editing and genome packaging. However, even randomly generated single-stranded nucleotide sequences have the capacity to form stable secondary structures and therefore, amongst the numerous secondary structures formed in large viral genomes only a few of these elements will likely be biologically relevant. While it is possible to identify functional elements through series of laboratory experiments, this is both excessively resource- and time-intensive, and therefore not always feasible. A more efficient approach involves the use of computational comparative analyses methods to study the signals of molecular evolution that are consistent with selection acting to preserve particular structural elements. In this study, I systematically deploy a collection of computationally-based molecular evolution detection methods to analyse the genomes of viruses belonging to a number of ssRNA viral families (*Alphaflexiviridae, Arteriviridae, Caliciviridae, Closteroviridae, Coronavirinae, Flaviviridae, Luteoviridae, Picornaviridae, Potyviridae, Togaviridae* and *Virgaviridae*), for evidence of selectively stabilised secondary structures. To identify potentially important structural elements the approach incorporates structure prediction data with signals of natural selection, sequence co-evolution and genetic recombination. In addition, auxiliary computational tools were used to; 1) quantitatively rank the identified structures in order of their likely biological importance, 2) plot co-ordinates of structures onto viral genome maps, and 3) visualise individual structures, overlaid with estimates from the molecular evolution analyses. I show that in many of these viruses purifying selection tends to be stronger at sites that are predicted to be base-paired within secondary structures, in addition to strong associations between base-paired sites and those that are complementarily co-evolving. Lastly, I show that in recombinant genomes breakpoint locations are weakly associated with co-ordinates of secondary structures. Collectively, these findings suggest that natural selection acting to maintain potentially functional secondary structures has been a major theme during the evolution of these ssRNA viruses.

Acknowledgments

I am grateful for having met and worked with the following people, who have made invaluable contributions towards the completion of this degree.

Dr. Gordon Harkins for his parent-like patience in supervising this PhD project and never growing tired of repeating yet again to “carry on”. Thanks Gordon.

Dr. Brejnev Muhire for sacrificing time with his beautiful wife and family to help me with the consecutive programming problem while still managing to pretend to be enjoying it.

Prof. Darren Martin for providing most of the ideas for this project and critically reviewing the manuscript. And crucially, keeping a sense of humour when I had lost mine.

Prof. Alan Christoffels for his encouragement to pursue my studies further, backing it up with financial support.

I would also like to extend my gratitude to the Poliomyelitis Research Foundation for providing supplementary funding without which I would have not been able to complete this degree.

And last but not least, my mother, Polya Tanova, for her support and sacrifice throughout this period. This thesis is dedicated to the loving memory of my father, Pavel Tanov, who inspired scientific thinking in our family.

Table of contents

List of tables

List of figures

Chapter 1: Introduction

1.1 Nucleic-acid secondary structure within virus genomes

Like the double-helical regions of DNA, single-stranded ribonucleic- (ssRNA) and deoxyribonucleic-acid (ssDNA) molecules have the potential to fold onto themselves to form higher-order structures that are stabilised by hydrogen bonding interactions of complimentary bases (Watson-Crick A:T and C:G, as well as the slightly less stable T:G wobble-pair). Whereas hydrogen bonds are responsible for maintaining pairing between nucleotides, dispersion forces and hydrophobic interactions are responsible for base stacking (Yakovchuk et al., 2006). Collectively, the sum of pairing interactions between bases, (as well as the inherently resulting single-stranded apical loop regions), are referred to as "secondary structures." Furthermore, the maintenance of these base-pairing interactions places constraints on the overall conformation of RNA molecules in three-dimensional space and is largely responsible for the "tertiary structures" of these molecules.

Naturally occurring secondary structure elements are known to play various regulatory roles in the biological processes of many organisms. For instance, stem-loop structures can mediate processing of the mRNA (Dominski et al., 1999) and during peptide synthesis, secondary structures can facilitate nucleic-acid – amino-acid interactions of transfer RNAs (RajBhandary, 1994; Sperling-Petersen et al., 2002) and ribosomal RNAs (Barciszewska et al., 2001; Brimacombe and Stiege, 1985). The broad and complex spectrum of functional elements formed in nature is in part due to the ability of RNA to explore the free-energy landscape and assume different stable conformations, driven by specific functional requirements. There are a number of physiochemical triggers that can contribute towards conformational changes in RNA structures, such as pH (Cromie et al., 2006), Mg^{2+} concentration (Nechooshtan et al., 2009), Na^{+} concentration (Chen and Znosko, 2013) and proximity to and orientation in relation to specific amino-acids (Mandal et al, 2004), as well as more mechanistic effectors, like variation in physiological temperature (Nocker et al., 2001) and the translation-induced melting of mRNA structures (Watts et al., 2009). However, in some cases high-energy barriers associated with particular conformations have to be overcome before alternative conformations can be assumed by a molecule. Such transitions can be facilitated by a group of proteins, called 'RNA chaperones', that have the

capacity to destabilise and re-anneal RNA strands (Herschlag et al., 1994; Rajkowitsch et al., 2007).

Viral genomes encode a plethora of functional elements involved in various stages of different virus life-cycles, such as, transcription (Li and Broyles, 1993; Modrof et al., 2003), translation (Bushell and Sarnow, 2002) and viral genome packaging (Catalano et al., 1995; Hutchinson et al., 2010). Apart from these functional motifs, the genomes of viruses also encode higher-order structural conformations that could potentially regulate specific genetic and biochemical pathways. In some viruses, such structures have been shown to direct several processes, including replication (Clyde and Harris, 2006), translation (Shen and Miller, 2007) and post-transcriptional RNA editing (Linnstaedt et al., 2009). However, the sheer abundance of stable secondary structures, along with their conformational alternatives, means that the actual importance of many of these structures remains unknown. Nevertheless, the extensive structure that is present in many viral genomes provides a foundation from which to further explore the role of these elements in shaping the evolutionary patterns of these genomes.

1.1.1 The impacts of genomic secondary structures on virus evolution

Viral genomes are some of the fastest evolving entities in biology, largely because of their short replication cycle and high fecundity (Duffy et al., 2008). The process of evolution in viruses is driven by several mechanisms, including: random mutation (Drake, 1999), recombination (Lai, 1992; Sztuba-Solińska et al., 2011), reassortment (Vijaykrishna et al., 2015), insertion and deletion (Cheynier et al., 2001), and gene amplification (Roth and Andersson, 2012). While these processes introduce genetic variation and molecular changes into genomes, very few of these changes will be advantageous and as a result natural selection will frequently act against the survival of genomes where changes occur in genetic elements such as genes, regulatory motifs or secondary structures that provide a fitness advantage to the organism: a phenomenon known as, negative selection (sometimes called purifying selection) (Charlesworth et al., 1993). Subsequently, it is expected that, given the high mutation and recombination frequencies in viruses, any widely conserved secondary structures that can be found in any contemporary viral genomes, will probably have been selectively maintained.

1.1.1.1 Mutation

The rate at which mutations in viruses arise may, in part, be influenced by the extent of secondary structure elements present in their genomes (Muhire et al., 2014). Analysis of viral genome sequences in evolution experiments indicate that ssDNA is more susceptible to oxidative damage, than double-stranded DNA (dsDNA) regions (Xia and Yuen, 2005). Although, there are many base modifications caused by oxidation, some of the more common types are the deamination of cytosine to form uracil and the conversion of guanine to 8-oxo-guanine, enabling it to base-pair with alanine. In addition, there is experimental evidence in vitro that the rate of cytosine deamination is strongly dependant on DNA secondary structure being notably slower (>100 fold) in double-stranded DNA as compared to single-stranded DNA (Frederico et al., 1990).

Mutations which arise in a functional secondary structural element might also be strongly favoured against by selection if the mutation impacts the stability of the element. Extremely low synonymous substitution rates which are consistent with strong purifying selection acting at the nucleotide level to prevent even substitutions that have no impact on encoded amino acid sequences, have been observed in the well-conserved and highly structured rev response element (RRE) region in the HIV-1 genome (Ngandu et al., 2008). Evidence exists, suggesting that natural selection may act to maintain some secondary structures within viral genomes (Cloete et al., 2014; Rubio et al., 2013). In a viral population of maize streak virus, an introduced mutation which disrupted a particular structure, spontaneously reverted at extremely high frequencies, restoring the initial structural conformation (Shepherd et al., 2006).

1.1.1.2 Genetic recombination

In addition to mutation, recombination is a significant contributor to the accumulation of genetic diversity within and among viral species. This is because it can involve the shuffling of large tracts of genome sequence enabling viruses to rapidly access greater areas of sequence space than is possible by the stepwise accumulation of random point mutations alone (Domingo and Holland, 1997). As a result, genetic recombination has the capacity to

accelerate both the rate at which advantageous mutations become fixed within, and maladaptive mutations are purged from, viral populations (Moya et al., 2000).

As is the case with point mutations, genomic secondary structures can potentially have two distinct impacts on the recombination dynamics of RNA viruses; firstly, it can demarcate preferred locations of recombination breakpoints, and secondly, the disruption of biologically important structures through recombination, can significantly reduce the fitness of recombinants and thus their survival probabilities. Recombination analysis in ssDNA geminiviruses has shown, that selection disfavours recombinants with disrupted intra-genomic interactions that include co-evolved protein-protein and protein-nucleotide, and nucleotide-nucleotide interactions (Martin et al., 2011).

In some RNA viruses genomic secondary structures have been shown to clearly play a role in directing genetic recombination such that it is far more likely to occur at certain genomic sites than others (Galetto et al., 2004). For example, in the study by Galetto et al., (2004) a hairpin structure located on the C2 portion of the gp120 envelope gene of HIV-1 was implicated with the formation of a recombination hot-spot at the loop of this structure. By varying the stability of the hairpin without altering its sequence, these authors showed that they could significantly alter recombination patterns occurring in the envelope gene. Notably, RNA viruses lacking a stable version of this hairpin, displayed drastically reduced rates of recombination in that region compared with sequences in which the hairpin loop was present. It has subsequently been discovered that recombination breakpoints arising during HIV replication have a very strong tendency to occur at paired nucleotide sites within genomic secondary structures (Simon-Loriere et al., 2010). To explain this phenomenon it has been hypothesised that stem-loop structures within the HIV genome, that are the sites of clearly defined recombination hot-spots, might promote template switching during the reverse transcription phase of the HIV life-cycle (Simon-Loriere et al. 2010). Additionally, the fact that inter-protein linkers and inter-domain sites within proteins are enriched in RNA secondary structures, might be an evolved mechanism that, besides facilitating the proper folding of HIV polyproteins during translation, might also ensure that recombinant HIV genomes will tend to express proteins where either entire proteins or entire sub-protein domains are inherited from the same parental virus (Simon-Loriere et al. 2010).

1.1.2 Computational prediction of nucleic-acid secondary structures

Computational methods of predicting RNA secondary structures can be generally classified into three categories; those using thermodynamic approaches, those using probabilistic approaches, and those using hybrid weighted thermodynamic+probabilistic approaches. Although there is no strong consensus in the field as to which is best, the thermodynamic approaches that compute minimum (or close to minimum) free energy structures of RNA molecules through dynamic programming are the most widely used (Zuker and Stiegler, 1981).

Given an isolated RNA sequence, thermodynamic methods tend to primarily focus on the inference of either the structure with the lowest estimated free energy, or computing samples of structures with folds that are near that of the structure with the minimum free energy folds (Zuker, 1989). Variations of this approach also exist for finding samples of structures within a prescribed energy range (Wuchty et al., 1999).

Non-deterministic probabilistic folding algorithms (Flamm et al., 2000) can produce ensembles of structures by repeatedly refolding sequences from randomly determined starting points. A much more elegant and efficient approach is the computation of the complete matrix of base pairing probabilities, which contains suitably weighted information about all possible secondary structures and therefore reduces the impact of model inaccuracies and over-simplifications on the final predicted structure (Semegni et al., 2011).

However, prediction algorithms based on thermodynamic and/or probabilistic models are not able to detect a certain group of structures called pseudoknots. In a pseudoknot, nucleotides within a loop of one stem-loop structure form base pairs with nucleotides outside the stem-loop structure. In recent years numerous biologically important examples of pseudoknots have been discovered. However, pseudoknots violate the simplified assumption made by most current energy minimisation approaches that all secondary structures will be perfectly nested within one another (Andronescu et al., 2010). Due to the massive numbers of potential pseudoknots that might occur within any given folded nucleic acid, these structures can be very difficult to infer computationally and as a result none of the most frequently used RNA/DNA secondary structure prediction methods even attempt to account for their occurrence.

1.1.3 Experimental secondary structure prediction

Experimental methods for the prediction of secondary structures exist that, unlike the *in silico* computational methods, provide a means of directly probing the structures of RNA molecules *in vivo*. Such approaches involve using experimentally derived data to determine whether individual nucleotides are folded. These include chemical probing methods including among others selective 2'-hydroxyl acylation analysed by primer extension (SHAPE; Wilkinson et al., 2006) and dimethyl sulphate based methods (Tijerina et al., 2007).

SHAPE is a popular approach that is able to determine whether nucleotides in the RNA sequence are base-paired or unpaired, from adduct formation of chemical reagents on individual nucleotide bases. Modifications are identified as stops during a primer extension reaction with reverse transcriptase, and are compared to the results obtained from an unmodified control to yield an accurate biophysical measurement of the base-pairing dynamics within an RNA molecule. Sites which are constrained due to base-pairing, display low SHAPE reactivity, whereas unpaired sites show more adduct formation and thus, high SHAPE reactivity. This SHAPE reactivity data can be used as an experimental correction to folding prediction algorithms to improve the accuracy of the predicted RNA structure (Deigan et al., 2009).

The SHAPE method has been employed to decode the global secondary structure of several single-stranded viruses with RNA genomes, including those of human immunodeficiency virus (HIV; Watts et al., 2009), hepatitis C virus (HCV; Mauger et al., 2015) and satellite tobacco mosaic virus (SMTV; Athavale et al., 2013). In addition, SHAPE reactivities have also been used to validate the results of other biophysical prediction techniques, including nuclear magnetic resonance (Xue et al., 2016) and cryo-electron microscopy approaches (Garmann et al., 2015).

Although the SHAPE method can determine the extent of base-pairing between specific nucleotides with relative accuracy, it does not reveal the actual pairing partners. While a SHAPE-directed thermodynamic prediction approach can contribute to improved average structure inference accuracy (depending on the structure) over an exclusively computational minimum free energy prediction, it has been criticised for not being able to unambiguously distinguish between the pairing states of the analysed bases (Sükösd et al., 2013).

An alternative chemical probing method has been recently proposed that can unambiguously distinguish between the pairing states of the analysed bases. This method, called RNA Proximity Ligation (RPL; Ramani et al., 2015), initially performs RNase digestion followed by re-ligation with T4 RNA Ligase I, forming chimeric molecules of RNA sequences that were initially involved in base-pairing interactions. Subsequent deep sequencing of the resulting fragments, as well as quantification of the relative abundance of specific intramolecular ligation junctions, provide pairwise contact maps that reflect the short- and long-range stem-loop and pseudoknot interactions of the RNA secondary structure.

Whereas these chemical and biophysical probing methods offer an improvement in secondary structure prediction accuracy, site-directed mutagenesis assays have proved to be a valuable tool in validating predicted base-pairing interactions (Burrill et al., 2013; Crary et al., 2003; Stewart et al., 2016). Such experiments provide the means to target specific sites of the nucleic-acid molecule by introducing changes that disrupt the stability of predicted secondary structures. The effect of the introduced mutations on the fitness (i.e. infectivity, replication, transmission etc.) of the mutant can then be compared with the relative characteristics of the wild-type strain for any significant differences that could highlight the biological importance of the proposed base-pairing interactions or secondary structures. Examples of functional secondary structures that have been successfully identified through such experiments (i.e. coupling mutagenesis assays with thermodynamic predictions) include; stem-loops involved in the replication cycle of ebola virus (Crary et al., 2003) and poliovirus (Burrill et al., 2013), and secondary structures in the HCV genome, involved in genome packaging (Stewart et al., 2016).

Despite the obvious advances in both computational and experimental methods of secondary structure prediction, the accurate determination of these structures remains a complex problem and, as yet, no single method is sensitive enough to provide a complete solution. Nevertheless, when results of the current prediction methods are used in tandem, they can provide a much clearer picture of the base-pairing interactions present within RNA molecules.

1.2 Positive-sense single-stranded RNA viruses

The single-stranded RNA (ssRNA) viruses can be divided into several groups: those with positive-sense genomes [(+)ssRNA], negative-sense genomes [(-)ssRNA], and retro-transcribing genomes (ssRNA-rt). This study will analyse selected viral genomes belonging to the (+)ssRNA group which, according to the latest taxonomic proposal by the International Committee on Virus Taxonomy (ICTV; Adams et al., 2017), are classified into three orders; the Nidovirales, Picornavirales and Tymovirales, comprised of 33 families, of which 20 cannot currently be assigned to any order.

Generally, the (+)ssRNA viruses have relatively small linear genomes ranging from ~ 4.8kb - ~32kb. Tombusviruses have some of the smallest and coronaviruses have the largest genomes in the group (Smith et al, 2013). Viruses in this group have wide-ranging host-specificity, infecting diverse forms of life including bacteria (Biebricher, 2008; Schindler, 1964), fungi (Wang et al., 2017; Zhang et al., 2014), plants (Gibbs et al., 2008; Revers and García, 2015), and animals (Choi and Chae, 2002; Goens, 2002; Salguero et al., 2005).

Despite displaying a high degree of structural diversity, the genomes of eukaryote-infecting ssRNA viruses share many common characteristics. In terms of their replication strategy, (+)ssRNA viruses initiate infection with the translation of the genomic RNA to produce the viral replicase and transcriptase proteins. These enzymes synthesise negative strand templates, amplify positive sense strands, and, in some cases, produce subgenomic RNA. More specifically, following the translation of the genomic RNA (serving as messenger RNA; mRNA) by the host cellular machinery, the RNA-dependant RNA polymerase (RdRp) transcribes viral genomic plus-strands into complimentary minus-strand RNA, forming double-stranded RNA (dsRNA) intermediates (Uchil and Satchidanandam, 2003). The newly produced minus-strands of the dsRNA molecules serve as templates to generate multiple copies of plus-strand RNA. The replication process then proceeds in an asymmetric fashion reflected by the accumulation of greater concentrations of plus-strands over minus-strands. Although the degree to which the molar ratio of plus-strands to minus-strands varies between different groups of viruses, it is a characteristic trait of the replication process of all RNA viruses (Buck, 1996). For example, the ratio of positive- to negative-strands produced in flaviviruses is ~10:1 (Chambers et al., 1990), ~50:1 in coronaviruses (van Marle et al., 1995), ~100:1 in brome mosaic virus (French and Ahlquist, 1987) and ~40:1 in citrus tristeza virus (Satyanarayana et al., 2002).

The major differences in the replication strategies of the (+)ssRNA viruses involve the mechanisms used for the synthesis and translation of mRNAs. The picornaviruses and caliciviruses, for example, initiate RNA synthesis via a primer dependant mechanism (Ahlquist et al., 2003). The 5'-end of these viral genomes is covalently linked to a terminal protein called VPg (virion protein genome-linked). The viral RdRp is expressed as a precursor protein that does not have polymerase activity, but rather functions as a protease performing specific proteolytic cleavages of the polyprotein encoded by the viral genome. The cleaved RdRp catalyses the initial attachment of VPg to the genome which then serves as the protein primer for RNA elongation.

By contrast, flaviviruses have a methylated nucleotide cap structure at the 5'-end of their genomes and therefore translation can be initiated de novo by RdRp binding at the 3'-end of the positive-strand genomic RNA and an entire genome is copied without dissociation. In turn, the produced negative-strand serves as a template to synthesize copies of the plus-strand genomic RNA (van Dijk et al., 2004).

In the potyvirus plant-infecting group on the other hand, the process of replication resembles the mechanism employed by the picornaviruses. The members of this family have a VPg protein, as in the picornavirus genomes, that is covalently linked to the 5'-end of the genome. The RdRp is recruited to the genome via an interaction with VPg and VPg is uridylyted. Replication of the minus-strand is then initiated at the 3'-end of the plus-strand genomic RNA (Wang et al., 2000).

Some (+)ssRNA viruses, such as members of the *Togaviridae* and *Arteriviridae* families, generate subgenomic RNAs that specify additional open reading frames to express structural and accessory proteins; however, in all cases the translation of the genomic RNA produces a RdRp molecule that is capable of copying the genome.

In addition, there are a number of conserved features and elements shared amongst many of the (+)ssRNA viral genomes. These include a partially conserved set of genes consisting of the RdRp, a chymotrypsin-like protease, called 3C, a superfamily 3 helicase (S3H) and VPg. Widely conserved protein sequences, such as those of the RdRps, have been primarily used in the past to aid in the phylogenetic classification of these viruses (Koonin, 1991).

1.2.1 (+)ssRNA viruses as disease causing agents

Viruses offer a useful experimental model to study the dynamics of evolution due to their relatively small size and rapid mutation rates, but more importantly they are also responsible for causing numerous diseases, in a wide range of life forms. Many of the viruses of belonging to the (+)ssRNA group cause major diseases in humans, animals and plants, with significant socio-economic impacts around the world.

Historically, there is evidence of poliovirus outbreaks going back thousands of years (Daniel and Robbins, 1997), as well as the well documented devastating epidemics of the late 19th- and early 20th-centuries in industrialised Europe and North America (Landsteiner, K Popper, 1909; Nolan et al., 1955). Poliovirus is a member of the *Picornaviridae* family of viruses and is the causative agent of poliomyelitis.

More recent global outbreaks of (+)ssRNA viruses include the on-going zika virus disease epidemic in the Americas, caused by the zika virus (Ali et al., 2017; Ikejezie et al., 2017), member of the *Flaviviridae* family, and the severe acute respiratory syndrome epidemic in the early 2000's in Asia (Lee, 2005), caused by SARS-coronavirus (SARS-CoV), which is a member of the *Coronaviridae* family of viruses.

It is also notable that, the (+)ssRNA plant-infecting viruses contribute significantly to losses in important food crops worldwide. The *Potyviridae* virus family is the second largest family of plant-infecting viruses after the *Geminiviridae* family (single-stranded DNA viruses) (Adams et al., 2017). The *Potyviridae* family is comprised of many agriculturally, economically and biologically important viral species. For example, maize dwarf mosaic- and sugarcane mosaic-potyviruses are distributed worldwide and are the cause of some of the most serious losses in maize crops globally (Meyer and Pataky, 2010), while the sweet potato feathery mottle potyvirus is one of the sweet potato-infecting viruses causing the most significant losses of this crop in Africa and India (Gibson et al., 2004). The *Luteoviridae* family is another group of (+)ssRNA viruses that cause severe disease in food crops worldwide. Barley yellow dwarf virus is a member of this group, causing yellow dwarf disease in cereals, and alongside wheat dwarf virus of the *Geminiviridae* family (Gauthier et al., 2017), it is globally regarded as the most important pathogen in wheat.

Since many of these viruses share common characteristics important for their continued survival, including mechanisms of replication, protein coding regions and structural features, identifying and characterising conserved structural elements can contribute to ongoing prevention and treatment strategies.

1.2.2 Characterised secondary structure elements in (+)ssRNA viruses

Considering the substantial biomedical and economic impacts of (+)ssRNA viruses around the world, it is not surprising that many of these viruses are well sampled and have been extensively analysed for the presence of functional secondary structures.

A common structure present in the 5' untranslated genome region (UTR) of all members of the *Picornaviridae* family of viruses, termed 'internal ribosome entry site' (IRES), is implicated with recruitment of ribosomal subunits to initiate translation (Martínez-Salas et al., 2015; Pelletier and Sonenberg, 1988). While the overall structure and sequence of the IRES varies amongst the different picornavirus genera it remains absolutely critical for the initiation of translation (Belsham, 2009). IRES elements have also been identified in other (+)ssRNA viruses including HCV (Honda et al., 1996), members of the *Dicistroviridae* family of viruses (Wilson et al., 2000), human immunodeficiency virus 2 (Herbreteau et al., 2005) and classical swine fever virus (Kolupaeva et al., 2000).

In the flaviviruses, a structural domain at the 5' UTR, termed 'stem-loop A' (SLA), directs the attachment of the RdRp during the replication of the viral RNA (Filomatori et al., 2006). Furthermore, it has been shown that the structure guides the addition of the 5' cap-structure required for binding of the eukaryotic initiation factors (Zhang et al., 2008). The sequence and folding conformation of SLA is also largely conserved across the flaviviruses (Lodeiro et al, 2009).

In the tobacco etch virus of the *Potyviridae* family, secondary structure elements forming in the capsid protein coding region as well as the adjacent 3' UTR, have been identified to play a role in recruiting the RdRp replication complex (Li et al., 1997) as well as regulating genome amplification (Haldeman-Cahill et al., 1998).

In coronaviruses belonging to the *Coronaviridae* family, secondary structures together with 'slippery' nucleotide sequences facilitate ribosomal frameshifting during translation. These

signals enable the ribosome to skip a stop codon at the N-terminus of the ORF1a gene and switch to the ORF1b reading frame so that extended ORF1a/b products are translated (Brierley and Dos Ramos, 2006; Giedroc and Cornish, 2009). In SARS-CoV, this 'slippery. region is characterised by a pseudoknotted element causing ribosomes to pause and shift into the ORF1b reading frame and proceed with translation of the ORF1a/b polypeptide complex (Plant and Dinman, 2008).

Collectively, these studies indicate that many (+)ssRNA viruses have evolved to incorporate structural elements in their genomes that complement and regulate various mechanisms crucial to their survival. However, there are still numerous uncharacterised secondary structures present in (+)ssRNA viral genomes, and therefore identifying and ordering those elements that are most likely to be biologically relevant remains an important consideration.

1.2.3 Evolution of (+)ssRNA viruses

Viruses are among the fastest evolving biological entities in the world. Single stranded RNA viruses evolve at a rate in the order of 10^{-3} substitutions/site/year (Hanada et al., 2004; Sanjuan et al., 2010). This is, approximately six orders of magnitude (i.e. one million times) more rapid than the rates of evolution of cellular host organisms, that are typically around 10^{-9} substitutions/site/year (Kumar and Subramanian, 2002). RNA viruses display very high short term rates of mutation (~10^{-4} and 10^{-6} mutations per nucleotide site per replication cycle; Duffy et al., 2008; Sanjuan et al., 2010) several fold greater than those of most DNA-based life forms with, for example, the estimated mutation rate of the human genome being only ~1.1 x 10^{-8} mutations per nucleotide site per replication cycle (Nachman and Crowell, 2000).

The high mutation rates of RNA viruses could be partly explained by the inherently error-prone RdRp mediated RNA synthesis mechanism employed by these viruses, and partly attributed to their short replicative cycles, the large numbers of replicons that are generated, and the selective proliferation of genomes containing adaptive mutations that, for example, mask these genomes from host immune systems. It is likely that most of the mutations arising in RNA virus genomes are at least slightly maladaptive (i.e. they decrease the fitness of the genomes in which they occur). Besides non-neutral mutations within genes that alter encoded amino acid sequences, other mutations that might impact fitness include those that disrupt

functional secondary structure interactions or regulatory motifs within viral genomes. In fact, replication error- and mutation in most RNA viruses are likely near the maximum that are compatible with viral viability (Holmes, 2003).

Recombination is another process that can rapidly generate genetic variability among ssRNA viral genomes. When this process occurs between viruses with a single genome segment, it is simply referred to as recombination, whereas the exchange of genomic segments between viruses with multi-segment genomes is termed reassortment. Recombination has been shown to readily occur *in vivo* in both; non-segmented genome ssRNA viruses including HCV (Reiter et al., 2011), mouse hepatitis virus (MHV; Baric et al., 1990) and poliovirus (Duggal et al., 1997), and in viruses with segmented genomes such as, *cucumber mosaic virus* (CMV; Pita and Roossinck, 2013) and brome mosaic virus (BMV; Bruyere et al., 2000; Urbanowicz et al., 2005). However, actual recombination frequencies occurring in nature may be underappreciated as a result of natural selection acting to remove hybrid genomes within which segments of sequence that are derived from different parents do not function optimally with one another. A positive correlation between mutation rates and the rates of recombination have been detected in several ssRNA viruses (Tromas et al., 2014), suggesting that high recombination frequencies are associated with selection favouring fast-paced error-prone replication. Although recombination has the potential to frequently yield defective genotypes, especially when the exchange is between distantly related species (Martin et al., 2005), it can sometimes provide major adaptive advantages during viral evolution. For example, it has been associated with host range expansion in plant- (Gibbs and Weiller, 1999; Glasa et al., 2005; Schoelz and Wintermantel, 1993) and animal-infecting viruses (Brown, 1997), the evasion of host immunity (Malim and Emerman, 2001), the evolution of anti-viral drug resistance (Nora et al., 2007), and increases in virulence (Khatchikian et al., 1989).

1.3 Thesis structure

In this study, I will perform a range of *in silico* analyses using a suite of computational methods and tools to identify and characterise predicted secondary structures within the genomes of 51 pathogenic (+)ssRNA virus species from a wide-selection of viral families (*Alphaflexiviridae, Arteriviridae, Caliciviridae, Closteroviridae, Coronavirinae, Flaviviridae, Luteoviridae, Picornaviridae, Potyviridae*, *Togaviridae* and *Virgaviridae*). In chapter two, I use a computational secondary structure prediction method to identify structural elements that are present within the viral genomes and employ several methods of

estimating selection to test for significant evidence of natural selection acting to conserve the predicted structures. Chapter three presents a ranking of the structures, based on the results obtained in chapter two, in order of their likely biological relevance and potentially important structures within several viral families are characterised and graphically presented. Chapter four investigates to what extent recombination patterns may be influenced by the predicted secondary structures of the investigated viral genomes.

Chapter 2: Effect of secondary structure on the evolutionary patterns of single-stranded RNA virus genomes

2.1 Abstract

Single-stranded RNA (ssRNA) viral genomes have the potential to form intricate secondary structures through Watson-Crick base-pairing between their component nucleotides. While several secondary structures formed through this type of interaction have been identified to play important biological roles in the replication cycles of numerous ssRNA viruses, many potentially functional structures remain 'hidden in plain sight'. It is, however, possible to use computational techniques to identify evolutionarily preserved, and hence likely functional, genomic secondary structures in virus genomes wherever full genome sequence data is available in public sequence databases for a large enough number of sufficiently diverse virus isolates. In this large scale comparative analysis, I identify the most conserved secondary structures within the genomes of several vertebrate- and plant-infecting ssRNA viruses, representing the families; *Alphaflexiviridae, Arteriviridae, Caliciviridae, Closteroviridae, Coronavirinae, Flaviviridae, Luteoviridae, Picornaviridae, Potyviridae, Togaviridae* and *Virgaviridae*. Furthermore, in an attempt to demonstrate that the predicted structures are likely to exist in the viral populations, I probe these genomes for evidence of natural selection acting to maintain the stability of the identified structures. In a large proportion of the genomes analysed here, there was strong evidence of both, 1) greater degrees of purifying selection acting on sites predicted to be paired than at sites predicted to be unpaired, and 2) lower than expected synonymous substitution rates at codons predicted to be base-paired than codons that were unpaired. In a smaller proportion of datasets, there was also significant evidence of complimentarily coevolving nucleotides that are also predicted to be base paired. Collectively, these results indicate that natural selection has, in part, selectively optimised the

stability of many of the structures predicted here, and therefore that, in at least some of these viruses, these structures are providing a fitness advantage.

2.2 Introduction

Apart from encoding proteins, viral genomes present a scaffold of conserved functional and regulatory motifs contributing to, amongst other processes, binding of transcription factors (Hou et al., 1994; Muller et al., 1988), translation (Poulin and Sonenberg, 2000), replication (Gunawardene et al., 2015) and genome packaging (Stewart et al., 2016). Among these regulatory motifs are secondary structural elements that are formed as a result of thermodynamically stable complimentary base-pairing, in single-stranded nucleic acid molecules.

In ssRNA viruses, biologically functional secondary structures tend to be well conserved, even amongst distantly related species. These include, the cis-acting replication elements (CREs) of flaviviruses (Tuplin et al., 2011), coronaviruses (Raman et al., 2003) and picornaviruses (Steil and Barton, 2009), the internal ribosome entry sites of pestiviruses (Fletcher and Jackson, 2002) and cap-independent translation elements (CITEs) of potyviruses (Carrington and Freed, 1990) and other plant-infecting ssRNA viruses (Kneller et al., 2006).

Although it is possible, through laboratory experiments, to determine whether particular secondary structures are of functional importance, a more efficient approach involves computational examination of the data for systematic evolutionary patterns in groups of related viral genome sequences. Besides identifying the most conserved nucleotide pairs that would form stable structures, selection acting to maintain the structural integrity of biologically relevant structures should display a signal that is easily detectable by computational methods. For example, it is expected that bases predicted to be paired within the coding regions should display decreased synonymous substitution rates (Gerber et al., 2001; Ngandu et al., 2008; Tuplin et al., 2004). Similarly, a greater degree of purifying selection would be expected at sites predicted to be base-paired (Muhire et al., 2014). Lastly, a greater frequency of complementarily co-evolving pairs of nucleotides should be detected within the paired regions of secondary structures (Pace et al., 1999).

Despite the fact that there are numerous functionally important structures documented in ssRNA viral genomes, many viruses might yet harbour biologically relevant elements that are still to be identified. In this chapter, I attempt to obtain evidence that such potentially important secondary structures do indeed exist in the genomes of the viruses tested here. I analysed a large number of full genome datasets for significant differences in the evolutionary patterns of their structured and non-structured regions using various computational methods. More specifically, I compare predicted paired and unpaired regions for significant differences in 1) the degree of purifying selection across the genome and 2) synonymous substitution rate estimates in coding regions. In addition, I analysed the datasets for evidence of complimentary coevolution and tested whether there was an association between pairs of nucleotides detected to be coevolving and those predicted to be base-paired.

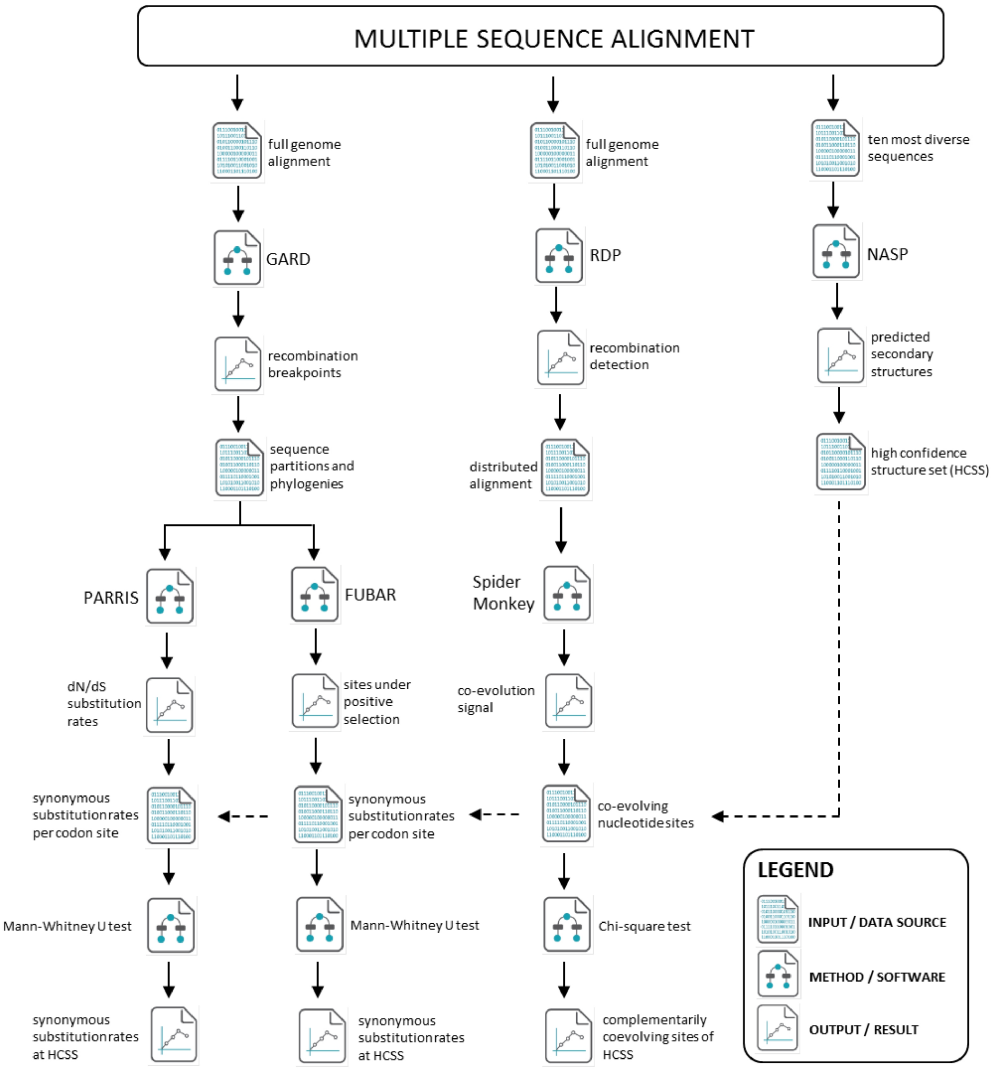


Figure 2.1 | Pipeline of molecular selection analyses

Program workflow of methods used to determine whether predicted secondary structure imposes selection constraints across the viral genomes tested. The dashed line indicates that the NASP high confidence structure set results were used in conjunction with the corresponding synonymous substitution rate and coevolution results to perform the statistical tests.

2.3 Materials and Methods

2.3.1 Sequence dataset preparation

Single-stranded positive sense RNA viral genome datasets were assembled from sequences obtained from the National Center for Biotechnology Information (NCBI) nucleotide database (http://www.ncbi.nlm.nih.gov/nuccore) during October, 2013. Sequence length query was set to cover at least 70% of the total length of the genome for each viral species.

The MAFFT webserver (Katoh and Standley, 2013), was used to align each dataset and the resulting alignments were further manually edited where necessary, using MEGA version 6.0.5 (Tamura et al., 2013). Additionally, IMPALE version 1.28 (http://www.cbio.uct.ac.za/~arjun/IMPALE_V1.28.zip) was used to further improve poorly aligned regions. Fifty one 'large' datasets were assembled, each consisting of between 22, and 1548 full genome sequences, ranging from ~ 4.4kb to ~29kb in length (Table 1). These 'large' datasets were used to calculate Tajima's *D* and Fu and Li's *F* statistics, detect genome-wide nucleotide coevolution, as well as detect patterns of recombination in each of the viral species. From every large-sized dataset, one representative sequence of the ten most divergent lineages were selected to form a set of 'small' datasets. These small datasets were used to computationally predict the coordinates of evolutionarily conserved secondary structures across the viral genomes.

Table 2.1 | Viral genome datasets used in this study

Family name and genus	Species and dataset name	Size[a]	Length[b]	Diversity[c]
Alphaflexiviridae				
Potexvirus	Pepino mosaic virus (PepMV)	75	6413	0.257
	Potato virus X (PVX)	34	6436	0.146
Arteriviridae				
Arterivirus	Porcine reproductive and respiratory syndrome virus (PRRSV)	261	15 538	0.054
Caliciviridae				
Norovirus	Norwalk virus (NV)	473	7575	0.051
Lagovirus	Rabbit haemorrhagic disease virus (RHDV)	46	7467	0.154
Vesivirus	Feline calicivirus (FCV)	23	7701	0.086

Table 2.1 Continued

Family name and genus	Species and dataset name	Size[a]	Length[b]	Diversity[c]
Closteroviridae				
Closterovirus	Citrus tristeza virus (CTV)	44	19 255	0.151
Coronaviridae				
Alphacoronavirus	Human coronavirus NL63 (HCoV-NL63)	124	29 840	0.008
Flaviviridae				
Flavivirus	Dengue virus Type 1 (DENV T1)	1548	11 299	0.056
	Dengue virus Type 2 (DENV T2)	1154	11 492	0.073
	Dengue virus Type 3 (DENV T3)	862	10 827	0.042
	Dengue virus Type 4 (DENV T4)	150	10 687	0.057
	Japanese encephalitis virus (JEV)	166	10 988	0.081
	West Nile virus (WNV)	702	11 029	0.252
	Tick-borne encephalitis virus (TBEV)	93	10 835	0.119
	Yellow fever virus (YFV)	62	10 823	0.102
Hepacivirus	Hepatitis C virus (HCV)	1294	9 365	0.1
Pestivirus	Bovine viral diarrhea virus Type 1 (BVDV-1)	43	12 267	0.177
	Bovine viral diarrhea virus Type 2 (BVDV-2)	29	12 285	0.094
	Classical swine fever virus (CSFV)	70	12 311	0.115
Hepevirus	Hepatitis E virus (HEV)	196	7212	0.291
Luteoviridae				
Luteovirus	Barley yellow dwarf virus (BYDV)	64	5827	0.19
Polerovirus	Potato leafroll virus (PLRV)	29	5987	0.256
	Sugarcane yellow leaf (ScYLV)	30	5612	0.13
Picornaviridae				
Aphthovirus	Foot-and-mouth disease virus (FMDV)	402	9720	0.115
Enterovirus	Human enterovirus A (HEV-A)	467	8455	4.024
	Human enterovirus B (HEV-B)	191	8352	0.98
	Human enterovirus C (HEV-C)	417	8730	0.281
Rhinovirus	Human rhinovirus A (HRV-A)	184	8240	0.31
	Human rhinovirus B (HRV-B)	67	8057	0.321
	Human rhinovirus C (HRV-C)	62	8185	0.43
Hepatitis A	Hepatitis A virus (HAV)	84	7890	0.125
Cardiovirus	Encephalomyocarditis virus (EMCV)	52	7460	0.084

Table 2.1 Continued

Family name and genus	Species and dataset name	Size[a]	Length[b]	Diversity[c]
Teschovirus	Porcine teschovirus (PTV)	57	7853	0.211
	Saffold virus (SAFV)	60	7382	0.343
Parechovirus	Human parechovirus (HPeV)	83	8344	0.085
Duck Hep A	Duck hepatitis A virus (DHAV)	110	7278	0.069
Potyviridae				
Potyvirus	Potato virus Y (PVY)	238	9773	0.067
	Soybean mosaic virus (SMV)	42	7710	0.039
	Turnip mosaic virus (TuMV)	219	9791	0.13
	Zucchini yellow mosaic virus (ZYMV)	55	9593	0.037
	Bean common mosaic virus (BCMV)	42	10086	0.356
	Bean yellow mosaic virus (BYMV)	40	9532	0.245
	Papaya ringspot virus (PRSV)	29	10326	0.058
	Plum pox virus (PPV)	105	9791	0.228
	Watermelon mosaic virus (WMV)	34	10041	0.146
	Sugarcane mosaic virus (SCMV)	22	9596	0.085
Togaviridae				
Rubivirus	Rubella virus (RUBV)	34	9875	0.228
Virgaviridae				
Tobamovirus	Cucumber green mottle mosaic virus (CGMMV)	32	6422	0.085
	Tobacco mosaic virus (TMV)	71	6395	0.029
Not assigned				
Sobemovirus	Rice yellow mottle virus (RYMV)	35	4457	0.078

[a] Number of sequences in dataset
[b] Longest genome in dataset
[c] Median pairwise-diversity (estimated using: https://indra.mullins.microbiol.washington.edu/DIVEIN).

For the estimation of synonymous substitution rates at codon sites, known genes from each of the large dataset alignments were separated and 104 'gene' datasets were assembled. For a number of well-sampled viral groups (DENV T1, DENV T2 and HCV), there were >1000 sequences publically available. Due to the computationally demanding nature of the methods used to estimate the synonymous substitution rates, the size of the excessively large

alignments were reduced to a representative set of no more than 500 sequences (sequences sharing more than ~95% nucleotide identity discarded as were those that were poorly aligned).

2.3.2 *In silico* prediction of evolutionarily conserved secondary structures

The computer program Nucleic Acid Structure Predictor (NASP; Semegni et al., 2011) was used to predict evolutionarily conserved base-pairing interactions within the small viral genome alignments. This method uses the hybrid-ss-min module of the UNAFold package (Markham and Zuker, 2008) to compute the over-all Gibbs free energy of a single-stranded nucleic acid sequence, yielding a list of Boltzmann probabilities of individual potential base pairings, the minimum free folding energy of each sequence in each dataset, and a consensus structure.

More specifically, NASP constructs a consensus pairing matrix, called the M matrix, by merging the pairing matrices for each sequence of the small dataset, by weighted summation. NASP uses a weighted sum method to calculate its M matrix, in order to avoid similar structures from closely related sequences contributing unfairly to the conservation score it calculates for each of the identified structures. The algorithm minimises the false positive prediction rate by applying recursive permutations to determine a set of substructures that at least 95% of the time make the overall structure more stable than a number of randomly generated alignments (100 permuted alignments were used here) with the same nucleotide composition. Collectively, these stable secondary structures comprise the 'high-confidence structure set' (HCSS), used for subsequent analyses. Paired nucleotides comprising the HCSS were analysed for association with signals of selection favouring the maintenance of these structures.

In order to mimic physiological conditions as closely as possible, the NASP parameters were set to fold sequences as linear RNA, and annealing temperature was set to 37^{o}C and 22^{o}C for vertebrate and plant viruses, respectively. Sodium and magnesium concentrations were set at 1M and 0M for both vertebrate and plant infecting viruses.

2.3.3 Tajima's *D* and Fu and Li's *F* neutrality tests comparing purifying selection at paired- and unpaired-sites

Structural elements that contribute to the fitness of the viral genome are expected to be preserved by selection acting to maintain such biologically functional nucleotide interactions. More specifically, in neutrality tests, such as Tajima's *D* (Tajima, 1989) and Fu and Li's *F* (Fu and Li, 1993) tests, it is expected that sites predicted to be base-paired will display stronger evidence of negative selection than those sites predicted to be unpaired.

Tajima's test is one of most widely used statistical tests of evolution neutrality at the sequence level. Tajima's D is a summary statistic that measures the difference between two estimates of genetic diversity: the mean number of pairwise differences and the number of segregating sites. These estimates are expected to be equal when evolution is neutral (i.e. there is no evidence of natural selection), resulting in a Tajima's D value of zero. A nonzero value for D implies departure from neutral evolution, where at least one condition for neutral evolution is violated. Sequences evolving under purifying selection are expected to have an excess of low frequency polymorphisms compared to those evolving under neutral selection and are, therefore, expected to yield negative values for Tajima's D statistic, whereas positive Tajima's D values would signify low levels of both, low and high frequency polymorphisms, possibly indicating positive selection.

In many ways Fu and Li's F test statistic shares much information with Tajima's D statistic, as it is based on the differences between the total number of mutations in external branches of the genealogy, and the average number of nucleotide differences between pairs of sequences.

Similarly to Tajima's test a zero F value would indicate neutral evolution, whereas a negative value signifies negative selection and a positive value positive selection. This test is more sensitive than Tajima's D statistic under the action of purifying selection (Fu and Li, 1993).

Since, in many of the datasets, there were fewer sites predicted to base-paired than sites predicted to be unpaired, a permutation test was used, involving the random selection of identical numbers of paired and unpaired-sites and the comparison of summary selection statistics between these paired- and unpaired-site datasets. Initially, a profile alignment of the small and large datasets was performed and coordinates of base-paired sites (HCSS coordinates) were mapped to the resulting profile alignment. It was then possible to split the profile alignment into 'paired'- (made of columns comprising sites within the HCSS

predicted to be paired) and 'unpaired-alignments' (comprising of columns of the HCSS predicted to be unpaired). The paired-alignments were retained, while 100 datasets were permuted, each consisting of n sites (n equal the number of sites in the paired alignment), randomly sampled from sites in the unpaired alignment. To compare the strength of negative selection, Tajima's D and Fu & Li's F statistics were computed for every set of paired- and permuted unpaired-alignments as implemented in EGGLib (Evolutionary Genetics and Genomics Library; De Mita and Siol, 2012) The p-value was computed as the proportion of times the Tajima or Fu and Li statistics for the paired alignment were lower than those of each of the 100 simulated datasets.

2.3.4 Codon-specific selection inference using *HyPhy* - an open- source platform for likelihood-based molecular evolution analysis

Hypothesis testing using Phylogenies (HyPhy; Kosakovsky Pond et al., 2005) is an open-source high level programming language providing a flexible and unified platform for analysing rates and patterns of sequence evolution of comparative genomic datasets, employing phylogenetic, machine learning and likelihood-based techniques. In addition to the broad array of default molecular analysis tools available within the HyPhy software package, the HyPhy batch language (HBL) allows for the customisation of existing analyses, methods and parameters. Furthermore, HyPhy is designed to be implemented on parallel computing environments, allowing it to efficiently process large and complex datasets. For this reason, all of the codon-specific analyses in this study employed the HyPhy platform and were executed via the HBL interface installed on parallel computing clusters at the South African National Bioinformatics Institute at the University of the Western Cape, the CSIR Centre for High Performance Computing, and the ICTS High Performance Computing Cluster at the University of Cape Town.

2.3.4.1 Screening the codon alignments for evidence of recombination prior to selection analysis

Although genetic recombination frequencies amongst (+)ssRNA viral families can vary dramatically, it has been shown to be a relatively common process in both vertebrate (Simon-Loriere and Holmes, 2011) and plant (Bujarski, 2013) infecting viral species.

Recombinant sequences will likely display patterns of variation inconsistent with the rest of the alignment, owing to the different evolutionary histories of the exchanged fragments, and this may produce misleading and incongruent results in many phylogeny-based methods for sequence analysis (Anisimova et al., 2003; Shriner et al., 2003). Instead of discarding recombinant sequences or assuming a single phylogenetic history for the entire alignment, it is possible to identify non-recombinant portions of the alignment in order to infer the correct phylogeny for each region.

The Genetic Algorithm for Recombination Detection (GARD; Kosakovsky Pond et al., 2006), part of the standard HyPhy package library, is designed to detect regions in the alignment that present evidence of phylogenies divergent from other segments. The method searches all possible locations for breakpoints in the alignment, inferring phylogenies of predicted non-recombinant fragments, and assesses goodness of fit by an information-based criterion, such as small sample Akaike Information Criterion (AIC), derived from a maximum likelihood model fit to each segment. The information from all fitted models is combined and assigned a level of support to the placement of break points as well as the different phylogenies among inferred non-recombinant segments. Furthermore, GARD has been shown to perform equally well screening alignments of both, high and low diversity and is not limited by the size of the alignment, such as several Markov Chain Monte Carlo-based methods (Husmeier and McGuire, 2002; Minin et al., 2005; Suchard et al., 2002).

Prior to testing the large datasets for evidence of genetic recombination, non-coding regions of the full genome datasets were discarded and 104 codon aligned gene datasets were constructed (Table 2.3). These codon alignments were screened using the GARD method and the resulting NEXUS files, containing the coordinates of the detected recombinant-free partitions along with their corresponding nucleotide sequence alignments and inferred phylogenies, served as input for both the estimation of synonymous substitution rates and the inference of complimentary coevolution (Figure 2.1).

2.3.4.2 Estimation of synonymous substitution rates in protein coding regions of the genome

It is expected that sequences comprising biologically functional secondary structures that also happen to fall within the coding regions of genomes might evolve in a way that reflects two

distinct layers of selection: (1) selection at the codon level favouring the preservation of amino acid sequences and (2) selection at the nucleotide sequence level favouring the maintenance of base pairing within the secondary structures. It is expected that these two distinct layers of selection would be reflected in variations in substitution rates at codon sites that contain nucleotides participating in base pairing interactions within biologically important secondary structures. Specifically, synonymous substitution rates at codon sites that contain nucleotides predicted to be paired are anticipated to be lower than at codons that are not involved in base-pairing interactions.

(a) PARRIS

The PARtitioning approach for Robust Inference of Selection (PARRIS; Scheffler et al., 2006) method was applied, to the mostly recombination-free gene nucleotide sequences inferred using the GARD method. PARRIS used each partition and tree in order to estimate synonymous substitution rates at individual codon sites.

Here, PARRIS uses a MG94 61x61 codon substitution matrix (Muse and Gaut 1994) and dual time-reversible model of evolution allowing independent rate distributions for both synonymous and non-synonymous rates. Since many (+)ssRNA viruses are known to be highly recombinogenic (Simon-Loriere and Holmes, 2011) the synonymous substitution rate parameter for each codon site in the alignment was obtained by allowing site-to-site variation, which accounted for undetected recombination events. PARRIS uses the recombination breakpoints detected by the GARD algorithm to partition the coding sequence alignments into segments, which are assumed to contain no further evidence of recombination. For each partition, individual tree topology and branch lengths were used, to avoid detection of recombination-induced false-positive signals of synonymous substitution.

(b) FUBAR

The Fast Unconstrained Bayesian AppRoximation for Inferring Selection (FUBAR; Murrell et al., 2013) method is a maximum likelihood method that identifies sites in protein coding alignments which may be under positive or purifying selection. FUBAR estimates synonymous and non-synonymous substitution rates at each codon site, where their ratio (ω) is a measure of the type of natural selection acting on the coding region. Values of ω of less

than 1 represent purifying selection, ω values greater than 1 signify diversifying selection, and values for ω not significantly different from 1 represent neutral evolution. A Markov chain Monte Carlo sampling procedure yields a set of samples where the site-specific posterior distribution of ω can be inferred by an empirical Bayesian procedure, thus identifying sites which are constrained and those which are evolving adaptively. Of interest here though is the utility of FUBAR to quantify rates of synonymous substitution at individual codon sites: rates which are anticipated to be lower within codon sites that contain nucleotides that are base-paired within secondary structures than in codons that contain only unpaired nucleotides.

2.3.4.3 Detection of genome wide co-evolution

The SPIDERMONKEY HyPhy algorithm (Poon et al., 2008) was used to detect whether pairs of sites within the alignment are evolving complimentarily (i.e. a substitution at a particular base is associated with a complementary substitution at some other site). The method extends a simple model of nucleotide substitution - HKY85 which employs a 4 X 4 transition matrix - to a model for independently evolving pairs of nucleotides using a 16 X 16 transition matrix, where the elements in the matrix represent the probability of changing from one pair of nucleotides to another (Muse, 1995). Every pair of nucleotides is compared against the Muse-modified model and an independent sites model – HKY85 - to test for evidence of coevolution using a likelihood ratio test (LRT).

When pairing in sites is favoured to maintain secondary structure within stem regions, the frequency with which these sites will change to unpaired state is expected to be lower than those predicted by the independent sites model. Similarly, in regions where pairing is favoured, the probability of change from an unpaired state to a paired state should be greater than the corresponding probability when sites are independent. A pairing parameter, λ, is introduced in order to capture these features. Rates which are considered to form pairing (Watson-Crick pairing, AT or CG) are multiplied by the pairing parameter, and those that cause changes from paired to unpaired state are multiplied by 1/ λ. Pairs of nucleotides which evolve complementarily should favour a $\lambda > 1$ when fitting the model to nucleotide sequence alignment and corresponding tree. By setting λ to 1 in the Muse-modified model, it is possible to perform a LRT comparing the maximum likelihood estimates (MLEs) obtained for each model. It is expected that the Muse model should produce significantly higher-

likelihood score than HKY85, where paired nucleotides seem to be coevolving. Where the rates of change to paired or unpaired states are largely similar, similar likelihood scores should be obtained for the two models. It is possible to obtain a p-value from the LRT, indicating whether the Muse paired character model fits significantly better to the data. P-values below 0.05 indicate that the Muse model provides a better fit than HKY85, thus implying evidence of complementary coevolution for the pairs of sites examined.

2.3.4.4 Improvement of co-evolution analysis based on recombination detection

Since recombination can affect the coevolution analysis in much the same way it is able to undermine the accuracy of the selection analysis, recombination had to be accounted for in our method. Specifically, if nucleotide coevolution detected from potentially recombined fragments is regarded to have occurred during the same evolutionary event, it may lead to a false positive signal. Each large dataset was analysed for recombination, using the Recombination Detection Program 4.16 (RDP; Martin et al., 2015), identifying recombinants and their corresponding parental sequences. All of the recombinant sequences were split into separate segments at the breakpoints identified by RDP and were appended to the rest of the alignment file as individual sequences, creating a 'recombination-free alignment'. A 100nt sliding window was moved a single nucleotide at a time along the recombination-free alignment, and at every step, the longest (where the length is interpreted as the least number of gap characters contained within that window) N sequences (N is the number of sequences in the original sequence alignment) from each window, were added to a separate alignment file. The resulting alignments were then used to infer the maximum likelihood trees using PHYlogenetic estimation using Maximum Likelihood 3.0 (PhyML; Guindon et al., 2010). Lastly, the SPIDERMONKEY algorithm was executed in the HyPhy environment for each separated alignment file along with its corresponding tree.

2.3.5 Test for association of sites predicted to be paired and having lower than expected synonymous substitution rates

The test uses the estimated synonymous substitution rates provided by the PARRIS and FUBAR methods for gene alignments that exclude regions of overlapping open reading

frames. It determines whether rates of synonymous substitutions (dS; i.e. nucleotide substitutions that do not change the encoded amino acid) at nucleotide sites that are predicted to be base-paired are significantly lower than those at nucleotide sites predicted to be unpaired. PARRIS and FUBAR are codon-based methods and therefore the individual nucleotides within each of the codons tested are assigned the same estimate. Codon positions have different evolutionary constraints, and while changes in the first and second codon positions generally cause changes in the encoded amino acid (i.e. they are non-synonymous), the third position is the least functionally constrained and most substitutions at this position do not cause changes in the encoded amino acid (i.e. they are synonymous). For this reason, only codons in which the third codon position was predicted to be base-paired were considered 'paired' for the purposes of the statistical test. A non-parametric Mann-Whitney U test was used to test whether synonymous substitution rates are lower than expected at sites predicted to be base-paired.

2.3.6 Testing for association between co-evolving sites and base-paired sites within structural elements

Complimentary coevolution was detected by comparing each site against every other site within a 100 nucleotide window, shifting it a single nucleotide at a time across the recombination-free alignments, invariant pairs of nucleotides were ignored. While a sliding window larger than 100nt could reveal distantly located co-evolving nucleotides in some of the larger structures predicted by NASP, the computational resources available for this study did not allow for this possibility.

In some of the datasets (HEV-B, HEV-C, ZYMV, CTV) many recombination events were detected, resulting in a large number of recombination-free sub-alignments (along with their associated phylogenetic trees) that had to be computationally analysed. Furthermore, the large size and length of some of the sequence alignments increase the computational complexity to a point where it is not feasible to analyse the data on a local implementation. This problem was addressed by a script that identifies the created sub-alignments and corresponding trees based on their indexed names, generating an array of submission scripts used to execute the complimentary coevolution analysis on a high-performance computer cluster. Lastly, a statistical test for association between coevolution and base-paring was performed within R (R Development Core Team, 2016) which implements a chi-squared test

of sites predicted to be paired versus those predicted to be unpaired and sites predicted to be coevolving versus sites predicted not to be coevolving.

2.4 Results

2.4.1 ssRNA viruses exhibit extensive genomic secondary structure

Computationally predicted, genome-wide secondary structure profiles were prepared for 51 datasets representing the families *Alphaflexiviridae, Arteriviridae, Caliciviridae, Closteroviridae, Coronavirinae, Flaviviridae, Luteoviridae, Picornaviridae, Potyviridae, Togaviridae* and *Virgaviridae* (Table 2.1). Computational secondary structure prediction revealed an abundance of structural elements across the selected genomes. The primary stage of the NASP prediction method, using UNAfold, initially identified between 362 (BYDV, 5.7kb) and 1000 (CTV, 19.3kb) potentially conserved structural elements in each of the analysed datasets. In order to focus the analysis on structures that are most likely to be biologically relevant, NASP performs a permutation test, where the program identifies those elements that have greater structural stability than structures generated by randomly shuffled sequences of identical nucleotide composition. This permutation analysis generates a subset of secondary structures that is referred to as the HCSS (high-confidence structure set). Despite the HCSS being a reduced subset of detected secondary structures, in 39/51 datasets there were at least 100 such conserved structural elements present, and in some datasets substantially higher numbers of conserved structures were detected. More specifically, in the HCV, FMDV, and FCV datasets, 450, 312, and 270 conserved structures were identified, respectively. Additionally, relatively high numbers of high-confidence structures were identified in the PepMV, DENV T4, TbEV, and BCMV datasets, where 242, 235, 232, and 224 elements were detected, respectively. In all subsequent analyses, only nucleotides predicted to form nucleotide bonds within the HCSS were regarded as being ‘paired’, and those referred to as ‘unpaired’ are bases predicted to be single-stranded.

2.4.2 Neutrality tests for elevated negative selection at nucleotide sites predicted to be paired

It is expected that nucleotides involved in base-pairing interactions would evolve under a greater degree of negative selection when compared to unpaired sites. It is also expected, that sequences evolving under negative selection will have reduced frequencies of mutations relative to those evolving neutrally, and therefore are expected to result in negative values for Tajima's D and Fu and Li's F site-frequency summary statistics. If purifying selection was indeed stronger at sites predicted to be paired, it would be expected that *D* and *F* statistics would produce lower values for datasets of only base-paired nucleotides than datasets containing only unpaired sites.

In 25/51 datasets, both the Tajima's D and Li and Fu's F statistic permutation tests showed evidence of paired sites within structured regions experiencing significantly stronger (P value ≤ 0.05) purifying selection than sites predicted to be unpaired. In a further 6/51 cases, either the D or F statistic test yielded significant evidence (P value ≤ 0.05) of paired sites experiencing stronger purifying selection than unpaired sites. In only 20/51 datasets, no significant evidence was detected of paired sites experiencing significantly stronger purifying selection than unpaired sites (Table 2.2).

Table 2.2 | Tajima's D and Fu and Li's F statistics for paired and unpaired genomic sites

Data set	Tajima's *D*			Fu and Li's *F*		
	Paired [a]	Permuted unpaired [b]	P value	Paired [c]	Permuted unpaired [d]	P value
Norwalk virus	-1.71	-1.43	<0.01	-3.23	-2.31	<0.01
Rabbit haemorrhagic disease virus	-0.77	-0.71	0.21	0.23	0.35	0.05
Feline calicivirus	0.14	0.21	0.03	1.54	1.56	0.03
Citrus Tristeza Virus	0.43	0.62	<0.01	0.45	0.71	<0.01
Dengue Virus Type 1	-1.02	-1.14	0.8	-2.85	-2.71	0.33
Dengue Virus Type 2	-0.77	-0.75	0.53	-3.26	-0.77	<0.01
Dengue Virus Type 3	-1.24	-1.18	0.08	-2.63	-2.31	0.14
Dengue Virus Type 4	-0.72	-0.93	0.88	-0.38	-0.46	0.58
Japanese Encephalitis Virus	-0.88	-0.88	0.47	-1.74	-1.71	0.06
West Nile Virus	-0.44	-0.47	0.58	1.27	1.27	0.54
Tick-borne Encephalitis Virus	-0.14	-0.14	0.45	-0.44	-0.64	0.96
Yellow Fever Virus	-0.57	-0.62	0.86	1.28	1.26	0.84
Hepatitis C	-0.13	0.01	<0.01	1.14	1.49	<0.01
Bovine viral diarrhea virus Type 1	-0.34	-0.22	<0.01	0.59	0.92	<0.01

Table 2.2 Continued

Data set	Tajima's *D*			Fu and Li's *F*		
	Paired [a]	Permuted unpaired [b]	P value	Paired [c]	Permuted unpaired [d]	P value
Bovine viral diarrhea virus Type 2	-0.26	-0.31	0.68	1.03	0.92	0.93
Classical swine fever virus	0.29	0.51	<0.01	1.59	1.67	<0.01
Hepatitis E	0.53	0.76	<0.01	0.47	0.76	0.01
Barley Yellow Dwarf Virus	0.21	0.47	<0.01	0.88	1.14	0.01
Potato leafroll virus	-0.8	-0.76	0.31	1.16	1.19	0.06
Sugarcane yellow leaf virus	0.11	0.86	0.02	0.78	1.23	0.03
PRRSV	-0.86	-0.89	0.69	1.15	1.15	0.59
Human coronavirus NL63	0.57	0.54	0.59	1.58	1.58	0.45
Foot-and-mouth disease virus	-0.18	-0.11	0.03	1.42	1.44	0.03
Enterovirus A	0.29	0.24	0.9	1.59	1.57	0.9
Enterovirus B	0.16	0.14	0.67	1.55	1.54	0.65
Enterovirus C	0.28	0.28	0.44	1.04	1.25	0.02
Rhinovirus A	0.06	0.14	<0.01	1.51	1.54	<0.01
Rhinovirus B	0.37	0.62	0.03	1.58	1.59	0.23
Rhinovirus C	0.92	0.85	0.91	1.73	1.71	0.91
Hepatitis A	0.32	0.28	0.62	0.34	0.62	0.01
Encephalomyocarditis virus	-1.17	-1.06	0.02	-0.55	-0.38	0.05
Saffold Virus	0.16	0.32	0.01	0.84	1.15	<0.01
Teschovirus	1	1.13	0.05	1.84	1.89	0.05
Parechovirus	0.05	0.18	<0.01	0.75	0.78	0.38
Duck Hepatitis A	-0.58	-0.55	0.08	1.27	1.27	0.48
Potato virus Y	1.45	1.49	0.32	1.99	2.01	0.06
Soybean mosaic virus	-0.59	-0.36	0.05	-0.41	-0.14	0.02
Turnip mosaic virus	0.26	0.48	0.04	1.13	1.66	0.03
Zucchini yellow mosaic virus	0.18	0.35	0.04	1.23	1.57	0.04
Bean common mosaic virus	-0.03	0.09	<0.01	1.47	1.52	<0.01
Bean yellow mosaic virus	0.11	0.51	<0.01	1.52	1.66	<0.01
Papaya ringspot virus	0.71	0.82	0.03	1.73	1.78	0.02
Plum pox virus	0.65	0.83	0.02	1.68	1.75	0.02
Watermelon mosaic virus	0.61	0.61	0.51	1.69	1.69	0.49
Sugarcane mosaic virus	0.77	0.89	0.03	1.61	1.75	0.03
Cucumber grn mottle msc virus	1.12	1.06	0.83	1.85	1.81	0.83
Tobacco mosaic virus	-0.91	-0.61	0.03	0.75	1.26	0.02
Pepino mosaic virus	1.44	1.48	0.25	2.03	2.05	0.23
Potato virus X	-0.13	-0.18	0.8	1.43	1.42	0.8
Rubella virus	-0.33	-0.02	<0.01	1.35	1.47	<0.01
Rice yellow mottle virus	0.24	0.39	0.02	1.49	1.62	0.03

[a] Tajima's *D* for paired-site alignments corresponding to the HCSS.
[b] Average Tajima's *D* for 100 permuted alignments sampled for the unpaired sites.
[c] Fu and Li's *F* for paired-site alignments corresponding to the HCSS.
[d] Average Fu and Li's *F* for 100 permuted alignments sampled for the unpaired sites.

Shaded values highlight datasets with significantly ($P<0.05$) lower site-frequency summary statistic at paired sites, while boxed values indicate datasets that produced marginally insignificant evidence ($0.05 \leq P \leq 0.1$).

In a further five datasets, the tests produced marginally significant evidence ($0.05 \leq P \leq 0.1$) of lower D or F values, at paired sites than at unpaired sites (Table 2.2). It is also notable, that a higher proportion of plant infecting viruses (13/18) showed significantly lower *D* and *F* statistics in their paired alignments, compared to vertebrate specific viruses (18/33). In addition, in 16 of the 20 datasets in which a significant difference was not detected, datasets constructed from paired sites still displayed a more negative value, for both *D* and *F* statistics, than datasets sampled from the pool of unpaired sites.

2.4.3 Lower synonymous substitution rates at paired sites might provide evidence of selection acting to conserve structural elements

The FUBAR analysis results were used to compare median synonymous substitution rates of nucleotides within the HCSS versus those predicted to be unpaired. It was evident that in only 29 of the 104 coding regions examined was the median synonymous substitution rate higher in bases predicted to be paired (Figure 2.2). Despite the fact that structure prediction merely provides a snapshot of a dynamic metastable global secondary structure which may have a number of alternate folding pathways depending on physiological conditions, these findings suggest that a large number of the predicted structural elements within coding regions likely do exist and strong selective pressure is acting to preserve the integrity of these structures. Furthermore, it is possible to statistically determine whether there is a significant difference between rate estimates of paired and unpaired codons (Section 2.4.4), where lower synonymous substitution rate estimates imply stronger evolutionary pressure to maintain primary sequence, and therefore preserve secondary structure.

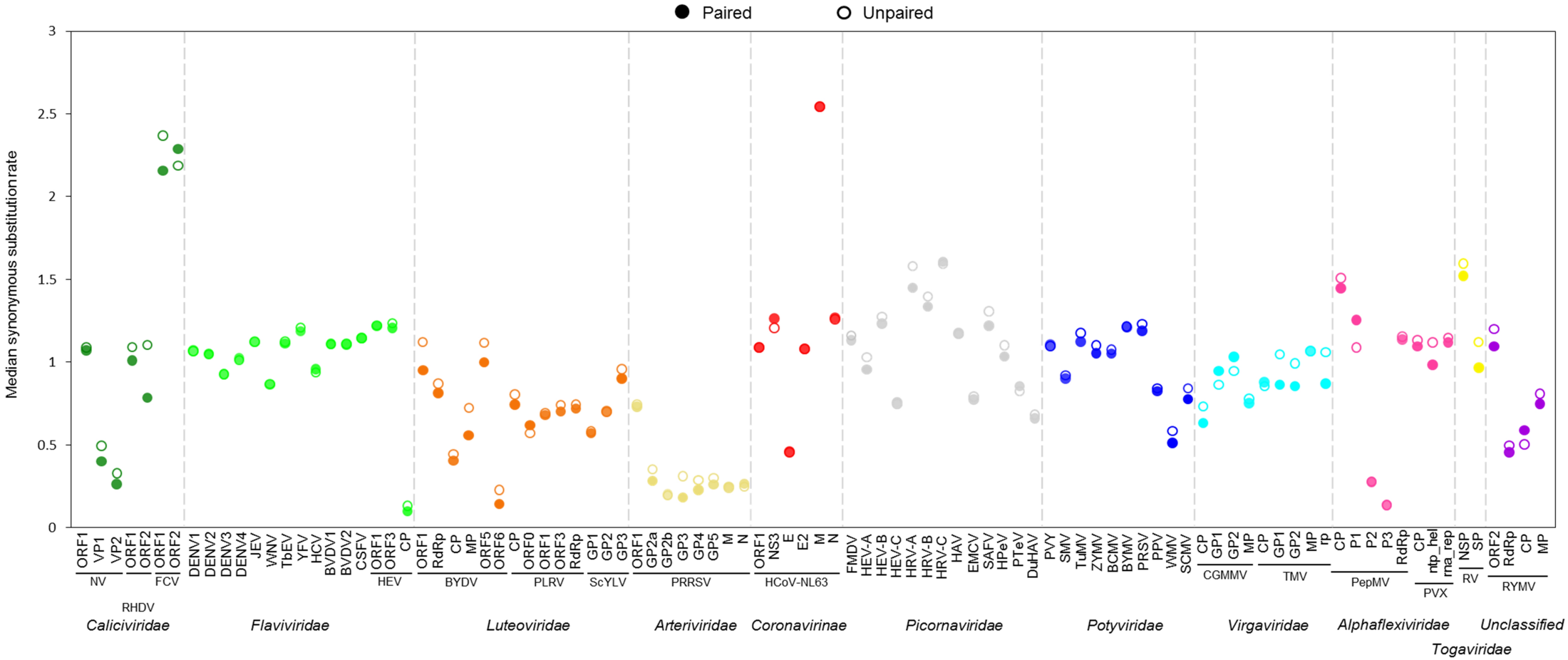


Figure 2.2 | Median synonymous substitution rate estimates of paired versus unpaired codon sites of positive sense ssRNA viruses

The scatter plot shows median synonymous substitution rates for predicted paired and unpaired sites in the coding regions of positive sense ssRNA viruses. Only rate estimates of the 3rd codon positions (i.e. synonymous sites) were considered. Dashed grey lines and colours distinguish viral families. Below the x-axis, coding regions are shown vertically while species and family names, respectively, are shown horizontally. Polyprotein alignments were analysed for the *Flaviviridae* (except HEV), *Picornaviridae* and *Potyviridae* families, therefore only the species name is shown. Rate estimates for paired codons are depicted by solid spheres, while hollow spheres are used to represent codons predicted to be unpaired.

2.4.4 Association of lower than expected synonymous substitution rates at sites predicted to be paired

A total of 107 gene coding alignments were analysed for large scale associations between paired third codon positions and lower than expected synonymous substitution rates (Table 2.3). Synonymous substitution rates of individual codons were estimated using the random effects likelihood selection analysis methods, PARRIS and FUBAR. Low synonymous substitution rates provide evidence of selection acting to preserve the underlying nucleotide sequence. Genera in the *Picornaviridae*, *Calicivirdae* and *Togaviridae* families showed strong associations between base-paired nucleotide positions and lower than expected synonymous substitution rates, which is somewhat expected since these viruses have been shown to harbour a number of substructures that are known to be both evolutionarily conserved and biologically relevant (Cloete et al., 2014; Simmonds et al., 2008; Victoria et al., 2009). Lower than expected synonymous substitution rates at sites predicted to be paired, is likely a consequence of selection acting to maintain base-pairing interactions in biologically functional secondary structures. Similar significant associations between paired third codon positions and lower than expected synonymous substitution rates were also detected in plant viruses of the *Luteoviridae*, *Potyviridae* and *Virgaviridae* families.

Table 2.3 | Association of codons predicted to be paired and lower than expected synonymous substitution rates

Family	Genus	Species	Gene name	PARRIS p value [a]	FUBAR p value [b]	Gene size [c]	Gene Len. [d]	Gene MPD [e]
Calicivirdae	Norovirus	Norwalk Virus	NV_ORF1	0.02	0.03	473	5097	0.051
			NV_VP1	<0.01	<0.01	473	1620	0.054
			NV_VP2	<0.01	0.02	473	804	0.052
	Lagovirus	Rabbit haemorrhagic disease	RHDV_ORF1	<0.01	<0.01	34	7032	0.085
			RHDV_ORF2	0.04	0.12	46	354	0.122
	Vesivirus	Feline calicivirus	FCV_ORF1	<0.01	<0.01	23	5292	0.295
			FCV_ORF2	0.91	0.96	23	2016	0.17
Closteroviridae	Closterovirus	Citrus tristeza virus	CTV_ORF_1a	<0.01	<0.01	43	9441	0.233
			CTV_RdRp	0.35	0.68	42	1449	0.157
			CTV_p6	0.54	0.49	44	153	0.083
			CTV_p13	0.12	0.15	44	357	0.083
			CTV_p18	0.21	0.55	44	501	0.062
			CTV_p20	0.72	0.72	44	549	0.081
			CTV_p23	0.13	0.12	44	630	0.085
			CTV_p25	0.02	0.01	44	669	0.064

Table 2.3 Continued

Family	Genus	Species	Gene name	PARRIS p value [a]	FUBAR p value [b]	Gene size [c]	Gene Len. [d]	Gene MPD [e]
Closteroviridae	Closterovirus	Citrus tristeza virus	CTV_p33	0.59	0.54	44	909	0.134
			CTV_p61	0.81	0.87	44	1608	0.09
			CTV_p65	<0.01	<0.01	44	1782	0.088
Flaviviridae	Flavivirus	Dengue Virus Type 1	DENV_T1_Pol	0.89	0.26	99	10176	0.043
		Dengue Virus Type 2	DENV_T2_Pol	0.88	0.79	100	10176	0.063
		Dengue Virus Type 3	DENV_T3_Pol	0.19	0.03	139	10173	0.012
		Dengue Virus Type 4	DENV_T4_Pol	0.78	0.92	134	10164	0.04
		Japanese Encephalitis Virus	JEV_Pol	0.61	0.31	1 64	10298	0.076
		West Nile Virus (WNV)	WNV_Pol	0.03	0.01	100	10305	0.169
		Tick-borne Encephalitis Virus	TbEV_Pol	<0.01	0.03	92	10239	0.119
		Yellow Fever Virus	YFV_Pol	<0.01	0.01	62	10236	0.102
	Hepacivirus	Hepatitis C	HCV_Pol	0.52	0.66	100	9030	0.052
	Pestivirus	BVDV Type 1	BVDV_T1_Pol	0.67	0.44	42	11718	0.167
		BVDV Type 2	BVDV_T2_Pol	0.88	0.31	28	11739	0.091
		Classical swine fever virus	CSFV_Pol	0.98	0.73	75	11694	0.115
	Hepevirus	Hepatitis E	HEV_ORF1	0.01	0.04	192	5169	0.347
			HEV_CP	<0.01	<0.01	196	1980	0.186
			HEV_ORF3	0.32	0.16	87	369	0.115
Luteoviridae	Luteovirus	Barley Yellow Dwarf Virus	BYDV_ORF1	<0.01	<0.01	64	1022	0.117
			BYDV_RdRp	<0.01	<0.01	64	1589	0.132
			BYDV_CP	0.22	0.17	64	600	0.133
			BYDV_MP	0.05	0.02	64	459	0.091
			BYDV_ORF5	<0.01	0.05	64	1965	0.106
			BYDV_ORF6	0.04	0.04	64	129	0.545
	Polerovirus	Potato leafroll virus	PLRV_CP	<0.01	<0.01	29	624	0.012
			PLRV_ORF0	0.37	0.31	29	741	0.031
			PLRV_ORF1	0.39	0.19	28	1917	0.029
			PLRV_ORF3	0.17	0.17	29	2948	0.026
			PLRV_RdRp	0.04	0.01	29	3192	0.024
		Sugarcane yellow leaf virus	ScYLV_GP1	0.26	0.39	30	768	0.101
			ScYLV_GP2	0.29	0.35	30	3213	0.059
			ScYLV_GP3	0.15	0.12	30	2028	0.06
Arteriviridae	Arterivirus	PRRSV	PRRSV_ORF1a	0.02	0.05	253	7566	0.061
			PRRSV_GP2a	0.35	0.13	256	768	0.042
			PRRSV_GP2b		0.55	257	219	0.035
			PRRSV_GP3	<0.01	<0.01	257	762	0.053
			PRRSV_GP4	0.34	0.21	258	534	0.047
			PRRSV_GP5	0.04	0.03	258	600	0.056
			PRRSV_M	0.68	0.83	258	522	0.025
			PRRSV_N	0.25	0.56	257	369	0.038
Coronavirinae	Alphacoronavirus	Human coronavirus NL63	HCoV-NL63_ORF1a	0.33	0.48	37	12180	0.008

Table 2.3 Continued

Family	Genus	Species	Gene name	PARRIS p value [a]	FUBAR p value [b]	Gene size [c]	Gene Len. [d]	Gene MPD [e]
Coronaviridae	Alphacoronavirus	Human coronavirus NL63	HCoV-NL63_E	0.21	0.48	40	231	0.005
			HCoV-NL63_E2	0.46	0.62	36	4038	0.018
			HCoV-NL63_M	0.02	0.24	40	678	0.005
			HCoV-NL63_N	0.27	0.87	40	1131	0.005
Picornaviridae	FMDV		FMDV_Pol	0.67	0.01	400	6987	0.116
	Enterovirus	A	HEV-A_Pol	0.15	0.62	462	6684	0.221
		B	HEV-B_Pol	<0.01	<0.01	191	6645	0.291
		C	HEV-C_Pol	0.01	0.02	416	6708	0.288
	Rhinovirus	A	HRV-A_Pol	0.56	0.24	143	6600	0.344
		B	HRV-B_Pol	0.02	0.08	53	6585	0.321
		C	HRV-C_Pol	0.54	0.36	53	6582	0.426
	Hepatitis A		HAV_Pol	0.55	0.47	84	6687	0.126
	Cardiovirus	Encephalomyocarditis virus	EMCV_Pol	0.02	<0.01	41	6945	0.074
		Saffold Virus	SAFV_Pol	<0.01	<0.01	60	6951	0.345
	Parechovirus		HPeV_Pol	0.08	0.65	82	6588	0.085
	Teschovirus		PTeV_Pol	<0.01	<0.01	46	6633	0.257
	Duck Hep A		DuHAV_Pol	0.01	<0.01	110	6762	0.114
Potyviridae	Potyvirus	Potato virus Y	PVY_Pol	0.52	0.31	203	9210	0.085
		Soybean mosaic virus	SMV_Pol	0.18	0.04	42	9201	0.039
		Turnip mosaic virus	TuMV_Pol	0.11	<0.01	218	9531	0.153
		Zucchini yellow mosaic virus	ZYMV_Pol	0.62	<0.01	55	9249	0.052
		Bean Common Mosaic Virus	BCMV_Pol	0.12	0.03	42	9666	0.103
		Bean Yellow Mosaic Virus	BYMV_Pol	0.55	0.33	39	9168	0.125
		Papaya Ringspot Virus	PRSV_Pol	<0.01	0.01	29	10056	0.14
		Plum pox virus	PPV_Pol	<0.01	<0.01	105	9441	0.094
		Watermelon mosaic virus	WMV_Pol	0.04	0.01	34	9660	0.081
		Sugarcane mosaic virus	SCMV_Pol	<0.01	<0.01	21	9627	0.211
Virgaviridae	Tobamovirus	Cucumber green mottle mosaic virus	CGMMV_CP	0.01	0.05	27	483	0.037
			CGMMV_GP1	0.41	0.54	27	4941	0.048
			CGMMV_GP2	0.23	0.33	28	3432	0.049
			CGMMV_MP	0.07	0.05	27	792	0.028
		Tobacco mosaic virus	TMV_CP	0.44	0.27	70	477	0.029
			TMV_GP1	<0.01	<0.01	63	4848	0.025
			TMV_GP2	<0.01	<0.01	71	3348	0.031
			TMV_MP	0.13	0.65	70	804	0.029
			TMV_rna_pol	0.02	0.03	71	1422	0.028
Alphaflexiviridae	Potexvirus	Pepino mosaic virus	PepMV_CP	0.03	0.02	75	711	0.179
			PepMV_P1	0.91	0.88	75	702	0.114
			PepMV_P2	0.12	0.36	75	369	0.074
			PepMV_P3	0.04	0.37	75	246	0.053
			PepMV_RdRp	0.08	0.03	68	4317	0.173
		Potato Virus X	PVX_CP	0.56	0.3	32	720	0.017

Table 2.3 Continued

Family	Genus	Species	Gene name	PARRIS p value [a]	FUBAR p value [b]	Gene size [c]	Gene Len. [d]	Gene MPD [e]
Alphaflexiviridae	Potexvirus	Potato Virus X	PVX_rna_rep_prot	0.11	0.05	32	4368	0.154
			PVX_ntp_hel	0.41	0.27	32	678	0.152
Togaviridae	Rubivirus	Rubella virus	RUBV_NSP	<0.01	<0.01	34	6348	0.062
			RUBV_SP	<0.01	<0.01	34	3189	0.064
Not asigned	Sobemovirus	Rice yellow mottle virus	RYMV_ORF2a	0.41	0.46	32	1818	0.055
			RYMV_RdRp	0.51	0.52	33	1512	0.061
			RYMV_CP	0.82	0.82	35	720	0.096
			RYMV_MP	0.62	0.52	35	471	0.121

[a] P value associated with PARRIS based Mann-Whitney U test.

[b] P value associated with FUBAR based Mann-Whitney U test.

[c] Number of sequences in gene alignment used to estimate synonymous substitution rates.

[d] Length of gene alignment in nucleotides.

[e] Median pairwise distance of sequences within the gene alignment.

Shaded values highlight gene datasets with significantly (P<0.05) lower than expected synonymous substitution rates at paired sites, and boxed values marginally insignificant ($0.05 \leq P \leq 0.1$).

Datasets for which the polyprotein coding regions were analysed, are suffixed with the "Pol" contraction, within the gene name category.

Overall, in 40/104 of the gene datasets both of the selection analysis methods showed significantly lower median substitution rates (multiple comparison-corrected Mann-Whitney U test, P value of <0.05) in codons located within structural elements as opposed to codons occupying regions predicted to be unpaired. While in 13 datasets, only one of the two methods presented significant evidence of lower than expected synonymous substitution rates at codons regarded as being paired. Additionally, eight datasets presented marginally insignificant results ($0.05 \leq P \leq 0.1$). It is notable that in only 4/19 plant-infecting species was there no significant evidence in any of their genes, by either of the two methods, that synonymous substitution rates are lower at paired codon sites than at unpaired codons.

2.4.5 Associations between sites predicted to be coevolving and sites predicted to be base-paired

Assuming that the predicted secondary structures are functionally relevant, it is expected that a mutation occurring at a nucleotide site involved in base-pairing would result in a

compensatory change at its corresponding base such that the structural integrity of secondary structure is preserved. Here, the 51 large datasets are analysed to detect pairs of coevolving bases (within a maximum distance of 100nt) and tested for significant associations between bases predicted to be paired and sites detected to be complementarily coevolving. Specifically, for each of the large datasets, a chi-square test of association was performed, where the first category is nucleotide sites predicted to be paired versus unpaired and the other category, sites predicted to be coevolving versus not coevolving.

In 15 of the 51 full genome datasets investigated, strong significant associations (multiple testing-corrected P values < 0.05) were found between constituent nucleotides of secondary structures and sites for which complementary coevolution was detected. In a further 4/51 datasets, only marginally insignificant evidence ($0.05 \leq P \leq 0.1$) was found (Table 2.4).

Table 2.4 Association between paired sites and complementarily coevolving sites

Family	Genus	Species	Chi-square value	*P* Value
Caliciviridae	Norovirus	Norwalk Virus	0.189	0.6
	Lagovirus	Rabbit haemorrhagic disease virus	0.7627	0.38
	Vesivirus	Feline calicivirus	3.4378	0.05
Closteroviridae	Closterovirus	Citrus Tristeza Virus (CTV)	46.8352	<0.01
Flaviviridae	Flavivirus	Dengue Virus Type 1	0.8744	0.07
		Dengue Virus Type 2	0.7172	0.39
		Dengue Virus Type 3	0.2904	0.59
		Dengue Virus Type 4	2.7655	0.09
		Japanese Encephalitis Virus	0.5291	0.46
		West Nile Virus (WNV)	0.5689	0.74
		Tick-borne Encephalitis Virus	0.8749	0.34
		Yellow Fever Virus	1.1784	0.27
	Hepacivirus	Hepatitis C	2.1146	0.62

Table 2.4 Continued

Family	Genus	Species	Chi-square value	*P* Value
Flaviviridae	Pestivirus	BVDV Type 1	14.7824	<0.01
		BVDV Type 2	4.2022	0.04
		Classical swine fever virus	1.5654	0.21
	Hepevirus	Hepatitis E	0.3736	0.54
Luteoviridae	Luteovirus	Barley Yellow Dwarf Virus	5.9392	0.01
	Polerovirus	Potato leafroll virus	3.4954	0.06
		Sugarcane yellow leaf virus	0.3608	0.54
Arteriviridae	Arterivirus	PRRSV	2.4184	0.11
Coronavirinae	Alphacoronavirus	Human coronavirus NL63	2.5874	0.11
Picornaviridae	FMDV		12.3293	0.98
	Enterovirus	A	0.2546	0.13
		B	0.8564	0.06
		C	0.7215	
	Rhinovirus	A	0.5563	0.08
		B	0.7332	0.39
		C	13.1246	<0.01
	Hepatitis A		0.0615	0.81
	Cardiovirus	EMCV	4.0216	0.04
		Saffold Virus	4.9976	0.02
	Parechovirus		3.5548	0.03
	Teschovirus		7.4094	<0.01
	Duck Hep A		8.8961	<0.01
Potyviridae	Potyvirus	Potato virus Y	0.0223	0.88
		Soybean mosaic virus	0.6974	0.41
		Turnip mosaic virus	2.3898	0.12
		Zucchini yellow mosaic virus	0.9099	0.34
		Bean Common Mosaic Virus	0.6929	0.41
		Bean Yellow Mosaic Virus	2.7694	0.09
		Papaya Ringspot Virus	18.2171	<0.01
		Plum pox virus	2.3868	0.16
		Watermelon mosaic virus	0.4338	0.51

Table 2.4 Continued

Family	Genus	Species	Chi-square value	*P* Value
Potyviridae	Potyvirus	Sugarcane mosaic virus	4.8101	0.02
Virgaviridae	Tobamovirus	Cucumber grn mottle msc virus	1.2332	0.27
		Tobacco mosaic virus	19.9027	<0.01
Alphaflexiviridae	Potexvirus	Pepino mosaic virus	2.2671	0.13
		Potato Virus X	0.5226	0.46
Togaviridae	Rubivirus	Rubella virus	8.3694	<0.01
Not assigned	Sobemovirus	Rice yellow mottle virus	0.0224	0.88

Shaded values highlight datasets with significant ($P<0.05$) association between sites detectably coevolving and base-paired nucleotides, whereas boxed values highlight marginally insignificant ($0.05 \leq P \leq 0.1$) associations.

These results provide evidence that in a substantial number of the genomes tested, selection is favouring the maintenance of structures within the HCSS. In ssDNA viral genomes, compelling evidence of nucleotide complementary co-evolution in secondary structure elements exists, implying that these structures are likely to substantially contribute to the fitness of viral genomes (Muhire et al., 2014).

The results reported here suggest that there are a number of specific base-pairs that are being selectively maintained, and therefore many of the predicted structural elements that have not yet been biologically validated may have as yet unconfirmed biological functions.

2.5 Discussion

2.5.1 Extensive secondary structure present in positive-sense ssRNA viral genomes

It is evident, from the secondary structure prediction analysis, that all of the viruses tested here likely possess extensive secondary structure that is frequently conserved among related genomes. In approx. 74% of the datasets, 100 or more conserved structures were detected. As is the case with well characterised structures in some ssRNA viruses, it is plausible that a certain proportion of the unknown elements detected here might serve particular functional roles during discrete stages of the viral life cycle. However, it has been proposed that, as an ordered ensemble, these uncharacterised structures might also collectively have a role in evading innate host defence mechanisms (Simmonds et al., 2004). Strong association was

found between extensive genome-wide secondary structure detected in aphthoviruses, flaviviruses and caliciviruses and the ability of these viruses to establish persistent long-term infections in their hosts (Simmonds et al., 2004). In accordance with those findings, species that are known to cause persistent infections in their hosts (FMDV, HCV, DENV, TbEV and FCV datasets), were amongst those datasets exhibiting the highest distribution of conserved RNA folding across their genomes. Although it is entirely possible that some of the highly structured genomes analysed here have indeed adopted such host-defence evasion mechanisms, I was not however, able to directly test for such a correlation. Instead, the analysis was focused on determining whether there are detectable signals of selection acting in a way that maintains the stability of the structures detected here.

2.5.2 Evidence of higher degree of purifying selection at sites predicted to be paired versus unpaired sites

Nucleotide sites that reside in functionally important secondary structures are expected to evolve under a greater degree of purifying selection than sites that are unpaired. In approximately half of the datasets analysed here, higher degrees of purifying selection was observed in the paired-alignments than in 95% or more of the permuted unpaired-alignments. Similarly, in several plant- and animal-infecting ssDNA viruses, the component nucleotides of the majority of structures predicted to facilitate replication, exhibited significantly lower nucleotide variability (Muhire et al., 2014). This is consistent with the hypothesis that if nucleotides predicted to be base-paired do indeed form part of biologically relevant structures they should display higher degrees of negative selection than sites predicted to be unpaired.

Interestingly, the proportion of datasets, showing strong evidence for paired sites experiencing higher degrees of purifying selection was higher for plant infecting viruses (13/18) than for vertebrate infecting viruses (18/33) (Table 2.2). Although analysing additional plant-based virus datasets could lower the proportion in those datasets, it is conceivable that due to the diversity of cell types and stronger antiviral responses in vertebrate hosts (Kamp et al., 2002), animal RNA viruses are experiencing higher mutation rates in general. It is also plausible that the lower physiological temperatures of plant hosts, relative to animals (with primarily warm-blooded mammal hosts being considered here), may promote greater stability of the structural folds in plant infecting virus genomes.

Collectively, the greater degree of negative selection at paired versus unpaired sites detected in 25/51 datasets, indicates that in a large proportion of the datasets analysed here selection is acting to preserve predicted base-pairing interactions.

2.5.3 Evidence of decreased synonymous substitution rates at sites predicted to be paired

It is expected that if selection is acting to favour the maintenance of biologically important structures, synonymous substitution rates in coding regions should be significantly lower at sites predicted to be paired than at those predicted to be unpaired. Synonymous substitution rates of individual codons were estimated using the random effects likelihood selection analysis methods, PARRIS and FUBAR. In approximately 45% of the gene datasets, both selection detection methods indicated that mean synonymous substitution rates at codons predicted to be involved in base pairing interactions were significantly lower than those codons residing in single-stranded regions. Furthermore, in 13 datasets, at least one of the two methods provided strong evidence of such association. Notably, in only 16 out of the 51 viral species tested, was there no evidence, in any of the gene datasets, that synonymous substitution rates at paired codons are significantly lower than those at unpaired codon sites.

While the distinct difference between synonymous substitution rates (as well as nucleotide polymorphisms) at paired and unpaired sites, indicates that the structures detected here are selectively maintained, it does not, however, confirm that they are indeed biologically relevant. The difference in synonymous substitution rates between single- and double-stranded regions could be partly due to the fact, that single-stranded nucleic acids are more susceptible to post-transcriptional chemical damage, such as oxidative deamination. Cellular metabolites and reactive oxygen species produced by the host, as a response to a viral infection, can induce mutations in the host cell as well as viral genomes. For instance, elevated nucleotide and amino acid mutation rates have been observed in Hepatitis C genomes, following exposure to ethanol in conjunction with virus-induced oxidative stress (Seronello et al., 2011). Similarly, nitrous oxide-induced oxidative stress has been shown to elevate mutation rates in Influenza viral genomes (Akaike, 2001). Another probable explanation for the elevated substitution rates at single stranded regions of animal-infecting viruses, is the activity of the mammalian lethal-mutagenesis immunity mechanism, mediated

by APOBEC3 proteins that preferentially mutate single-stranded viral RNA regions in an attempt to inactivate the virus (Bishop et al., 2004; Willems and Gillet, 2015).

Nevertheless, the results presented here, along with the results of the Tajima's D- and Fu and Li's F-based tests, suggest that at both the nucleotide and the amino acid level, universally conserved structural elements have likely had a significant effect on the evolution of these viral genomes. Since the fitness of many viruses is in part, dependent on the stability and distribution of secondary structures along their genomes (Davis et al., 2008; Simmonds et al., 2004), it is very likely that the predicted structures forming the HCSSs do actually exist and that they may be contributing to the fitness of the viral genomes within which they reside.

2.5.4 Evidence of genome-wide complimentary coevolution at sites predicted to be paired

Nucleotide coevolution of the secondary structures of RNA molecules is well known (Pace et al., 1999). Acquired mutations at nucleotide sites predicted to be involved in base-pairing interactions (within biologically important secondary structures), are expected to be associated with compensatory mutations that maintain the overall conformation and stability of the structure (Muhire et al., 2014). In recent years, the presence of such nucleotide coevolution interaction-networks have become the basis of several methods used to predict conserved secondary structures (Bernhart et al., 2008; De Leonardis et al., 2015), in addition to being employed in improving the accuracy of multiple sequence alignment programs (Nawrocki and Eddy, 2013).

Here, I tested for associations (within a maximum genomic distance of 100nt) between sites predicted to be paired within the HCSSs and pairs of sites that are detected to be co-evolving complimentarily. Specifically, for every dataset, a chi-squared test was performed with sites predicted to be paired versus unpaired sites as the one category, and sites detected to be coevolving versus not coevolving in the other. Only 19 out of the 51 datasets analysed, presented significant evidence consistent with the hypothesis that in order for a mutation at a paired sites to be tolerated, it has to be accompanied by a complimentary mutation in a reciprocal site, such that it maintains base-pairing. In a further four datasets (BYMV, PLRV, DENV T4 and HEV-B), where marginally insignificant associations were detected, the power of the test could be improved as more full-genome sequences become available (BYMV and

PLRV datasets contained 39 and 29 full genome sequences, respectively). Interestingly, out of the 33 datasets for which there was no compelling evidence (P value > 0.05) of association between structured regions and complimentarily coevolving bases, 13 datasets also displayed strong evidence of selection disfavouring synonymous substitutions at base-paired sites within coding regions in the PARRIS and FUBAR based tests (Table 2.5). It is plausible that in some of these datasets, selection acting against substitutions in paired regions constraints sequence variability at these sites, and this results in insufficient signal necessary to detect complementary coevolution interactions.

Although the overwhelming majority of viruses did not present evidence that complimentary coevolution is preferentially occurring at paired sites, it indicates that in at least some of the species analysed here, potentially biologically relevant base-pairing is selectively maintained.

2.6 Conclusions

In this chapter, I found that natural selection appears to be operating in a way that favours the maintenance of base-paired nucleotides across the genomes of several ssRNA viral species. Specifically, I found significant evidence, in 33/51 datasets, of lower than expected synonymous substitution rates at codon sites predicted to be paired, as compared to unpaired sites, in at least one of their coding regions (Table 2.5). In 31/51 datasets I detected evidence of significantly stronger purifying selection acting on bases predicted to be paired than unpaired sites, with at least one of the neutrality tests used (Tajima's D and Fu and Li's F statistics) (Table 2.5). Although the results of the complimentary coevolution tests were not as conclusive, I nevertheless detected evidence in 15/51 datasets, of significant associations between complimentarily coevolving nucleotides and sites predicted to be base-paired. Overall, in ten of the datasets analysed here I found significant evidence, consistent with selection acting to maintain predicted base-pairing, in all three of the tests used (Table 2.5). What makes the evidence of selection favouring the maintainace of base-pairs in these ten datasets even more compelling, is that they showed significant results in all of the methods used, for each category of tests.

However, there are limitations to methods employed here, namely; (1) the secondary structure prediction only takes into account one of several possible metastable global folds which may occur under various physiological conditions, (2) only 100 permutations were used during secondary structure prediction in NASP due to the computationally demanding nature of the algorithm, and (3) because if the low number of full genome sequences available for some of the viral species studied, the synonymous substitution rate estimates and inferences of complimentarily coevolving sites may not be representative of the actual rates in the viral population. These issues could easily be mitigated in future studies as more data becomes available and processing power is more readily available.

Although significant evidence of selection acting to maintain base-pairing, for all three tests, was not detected in every dataset, the results nevertheless provide an indication that conserved secondary structure elements have clearly had an impact on the evolutionary patterns in some of these viruses. Furthermore, it indicates that a large proportion of the structures predicted here likely do exist in the natural populations of these ssRNA viruses.

Table 2.5 | Summary of results in Chapter 2

Viral family	Species	Purifying selection at paired sites [a]	dS at paired sites < dS at unpaired sites [b]	Complimentary coevolution at paired sites [c]
Alphaflexiviridae	Pepino mosaic virus	-	+*	-
	Potato virus X	-	-	-
Arteriviridae	Porcine reproductive and respiratory syndrome virus	-	+	-
Caliciviriade	Norwalk virus	+	+	-
	Rabbit haemorrhagic disease virus	+*	+	-
	Feline calicivirus	+	+	+
Closteroviridae	Citrus tristeza virus	+	+	+
Coronaviridae	Human coronavirus NL63	-	+*	-
Flaviviridae	Dengue Virus Type 1	-	-	-
	Dengue Virus Type 2	+*	-	-
	Dengue Virus Type 3	-	+*	-
	Dengue Virus Type 4	-	-	-
	Japanese encephalitis virus	-	-	-
	West Nile virus	-	+	-
	Tick-borne encephalitis virus	-	+	-
	Yellow Fever Virus	-	+	-
	Hepatitis C virus	+	-	-
	Bovine viral diarrhea virus Type 1	+	-	+
	Bovine viral diarrhea virus Type 2	-	-	+
	Classical swine fever virus	+	-	-
	Hepatitis E virus	+	+	-

Table 2.5 Continued

Viral family	Species	Purifying selection at paired sites [a]	dS at paired sites < dS at unpaired sites [b]	Complimentary coevolution at paired sites [c]
Luteoviridae	Barley yellow dwarf virus	+	+	+
	Potato leafroll virus	-	+	-
	Sugarcane yellow leaf virus	+	-	-
Picornaviridae	Foot-and-mouth disease virus	+	+*	-
	Human enterovirus A	-	-	-
	Human enterovirus B	-	+	-
	Human enterovirus C	+*	+	-
	Human rhinovirus A	+	-	-
	Human rhinovirus B	+*	+*	-
	Human rhinovirus C	-	-	+
	Hepatitis A virus	+*	-	-
	Encephalomyocarditis virus	+	+	+
	Saffold virus	+	+	+
	Human parechovirus	+*	-	+
	Teschovirus	+	+	+
	Duck Hep A	-	+	+
Potyviridae	Potato virus Y	-	-	-
	Soybean mosaic virus	+	+*	-
	Turnip mosaic virus	+	+*	-
	Zucchini yellow mosaic virus	+	+*	-

Table 2.5 Continued

Viral family	Species	Purifying selection at paired sites [a]	dS at paired sites < dS at unpaired sites [b]	Complimentary coevolution at paired sites [c]
Potyviridae	Bean common mosaic virus	+	+*	-
	Bean yellow mosaic virus	+	-	-
	Papaya ringspot virus	+	+	+
	Plum pox virus	+	+	-
	Watermelon mosaic virus	-	+	-
	Sugarcane mosaic virus	+	+	+
Togaviridae	Rubella virus	+	+	+
Virgaviridae	Cucumber green mottle mosaic virus	-	+	-
	Tobacco mosaic virus	+	+	+
Not assigned	Rice yellow mottle virus	+	-	-

[a] Datasets where purifying selection was significantly stronger at sites predicted to be paired than at unpaired sites, based on both *D* and *F* statistics, are designated a "+" sign.

[b] Datasets where at least one of the gene alignments analysed, presented evidence of synonymous substitution rates being significantly lower at paired sites than unpaired sites, based on both PARRIS and FUBAR estimates, are designated a "+" sign.

[c] Datasets where statistically significant associations between sites predicted to be paired and complimentary coevolving sites were detected, are designated a "+" sign.

* Datasets where at least one of the methods presented significant results.

Shaded rows highlight those datasets where all three tests produced significant results.

In this Chapter, evidence was provided that conserved secondary structures within the genomes of some (+)ssRNA viruses appear to be selectively maintained. Based on these results, in Chapter 3, I examine the computationally predicted structures in significant detail in an attempt to identify and characterise potentially functional elements.

Chapter 3: Ranking, visualising and identifying potentially important structural elements

3.1 Abstract

While the NASP secondary structure prediction method provides an objective means of identifying the most evolutionarily conserved base-pairing interactions in diverse viral lineages, it is unable to automatically extrapolate the biological relevance of particular predicted structures. A possible way of circumventing this problem is to overlay the distribution of values obtained from the selection tests, in Chapter 2, across the predicted structures, as a means to quantitatively order those structures having the highest probability of being biologically functional. Specifically, it is possible to identify the most biologically interesting predicted structures by ranking the structures based on: (1) the degree to which they are conserved across the analysed genomes; (2) the degree to which synonymous substitution rates are constrained at codon sites containing nucleotides that are predicted to be base-paired and; (3) the degree to which nucleotides predicted to be base-paired coevolve with one another. Using these rankings, potentially biologically functional structural analogues were identified in several of the viral lineages. Furthermore, well characterised functional structures that have been previously identified in the literature were amongst the highest ranked structures identified. The results here demonstrate that this approach could provide a reliable means of 1) highlighting what are potentially the most interesting structures in (+)ssRNA virus genomes, and 2) proposing what the functional roles of these structures might be.

3.2 Introduction

Given that we detected genome-wide evidence for selection acting to maintain structural elements within several of the datasets, we used the results obtained from the selection analyses to quantitatively rank the detected structures. Secondary structures were ranked according to their likely biological relevance based on three criteria: 1) the degree of conservation detected by the NASP analysis, 2) the degree of constraint of synonymous substitution rates on paired nucleotides within coding regions, and 3) the degree to which paired nucleotides complementarily coevolve with one another. Consensus ranking was

achieved by mapping the scores in the three scoring categories (degrees of conservation indicated by NASP, synonymous substitution rates determined by FUBAR and complimentary coevolution likelihood ratio test p-values determined by SPIDERMONKEY) to the list of predicted NASP structures and choosing the minimum rank of the three scores as the rank for that structure. In the case where two or more structures had the same score, the average of the ranks of the three categories is assigned to the tied structures and placed in ascending order on the rank list. The motivation behind using the minimum rank instead of the weighted average of the results of the three methods is that the relative powers of the three tests performed are not consistent across all structures with the tests being differentially influenced by (1) whether structures are in coding or non-coding regions (the synonymous substitution rate test is restricted to analysing structures in coding regions), (2) the degree of nucleotide conservation within the structures (decreased diversity increases the power of synonymous substitution rate test but decreases the power of the coevolution test), and (3) the number of full genome sequences analysed (the power of the coevolution test is more strongly impacted by the number of sequences analysed than is the power of the synonymous substitution rate test).

3.3 Materials and Methods

3.3.1 Ranking and visualisation of discrete structures

The approach of ranking secondary structures in order of their biological functionality is implemented in DOOSS (Data Overlaid On Secondary Structures; Golden and Martin, 2013). DOOSS ranks structures using a Mann Whitney U test that compares the distribution of data values (either synonymous substitution rates or complementary coevolution p-values) for an individual structure to the distribution of data values of all other structures

To visualise individual elements, DOOSS maps the coordinates of the evolutionarily conserved structures (HCSS data) on the small datasets (used for prediction of the HCSS) and employs VARNA (Visual Applet for RNA; Darty et al., 2009) to graphically present secondary structures and their component nucleotides. It further allows profile alignment of the gene codon-alignment and mapping of corresponding synonymous substitution rates to the structures, so that the inferred rates are overlaid on to the corresponding sites. Similarly,

coevolution data can be mapped on to the structures and discrete complementary coevolution interactions can be visualised.

3.3.2 Ranking based on synonymous substitution constraints on predicted secondary structures

Predicted structures were ranked based on the degree of synonymous substitution constraint at each codon site in a structure of interest, using the Mann-Whitney U test implemented in DOOSS. This method compares the distribution of rates in the structure to the distribution of rates throughout the entire coding region of the genome. The motivation for using this particular test, instead of ranking the predicted structures based on their median associated dS values alone, was that it considers the relative ordering of p-values and accounts for variations in the number of sites found within the different structures that are being compared, hence avoiding bias towards smaller structures consisting of codons with low substitution rates.

Individual structural elements were ranked by comparing the distribution of data values corresponding to the complete list of predicted structural elements for each of the datasets against all data values for the same dataset using a Mann-Whitney U test. DOOSS supports ranking of structures by their one-dimensional data values (e.g. synonymous substitution rates) or their two-dimensional data values (e.g. coevolution p-values) to assist in the identification of structures which are most likely to be biologically functional.

This test generates a z-score which gives an indication of whether a particular structure lies at an extreme of the distribution of all data values (such as substitution rates) being analysed. For example, when considering synonymous substitution rates, a large negative z-score for a particular structural element means that the median synonymous substitution rate for codons within the structural element region is significantly lower than those for most other codons, whereas a z-score close to zero indicates that the structural element does not contain codons with synonymous substitution rates that are significantly different from the rest of the codons in the analysed dataset. Structural elements with high or low associated z-scores are typically the most interesting. Although the p-values obtained by this approach are not statistically accurate, they nevertheless, provide a valuable means of ranking structures based on their likely biological relevance.

3.3.3 Ranking based on paired sites predicted to be co-evolving

Structures were ranked based on the degree of complimentary co-evolving nucleotide pairs they displayed (e.g. an A to G transition at one site coupled with a T to C transition at a second site). Such coevolution may be acting to maintain the shape of secondary structures because these structures are functionally important.

Scores were obtained by comparing the SPIDERMONKEY likelihood ratio test p-values for every set of base-paired nucleotides within a NASP predicted structure, to the list containing all the SPIDERMONKEY LRT derived p-values for predicted base-paired nucleotides within the consensus fold of the genome. Ranking of the structural elements based on p-values was performed using the same test that was used for ranking the structures based on synonymous substitution rates described in section 3.4.1 above.

Structural elements with p-values approaching 0 ($P < 0.05$) were considered to display significant evidence of coevolution between base-paired nucleotides. The z-scores associated with these p-values provided directionality to the p-value, indicating whether the structural elements contained significantly more evidence of complementarily coevolving nucleotide pairs (relatively low z-score), or more evidence of non-complementarily coevolving nucleotide pairs (relatively high z-score).

3.3.4 Consensus ranking of constraints on structural elements

Resulting rank tables from DOOSS were collated into a single file and consensus ranking was achieved by mapping the scores in the three scoring categories (degrees of conservation indicated by NASP, synonymous substitution rates determined by FUBAR and complimentary coevolution likelihood ratio test p-values determined by SPIDERMONKEY) to the list of predicted NASP structures and choosing the minimum rank of the three scores as the rank for that structure. In cases where two or more structures had the same score, the average of the ranks of the three categories was assigned to the tied structures and placed in ascending order on the rank list. As mentioned above, the motivation behind using the minimum rank instead of the weighted average of the three criteria was that the three tests are unevenly influenced by features of the sequences being analysed such that a simple average of the ranks would bias the analysis against detecting, for example, small structures in highly

conserved non-coding regions or large highly conserved structures in diverse parts of the genome where there were too few available full-genome sequences to reliably detect complementary coevolution.

3.3.5 Mapping highest ranked predicted structural elements

A computer program, StructureMap (Muhire et al., 2014), which plots HCCS elements across the full genome map, was used to graphically display the locations of highly ranked structural elements. StructureMap uses the coordinates of the consensus HCSS rank list and represents these by vertical lines drawn along the length of the genome. The program draws all of the predicted locations of structures of the HCSS across the genome. Additionally, a user-defined structure list and colour scheme can be used to demarcate specific structures. For each structure map, the coordinates of protein coding regions were mapped to the graphic based on the gene coordinates of the first sequence in each of the small datasets. Structure maps drawn for different virus species were scaled according to the length of the genomes such that structure maps for related species could be accurately compared for identification of positionally homologous structures in different species. This approach enabled the identification of likely biologically functional structural homologues between some of the 51 different independently analysed datasets.

3.4 Results

3.4.1 Identifying potentially important conserved secondary structures within (+)ssRNA viral genomes

Using the NASP method we detected evidence of extensively conserved base-pairings across the majority of the datasets analysed. Natural selection analysis on coding regions revealed evidence that in a substantial number of viruses, a strong association between lower synonymous substitution rates and predicted paired sites exists, along with significant associations detected between complementarily coevolving nucleotides that are part of conserved base-pairing interactions. These results provided motivation to further explore the locations and arrangement of these structures within the viral genomes. Using the consensus ranking results, structures of the HCSS were graphically mapped and co-ordinates of top ranking and potentially biologically important structures were represented on their respective

genomes. I also examine individual predicted structures in several of the datasets and compare experimentally validated structures and their uncharacterised, but evolutionarily conserved, analogues that are likely to serve similar biological functions.

3.4.1.1 Potexvirus

The potexvirus genus belongs to the *Alphaflexiviridae* family of (+)ssRNA viruses. Here I examined the potexvirus species potato virus X (PVX) and pepino mosaic virus (PepMV) for potentially functional secondary structure elements. A well-studied structural element was identified in the PVX dataset (location 32-106), previously reported to play important roles in PVX life cycle, such as translation (Miller et al., 1999), viral packaging (Kwon et al., 2005) and cell-to-cell movement (Lough et al., 2006). I named this structure PX1A. There are several nucleotide and structural features in PX1A of functional importance for PVX, namely, an unpaired *ACCA* motif, a terminal *GAAA* tetraloop and an *AUG* start codon. Within the PepMV dataset, a well conserved 71nt structure was identified, named PX1B, straddling the 5' UTR and the RdRp gene at a location similar (39-102nt) to that of PX1A (Figure 3.1). Despite the relatively low sequence identity (~37%) between the two structures, PX1B bears a close conformational resemblance to PX1A (Figure 3.2), sharing all of the important structural features. It is also notable that, in the central stem regions of both structures, there are a number of bases predicted to be significantly (p value < 0.05) complimentarily coevolving. PX1A is positioned in the top ten structures (9/118) of the consensus ranking of the PVX dataset, while PX1B ranks relatively high in the NASP PepMV conservation score (16/242) but places lower in the PepMV consensus ranking at 40/242.

Potexvirus

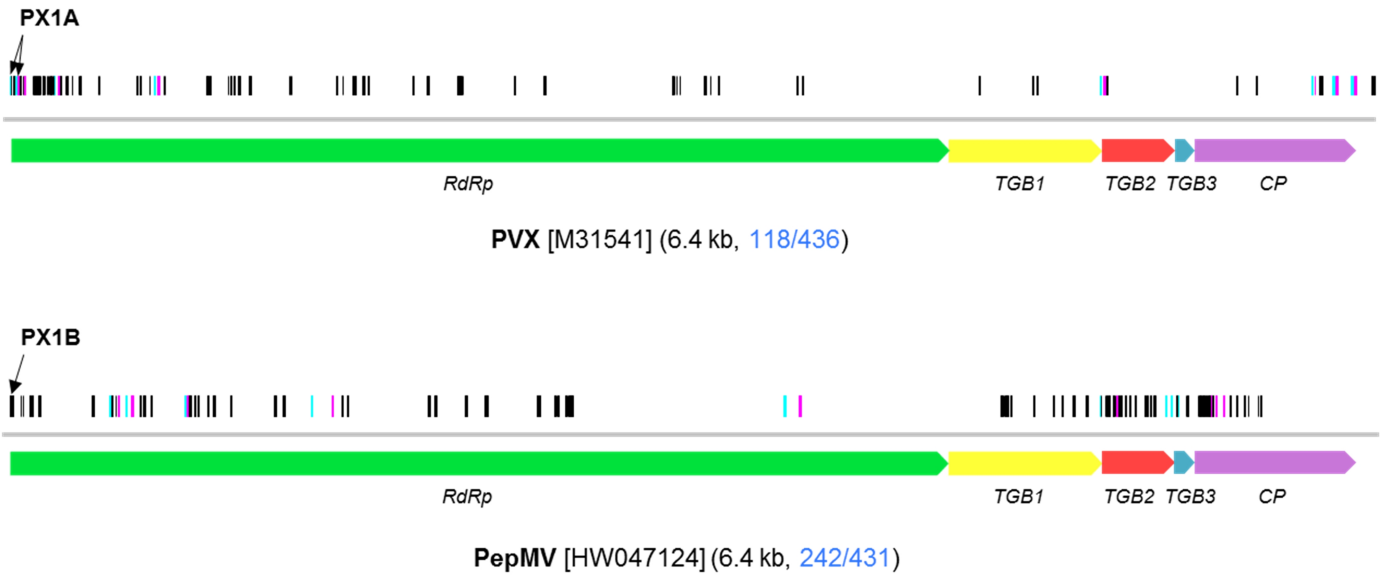


Figure 3.1 | Secondary structure maps of Potexvirus genomes.

The full length of the genome is represented by a grey horizontal line in a 5' to 3' direction, from left to right. Vertical lines above the genome represent the coordinates of identified structural elements within the 'high confidence structure sets' (HCSS). Only the 50 highest consensus ranking structures are shown in order to more clearly differentiate individual elements. The ten highest consensus ranking structures are shown in cyan and magenta vertical lines, representing the two complementary parts of the stem structure, the 5' and 3' stem respectively. All remaining structures are plotted as black vertical lines. Arrows above the vertical lines represent structures of interest discussed further in text. Coloured bars below the genome represent the coding regions. Numbers in brackets indicate the length of the genomes in kilobases as well as the ratio of the number of high confidence structures over the total number of predicted structures, in blue.

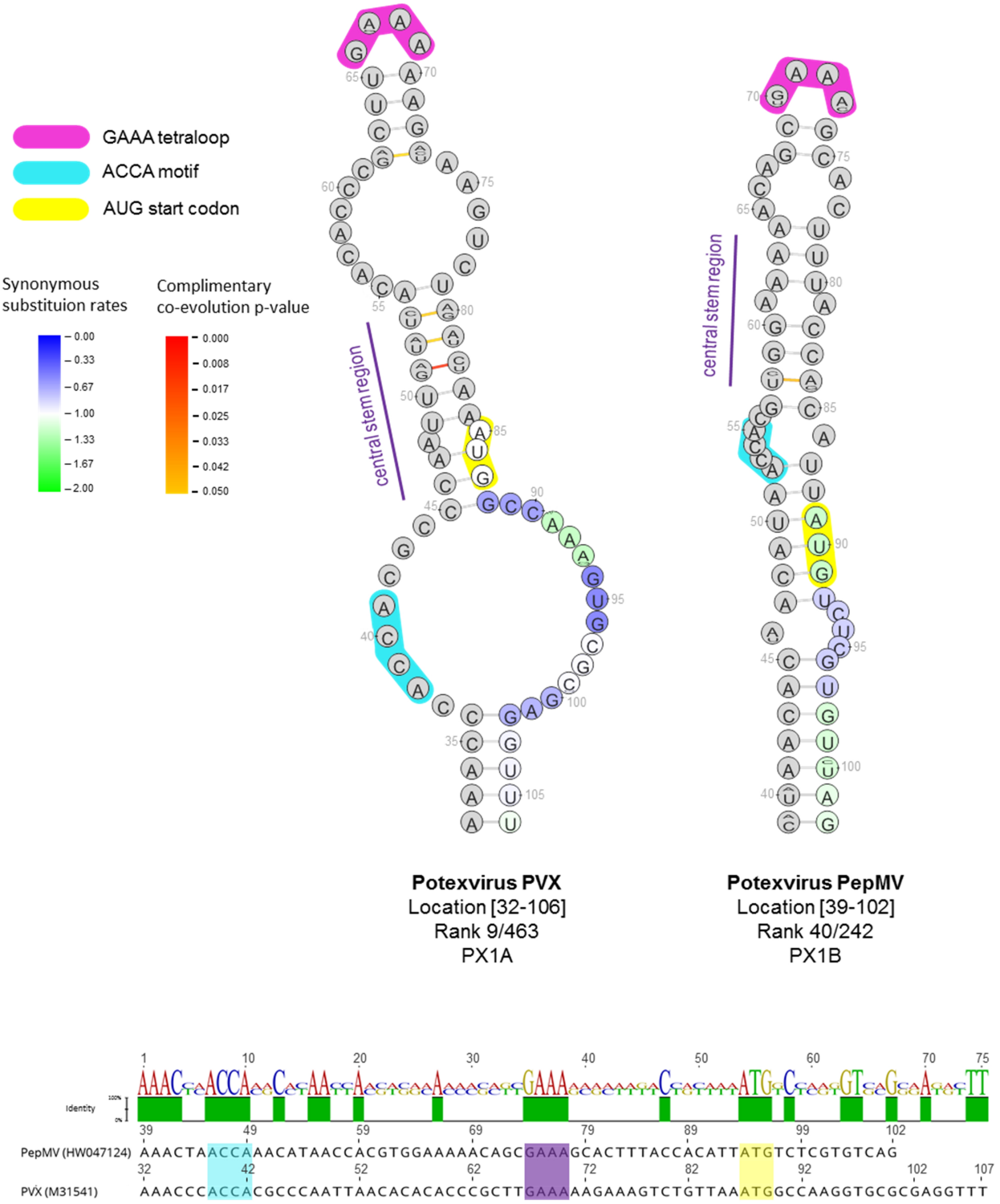


Figure 3.2 | Secondary structure associated with the 5' end of *Potexvirus* genome.

A secondary structure element associated with replication was predicted at the beginning of the genome in the Potexvirus datasets, PVX and PepMV. A common GAAA terminal tetraloop is highlighted in purple and the ACCA motif is highlighted in cyan, the AUG start codon is highlighted in yellow. Range of blue and green colours within nucleotides are associated with the synonymous substitution rate for that position. Lines connecting nucleotide bases, ranging from orange to red, indicate significant (p value < 0.05) complimentarily co-evolving base pairs. Below each structure, the genera and species, the location in the NASP alignment with gaps, the consensus rank, and allotted name are listed. The structures' sequence alignment, shows the position of each structure on the respective genome (without gaps), in addition to highlighting common functionally important regions. The sequence logo and identity plot show regions of nucleotide similarity between the two structures.

3.4.1.2 Caliciviridae

Caliciviridae is a family of viruses, currently containing the Lagovirus, Nebovirus, Norovirus, Sapovirus and Vesivirus genera. I examined the genomes of human norovirus (NV; Norovirus genus), rabbit haemorrhagic disease virus (RHDV; Lagovirus genus) and feline calicivirus (FCV; Vesivirus genus), for evidence of biologically relevant conserved secondary structures. Two separate conserved structural elements were identified in all three species analysed here. These structures shared structural and specific motif similarities, despite their genomes being somewhat divergent (~58% pairwise sequence identity shared between the three datasets).

A secondary structure in FCV, consensus ranked 5th out of 270 HCSS structures, termed CL1A (Figure 3.4), was identified at the 5' end of the 2A protein coding region (Figure 3.3). The structure contains a conserved *UCUUC* motif on its 3'-side of the stem, proposed to be a polypyrimidine tract binding protein (PTB) binding site (Vashist et al., 2012). Secondary structures sharing the same *UCUUC* motif as CL1A on their 3' stems and in a similar genomic region were identified in the RHDV and NV datasets, in spite of the low genomic sequence identity between the genomes of the *Caliciviridae* genera analysed. The analogue structure identified in the NV genome was termed CL1B (Figure 3.4) and was ranked 21st out of 91 HCSS structures in the consensus ranking, whereas the analogue structure in RHDV was named CL1C (Figure 3.4) and ranked 7th out of 77 HCSS consensus ranked structures.

Another conserved structure was identified in the 5' half of the RdRp gene in all three datasets. It has been suggested that the structure is involved in templating VPg uridilation in hepatoviruses (Simmonds et al., 2008) and has been reported to be conserved in NV genomes (Victoria et al., 2009), however it has not yet been reported in other calicivirus genomes. The structure identified in the NV genome was named CL2A (Figure 3.5), and ranks 9th out of 91 HCSS structures in the consensus ranking. The corresponding structure in FCV ranks 44th out of 270 HCSS structures in the consensus ranking and was named CL2B, in the RHDV dataset the analogous structure is termed CL2C and ranks 49th out of 77 consensus ranked HCSS structures (Figure 3.5). There is an *AAACA\C\G* motif located in a single-stranded region in all of the identified structures. In CL2A it is located in an internal-loop, whereas in CL2B and CL2C it is located in the terminal stem-loops (Figure 3.5).

Caliciviridae

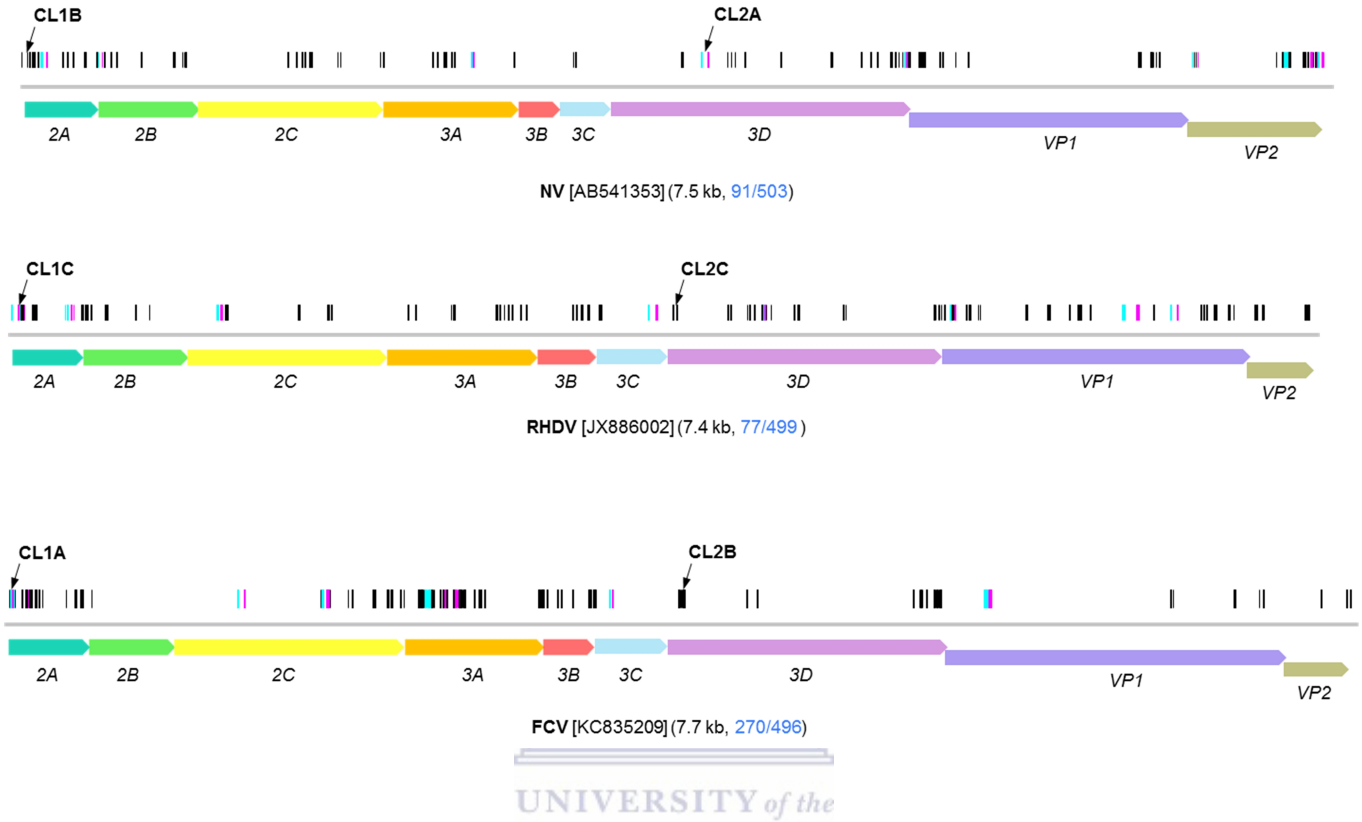


Figure 3.3 | Secondary structure maps of *Caliciviridae* genomes.

Vertical lines represent the coordinates of identified structural elements within the 'high confidence structure sets' (HCSS). Only the 50 highest consensus ranking structures are shown for legibility purposes. The ten highest consensus ranking structures are shown in cyan and magenta, representing the two complementary parts of the stem sequence, the 5' and 3' stems respectively. All remaining structures are represented in black. Arrows at the 5' end and the start of the 3D protein coding region indicate proposed common biologically functional structures in Caliciviruses. Numbers in brackets indicate the length of the genomes in kilobases as well as the ratio of the high confidence structures over the total number of predicted structures.

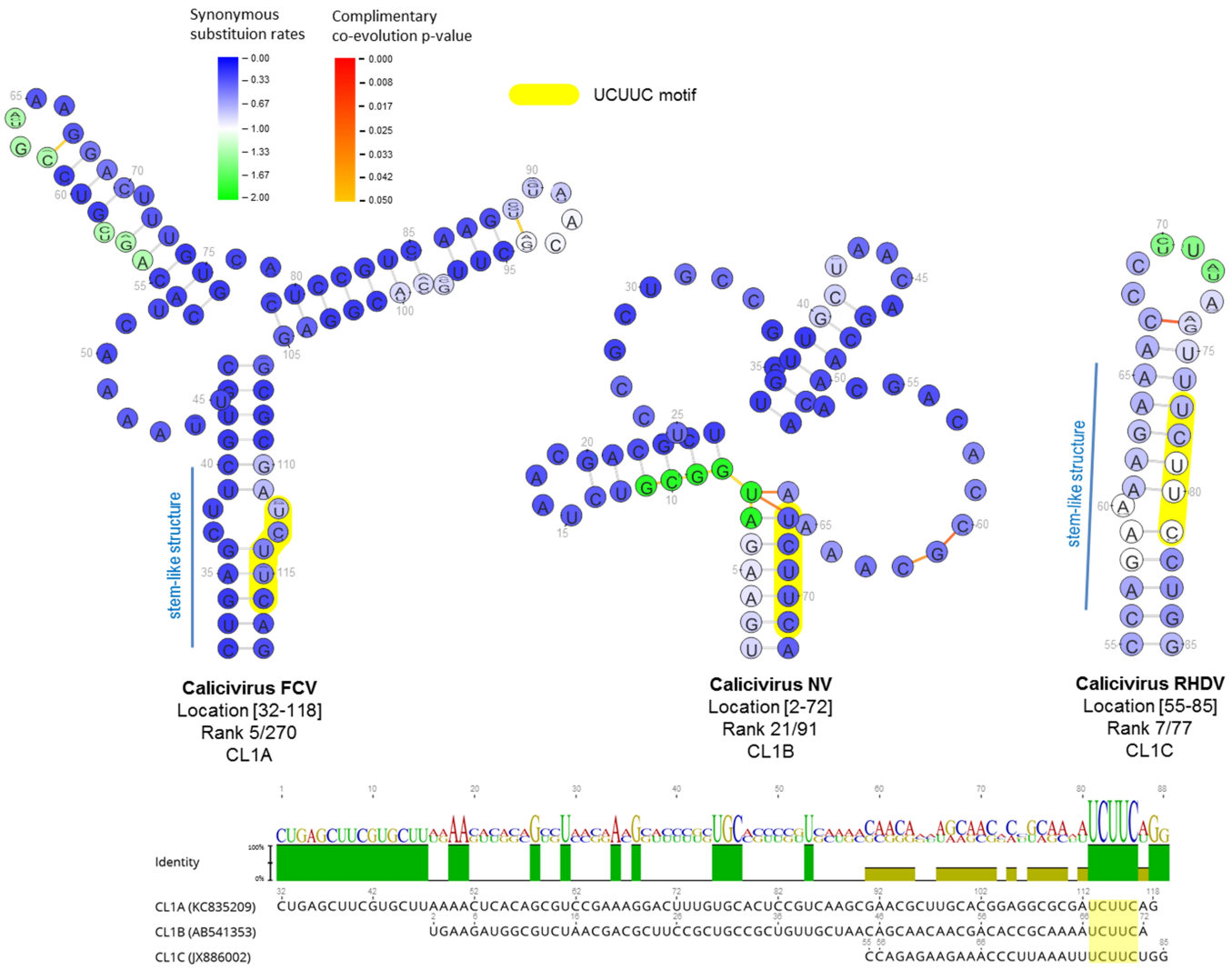


Figure 3.4 | Secondary structures within the 5' region of 2A protein in *Caliciviridae* genomes.

A secondary structure element associated with polypyrimidine tract binding was predicted at the 5' end of the 2A protein in the Calicivirus datasets, FCV, NV and RHDV. A common UCUUC motif is highlighted in yellow. Range of blue and green colours within nucleotides are associated with the synonymous substitution rate for that position. Lines connecting nucleotide bases, ranging from orange to red, indicate significant (p value < 0.05) complimentarily co-evolving base pairs. Below each structure, the genera and species, the location in the NASP alignment with gaps, the consensus rank, and allotted name are listed. The structures' sequence alignment, shows the position of each structure on the respective genome (without gaps), in addition to highlighting common functionally important regions. The sequence logo and identity plot show regions of similarity between the structures.

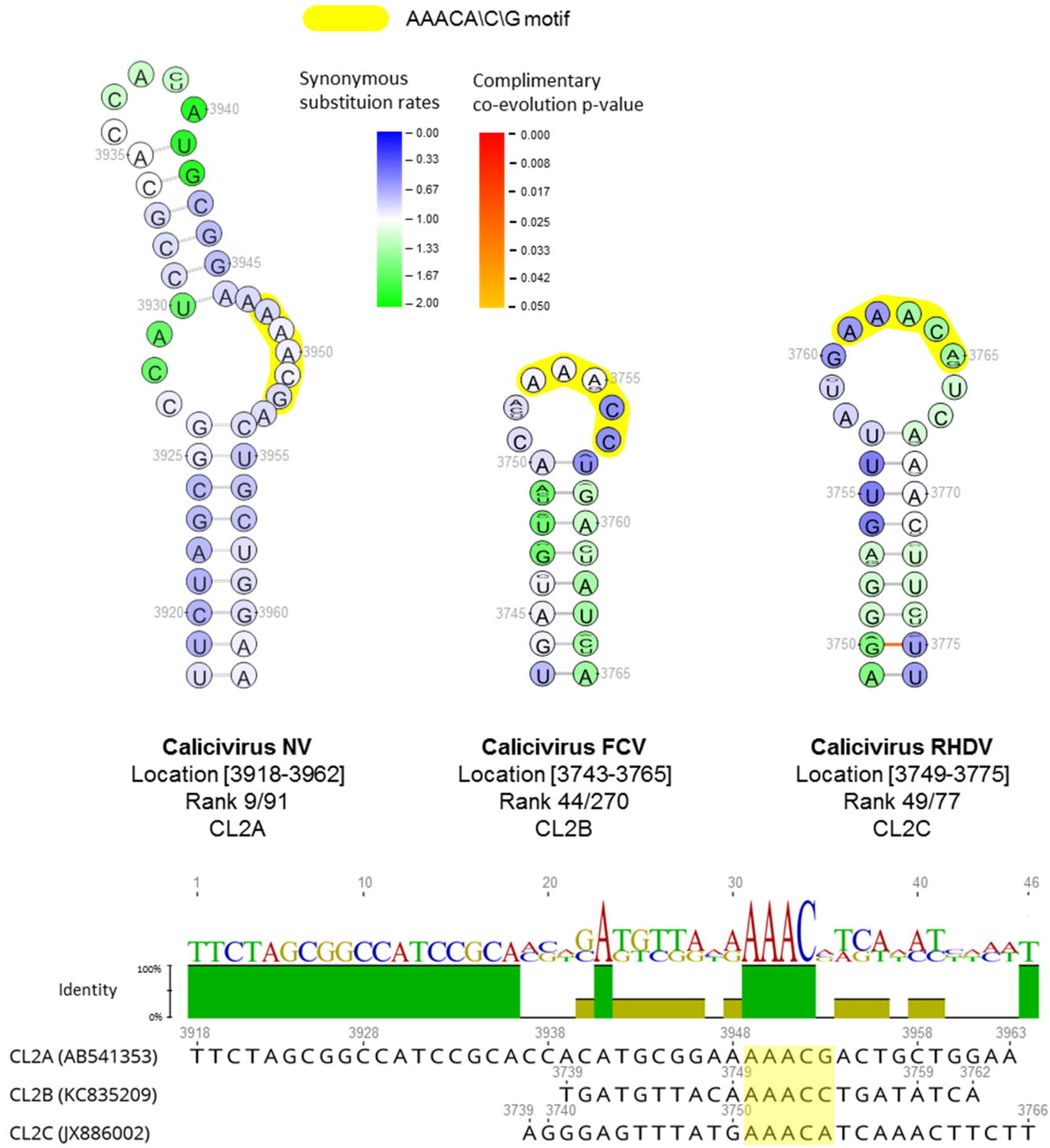


Figure 3.5 | Secondary structures within the 5' region of 3D (RdRp) protein in *Caliciviridae* genomes.

A secondary structure element associated with templating VPg uridilation was predicted at the 5' end of the 3D protein in the Calicivirus datasets, FCV, NV and RHDV. A common AAACA\C\G motif is highlighted in yellow. Range of blue and green colours of nucleotides are associated with the synonymous substitution rates of the codons within which they reside. Lines connecting nucleotide bases, ranging from orange to red, indicate significant (p value < 0.05) complimentarily co-evolving base pairs. Below each structure, the genera and species, the location in the NASP alignment with gaps, the consensus rank, and allotted name are listed. The structures' sequence alignment, shows the position of each structure on the respective genome (without gaps), in addition to highlighting common functionally important regions. The sequence logo and identity plot show regions of similarity between the structures.

3.4.1.3 Flaviviridae

The *Flaviviridae* family of viruses are classified into four genera namely, flavivirus, hepacivirus, pegivirus and pestivirus. The largest of the four genera, the flaviviruses, consists of more than 70 species and contains many important human and veterinary pathogens such as; dengue virus, Japanese encephalitis virus, West Nile virus and yellow fever virus (Kuno et al., 1998). I analysed the genomes of species belonging to the flaviviurs, hepacivirus and pestivirus genera for the presence of conserved secondary structure elements.

Various conserved structures were detected in the genomes of dengue viruses. In the dengue virus type 2 (DENV T2) dataset, a structure straddling the 5' UTR and capsid gene coding region (location 79-100) was identified (Figure 3.6). The structure was ranked 15th out of 195 HCSS structures in the consensus ranking and was named DV1A (Figure 3.7). This is a known functional element which enhances AUG-codon selection during translation in DENV2 (Clyde and Harris, 2006). The structure is composed of a stem (or stem-like region) harbouring the AUG start codon (in its 3' arm) and a conserved *GCAGA* terminal pentaloop. Analogous highly ranked structures, sharing a similar genomic location and structural features as DV1A, were identified in the dengue virus type 1 (DENV T1), dengue virus type 3 (DENV T3) and dengue virus type 4 (DENV T4) datasets, and were named DV1B, DV1C and DV1D, respectively (Figure 3.7). Structure DV1B ranked 4th out of 151 HCSS structures, DV1C ranked 1st out of 159 structures in the HCSS, and DV1D was ranked 152nd out of 235 HCSS structures. In contrast to the other three structures identified, structure DV1D included two internal-loops in its stem, but nevertheless contained the AUG start-codon in its 3' stem as well as the conserved *GCAGA* pentaloop.

Flaviviridae

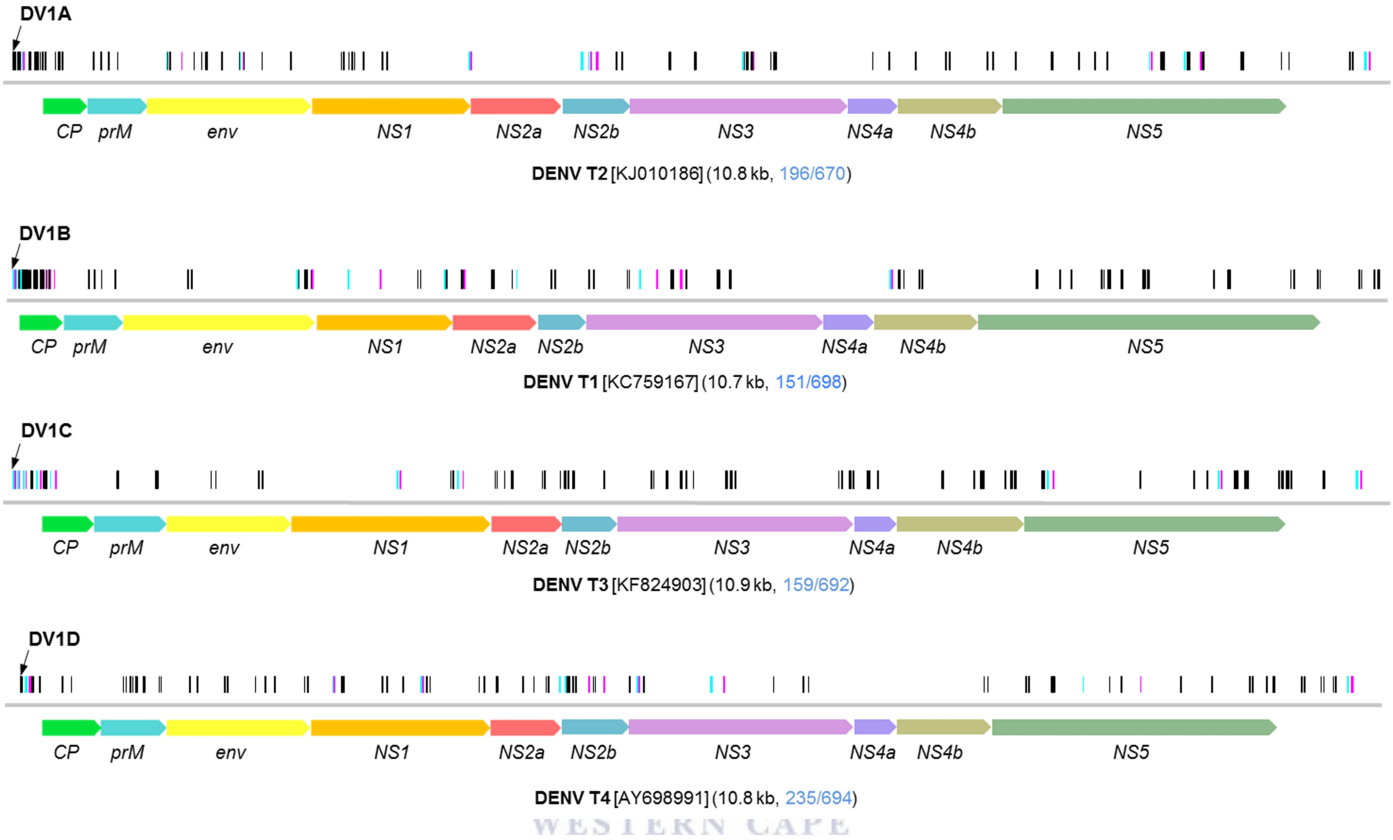


Figure 3.6 | Secondary structure maps of four types of dengue virus genomes.

The full length of the genome is represented by a grey horizontal line in a 5' to 3' direction, from left to right. Vertical lines above the genome represent the coordinates of identified structural elements within the 'high confidence structure sets' (HCSS). Only the 50 highest consensus ranking structures are shown in order to more clearly differentiate individual elements. The ten highest consensus ranking structures are shown in cyan and magenta vertical lines, representing the two complementary parts of the stem structure, the 5' and 3' stems respectively. All remaining structures are plotted as black vertical lines. Arrows above the vertical lines represent structures of interest discussed in text. Coloured bars below the genome represent the coding regions. Numbers in brackets indicate the length of the genomes in kilobases as well as the ratio of the number of high confidence structures over the total number of predicted structures, in blue.

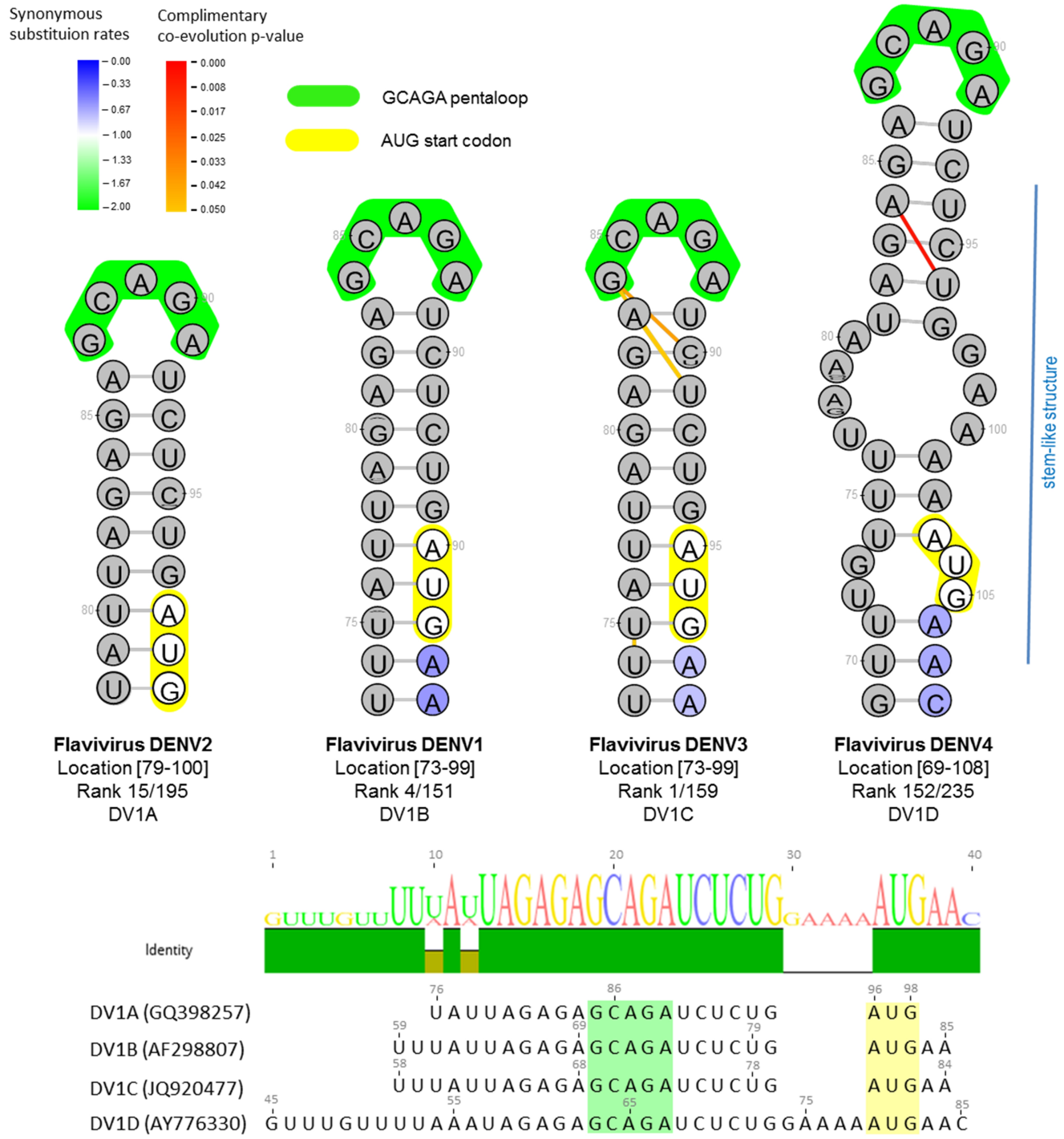


Figure 3.7 | Secondary structures within the 5' UTR region of dengue virus genomes.

A secondary structure element associated with start codon selection was predicted at the 5' UTR end of the Flavivirus datasets, DENV1, DENV2, DENV3 and DENV4. A common GCAGA pentaloop is highlighted in green, and the AUG start codon is highlighted in yellow. Range of blue and green colours within nucleotides are associated with the synonymous substitution rate for that position. Lines connecting nucleotide bases, ranging from orange to red colour, indicate significant (p value < 0.05) complimentarily co-evolving base pairs. Below each structure, the genera and species, the location in the NASP alignment with gaps, the consensus rank, and allotted name are listed. The structures' sequence alignment, shows the position of each structure on the respective genome (without gaps), in addition to highlighting common functionally important regions. The sequence logo and identity plot show regions of similarity between the structures.

3.4.1.4 Picornaviridae

The *Picornaviridae* viral family is comprised of 80 species grouped into 35 genera. Several genera in this family, namely, aphthovirus, cardiovirus, enterovirus and rhinovirus, can cause a wide range of illnesses in humans and other mammals. In humans, species of the enterovirus genus are responsible for diseases such as poliomyelitis, aseptic meningitis, conjunctivitis and gastroenteritis (Yin-Murphy and Almond, 1996).

In picornaviruses, translation and replication processes are regulated by, amongst other factors, three distinct well-conserved structured regions, specifically; an internal ribosome entry site (IRES) - composed of several structure domains located in the 5'UTR (Filbin and Kieft, 2009; Pelletier and Sonenberg, 1988), a cis-acting replicating element (CRE) required for the efficient synthesis of new + ve sense RNA strands, found in the polyprotein coding region (Goodfellow et al., 2000), and the origin of replication (OriR), required for negative strand synthesis, located in the 3'UTR (van Ooij et al., 2006).

In each of the three enterovirus types (A, B, C) analysed here, previously described functional elements contributed a large proportion of the ten highest consensus ranked secondary structures. For example, in enterovirus A, structures ranked 1st, 4th, 7th and 10th formed part of the IRES domains (Thompson and Sarnow, 2003), whereas structures ranked 2nd and 5th were elements of the CRE (Paul et al., 2000). In enterovirus B, the IRES domains were represented by structures ranked 3rd, 5th, 7th and 9th (Bailey and Tapprich, 2007), the CRE structure ranked 8th (Cordey et al., 2008), while the top ranked structure was part of the OriR element. Similarly, in the eneterovirus C dataset, the 1st, 3rd and 5th consensus ranked structures were located in the IRES region (Malnou et al., 2002), whereas the structures ranked 2nd and 8th represented elements of the CRE (Murray and Barton, 2003).

Although the genomes of the enterovirus species analysed here differ in length and gene coordinates, previously unreported structures were detected at similar genomic sites within each of the enterovirus datasets. In enteroviruses B and C the structures were well conserved, consensus ranked 13th and 7th respectively, and showed similar topology, but were divergent in both their nucleotide and amino acid sequences. In contrast, the structure identified in the enterovirus A dataset was neither very well conserved within its dataset, ranking 155th in the consensus rank, nor was it conformationally similar to the structures detected in the enterovirus B and C datasets. Interestingly however, overall its constituent nucleotides presented lower synonymous substitution rates in comparison to the two other structures. The

structural element identified in the enterovirus A dataset is located at position 5109 – 5176, at the N-terminus of the 3A protein (Figure 3.8), 17 nucleotides downstream of the 2C gene border and was termed EVA1 (Figure 3.9).

The corresponding structure detected in the enterovirus B dataset, named EVA2 (Figure 3.9), is located at the start of protein 3A, 28 nucleotides downstream of the 2C gene border, at location 5059 – 5101 (Figure 3.8). The structure identified in enterovirus C straddles the gene border of the 2C and 3A proteins, at location 5069 – 5110 (Figure 3.8), this structure was termed EVA3 (Figure 3.9).

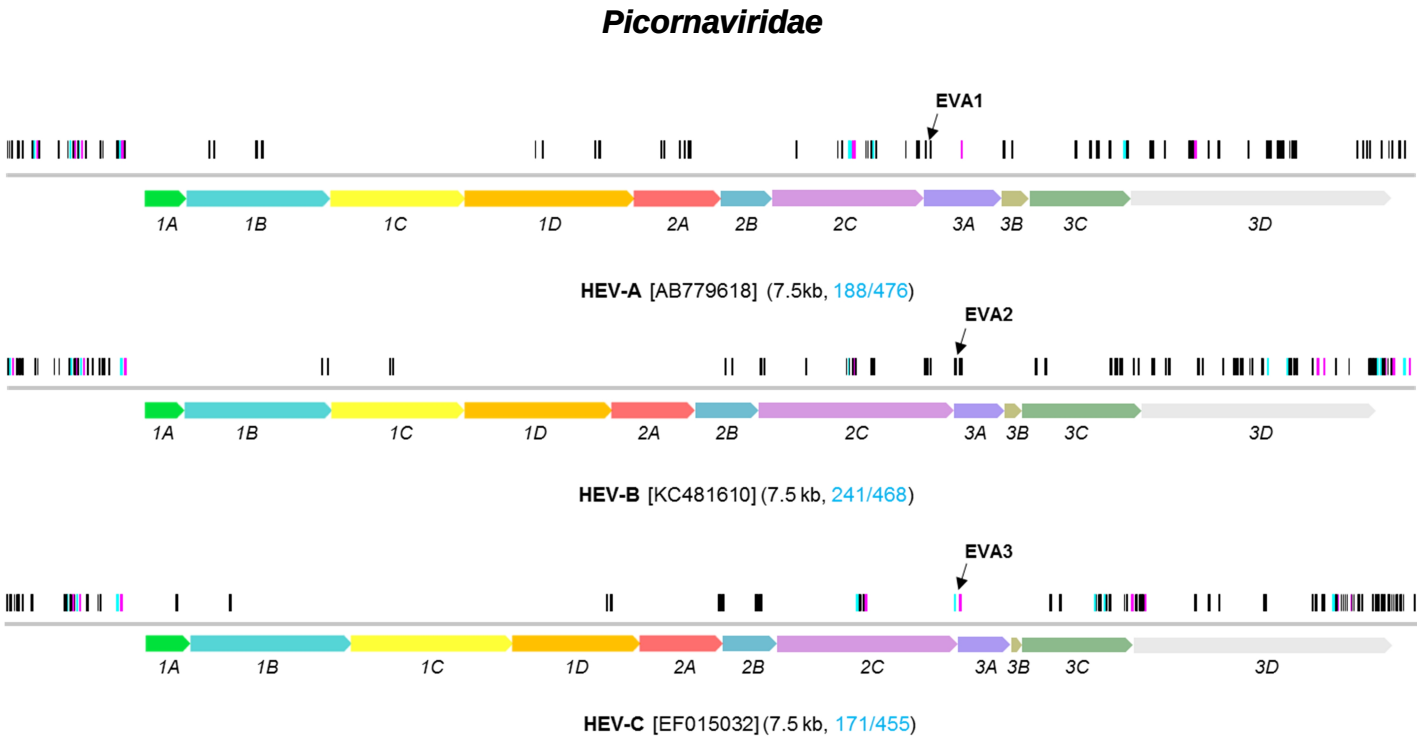


Figure 3.8 | Secondary structure maps of enterovirus genomes.

The full length of the genome is represented by a grey horizontal line in a 5' to 3' direction, from left to right. Vertical lines above the genome represent the coordinates of identified structural elements within the 'high confidence structure sets' (HCSS). Only the 50 highest consensus ranking structures are shown in order to more clearly differentiate individual elements. The ten highest consensus ranking structures are shown in cyan and magenta vertical lines, representing the two complementary parts of the stem structure, the 5' and 3' stems respectively. All remaining structures are plotted as black vertical lines. Arrows above the vertical lines represent structures of interest discussed in text. Coloured bars below the genome represent the coding regions. Numbers in brackets indicate the length of the genomes in kilobases as well as the ratio of the number of high confidence structures over the total number of predicted structures, in blue.

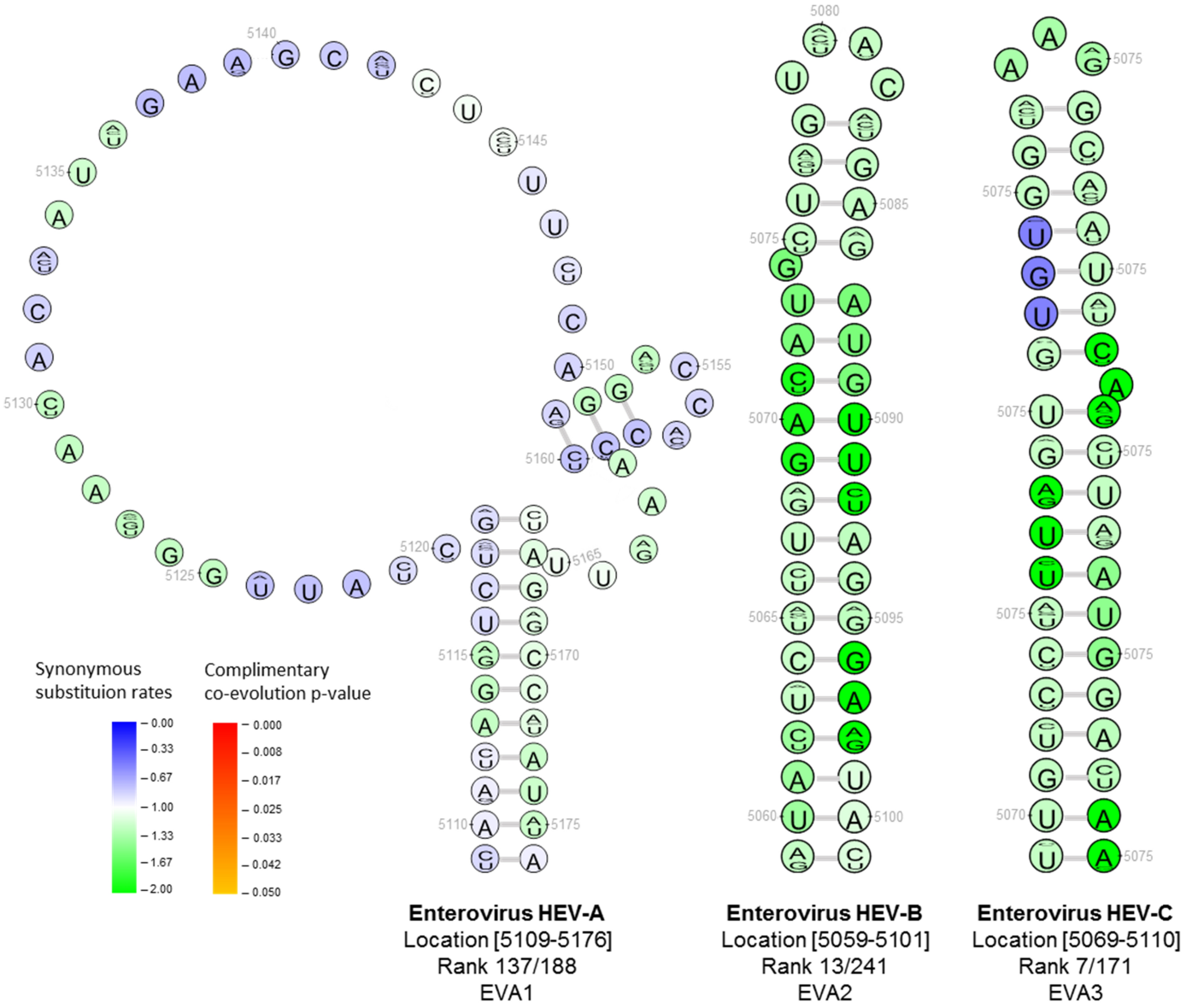


Figure 3.9 | Secondary structures at 2C/3A region of enterovirus genomes.

A secondary structure element associated with polyprotein cleavage was predicted at the 2C/3A protein border region of the Enterovirus datasets, HEV-A, HEV-B and HEV-C. Range of blue and green colours within nucleotides are associated with the synonymous substitution rate for that position. Lines connecting nucleotide bases, ranging from orange to red in colour, indicate significant evidence (p value < 0.05) of complimentarily co-evolving base pairs. Below each structure, the genera and species, the location in the NASP alignment with gaps, the consensus rank, and allotted name are listed.

3.4.1.5 Potyvirus

A previously uncharacterised but well conserved stem-loop structure was identified at the 3' end of the capsid protein coding regions of the ZYMV, SMV, PVY, WMV and BCMV datasets analysed here and was termed PVA1, PVA2, PVA3, PVA4 and PVA5, respectively (Figure 3.12). Structure PVA1 ranked 6/31 in the HCSS consensus ranking, PVA2 was ranked 13/144, PVA3 was ranked 22/88, PVA4 ranked 3/67 and PVA5 ranked 51/224. Although the five datasets share 68% sequence similarity, the structures display a nearly identical stem sequence, with the exception of PVA5 which presented an internal loop within the stem, and all five structures presented a conserved *AUGCC pentanucleotide* in their apical loops. All of the structures were located at the 3'-end of the CP coding region, approx. 200nt upstream of the start of the 3'UTR. In the smaller genomes of ZYMV, SMV and PVY, structures PV1A (9110-9131nt), PVA2 (9060-9085nt) and PVA3 (9103-9128nt) were positionally analogous to each other, whereas in the larger genomes of the WMV and BCMV datasets, both PVA4 and PVA5, were located approx. 400nt downstream of the other structures, at 9520-9545nt and 9517-9564nt, respectively (Figure 3.10).

Another highly conserved structure, present in two of the potyvirus datasets, was identified in the C-terminus of the TuMV and PPV capsid coding region, termed PVB1 (9271-9302nt) and PVB2 (9254-9285nt) respectively (Figure 3.13). Structure PVB1 ranked 10th out of 162 HCSS structures in the consensus ranking, and PVB2 ranked 15th out of 106 structures in the HCSS consensus rank. The structures had identical nucleotide length and contained an almost completely conserved stem-loop sequence that was also positionally analogous in both genomes.

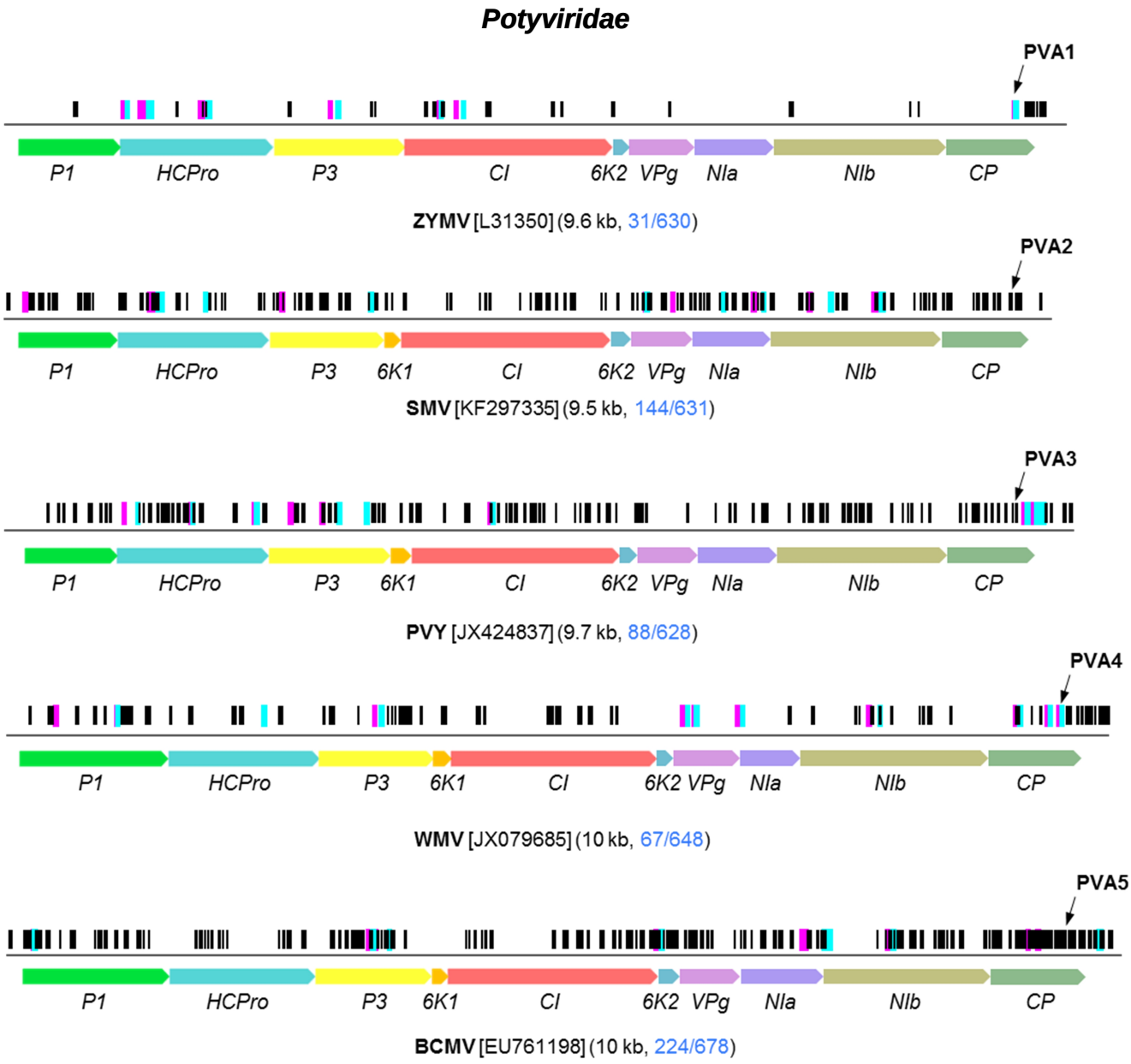


Figure 3.10 | Secondary structure maps of potyvirus genomes.

The full length of the genome is represented by a grey horizontal line in a 5' to 3' direction, from left to right. Vertical lines above the genome represent the coordinates of identified structural elements within the 'high confidence structure sets' (HCSS). Only the 50 highest consensus ranking structures are shown in order to more clearly differentiate individual elements. The ten highest consensus ranking structures are shown in cyan and magenta vertical lines, representing the two complementary parts of the stem structure, the 5' and 3' stems respectively. All remaining structures are plotted as black vertical lines. Arrows above the vertical lines represent structures of interest discussed in text. Coloured bars below the genome represent the coding regions. Numbers in brackets indicate the length of the genomes in kilobases as well as the ratio of the number of high confidence structures over the total number of predicted structures, in blue.

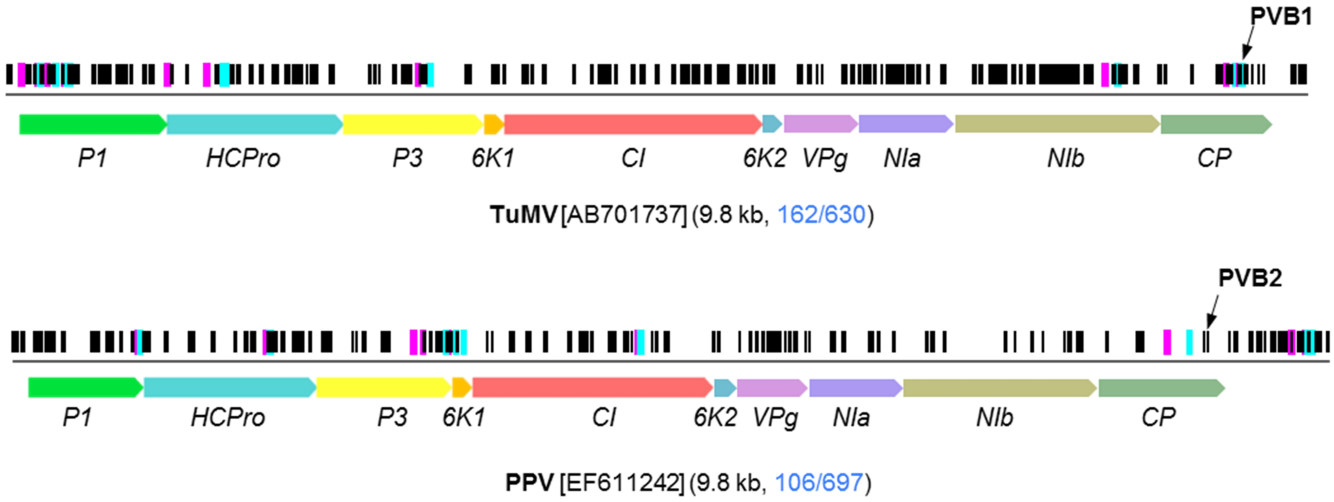


Figure 3.11 | Secondary structure maps of potyvirus genomes.

The full length of the genome is represented by a horizontal line in a 5' to 3' direction, from left to right. Vertical lines above the genome represent the coordinates of identified structural elements within the 'high confidence structure sets' (HCSS). Only the 50 highest consensus ranking structures are shown in order to more clearly differentiate individual elements. The ten highest consensus ranking structures are shown in cyan and magenta vertical lines, representing the two complementary parts of the stem structure, the 5' and 3' stems respectively. All remaining structures are plotted as black vertical lines. Arrows above the vertical lines represent structures of interest discussed in text. Coloured bars below the genome represent the coding regions. Numbers in brackets indicate the length of the genomes in kilobases as well as the ratio of the number of high confidence structures over the total number of predicted structures, in blue.

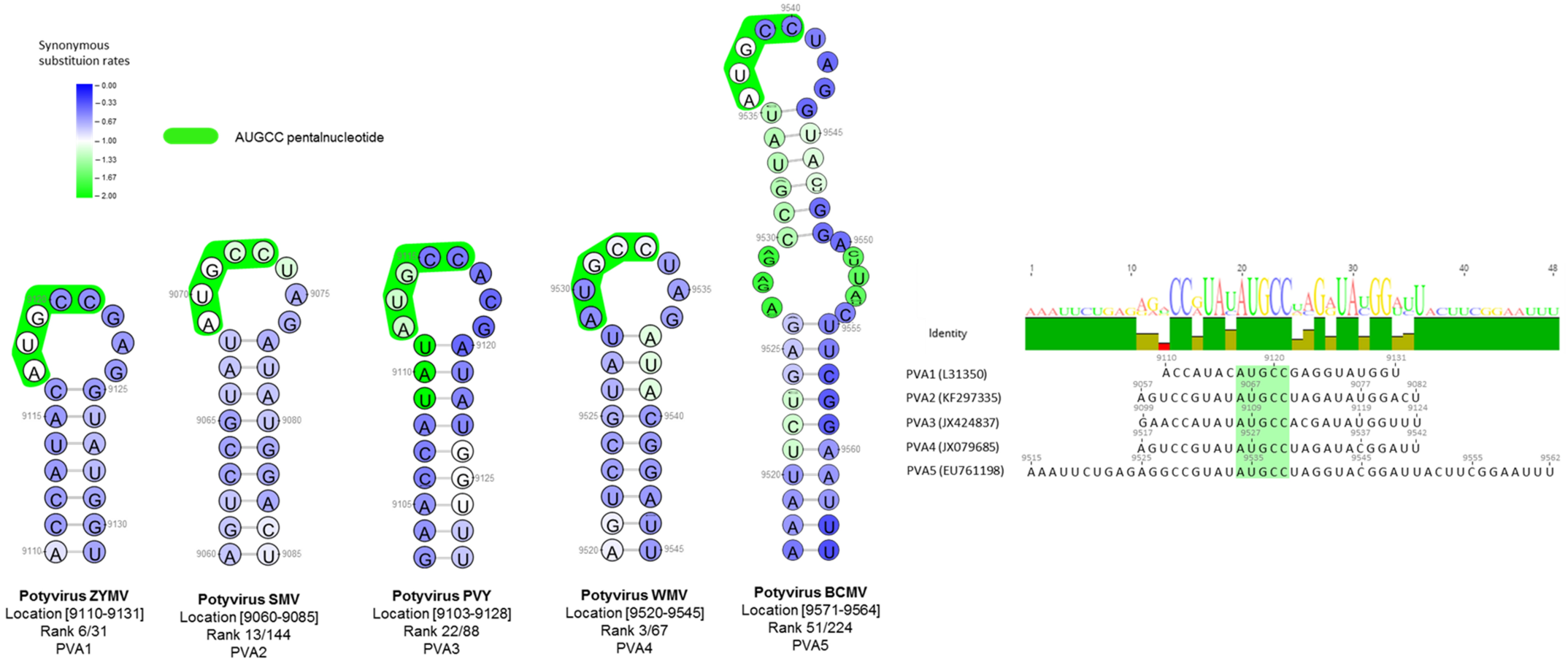


Figure 3.12 | Secondary structures within the capsid protein 3' region of potyvirus genomes.

A secondary structure element associated with genome replication was predicted at the capsid protein 3' region of the Potyvirus datasets, ZYMV, SMV, PVY, WMV and BCMV. A common AUGCC pentanucleotide is highlighted in green. Range of blue and green colours within nucleotides are associated with the synonymous substitution rate for that position. Below each structure, the genera and species, the location in the NASP alignment with gaps, the consensus rank, and allotted name are listed. The structures' sequence alignment, shows the position of each structure on the respective genome (without gaps), in addition to highlighting common functionally important regions. The sequence logo and identity plot show regions of similarity between the structures.

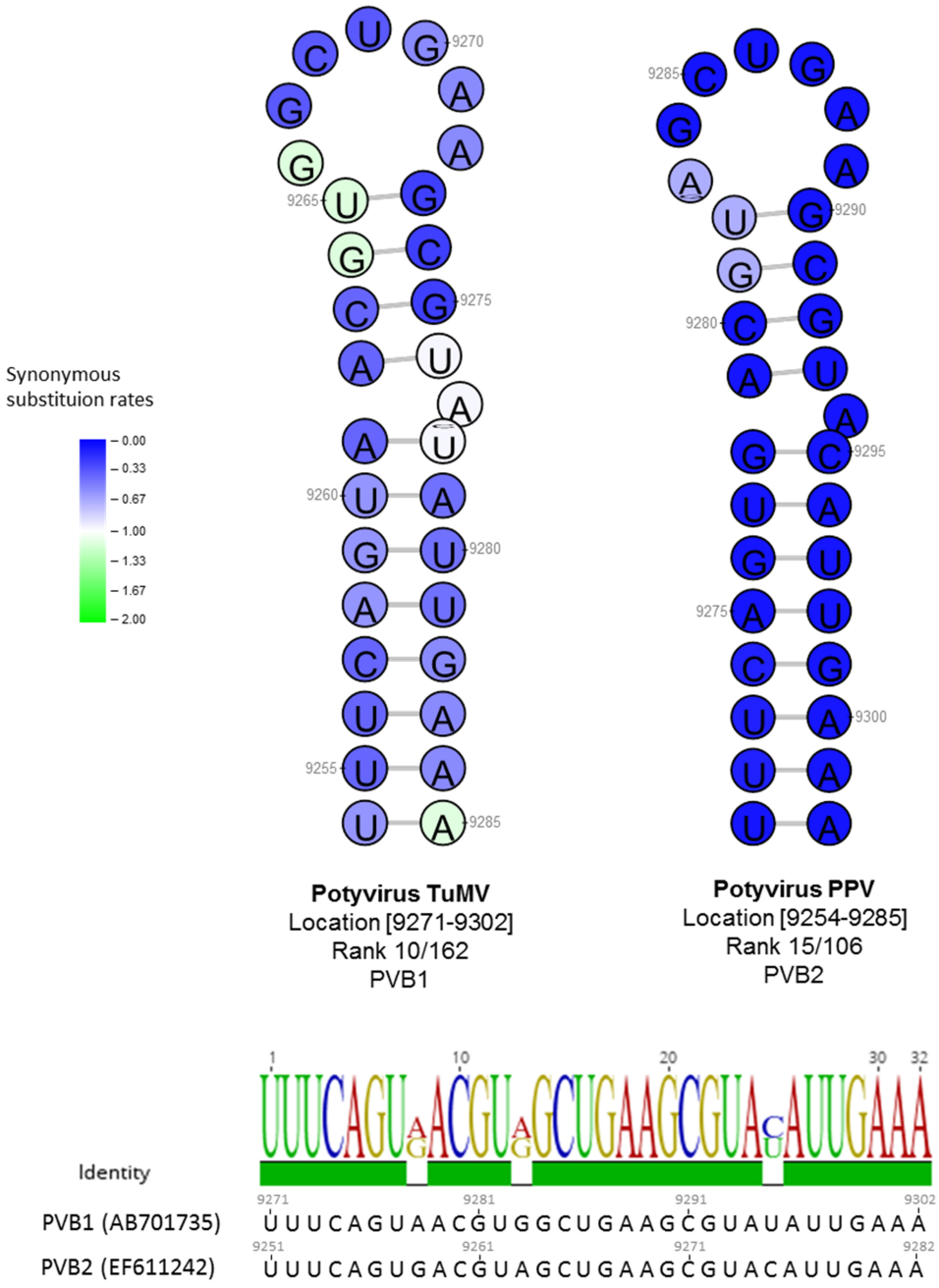


Figure 3.13 | Secondary structures within the capsid protein 3' region of potyvirus genomes.

A secondary structure element associated with genome replication was predicted at the capsid protein 3' region of the Potyvirus datasets, TuMV and PPV. Range of blue and green colours within nucleotides are associated with the synonymous substitution rate for that position. Below each structure, the genera and species, the location in the NASP alignment with gaps, the consensus rank, and allotted name are listed. The structures' sequence alignment, shows the position of each structure on the respective genome (without gaps), in addition to highlighting common functionally important regions. The sequence logo and identity plot show regions of similarity between the structures.

3.5 Discussion

3.5.1 Potexvirus

Presently, except for PVX (Miller et al., 1998) and bamboo mosaic virus (BaMV; Chen et al., 2010) there are no experimental reports describing biological relevance of secondary structures in the 5' regions of other potexvirus genomes. In PVX, a secondary structure, named SL1 (stem-loop 1), in the 5' end of the genome has been reported to play a critical role in the life cycle of PVX within infected host plants. It has been shown that SL1 is essential for the accumulation of plus-sense RNA, and that a stable central stem is of greater biological importance than the nucleotide composition of that region (Miller et al., 1998). Various features of a stable SL1 may promote translation or replication of genomic RNA. PVX transcripts containing structurally disruptive modifications of the terminal loop and stem regions have been shown to revert to wild-type structural conformation after several passages (Miller et al., 1999). Additionally, SL1 has been identified to play a role in the formation of virus-like particles by providing the specific recognition site for the capsid protein (CP) subunit (Kwon et al., 2005) that initiates viral packaging. This theory is supported by reports that SL1 is also involved in cell-to-cell movement in PVX (Lough et al., 2006) for which CP is required for sequestering RNAs from the viral replicase. It is also possible that SL1 might provide a specific binding site for host factors that play a role in modulating the virus replication cycle by interacting with viral proteins and other host factors. This particular structure has also been implicated as the origin of assembly in PVX (Kwon et al., 2005). The *ACCA* motif is repeated five times across the 5' UTR and overlaps the SL1 (Kim and Hemenway, 1996). These motifs bind a 54 kDa protein (p54) which is also important for the replication of PVX (Kim et al., 2002).

The SL1 structure identified in the PVX dataset (termed PX1A) presented all of the stable structural features described above and ranked 9/463, amongst structures comprising the HCSS. The PepMV dataset contained a structure (named PX1B) resembling the features of PX1A, at the same 5' UTR bordering RdRp gene location. The PX1B structure is positioned relatively lower in the consensus rank (40/242), largely due to a low rank in the synonymous substitution category, with only a small proportion of its constituent nucleotides being part of the coding portion of the structure (FUBAR only considers coding regions in its estimation of synonymous substitution rates).

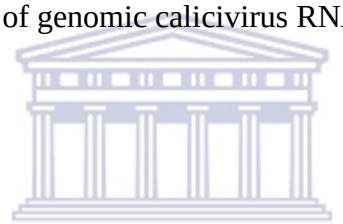
Since PVX and PepMV are demarcated as individual species within the potexvirus genus, their genomes have only a relatively low degree of nucleotide sequence conservation (~66% sequence identity). However, despite the difference in nucleotide composition, considering the structural similarities between these two elements it is likely that they share similar functional roles in their respective genomes.

3.5.2 Caliciviridae

Calicivirus genomes have been analysed previously for the presence and function of secondary structures. Earlier studies of replication kinetics in Murine norovirus (MNV) highlighted lowered synonymous substitution rates at regions predicted to be structured and demonstrated, through mutational analysis, that structural elements located in the 5' and 3' UTRs are of significance in the replication of MNV (Bailey et al., 2010; Simmonds et al., 2008). In the FCV dataset analysed here, an *UCUUC* motif [shown to be a polypyrimidine tract-binding protein (PTB) binding site (Yuan et al., 2002)] at the 5' end of the 2A protein (approx. 60 bases downstream of the *AUG* start codon) was nested within the stem of a well conserved structural element, named CL1A. Protein binding experiments have shown that PTB is essential for translation initiation in picornaviruses such as encephalomyocarditis virus (Kaminski et al., 1995) and foot-and-mouth disease virus (Hunt and Jackson, 1999) as well as the hepatitis C flavivirus (Gosert et al., 2000). While there were several pyrimidine-rich repeats in the 5' region of the 2A protein (eg. *UCUUU*, *UCCUUCUU*, *UCUUCC*) present in FCV dataset, experiments have shown that PTB has a specific preference to interact with the *UCUUC* sequence (Yuan et al., 2002).

In contrast with FCV, the other two calicivirus datasets, NV and RHDV do not present such saturated regions of pyrimidine rich repeats in their 5' ends, but in both datasets the *UCUUC* motif is well conserved and is similarly located only a few dozen bases downstream of the *AUG* start codon. The motif is nested within high ranking secondary structures in NV and RHDV, termed CL1B and CL1C, respectively. These structural elements have not yet been analysed to determine their function. It is likely, however, that CL1B and CL1C play a similar role to that of CL1A in the stimulation of translation and replication by serving as binding sites for PTB.

A secondary structure previously proposed to play a role in the uridylylation of VPg in NV (Victoria et al., 2009) was identified at the 5' end of the RdRp coding region of the NV dataset (3918-3963nt). The structure in our dataset, named CL2A, contained an *AAACG* motif embedded in its internal loop. Mutagenesis studies in polio- and rhinoviruses have revealed an *AAACA\C\G* motif in the loop regions of conserved CRE-like sequences (Rieder et al., 2000; Yin et al., 2003) that is important for RNA replication, while another study involving polioviruses reports that uridylylation of VPg serves as a primer for initiation of RdRp replication of genomic RNA (Ahlquist et al., 2003). Stem-loop structures, analogous to CL2A, in the FCV and RHDV datasets, named CL2B and CL2C respectively, presented *AAACA/C/G* conserved motifs in their terminal loops. Given the similarity in genomic location, the presence of a conserved *AAACA\C\G* motif, as well as the similarity in overall topology of these structures (despite the low sequence identity between the sequences), it is likely that CL2B and CL2C, like CL2A, will be capable of uridylylation of VPg, and in turn act as a primer for the replication of genomic calicivirus RNA (Högbom et al., 2009).



3.5.3 Flaviviridae

Dengue viruses, and other mosquito-borne flavivirus genomes are known to have multiple in-frame *AUG* codons at the start of their polyprotein coding regions, which may have a negative impact on the efficiency of translation. In a suboptimal *AUG* context, where particular primary sequence elements that modulate start-site recognition are absent or disordered, structural elements at the beginning of coding regions have been shown to enhance translation initiation of eukaryotic (Kochetov et al., 2005; Kozak, 1990) and other viral mRNAs (Clyde and Harris, 2006; Hwang and Su, 1999) by facilitating start-codon recognition. In mammalian mRNA, hairpin structures have been shown to increase translation from non-*AUG* initiator codons (Kozak, 1989; Takahashi et al., 2005).

The structure I termed DV1A, detected in the DENV T2 dataset has been computationally identified previously [named cHP, (Clyde and Harris, 2006)], and validated experimentally to be a functional element essential for efficient viral RNA synthesis in cultured human and mosquito cells (Clyde et al., 2008). While analogous secondary structures to DV1A have been computationally identified in other flaviviruses (Clyde and Harris, 2006), their role in the viral life cycle is yet to be investigated. I detected nearly identical structural elements in

the DENV T1, DENV T3 and DENV T4 datasets and these structures were named DV1B, DV1C and DV1D, respectively. Considering the similarities in structural topology and genomic location, coupled with conserved features such as the *GCAGA terminal pentaloop* and the *AUG* start-codon at their 3' stems, it is likely that DV1B, DV1C and DV1D play a similar biological role to that of DV1A (cHP) in facilitating start-codon recognition in the absence of a favourable initiation context.

The positive (+) sense RNA genomes of flaviviruses are translated as a single polyprotein that is posttranslationally cleaved by cellular host and viral proteases into three structural and seven non-structural proteins (Chambers et al., 1990). Proteases have very specific and complex active sites that present affinity for a particular substrate, be it a single amino acid or a specific set of codons (Garcia et al., 1999; López-Otín and Bond, 2008). In addition to conserved amino acid motifs at cleavage sites, conserved RNA structures detected at splice locations have been postulated to play regulatory roles in the splicing process in diverse viral groups of ssRNA viruses including retroviruses (Jacquenet et al., 2001; Mueller et al., 2014), alphaviruses (Yu et al., 1998), adenoviruses (Chebli et al., 1989) and orthomyxoviruses (Dela-Moss et al., 2014). No such structures have been yet identified at junctions of the polyprotein genes in the genomes of flaviviruses. After examining the splice sites of the dengue virus datasets for structural features, many of the junctions fell within some form of structured region [large irregular structures (>300nt) and long single stranded loop-regions], but these regions were neither conserved nor highly placed in the consensus rankings of their respective datasets. However, at one of the gene boundaries [N-terminal of the capsid protein (CP)], conserved stem-loop structures that were ordered relatively high in the consensus ranking were present. The structures in the different dengue virus types (1-4) show conformational similarity, but are divergent in their primary sequence. The four structural elements were termed DV2A, DV2B, DV2C and DV2D in DENV T1, DENV T2, DENV T3 and DENV T4, respectively. In each of the viruses the proposed splice site was located in the 3' arm of the stem region.

One of the ways secondary structures are able to regulate the splicing process is by concealing or exposing cleavage-protein binding-sites within paired helices, alternatively they can modulate the three dimensional proximity of important elements and motifs (Warf and Berglund, 2010). It is plausible that, in dengue viruses, these structures may have a role in regulating splicing efficiency by sequestering the cleavage site in their paired regions. A

factor increasing the plausibility of this proposition is that the structural proteins (and in particular the CP and its precursors) are the first to be proteolyticaly processed in mosquito-born flaviviruses (Mukhopadhyay et al., 2005).

3.5.4 Picornaviridae

Due to picornaviruses being amongst the smallest known viruses (18-30nm), it was initially believed that the picornavirus genome remained single-stranded and served merely as a template for polyprotein synthesis. It has since been shown that many (+)ssRNA viruses, including those of the *Picornaviridae* family, contain preserved global secondary structure across their genomes, termed genome-scale ordered RNA structure (Simmonds et al., 2004). Evolutionarily conserved genome-wide secondary structure potentially optimises the fitness of these viral groups by placing a constraint on viral mutation and substitution rates, leading to a low rate of divergence from successful genotypes. Additionally, global RNA structure may be involved in aiding the virus to escape host defences by mimicking cellular structured RNAs, thus preventing detection by pattern-recognition receptors and establishing persistence in the host (Goodbourn et al., 2000; Witteveldt et al., 2014).

Discrete genomic structures have been well characterised and conserved secondary structures that play essential roles in regulating translation (Filbin and Kieft, 2009), replication (Zoll et al., 2009) and virion assembly (Shakeel et al., 2017) have been identified. These structures were among the highest consensus ranked structures in each of the enterovirus datasets analysed here, highlighting both the biological relevance of these elements and validating the computational structural ranking approach applied here.

Despite the existence of extensive RNA secondary structure across the eneterovirus genomes, only one potentially functional element was identified in all three species, at the 2C/3A cleavage region: structures named EVA1, EVA2 and EVA3 in enterovirus A, B and C, respectively. Structurally, EVA2 and EVA3 have similar stem-loop conformations with each having a 1nt buldge in their stems. However the nucleotide compositions of EVA2 and EVA3 are very different and no specific conserved features or motifs were identified. The element assigned the EVA1 moniker, is found at the same location but has a different fold, compared to the other two structures, in that the stem-loop structure contains two internal bulges, a large 30nt loop and a smaller 5nt loop.

In picornaviruses the 2A protease catalyses the primary cleavage of the polyprotein co-transaltionally, separating the structural and non-structural proteins at the 1D/2A splice site. Subsequent cleavages of the rest of the polyprotein are facilitated by the 3C protease. Cleavage specificity studies have shown that the 2C/3A and the 2A/2B junctions, are the most efficiently cleaved splice sites by the 3C protease (Schultheiss et al., 1995). Despite the low sequence homology between EVA1, EVA2 and EVA3, their locations at the 2C/3A junction suggests they might be involved in regulating cleavage of the polyprotein.

RNA secondary structures have also been shown to facilitate the distribution and rate of genetic recombination at specific genomic regions in a number of viruses, such as; turnip crinckle virus (Nagy et al., 1999), human immunodeficiency virus (Galetto et al., 2004; Simon-Loriere et al., 2010), poliovirus (Dedepsidis et al., 2010; Runckel et al., 2013) and potato virus x (Draghici and Varrelmann, 2010). However, in enteroviruses, a study by Simmonds and Welch, 2006, found no evidence of correlation between genomic secondary structures and favoured breakpoint locations. However, in that study, the full spectrum of structural elements that exist was not fully appreciated and only previously described structures were considered when testing for associations.

It is not inconceivable that the structures in the enterovirus species described here are involved in directing recombination. Recombination breakpoint-clustering analysis in picornaviruses (Heath et al., 2006) report breakpoint clusters at or near the 2C/3A gene border in enterovirus A, B and C species.

It is plausible the structures identified here might have a role in regulating the splicing processes or perhaps serve as recombination ‘beacons’. In addition, analysing the genomes of specific enterovirus serotypes individually may provide a more accurate account of conserved secondary structures in distinct viral lineages, providing possible targets for functional analysis studies.

3.5.5 Potyvirus

Most (+)ssRNA viruses, and almost all RNA plant viruses, contain specific sequences and structural elements at the 3’ end of their genomes that recruit RNA-dependant RNA-polymerases to initiate synthesis of full length complementary minus-strands during RNA replication (Ahlquist et al., 2003; Hyodo and Okuno, 2014). In addition, in some plant

viruses, structures at the 3' end of the genome have been shown to act as transcriptional repressors, functioning via RNA-RNA interactions, leading to suppression of minus-strand RNA synthesis and thus regulating the ratio of positive- and minus-strands accumulated in the host cell during active replication (Sun et al., 2005; Zhang et al., 2004).

Unlike many other (+)ssRNA viruses in which the structural features of the 3' ends of their genomes have been extensively examined, in potyviruses functional analysis studies characterising the roles of secondary structures the 3'-terminal genomic regions are lacking. Nevertheless, studies have shown the existence of secondary structures located at both the CP and 3' UTR regions of tobacco etch potyvirus (TEV) which are essential for genome replication (Haldeman-Cahill et al., 1998; Mahajan et al., 1996).

Even though PVY and TEV belong to the same potyvirus supergroup, supergroup A (Gibbs and Ohshima, 2010), they only share ~60% pairwise sequence identity. The species ZYMV, WMV, SMV and BCMV belong to supergroup B (Gibbs and Ohshima, 2010). Despite the conserved structural elements identified in our datasets not sharing any primary sequence homology with the structure identified in TEV (named region A in: Haldeman-Cahill et al., 1998), they were positionally analogous to the TEV structure, and were located approx. 200nt upstream of the CP stop codon. Given that these well conserved structures are located in the biologically relevant 3'-end of their respective genomes, they are likely to be important factors for the positioning of the replication complex. What makes the conservation of these structures even more intriguing is that, among the potyvirus species, the CP region along with the 3'UTR are some of the most variable genomic regions: once being used as the defacto taxonomic indicators (Shukla and Ward, 1988), prior to full-genome phylogenetic classification (Gibbs and Ohshima, 2010).

Another almost completely conserved structure, named PVB, was identified in the CP coding region of two subgroup A potyvirus species, PPV and TuMV. This structure did not, however, share any sequence or structural similarity with PVA and there were no analogous structures identified in any of the other potyvirus species tested. Similarly to the conserved structure identified in the other potyvirus datasets, this element was located relatively close to the end of the CP gene, approx. 300nt upstream of the 3'UTR. It is likely that this is a biologically relevant structure contributing to the assembly of replicase proteins in this region.

It is also notable that in all of the potyvirus datasets, extensive secondary structure folding was predicted in the CP and 3'UTR genomic regions. Moreover, structures located in these regions, although not always conserved across all the species analysed here, constituted a large proportion (30% - 60%; Supplementary Table 1) of the ten highest ranking structures in each of the potyvirus datasets.

3.6 Conclusion

Although it was apparent, through our analysis, that in many of the species tested, selection is acting to maintain certain structural regions, it is also important to note that there are certain limitations to our study. For example, our model assumes thermodynamic equilibrium of the RNA molecule, when in reality it is likely that these molecules adopt a multitude of thermodynamically stable conformations, and this is especially true for longer sequences such as the ones tested here. While the structures reported here are the most optimal folds based on the parameters provided to the predictor, it is not possible to report with absolute certainty that these are the native structures present in these viral genomes, or that actually occurring biologically functional structures have not possibly been omitted. In addition, by excluding prediction of pseudoknoted structures from our analysis, potentially important biological RNA-RNA interactions may have been overlooked.

It will not be possible to definitively confirm the biological importance of any of the specific structures that I have highlighted here without laboratory validation. However, given that the biologically functional structures that have been identified in various studies do indeed rank highly in the analyses I have performed, it is very likely that many (if not all) of the predicted conserved structures presented here might be worthy of more detailed experimental evaluation.

Chapter 4: Influence of predicted secondary structures on recombination patterns in (+)ssRNA virus genomes

4.1 Abstract

Genetic recombination allows viral genomes to explore a significantly greater sequence space than would be possible by mutation alone. However, this process can also cause undesirable damage to coevolved sequence-specific interactions between specific components of their genomes. It has been shown that recombinants found in nature tend to produce chimeric proteins that are predicted to have a degree of folding disruption that is significantly lower than would be expected if recombination breakpoints were random and all chimeric proteins were equally functional; suggesting that natural selection is acting against recombinants expressing chimeric proteins with disrupted intra-protein amino acid interactions. Another biologically relevant class of co-evolved sequence specific interactions that could potentially be disrupted by recombination, are base-pairings within genomic secondary structures. In this study, I identified the locations of detectable recombination breakpoints in 51 (+)ssRNA full genome datasets and analysed the distribution patterns of these breakpoints to test whether there were associations between 1) breakpoint locations and the locations of protein coding regions and 2) the locations of breakpoints and the co-ordinates of predicted secondary structures. Overall, I observed strong evidence of recombination breakpoints regularly falling at the edges of protein coding regions, and found only weak evidence for breakpoints preferentially occurring at structured regions of the genome. The results from these tests suggest that. In the (+)ssRNA viruses at least, secondary structures are not as important an influence on recombination patterns as are the genomic features and/or selective forces that respectively favour the occurrence of breakpoints and/or the survival of recombinants with breakpoints that fall either outside of genes or close to gene boundaries.

4.2 Introduction

Besides point mutations arising during error-prone replication, genetic recombination is another contributing factor to the genetic variability of many (+)ssRNA viruses. Recombination involves the exchange of genetic material between genomes, referred to as parental sequences, to form a progeny virus, referred to as a recombinant sequence. A

number of molecular events have been proposed to contribute to the formation of viral RNA recombinants. These include genetic exchange between members of the same species, between viruses of different species co-infecting the same hosts or between viruses and their hosts.

It is widely accepted that the actual mechanism of sequence transfer during recombination in (+)ssRNA viruses involves a copy-choice mechanism, where the viral RNA polymerase complex switches templates during synthesis of nascent strands (Galetto et al., 2006). Template switching between related RNA templates is referred to as homologous recombination. This type of recombination is guided by high sequence similarity, where a sequence segment in one of the parental strands is replaced by a homologous segment from another parental strand. In contrast, non-homologous recombination involves the swapping of segments with low sequence similarity or the insertion of a sequence at different, 'incompatible' sites of the genome.

During non-homologous recombination, co-evolved intra-genome interaction networks are more likely to be disrupted, than they are during homologous recombination (Jain et al., 1999; Liu et al., 1999; Moreno et al., 2004). These networks include, but are not limited to; amino acid-amino acid interactions within protein folds, protein-protein interactions, interactions between proteins and specific RNA sequences, and interactions between base-paired nucleotides forming the basis of genomic secondary structures. Such recombination-induced disruptions have been inferred (Lefeuvre et al., 2007; Simon-Loriere et al., 2009), from the observation that there are generally higher degrees of intra-protein amino acid interaction disruptions in chimeric proteins that are expressed by randomly generated recombinants than there are in recombinants that occur naturally. It is expected, therefore, that recombinants where such intra-genome interactions have remained undisrupted will have chance of surviving and replicating, than those in which these interactions have been disrupted. Recently it has been shown, that in HIV-1M, selection might be acting to disfavour the survival of recombinant viruses in which base-pairing interactions within biologically relevant secondary structures have been disrupted (Golden et al., 2014).

While disruption of secondary structures due to recombining RNA molecules may have deleterious effects on the fitness of the recombinant, in some RNA viruses genomic secondary structures clearly play a role in directing genetic recombination such that it is far more likely to occur at certain genomic sites than it is at others. In the 2C and 3D gene

regions of polioviruses, for example, most recombination junctions occur in regions containing secondary structures that are similar between the recombining partners. It is proposed that when the poliovirus polymerase, in conjunction with the nascent negative strand, reaches such structured regions, the 3' end of the nascent negative strand may become detached from the initial template molecule and then re-attach to a homologous element on a second molecule which then becomes the new template, resulting in the synthesis of recombinant RNA (Dedepsidis et al., 2010). In HIV, clusters of recombination breakpoints at specific locations colocalise with stable stem-loop structures (Dykes et al., 2004; Galetto et al., 2004; Galli et al., 2008). It is thought that, as is the case in polioviruses, secondary structures in HIV may promote template-switching by causing the reverse transcriptase molecule, in combination with the growing nascent strand, to dissociate from a template as it stalls at the base of stem-loop regions. It is conceivable that some of the evolutionarily conserved secondary structures predicted to occur in the wide array of (+)ssRNA genomes analysed here could place similar constraints on the breakpoint distributions found in these genomes.

In this study I detect recombination breakpoint distributions in 51 (+)ssRNA full-genome datasets, to determine whether breakpoints preferentially occur at specific sites along these genomes. I use permutation based tests to determine whether the recombination breakpoints detected occur significantly more often than could be accounted for by chance at 1) the borders of genes and 2) at sites predicted, with high degree of probability, to be base paired.

4.3 Materials and Methods

4.3.1 Recombination detection and dataset preparation

Recombination detection within each of the 51 (+)ssRNA virus full-genome datasets, was performed using seven recombination methods implemented in RDP4 (Martin et al., 2015): BOOTSCAN (Salminen et al., 1995), GENECONV (Padidam et al., 1999), MAXCHI (Smith, 1992), CHIMAERA (Posada and Crandall, 2001), SISCAN (Gibbs et al., 2000), 3SEQ (Boni et al., 2007), and the RDP method (Martin and Rybicki, 2000). Sequences sharing more than 95% similarity with other genomes were excluded from the initial recombination scan in order to optimise detection by minimising the number of genome comparisons the program needed to make. Following this preliminary recombination

detection phase, potential recombinant and parental sequences (or rather, their reasonably close relatives), and breakpoint locations were identified and were manually checked for phylogenetic incongruences using the recombination signal analysis tools integrated in RDP4. Those recombination events identified by at least three of the detection methods with high degree of certainty ($P < 0.05$), were regarded 'accepted'. Datasets containing at least ten accepted recombination events (30/51) were selected for subsequent analysis to test for evidence of (1) detected breakpoints co-localising with predicted structures of the HCSSs and (2) breakpoints clustering at the boundaries of genes, as opposed to the middle of the genes (Figure 4.1).

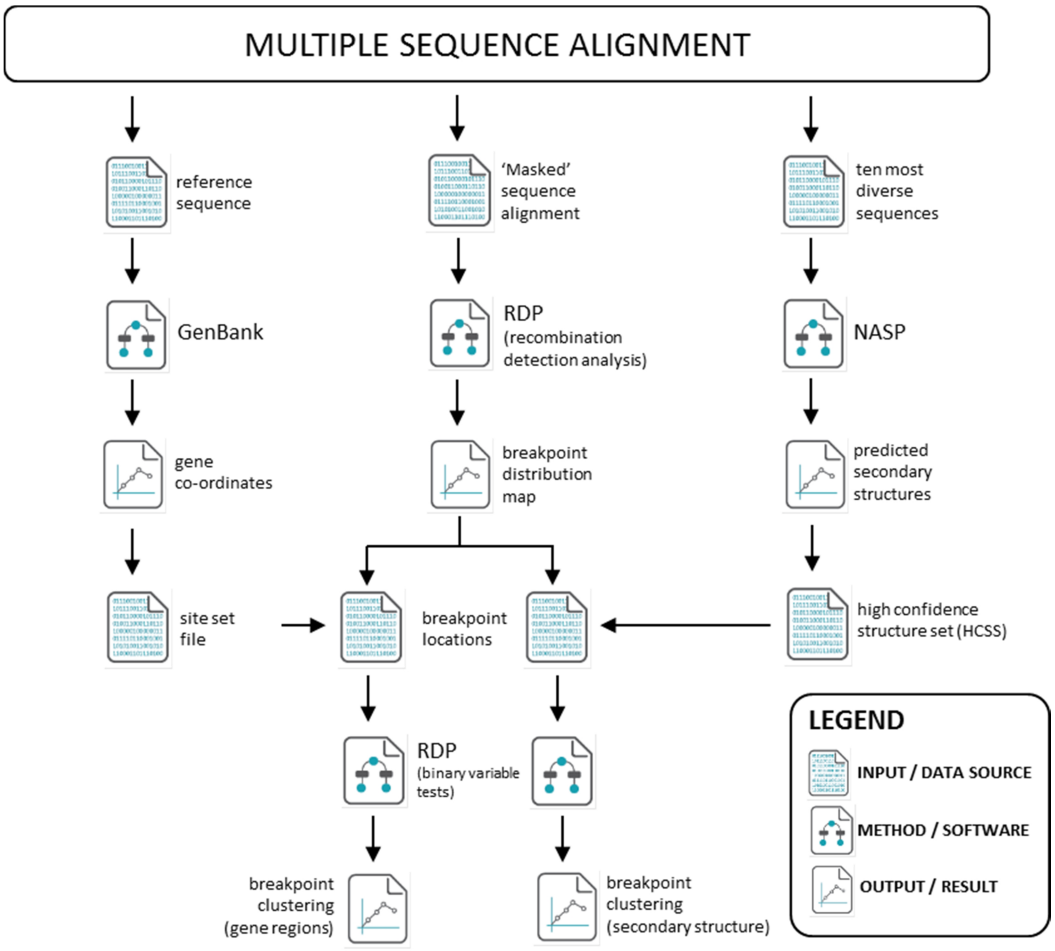


Figure 4.1 | Recombination detection and analysis workflow.

For each dataset, recombination events were detected using RDP4 (Martin et al, 2015) and breakpoint distribution plots constructed. To test whether there was a tendency for breakpoints to occur at the edges of genes, GenBank gene co-ordinates were mapped to the breakpoint plots and permutation tests were performed. Predicted HCCS co-ordinates were overlaid onto the breakpoint plots and binary variable tests were performed to determine whether there were associations between secondary structures and breakpoint locations.

4.3.2 Genome-wide breakpoint distribution maps

Breakpoint distribution maps were constructed, containing all positively identified unambiguous recombination events, identifying regions of high and low concentration of recombination breakpoints, referred to as, 'hot- and cold-spots', respectively. The breakpoint distributions were used to build breakpoint density plots for each of the datasets by sliding a 200nt window one nucleotide at a time along the sequences, counting all of the identified breakpoint positions within every window and plotting the sum at the central position of each window. A permutation test was performed to test the significance of the identified hot and cold spots. The random access memory available to the workstation is the limiting factor to the number of permutations that can be performed. These ranged from 4300 to 6500 permutations, depending on the size of the dataset analysed. The permutation test accounted for the fact that recombination breakpoints are both more easily, and more accurately, detectable in genome regions where sequences are more diverse than in regions of high similarity. When compiling the permuted datasets, beginning breakpoint positions were randomised and ending breakpoint positions were placed the same number of variable nucleotide sites downstream from the beginning breakpoint positions as in the real datasets (Heath et al., 2006). Significant recombination clusters are in those 'windows' that contain more breakpoints than were contained within the corresponding 'windows' of more than 99% of the permuted sequences.

4.3.3 Test for association between breakpoint distribution and gene coordinates

Conserved recombination breakpoint patterns have been detected in a number of ssRNA- (Heath et al., 2006; Lukashev et al., 2003) and ssDNA-viruses (Lefeuvre et al., 2009; Lukashev et al., 2008; Martin et al., 2011) where, during sequence exchanges, it is apparent that coding modules tend to be preserved largely intact. It is likely that viable recombinant progeny survive as a result of recombination events that minimise disturbances of the arrangements of interacting compatible modules, thus avoiding disruptions of vital intragenome interaction networks (Escriu et al., 2007; Martin et al., 2005). I set out to test the selected datasets for evidence of association between breakpoint dense regions and locations of protein coding domains. Gene coordinates for each of the datasets were mapped to the breakpoint distribution plots and a binary variable test was used to establish whether the

beginning and ending 5% of gene domains contained significantly more or fewer detectable breakpoints than those collectively observed in the remaining regions of the genes.

4.3.4 Tests for association between locations of detected recombination breakpoints and predicted secondary structures

It has been suggested that genomic secondary structures facilitate template switching events during recombination in ssRNA (Dedepsidis et al., 2010; Makino et al., 1986; Runckel et al., 2013) and retro- (Galetto et al., 2006; Galli et al., 2008) viruses, where higher concentration of breakpoint junctions have been associated with stable secondary structure regions. I tested each of the selected datasets to determine whether the coordinates of structural elements within the HCSSs corresponded with locations of detected recombination breakpoints. For every dataset, the HCSS genome-wide secondary structure coordinates were mapped to its breakpoint distribution map and a Fisher's exact test was applied to test whether recombination breakpoints occurred more frequently than could be accounted for by chance at sites that NASP had predicted were base-paired.

4.4 Results

4.4.1 Genome-wide breakpoint distribution in ssRNA viruses

Significant evidence of recombination breakpoint hot- and/or -cold-spots were detected in all but two (DENV T2 and BCMV) of the datasets investigated.

4.4.1.1 Picornaviridae

Breakpoint clustering observed in datasets of the *Picornaviridae* family appeared to be non-randomly distributed and is in alignment with previous findings (Heath et al., 2006; Simmonds and Welch, 2006). Specifically, in the FMDV, HEV-B, HEV-C, PTeV and SAFV datasets, two significant recombination hot-spots were detected on either side of the structural genome region, comprising the VP2, VP3 and VP1 genes. In addition, the FMDV, HEV-B, HEV-C and PTeV datasets displayed significant breakpoint cold-spots within areas of the structural region. Another common hotspot, between the junction of the 3A and 3B genes, was detected in the HEV-B, HEV-C, HRV-A, HRV-C, PTeV and HPeV datasets (Figure 4.2). Decreased recombination breakpoint frequencies within the structural protein genes have been reported previously in picornaviruses (Heath et al., 2006), geminiviruses (Lefeuvre et al., 2009) and human immunodeficiency viruses (Fan et al., 2007), implying that disruption

of structural protein genes by recombination may be generally less tolerable than that of other gene types.

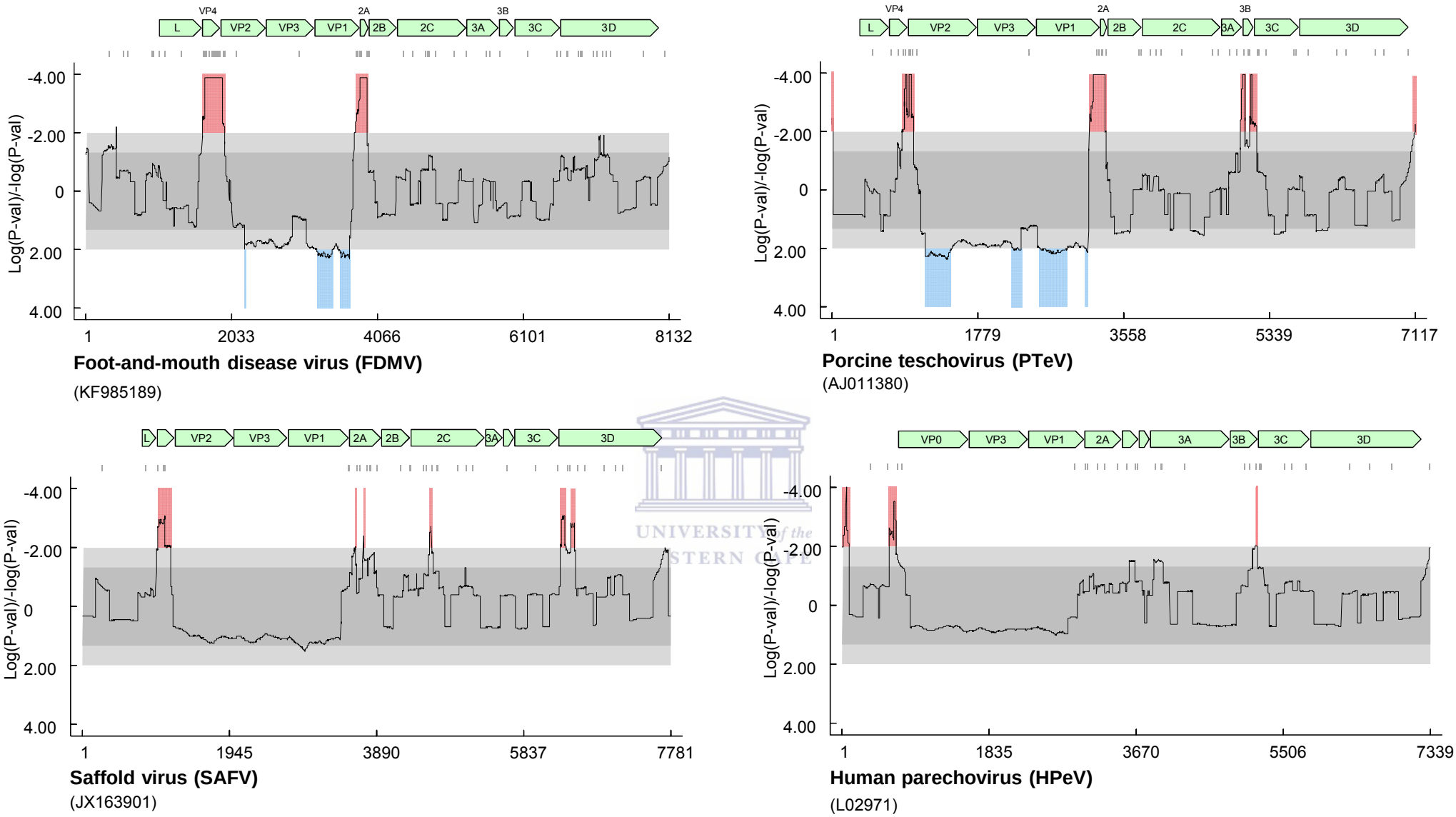


Figure 4.2 | Breakpoint distribution plots of *Picornaviridae* datasets The distribution of recombination breakpoints detected within FMDV, PTeV, SAFV and HPeV datasets. Estimated breakpoint positions are indicated by small vertical lines at the top of the graph. Light and dark grey areas, indicate local 99% and 95% breakpoint clustering thresholds, respectively. Red areas indicate recombination hot-spots, while blue areas represent recombinational cold-spots. Gene maps, in green, are drawn to scale in relation to the sequence of interest (accession number in brackets below each figure).

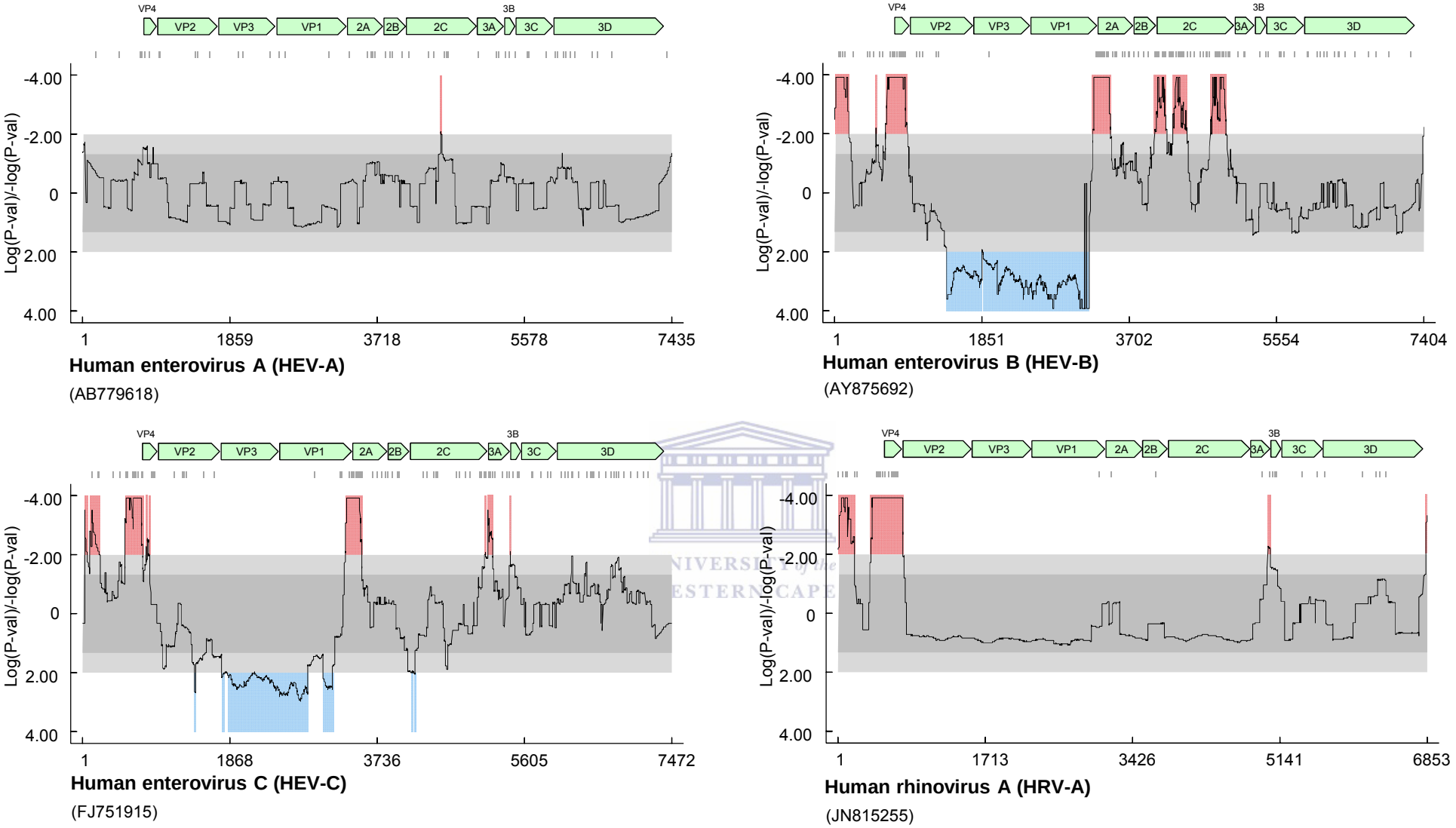


Figure 4.2 Continued | Breakpoint distribution plots of *Picornaviridae* datasets The distribution of recombination breakpoints detected within HEV-A, HEV-B, HEV-C and HRV-A datasets. Estimated breakpoint positions are indicated by small vertical lines at the top of the graph. Light and dark grey areas, indicate local 99% and 95% breakpoint clustering thresholds, respectively. Red areas indicate recombination hot-spots, while blue areas represent recombinational cold-spots. Gene maps, in green, are drawn to scale in relation to the sequence of interest (accession number in brackets below each figure).

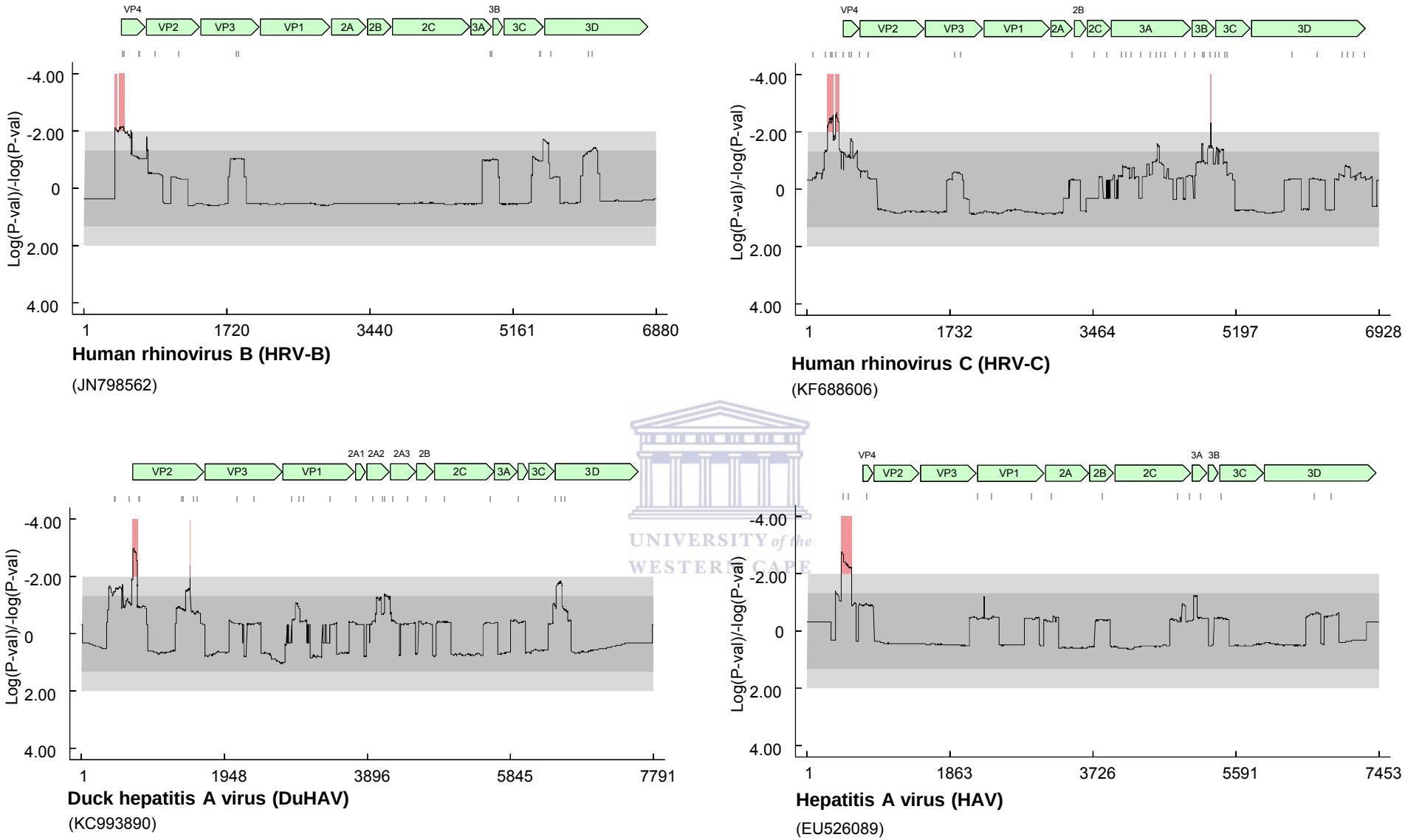


Figure 4.2 Continued | Breakpoint distribution plots of *Picornaviridae* datasets The distribution of recombination breakpoints detected within HRV-B, HRV-C, DuHAV and HAV datasets. Estimated breakpoint positions are indicated by small vertical lines at the top of the graph. Light and dark grey areas, indicate local 99% and 95% breakpoint clustering thresholds, respectively. Red areas indicate recombination hot-spots, while blue areas represent recombinational cold-spots. Gene maps, in green, are drawn to scale in relation to the sequence of interest (accession number in brackets below each figure).

4.4.1.2 Flaviviridae

Besides the DENV T2 dataset, significant recombination hot-spots were detected in all of the *Flaviviridae* datasets tested here. Most notably, significant breakpoint clustering was evident at the border of the E and NS1 genes in DENV T1, DENV T3 and JEV datasets. Similarly, a significant recombination hot-spot was detected at the border of the E1 and E2 genes in the HCV dataset. Another significant recombination hot-spot was found between the borders of the NS3 and NS4A genes in the DENV T1 and JEV datasets, and although a similar result was observed in the DENV T3 and TbEV datasets, these were not regarded as significant for the purposes of this study (these genomes contained more breakpoint positions than the maximum detected in more than 95% of the permuted density plots but did not exceed the 99% significance threshold). Decreased recombination breakpoint densities were also evident in the structural C (capsid) protein coding genes of the DENV T1, DENV T2, DENV T3 and TbEV datasets.

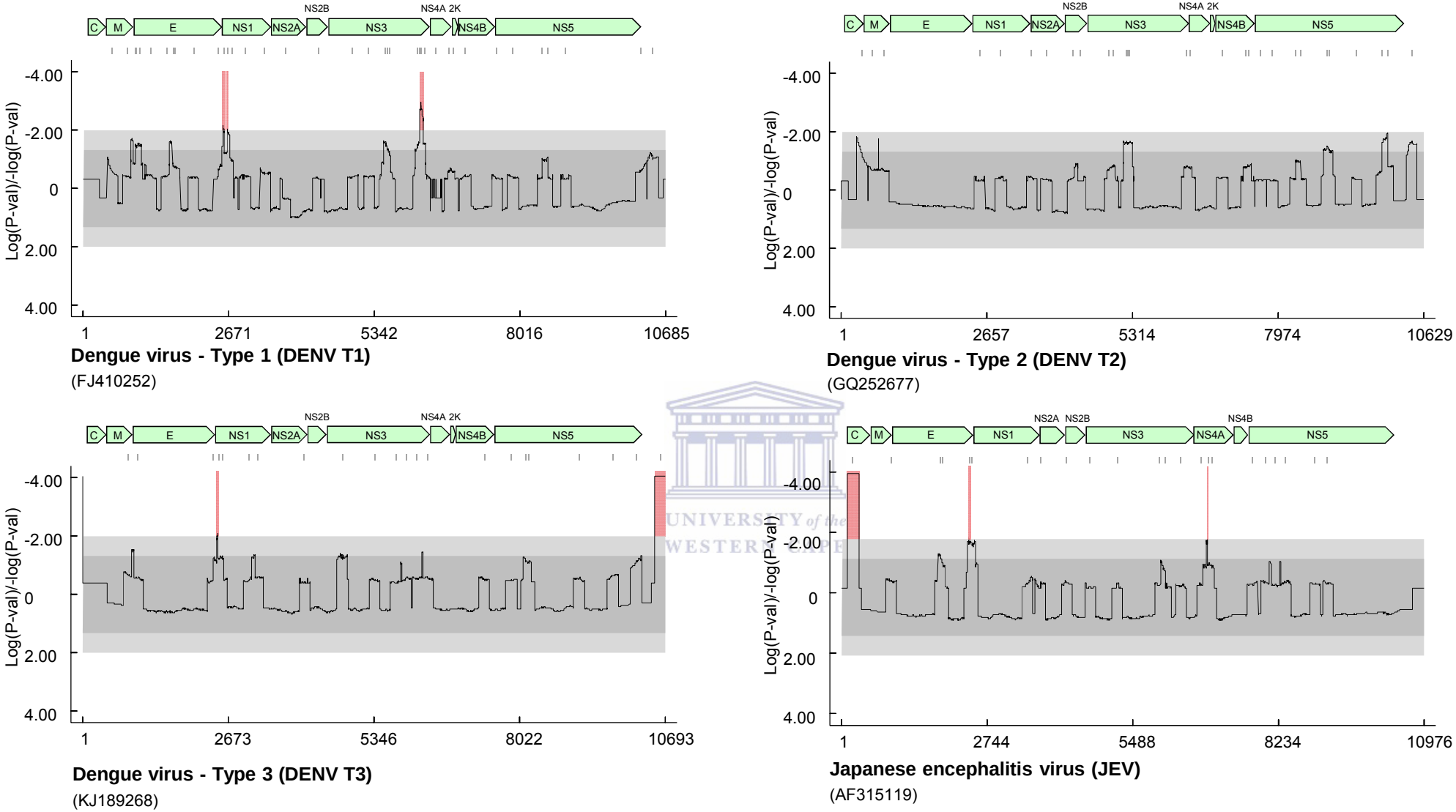


Figure 4.3 | Breakpoint distribution plots of *Flaviviridae* datasets The distribution of recombination breakpoints detected within,DENV T1, DENV T2, DENV T3 and JEV datasets. Estimated breakpoint positions are indicated by small vertical lines at the top of the graph. Light and dark grey areas, indicate local 99% and 95% breakpoint clustering thresholds, respectively. Red areas indicate recombination hot-spots, while blue areas represent recombinational cold-spots. Gene maps, in green, are drawn to scale in relation to the sequence of interest (accession number in brackets below each figure).

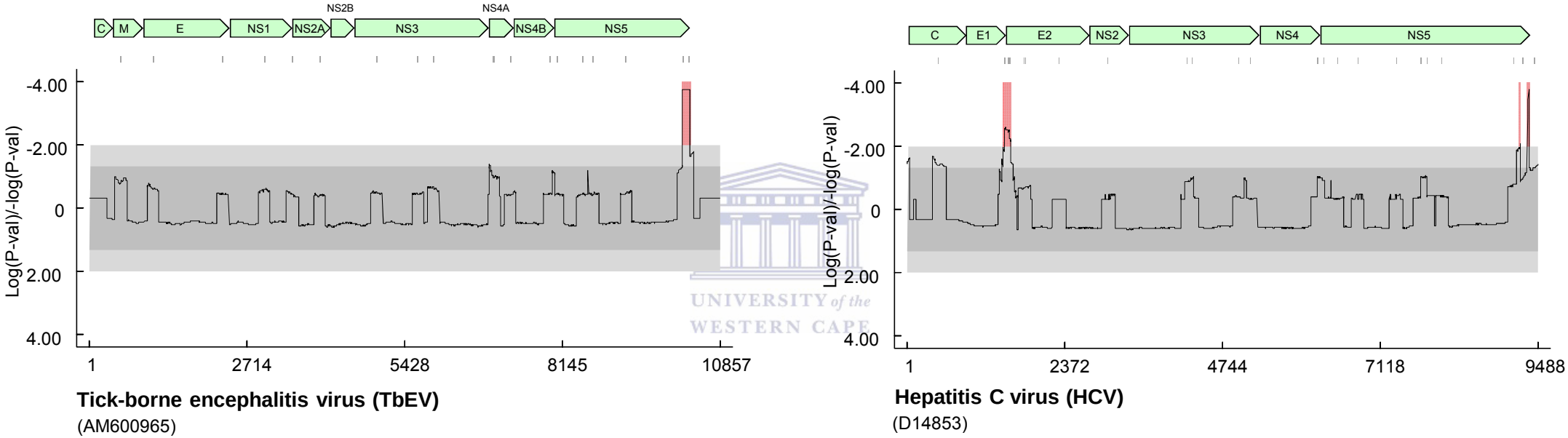


Figure 4.3 Continued | Breakpoint distribution plots of *Flaviviridae* datasets The distribution of recombination breakpoints detected within TbEV and HCV datasets. Estimated breakpoint positions are indicated by small vertical lines at the top of the graph. Light and dark grey areas, indicate local 99% and 95% breakpoint clustering thresholds, respectively. Red areas indicate recombination hot-spots, while blue areas represent recombinational cold-spots. Gene maps, in green, are drawn to scale in relation to the sequence of interest (accession number in brackets below each figure).

4.4.1.3 Potyviridae

The PVY, SMV, BYMV and TuMV datasets of the *Potyviridae* family appeared to have a conserved recombination hot-spot at the C-terminus of the HcPro gene. Another significant hot-spot of recombination was detected at the N-terminus of the CP gene in the PVY, SCMV, WMV and BYMV datasets. It is also notable that an absence of recombination breakpoints within the structural CP genes is apparent in four of the eight datasets analysed here (PVY, WMV, PRSV and BCMV). Although I did not specifically test whether recombination breakpoints tend to occur less frequently within the structural genes, it is plausible that natural selection is acting to avoid disrupting these genes and any intra-genome interactions they may participate in (Heath et al., 2006; Lukashev et al., 2005). In the PRSV and TuMV datasets there was a significant recombination hot-spot towards the 3'-end of the VPg gene, and although the PVY and BCMV datasets displayed breakpoint clusters in the same region, these were not significant.

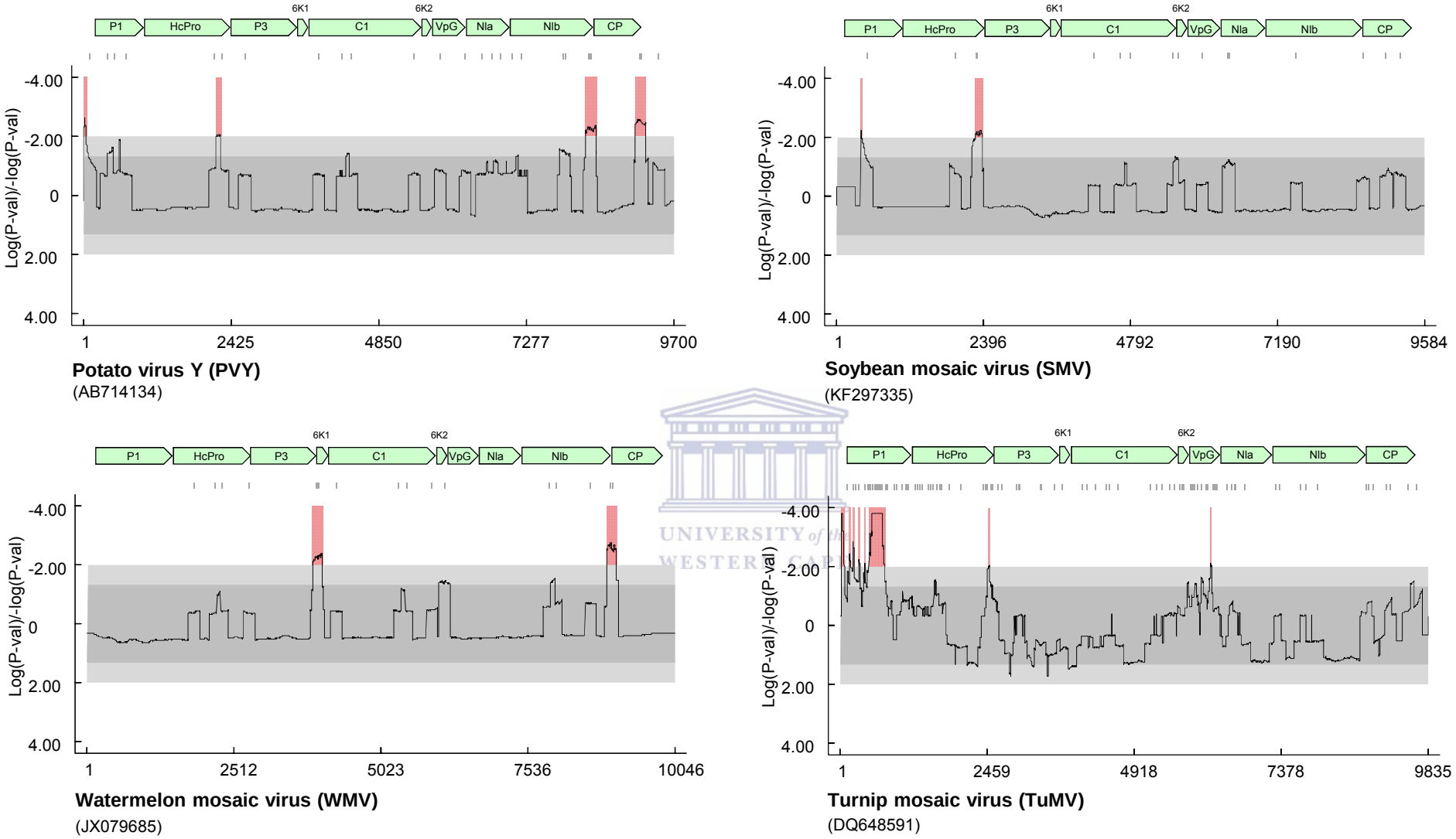


Figure 4.4 | Breakpoint distribution plots of *Potyviridae* datasets The distribution of recombination breakpoints detected within PVY, SMV, WMV and TuMV datasets. Estimated breakpoint positions are indicated by small vertical lines at the top of the graph. Light and dark grey areas, indicate local 99% and 95% breakpoint clustering thresholds, respectively. Red areas indicate recombination hot-spots, while blue areas represent recombinational cold-spots. Gene maps, in green, are drawn to scale in relation to the sequence of interest (accession number in brackets below each figure).

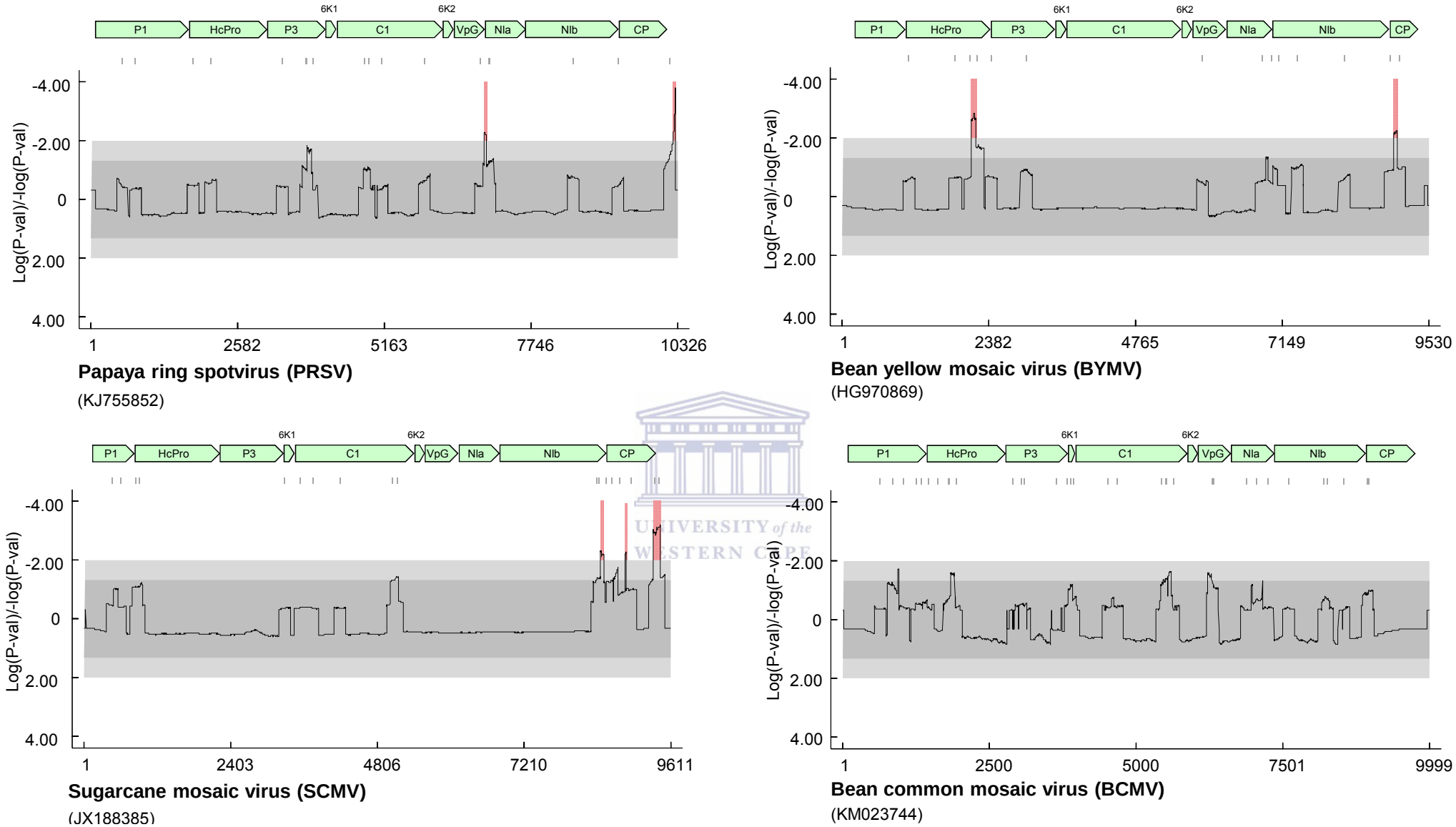


Figure 4.4 | Breakpoint distribution plots of *Potyviridae* datasets The distribution of recombination breakpoints detected within PRSV, BYMV, SCMV and BCMV datasets. Estimated breakpoint positions are indicated by small vertical lines at the top of the graph. Light and dark grey areas, indicate local 99% and 95% breakpoint clustering thresholds, respectively. Red areas indicate recombination hot-spots, while blue areas represent recombinational cold-spots. Gene maps, in green, are drawn to scale in relation to the sequence of interest (accession number in brackets below each figure).

4.4.1.4 Arteriviridae

In the only dataset of the *Arteriviridae* family of viruses analysed here, PRRSV, there were three significant recombination breakpoint clusters detected. The first was at the ribosomal frameshift site of the 1a' and 1a genes, the second within the 1b gene and thirdly, at the overlap of the GP3 and GP4 genes.

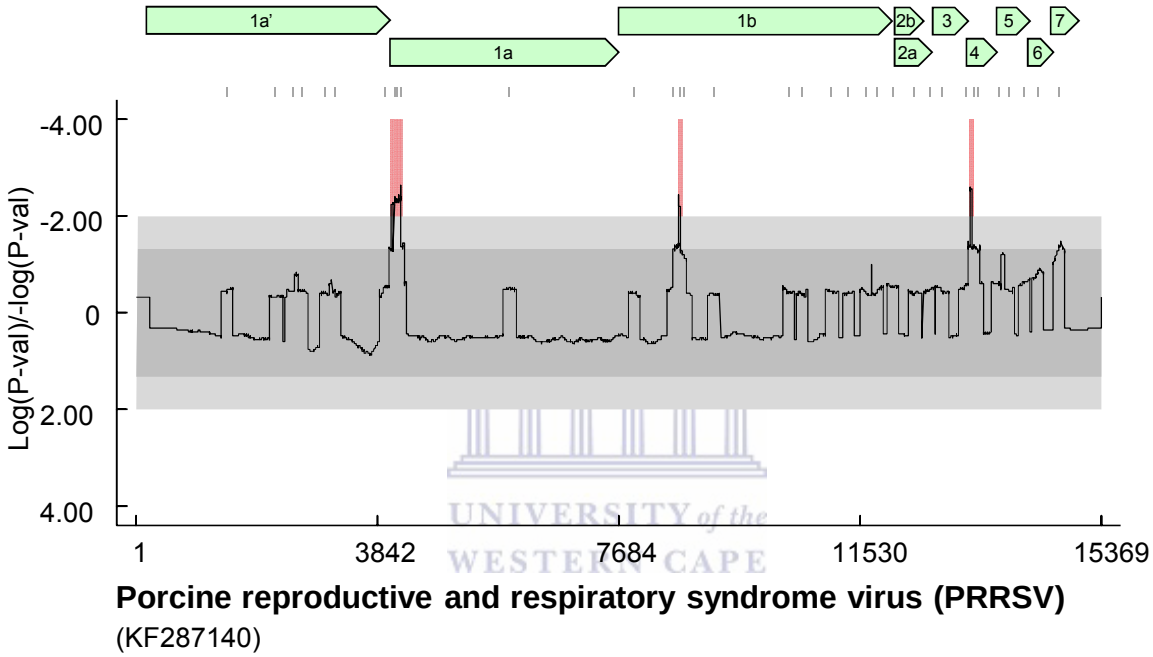


Figure 4.5 | Breakpoint distribution plot of an *Arterivirdae* dataset The distribution of recombination breakpoints detected within the PRRSV dataset. Estimated breakpoint positions are indicated by small vertical lines at the top of the graph. Light and dark grey areas, indicate local 99% and 95% breakpoint clustering thresholds, respectively. Red areas indicate recombination hot-spots, while blue areas represent recombinational cold-spots. Gene maps, in green, are drawn to scale in relation to the sequence of interest (accession number in brackets below each figure).

4.4.1.5 Closteroviridae

The CTV dataset contained four significant breakpoint clusters, two in close proximity to each other within the middle part of the 1a gene, another one at the C-terminus of the 1b gene, and lastly, one at the overlap of the p13 and p20 genes.

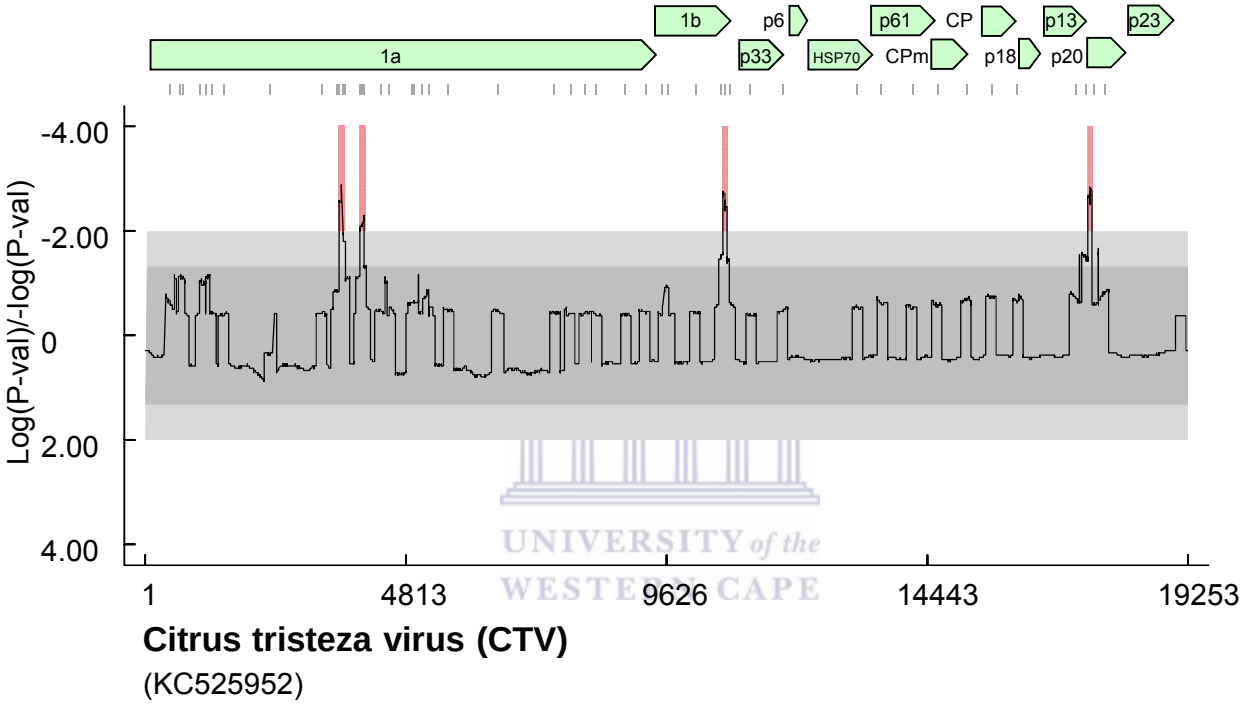


Figure 4.6 | Breakpoint distribution plot of a *Closteroviridae* dataset Distribution of recombination breakpoints detected within the CTV dataset. Estimated breakpoint positions are indicated by small vertical lines at the top of the graph. Light and dark grey areas, indicate local 99% and 95% breakpoint clustering thresholds, respectively. Red areas indicate recombination hot-spots, while blue areas represent recombinational cold-spots. Gene maps, in green, are drawn to scale in relation to the sequence of interest (accession number in brackets below each figure).

4.4.1.6 Orthohepeviridae

The recombination density plot of the HEV dataset shows a single significant recombination hot-spot at the region between the NSP and CP genes. This is also the N-terminus of the overlapping HP gene.

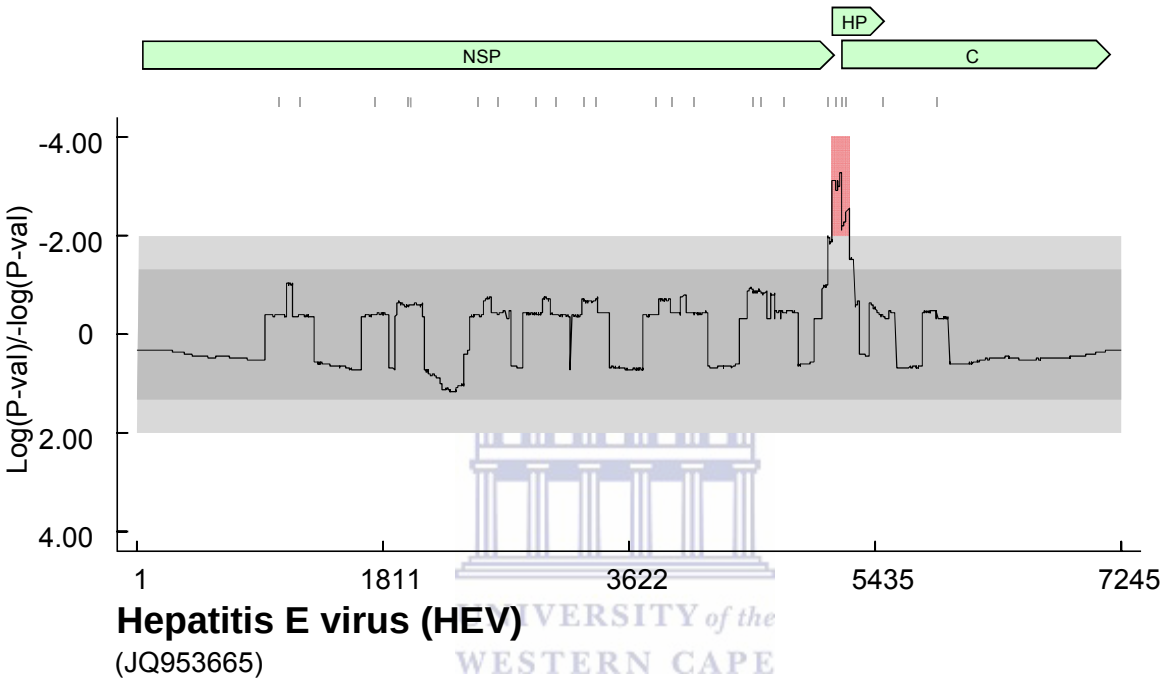


Figure 4.7 | Breakpoint distribution plot of an *Orthohepeviridae* dataset The distribution of recombination breakpoints detected within the HEV dataset. Estimated breakpoint positions are indicated by small vertical lines at the top of the graph. Light and dark grey areas, indicate local 99% and 95% breakpoint clustering thresholds, respectively. Red areas indicate recombination hot-spots, while blue areas represent recombinational cold-spots. Gene maps, in green, are drawn to scale in relation to the sequence of interest (accession number in brackets below each figure).

4.4.1.7 Luteoviridae

In the BYDV dataset there was evidence of significant breakpoint clustering towards the C-terminus of the gp1 gene. Another significant hot-spot of recombination was detected in the non-coding region between the gp6 and gp7 genes.

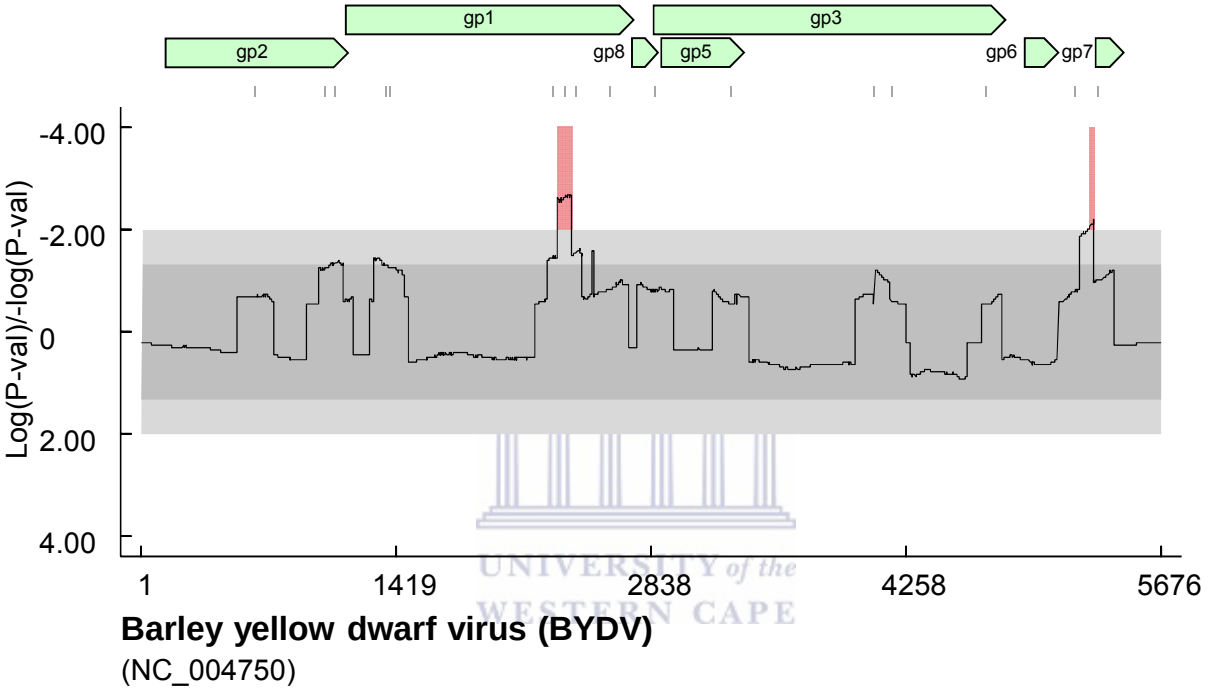


Figure 4.8 | Breakpoint distribution plots of a *Luteoviridae* dataset The distribution of recombination breakpoints detected within the BYDV dataset. Estimated breakpoint positions are indicated by small vertical lines at the top of the graph. Light and dark grey areas, indicate local 99% and 95% breakpoint clustering thresholds, respectively. Red areas indicate recombination hot-spots, while blue areas represent recombinational cold-spots. Gene maps, in green, are drawn to scale in relation to the sequence of interest (accession number in brackets below each figure).

4.4.2 Association between breakpoint location and gene boundaries

It was evident, from the majority of the breakpoint distribution plots, that there was a tendency for detectable recombination hot-spots to occur either at the edges of genes, or within non-coding intergenic regions, rather than within the middle of genes. For example, in the *Potyviridae* datasets, 13/16 significant breakpoint hot-spots appeared to occur at, or near, the boundaries of coding regions. Similar results were found in the *Picornaviridae* (28/34) and *Flaviviridae* (9/10) datasets. To test whether there were greater or fewer numbers of recombination breakpoints at the edges of genes than could be accounted for by chance, gene coordinates for each dataset were mapped to the corresponding permuted breakpoint density plots and the number of breakpoints occurring within the beginning and ending 5% of gene coding regions, rather than in the remaining 90%, were compared using a binary variable test. I detected a significant tendency ($P \leq 0.05$), in 16/30 of the datasets analysed here, for breakpoints to occur within the edges of their genes (beginning and ending 5%), versus the middle parts (middle 90%) (Table 4.1). This tendency was most obvious in the *Picornaviridae* and *Flaviviridae* datasets, where in 8/12 and 4/6 of the datasets tested (67% in both cases), recombination breakpoints occurred significantly ($P \leq 0.05$) more often at the edges of genes versus the remaining part of the genes. In the *Potyviridae* datasets however, only two of the eight datasets (WMV and SCMV – 25%), showed significant clustering ($P \leq 0.05$) of breakpoints at the boundaries of gene regions as compared to the middle of the genes. Both, the *Orthohepeviridae* and *Luteoviridae* datasets (HEV and BYDV, respectively) displayed significant ($P \leq 0.05$) localization of recombination breakpoints at the edges of genes, while the *Arteriviridae* and *Closteroviridae* datasets (PRRSV and CTV, respectively) showed no such evidence ($P > 0.05$).

Table 4.1 | Recombination based tests for clustering of detected breakpoints at the edges of genes

Family	Species	No. of sequences [a]	Recomb. events [b]	Breakpoints at edges of genes vs middle parts [c]
Arteriviridae	Porcine reproductive and respiratory syndrome virus (PRRSV)	261	22/72	0.238
Closteroviridae	Citrus tristeza virus (CTV)	44	34/98	0.507
Flaviviridae	Dengue virus - Type 1 (DENV T1)	1548	27/101	<0.001
	Dengue virus - Type 2 (DENV T2)	1154	20/65	0.288

Table 4.1 Continued

Family	Species	No. of sequences [a]	Recomb. events [b]	Breakpoints at edges of genes vs middle parts [c]
Flaviviridae	Dengue virus - Type 3 (DENV T3)	862	13/34	0.331
	Japanese encephalitis virus (JEV)	166	13/27	0.028
	Tick-borne encephalitis virus (TbEV)	93	11/22	0.032
	Hepatitis C virus (HCV)	555	19/71	0.026
Orthohepeviridae	Hepatitis E virus (HEV)	196	12/21	0.018
Luteoviridae	Barley yellow dwarf virus (BYDV)	89	24/53	0.05
Picornaviridae	Foot-and-mouth disease virus (FMDV)	402	56/170	0.05
	Human enterovirus A (HEV-A)	399	52/149	0.001
	Human enterovirus B (HEV-B)	244	136/331	0.05
	Human enterovirus C (HEV-C)	417	111/397	0.027
	Human rhinovirus A (HRV-A)	184	33/86	0.884
	Human rhinovirus B (HRV-B)	67	10/19	0.024
	Human rhinovirus C (HRV-C)	62	23/58	0.721
	Hepatitis virus A (HAV)	84	11/16	0.219
	Saffold virus (SAFV)	60	26/92	0.034
	Porcine teschovirus (PTeV)	57	35/82	<0.001
	Human parechovirus (HPeV)	83	24/81	0.592
	Duck hepatitis A virus (DuHAV)	110	20/27	0.035
Potyviridae	Bean common mosaic virus (BCMV)	42	27/58	0.781
	Bean yellow mosaic virus (BYMV)	40	12/48	0.214
	Potato virus Y (PVY)	238	27/65	0.731
	Soybean mosaic virus (SMV)	42	13/30	0.865
	Turnip mosaic virus (TuMV)	219	76/120	0.276
	Papaya ringspot virus (PRSV)	29	10/27	0.841
	Watermelon mosaic virus (WMV)	34	18/49	0.013
	Sugarcane mosaic virus (SCMV)	22	11/33	0.023

[a] Total number of sequences used for recombination breakpoint detection.

[b] The number of accepted recombination events over the total number of recombination events detected.

[c] The p value associated with breakpoints occurring at beginning and ending 5% of genes versus the middle parts of genes.

Datasets displaying significant evidence (P <0.05) of breakpoints preferentially clustering at gene borders are highlighted in grey.

4.4.3 Influence of secondary structure on recombination breakpoint distributions

The breakpoint distribution analysis results suggest that in many of the ssRNA virus datasets, the distribution of detected recombination breakpoints is non-random, showing breakpoints to preferentially occur at the edges of protein coding regions. While it is not obvious what the mechanism responsible for these patterns may be, there are reports in enteroviruses

(Dedepsidis et al., 2010) and potyviruses (Draghici and Varrelmann, 2010) suggesting correlation between positions of conserved secondary structure and breakpoint locations, where structural elements, and in particular stem regions, may play an important role in promoting and facilitating recombination. I set out to test whether detected recombination breakpoints co-localised with regions of the HCSSs that were predicted to be base paired.

However, I found there was no convincing evidence from the datasets tested here, that there is a strong association between the coordinates of predicted RNA structural elements and recombination breakpoint distributions. Except for the TuMV dataset, none of the other alignments yielded a significant ($P \leq 0.05$) association between predicted structured regions and breakpoint locations. Notably, the JEV, HAV and PVY datasets produced marginally significant p value's ($0.05 \leq P \leq 0.1$) for these tests.

Table 4.2 | Recombination based tests for association between detected breakpoint locations and co-ordinates of predicted secondary structures

Family	Species	No. of sequences [a]	Recomb. events [b]	Breakpoint co-localisation with secondary structure [c]
Arteriviridae	Porcine reproductive and respiratory syndrome virus (PRRSV)	261	22/72	0.63
Closteroviridae	Citrus tristeza virus (CTV)	44	34/98	0.95
Flaviviridae	Dengue virus - Type 1 (DENV T1)	1548	27/101	0.72
	Dengue virus - Type 2 (DENV T2)	1154	20/65	0.26
	Dengue virus - Type 3 (DENV T3)	862	13/34	0.44
	Japanese encephalitis virus (JEV)	166	13/27	0.09
	Tick-borne encephalitis virus (TbEV)	93	11/22	0.52
	Hepatitis C virus (HCV)	555	19/71	0.89
Orthohepeviridae	Hepatitis E virus (HEV)	196	12/21	0.21
Luteoviridae	Barley yellow dwarf virus (BYDV)	89	24/53	0.34
Picornaviridae	Foot-and-mouth disease virus (FMDV)	402	56/170	0.99
	Human enterovirus A (HEV-A)	399	52/149	0.78
	Human enterovirus B (HEV-B)	244	136/331	0.23
	Human enterovirus C (HEV-C)	417	111/397	0.65
	Human rhinovirus A (HRV-A)	184	33/86	0.95
	Human rhinovirus B (HRV-B)	67	10/19	0.72
	Human rhinovirus C (HRV-C)	62	23/58	0.64
	Hepatitis virus A (HAV)	84	11/16	0.08
	Saffold virus (SAFV)	60	26/92	0.54

Table 4.2 Continued

Family	Species	No. of sequences [a]	Recomb. events [b]	Breakpoint co-localisation with secondary structure [c]
Picornaviridae	Porcine teschovirus (PTeV)	57	35/82	0.83
	Human parechovirus (HPeV)	83	24/81	0.96
	Duck hepatitis A virus (DuHAV)	110	20/27	0.83
Potyviridae	Bean common mosaic virus (BCMV)	42	27/58	0.25
	Bean yellow mosaic virus (BYMV)	40	12/48	0.74
	Potato virus Y (PVY)	238	27/65	0.06
	Soybean mosaic virus (SMV)	42	13/30	0.85
	Turnip mosaic virus (TuMV)	219	76/120	0.03
	Papaya ringspot virus (PRSV)	29	10/27	0.93
	Watermelon mosaic virus (WMV)	34	18/49	0.75
	Sugarcane mosaic virus (SCMV)	22	11/33	0.71

[a] Total number of sequences used for recombination breakpoint detection.

[b] The number of accepted recombination events over the total number of recombination events detected.

[c] p value associated with the binary variable test for association between locations of detected breakpoints and co-ordinates of predicted secondary structures.

4.5 Discussion

4.5.1 Tendency for recombination breakpoints to fall at the edges of genes

Generally, the genomes of (+)ssRNA viruses contain single polyprotein encoding ORFs and, besides the 5' and 3' UTRs, these genomes lack extensive non-coding intergenic regions. Therefore, breakpoints detected within the UTRs were discarded and I specifically tested whether recombination breakpoints tend to occur more often near the boundaries (5% at either end) of demarcated gene regions instead of in the middle parts (90% middle region) of genes, than could be accounted for by chance. Significant associations (P values ≤ 0.05), between the detected locations of breakpoints occurring at the edges of genes versus the middle parts, were observed in a large proportion (53%) of the datasets tested. It is notable that significant associations were detected in datasets, with both, few (eg. TbEV, HEV and HRV-B) as well as many (eg. FMDV, HEV-A, HEV-C) recombination breakpoints, suggesting that purifying selection may be purging recombinants with breakpoints that fall within the middle parts of coding regions. Other studies of recombination patterns in

(+)ssRNA viruses (Fu and Baric, 1994; Pagan and Holmes, 2010), have also shown that recombination breakpoints tend to occur at gene boundaries rather than within the central regions of genes, suggesting that viable exchanges of genetic material amongst viral genomes involves a transfer of intact or nearly intact gene "modules". This supports the hypothesis (Botstein, 1980; Dolja and Carrington, 1992), that modular evolution has been a major mechanism of promoting the genetic mosaics seen amongst many recombinant vertebrate and plant ssRNA viruses. Although it was apparent in many of the datasets, that breakpoint patterns are non-randomly distributed, it is not clear whether there is either specific selective processes acting on these genomes, or particular genetic features of these genomes are responsible for the observed breakpoint distributions. It is, for example conceivable that there are conserved structural features of these genomes that might be facilitating recombination hot-spots at particular genome locations and, in so doing, these features might shepherd breakpoints to particular genome sites where they will be most likely to yield viable recombinant progeny (Figlerowicz, 2000; Simon-Loriere et al., 2010).

4.5.2 Weak evidence that recombination breakpoints preferentially occur within genomic secondary structures

While it is evident that there is likely a common mechanism responsible for the recombination patterns observed among many of the (+)ssRNA genomes, it is not obvious what this mechanism might be. Studies have suggested that the locations of cross-over junctions in some ssRNA viruses (Carpenter et al., 1995; Cascone et al., 1993; Nagy et al., 1999) preferentially occur at, or in close proximity to, conserved secondary structures and depend largely on the stability and conformation of these elements.

I tested whether recombination breakpoints occurred more frequently than could be accounted for by completely random recombination, at sites that were predicted to be base-paired. The results presented here however, do not provide any convincing evidence that there is an association between predicted RNA structural elements and breakpoint clustering in the (+)ssRNA genomes tested here (Table 4.2). The TuMV alignment was the only dataset that showed a significant association ($P \leq 0.05$) between recombination breakpoints and secondary structures, and although marginally insignificant associations ($0.05 \leq P \leq 0.1$) between these variables were detected for the JEV, HAV and PVY datasets, it is unlikely that the distribution of genomic secondary structures is a major determinant of the patterns of

recombination breakpoints observed here. Similar results have been reported in other ssRNA- (Simmonds and Welch, 2006) and ssDNA-viruses (Martin et al., 2011). Since a number of structural elements in these viruses facilitate other important biological processes, it is plausible that selection might be strongly disfavouring recombinants where the integrity of functional secondary structures is compromised by recombination breakpoints within these structures. It is important to note that results presented here do not necessarily mean that local RNA secondary structure cannot/does not facilitate recombination, but rather that such structures do not account overwhelmingly for the recombination breakpoints detected in this study.

4.6 Conclusions

Although the numbers of detectable recombination events varied amongst the datasets tested here, these numbers are consistent with previous studies reporting evidence of high recombination rates in viruses of the families *Picornaviridae* (Savolainen-Kopra and Blomqvist, 2010), *Potyviridae* (Gibbs and Ohshima, 2010), and *Luteoviridae* (Pagan and Holmes, 2010), and relatively lower recombination frequencies in the family *Flaviviridae* (Taucher et al., 2010).

Nevertheless, irrespective of the frequency of recombination detected within members of these viral families, breakpoint patterns appeared to be broadly conserved. While recombination detection analysis showed that recombination breakpoints preferentially clustered at specific sites along (+)ssRNA virus genomes, it was not immediately obvious what mechanism is responsible for this observation. I compared breakpoint distributions against genome coordinates with high base-pairing probabilities, showing that, despite reports to the contrary in other ssRNA viruses, there is no clear evidence of statistically significant association between the breakpoint patterns I detected and predicted secondary structures of the HCSSs that I inferred. A solitary significant association was detected in the TuMV dataset and marginally insignificant evidence was observed in three other datasets (PVY, HAV and JEV). This might be an indication of selection acting to preserve biologically functional base-pair interactions in these viruses and this could be specifically tested for in future analysis. It is important to note however, that in many of the datasets, relatively few recombination events were detected (12/30 datasets with < 20), which may have

underpowered the strength of our tests and future analysis considering larger datasets could prove useful in detecting such associations.

While it seems that secondary structures have no profound effects on recombination breakpoint patterns, I showed that the recombination patterns in (+)ssRNA viruses were strongly influenced by the organization of viral genomes. Specifically, among many of the analysed datasets, there was a significant tendency for recombination breakpoints to cluster at gene boundaries. When breakpoints fell within coding regions they preferentially occurred within the ending 5% of genes (Table 4.1): a pattern consistent with the hypothesis that recombination breakpoints that fall in the middle of genes have a greater probability of yielding dysfunctional chimaeric proteins, than breakpoints that fall at the edges of genes (Bonnet et al., 2005; Voigt et al., 2002).

It still remains unclear what general mechanistic factors, if any, are responsible for the recombination breakpoints observed here, but future studies, perhaps focusing on specific regions of conserved base-pairing and cross-over junctions, could expose such elements influencing preferred breakpoint locations. In addition, the breakpoint clustering locations presented here could serve as a platform to infer potential intra-genome interactions, since it is conceivable that the recombination patterns detected here are influenced by such intra-genome relationships.

Chapter 5: Conclusions and recommendations

5.1 Summary of findings

Secondary structure elements in nucleic acid molecules are encoded as a complementary layer of information to the underlying primary sequence. In viruses, many of these structures have been shown to contribute significantly to the evolved mechanisms these organisms employ to replicate and evade the host immune systems. Therefore, those structures that infer a fitness advantage are expected to produce clear signals of selection consistent with their maintenance. In this study I used a battery of bioinformatics analyses to assess to what degree secondary structures have impacted the evolution of ssRNA viral genomes across several families of viruses.

5.1.1 Conserved secondary structures in ssRNA viruses

Although it is not possible to determine the actual function of the predicted structures through a purely computational approach, the combination of selection tests provide a reliable indication of the structures that are most likely to be biologically relevant.

The secondary structure prediction program NASP provided the first line of evidence, with high-confidence, that the genomes of these viruses contained numerous evolutionarily conserved (and therefore potentially functional) secondary structures. Although the number of predicted structures varied amongst the different datasets (1000 structures detected in CTV, 67 in CGMV) there was sufficient data to conduct the necessary evolutionary analyses in all the datasets. To analyse to what degree the predicted structures have constrained the patterns of evolution in these viruses a combination of selection tests were used to estimate, synonymous substitution rates, the degree of purifying selection, and complementary co-evolution interactions. This large-scale analysis revealed that in a large proportion of the data sets synonymous substitution rates in the coding regions of these viruses were significantly lower than expected at sites predicted to be paired, suggesting that the maintenace of structures in those viruses likely has priority over the evolution of the underlying protein coding sequence. Similarly, the results of the Tajima's D- and Fu and Li F-based tests on full-length genome sequences (including coding and non-coding regions) showed that in many of the datasets there was significant evidence of stronger purifying selection at paired sites than

at unpaired sites. In contrast, only about a third of the datasets displayed strong evidence that natural selection favours complementary co-evolution of nucleotides predicted to be base-paired. While these results show that in many of the datasets selectively maintained, and therefore probably functional, secondary structures do occur, the abundance of less conserved elements in some of the datasets suggests that many of those less conserved structures are likely not directly involved in specific mechanisms or processes but rather function as a collective organised unit. It has been shown that the highly structured genome of some ssRNA viruses, termed Genome-scale Ordered RNA Structure (GORS), is associated with their capacity to establish long-term persistence in their hosts (Simmonds et al., 2004). Viruses causing chronic disease in mammalian hosts (among many of them analysed here include flaviviruses, caliciviruses and picornaviruses) have been shown to adopt pseudo-globular conformations, and it is possible that these extensive genome-wide structures have evolved as a way to counteract host defences by mimicking the host's structured RNAs (rRNAs, tRNAs) thus avoiding detection by RNA interference (RNAi) factors (Davis et al., 2008).

Furthermore, the results of the molecular selection tests above provided a means to quantitatively rank the predicted structures in order of their likely biological importance. Using these lists, I identified specific structures that might be functionally important in these viruses. More importantly, many well characterised structures were amongst the highest ranking structures (Supplementary Table 1) which provided further validation that the approach employed in this study can be a reliable means of identifying potentially functional structures.

In addition, I used the HCCS lists to test whether the recombination patterns detected in these viruses are influenced by the predicted secondary structures. The results showed that in only one of the datasets there was significant association between the observed breakpoint locations and the co-ordinates of secondary structures, whereas there was evidence in several datasets that breakpoints locations were selectively constrained by the coordinates of protein coding regions. More specifically, the number of breakpoints detected across the genomes were significantly higher at the edges of genes than the remainder of those genes. Based on these results, protein crystallisation data could be employed to further test whether the observed recombination evens are less disruptive of amino-acid─amino-acid protein fold interactions than could be accounted for by chance.

Collectively, these results provide evidence that natural selection acting to preserve important molecular interactions within biologically functional secondary structures has at least weakly constrained the evolution of some (+)ssRNA viruses.

5.2 Major challenges

The process of evolution is constrained by fundamental physiochemical principals, among which is thermodynamics. Using thermodynamics alone it is possible to correctly predict around 70% of pairing interactions in a sequence of no more than 700 nucleotides (MATHEWS, 2004). However, many larger ssRNA and ssDNA molecules have the capacity to fold into several alternate metastable structures and the conformational space is enormous. For example, a single stranded nucleic-acid molecule of length *n* can fold into 1.8^n alternate secondary structures (Zuker and Sankoff, 1984) and therefore absolutely accurate prediction of existing secondary structure conformations is challenging. Another limitation of the minimum free-energy prediction approach is the exclusive use of thermodynamic rules. While this provides a relatively accurate depiction of existing nucleotide pairing interactions, the exact conformation of functional structures is likely to be influenced by interactions with proteins and other molecules. This is an inherent problem due to our limited understanding of the dynamics between structure and evolution at the molecular level.

A major limiting factor in many of the analyses performed in this study was the computational complexity inherent in resolving the secondary structure profiles for some of the larger datasets. The time complexity of NASP is $O(NPL^3)$ where L is the length, N the number of sequences in the alignment and P the number of permutation tests (Semegni et al., 2011). This translated into approx. 144hr required to analyse datasets comprised of 10 sequences of 10kb each, using eight processors on a parallel computing cluster.

5.3 Outlook

Identifying potentially functional structures among numerous predicted candidates, by mining the literature for characterised structures and comparing it to the computational output, can be a time-consuming and laborious task. A more efficient and automated approach is needed for finding homologous functional structures among divergent sequences, that could rely on both evolutionary information and folding free-energy similarities. While several algorithms that

compare secondary structures have been developed, the analyses are generally limited to relatively short and small datasets, and rely on external folding methods for predicting structures, which may not be entirely accurate (Allali and Sagot, 2008; Dulucq and Tichit, 2003; Hofacker et al., 1994; Schirmer et al., 2014). Moreover, they are not able to separate, and compare, individual elements from the global-structure profile. Besides identifying potentially functional homologues, automated methods for accurate large-scale comparison of evolutionarily preserved structures could be useful in guiding taxonomic classification of phylogenetically related virus sequences where unassigned and newly discovered viruses can be assigned to taxonomic groups based on common secondary structures.

As has been demonstrated in this thesis, incorporating evolutionary information to constrain computational prediction methods can contribute significantly to refining the empirical thermodynamic prediction parameters that are currently available, leading to more accurate, and therefore more biologically meaningful, predictions of secondary structure.

References

Adams, M.J., Lefkowitz, E.J., King, A.M.Q., Harrach, B., Harrison, R.L., Knowles, N.J., Kropinski, A.M., Krupovic, M., Kuhn, J.H., Mushegian, A.R., Nibert, M., Sabanadzovic, S., Sanfaçon, H., Siddell, S.G., Simmonds, P., Varsani, A., Zerbini, F.M., Gorbalenya, A.E., Davison, A.J., 2017. Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2017). Arch. Virol. 162, 2505–2538.

Ahlquist, P., Noueiry, A.O., Lee, W., David, B., Dye, B.T., 2003. Host Factors in Positive-Strand RNA Virus Genome Replication. J. Virol. 77, 8181–8186.

Akaike, T., 2001. Role of free radicals in viral pathogenesis and mutation. Rev. Med. Virol. 11, 87–101.

Ali, S., Gugliemini, O., Harber, S., Harrison, A., Houle, L., Ivory, J., Kersten, S., Khan, R., Kim, J., LeBoa, C., Nez-Whitfield, E., O'Marr, J., Rothenberg, E., Segnitz, R.M., Sila, S., Verwillow, A., Vogt, M., Yang, A., Mordecai, E.A., 2017. Environmental and Social Change Drive the Explosive Emergence of Zika Virus in the Americas. PLoS Negl. Trop. Dis.

Allali, J., Sagot, M.F., 2008. A multiple layer model to compare RNA secondary structures. Softw. - Pract. Exp. 38, 775–792.

Andronescu, M.S., Pop, C., Condon, A.E., 2010. Improved free energy parameters for RNA pseudoknotted secondary structure prediction. RNA 16, 26–42.

Anisimova, M., Nielsen, R., Yang, Z., 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164, 1229–36.

Athavale, S.S., Gossett, J.J., Bowman, J.C., Hud, N. V., de Williams, L., Harvey, S.C., 2013. In Vitro Secondary Structure of the Genomic RNA of Satellite Tobacco Mosaic Virus. PLoS One 8.

Bailey, D., Karakasiliotis, I., Vashist, S., Chung, L.M.W., Reese, J., McFadden, N., Benson, A., Yarovinsky, F., Simmonds, P., Goodfellow, I., 2010. Functional Analysis of RNA Structures Present at the 3' Extremity of the Murine Norovirus Genome: the Variable Polypyrimidine Tract Plays a Role in Viral Virulence. J. Virol. 84, 2859–2870.

Bailey, J.M., Tapprich, W.E., 2007. Structure of the 5’ Nontranslated Region of the Coxsackievirus B3 Genome: Chemical Modification and Comparative Sequence Analysis. J. Virol. 81, 650–668.

Barciszewska, M.Z., Szymański, M., Erdmann, V. a, Barciszewski, J., 2001. Structure and functions of 5S rRNA. Acta Biochim. Pol. 48, 191–8.

Baric, R.S., Fu, K., Schaad, M.C., Stohlman, S.A., 1990. Establishing a genetic recombination map for murine coronavirus strain A59 complementation groups. Virology 177, 646–656.

Belsham, G.J., 2009. Divergent picornavirus IRES elements. Virus Res. 139, 183–192.

Bernhart, S.H., Hofacker, I.L., Will, S., Gruber, A.R., Stadler, P.F., 2008. RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinformatics 9, 474.

Biebricher, C.K., 2008. Mutation, Competition, and Selection as Measured with Small RNA Molecules. In: Origin and Evolution of Viruses. pp. 65–85.

Bishop, K.N., Holmes, R.K., Sheehy, A.M., Davidson, N.O., Cho, S.J., Malim, M.H., 2004. Cytidine deamination of retroviral DNA by diverse APOBEC proteins. Curr. Biol. 14, 1392–1396.

Boni, M.F., Posada, D., Feldman, M.W., 2007. An exact nonparametric method for inferring mosaic structure in sequence triplets. Genetics 176, 1035–1047.

Bonnet, J., Fraile, A., Sacristán, S., Malpica, J.M., García-Arenal, F., 2005. Role of recombination in the evolution of natural populations of Cucumber mosaic virus, a tripartite RNA plant virus. Virology 332, 359–368.

Botstein, D., 1980. A theory of modular evolution for bacteriophages. Ann. N. Y. Acad. Sci. 354, 484–491.

Brierley, I., Dos Ramos, F.J., 2006. Programmed ribosomal frameshifting in HIV-1 and the SARS-CoV. Virus Res. 119, 29–42.

Brimacombe, R., Stiege, W., 1985. Structure and function of ribosomal RNA. Biochem. J. 229, 1–17.

Brown, D.W.G., 1997. Threat to humans from virus infections of non-human primates. Rev.

Med. Virol.

Bruyere, A., Wantroba, M., Flasinski, S., Dzianott, A., Bujarski, J.J., 2000. Frequent Homologous Recombination Events between Molecules of One RNA Component in a Multipartite RNA Virus. J. Virol. 74, 4214–4219.

Buck, K.W., 1996. Comparison of The Replication of Positive-Stranded Rna Viruses of Plants and Animals. Adv. Virus Res. 47, 159–251.

Bujarski, J.J., 2013. Genetic recombination in plant-infecting messenger-sense RNA viruses: overview and research perspectives. Front. Plant Sci. 4, 68.

Burrill, C.P., Westesson, O., Schulte, M.B., Strings, V.R., Segal, M., Andino, R., 2013. Global RNA Structure Analysis of Poliovirus Identifies Conserved RNA Structure Involved in Viral Replication and Infectivity. J. Virol. 87, 11670–83.

Bushell, M., Sarnow, P., 2002. Hijacking the translation apparatus by RNA viruses. J. Cell Biol.

Carpenter, C., Oh, J., Zhang, C., Simon, A., 1995. Involvement of a stem-loop structure in the location of junction sites in viral RNA recombination. J. Mol. Biol. 435, 214–219.

Carrington, J.C., Freed, D.D., 1990. Cap-independent enhancement of translation by a plant potyvirus 5’ nontranslated region. J. Virol. 64, 1590–1597.

Cascone, P.J., Haydar, T.F., Simon, a E., 1993. Sequences and structures required for recombination between virus-associated RNAs. Science 260, 801–5.

Catalano, C.E., Cue, D., Feiss, M., 1995. Virus DNA packaging: the strategy used by phage λ. Mol. Microbiol.

Chambers, T.J., Hahn, C.S., Galler, R., Rice, C.M., 1990. Flavivirus Genome Organization, Expression, and Replication. Annu. Rev. Microbiol. 44, 649–688.

Charlesworth, B., Morgan, M.T., Charlesworth, D., 1993. The effect of deleterious mutations on neutral molecular variation. Genetics.

Chebli, K., Gattoni, R., Schmitt, P., Hildwein, G., Stevenin, J., 1989. The 216-nucleotide intron of the E1A pre-mRNA contains a hairpin structure that permits utilization of unusually distant branch acceptors. Mol. Cell. Biol. 9, 4852–61.

Chen, S.C., Desprez, A., Olsthoorn, R.C.L., 2010. Structural homology between bamboo mosaic virus and its satellite RNAs in the 5??? untranslated region. J. Gen. Virol. 91, 782–787.

Chen, Z., Znosko, B.M., 2013. Effect of sodium ions on RNA duplex stability. Biochemistry 52, 7477–7485.

Cheynier, R., Kils-H??tten, L., Meyerhans, A., Wain-Hobson, S., 2001. Insertion/deletion frequencies match those of point mutations in the hypervariable regions of the simian immunodeficiency virus surface envelope gene. J. Gen. Virol. 82, 1613–1619.

Choi, C., Chae, C., 2002. Localization of classical swine fever virus in male gonads during subclinical infection. J. Gen. Virol. 83, 2717–2721.

Cloete, L.J., Tanov, E.P., Muhire, B.M., Martin, D.P., Harkins, G.W., 2014. The influence of secondary structure, selection and recombination on rubella virus nucleotide substitution rate estimates. Virol. J. 11, 166.

Clyde, K., Barrera, J., Harris, E., 2008. The capsid-coding region hairpin element (cHP) is a critical determinant of dengue virus and West Nile virus RNA synthesis. Virology 379, 314–323.

Clyde, K., Harris, E., 2006. RNA secondary structure in the coding region of dengue virus type 2 directs translation start codon selection and is required for viral replication. J. Virol. 80, 2170–82.

Cordey, S., Gerlach, D., Junier, T., Zdobnov, E.M., Kaiser, L., Tapparel, C., 2008. The cis-acting replication elements define human enterovirus and rhinovirus species. RNA 14, 1568–1578.

Crary, S.M., Towner, J.S., Honig, J.E., Shoemaker, T.R., Nichol, S.T., 2003. Analysis of the role of predicted RNA secondary structures in Ebola virus replication. Virology 306, 210–218.

Cromie, M.J., Shi, Y., Latifi, T., Groisman, E.A., 2006. An RNA Sensor for Intracellular Mg2+. Cell 125, 71–84.

Daniel, T., Robbins, F., 1997. A history of polyomyelitis. University of Rochester Press, Rochester, New York.

Darty, K., Denise, A., Ponty, Y., 2009. VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics 25, 1974–1975.

Davis, M., Sagan, S.M., Pezacki, J.P., Evans, D.J., Simmonds, P., 2008. Bioinformatic and Physical Characterizations of Genome-Scale Ordered RNA Structure in Mammalian RNA Viruses. J. Virol. 82, 11824–11836.

De Leonardis, E., Lutz, B., Ratz, S., Cocco, S., Monasson, R., Schug, A., Weigt, M., 2015. Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. Nucleic Acids Res. 43, 10444–10455.

De Mita, S., Siol, M., 2012. EggLib: processing, analysis and simulation tools for population genetics and genomics. BMC Genet. 13, 27.

Dedepsidis, E., Kyriakopoulou, Z., Pliaka, V., Markoulatos, P., 2010. Correlation between recombination junctions and RNA secondary structure elements in poliovirus Sabin strains. Virus Genes 41, 181–191.

Deigan, K.E., Li, T.W., Mathews, D.H., Weeks, K.M., 2009. Accurate SHAPE-directed RNA structure determination. Proc. Natl. Acad. Sci. 106, 97–102.

Dela-Moss, L.I., Moss, W.N., Turner, D.H., 2014. Identification of conserved RNA secondary structures at influenza B and C splice sites reveals similarities and differences between influenza A, B, and C. BMC Res. Notes 7, 22.

Dolja, V., Carrington, J., 1992. Evolution of positive-strand RNA viruses. Semin. Vor. 3, 315–326.

Domingo, E., Holland, J.J., 1997. RNA VIRUS MUTATIONS AND FITNESS FOR SURVIVAL. Annu. Rev. Microbiol. 51, 151–178.

Dominski, Z., Zheng, L.X., Sanchez, R., Marzluff, W.F., 1999. Stem-loop binding protein facilitates 3’-end formation by stabilizing U7 snRNP binding to histone pre-mRNA. Mol. Cell. Biol. 19, 3561–3570.

Draghici, H.K., Varrelmann, M., 2010. Evidence for similarity-assisted recombination and predicted stem-loop structure determinant in potato virus X RNA recombination. J. Gen. Virol. 91, 552–562.

Drake, J.W., 1999. The distribution of rates of spontaneous mutation over viruses,

prokaryotes, and eukaryotes. In: Annals of the New York Academy of Sciences. pp. 100–107.

Duffy, S., Shackelton, L.A., Holmes, E.C., 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. 9, 267–276.

Duggal, R., Cuconati, a, Gromeier, M., Wimmer, E., 1997. Genetic recombination of poliovirus in a cell-free system. Proc. Natl. Acad. Sci. U. S. A. 94, 13786–91.

Dulucq, S., Tichit, L., 2003. RNA secondary structure comparison: Exact analysis of the Zhang-Shasha tree edit algorithm. Theor. Comput. Sci. 306, 471–484.

Dykes, C., Balakrishnan, M., Planelles, V., Zhu, Y., Bambara, R.A., Demeter, L.M., 2004. Identification of a preferred region for recombination and mutation in HIV-1 gag. Virology 326, 262–279.

Escriu, F., Fraile, A., García-Arenal, F., 2007. Constraints to genetic exchange support gene coadaptation in a tripartite RNA virus. PLoS Pathog. 3, 0067–0074.

Fan, J., Negroni, M., Robertson, D.L., 2007. The distribution of HIV-1 recombination breakpoints. Infect. Genet. Evol. 7, 717–723.

Figlerowicz, M., 2000. Role of RNA structure in non-homologous recombination between genomic molecules of brome mosaic virus. Nucleic Acids Res. 28, 1714–23.

Filbin, M.E., Kieft, J.S., 2009. Toward a structural understanding of IRES RNA function. Curr. Opin. Struct. Biol.

Filomatori, C. V., Lodeiro, M.F., Alvarez, D.E., Samsa, M.M., Pietrasanta, L., Gamarnik, A. V., 2006. A 5′ RNA element promotes dengue virus RNA synthesis on a circular genome. Genes Dev. 20, 2238–2249.

Flamm, C., Fontana, W., Hofacker, I.L., Schuster, P., 2000. RNA folding at elementary step resolution. RNA 6, 325–38.

Fletcher, S.P., Jackson, R.J., 2002. Pestivirus internal ribosome entry site (IRES) structure and function: elements in the 5' untranslated region important for IRES function. J. Virol. 76, 5024–33.

Frederico, L.A., Shaw, B.R., Kunkel, T.A., 1990. A Sensitive Genetic Assay for the Detection of Cytosine Deamination: Determination of Rate Constants and the Activation

Energy. Biochemistry 29, 2532–2537.

French, R., Ahlquist, P., 1987. Intercistronic as well as terminal sequences are required for efficient amplification of brome mosaic virus RNA3. J. Virol. 61, 1457–65.

Friebe, P., Harris, E., 2010. Interplay of RNA Elements in the Dengue Virus 5' and 3' Ends Required for Viral RNA Replication. J. Virol. 84, 6103–6118.

Fu, K., Baric, R.S., 1994. Map locations of mouse hepatitis virus temperature-sensitive mutants: confirmation of variable rates of recombination. J. Virol. 68, 7458–66.

Fu, Y.X., Li, W.H., 1993. Statistical tests of neutrality of mutations. Genetics 133, 693–709.

Galetto, R., Giacomoni, V., Véron, M., Negroni, M., 2006. Dissection of a circumscribed recombination hot spot in HIV-1 after a single infectious cycle. J. Biol. Chem. 281, 2711–2720.

Galetto, R., Moumen, A., Giacomoni, V., Véron, M., Charneau, P., Negroni, M., 2004. The structure of HIV-1 genomic RNA in the gp120 gene determines a recombination hot spot in vivo. J. Biol. Chem. 279, 36625–36632.

Galli, A., Lai, A., Corvasce, S., Saladini, F., Riva, C., Dehò, L., Caramma, I., Franzetti, M., Romano, L., Galli, M., Zazzi, M., Balotta, C., 2008. Recombination analysis and structure prediction show correlation between breakpoint clusters and RNA hairpins in the pol gene of human immunodeficiency virus type 1 unique recombinant forms. J. Gen. Virol. 89, 3119–3125.

Garcia, K.C., Teyton, L., Wilson, I.A., 1999. Structural basis of T cell recognition. Annu. Rev. Immunol. 17, 369–97.

Garmann, R.F., Gopal, A., Athavale, S.S., Knobler, C.M., Gelbart, W.M., Harvey, S.C., 2015. Visualizing the global secondary structure of a viral RNA genome with cryo-electron microscopy. RNA 21, 877–86.

Gauthier, K., Abt, I., Deshoux, M., Marchat, M., Temple, C., 2017. Epidemiology of barley yellow dwarf and wheat dwarf diseases. Madrid.

Gerber, K., Wimmer, E., Paul, A. V, 2001. Biochemical and genetic studies of the initiation of human rhinovirus 2 RNA replication: identification of a cis-replicating element in the coding sequence of 2A(pro). J. Virol. 75, 10979–90.

Gibbs, A.J., Ohshima, K., 2010. Potyviruses and the Digital Revolution. Annu. Rev. Phytopathol. 48, 205–223.

Gibbs, A.J., Ohshima, K., Phillips, M.J., Gibbs, M.J., 2008. The prehistory of potyviruses: Their initial radiation was during the dawn of agriculture. PLoS One 3.

Gibbs, M.J., Armstrong, J.S., Gibbs, A.J., 2000. Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics 16, 573–582.

Gibbs, M.J., Weiller, G.F., 1999. Evidence that a plant virus switched hosts to infect a vertebrate and then recombined with a vertebrate-infecting virus. Proc. Natl. Acad. Sci. U. S. A. 96, 8022–7.

Gibson, R.W., Aritua, V., Byamukama, E., Mpembe, I., Kayongo, J., 2004. Control strategies for sweet potato virus disease in Africa. In: Virus Research. pp. 115–122.

Giedroc, D.P., Cornish, P. V., 2009. Frameshifting RNA pseudoknots: Structure and mechanism. Virus Res. 139, 193–208.

Glasa, M., Paunovic, S., Jevremovic, D., Myrta, A., Pittnerová, S., Candresse, T., 2005. Analysis of recombinant Plum pox virus (PPV) isolates from Serbia confirms genetic homogeneity and supports a regional origin for the PPV-Rec subgroup. Arch. Virol. 150, 2051–2060.

Goens, S.D., 2002. The evolution of bovine viral diarrhea: A review. Can. Vet. J.

Golden, M., Martin, D., 2013. DOOSS: A tool for visual analysis of data overlaid on secondary structures. Bioinformatics 29, 271–272.

Golden, M., Muhire, B.M., Semegni, Y., Martin, D.P., 2014. Patterns of recombination in HIV-1M are influenced by selection disfavouring the survival of recombinants with disrupted genomic RNA and protein structures. PLoS One 9, e100400.

Goodbourn, S., Didcock, L., Randall, R.E., 2000. Interferons: Cell signalling, immune modulation, antiviral responses and virus countermeasures. J. Gen. Virol.

Goodfellow, I., Chaudhry, Y., Richardson, A., Meredith, J., Almond, J.W., Barclay, W., Evans, D.J., 2000. Identification of a cis-acting replication element within the poliovirus coding region. J. Virol. 74, 4590–600.

Gosert, R., Chang, K.H., Rijnbrand, R., Yi, M., Sangar, D. V, Lemon, S.M., 2000. Transient

expression of cellular polypyrimidine-tract binding protein stimulates cap-independent translation directed by both picornaviral and flaviviral internal ribosome entry sites In vivo. Mol Cell Biol 20, 1583–1595.

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

Gunawardene, C.D., Jaluba, K., White, K.A., 2015. Conserved Motifs in a Tombusvirus Polymerase Modulate Genome Replication, Subgenomic Transcription, and Amplification of Defective Interfering RNAs. J. Virol. 89, 3236–3246.

Haldeman-Cahill, R., Daros, J.A., Carrington, J.C., 1998. Secondary structures in the capsid protein coding sequence and 3' nontranslated region involved in amplification of the tobacco etch virus genome. J Virol 72, 4072–9.

Hanada, K., Suzuki, Y., Gojobori, T., 2004. A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. Mol. Biol. Evol. 21, 1074–1080.

Heath, L., van der Walt, E., Varsani, A., Martin, D.P., 2006. Recombination Patterns in Aphthoviruses Mirror Those Found in Other Picornaviruses. J. Virol. 80, 11827–11832.

Herbreteau, C.H., Weill, L., Décimo, D., Prévôt, D., Darlix, J.-L., Sargueil, B., Ohlmann, T., 2005. HIV-2 genomic RNA contains a novel type of IRES located downstream of its initiation codon. Nat. Struct. Mol. Biol. 12, 1001–7.

Herschlag, D., Khosla, M., Tsuchihashi, Z., Karpel, R.L., 1994. An RNA chaperone activity of non-specific RNA binding proteins in hammerhead ribozyme catalysis. EMBO J. 13, 2913–24.

Hofacker, I., Fontana, F., Stadler, L., Bonhoeffer, M., Schuster, P., 1994. Fast folding and comparison of RNA secondary structures. Chem. Mon. 125, 167–188.

Högbom, M., Jäger, K., Robel, I., Unge, T., Rohayem, J., 2009. The active form of the norovirus RNA-dependent RNA polymerase is a homodimer with cooperative activity. J. Gen. Virol. 90, 281–291.

Holmes, E.C., 2003. Error thresholds and the constraints to RNA virus evolution. Trends

Microbiol.

Honda, M., Ping, L.H., Rijnbrand, R.C., Amphlett, E., Clarke, B., Rowlands, D., Lemon, S.M., 1996. Structural requirements for initiation of translation by internal ribosome entry within genome-length hepatitis C virus RNA. Virology 222, 31–42.

Hou, J., Baichwal, V., Cao, Z., 1994. Regulatory elements and transcription factors controlling basal and cytokine-induced expression of the gene encoding intercellular adhesion molecule 1. Proc. Natl. Acad. Sci. U. S. A. 91, 11641–5.

Hunt, S.L., Jackson, R.J., 1999. Polypyrimidine-tract binding protein (PTB) is necessary, but not sufficient, for efficient internal initiation of translation of human rhinovirus-2 RNA. RNA 5, 344–59.

Husmeier, D., McGuire, G., 2002. Detecting recombination with MCMC. Bioinformatics 18, S345–S353.

Hutchinson, E.C., von Kirchbach, J.C., Gog, J.R., Digard, P., 2010. Genome packaging in influenza A virus. J. Gen. Virol.

Hwang, W.L., Su, T.S., 1999. The encapsidation signal of hepatitis B virus facilitates preC AUG recognition resulting in inefficient translation of the downstream genes. J. Gen. Virol. 80, 1769–1776.

Hyodo, K., Okuno, T., 2014. Host factors used by positive-strand RNA plant viruses for genome replication. J. Gen. Plant Pathol.

Ikejezie, J., Shapiro, C.N., Kim, J., Chiu, M., Almiron, M., Ugarte, C., Espinal, M.A., Aldighieri, S., 2017. Zika Virus Transmission — Region of the Americas, May 15, 2015–December 15, 2016. MMWR. Morb. Mortal. Wkly. Rep. 66, 329–334.

Jacquenet, S., Ropers, D., Bilodeau, P.S., Damier, L., Mougin, a, Stoltzfus, C.M., Branlant, C., 2001. Conserved stem-loop structures in the HIV-1 RNA region containing the A3 3' splice site and its cis-regulatory element: possible involvement in RNA splicing. Nucleic Acids Res. 29, 464–78.

Jain, R., Rivera, M.C., Lake, J.A., 1999. Horizontal gene transfer among genomes: The complexity hypothesis. Proc. Natl. Acad. Sci. 96, 3801–3806.

Kaminski, a, Hunt, S.L., Patton, J.G., Jackson, R.J., 1995. Direct evidence that

polypyrimidine tract binding protein (PTB) is essential for internal initiation of translation of encephalomyocarditis virus RNA. RNA.

Kamp, C., Wilke, C.O., Adami, C., Bornholdt, S., 2002. Viral evolution under the pressure of an adaptive immune system - optimal mutation rates for viral escape. Complexity 8, 5.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.

Khatchikian, D., Orlich, M., Rott, R., 1989. Increased viral pathogenicity after insertion of a 28S ribosomal RNA sequence into the haemagglutinin gene of an influenza virus. Nature 340, 156–7.

Kim, K.-H., Kwon, S.-J., Hemenway, C., 2002. Cellular Protein Binds to Sequences near the 5′ Terminus of Potato virus X RNA That Are Important for Virus Replication. Virology 301, 305–312.

Kim, K.H., Hemenway, C., 1996. The 5' nontranslated region of potato virus X RNA affects both genomic and subgenomic RNA synthesis. J. Virol. 70, 5533–40.

Kneller, E.L.P., Rakotondrafara, A.M., Miller, W.A., 2006. Cap-independent translation of plant viral RNAs. Virus Res. 119, 63–75.

Kochetov, A. V., Sarai, A., Rogozin, I.B., Shumny, V.K., Kolchanov, N.A., 2005. The role of alternative translation start sites in the generation of human protein diversity. Mol. Genet. Genomics 273, 491–496.

Kolupaeva, V.G., Pestova, T. V, Hellen, C.U., 2000. Ribosomal binding to the internal ribosomal entry site of classical swine fever virus. RNA 6, 1791–1807.

Koonin, E. V., 1991. The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. J. Gen. Virol. 72, 2197–2206.

Kosakovsky Pond, S.L., Frost, S.D.W., Muse, S. V., 2005. HyPhy: Hypothesis testing using phylogenies. Bioinformatics 21, 676–679.

Kosakovsky Pond, S.L., Posada, D., Gravenor, M.B., Woelk, C.H., Frost, S.D.W., 2006. GARD: a genetic algorithm for recombination detection. Bioinformatics 22, 3096–8.

Kozak, M., 1989. The scanning model for translation: An update. J. Cell Biol.

Kozak, M., 1990. Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. Proc. Natl. Acad. Sci. 87, 8301–8305.

Kumar, S., Subramanian, S., 2002. Mutation rates in mammalian genomes. Proc. Natl. Acad. Sci. U. S. A. 99, 803–808.

Kuno, G., Chang, G.J., Tsuchiya, K.R., Karabatsos, N., Cropp, C.B., 1998. Phylogeny of the genus Flavivirus. J. Virol. 72, 73–83.

Kwon, S.-J., Park, M.-R., Kim, K.-W., Plante, C. a, Hemenway, C.L., Kim, K.-H., 2005. cis-Acting sequences required for coat protein binding and in vitro assembly of Potato virus X. Virology 334, 83–97.

Lai, M.M., 1992. RNA recombination in animal and plant viruses. Microbiol. Rev. 56, 61–79.

Landsteiner, K Popper, E., 1909. Übertragung der Poliomyelitis acuta auf Affen. Immunitätsforsch 2, 377–390.

Lefeuvre, P., Lett, J.-M., Varsani, A., Martin, D.P., 2009. Widely Conserved Recombination Patterns among Single-Stranded DNA Viruses. J. Virol. 83, 2697–2707.

Lefeuvre, P., Lett, J.M., Reynaud, B., Martin, D.P., 2007. Avoidance of protein fold disruption in natural virus recombinants. PLoS Pathog. 3, 1782–1789.

Li, J., Broyles, S.S., 1993. The DNA-dependent ATPase activity of vaccinia virus early gene transcription factor is essential for its transcription activation function. J. Biol. Chem. 268, 20016–20021.

Li, X.H., Valdez, P., Olvera, R.E., Carrington, J.C., 1997. Functions of the tobacco etch virus RNA polymerase (NIb): subcellular transport and protein-protein interaction with VPg/proteinase (NIa). J. Virol. 71, 1598–607.

Lin, X., Thorne, L., Jin, Z., Hammad, L.A., Li, S., Deval, J., Goodfellow, I.G., Kao, C.C., 2015. Subgenomic promoter recognition by the norovirus RNA-dependent RNA polymerases. Nucleic Acids Res. 43, 446–460.

Linnstaedt, S.D., Kasprzak, W.K., Shapiro, B.A., Casey, J.L., 2009. The fraction of RNA that folds into the correct branched secondary structure determines hepatitis delta virus type 3 RNA editing levels. RNA 15, 1177–1187.

Liu, L., Pinner, M.S., Davies, J.W., Stanley, J., 1999. Adaptation of the geminivirus bean yellow dwarf virus to dicotyledonous hosts involves both virion-sense and complementary-sense genes. J. Gen. Virol. 80, 501–506.

López-Otín, C., Bond, J.S., 2008. Proteases: Multifunctional enzymes in life and disease. J. Biol. Chem.

Lough, T.J., Lee, R.H., Emerson, S.J., Forster, R.L., Lucas, W.J., 2006. Functional analysis of the 5' untranslated region of potexvirus RNA reveals a role in viral replication and cell-to-cell movement. Virology 351, 455–465.

Lukashev, A.N., Ivanova, O.E., Eremeeva, T.P., Iggo, R.D., 2008. Evidence of frequent recombination among human adenoviruses. J. Gen. Virol. 89, 380–388.

Lukashev, A.N., Lashkevich, V.A., Ivanova, O.E., Koroleva, G.A., Hinkkanen, A.E., Ilonen, J., 2003. Recombination in circulating enteroviruses. J. Virol. 77, 10423–31.

Lukashev, A.N., Lashkevich, V.A., Ivanova, O.E., Koroleva, G.A., Hinkkanen, A.E., Ilonen, J., 2005. Recombination in circulating Human enterovirus B: Independent evolution of structural and non-structural genome regions. J. Gen. Virol. 86, 3281–3290.

Mahajan, S., Dolja, V. V, Carrington, J.C., 1996. Roles of the sequence encoding tobacco etch virus capsid protein in genome amplification: requirements for the translation process and a cis-active element. J. Virol. 70, 4370–9.

Makino, S., Keck, J.G., Stohlman, S. a, Lai, M.M., 1986. High-frequency RNA recombination of murine coronaviruses. J. Virol. 57, 729–737.

Malim, M.H., Emerman, M., 2001. HIV-1 sequence variation: Drift, shift, and attenuation. Cell.

Malnou, C.E., Poyry, T.A.A., Jackson, R.J., Kean, K.M., 2002. Poliovirus Internal Ribosome Entry Segment Structure Alterations That Specifically Affect Function in Neuronal Cells: Molecular Genetic Analysis. J. Virol. 76, 10617–10626.

Mandal, M., 2004. A Glycine-Dependent Riboswitch That Uses Cooperative Binding to Control Gene Expression. Science (80-.). 306, 275–279.

Markham, N.R., Zuker, M., 2008. UNAFold: Software for nucleic acid folding and hybridization. Methods Mol. Biol. 453, 3–31.

Martin, D., Rybicki, E., 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics 16, 562–563.

Martin, D.P., Lefeuvre, P., Varsani, A., Hoareau, M., Semegni, J.Y., Dijoux, B., Vincent, C., Reynaud, B., Lett, J.M., 2011. Complex recombination patterns arising during geminivirus coinfections preserve and demarcate biologically important intra-genome interaction networks. PLoS Pathog. 7.

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.

Martin, D.P., Van Walt, E. Der, Posada, D., Rybicki, E.P., 2005. The evolutionary value of recombination is constrained by genome modularity. PLoS Genet. 1, 0475–0479.

Martínez-Salas, E., Francisco-Velilla, R., Fernandez-Chamorro, J., Lozano, G., Diaz-Toledano, R., 2015. Picornavirus IRES elements: RNA structure and host protein interactions. Virus Res. 206, 62–73.

Matthews, D.H., 2004. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. RNA 10, 1178–1190.

Mauger, D.M., Golden, M., Yamane, D., Williford, S., Lemon, S.M., Martin, D.P., Weeks, K.M., 2015. Functionally conserved architecture of hepatitis C virus RNA genomes. Proc. Natl. Acad. Sci. 201416266.

Meyer, M.D., Pataky, J.K., 2010. Increased Severity of Foliar Diseases of Sweet Corn Infected with Maize Dwarf Mosaic and Sugarcane Mosaic Viruses. Plant Dis. 94, 1093–1099.

Miller, E.D., Kim, K.H., Hemenway, C., 1999. Restoration of a stem-loop structure required for potato virus X RNA accumulation indicates selection for a mismatch and a GNRA tetraloop. Virology 260, 342–353.

Miller, E.D., Plante, C. a, Kim, K.H., Brown, J.W., Hemenway, C., 1998. Stem-loop structure in the 5’ region of potato virus X genome required for plus-strand RNA accumulation. J. Mol. Biol. 284, 591–608.

Minin, V.N., Dorman, K.S., Fang, F., Suchard, M.A., 2005. Dual multiple change-point model leads to more accurate recombination detection. Bioinformatics 21, 3034–42.

Modrof, J., Becker, S., Mühlberger, E., 2003. Ebola Virus Transcription Activator VP30 Is a Zinc-Binding Protein. J. Virol. 77, 3334–3338.

Moreno, I.M., Malpica, J.M., Díaz-Pendón, J.A., Moriones, E., Fraile, A., García-Arenal, F., 2004. Variability and genetic structure of the population of watermelon mosaic virus infecting melon in Spain. Virology 318, 451–460.

Moya, A., Elena, S.F., Bracho, A., Miralles, R., Barrio, E., 2000. The evolution of RNA viruses: A population genetics view. Proc. Natl. Acad. Sci. U. S. A. 97, 6967–73.

Mueller, N., van Bel, N., Berkhout, B., Das, A.T., 2014. HIV-1 splicing at the major splice donor site is restricted by RNA structure. Virology 468, 609–620.

Muhire, B.M., Golden, M., Murrell, B., Lefeuvre, P., Lett, J., Gray, A., Poon, A.Y.F., Ngandu, N.K., Semegni, Y., Tanov, E.P., Monjane, A.L., Harkins, G.W., Varsani, A., Shepherd, D.N., Martin, D.P., 2014. Evidence of pervasive biologically functional secondary structures within the genomes of eukaryotic single-stranded DNA viruses. J. Virol. 88, 1972–89.

Mukhopadhyay, S., Kuhn, R.J., Rossmann, M.G., 2005. A structural perspective of the flavivirus life cycle. Nat. Rev. Microbiol. 3, 13–22.

Müller, M.M., Gerster, T., Schaffner, W., 1988. Enhancer sequences and the regulation of gene transcription. Eur. J. Biochem.

Murray, K.E., Barton, D.J., 2003. Poliovirus CRE-Dependent VPg Uridylylation Is Required for Positive-Strand RNA Synthesis but Not for Negative-Strand RNA Synthesis. J. Virol. 77, 4739–4750.

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S.L., Scheffler, K., 2013. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol. Biol. Evol. 30, 1196–205.

Muse, S. V, 1995. Evolutionary analyses of DNA sequences subject to constraints of secondary structure. Genetics 139, 1429–39.

Nachman, M.W., Crowell, S.L., 2000. Estimate of the mutation rate per nucleotide in humans. Genetics 156, 297–304.

Nagy, P.D., Pogany, J., Simon, A.E., 1999. RNA elements required for RNA recombination

function as replication enhancers in vitro and in vivo in a plus-strand RNA virus. EMBO J. 18, 5653–5665.

Nawrocki, E.P., Eddy, S.R., 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29, 2933–2935.

Nechooshtan, G., Elgrably-Weiss, M., Sheaffer, A., Westhof, E., Altuvia, S., 2009. A pH-responsive riboregulator. Genes Dev. 23, 2650–2662.

Ngandu, N.K., Scheffler, K., Moore, P., Woodman, Z., Martin, D., Seoighe, C., 2008. Extensive purifying selection acting on synonymous sites in HIV-1 Group M sequences. Virol. J. 5, 160.

Nocker, A., Hausherr, T., Balsiger, S., Krstulovic, N.P., Hennecke, H., Narberhaus, F., 2001. A mRNA-based thermosensor controls expression of rhizobial heat shock genes. Nucleic Acids Res. 29, 4800–7.

Nolan, J.P., Wilmer, B.H., Melnick, J.L., 1955. Poliomyelitis — Its Highly Invasive Nature and Narrow Stream of Infection in a Community of High Socioeconomic Level. N Engl J Med 253, 945–954.

Nora, T., Charpentier, C., Tenaillon, O., Hoede, C., Clavel, F., Hance, A.J., 2007. Contribution of Recombination to the Evolution of Human Immunodeficiency Viruses Expressing Resistance to Antiretroviral Treatment. J. Virol. 81, 7620–7628.

Pace, N.R., Thomas, B.C., Woese, C.R., 1999. Probing RNA Structure, Function, and History by Comparative Analysis. RNA World 113–142.

Padidam, M., Sawyer, S., Fauquet, C.M., 1999. Possible Emergence of New Geminiviruses by Frequent Recombination. Virology 265, 218–225.

Pagan, I., Holmes, E.C., 2010. Long-Term Evolution of the Luteoviridae: Time Scale and Mode of Virus Speciation. J. Virol. 84, 6177–6187.

Paul, A. V, Rieder, E., Kim, D.W., van Boom, J.H., Wimmer, E., 2000. Identification of an RNA hairpin in poliovirus RNA that serves as the primary template in the in vitro uridylylation of VPg. J. Virol. 74, 10359–70.

Pelletier, J., Sonenberg, N., 1988. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. Nature 334, 320–325.

Pita, J.S., Roossinck, M.J., 2013. Fixation of Emerging Interviral Recombinants in Cucumber Mosaic Virus Populations. J. Virol. 87, 1264–1269.

Plant, E.P., Dinman, J.D., 2008. The role of programmed-1 ribosomal frameshifting in coronavirus propagation. Front. Biosci. 13, 4873–81.

Poon, A.F.Y., Lewis, F.I., Frost, S.D.W., Kosakovsky Pond, S.L., 2008. Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models. Bioinformatics 24, 1949–50.

Posada, D., Crandall, K.A., 2001. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. Proc. Natl. Acad. Sci. 98, 13757–13762.

Poulin, F., Sonenberg, N., 2000. Mechanism of Translation Initiation in Eukaryotes. Madame Curie Biosci. Database 1–33.

Proutski, V., Gaunt, M.W., Gould, E.A., Holmes, E.C., 1997. Secondary structure of the 3'-untranslated region of yellow fever virus: Implications for virulence, attenuation and vaccine development. J. Gen. Virol. 78, 1543–1549.

R Development Core Team, 2016. R: A Language and Environment for Statistical Computing. R Found. Stat. Comput. Vienna Austria 0, {ISBN} 3-900051-07-0.

RajBhandary, U.L., 1994. Initiator transfer RNAs. J. Bacteriol.

Rajkowitsch, L., Chen, D., Stampfl, S., Semrad, K., Waldsich, C., Mayer, O., Jantsch, M.F., Konrat, R., Bläsi, U., Schroeder, R., 2007. RNA Chaperones, RNA Annealers and RNA Helicases. RNA Biol. 4, 118–130.

Raman, S., Bouma, P., Williams, G.D., Brian, D.A., 2003. Stem-loop III in the 5' untranslated region is a cis-acting element in bovine coronavirus defective interfering RNA replication. J. Virol. 77, 6720–30.

Ramani, V., Qiu, R., Shendure, J., 2015. High-throughput determination of RNA structure by proximity ligation. Nat. Biotechnol. 33, 980–984.

Reiter, J., Pérez-Vilaró, G., Scheller, N., Mina, L.B., Díez, J., Meyerhans, A., 2011. Hepatitis C virus RNA recombination in cell culture. J. Hepatol. 55, 777–783.

Revers, F., García, J.A., 2015. Molecular biology of potyviruses. Adv. Virus Res. 92, 101–199.

Rieder, E., Paul, A. V, Wook Kim, D., Van Boom, J.H., Wimmer, E., 2000. Genetic and Biochemical Studies of Poliovirus cis-Acting Replication Element cre in Relation to VPg Uridylylation. J. Virol. 74, 10371–10380.

Romero, T. a, Tumban, E., Jun, J., Lott, W.B., Hanley, K. a, 2006. Secondary structure of dengue virus type 4 3' untranslated region: impact of deletion and substitution mutations. J. Gen. Virol. 87, 3291–3296.

Roth, J.R., Andersson, D.I., 2012. Poxvirus use a "gene accordion" to tune out host defenses. Cell 150, 671–672.

Rubio, L., Guerri, J., Moreno, P., 2013. Genetic variability and evolutionary dynamics of viruses of the family Closteroviridae. Front. Microbiol. 4.

Runckel, C., Westesson, O., Andino, R., DeRisi, J.L., 2013. Identification and Manipulation of the Molecular Determinants Influencing Poliovirus Recombination. PLoS Pathog. 9.

Salguero, F.J., Sánchez-Martín, M.A., Díaz-San Segundo, F., De Avila, A., Sevilla, N., 2005. Foot-and-mouth disease virus (FMDV) causes an acute disease that can be lethal for adult laboratory mice. Virology 332, 384–396.

Salminen, M., Carr, K., Burke, D., McCutchan, F., 1995. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. AIDS Res Hum Retroviruses 11, 1423–1425.

Sanjuan, R., Nebot, M.R., Chirico, N., Mansky, L.M., Belshaw, R., 2010. Viral Mutation Rates. J. Virol. 84, 9733–9748.

Satyanarayana, T., Gowda, S., Ayllón, M.A., Albiach-Martí, M.R., Rabindran, S., Dawson, W.O., 2002. The p23 protein of citrus tristeza virus controls asymmetrical RNA accumulation. J. Virol. 76, 473–83.

Savolainen-Kopra, C., Blomqvist, S., 2010. Mechanisms of genetic variation in polioviruses. Rev. Med. Virol.

Scheffler, K., Martin, D.P., Seoighe, C., 2006. Robust inference of positive selection from recombining coding sequences. Bioinformatics 22, 2493–9.

Schindler, J., 1964. RNA-containing bacteriophages. Folia Microbiol. (Praha). 9, 312–320.

Schirmer, S., Ponty, Y., Giegerich, R., 2014. Introduction to RNA secondary structure

comparison. Methods Mol. Biol. 1097, 247–73.

Schoelz, J., Wintermantel, W., 1993. Expansion of Viral Host Range through Complementation and Recombination in Transgenic Plants. Plant Cell 5, 1669–1679.

Schultheiss, T., Sommergruber, W., Kusov, Y., Gauss-Müller, V., 1995. Cleavage specificity of purified recombinant hepatitis A virus 3C proteinase on natural substrates. J. Virol. 69, 1727–1733.

Semegni, J.Y., Wamalwa, M., Gaujoux, R., Harkins, G.W., Gray, A., Martin, D.P., 2011. NASP: a parallel program for identifying evolutionarily conserved nucleic acid secondary structures from nucleotide sequence alignments. Bioinformatics 27, 2443–5.

Seronello, S., Montanez, J., Presleigh, K., Barlow, M., Park, S.B., Choi, J., 2011. Ethanol and reactive species increase basal sequence heterogeneity of hepatitis C virus and produce variants with reduced susceptibility to antivirals. PLoS One 6.

Shakeel, S., Dykeman, E.C., White, S.J., Ora, A., Cockburn, J.J.B., Butcher, S.J., Stockley, P.G., Twarock, R., 2017. Erratum: Genomic RNA folding mediates assembly of human parechovirus. Nat. Commun. 8, 83.

Shen, R., Miller, W.A., 2007. Structures required for poly(A) tail-independent translation overlap with, but are distinct from, cap-independent translation and RNA replication signals at the 3′ end of Tobacco necrosis virus RNA. Virology 358, 448–458.

Shepherd, D.N., Martin, D.P., Varsani, A., Thomson, J.A., Rybicki, E.P., Klump, H.H., 2006. Restoration of native folding of single-stranded DNA sequences through reverse mutations: An indication of a new epigenetic mechanism. Arch. Biochem. Biophys. 453, 106–120.

Shriner, D., Nickle, D.C., Jensen, M.A., Mullins, J.I., 2003. Potential impact of recombination on sitewise approaches for detecting positive natural selection. Genet. Res. 81, 115–21.

Shukla, D.D., Ward, C.W., 1988. Amino Acid Sequence Homology of Coat Proteins as a Basis for Identification and Classification of the Potyvirus Group. J. gen. Virol 69, 2703–2710.

Simmonds, P., Karakasiliotis, I., Bailey, D., Chaudhry, Y., Evans, D.J., Goodfellow, I.G.,

2008. Bioinformatic and functional analysis of RNA secondary structure elements among different genera of human and animal caliciviruses. Nucleic Acids Res. 36, 2530–2546.

Simmonds, P., Tuplin, A., Evans, D.J., 2004. Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence. RNA 10, 1337–51.

Simmonds, P., Welch, J., 2006. Frequency and Dynamics of Recombination within Different Species of Human Enteroviruses. J. Virol. 80, 483–493.

Simon-Loriere, E., Galetto, R., Hamoudi, M., Archer, J., Lefeuvre, P., Martin, D.P., Robertson, D.L., Negroni, M., 2009. Molecular mechanisms of recombination restriction in the envelope gene of the human immunodeficiency virus. PLoS Pathog. 5.

Simon-Loriere, E., Holmes, E.C., 2011. Why do RNA viruses recombine? Nat. Rev. Microbiol. 9, 617–626.

Simon-Loriere, E., Martin, D.P., Weeks, K.M., Negroni, M., 2010. RNA Structures Facilitate Recombination-Mediated Gene Swapping in HIV-1. J. Virol. 84, 12675–12682.

Smith, J.M., 1992. Analyzing the mosaic structure of genes. J. Mol. Evol. 34, 126–129.

Sperling-Petersen, H.U., Laursen, B.S., Mortensen, K.K., 2002. Initiator tRNAs in Prokaryotes and Eukaryotes. Encycl. Life Sci.

Steil, B.P., Barton, D.J., 2009. Cis-active RNA elements (CREs) and picornavirus RNA replication. Virus Res. 139, 240–252.

Stewart, H., Bingham, R.J., White, S.J., Dykeman, E.C., Zothner, C., Tuplin, A.K., Stockley, P.G., Twarock, R., Harris, M., 2016. Identification of novel RNA secondary structures within the hepatitis C virus genome reveals a cooperative involvement in genome packaging. Sci. Rep. 6, 22952.

Suchard, M.A., Weiss, R.E., Dorman, K.S., Sinsheimer, J.S., 2002. Oh brother, where art thou? A bayes factor test for recombination with uncertain heritage. Syst. Biol. 51, 715–728.

Sükösd, Z., Swenson, M.S., Kjems, J., Heitsch, C.E., 2013. Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions. Nucleic Acids Res. 41, 2807–

2816.

Sun, X., Zhang, G., Simon, A.E., 2005. Short internal sequences involved in replication and virion accumulation in a subviral RNA of turnip crinkle virus. J. Virol. 79, 512–24.

Sztuba-Solińska, J., Urbanowicz, A., Figlerowicz, M., Bujarski, J.J., 2011. RNA-RNA Recombination in Plant Virus Replication and Evolution. Annu. Rev. Phytopathol. 49, 415–443.

Tajima, F., 1989. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics 123, 585–595.

Takahashi, H., Yamaji, M., Hosaka, M., Kishine, H., Hijikata, M., Shimotohno, K., 2005. Analysis of the 5′ end structure of HCV subgenomic RNA replicated in a Huh7 cell line. Intervirology 48, 104–111.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol. Biol. Evol. 30, 2725–9.

Taucher, C., Berger, A., Mandl, C.W., 2010. A trans-complementing recombination trap demonstrates a low propensity of flaviviruses for intermolecular recombination. J. Virol. 84, 599–611.

Thompson, S.R., Sarnow, P., 2003. Enterovirus 71 contains a type I IRES element that functions when eukaryotic initiation factor eIF4G is cleaved. Virology 315, 259–266.

Tijerina, P., Mohr, S., Russell, R., 2007. DMS footprinting of structured RNAs and RNA–protein complexes. Nat. Protoc. 2, 2608–2623.

Tromas, N., Zwart, M.P., Maïté, P., Elena, S.F., 2014. Estimation of the in vivo recombination rate for a plant RNA virus. J. Gen. Virol. 95, 724–732.

Tuplin, A., Evans, D.J., Buckley, A., Jones, I.M., Gould, E.A., Gritsun, T.S., 2011. Replication enhancer elements within the open reading frame of tick-borne encephalitis virus and their evolution within the Flavivirus genus. Nucleic Acids Res. 39, 7034–7048.

Tuplin, A., Evans, D.J., Simmonds, P., 2004. Detailed mapping of RNA secondary structures in core and NS5B-encoding region sequences of hepatitis C virus by RNase cleavage and novel bioinformatic prediction methods. J. Gen. Virol. 85, 3037–3047.

Uchil, P.D., Satchidanandam, V., 2003. Architecture of the flaviviral replication complex: Protease, nuclease, and detergents reveal encasement within double-layered membrane compartments. J. Biol. Chem. 278, 24388–24398.

Urbanowicz, A., Alejska, M., Formanowicz, P., Blazewicz, J., Figlerowicz, M., Bujarski, J.J., 2005. Homologous crossovers among molecules of brome mosaic bromovirus RNA1 or RNA2 segments in vivo. J. Virol. 79, 5732–5742.

van Dijk, A.A., Makeyev, E. V., Bamford, D.H., 2004. Initiation of viral RNA-dependent RNA polymerization. J. Gen. Virol.

van Marle, G., Luytjes, W., van der Most, R.G., van der Straaten, T., Spaan, W.J., 1995. Regulation of coronavirus mRNA transcription. J. Virol. 69, 7851–6.

van Ooij, M.J.M., Glaudemans, D.H.R.F., Heus, H.A., van Kuppeveld, F.J.M., Melchers, W.J.G., 2006. Structural and functional integrity of the coxsackievirus B3 oriR: Spacing between coaxial RNA helices. J. Gen. Virol. 87, 689–695.

Vashist, S., Urena, L., Chaudhry, Y., Goodfellow, I., 2012. Identification of RNA-Protein Interaction Networks Involved in the Norovirus Life Cycle. J. Virol. 86, 11977–11990.

Victoria, M., Colina, R., Miagostovich, M.P., Leite, J.P., Cristina, J., 2009. Phylogenetic prediction of cis-acting elements: a cre-like sequence in Norovirus genome? BMC Res. Notes 2, 176.

Vijaykrishna, D., Mukerji, R., Smith, G.J.D., 2015. RNA Virus Reassortment: An Evolutionary Mechanism for Host Jumps and Immune Evasion. PLoS Pathog. 11.

Voigt, C.A., Martinez, C., Wang, Z.-G., Mayo, S.L., Arnold, F.H., 2002. Protein building blocks preserved by recombination. Nat. Struct. Biol.

Wang, S., Ongena, M., Qiu, D., Guo, and L., 2017. Fungal Viruses: Promising Fundamental Research and Biological Control Agents of Fungi. SM Virol 2, 1011–1016.

Wang, X., Ullah, Z., Grumet, R., 2000. Interaction between Zucchini Yellow Mosaic Potyvirus RNA-Dependent RNA Polymerase and Host Poly-(A) Binding Protein. Virology 275, 433–443.

Warf, M.B., Berglund, J.A., 2010. Role of RNA structure in regulating pre-mRNA splicing. Trends Biochem. Sci.

Watts, J.M., Dang, K.K., Gorelick, R.J., Leonard, C.W., Bess, J.W., Swanstrom, R., Burch, C.L., Weeks, K.M., 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. Nature 460, 711–6.

Wilkinson, K.A., Merino, E.J., Weeks, K.M., 2006. Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. Nat. Protoc. 1, 1610–1616.

Willems, L., Gillet, N.A., 2015. APOBEC3 interference during replication of viral genomes. Viruses.

Wilson, J.E., Powell, M.J., Hoover, S.E., Sarnow, P., 2000. Naturally Occurring Dicistronic Cricket Paralysis Virus RNA Is Regulated by Two Internal Ribosome Entry Sites. Mol. Cell. Biol. 20, 4990–4999.

Witteveldt, J., Blundell, R., Maarleveld, J.J., Mcfadden, N., Evans, D.J., Simmonds, P., 2014. The influence of viral RNA secondary structure on interactions with innate host cell defences. Nucleic Acids Res. 42, 3314–3329.

Wuchty, S., Fontana, W., Hofacker, I.L., Schuster, P., 1999. Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers.

Xia, X., Yuen, K.Y., 2005. Differential selection and mutation between dsDNA and ssDNA phages shape the evolution of their genomic AT percentage. BMC Genet. 6, 20.

Xue, Y., Gracia, B., Herschlag, D., Russell, R., Al-Hashimi, H.M., 2016. Visualizing the formation of an RNA folding intermediate through a fast highly modular secondary structure switch. Nat. Commun. 7, ncomms11768.

Yakovchuk, P., Protozanova, E., Frank-Kamenetskii, M.D., 2006. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. Nucleic Acids Res. 34, 564–574.

Yin-Murphy, M., Almond, J.W., 1996. Picornaviruses, Medical Microbiology.

Yin, J., Paul, A. V, Wimmer, E., Rieder, E., 2003. Functional dissection of a poliovirus cis-acting replication element [PV-cre(2C)]: analysis of single- and dual-cre viral genomes and proteins that bind specifically to PV-cre RNA. J. Virol. 77, 5152–5166.

Yu, Q., Pecchia, D.B., Kingsley, S.L., Heckman, J.E., Burke, J.M., 1998. Cleavage of highly

structured viral RNA molecules by combinatorial libraries of hairpin ribozymes. The most effective ribozymes are not predicted by substrate selection rules. J. Biol. Chem. 273, 23524–23533.

Yuan, X., Davydova, N., Conte, M.R., Curry, S., Matthews, S., 2002. Chemical shift mapping of RNA interactions with the polypyrimidine tract binding protein. Nucleic Acids Res. 30, 456–62.

Zhang, B., Dong, H., Zhou, Y., Shi, P.-Y., 2008. Genetic interactions among the West Nile virus methyltransferase, the RNA-dependent RNA polymerase, and the 5' stem-loop of genomic RNA. J. Virol. 82, 7047–7058.

Zhang, G., Zhang, J., Simon, A.E., 2004. Repression and derepression of minus-strand synthesis in a plus-strand RNA virus replicon. J Virol 78, 7619–7633.

Zhang, R., Liu, S., Chiba, S., Kondo, H., Kanematsu, S., Suzuki, N., 2014. A novel single-stranded RNA virus isolated from a phytopathogenic filamentous fungus, Rosellinia necatrix, with similarity to hypo-like viruses. Front. Microbiol. 5.

Zoll, J., Heus, H.A., van Kuppeveld, F.J.M., Melchers, W.J.G., 2009. The structure-function relationship of the enterovirus 3′-UTR. Virus Res. 139, 209–216.

Zuker, M., 1989. Computer prediction of RNA structure. Methods Enzymol. 180, 262–288.

Zuker, M., Sankoff, D., 1984. RNA secondary structures and their prediction. Bull. Math. Biol. 46, 591–621.

Zuker, M., Stiegler, P., 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 9, 133–148.

Supplementary Material

Supplementary Table 1 | Consensus ranking of high confidence structure sets (HCSSs)

Previously characterised structures are highlighted in yellow, whereas the structures identified and discussed in this study are highlighted in green.
Please note that the complete list of rankings is prohibitively large to insert here in full (over 200 pages), therefore only the ten highest consensus ranked structures of each dataset are tabulated. For completeness of referencing, structures that are discussed in this study but rank outside the top ten structures are appended to the end of the list of the corresponding dataset.
Please find spreadsheet file containing the complete consensus-rankings of all datasets here: goo.gl/uEsXz8

Cons. Rank [a]	NASP rank / ID [b]	Structure coordinates including gaps				Syn. Sub. Rank [c]	Coevo. Rank [d]	Biological function	Reference
		5' stem		3' stem					
Potexvirus PepMV									
1	1	574	585	603	614	45	131		
2	22	5426	5435	5697	5706	1	30		
3	132	1442	1449	1540	1547	152	1		
4	2	853	861	865	873	56	170		
5	171	5490	5493	5638	5641	5	2		
6	206	5454	5460	5661	5667	2	8		
7	3	3647	3659	3719	3731	142	88		
8	13	5120	5126	5199	5205	3	161		
9	208	5486	5488	5643	5645	6	3		
10	4	499	507	542	550	87	50		
40	16	39	45	96	102	60	43		Figure 3.2, PX1B
Potexvirus PVX									
1	1	6234	6244	6248	6258	213	216		
2	18	62	65	70	73	22	1		
3	31	6136	6141	6148	6153	1	48		
4	2	709	718	724	733	64	102		
5	126	56	57	76	77	235	2		
6	339	6133	6134	6157	6158	2	49		
7	3	239	245	259	265	18	57		
8	13	5144	5152	5159	5167	3	50		
9	86	32	36	102	106	51	3	Movement and viral replication	Lough et al. 2006 Figure 3.2, PX1A
10	4	6319	6330	6336	6347	175	187		

[a] Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b] NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID in NASP.
[c] Syn. Sub. ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d] Coevo. ranking is a ranking based on complementarily coevolution. NA indicates structures which had no variable sites.

Cons. Rank [a]	NASP rank / ID [b]	Structure coordinates including gaps				Syn. Sub. Rank [c]	Coevo. Rank [d]	Biological function	Reference
		5' stem		3' stem					
Vesivirus FCV									
1	1	2403	2415	2491	2503	12	78		
2	34	2387	2391	2560	2564	1	89		
3	55	1308	1315	1345	1352	67	1		
4	2	1787	1803	1820	1836	69	59		
5	29	32	36	114	118	8	2	PTB binding site	Vashist et al, 2012 Figure 3.4, CL1A
6	61	2380	2385	2566	2571	2	49		
7	3	11	18	22	29	44	232		
8	11	2395	2400	2552	2557	3	9		
9	267	3433	3437	3450	3454	140	3		
10	4	5575	5595	5599	5619	188	137		
44	157	3836	3839	3854	3857	192	19		Figure 3.5, CL2B
Norovirus NV									
1	1	7466	7475	7488	7497	14	47		
2	33	2594	2599	2607	2612	34	1		
3	66	7275	7279	7500	7504	1	19		
4	2	116	124	143	151	13	46		
5	68	447	451	466	470	46	2		
6	75	7283	7286	7461	7464	2	20		
7	3	7292	7300	7429	7437	3	21		
8	81	6745	6749	6776	6780	7	3		
9	4	3918	3926	3954	3962	78	79	Plays a role in VPg uridilation	Lin et al, 2015 Figure 3.5, CL2A
10	18	5088	5094	5099	5105	4	41		
21	20	2	8	66	72	17	29		Figure 3.4, CL1B
Lagovirus RHDV									
1	1	1170	1183	1188	1201	58	65		
2	2	323	333	344	354	1	3		
3	6	6291	6306	6372	6387	56	1		
4	27	10	17	43	50	2	29		
5	49	5317	5323	5343	5349	35	2		
6	3	6561	6569	6601	6609	52	61		
7	68	55	59	81	85	3	24		Figure 3.4, CL1C
8	4	3608	3618	3651	3661	64	68		
9	42	4257	4262	4269	4274	4	30		
10	74	311	316	361	366	8	4		
49	46	3749	3757	3768	3776	41	22		Figure 3.5, CL2C

[a] Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.

[b] NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID in NASP.

[c] Syn. Sub. ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.

[d] Coevo. ranking is a ranking based on complementarily coevolution. NA indicates structures which had no variable sites.

Cons. Rank [a]	NASP rank / ID [b]	Structure coordinates including gaps				Syn. Sub. Rank [c]	Coevo. Rank [d]	Biological function	Reference
		5' stem		3' stem					
Closterovirus CTV									
1	1	18318	18330	18335	18347	154	148		
2	170	12525	12537	12853	12865	123	1		
3	269	18760	18762	19335	19337	1	96		
4	2	18775	18787	19302	19314	9	74		
5	83	8113	8123	8180	8190	142	2		
6	85	18764	18768	19326	19330	2	72		
7	3	16192	16203	16208	16219	257	263		
8	73	112	117	139	144	3	24		
9	137	196	200	220	224	22	3		
10	4	17550	17560	17564	17574	120	249		
Alphacoronavirus NL63									
1	1	26225	26241	26313	26329	308	NA		
2	373	2347	2353	2734	2740	1	NA		
3	2	2668	2684	2717	2733	103	NA		
4	83	722	731	1397	1406	2	NA		
5	3	19918	19933	20226	20241	318	NA		
6	293	714	720	1407	1413	3	NA		
7	4	5232	5247	5253	5268	159	NA		
8	374	2332	2338	2778	2784	4	NA		
9	5	2150	2163	2204	2217	26	NA		
10	142	2238	2247	3181	3190	5	NA		
Flavivirus DENV T1									
1	1	3437	3451	3590	3604	25	33		
2	114	4958	4966	5094	5102	12	1		
3	136	144	148	332	336	1	35		
4	2	73	83	89	99	71	133		Figure 3.7, DV1B
5	18	131	141	362	372	2	43		
6	84	2686	2693	2938	2945	139	2		
7	3	2289	2300	2406	2417	16	113		
8	68	3999	4010	5279	5290	77	3		
9	147	105	110	395	400	3	48		
10	4	6910	6919	6924	6933	70	132		

[a] Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b] NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID in NASP.
[c] Syn. Sub. ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d] Coevo. ranking is a ranking based on complementarily coevolution. NA indicates structures which had no variable sites.

Cons. Rank [a]	NASP rank / ID [b]	Structure coordinates including gaps				Syn. Sub. Rank [c]	Coevo. Rank [d]	Biological function	Reference
		5' stem		3' stem					
Flavivirus DENV T2									
1	1	8986	8995	8999	9008	86	1		
2	58	149	152	161	164	1	58		
3	115	9267	9275	9390	9398	85	115		
4	2	10679	10691	10711	10723	183	2	Required for viral RNA replication	Friebe and Harris, 2010
5	86	1281	1286	1399	1404	72	86		
6	169	4592	4597	4609	4614	2	169		
7	3	4533	4549	4648	4664	38	3		
8	18	5801	5808	5880	5887	66	18		
9	41	1859	1864	1881	1886	3	41		
10	4	3650	3659	3663	3672	179	4		
15	6	79	86	92	100	118	6	Required for viral RNA replication	Clyde and Harris, 2006 Figure 3.7, DV1A
Flavivirus DENV T3									
1	1	73	83	89	99	107	125		Figure 3.7, DV1C
2	8	162	170	176	184	1	46		
3	56	3577	3584	3617	3624	139	1		
4	2	371	380	411	420	143	37		
5	17	114	119	125	130	2	47		
6	34	8218	8226	8263	8271	115	2		
7	3	10646	10658	10678	10690	150	152		
8	38	262	269	291	298	3	48		
9	154	9561	9567	9584	9590	61	3		
10	4	3100	3109	3122	3131	127	136		
Flavivirus DENV T4									
1	1	10593	10606	10626	10639	219	228	Modulate viral replication	Romero et al, 2006
2	47	4415	4427	4605	4617	25	1		
3	210	156	160	196	200	1	176		
4	2	3277	3286	3291	3300	193	97		
5	163	2583	2588	2595	2600	2	177		
6	232	8506	8513	8960	8967	128	2		
7	3	5565	5579	5668	5682	18	81		
8	13	167	175	181	189	3	178		
9	169	4373	4383	4719	4729	4	3		
10	4	4984	4992	4999	5007	157	44		
152	102	69	71	106	108	68	133		Figure 3.7, DV1D

[a] Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.

[b] NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID in NASP.

[c] Syn. Sub. ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.

[d] Coevo. ranking is a ranking based on complementarily coevolution. NA indicates structures which had no variable sites.

Cons. Rank [a]	NASP rank / ID [b]	Structure coordinates including gaps				Syn. Sub. Rank [c]	Coevo. Rank [d]	Biological function	Reference
		5' stem		3' stem					
Flavivirus JEV									
1	1	218	230	234	246	7	107		
2	34	142	147	266	271	1	112		
3	109	8625	8637	9170	9182	58	1		
4	2	10917	10926	10949	10958	141	146		
5	7	200	207	258	265	2	117		
6	80	8742	8749	9162	9169	44	2		
7	3	10717	10726	10735	10744	142	52		
8	30	156	162	191	197	3	101		
9	94	3203	3210	3471	3478	126	3		
10	4	167	174	179	186	4	102		
Flavivirus WNV									
1	1	4238	4249	4704	4715	88	17		
2	32	5692	5703	6083	6094	58	1		
3	70	3530	3537	3619	3626	1	21		
4	2	6975	6986	7120	7131	40	57		
5	3	3558	3568	3585	3595	2	22		
6	15	4904	4914	5222	5232	82	2		
7	84	288	295	371	378	3	23		
8	89	4899	4903	5234	5238	85	3		
9	4	6959	6971	7134	7146	54	70		
10	23	116	123	128	135	4	24		
Flavivirus TBEV									
1	1	10848	10861	10908	10921	209	57		
2	112	9238	9244	9661	9667	62	1		
3	158	2	9	359	366	1	159		
4	2	9160	9171	9175	9186	83	51		
5	32	5204	5214	5230	5240	2	61		
6	104	6321	6326	6626	6631	54	2		
7	3	130	138	144	152	6	139		
8	6	209	217	246	254	3	187		
9	195	9220	9224	9678	9682	68	3		
10	4	10216	10223	10229	10236	36	29		

[a] Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.

[b] NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID in NASP.

[c] Syn. Sub. ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.

[d] Coevo. ranking is a ranking based on complementarily coevolution. NA indicates structures which had no variable sites.

Cons. Rank [a]	NASP rank / ID [b]	Structure coordinates including gaps				Syn. Sub. Rank [c]	Coevo. Rank [d]	Biological function	Reference
		5' stem		3' stem					
Flavivirus YFV									
1	1	4572	4584	4589	4601	50	103		
2	12	1	10	254	263	1	27		
3	150	9590	9597	9628	9635	20	1		
4	2	10808	10820	10844	10856	155	42	Important in virus pathogenicity	Proutski et al, 1997
5	121	11	17	246	252	2	35		
6	140	9321	9328	9554	9561	117	2		
7	3	78	87	91	100	156	165		
8	46	4749	4755	4760	4766	139	3		
9	71	20	25	238	243	3	36		
10	4	723	735	754	766	66	63		
Hepacivirus HCV									
1	1	9294	9314	9319	9339	162	405		
2	224	317	323	749	755	1	374		
3	261	7421	7431	7818	7828	413	1		
4	2	9398	9412	9418	9432	56	395		
5	310	7498	7504	7673	7679	358	2		
6	360	326	330	742	746	2	373		
7	3	393	403	422	432	11	384		
8	152	7506	7513	7665	7672	333	3		
9	436	367	369	507	509	3	166		
10	4	6051	6063	6116	6128	396	137		
Pestivirus BVDV-1									
1	1	7363	7374	7411	7422	16	254		
2	167	399	406	517	524	1	99		
3	399	11743	11751	11826	11834	335	1		
4	2	8989	9000	9018	9029	168	286		
5	324	8295	8300	8505	8510	346	2		
6	446	760	764	811	815	2	219		
7	3	3	12	25	34	414	297		
8	44	775	780	794	799	3	311		
9	248	6001	6011	6124	6134	123	3		
10	4	5111	5120	5174	5183	57	147		

[a] Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b] NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID in NASP.
[c] Syn. Sub. ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d] Coevo. ranking is a ranking based on complementarily coevolution. NA indicates structures which had no variable sites.

Cons. Rank [a]	NASP rank / ID [b]	Structure coordinates including gaps				Syn. Sub. Rank [c]	Coevo. Rank [d]	Biological function	Reference
		5' stem		3' stem					
Pestivirus BVDV-2									
1	1	273	282	286	295	246	136		
2	51	156	164	407	415	1	33		
3	70	4423	4433	4446	4456	119	1		
4	2	6515	6524	6531	6540	2	204		
5	106	11167	11176	11661	11670	134	2		
6	3	1	9	23	31	247	142		
7	242	11254	11257	11262	11265	3	205		
8	260	5003	5011	5222	5230	15	3		
9	4	6288	6296	6302	6310	190	115		
10	52	11446	11459	11616	11629	184	4		
Pestivirus CSFV									
1	1	259	267	273	281	179	91		
2	171	6386	6401	6933	6948	146	1		
3	180	50	54	391	395	1	63		
4	2	12159	12172	12178	12191	180	191		
5	118	6507	6512	6674	6679	141	2		
6	145	4509	4519	4759	4769	2	169		
7	3	9108	9119	9248	9259	93	114		
8	4	6514	6524	6662	6672	122	3		
9	84	57	64	381	388	3	73		
10	90	6491	6498	6707	6714	112	4		
Hepevirus HEV									
1	1	6309	6318	6445	6454	4	10		
2	44	6277	6284	6540	6547	29	1		
3	56	5235	5245	5276	5286	1	72		
4	2	4146	4155	4159	4168	68	34		
5	22	6324	6333	6432	6441	2	9		
6	99	6272	6276	6551	6555	35	2		
7	3	3984	3992	3997	4005	114	54		
8	144	6264	6269	6558	6563	37	3		
9	152	5380	5385	5482	5487	3	38		
10	4	1065	1078	1092	1105	42	109		

[a] Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.

[b] NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID in NASP.

[c] Syn. Sub. ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.

[d] Coevo. ranking is a ranking based on complementarily coevolution. NA indicates structures which had no variable sites.

Cons. Rank [a]	NASP rank / ID [b]	Structure coordinates including gaps				Syn. Sub. Rank [c]	Coevo. Rank [d]	Biological function	Reference
		5' stem		3' stem					
Luteovirus BYDV									
1	1	1759	1775	1794	1810	66	75		
2	23	3474	3481	4192	4199	104	1		
3	86	2478	2484	2730	2736	1	68		
4	2	195	211	245	261	112	62		
5	30	2518	2523	2694	2699	2	98		
6	109	3483	3486	4187	4190	109	2		
7	3	3186	3198	3237	3249	13	108		
8	5	2598	2607	2664	2673	3	84		
9	52	3490	3501	3918	3929	115	3		
10	4	1430	1440	1450	1460	101	4		
Polerivirus PLRV									
1	1	5087	5096	5101	5110	39	91		
2	108	4618	4621	4662	4665	1	15		
3	175	1202	1206	1298	1302	132	1		
4	2	2075	2084	2092	2101	50	101		
5	17	5505	5512	5517	5524	31	2		
6	120	5115	5121	5177	5183	2	28		
7	3	3291	3299	3305	3313	107	141		
8	9	4402	4409	4419	4426	141	3		
9	81	4864	4870	4909	4915	3	4		
10	4	3764	3773	3936	3945	102	53		
Polerivirus ScYLV									
1	1	3782	3792	3988	3998	154	104		
2	83	5043	5051	5232	5240	114	1		
3	108	575	584	590	599	1	84		
4	2	626	637	858	869	33	7		
5	103	2113	2119	2193	2199	2	76		
6	123	1684	1689	1729	1734	91	2		
7	3	875	885	917	927	70	95		
8	93	4490	4498	4509	4517	134	3		
9	98	1994	2000	2044	2050	3	88		
10	4	4	15	25	36	165	55		

[a] Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.

[b] NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID in NASP.

[c] Syn. Sub. ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.

[d] Coevo. ranking is a ranking based on complementarily coevolution. NA indicates structures which had no variable sites.

Cons. Rank	NASP rank / ID	Structure coordinates including gaps				Syn. Sub. Rank	Coevo. Rank	Biological function	Reference
		5' stem		3' stem					
Aphthovirus FMDV									
1	1	7998	8008	8015	8025	33	242		
2	216	4267	4269	4581	4583	1	190		
3	304	6031	6035	8026	8030	3	1		
4	2	1418	1428	1434	1444	207	94		
5	48	4274	4279	4547	4552	2	278		
6	295	6472	6476	6853	6857	89	2		
7	3	2618	2636	2657	2675	261	214		
8	167	5087	5090	5230	5233	200	3		
9	4	5824	5836	5841	5853	158	107		
10	45	4181	4191	4629	4639	4	43		
Enterovirus HEV-A									
1	1	598	610	615	627	157	124	IRES domain	Thompson and Sarnow, 2003
2	5	4496	4503	4529	4536	1	149	CRE element	Paul et al, 2000
3	154	4625	4634	5098	5107	42	1		
4	2	334	342	363	371	158	133	IRES domain	Thompson and Sarnow, 2003
5	11	4504	4508	4523	4527	2	139	CRE element	Paul et al, 2000
6	71	5967	5973	6349	6355	123	2		
7	3	386	393	398	405	159	66	IRES domain	Thompson and Sarnow, 2003
8	89	4511	4512	4520	4521	3	181		
9	186	5977	5981	6342	6346	108	3		
10	4	146	154	161	169	160	93	IRES domain	Thompson and Sarnow, 2003
157	108	5043	5046	5052	5055	77	109		Figure 3.9, EVA1
Enterovirus HEV-B									
1	1	7462	7473	7490	7501	204	NA	Negative strand synthesis	van Ooij et al, 2006
2	16	7319	7328	7404	7413	1	NA		
3	2	606	618	623	635	205	NA	IRES domain	Bailey and Tapprich, 2006
4	202	6732	6737	7032	7037	2	NA		
5	3	10	18	26	34	206	NA	IRES domain	Bailey and Tapprich, 2006
6	71	6834	6842	7000	7008	3	NA		
7	4	338	347	366	375	207	NA	IRES domain	Bailey and Tapprich, 2006
8	9	4491	4498	4524	4531	4	NA	CRE element	Cordey et al, 2008
9	5	390	398	404	412	208	NA	IRES domain	Bailey and Tapprich, 2007
10	55	7338	7342	7363	7367	5	NA		
13	7	5059	5073	5087	5101	113	NA		Figure 3.9, EVA2

[a] Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b] NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID in NASP.
[c] Syn. Sub. ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d] Coevo. ranking is a ranking based on complementarily coevolution. NA indicates structures which had no variable sites.

Cons. Rank [a]	NASP rank / ID [b]	Structure coordinates including gaps				Syn. Sub. Rank [c]	Coevo. Rank [d]	Biological function	Reference
		5' stem		3' stem					
Enterovirus HEV-C									
1	1	595	609	613	627	148	Na	IRES domain	Malnou et al, 2002
2	6	4557	4563	4593	4599	1	Na	CRE element	Murray and Barton, 2003
3	2	384	392	396	404	149	Na	IRES domain	Malnou et al, 2002
4	36	7093	7098	7189	7194	2	Na		
5	3	332	340	361	369	150	Na	IRES domain	Malnou et al, 2002
6	131	7100	7104	7122	7126	3	Na		
7	4	5069	5080	5099	5110	120	Na		Figure 3.9, EVA3
8	97	4548	4554	4600	4606	4	Na	CRE element	Murray and Barton, 2003
9	5	5869	5878	6019	6028	22	Na		
10	29	5821	5825	6090	6094	5	Na		
Enterovirus DHAV									
1	1	239	253	326	340	82	25		
2	2	4027	4039	4055	4067	1	81		
3	62	1683	1691	1833	1841	7	1		
4	19	3953	3970	4088	4105	2	28		
5	32	7617	7624	7766	7773	94	2		
6	3	13	24	57	68	83	95		
7	8	7688	7695	7743	7750	86	3		
8	83	1253	1258	1265	1270	3	96		
9	4	470	479	524	533	84	5		
10	39	487	490	497	500	97	4		
Enterovirus HRV-A									
1	1	594	616	620	642	58	244		
2	4	3380	3391	3413	3424	1	165		
3	102	3635	3643	3883	3891	53	1	CRE element	Tapparel et al, 2007
4	2	7047	7063	7069	7085	94	292		
5	50	7107	7114	7144	7151	2	279		
6	245	3668	3671	3852	3855	110	2		
7	3	333	341	362	370	275	160		
8	199	3699	3703	3783	3787	109	3		
9	229	7100	7104	7155	7159	3	280		
10	19	7034	7045	7160	7171	4	281		

[a] Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.

[b] NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID in NASP.

[c] Syn. Sub. ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.

[d] Coevo. ranking is a ranking based on complementarily coevolution. NA indicates structures which had no variable sites.

Cons. Rank [a]	NASP rank / ID [b]	Structure coordinates including gaps				Syn. Sub. Rank [c]	Coevo. Rank [d]	Biological function	Reference
		5' stem		3' stem					
Enterovirus HRV-B									
1	1	285	298	423	436	81	29		
2	7	502	510	524	532	85	1		
3	16	2357	2363	2370	2376	1	44		
4	2	606	620	624	638	8	17		
5	4	2343	2356	2381	2394	2	21		
6	24	496	501	546	551	92	2		
7	3	347	355	376	384	82	8		
8	41	511	513	520	522	96	3		
9	93	734	739	749	754	3	45		
10	49	1206	1211	1229	1234	69	4		
Rhinovirus HRV-C									
1	1	588	607	621	640	12	88		
2	8	853	858	895	900	1	135		
3	102	4724	4734	6407	6417	120	1		
4	2	860	869	884	893	2	149		
5	11	5047	5056	5291	5300	99	2		
6	3	332	340	361	369	135	19		
7	20	5087	5096	5280	5289	103	3		
8	110	467	472	1076	1081	3	76		
9	4	383	389	394	400	136	75		
10	29	5909	5917	6339	6347	74	4		
Hepatovirus HAV									
1	1	7349	7360	7384	7395	7	69		
2	111	1178	1189	1403	1414	91	1		
3	112	6014	6018	6086	6090	1	104		
4	2	629	639	716	726	106	8		
5	36	6020	6026	6079	6085	2	96		
6	87	1100	1111	1979	1990	105	2		
7	3	7423	7431	7444	7452	11	111		
8	22	7409	7414	7461	7466	3	106		
9	98	6236	6247	6317	6328	60	3		
10	4	2	14	31	43	107	19		

[a] Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.

[b] NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID in NASP.

[c] Syn. Sub. ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.

[d] Coevo. ranking is a ranking based on complementarily coevolution. NA indicates structures which had no variable sites.

Cons. Rank [a]	NASP rank / ID [b]	Structure coordinates including gaps				Syn. Sub. Rank [c]	Coevo. Rank [d]	Biological function	Reference
		5' stem		3' stem					
Cardiovirus EMCV									
1	1	6304	6320	6325	6341	19	62		
2	5	406	413	437	444	114	1		
3	13	4180	4189	4285	4294	1	43		
4	2	6952	6964	6969	6981	9	119		
5	19	6709	6715	6723	6729	10	2		
6	73	4040	4046	4068	4074	2	71		
7	3	346	357	361	372	113	7		
8	61	4087	4095	4113	4121	3	73		
9	120	2829	2839	2890	2900	31	3		
10	4	6816	6826	6922	6932	13	58		
Cardiovirus SAFV									
1	1	21	38	43	60	46	4		
2	8	13	20	64	71	51	1		
3	56	7625	7632	8053	8060	1	60		
4	2	710	724	843	857	47	14		
5	47	947	953	978	984	68	2		
6	58	7646	7651	8037	8042	2	59		
7	3	2	12	75	85	48	5		
8	15	4242	4248	4261	4267	3	62		
9	31	1425	1439	1447	1461	15	3		
10	4	7855	7866	7876	7887	34	35		
Teschovirus PTV									
1	1	6245	6259	6263	6277	30	80		
2	71	6749	6753	7060	7064	1	116		
3	139	1633	1642	1995	2004	39	1		
4	2	3335	3346	3353	3364	119	63		
5	26	1578	1589	1602	1613	34	2		
6	44	6755	6759	7054	7058	2	115		
7	3	4232	4239	4250	4257	82	128		
8	5	6764	6772	7042	7050	3	114		
9	125	669	676	683	690	68	3		
10	4	6545	6555	6654	6664	42	88		

[a] Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b] NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID in NASP.
[c] Syn. Sub. ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d] Coevo. ranking is a ranking based on complementarily coevolution. NA indicates structures which had no variable sites.

Cons. Rank [a]	NASP rank / ID [b]	Structure coordinates including gaps				Syn. Sub. Rank [c]	Coevo. Rank [d]	Biological function	Reference
		5' stem		3' stem					
Parechovirus HPeV									
1	1	6966	6978	7040	7052	72	52		
2	51	1390	1394	1436	1440	1	12		
3	132	5649	5661	6884	6896	144	1		
4	2	3730	3740	3767	3777	24	47		
5	88	6932	6936	7139	7143	2	115		
6	119	6091	6100	6804	6813	146	2		
7	3	13	24	61	72	148	71		
8	46	1398	1404	1428	1434	3	15		
9	158	6238	6245	6405	6412	145	3		
10	4	619	625	641	647	149	22		
Potyvirus PVY									
1	1	9402	9411	9419	9428	85	87		
2	42	1080	1089	1206	1215	11	1		
3	61	2880	2886	3033	3039	1	7		
4	2	1691	1700	1705	1714	16	54		
5	5	9310	9318	9381	9389	2	73		
6	26	9255	9261	9271	9277	14	2		
7	3	9335	9343	9353	9361	4	74		
8	27	2587	2598	3278	3289	3	30		
9	33	4401	4408	4426	4433	35	3		
10	4	2259	2269	2286	2296	81	51		
22	12	9103	9111	9120	9128	9	75		Figure 3.12, PVA3
Potyvirus SMV									
1	1	1342	1353	1442	1453	104	24		
2	19	7366	7375	7560	7569	1	23		
3	101	201	212	1846	1857	60	1		
4	2	5862	5871	5876	5885	111	44		
5	23	6118	6125	6580	6587	79	2		
6	41	8000	8008	8022	8030	2	88		
7	3	2546	2560	3351	3365	133	87		
8	8	1382	1388	1414	1420	117	3		
9	73	6856	6861	6947	6952	3	37		
10	4	7956	7966	8035	8045	16	59		
13	5	9060	9068	9077	9085	12	91		Figure 3.12, PVA2

[a] Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b] NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID in NASP.
[c] Syn. Sub. ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d] Coevo. ranking is a ranking based on complementarily coevolution. NA indicates structures which had no variable sites.

Cons. Rank [a]	NASP rank / ID [b]	Structure coordinates including gaps				Syn. Sub. Rank [c]	Coevo. Rank [d]	Biological function	Reference
		5' stem		3' stem					
Potyvirus TuMV									
1	1	239	250	254	265	41	95		
2	128	8261	8271	8362	8372	148	1		
3	141	116	123	485	492	1	135		
4	2	441	454	459	472	81	41		
5	95	1510	1517	1632	1639	142	2		
6	140	3097	3106	3186	3195	2	34		
7	3	9174	9183	9244	9253	12	127		
8	63	313	319	376	382	3	104		
9	108	1215	1225	1648	1658	159	3		
10	4	9271	9278	9295	9302	13	136		Figure 3.13, PVB1
Potyvirus ZYMV									
1	1	1268	1278	1283	1293	26	29		
2	8	2937	2946	3006	3015	1	13		
3	12	1221	1230	1325	1334	20	1		
4	2	1067	1076	1105	1114	28	6		
5	20	4073	4082	4140	4149	27	2		
6	25	9110	9116	9125	9131	2	18		Figure 3.12, PVA1
7	3	1255	1264	1298	1307	19	3		
8	6	1772	1779	1850	1857	3	7		
9	4	3922	3931	3938	3947	29	14		
10	24	1786	1791	1837	1842	4	19		
Potyvirus BCMV									
1	1	5849	5863	5901	5915	76	131		
2	145	7177	7184	7410	7417	204	1		
3	167	9208	9214	9858	9864	1	73		
4	2	7942	7954	7987	7999	142	119		
5	211	7170	7175	7419	7424	212	2		
6	222	3324	3329	3452	3457	2	78		
7	3	237	246	252	261	42	43		
8	95	7198	7206	7370	7378	182	3		
9	96	9295	9303	9848	9856	3	75		
10	4	3261	3272	3299	3310	40	8		
51	19	9517	9526	9555	9564	63	123		Figure 3.12, PVA5

[a] Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b] NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID in NASP.
[c] Syn. Sub. ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d] Coevo. ranking is a ranking based on complementarily coevolution. NA indicates structures which had no variable sites.

Cons. Rank [a]	NASP rank / ID [b]	Structure coordinates including gaps				Syn. Sub. Rank [c]	Coevo. Rank [d]	Biological function	Reference
		5' stem		3' stem					
Potyvirus BYMV									
1	1	868	882	1335	1349	131	29		
2	99	4394	4400	4445	4451	120	1		
3	104	9312	9317	9401	9406	1	10		
4	2	1035	1048	1178	1191	126	74		
5	43	2639	2647	3199	3207	111	2		
6	105	9323	9328	9392	9397	2	19		
7	3	5413	5422	5427	5436	67	118		
8	27	3569	3575	3644	3650	36	3		
9	51	6429	6434	6443	6448	3	92		
10	4	3597	3607	3618	3628	16	99		
Potyvirus PRSV									
1	1	9579	9589	9606	9616	50	78		
2	64	9891	9896	10277	10282	1	34		
3	163	117	122	225	230	37	1		
4	2	9424	9435	9564	9575	32	79		
5	26	1597	1604	1626	1633	132	2		
6	27	9880	9886	10288	10294	2	23		
7	3	9488	9497	9532	9541	24	130		
8	130	10055	10058	10179	10182	3	82		
9	144	6863	6871	7020	7028	155	3		
10	4	5881	5890	5896	5905	29	60		
Potyvirus PPV									
1	1	1882	1890	1908	1916	84	38		
2	25	8577	8584	8741	8748	88	1		
3	37	2975	2983	3347	3355	1	37		
4	2	3231	3238	3289	3296	44	81		
5	4	9494	9502	9648	9656	2	20		
6	9	9601	9608	9617	9624	100	2		
7	3	924	933	940	949	52	56		
8	26	4637	4646	4662	4671	98	3		
9	49	3044	3051	3214	3221	3	54		
10	85	9504	9507	9639	9642	4	31		
15	6	9254	9261	9278	9285	10	87		Figure 3.13, PVB2

[a] Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b] NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID in NASP.
[c] Syn. Sub. ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d] Coevo. ranking is a ranking based on complementarily coevolution. NA indicates structures which had no variable sites.

Cons. Rank [a]	NASP rank / ID [b]	Structure coordinates including gaps				Syn. Sub. Rank [c]	Coevo. Rank [d]	Biological function	Reference
		5' stem		3' stem					
Potyvirus WMV									
1	29	7856	7863	7957	7964	40	1		
2	14	444	456	2337	2349	1	2		
3	1	9520	9528	9537	9545	2	45		Figure 3.12, PVA4
4	2	6268	6277	6282	6291	46	55		
5	53	6655	6664	6704	6713	43	3		
6	3	9196	9206	9234	9244	45	4		
7	66	9473	9483	9554	9564	3	31		
8	4	3349	3359	3412	3422	12	6		
9	39	999	1004	1010	1015	4	33		
10	12	6159	6168	6201	6210	48	5		
Potyvirus SCMV									
1	1	9013	9020	9025	9032	66	139		
2	28	9101	9109	9593	9601	1	117		
3	92	1175	1179	1222	1226	99	1		
4	2	1188	1197	1202	1211	111	3		
5	41	1183	1187	1214	1218	131	2		
6	57	2760	2770	3137	3147	2	108		
7	3	9033	9039	9090	9096	35	65		
8	26	9263	9270	9345	9352	3	88		
9	4	7235	7241	7302	7308	27	32		
10	137	9258	9262	9588	9592	4	119		
Rubivirus RUBV									
1	1	9304	9316	9320	9332	32	55		
2	29	8004	8014	8534	8544	2	1		
3	70	7930	7941	8596	8607	1	2		
4	2	9576	9588	9601	9613	9	44		
5	3	5875	5886	5891	5902	60	71		
6	20	8146	8156	8374	8384	3	15		
7	74	2170	2179	2184	2193	29	3		
8	4	6483	6492	6507	6516	43	35		
9	32	6329	6338	6364	6373	59	4		
10	72	9477	9481	9505	9509	4	9		

[a] Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b] NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID in NASP.
[c] Syn. Sub. ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d] Coevo. ranking is a ranking based on complementarily coevolution. NA indicates structures which had no variable sites.

Cons. Rank [a]	NASP rank / ID [b]	Structure coordinates including gaps				Syn. Sub. Rank [c]	Coevo. Rank [d]	Biological function	Reference
		5' stem		3' stem					
Sobemovirus RYMV									
1	1	895	911	925	941	157	166		
2	107	4092	4101	4121	4130	93	1		
3	142	3557	3561	3597	3601	1	85		
4	2	2455	2471	2479	2495	158	41		
5	13	2539	2548	2860	2869	2	39		
6	94	4218	4224	4277	4283	171	2		
7	3	2647	2658	2680	2691	11	34		
8	52	1184	1189	1276	1281	143	3		
9	63	2551	2557	2699	2705	3	14		
10	4	753	763	820	830	34	104		
Arterivirus PRRSV									
1	1	9579	9589	9606	9616	50	78		
2	64	9891	9896	10277	10282	1	34		
3	163	117	122	225	230	37	1		
4	2	9424	9435	9564	9575	32	79		
5	26	1597	1604	1626	1633	132	2		
6	27	9880	9886	10288	10294	2	23		
7	3	9488	9497	9532	9541	24	130		
8	130	10055	10058	10179	10182	3	82		
9	144	6863	6871	7020	7028	155	3		
10	4	5881	5890	5896	5905	29	60		

[a] Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b] NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID in NASP.
[c] Syn. Sub. ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d] Coevo. ranking is a ranking based on complementarily coevolution. NA indicates structures which had no variable sites.