THE IDENTIFICATION OF BIOLOGICALLY IMPORTANT SECONDARY STRUCTURES IN DISEASE-CAUSING RNA VIRUSES

by Emil P. Tanov

Dr. Gordon W. Harkins, Supervisor

Prof. Alan Christoffels, Co-Supervisor

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

22 November, 2012

UNIVERSITY of the
WESTERN CAPE

ACKNOWLEDGEMENTS

Before I thank my supervisors, colleagues and peers for their invaluable input and support during the last two years, I would like to thank my parents, for theirs, during the last 26.

I was fortunate to work with an amazing group of people on this project and would like to express my eternal gratitude to Dr. Darren P. Martin, Michael Golden, Brejnev Muhire, my supervisor Dr. Gordon W. Harkins, co-supervisor Prof. Alan Christoffels for their guidance and support.

Last but not least, I would like to thank my family and friends who have been a positive influence in my life during this time and have unknowingly contributed to the completion of this work, helping me maintain an acceptable level of deviation from sanity.

"The beginning of the end of any human is the moment he starts to believe that he knows it all"

Anonymous

ABSTRACT

Viral genomes consist of either deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). The viral RNA molecules are responsible for two functions, firstly, their sequences contain the genetic code, which encodes the viral proteins, and secondly, they may form structural elements important in the regulation of the viral life-cycle. Using a host of computational and bioinformatics techniques we investigated how predicted secondary structure may influence the evolutionary dynamics of a group of single-stranded RNA viruses from the *Picornaviridae* family. We detected significant and marginally significant correlations between regions predicted to be structured and synonymous substitution constraints in these regions, suggesting that selection may be acting on those sites to maintain the integrity of certain structures. Additionally, coevolution analysis showed that nucleotides predicted to be base paired, tended to co-evolve with one another in a complimentary fashion in four out of the eleven species examined. Our analyses were then focused on individual structural elements within the genome-wide predicted structures. We ranked the predicted secondary structural elements according to their degree of evolutionary conservation, their associated synonymous substitution rates and the degree to which nucleotides predicted to be base-paired coevolved with one another. Top ranking structures coincided with well characterised secondary structures that have been previously described in the literature. We also assessed the impact that genomic secondary structures had on the recombinational dynamics of picornavirus genomes, observing a strong tendency for recombination breakpoints to occur in non-coding regions. However, convincing evidence for the association between the distribution of predicted RNA structural elements and breakpoint clustering was not detected.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Structural elements in viral genomes

Viruses share many homologous features and infect all forms of cellular life, suggesting that they have a common ancestor that likely existed very early in the evolution of life, perhaps before even the origin of the "three domains" of life (Forterre 2006). A multitude of viral species have been isolated from a wide range of hosts, ranging from large marine mammals in the Pacific Ocean (Smith et al. 1979); Rivera et al. (2010) to bacteria in central Sahara (Fancello et al. 2012), and some have even been found to infect other viruses (La Scola et al. 2008).

The viral particle, or virion, is made up of nucleic acid genome (either DNA or RNA) encoding the viral proteins and a protein coat encapsulating the viral genome. However, the size, composition and structure of their genomes vary greatly between different viral species. Viral genomes may be double-stranded or single-stranded and orientated in a linear or circular configuration. Single-stranded genomes, such as the RNA genomes of the *Picornaviridae* family, have the potential to form a greater variety of structures than the more rigid structural conformation of dsDNA. The RNA molecules in single-stranded genomes are able to assume secondary and tertiary structuresby hydrogen bonding between guanines (G) and cytosines (C), and adenines (A) and uracil (U) bases and the less stable G-U base pairs (Watson and Crick 1953). In the horizontal plane of the bases, hydrogen bond interactions are responsible for maintaining pairing, while dispersion forces and hydrophobic interactions are responsible for the base stacking effect in the perpendicular plane of the structure (Yakovchuk et al. 2006). Whereas the secondary structure of single-stranded nucleic acids describes the set of hydrogen bond interactions between base pairs within the sequence, its tertiary structural arrangement in three-dimensional space. The base-pairing of complementary nucleotides in the DNA double helix is an example of secondary structure while the A, B, and Z conformations of double-stranded DNA are examples of tertiary structure (Richmond and Davey 2003). In the case of viruses with single-stranded nucleic acid genomes, there are many examples of functionally important genomic secondary structures that have been conserved during the course of viral evolution. In the absence of selection, a few random mutations are sufficient to disrupt structural motifs (Fontana et al. 1993). Thus, whenever conserved structures are evident, it is likely that these potentially have

a biological purpose that confers some selective advantage. Conserved structures found in genome regions within non-translated sequences of viral genomes are particularly common and in many cases have been shown to play important roles in genome replication and the control of gene expression (Jayan and Casey 2005, Kieft et al. 2001). Biologically functional secondary structures can also occur in coding regions (Hofacker et al. 2004, Pollard and Malim 1998) where their evolutionary maintenance can place strong constraints on the evolution of encoded protein.

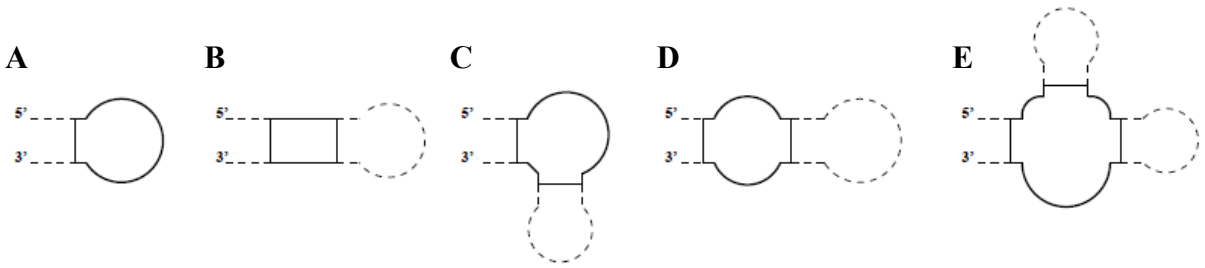


Figure 1. Morphology types of secondary structure. Secondary structures can be divided into five distinct types which form the basis of the additive energy model. These are (A) hairpin loop, (B) stem, (C) bulge, (D) internal loop and (E) multiloop. Dashes are representative of nucleotide bases, whereas the solid line highlights the structure

1.2 Secondary structure prediction

RNA secondary structure motifs can be determined using either computational or experimental methods. The algorithms used for the prediction of RNA secondary structures are typically based on thermodynamic rules. The most widely used methods compute a single minimum free energy structure through dynamic programming (Zuker and Stiegler 1981). There are however, a number of different approaches to computationally infer secondary structures (Gultyaev et al. 1995) and there is no real consensus in the field as to which is best. So far methods based on the kinetic analysis of self-organising molecules (Mironov et al. 1985) have not proved to be significantly better than minimum free energy based thermodynamic folding methods.

Because of over-simplification of the energy models and inaccuracies and approximations in the measured parameters (thermodynamic increments for; G-U pairs, mismatches and loop regions) a high degree of uncertainty exists regarding the reliability of the exact base-pairings identified by these methods. In cases where the correct structure is known, it has been found that only between 40% and 80% of the predicted base-pairs actually exist (Doshi et al. 2004). As a consequence of such uncertainty in the accuracy of predicted base-pairings,

thermodynamic predictions of base pairing probabilities are an ideal starting point in comparative studies aimed at testing the biological importance or evolutionary impacts of predicted secondary structures. Rather than focusing on the inference of the structure with the lowest estimated free energy, there are variations of such folding algorithms for computing samples of suboptimal folds (Zuker 1989) or even all structures within a prescribed energy range (Wuchty et al. 1999). Similarly, non-deterministic kinetic folding algorithms (Flamm et al. 2000) can produce ensembles of structures by repeatedly refolding the sequences from randomly determined starting points. A much more elegant and efficient solution 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 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.

Because of such problems with the accurate computational prediction of secondary structures within individual sequences a number of alternative approaches have been developed to more accurately infer secondary structures. These include (1) using experimentally determined data to determine whether individual nucleotides are folded (achievable, for example, using selective 2'-hydroxyl acylation analysed by primer extension, or simply, SHAPE; (Wilkinson et al. 2006) and (2) methods that, rather than attempting to accurately fold individual sequences, utilise information on evolutionary conservation or nucleotide co-variation drawn from comparative analyses of multiple homologous sequences.

Box 1: SHAPE

Despite advances in computational methods for the prediction of secondary structures within a given single-stranded RNA virus genome (or any other single-stranded RNA sequence) their accuracy remains a concern (Deigan et al. 2009). It is possible to use an experimental approach called SHAPE (Wilkinson et al. 2006), which is able to report whether nucleotides in the RNA sequence are base-paired or unpaired, based on adduct formation of chemical reagents on individual nucleotide bases. Modifications are identified as stops during a primer extension reaction with reverse transcriptase and compared to the results from an unmodified control to yield an accurate biophysical measurement of the RNA dynamics within the sequence of interest. 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 then be used as an experimental correction to a folding prediction algorithm to obtain highly accurate models of the RNA folding within the sequence (Deigan et al. 2009).

Traditionally, investigations of viral structural elements have focused their attention on a few well-conserved individual structures located within the non-coding region of viral genomes. Many such structures are known to have key functions during viral replication and gene expression (Lu and Wimmer 1996) including, for example, internal ribosomal entry sites (IRESs) in Poliovirus (Pelletier and Sonenberg 1988) and the virion strand origin of replication in circular single-stranded DNA viruses (Ravetch et al. 1977)

Viewing structures and their function in an isolated manner is changing towards a more 'global' perspective due to increasing evidence that there exist critically functional long-range interactions in many positive-strand RNA viruses (Khromykh et al. 2001, Zhang et al. 2008) and retroviruses (Abbink and Berkhout 2003, Ooms et al. 2007). Recently, a study presented evidence suggesting correlation between the extent of secondary structure found in several positive-stranded viruses and their degree of viral persistence during infections (Davis et al. 2008). Structure formation of the entire length of their genomes was predicted using free energy minimisation techniques, observing increased persistence during infections of those viruses which contained more overall genome wide secondary structure, termed Genome-Scale Ordered RNA Structure (GORS), as compared to those that had less overall evidence of secondary structure. Atomic microscopy analyses have shown that viruses are

able to adopt pseudo-globular conformations (Davis et al. 2008), 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) (Simmonds et al. 2004), thus avoiding detection by RNA interference (RNAi) factors. RNAi is a gene regulation system essential for maintaining the integrity of the host cell genome. Small interfering RNAs (siRNAs) are synthesised by the host to recognise and bind to specific complementary sequences on the viral genomes, targeting them for destruction or inhibiting their translation. Some viruses (e.g. tombusviruses) encode proteins which can bind to these siRNAs and inhibit their ability to form RNA-induced silencing complexes (RISC) (Ye et al. 2003), whereas others (such as flaviviruses and aphthoviruses) rely on elaborate secondary structures at these target sites to decrease the efficiency of RNAi binding (Shao et al. 2007, Simmonds et al. 2004). In addition, it is believed viral genomes form dynamic meta-stable structures which are able to readily accommodate changes of conformation due to environmental pressures (Simmonds et al. 2004). It is in fact been shown that viruses may be able to restrict accessibility of RNAi factors to target sites on their genomes by rapidly evolving the conformation of their structures (Tafer et al. 2008). The role of these mechanisms of sequestering viral sequences from interaction with siRNAs may have clinical relevance, because synthetic siRNAs have potential applications as antiviral therapeutics.

Besides the influences of over-all degrees of genomic secondary structure on the long term survival of single stranded RNA viruses during chronic infections, the genome-scale arrangements of secondary structural elements likely also has a crucial impact on how viruses express their genes. In HIV-1, for example, SHAPE analyses has been used to propose a genome-scale secondary structure model (Watts et al. 2009), where it was found that sequences encoding both the inter-protein linkers within polyproteins and inter-domain regions within individual proteins, contained more structured regions than could be accounted for by chance as compared to the rest of the genome. These structures are apparently involved in ribosomal pausing occurring at inter-protein linkers and inter–domain sites so as to enable functionally distinct parts of proteins to fold in an independent manner. During translation, ribosomal pausing seems to prevent the interference of parts of the protein that have already been translated with those parts that are still to be translated (Willis 1993).

1.3 The evolutionary impacts of genomic secondary structure

Viruses are constantly trying to escape the host's defences by employing various strategies. The accumulation of mutations in their genomes is one of the ways viruses are able to alter their "appearance" to the host's immune system. However, the accumulation of mutations has to occur at rates that enable viruses to retain their viability and genetic identity. Maintaining this delicate equilibrium is vital for viruses, and it strongly influences their pathogenicity and replication success.

In addition to the accumulation of mutations, recombination plays a vital role in amassing genetic diversity amongst RNA viruses. It enables them to rapidly access greater areas of sequence space than is possible by the stepwise accumulation of point mutations alone (Domingo and Holland 1997). This helps to facilitate both the fixation of advantageous mutations and the purging of deleterious mutations from viral populations (Moya et al. 2000). In many of the RNA viruses, these evolutionary mechanisms contribute to the evasion of immune responses and the development of drug resistance (Johnson and Desrosiers 2002).

It is expected that both the mutational and recombinational dynamics of RNA viruses could be strongly influenced by their genomic secondary structures. The rate at which mutations in viruses arise may, in part, be influenced by the extent of secondary structure elements present in their genomes. Analysis of viral genome sequences in evolution experiments indicate that ssDNA is more susceptible to oxidative damage, than regions where DNA is in the double-stranded state (Xia and Yuen 2005). While 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 structure, the rate of cytosine deamination is notably slower (>100 fold) in double-stranded DNA as compared to single-stranded DNA (Frederico et al. 1990). Mutations which arise in a population may eventually become fixed due to the probable selective pressure of secondary structure on the underlying nucleotide sequence. Extremely low synonymous substitution variation rate was observed in the well-conserved and highly structured rev response element (RRE) region in HIV-1, displaying evidence of purifying selection acting on those sites (Ngandu et al. 2008). Evidence exists, suggesting that natural selection may act to maintain some secondary structures within viral genomes. In a viral population of maize streak virus, a mutation introduced within a particular structure,

disrupting pairing, was reverted, restoring the initial structural conformation (Shepherd et al. 2006).

Genomic secondary structures can potentially also have two distinct impacts on the recombination dynamics of RNA viruses, both determining where recombination events are most likely to occur, and determining which recombinants that arise are most likely to survive. 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. A study (Galetto et al. 2004), implicated a hairpin structure located on the C2 portion of the gp120 envelope gene of HIV-1 with a recombination hot-spot at the loop of this structure. By varying the stability of the hairpin without altering its sequence, they showed that they could significantly alter recombination patterns occurring in the envelope gene. Additionally, on an RNA lacking a stable hairpin, recombination rates in that region fell drastically in comparison with sequences in which the hairpin loop was present. It has since been found 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). 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. 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).

It has been suggested that when two parental sequences share a similar secondary structure, it predisposes these parental sequences to base pair with one another within these structures to form heteroduplexes (Dedepsidis et al. 2010). In the 2C and 3D gene regions of polioviruses, most recombination junctions occur in regions containing secondary structure that is similar between the recombining partners. It is likely that when the poliovirus RdRp in conjunction with the nascent negative strand reaches such heteroduplex regions, the 3' end of the nascent negative strand may become detached from the initial template molecule and then re-attach to the second molecule within the heteroduplex which then becomes the new template such that

the newly synthesised RNA molecule will be a recombinant of the two heteroduplex forming molecules.

Recombination events occurring outside well-defined regions in the genome are likely to produce recombinants that are less fit that the parental viruses (Teterina et al. 2006). Recombination breakpoint patterns may well determine which recombinants survive because for a recombinant to be viable it may be important that it does not have disrupted biologically-important secondary structures or does not have any novel structures, which were not originally present in the parental sequences (Martin et al. 2005).

1.4 Investigating the evolutionary impacts of secondary structures within picornavirus genomes

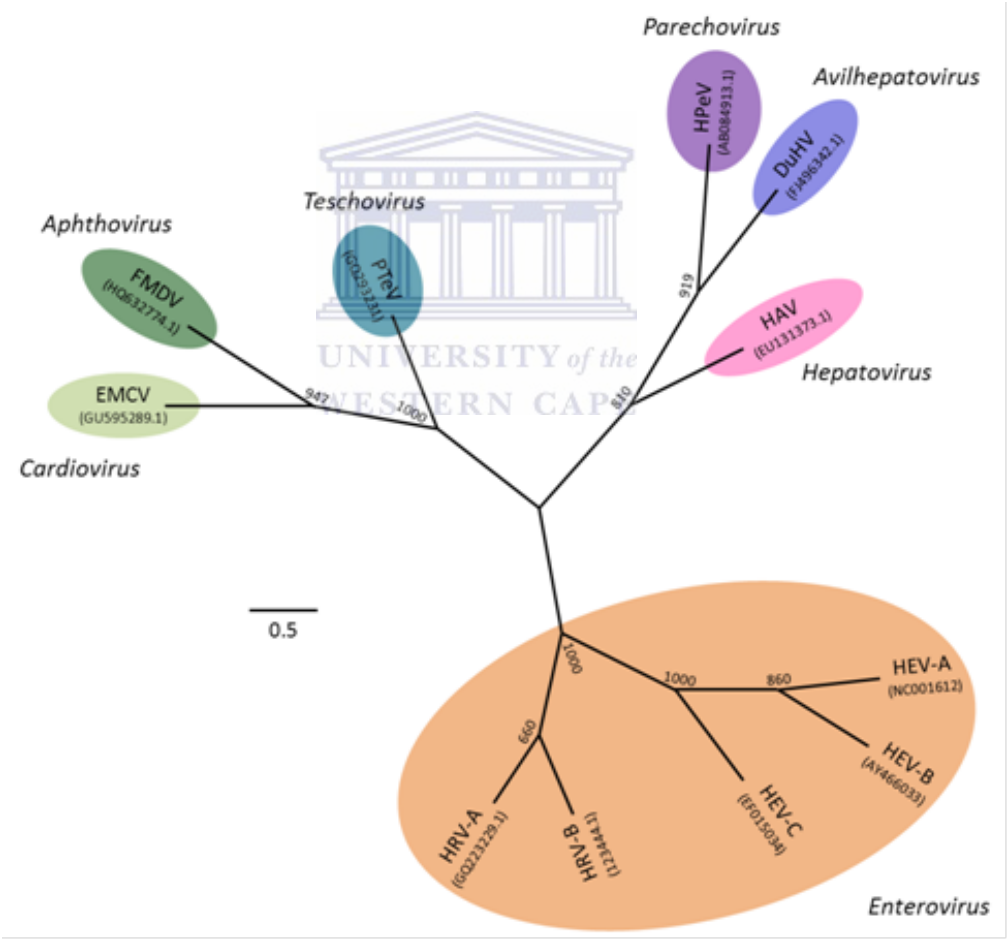


Figure 2. Piconrnaviruses phylogenetic tree Maximum likelihood tree inferred based on Tamura-Nei model of nucleotide substitution, as implemented in MEGA5 (Tamura, et al., 2011), using the polyprotein gene sequence of 11 picornavirus species used in this study [Foot-and-mouth disease virus (FMDV), Encephalomyocarditis virus (EMCV), Porcine teschovirus (PTeV), Human parechovirus (HPeV), Duck hepatitis A virus (DuHV), Hepatitis A virus (HAV), Human Enterovirus A (HEV-A), Human Enterovirus B (HEV-B), Human Enterovirus C (HEV-C), Human rhinovirus A (HRV-A), Human rhinovirus B (HRV-B)]. ICTV defined genera are highlighted by coloured ovals. Numbers at branch points provide support values from 1000 non-parametric bootstraps. The scale bar represents 0.5 nucleotide substitutions per site.

Picornaviruses rank among the smallest of all RNA viruses with single-stranded +sense RNA genomes between 7000 and 8000 nucleotides long. The most prominent species within this family include Rhinovirus, Poliovirus, Human hepatitis A virus, and Foot-and-mouth disease virus. All picornaviruses share a common genome arrangement with their genomes being partitioned into three discrete regions; the 5' UTR, the polyprotein open reading frame (ORF) and the 3'UTR. Whereas the 3' genome termini of picornaviruses are polyadelyated their 5' ends are covalently attached to a small virus encoded protein called VPg (virion protein, genome). The polyprotein ORF is directly translated from the genome and is both co-translationally and post-translationally cleaved by viral encoded proteases to produce up to four structural proteins (1A, 1B, 1C and 1D which collectively make up the viral capsid) and up to eight non-structural proteins (L, 2A, 2B, 2C, 3A, 3B, 3C and 3D; Figure 3).

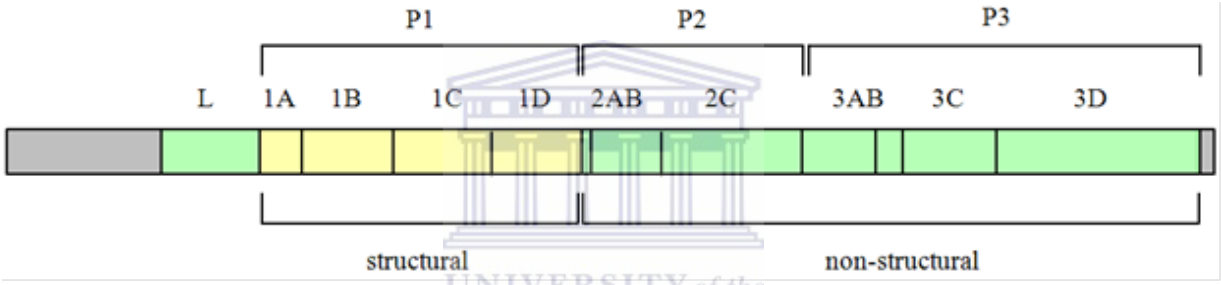


Figure 3. Genome organisation of a typical picornavirus genome Most members of this family encode four capsid proteins (P1) followed by the non-structural proteins (P2 and P3). The species of some picornavirus genera have a non-structural Leader (L) protein upstream of P1. The grey areas on the extreme left and right sides of the schematic, represent 5' UTR and 3' UTR respectively. The genome organisation shown according to the L-P1-P2-P3 structural scheme (van Regenmortel et al., 2000). Gene regions are drawn to scale with respect to *Aphthovirus*, Foot-and-mouth disease virus, isolate GU931682.1

Picornavirus genomes accumulate between 10^{-5} to 10^{-3} mutations per nucleotide per replication cycle and as a result, are capable of very high evolution rates (Jenkins et al. 2002). Despite this some regions of picornavirus genomes are highly conserved. These include sequences either comprising functional nucleotide sequence motifs or encoding functionally-important protein domains. Conserved genome regions also include sequences that fold into functionally important secondary structures (Pilipenko et al. 1989, Witwer et al. 2001). In fact, the need to preserve functionally important genomic regions in the face of extremely high mutation rates may at least in part explain why picornaviruses have such small genomes: put simply, at a given mutation rate per replication cycle smaller genomes have a lower chance of mutating than bigger genomes. For example, HAV, with a genome size of ~8000

nucleotides and a mutation rate of 10^{-3} - 10^{-4} mutations per site per replication cycle (Moratorio et al. 2007), will only undergo 10 mutation events every 7 replication cycles.

So far, a number of biologically-functional secondary structures in picornaviruses have been determined that play a role during the various stages of the viral life-cycle. The best studied of these are the structural elements that reside within a ~450 long region of the 5' UTR and constitute what is known as the internal ribosome entry site (IRES; (Pelletier and Sonenberg 1988). This highly structured region is critical for the initiation of translation. For example, the loop regions of a particular stem loop structure (designated K in EMCV) binding the host protein complex that recognises AUG start codons (called eukaryotic initiation factor 2/2B or eIF-2/2B)(Duke et al. 1992). More recently, the structure of another, 5 'UTR structure, the 33-nt cis-acting replication element (*cre*) stem-loop structure of human rhinovirus 14 was determined by nuclear magnetic resonance (NMR) in solution conditions optimal for uridylation. They showed four nucleotides within its stem region were crucial for VPg uridylation and viral RNA replication (Thiviyanathan et al. 2004).

Given that evidence already exists of some biologically functional secondary structures within picornavirus genomes, the species in this family are excellent candidates for analysing the impact on virus evolution of secondary structures on a genome-wide scale. Another feature of the picornavirus family that makes it useful for analysing the evolutionary impact of genomic secondary structures is that large numbers of full genome sequences are available within public sequence databases. Also crucial in this regard is the fact that these sequences are diverse enough to ensure that the various tests we intended to employ would have sufficient power to detect the alterations in evolutionary dynamics that we expected to be associated with presence of secondary structures.

The objective of this project was to carry out analyses to identify secondary structures within these virus genomes that have the highest probability of being biologically relevant. Besides identifying the most evolutionarily conserved secondary structures we used additional analyses to identify (1) evidence of purifying selection pressures on synonymous sites within protein coding regions - such pressures are expected when secondary structures occur within protein coding regions; (2) evidence that sites predicted to be paired within secondary structures are co-evolving; and (3) evidence that recombination that naturally occurs between virus genomes has tended to preserve the secondary structures more than would be expected if recombination events were distributed randomly across viral genomes. A range of

statistical approaches were used to test for evidence and quantification of the relationship between these sites and the test for associations of secondary structures with selection on synonymous sites, site coevolution and preservation of structural base-pairing within recombinants – tests that. Besides attempting to provide evidence for the existence of predicted structures, we were also able to order them, using the data generated, indicating their evolutionary, and therefore, also, their biological significance.

2. MATERIALS AND METHODS

2.1 Data preparation

Picornavirus genomic sequences were retrieved from the National Center for Biotechnology Information (NCBI) nucleotide database (http://www.ncbi.nlm.nih.gov/nuccore) during April, 2011. The sequence length query used, covered at least 75% of the total genome size, therefore including submissions of polyprotein genes lacking the 5' UTR and polyA-tail regions. The sequence sets for each picornavirus species were separately aligned using MUSCLE (Edgar 2004) and the resulting alignments were further edited manually, using Se-Al v2.0a (http://tree.bio.ed.ac.uk/software/seal/).

To create datasets for analysis, groups of sequences were selected from the alignments in which the most divergent pair of sequences were no less than 75% similar. By using this minimum degree of similarity, we were able to preserve the alignment accuracy while ensuring sufficient signal for our downstream analyses. This process yielded 11 large datasets, each containing between 46 and 313 full genome sequences (Table 1).

Table 1. Virus sequence alignments analysed in the current study

Virus Genus	Dataset name	No.[a]	Length (nt)[b]	Divergence[c]
Aphthovirus	Foot-and-mouth disease virus (FMDV)	313	8769	0.230
Enterovirus	Human enterovirus A (HEV-A)	231	7494	0.247
	Human enterovirus B (HEV-B)	150	7619	0.244
	Human enterovirus C (HEV-C)	289	7737	0.277
Rhinovirus	Human rhinovirus A (HRV-A)	92	7266	0.312
	Human rhinovirus B (HRV-B)	44	7317	0.298
Hepatovirus	Hepatitis A virus (HAV)	50	7558	0.266
Cardiovirus	Encepahlomyocarditis virus (EMCV)	64	8586	0.225
Teschovirus	Porcine teschovirus (PTeV)	49	7179	0.264
Parechovirus	Human parechovirus (HPeV)	52	7345	0.246
Avihepatovirus	Duck hepatitis A virus (DuHV)	46	7881	0.292

[a] Number of sequences in the dataset

[b] Length of the aligned sequences in each dataset (i.e. the alignment length including gaps)

[c] Mean pair-wise Jukes-Cantor corrected distance within group of complete genome sequences

From each of these large datasets, the 30 most distantly related sequences were selected, to form intermediate-sized datasets. Each of these contained a group of sequences that were representative of the entire breadth of diversity evident in the large datasets. These intermediate datasets were subsequently used for the analysis of purifying selection acting at synonymous sites (section 2.3.1 below).

From each of the intermediate-sized datasets, ten representatives of the ten most divergent sequence lineages were selected to form a set of small datasets. These small datasets were used both for the computational prediction of genome-wide secondary structures and the identification of evolutionarily conserved structural elements (section 2.2 below).

2.2 Computational prediction of genome wide secondary structure

The complete genomic secondary structure of each virus was obtained using Nucleic Acid Structure Predictor (NASP; (Semegni et al. 2011). NASP identifies secondary structures which may be evolutionary conserved with the lowest false positive rate possible. The NASP algorithm uses minimum free energy (MFE) estimates, provided by the UNAFOLD nucleic acid folding program, hybrid-ss (Markham and Zuker 2008), to predict secondary structures and generate a consensus base-pairing matrix – called the M matrix. NASP scans through M

to progressively identify the most evolutionary conserved base-paired nucleotides. The consensus matrix provides the most evolutionarily conserved structure for the whole alignment, consecutive non-zero values in the anti-diagonal of M, show positions of base-pairing and can be summed to yield a "conservation score" for a discrete sub-structure (such as a stem) within the entire folded molecule. The output of this algorithm provides CT files for the structure and estimates of the minimum free energy of the structure. Determination of whether significant unaccounted for structure remains within the sequences involved shuffling the original sequences 100 times, folding each and obtained 100 minimum free energy scores: the probability of unaccounted for structure remaining in the alignment is equivalent to the proportion of shuffled sequences with a higher MFE than that of the real sequence. The consensus p-value (combined p-value) was calculated from all p-values of simulated sequences and if it was less than 0.05, the structure with the highest score was fixed and the remaining positions shuffled (avoiding the fixed stem). The process was repeated several times to identify other structures until a p-value > 0.05 was obtained. The NASP analysis were set up to treat sequences as linear RNA, an annealing temperature of 37^{o}C with sodium and magnesium concentrations set 1M and 0M, respectively. The analysis was performed separately, using each of the small datasets.

2.3 Testing whether base-paired nucleotides in coding regions tend to occur in codons with lower than expected synonymous substitution rates

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 anticipated that these two distinct layers of selection would be reflected in codon sites that contain nucleotides that participate in base pairing interactions within biologically important secondary structures, having lower synonymous substitution rates compared to codon sites that contain nucleotides that do not participate in base-pairing interactions.

2.3.1 Estimation of synonymous substitution rates across coding region

This analysis aims at identifying highly-conserved nucleotide sites within coding regions. An aim was to calculate nucleotide synonymous substitution rates and identify sites with lower-

than-expected synonymous substitution rates. These are assumed to be under strong purifying selection pressured at the nucleotide level and are probably conserved for a biological purpose – perhaps due to a need to preserve base-pairing within secondary structures.

This analyses was performed on codon re-aligned (using MUSCLE) coding region sequences extracted from all of the intermediate sized datasets (the datasets containing 30 sequences representative of the diversity found in the large dataset). Recombination breakpoints within these coding region sequences were inferred using the GARD method (A Genetic Algorithm for Recombination Detection; (Kosakovsky Pond et al. 2006). GARD outputs a separate tree topology for each partition of the alignment. Each partition and tree was then used in the PARtitioning approach for Robust Inference of Selection (PARRIS; (Scheffler et al. 2006) analysis in order to infer synonymous substitution rates at each codon site.

In this study, PARRIS uses 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. 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, an individual tree topology and branch lengths were used in order to avoid the false inference of synonymous substitution in datasets displaying evidence of recombination. Generating trees from datasets in which recombination is present may lead to misleading branch lengths and tree topologies (Anisimova et al. 2003).

2.3.2 Testing for associations between synonymous substitution rates and secondary structure

It was possible to categorise the codon sites into more-constrained (ds < 1) or less-constrained (ds > 1) categories based on their synonymous substitution rates. The synonymous substitution rate of 1 was chosen as the cut-off, based on the average number of synonymous substitutions per codon across the coding region of the genomes. Codon sites were further classified based on the paired number of nucleotides they contained, using the paired site co-ordinates predicted by NASP. In order to test for association between the

degree of constraint on synonymous substitution and the number of paired nucleotides within codons, a chi-squared test was used.

2.4 Analysis of co-evolving nucleotide sites

By identifying co-evolving nucleotide sites within an alignment of RNA sequences, it will potentially be possible to directly detect evolutionary (and hence probably also functional), constraints on the evolution of sequences within genome regions that display secondary structure.

2.4.1 Testing whether base-paired nucleotides tend to coevolve with one another

The SPIDERMONKEY HYPHY script (Poon et al. 2008) was used to detect whether pairs of sites within the alignment are evolving in a way that they are constrained to form a stem structure. 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 investigate for evidence of altered nucleotide substitution patterns between paired and unpaired regions. 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 were multiplied by $1/\lambda$. 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 1in the Muse-modified model, it is possible to perform a likelihood ratio test (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 displaying evidence of complementary coevolution for the pairs of sites examined (Muse 1995).

2.4.2 Improvement of coevolution 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, it had to be accounted for in our method. When searching for possible sites which may be evolving in lock-step it will regard nucleotide changes from exchanged fragments as though they occurred during the same evolutionary event, thus detecting false signal. Recombination within picornavirus genomes is well documented (King et al. 1985, Lukashev 2010, Simmonds 2006) and therefore prior to the co-evolution analysis the full datasets were analysed for evidence of recombination events. Recombination detection was performed on every large dataset, using the Recombination Detection Program 4.16 (RDP; (Martin et al. 2010), identifying recombinants and their corresponding parental sequences. All of the recombinant sequences were split at the breakpoints identified by RDP and added to the rest of the alignment file as separate sequences, creating a recombination-free alignment. The resulting alignment contained all of the original dataset sequences with all of the separated sequences appended, creating an expanded dataset. A 125nt sliding window was then moved 1nt at a time, across this expanded alignment, selecting the N longest sequences from each window, where the length is interpreted as the least number of gap characters contained. Separate alignment files were created for the first 125 nucleotides from each sequence in each window. 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). The SPIDERMONKEY algorithm was then executed in the HYPHY environment for each separated alignment file and corresponding tree.

2.4.3 Testing for associations between coevolving sites and those predicted to be paired within secondary structures

The SPIDERMONKEY results were placed into a coevolution matrix in which every element represented a nucleotide pair considered for evidence of co-variation. This matrix was then compared to the consensus pairing matrix, obtained from the NASP analysis, by combining

the matrices to produce a reference matrix. It was then possible to classify each element in the reference matrix into one of four categories (paired and coevolving, paired and not coevolving, unpaired and coevolving, unpaired and not co-evolving; Figure 4) so that a chi-squared test could be used to test for evidence that sites are co-evolving in such a manner that secondary structures are maintained. If a nucleotide pair associated p-value in the coevolution matrix was < 0.05, it was considered to be coevolving, otherwise it was regarded as not coevolving. In the pairing matrix, those elements with an entry greater than 0 were deemed to be involved in base-pairing, else, they were regarded as unpaired sites.

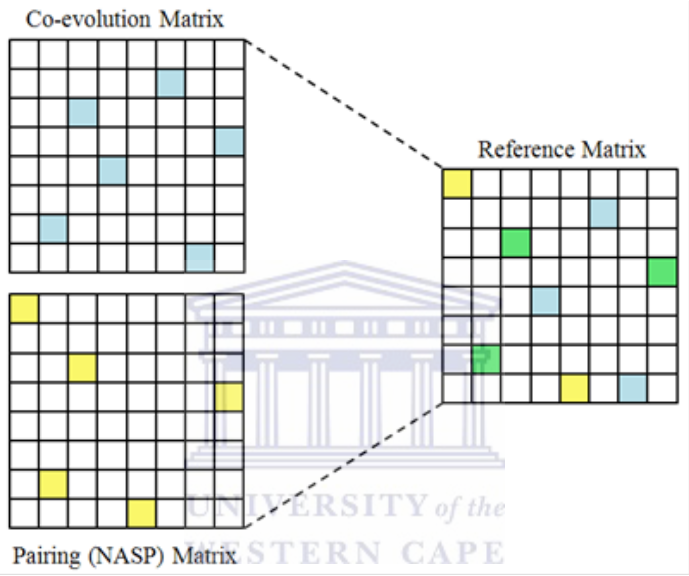


Figure 4. Testing degree of association between paired and co-evolving sites Cells shaded in green represent predicted paired nucleotide , which also show evidence of co-evolution

2.5 Testing whether base-paired sequences evolve as though double-stranded

It is plausible that sites that are paired within secondary structures might evolve in different way to sites that are unpaired. For example, sites that are unpaired may be sensitive to different mutagenic pressures than sites that are paired. Also, paired sites that are evolving under selective pressures favouring the maintenance of base-pairing may accumulate patterns of mutation that are more similar to those expected for double-stranded RNA than those expected for single-stranded RNA. Specifically, double-stranded RNA is expected to evolve in a way that is best described by a six-rate non-reversible nucleotide substitution model where rates of complementary substitutions (for example G to A and C to U) are similar to one another. The reason for this is that nucleotide sequence specific mutagenic processes

(i.e. processes that target the four different nucleotides differentially) will target both strands of double-stranded RNA such that, for example, a G to A substitution at a particular site on one strand will be reflected by a C to U substitution at that site on the opposite strand. Similarly, C to U substitutions on one strand will be reflected by G to A substitutions on the other. Therefore, regardless of how different the mutagenic processes responsible for changing G's to A's are from those responsible for changing C's to U's, in double-stranded RNA these rates should be conflated and indistinguishable from one another. Conversely, in single-stranded RNA the rates of G to A and C to U substitution should accurately reflect the characteristics of the processes that yield such mutations, and should therefore be free to occur at rates that are independent of one another. Therefore, whereas the evolution of double-stranded RNA might perhaps be best described by a six-rate non-reversible complementary nucleotide substitution model (i.e one where G to A substitutions are constrained to occur at the same rate as C to U substitutions), the evolution of single stranded RNA might be best described by a non-reversible 12-rate nucleotide substitution model (where all 12-possible nucleotide substitutions are free to occur at different rates) (Knies et al. 2008).

To test which of these two evolutionary models best described the evolution of base-paired and non-base paired sites within picornavirus genomes, was split each of the large dataset alignments into two parts: One part, called the "paired alignment", containing alignment columns corresponding with sites identified by NASP as being those with the highest probability of being paired (i.e. those identified with the NASP 0.05 permutation p-value cut-off) and the other, called the "unpaired alignment" containing the remaining alignment columns from the large dataset alignment. It is important to stress that the NASP 0.05 p-value cut-off used to split the large dataset alignment does not indicate that there is a >95% chance that sites in the paired alignment being paired. It instead indicates that there is a >5% probability that the set of nucleotides in each sequence of the "unpaired" alignment formed no substantially-conserved secondary structures. Therefore, it is very probable that whereas some of the genomic sites included in the "paired alignment" are not actually paired in all of the sequences in this alignment, many of the sites included in the "unpaired" alignment are likely paired in at least some of the sequences in this alignment.

The HYPHY script used to perform the nucleotide substitution model tests on each of the paired and unpaired alignments took as input a sequence alignment and a rooted phylogenetic

tree describing the evolutionary relationships of the sequences in the alignment. In order to construct a rooted phylogenetic tree for each large dataset an outgroup sequence, representing the nearest relative of the sequences in the dataset that was excluded from the dataset because it shared <75% identity with the sequences in the dataset, was aligned to the sequences in the large datasets, PHYML ver. 3.0 was then used (with the HKY nucleotide substitution model) to infer a maximum likelihood tree for each of the large dataset plus outlier alignments. The trees were then rooted on the outlier sequences, after which the outlier sequence branch was removed. Because the standard-six rate and general time reversible six-rate models are special cases of the non-reversible 12-rate model of nucleotide substitution it is possible to perform a LRT to directly compare the fit of the models. The p-value obtained can be used to decide whether the six rate models can be rejected in favour of the 12 rate model. The test also produces Akaike information criterion (AIC; Akaike, 1974) scores for the rates and parameters for each model. The absolute AIC score is not meaningful on its own but the differences in scores between models can be used as a guideline. The preferred model would be the one with the lowest AIC scores. Models which would provide insight into the evolutionary rates of these viruses were chosen, however, the best-fit model is not necessarily a good model, it is simply the best out of the selected models.

2.6 Ranking of structures

Although NASP provides ranking for the most evolutionarily-conserved predicted structures we performed ranking of the predicted secondary structures according to their support based on the degree of synonymous substitution constraint and complementary-evolution, as well as the consensus of all three (NASP, PARRIS, SPIDERMONKEY) analysis approaches.

2.6.1 Ranking based on constraints on synonymous substitution rates

Predicted structures were ranked based on the degree of synonymous substitution constraint at each codon site in the structure of interest, using the Wilcoxon rank-sum test implemented in DOOSS 1.0 (Golden and Martin 2013). 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 a structure, hence avoiding bias towards smaller structures consisting of codons with low substitution rates.

The p-value obtained from the Wilcoxon rank-sum test, was used to rank the structures, where (1) a low p-value and a corresponding negative z-score indicated a structure which contains significantly more low dS codons than expected and (2) a low p-value and a corresponding positive z-score is indicative of unconstrained codons, or higher dS scores than were expected by chance.

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 the likely biological relevance.

2.6.2 Ranking of structures based on degrees of complementary coevolution

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 the 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 2.6.1 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).

2.6.3 Consensus ranking

Consensus ranking was achieved by mapping the scores in the three scoring categories (degrees of conservation indicated by NASP, synonymous substitution rates determined by PARRIS 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 3 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 three criteria is that the contribution of the various tests is largely uneven which could influence the consensus ranking unfairly.

2.7 Recombination detection analysis

RDP 4.16 was used to identify and characterise individual recombination events evident within the different large picornavirus datasets. While a number of other programs have been written to carry out these tasks (Drouin et al. 1999, Posada and Crandall 2001), RDP is a single highly automated analysis tool that simultaneously uses a range of different recombination detection methods to both detect and characterise the recombination events that are evident within a sequence alignment without any prior user indication of a non-recombinant set of reference sequences.

2.7.1 Breakpoint distribution analysis

In order to visualise the distribution of breakpoint positions evident within the various picornavirus genomes analysed here, a breakpoint map for each large dataset was compiled, containing the positions of all clearly-identified breakpoints for every unambiguously unique recombination event. Breakpoint density plots were then constructed from these maps (Figure 9), by using a sliding window of 200 nucleotides moving one nucleotide at a time and counting the number of detected breakpoints at each frame, plotting that number at the central window position. In order to determine whether the breakpoint clustering within each window was statistically significant, a permutation test was used. Globally significant breakpoint clusters were considered as those windows within the breakpoint density plot that contained more breakpoint positions than the maximum found in more than 95% of the 8000 permuted breakpoint density plots. Locally significant breakpoint clusters are identified as those windows within the plot that contained more breakpoint positions than more than 99% of windows at the identical location in the permuted density plots. This 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 genome regions where sequences are less diverse. 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).

2.7.2 Permutation test of association between recombination breakpoint clustering and the locations of secondary structures

There is growing evidence that favoured recombination breakpoint positions observable in many single-stranded viral genomes are influenced by nucleotide base pairing within their thermodynamically most favourable folded structures (Draghici and Varrelmann 2010, Simon-Loriere et al. 2010). The consensus fold matrix (M matrix) from NASP was used to compile a list of structured and unstructured regions with relative support for each site. Sites regarded as being structured were those that had a cumulative NASP score greater than 0 and those that were predicted to be unpaired (NASP score 0) were classified as unstructured. The test aimed to determine whether recombination breakpoints are significantly more or less clustered within structured or unstructured genome sites. The observed breakpoint

distributions were compared with breakpoint distributions determined for 8000 permuted datasets (constructed as described above).

2.7.3 Test comparing recombination breakpoint densities between different gene regions

In order to test for clustering of recombination breakpoints in different genome regions, we mapped gene boundary positions to the large dataset alignments and used the same permutation test described above to determine whether: (1) there were more or fewer breakpoints detectable in the intergenic regions than within gene sites than can be accounted for by chance; (2) individual genes contained significantly more/fewer breakpoints than the rest of the genes present; (3) the beginning and ending 25%, 12.5%, and 5% of all genes, contained significantly more/fewer detectable breakpoints than those collectively observed in the remaining middle section of the genes.

2.7.4 Secondary structure disruption test

RDP provides statistical evidence for the occurrence of recombination events, identifies likely parental sequences, and estimates the positions of recombination break points. Using this information, string concatenation methods were used to construct the recombinant sequences by joining the regions of the genome donated by the major parent and the part originating from the minor parent.

Secondary structure disruption test was designed to determine whether recombinant picornavirus genomes arising in nature tended to have lower degrees of predicted fold disruption than randomly generated recombinants. The test involved using parental sequences identified by RDP to reconstruct a series of simulated recombinant sequences corresponding to each of the recombination events detected by RDP. For each detected recombination event 31 recombinants were simulated and secondary structures were predicted using UNAFOLD (as described above) for the parental, simulated recombinant sequences. Whereas one of these recombinants had the exact same breakpoint positions as the real recombinant identified by RDP, 30 others had randomly located 5' breakpoint position and a 3' breakpoint position exactly the same number of variable nucleotide positions downstream of the 5' breakpoint as in the simulated recombinant with breakpoints in the same position as in the real recombinant. The secondary structure predictions indicated two categories of sites: (1) base paired sites in parental sequences that were missing in the simulated recombinant sequences

and (2) base paired sites absent in both parental sequences that were present in the simulated recombinant sequences (Figure 5). For each of the simulated recombinants the numbers of category (1) and (2) sites were used in a permutation test to determine whether across each of the large picornavirus datasets recombinants simulated with real breakpoint positions tended to have fewer sites in categories (1) and (2) than those simulated with randomised breakpoint positions.

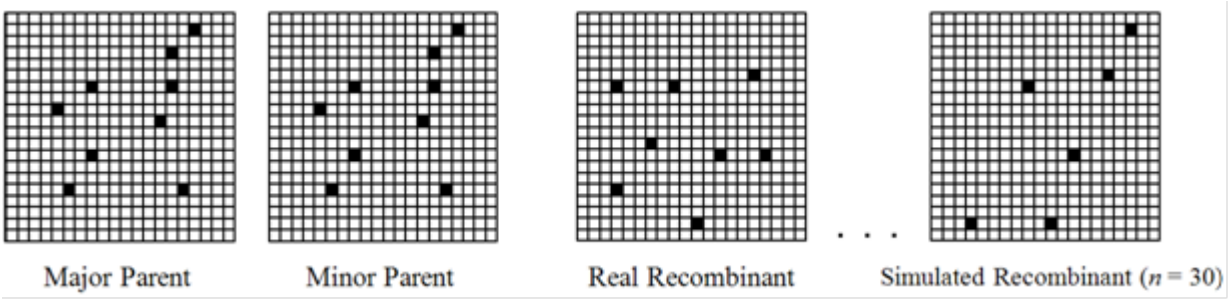


Figure 5. Pairing probability matrices for parental and recombinant sequences Shaded blocks represent paired nucleotides within the genomes.

Specifically, for each of the large datasets, the total number of category (1) and (2) sites are separately summed respectively obtaining a disruption score (1) and (2) for all of the recombinants simulated with real breakpoint positions. For each of 1000 permuted datasets, disruption scores (1) and (2) were calculated by summing the total numbers of category (1) and (2) sites in simulated recombinants with randomised breakpoint positions where one each of these recombinants is randomly selected from the 30 simulated for each real recombination event. The total number of times that the disruption scores (1) and (2) determined for recombinants simulated with the real breakpoints were higher than those computed for the permuted datasets divided by the number of permuted datasets (in this case 1000) represented the probability that the real recombinants had an estimated degree of single stranded RNA folding disruption that was not lower than that expected if the survival of recombinants was in no way influenced by the effects of recombination on secondary structure. Whereas, the permutation test involving the category (1) sites indicated the influence of base pairs broken by recombination, the test involving the category (2) sites indicated the influence of aberrant base pairing caused by recombination.

3. RESULTS AND DISCUSSION

It is known that at least some regions of some picornavirus genomes form defined secondary structure elements that have important functions during viral replication. Although the sizes and locations of these "cis-active RNA elements" (CREs) vary from one species to the next, at least some of these elements have equivalent functions across the different picornavirus genera (Pilipenko et al. 1989, Simmonds et al. 2008).

3.1 Testing for an association between constraints on synonymous substitution rates and NASP predicted base-pairing Several studies (Reynolds et al. 1995, Tuplin et al. 2004, Yang et al. 2008) have proposed that extensive base pairing might underlie the greater than expected frequency of invariant synonymous sites observed in the genomes of some RNA viruses. Our aim was to investigate whether lower synonymous substitution rates in the coding sequences of picornavirus genomes were also attributable to nucleotide sequence conservation that is driven by evolutionary constraints imposed by biologically-important secondary structures. On a genome-wide scale we only detected obvious associations between lower than expected synonymous substitution rates in the coding regions and sites predicted to be base paired by NASP in three of the eleven datasets: FMDV, HEV-A, and HEV-C (Table 2), whereas, HEV-B shows marginally insignificant association.

Table 2. Table representing statistical support for association between amount of paired sites within codons of the coding region and lower than expected synonymous substitution rates

Dataset	*n*[a]	Diversity (%)[b]	p-value
FMDV	30	77	**0.05**
HEV-A	30	75	**0.02**
HEV-B	30	75	0.06
HEV-C	30	74	**0.04**
HRV-A	30	70	0.32
HRV-B	30	70	0.44
EMCV	30	71	0.41
HPeV	30	72	0.39
PTeV	30	73	0.08
HAV	30	72	0.33

[a] Number of sequences

[b] Percentage similarity between the two most distantly related sequences in the dataset

However, discrete regions which were identified as having decreased synonymous substitution rates coincided with experimentally determined/computationally predicted CREs identified in previous studies (Figure 6). The loop regions of these structures in the Aphthoviruses, Cardioviruses, and Hepatoviruses display a variety of sequences and sizes but nonetheless contain an AAAC motif that is characteristic of all picornavirus CREs (Steil and Barton 2009), and is essential for their functioning. The stem regions of these structures also differed in length, containing various internal loops and bulges, showing the functional importance of the loop region with its conserved adenosine residues (as opposed to the variable composition of the rest of the structure).

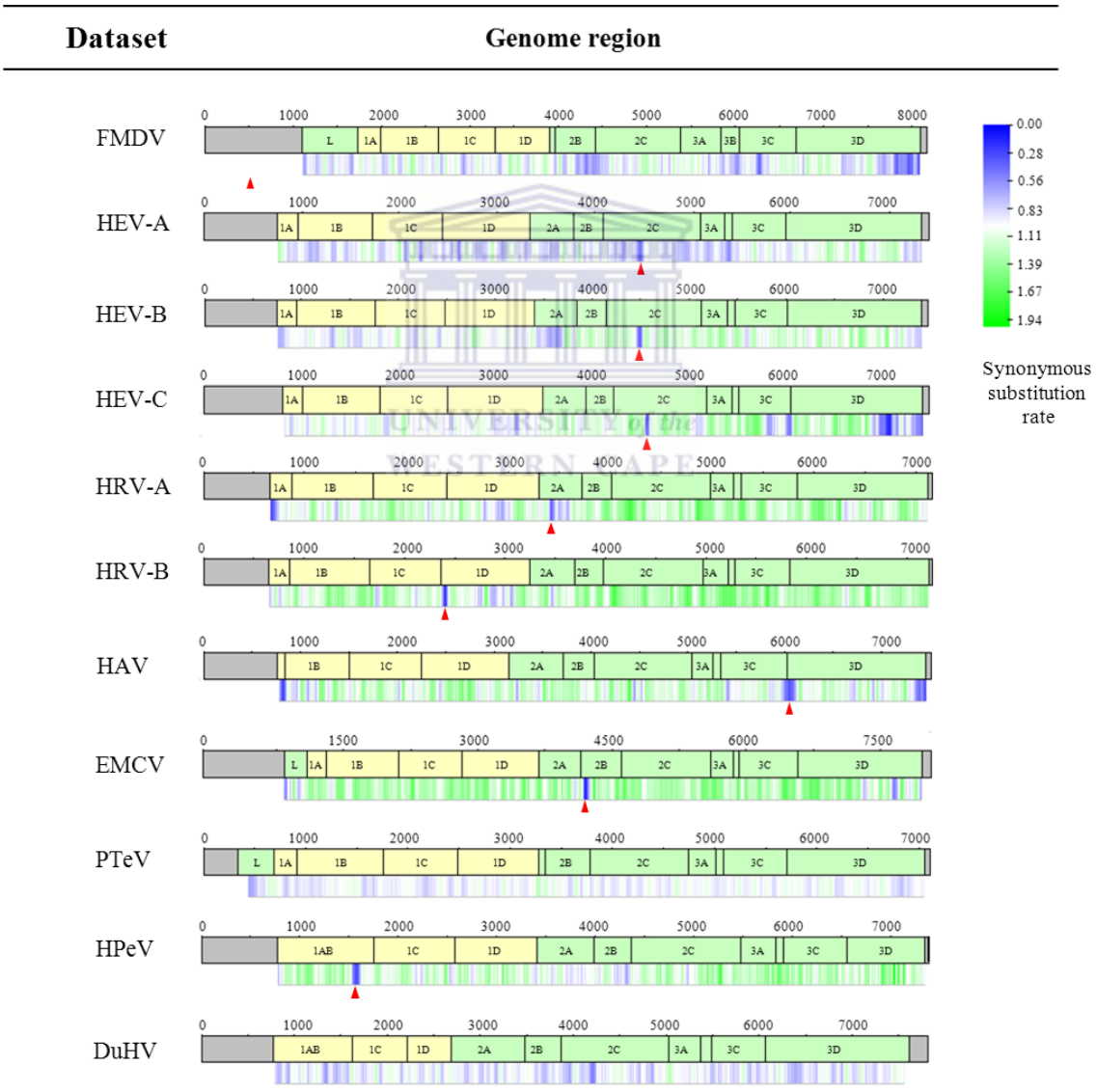


Figure 6. Estimated synonymous substitution rates across coding region of members of Picornaviridae Gene regions shown to scale with respect to their own genus. Nucleotide positions are shown above the gene maps; while the vertical lines below each map indicate site-to site variations in synonymous substitution rates (see colour key). The red arrows indicate the approximate position of previously identified CREs (note no CREs have yet been identified for PTeV and DuHV). The grey areas on the extreme left and right sides of the drawing, represent 5' UTR and 3' UTR respectively, whereas the yellow segments represent the structural coding region and the green areas represent the non-structural regions.

Decreased synonymous substitution rates were also detected within the highly conserved nucleotide sequence motif, AACCCTGGGCCC, found in the aphtho-, cardio-, tescho- and avihepatoviruses. This motif is found in the 2A gene where, during translation, it induces co-translational cleavage of the polyprotein by a process known as ribosomal "skipping". This process involves inhibition of the peptidyl which, by preventing the ribosome from linking a new inserted amino acid to the nascent amino acid chain, causes release of the chain. Translation is then restarted on a proline residue, initiating translation and elongation of the new amino acid sequence.

3.2 Testing whether base-paired nucleotides tend to coevolve with one another

There are well-known examples of coevolution between sites that form biologically-functional RNA secondary structures within picornavirus genomes (Fernandez et al. 2011) (Fernandez-Miragall et al. 2006, Martinez-Salas and Fernandez-Miragall 2004). Besides complementary co-evolution between pairs of nucleotides that are base-paired within these structures (i.e. where an A to C change at one site is coupled with a T to G change at another site), other sites that interact more distantly to maintain the conformation of structurally important regions of the viral genome also co-evolve (albeit not necessarily in a complementary fashion). Complementary coevolution between sites was detected using SPIDERMONKEY, and tested for associations between these and sites which were predicted by NASP to be base-paired. Such an association was detected in only four of the eleven analysed datasets (FMDV, HEV-B, -C and EMCV; Table 3). Some biologically-significant coevolutionary interactions in picornavirus genomes are not between nucleotides paired within secondary structures (Fernandez et al., 2011) and we also tested whether nucleotides which did not participate base-pairing interactions, showed any evidence of non-complementary coevolution. Only one dataset (PTeV) showed a marginally significant association between sites that were non-complementarily co-evolving and sites that were part of unstructured genome region.

Table 3. Statistical support for associations between nucleotide base-pairing and complementary coevolution

Dataset	Paired site coevolution p-value	Unpaired site coevolution p-value
FMDV	**0.032**	0.483
HEV-A	0.413	0.751
HEV-B	**0.001**	0.695
HEV-C	**0.001**	0.148
HRV-A	0.265	0.569
HRV-B	0.388	0.259
EMCV	**0.022**	0.891
HPeV	0.621	0.882
PTeV	0.819	0.063
HAV	0.668	0.175
DuHV	0.749	0.512

The association on a genome-wide scale between base-pairing and complementary coevolution that was detected in some of the picornavirus datasets, suggests that a substantial number of the predicted base-pairs in these datasets form part of biologically-important secondary structures that have likely been purposefully preserved by natural selection. It indicates that when structurally disruptive mutations occur these are compensated for by complementary mutations that restore base-pairing. Similarly, selection may favour the mis-pairing of certain nucleotides in order to maintain the overall shape of some biologically relevant secondary structures (Shepherd et al. 2006). These co-evolution signals should, however, be much harder to detect since there are many more potential mis-pairing interactions than there are potential base-pairing interactions (Shang et al. 2012). Another reason that the datasets vary with respect to the apparent association between variation in the results we observed, could be partially caused by the underlying structural features of the genome-wide structure, as base-pairings adjoining internal and apical loops, or bulges have been shown to be under different selective constraints when compared to internal base-pairings (Tian et al. 2008).

The fact that three of the four datasets for which we detected significant associations between predicted base pairing and complementary coevolution also happen to be the largest datasets that we examined (all have >150 sequences; Table 1) may also be significant. It indicates that it is likely that our genome-wide association test likely lacked power when applied to the seven datasets with 100 or fewer sequences (HRV-A, HRV-B, EMCV, HPeV, PTeV, HAV, and DuHV. Therefore in section 3.3.1 we employ a more focused approach to test for these

associations within individual NASP predicted genomic sub-structures and use the results of these tests to rank these sub-structures in order of their likely biological importance.

3.3 Test whether paired sites evolve as though double-stranded

Nucleotide substitutions in the base-paired regions of the viral genome might only be selectively tolerable if they are coupled with complementary compensatory mutations that restore base pairing. While this might result in nucleotide substitutions becoming fixed at a lower frequency in structured genome regions than in non-structured regions (Simmonds and Smith 1999), it is also plausible that the patterns of substitutions that do eventually become fixed within the structured regions might be better reflected by a six-rate complementary nucleotide substitution model such as 6NREV than either a reversible six-rate model such as GTR or a 12-rate non-reversible model such as 12NREV. Alternatively 12NREV, might provide the best description for the evolution of single stranded genome regions that are predicted to not be involved in base-pairing

Our likelihood ratio tests of model fit, however, showed that regardless of whether sites were predicted to be paired or unpaired the most general 12NREV model fit the data significantly better ($P < 0.05$) than either of the 'nested' six-rate models (Table 4). Using the AIC score to indicate which of the six-rate models fit the data best indicated that for the genome regions predicted to be paired the 6NREV model only fitted the data better than GTR in only two instances (the HRV-A and HRV-B). However, in the same test of sites predicted to be unpaired the 6NREV model was also only predicted to be a better fit than GTR in with these same two datasets.

Table 4. Six versus twelve rate nucleotide substitution model LRT p-values and AIC ranking

Dataset	Paired			Unpaired		
	LRT			LRT		
	GTR vs 12NREV	6NREV vs 12NREV	AIC: GTR vs 6NREV	GTR vs 12NREV	6NREV vs 12NREV	Lowest AIC: GTR vs 6NREV
FMDV	$<1.0x10^{-16}$	$<1.0x10^{-16}$	GTR	$<1.0x10^{-16}$	$<1.0x10^{-16}$	GTR
HEV-A	0.005	$<1.0x10^{-16}$	GTR	$1.11x10^{-16}$	$<1.0x10^{-16}$	GTR
HEV-B	$<1.0x10^{-16}$	$<1.0x10^{-16}$	GTR	$<1.0x10^{-16}$	$<1.0x10^{-16}$	GTR
HEV-C	$<1.0x10^{-16}$	$<1.0x10^{-16}$	GTR	$<1.0x10^{-16}$	$<1.0x10^{-16}$	GTR
HRV-A	$<1.0x10^{-16}$	$<1.0x10^{-16}$	6NREV	$<1.0x10^{-16}$	$<1.0x10^{-16}$	6NREV
HRV-B	$<1.0x10^{-16}$	$<1.0x10^{-16}$	6NREV	$<1.0x10^{-16}$	$<1.0x10^{-16}$	6NREV
EMCV	$3.23x10^{-10}$	$<1.0x10^{-16}$	GTR	$5.75x10^{-4}$	$<1.0x10^{-16}$	GTR
HPeV	$<1.0x10^{-16}$	$<1.0x10^{-16}$	GTR	$<1.0x10^{-16}$	$<1.0x10^{-16}$	GTR
PTeV	$7.52x10^{-6}$	$<1.0x10^{-16}$	GTR	0.021	$<1.0x10^{-16}$	GTR
HAV	$2.22x10^{-16}$	$<1.0x10^{-16}$	GTR	$4.54x10^{-7}$	$<1.0x10^{-16}$	GTR
DuHV	$<1.0x10^{-16}$	$<1.0x10^{-16}$	GTR	$2.80x10^{-9}$	$<1.0x10^{-16}$	GTR

The overwhelming support for 12NREV over the two six-rate substitution models is, unsurprisingly, strongly indicative of the fact that picornaviruses have single stranded RNA genomes. The fact that the evolution of sites in these genomes that are predicted to be base-paired is also best described by 12NREV implies that complementary compensatory substitutions favouring the maintenance of secondary structure have not been pervasive enough to produce patterns of complementary substitution that favour the application of the 6NREV model over the 12NREV model.

Although we detected evidence that the 6NREV model was a better fit that the GTR model when it came to describing the evolution of sites in the HRV-A and HRV-B genomes that are predicted to be base paired is also not compelling evidence of the paired regions of these viruses are evolving as though they were double stranded. The reason for this is that the evolution of sites in these genomes that are predicted to be unpaired is also better described by the 6NREV model than the GTR model – i.e. the relative fit of these two six-rate substitution models does not appear to reflect differences in the evolution of these two genome regions.

3.4 Ranking of predicted structures

3.4.1 Ranking and identification of individual structures

Given that we detected evidence of pervasive biologically relevant secondary structure within many picornavirus genomes, the detected structures were ranked according to their likely biological relevance based on three criteria: 1) their degree of conservation as indicated by their NASP ranking, 2) the degree of impact on structures within coding regions seemed to have on synonymous substitution rates and 3) the degree to which predicted base-pairs within structures co-evolved with one another. A consensus ranking on all three criteria was used, and revealed that some well-characterised experimentally-determined structures were amongst those with the highest ranks. In the FMDV dataset, for example, eight of the top 30 structures in the consensus rank fell within the 3Dpol gene, corresponding to regions in the gene with decreased synonymous substitution rates.

Three of the top ten FMDV structures in the consensus rank were within the 5' UTR, corresponding with IRES and CRE structural elements essential for *Aphthovirus* replication and gene expression (Bassili et al. 2004, Fernandez-Miragall and Martinez-Salas 2003). Instances where structures predicted here to have important biological functions have been previously suggested or experimentally verified to actually have some function are indicated with an associated reference to such claims in the "Reference" column of the consensus ranking tables for each dataset (Supplementary Table 2).

The highest ranking structures in Enterovirus species (HEV-A, -B and –C) were largely similar to one another, featuring the CREs located in the 2C gene (Steil and Barton, 2008), along with their component sub-structures, within the top ten of the consensus rankings of all three related datasets. The Human rhinovirus A CRE (Rfam ID: RF00220) motif located in 2A ranked seventh in the consensus ranking, and two of the top three structures were well characterised IRES structural elements (Kistler et al. 2007).

In HRV-B the cis-replicating element (HRV14 cre) is first in the consensus ranking of this dataset with predicted base-pairs within the structure displaying an extraordinary degree of complementary coevolution. Curiously, amongst the nucleotides comprising this structure there was also substantial evidence of complementary coevolution both with sites within the structure that were not predicted to be base paired, and with sites located more distantly in the

genome. It has been found that compensatory mutations in the HRV14 CRE appear to partially rescue lethal mutations in the 3' UTR, which may provide an explanation for the long range interactions detected amongst the constituent nucleotides of this structural element (McKnight and Lemon 1998).

Substructural elements of the HAV IRES (RF00228) made up four of the top ten structures in the consensus ranking of this dataset with the HAV CRE ranking 14th. Four more of the top 30 structures in this list are found in the 3C and 3D gene regions and have also previously been reported (Kusov and Gauss-Muller 1997).

The well conserved teschovirus IRES elements T1, T4 and T5 (Witwer et al. 2001) were all amongst the top 20 in the consensus ranking of this dataset (placing 9th, 14th, and 19th respectively). We can only speculate on the function and role of all the high scoring "unknown" structures. Although it is beyond the scope of this study to investigate how and to what extent, these structural elements influence viral processes, it is hoped that the rankings presented here (Supplementary Tables 2 A to H) will encourage other researchers to do so.

3.4.2 Detailed characterisation of a structure within the HEV-C genome

NASP41 (Figure 7) positioned in the 2C gene region has been previously identified using thermodynamic models of secondary structure prediction (Goodfellow et al. 2000; Rfam ID RF00048).

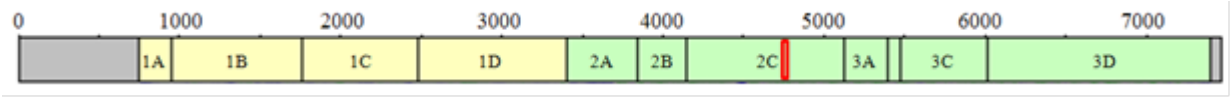


Figure 7. Position of NASP41 in the HEV-C genome Relative position and size of the NASP predicted structure in the HEV-C genome, indicated by the red box

The structure has been experimentally confirmed and its function is well documented (Goodfellow et al. 2000, Paul et al. 2000) however, it has not been considered for investigation by alternate computational methods. NASP41 and the resulting substructures: NASP115 and NASP179, ranked within the top six predicted structures of the synonymous substitution rate ranking. Additionally, NASP41, NASP115 and NASP179 ranked 7th, 18th and 10th out of a total of 473 structures, respectively, in the consensus based ranking (see Supplementary Table 2D) providing a good indication that NASP41 may be maintained by selection.

Visualisation of the 61nt long predicted structure (Figure 8) with DOOSS v1.0 shows that the majority of sites predicted to be base-paired are highly conserved, as indicated by the sequence logos for each base. Sites which show low levels of conservation, tend to be coevolving complementarily, indicated on the graphic, by red lines across nucleotide sites. The strong selection acting to maintain the overall structure suggests that this structure exists and it is an indication of its biological relevance.

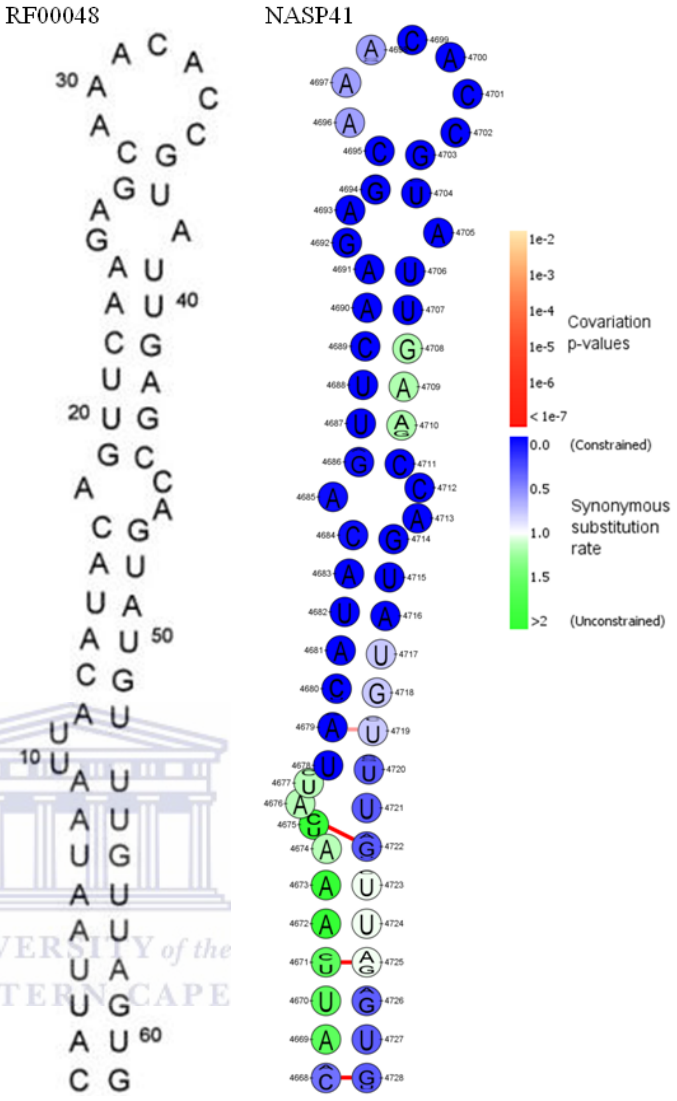


Figure 8. Visualisation of NASP 41 The NASP predicted structure is visualised alongside RF00048 (Goodfellow et al., 2004).

Nucleotides 4675 and 4722 show that they are evolving complementarily ($p = 4.34 \times 10^{-5}$) but are predicted to be unpaired by NASP. This mis-pairing may have been wrongly inferred by NASP given the evidence present. The possible pairing between nucleotides 4675 and 4722 is further supported by the pairing of these two bases observed in RF00048. The complementary coevolution of the less well conserved nucleotides suggests that the maintenance of the secondary structure significantly contributes to the low synonymous substitution rates observed within the codons of this structure. The relatively high ranking of this structure in the complementary coevolution rankings should not be at all surprising as many of the co-variance signals detected were long distance interactions, suggesting that tertiary folding conformation may also be acting to maintain the stability and functional relevance of this structure.

3.5 Recombination detection analysis

3.5.1 Breakpoint distribution analysis identifying hot- and cold-spots

In most of the datasets investigated, recombination hot-spots and/or -cold-spots were apparent and appeared to be non-randomly distributed. There were insufficient recombination events detected for the DuHV (8 detected events) and HAV (15 detected) datasets, for us to detect evidence of recombination breakpoint clustering and therefore, figures for these have not been included in this part of the analysis. Rather then indicating that the viral genomes in these two datasets recombine less than those of other picornaviruses, this result simply reflects the fact that these two datasets both had low numbers of sequences with relatively low degrees of diversity. The recombination breakpoint distributions observed in the FMDV, HEV-A, HAV-B, HAV-C, EMCV and PTeV datasets are in close agreement with those reported previously for picornaviruses (Heath et al. 2006, Simmonds 2006). Specifically, all of these six groups of viruses have strikingly similar recombination breakpoint patterns with the FMDV, HEV-B, HEV-C, PTeV genomes all displaying significant recombination breakpoint cold-spots within their 1B, 1C and 1D genes, and the FMDV, HEV-A, HEV-B, HEV-C, PTeV and EMCV displaying two recombination hot-spots on either side of the P1 genome region (indicated in yellow in Figure 9 below).

The presence of the 1A protein within these well-defined boundaries was somewhat unexpected/interesting as it is not exposed on the capsid surface, therefore being more evolutionary flexible at sequence level. These findings (well defined recombination boundaries) are also in accord with phylogenetic-compatibility analyses of (Simmonds 2006) that showed extensive phylogenetic incongruence between the structural protein encoding P1 genome regions of HEV-A, HEV-B and HEV-C viruses (i.e. genes 1A, 1B, 1C and 1D) and the remainder of their genomes. With each of the datasets containing various different serotypes such recombination patterns are suggestive/indicative of inter- as well as intra-typic recombination playing an important role in the evolutionary dynamics of these viruses. This type of partitioning of the structural and non-structural protein coding regions is reminiscent of the type of component swapping or re-assortment (also called pseudorecombination) that occurs in viruses with multi-component genomes. In addition evolution seems to have clustered the structural and non-structural genes to further facilitate the convenient swapping of complete structural protein encoding gene cassettes between genomes.

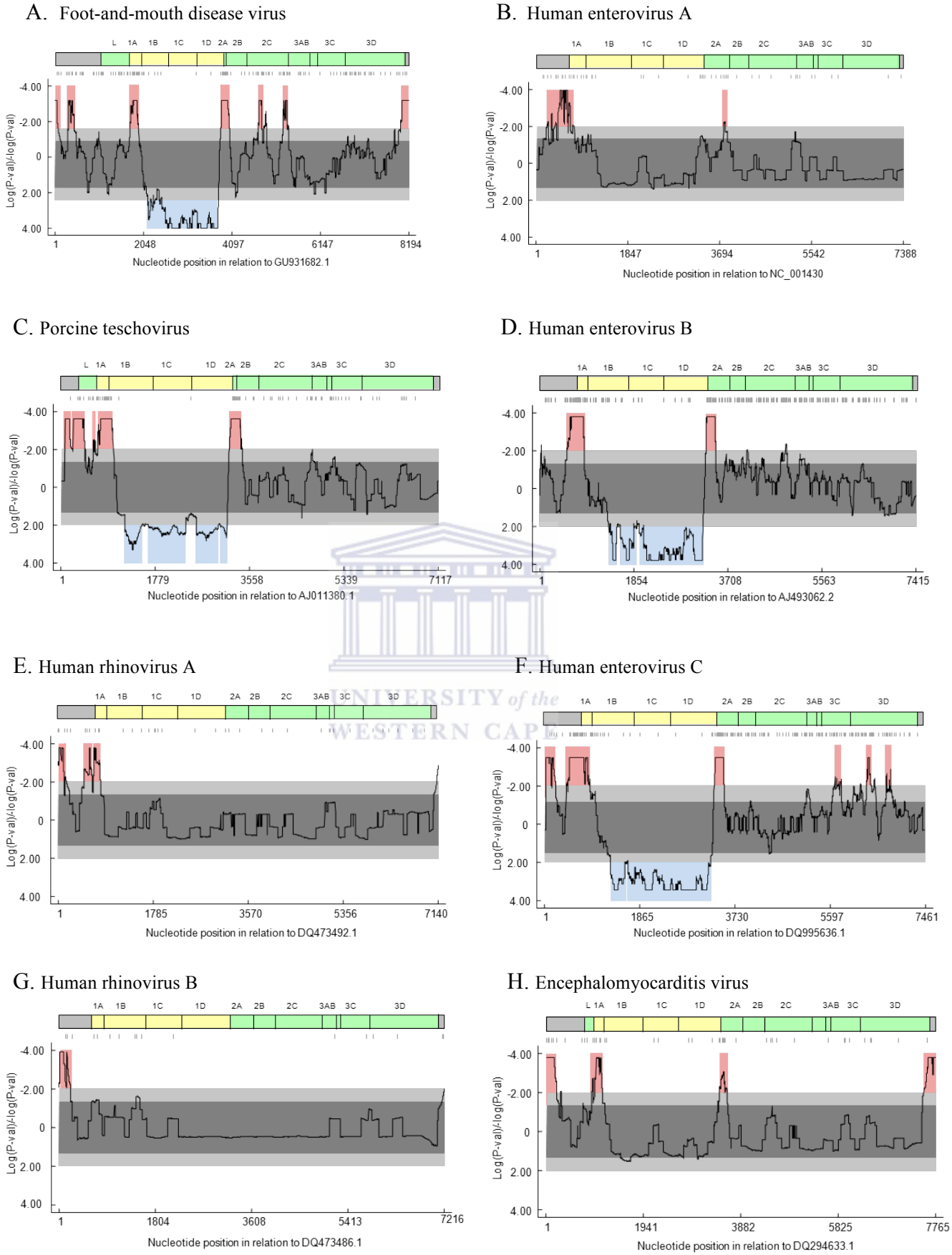


Figure 9. Breakpoint distribution plots The distribution of recombination breakpoints detected within (A) FMDV, (B) HEV-A, (C) PTeV, (D) HEV-B (E) HRV-A, (F) HEV-C, (G) HRV-B and (H) EMCV. Estimated breakpoint positions are indicated by small vertical lines at the top of the graph. Light and dark grey areas respectively indicate local 99% and 95% breakpoint clustering thresholds. Red areas indicate recombination hot-spots, while blue areas represent recombinational cold-spots. Gene maps above the graphs are drawn to scale in relation to the sequence of interest, indicating noncoding regions (grey) structural protein encoding genes (yellow) and non-structural protein encoding genes (green).

For the HEV-A, HEV-B and HEV-C datasets hot-spots are detected closer to the 5' end of the P1 region and breakpoint hot-spots are observed between the boundaries of 2A and 2B ORFs in Enterovirus A and in the 3C and 3D gene regions of Enterovirus C. The significant breakpoint hot-spots in PTeV and EMCV closely resemble those detected in FMDV, where the 5' UTR also contains significant evidence of breakpoint hot-spots (possibly due to the L-protein located ahead of 1A in the above mentioned genera).

The HRV-A and -B recombination breakpoint distributions also have elements consistent with the other datasets investigated, both showing significant breakpoint clustering in the 5' UTR with HRV-A also displaying a significant hot-spot at the 5' end of the P1 region.

Potentially, the reason that the rhinoviruses lack any evidence of recombination hotspots at the 3' ends of their P1 regions is that, compared with the other piconaviruses, the rhinoviruses have distinctly different patterns of sequence divergence in their structural protein encoding and non-structural protein encoding genome regions. It has been recently shown (McIntyre et al., 2010), that in HRV-A, HRV-B and HRV-C, there are substantially greater degrees of sequence divergence found throughout the coding regions of the genome (for example, an average of 32% divergence at the amino acid level across the entire coding region for any two HRV-A sequences). This is in contrast with the other picornaviruses where the non-structural protein encoding genome regions show markedly less sequence divergence among serotypes (<12% at amino acid level for the structural proteins) (McIntyre et al., 2010). It has been proposed that this very restriction in variability within the non-structural protein genes increases the likelihood of viruses productively exchanging these genes as intact modules to yield biologically viable recombinant progeny (Simmonds 2006). While recombination throughout picornavirus genomes may occur at similar rates (Heath et al. 2006), generally, recombinants that exchange non-modular pieces of sequence that end up not functioning as well within their new genomic backgrounds as they did within their original genomes will be less fit than the parental viruses and will therefore generally never survive for long enough in nature to be sampled and sequenced (van Rensburg et al. 2004). It is possible therefore, that the greater degrees of diversity found in rhinovirus genomes may either directly inhibit recombination between them or it may restrict the viability of whatever recombinant rhinoviruses arise in nature.

3.5.2 Influence of secondary structure on recombination breakpoint distributions

While it is evident that the recombination patterns observed amongst many of the picornavirus genomes are non-random, it is not obvious what mechanism(s) may be responsible for these non-random breakpoint distributions. It has been hypothesised (Simmonds and Welch 2006), that clustering of recombination events could be facilitated by preserved biochemical and/or secondary structural elements found at the boundaries between the structural and non-structural protein coding sequences. Convincing evidence that there is an association between predicted RNA structural elements and breakpoint clustering in picornavirus genomes was, however, not strongly supported by our analyses (Table 5). Although a marginal associations between breakpoint locations and predicted structural elements (i.e. with a p-value between 0.1 and 0.05) were detected for the FMDV, HEV-C and HRV-A datasets it is unlikely that the distribution of genomic secondary structures is as big a determinant of recombination breakpoint patterns in picornaviruses as it is in viruses such as HIV (Simon-Loriere et al. 2010) and other viruses where these associations have been detected. We tested whether recombination breakpoints occurred more frequently than could be accounted for by chance at sites that NASP had predicted were base-paired. Although other studies have reported clear association between secondary structure and breakpoint clustering (Draghici and Varrelmann 2010, Duch et al. 2004), we did not detect any clear tendency that breakpoints co-localised with regions predicted to be structured (Table 5). Although we did observe marginally insignificant p-values (between 0.05 and 0.10) for HRV-A, HEV-C and FMDV, this could partly be accounted to the fact that the HEV-C and FMDV datasets contain more evidence of breakpoint clustering in general, than the rest of the datasets.

Table 5. Table of associated p-values for testing correlation between breakpoint clustering and structured regions in picornavirus genomes

Dataset	p - value
FMDV	0.068
HEV-A	0.214
HEV-B	0.198
HEV-C	0.062
HRV-A	0.092
HRV-B	0.433
HAV	0.226
PTeV	0.319
EMCV	0.225
HPeV	0.442
DuHV	0.174

It is important to note that results presented here do not necessarily mean that local RNA secondary structure cannot/does not facilitate recombination, but simply, it does not account overwhelmingly for the recombination hot-spots detected in this study.

3.5.3 Breakpoint clustering in gene region

Recombination studies of other positive-sense single-stranded RNA viruses (Fu and Baric 1994, Pagan and Holmes 2010), have shown that recombination breakpoints tend to occur at gene boundaries rather than within the central regions of genes, suggesting that successful exchanges of genetic material amongst viral genomes have tended to involve transfers of intact or almost intact genes.

Virus	Gene regions [a]	p- value [b] (end 25% | mid 75%)
FMDV		0.05 | 0.94
HEV-A		0.04 | 0.89
HEV-B		0.03 | 0.98
HEV-C		0.01 | 0.99
HRV-A		0.86 | 0.24
HRV-B		0.49 | 0.79
HAV		0.54 | 0.72
EMCV		0.03 | 0.82
PTeV		0.01 | 0.99
HPeV		0.01 | 0.96

Figure 10. Breakpoint clustering at gene boundary regions [a] Genomes not drawn to scale relative to each other; however gene regions are drawn to scale with respect to their own genera. The blue coloured bars represent the end 12.5% of genes, while the orange parts represent the rest of the coding sequence in each gene. Grey regions represent the genomic UTRs. [b] p-values calculated based on number of breakpoints per 100nt in each gene region

Although, the genomes of picornaviruses contain a single polyprotein encoding ORF and lack extensive non-coding intergenic regions, we detected a tendency for recombination breakpoints to occur within the beginning and ending 12.5% of genes rather than in the middle 75% of genes (Figure 10). This tendency was significant ($P \leq 0.05$) for all but the HRV-A, HRV-B and HAV datasets. The results observed in the HRV-A, HRV-B and HAV datasets, may be explained by the fact that UTRs were not included in the calculation of these scores, while these specific genera show a high proportion of breakpoints in those very regions (Figure 9). The results we obtained when we compared the breakpoint clustering at the distal 50% and 10% of genes to the rest of the gene regions were more varied with only 3 and 5 datasets respectively, returning significant p-values (≤ 0.05) The results observed for the end 50% and 10% analyses are not entirely unexpected as we detected a significantly higher proportion of the breakpoints occur within the end 25% compared to the middle parts of genes (Figure 10), confirmed by the higher number of significant p-values for the end

regions in the 10% test relative to the 50% test. Detecting fewer breakpoints in the end 10%, as compared to the end 25% of genes may be accounted for by the relatively small size of the *Picornaviridae* genomes, making the confinement of breakpoints to these short regions (15 – 40nt) less likely to be observed in natural recombinants. Additionally, the test performed does not weigh the contributions to the total score of each individual gene, but rather treats all of the middle and end regions of every gene, in combination, as a single component.

These results are in agreement with previous reports (Bonnet et al. 2005, Voigt et al. 2002), supporting the hypothesis that recombination breakpoints that fall within genes increase the probability of disrupting the folding of the proteins encoded by recombinant genes.

3.5.4 Secondary structure fold disruption

The fitness of many viral species is, in part, dependant on the stability and distribution of secondary structure elements along their genomes (Davis et al. 2008, Simmonds et al. 2004). By using permutations tests, the aim was to determine if natural selection favoured the preservation within recombinant genomes of predicted base pairing interactions observed in the genomes of their parents.

The results of the tests applied here (Table 6) only provided strong evidence that there is a significant difference in disruption of secondary structures between the real and simulated recombinants in the HAV dataset. There was, however, also marginal evidence (p-values between 0.05 and 0.1) of such a trend in the FMDV, HEV-C and HRV-B datasets. This indicates that the observed recombination events detected in these datasets have tended to avoid disruption of predicted secondary structure to a greater degree than can be accounted for if recombination breakpoints were randomly distributed.

Table 6. Folding disruption associated p-values

Dataset	p-value
FMDV	0.062
HEV-A	0.368
HEV-B	0.898
HEV-C	0.067
HRV-A	0.519
HRV-B	0.059
HAV	**0.021**
EMCV	0.345
PTeV	0.443
HPeV	0.525
DuHV	0.177

It is interesting that for some of the other datasets such as HEV-B, HRV-A, and HPeV the predicted degree of folding disruption in recombinants was in fact higher than would be expected by chance (p-values $\geq$ 0.05). Although it is plausible that with larger datasets containing evidence of additional recombination breakpoints, many of the picornavirus groups with associated p-values $\leq$ 0.05 could eventually yield evidence that recombinants display a significant tendency to avoid secondary structure disruption, it is unlikely that this would be the case for datasets with associated p-values $\geq$ 0.05.

4. CONCLUSION

The purpose of this study was to quantify the evolutionary impacts of secondary structures within the genomes of picornaviruses. It was assumed that the degree to which secondary structural elements are evolutionary conserved should be directly correlated with the degree of biological importance during the viral life-cycle. It was therefore attempted to quantify the various selection pressures acting to maintain these structures, using these measurements firstly to confirm that the overall secondary structures of these genomes are biologically relevant and, secondly, to identify specific structural elements that appear to have particularly important biological functions.

The secondary structural elements that were most evolutionarily conserved within individual picornavirus species were computationally identified on a genome-wide scale. Significant evidence was discovered, of far greater degrees of predicted secondary structure within all of the examined picornavirus genomes than existed in randomised sequences of identical length and with identical nucleotide contents to the real picornavirus sequences.

We determined rates of synonymous substitution across the coding regions of the analysed picornavirus genomes (Table 1), and tested whether genomic sites which were predicted to be base-paired had lower-than-expected synonymous substitution rates. Although significant genome-wide associations between lower-than-expected synonymous substitution rates and degrees of base-pairing were only detected in three out of eleven of the datasets (Table 2). We found significant and marginally insignificant correlations for the FMDV and PTeV datasets respectively, which is consistent with the previous detection of extensive GORS in the genomes of viruses belonging to these species (Simmonds et al., 2004). However, whereas it was previously found that *Enteroviruses* (represented here by HEV-A, HEV,-B , HEV-C) and *Hepatoviruses* (represented here by HAV) lack evidence of GORS (Simmonds et al. 2004), we found evidence of genome wide associations between base paired sites and lower than expected synonymous substitution rates in all of the enterovirus datasets examined.

Using a model-based maximum likelihood method we found evidence on a genome-wide scale indicating that nucleotides predicted to be base paired, tended to co-evolve with one another in a complementary fashion in four out of the eleven species examined. Again, these included two enterovirus species (HEV-B and HEV-C) and FMDV, the same species with genome-wide evidence that coding sites that are base paired within secondary structures tend to have lower than expected synonymous substitution rates.

The analyses were then focused on individual structural elements within the genome-wide predicted structures. Predicted secondary structural elements were ranked according to their degree of evolutionary conservation, their associated synonymous substitution rates and the degree to which nucleotides predicted to be base-paired co-evolved with one another. Top ranking structures coincided with well characterised secondary structures that have been previously described in the literature (Supplementary Table 2). For example, coding regions with lower than expected synonymous substitution rates correlated with regions containing previously described (experimentally derived and computationally predicted) regulatory secondary structural elements called CREs (Figure 6). Similarly, significant signals ($p \leq 0.05$) of complementary coevolution between sites predicted to be base paired was found in previously proposed/experimentally, determined structures such as that indicated in Figure 8.

Next, the impact that genomic secondary structures had on the recombinational dynamics of picornavirus genomes was examined.

Recombination detection resulted in construction of breakpoint density plots from which we were able to identify locations of recombinational hot- and cold-spots along the viral genomes. These were compared against gene coordinates and pairing probability matrices for each of the analysed species in order to test if secondary structure had an effect on the recombinational patterns observed. Marginally insignificant evidence was observed for such an association in only two of our largest datasets (FMDV and HEV-C; Table 5), likely due to fact that these particular datasets contained more evidence of recombination as a whole and more evidence of recombination breakpoint clustering than any of the other datasets examined. We also tested whether selection favouring recombinant genomes with minimally disrupted secondary structures might influence picornavirus recombination patterns but only found marginal evidence for this is in the HAV dataset (Table 6). Although secondary structures seems to have no profound effects on recombination breakpoint patterns in the picornaviruses we showed that these patterns were very strongly influenced by the

distributions of coding and non-coding regions within picornavirus genomes. Across almost all the analysed datasets recombination breakpoints tended to cluster very significantly within non-coding regions. When breakpoints fell within coding regions they tended to occur within the ending 12.5% of genes (Figure 8): a pattern consistent with the hypothesis that recombination breakpoints that fall in the middle of genes have a greater probability of yielding genes that will encode dysfunctional chimaeric proteins than breakpoints that fall at the edges of genes (Bonnet et al. 2005, Voigt et al. 2002).

In conclusion, we have presented various lines of evidence that selection is acting to maintain structures within these viral genomes, in turn leading us to believe that some of the predicted structures do indeed exist, having an effect on the fitness of these viruses. However, their functional importance would have to be verified by biological experiments.

5. REFERENCES

Abbink TE, Berkhout B. 2003. A novel long distance base-pairing interaction in human immunodeficiency virus type 1 RNA occludes the Gag start codon. J Biol Chem 278: 11601-11611.

Andronescu MS, Pop C, Condon AE. 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: 1229-1236.

Bassili G, Tzima E, Song Y, Saleh L, Ochs K, Niepmann M. 2004. Sequence and secondary structure requirements in a highly conserved element for foot-and-mouth disease virus internal ribosome entry site activity and eIF4G binding. J Gen Virol 85: 2555-2565.

Bonnet J, Fraile A, Sacristan S, Malpica JM, Garcia-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.

Davis M, Sagan SM, Pezacki JP, Evans DJ, Simmonds P. 2008. Bioinformatic and physical characterizations of genome-scale ordered RNA structure in mammalian RNA viruses. J Virol 82: 11824-11836.

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 KE, Li TW, Mathews DH, Weeks KM. 2009. Accurate SHAPE-directed RNA structure determination. Proc Natl Acad Sci U S A 106: 97-102.

Domingo E, Holland JJ. 1997. RNA virus mutations and fitness for survival. Annu Rev Microbiol 51: 151-178.

Doshi KJ, Cannone JJ, Cobaugh CW, Gutell RR. 2004. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. BMC Bioinformatics 5: 105.

Draghici HK, 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.

Drouin G, Prat F, Ell M, Clarke GD. 1999. Detecting and characterizing gene conversions between multigene family members. Mol Biol Evol 16: 1369-1390.

Duch M, Carrasco ML, Jespersen T, Aagaard L, Pedersen FS. 2004. An RNA secondary structure bias for non-homologous reverse transcriptase-mediated deletions in vivo. Nucleic Acids Res 32: 2039-2048.

Duke GM, Hoffman MA, Palmenberg AC. 1992. Sequence and structural elements that contribute to efficient encephalomyocarditis virus RNA translation. J Virol 66: 1602-1609.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.

Fancello L, Trape S, Robert C, Boyer M, Popgeorgiev N, Raoult D, Desnues C. 2012. Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. ISME J.

Fernandez-Miragall O, Martinez-Salas E. 2003. Structural organization of a viral IRES depends on the integrity of the GNRA motif. RNA 9: 1333-1344.

Fernandez-Miragall O, Ramos R, Ramajo J, Martinez-Salas E. 2006. Evidence of reciprocal tertiary interactions between conserved motifs involved in organizing RNA structure essential for internal initiation of translation. RNA 12: 223-234.

Fernandez N, Fernandez-Miragall O, Ramajo J, Garcia-Sacristan A, Bellora N, Eyras E, Briones C, Martinez-Salas E. 2011. Structural basis for the biological relevance of the invariant apical stem in IRES-mediated translation. Nucleic Acids Res 39: 8572-8585.

Flamm C, Fontana W, Hofacker IL, Schuster P. 2000. RNA folding at elementary step resolution. RNA 6: 325-338.

Fontana W, Konings DA, Stadler PF, Schuster P. 1993. Statistics of RNA secondary structures. Biopolymers 33: 1389-1404.

Forterre P. 2006. The origin of viruses and their possible roles in major evolutionary transitions. Virus Res 117: 5-16.

Frederico LA, Kunkel TA, Shaw BR. 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry 29: 2532-2537.

Fu K, Baric RS. 1994. Map locations of mouse hepatitis virus temperature-sensitive mutants: confirmation of variable rates of recombination. J Virol 68: 7458-7466.

Galetto R, Moumen A, Giacomoni V, Veron 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.

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

Goodfellow I, Chaudhry Y, Richardson A, Meredith J, Almond JW, Barclay W, Evans DJ. 2000. Identification of a cis-acting replication element within the poliovirus coding region. J Virol 74: 4590-4600.

Guindon S, Dufayard JF, 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.

Gultyaev AP, van Batenburg FH, Pleij CW. 1995. The computer simulation of RNA folding pathways using a genetic algorithm. J Mol Biol 250: 37-51.

Heath L, van der Walt E, Varsani A, Martin DP. 2006. Recombination patterns in aphthoviruses mirror those found in other picornaviruses. J Virol 80: 11827-11832.

Hofacker IL, Stadler PF, Stocsits RR. 2004. Conserved RNA secondary structures in viral genomes: a survey. Bioinformatics 20: 1495-1499.

Jayan GC, Casey JL. 2005. Effects of conserved RNA secondary structures on hepatitis delta virus genotype I RNA editing, replication, and virus production. J Virol 79: 11187-11193.

Jenkins GM, Rambaut A, Pybus OG, Holmes EC. 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J Mol Evol 54: 156-165.

Johnson WE, Desrosiers RC. 2002. Viral persistence: HIV's strategies of immune system evasion. Annu Rev Med 53: 499-518.

Khromykh AA, Meka H, Guyatt KJ, Westaway EG. 2001. Essential role of cyclization sequences in flavivirus RNA replication. J Virol 75: 6719-6728.

Kieft JS, Zhou K, Jubin R, Doudna JA. 2001. Mechanism of ribosome recruitment by hepatitis C IRES RNA. RNA 7: 194-206.

King AM, McCahon D, Saunders K, Newman JW, Slade WR. 1985. Multiple sites of recombination within the RNA genome of foot-and-mouth disease virus. Virus Res 3: 373-384.

Kistler A, Avila PC, Rouskin S, Wang D, Ward T, Yagi S, Schnurr D, Ganem D, DeRisi JL, Boushey HA. 2007. Pan-viral screening of respiratory tract infections in adults with and without asthma reveals unexpected human coronavirus and human rhinovirus diversity. J Infect Dis 196: 817-825.

Knies JL, Dang KK, Vision TJ, Hoffman NG, Swanstrom R, Burch CL. 2008. Compensatory evolution in RNA secondary structures increases substitution rate variation among sites. Mol Biol Evol 25: 1778-1787.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. 2006. GARD: a genetic algorithm for recombination detection. Bioinformatics 22: 3096-3098.

Kusov YY, Gauss-Muller V. 1997. In vitro RNA binding of the hepatitis A virus proteinase 3C (HAV 3Cpro) to secondary structure elements within the 5' terminus of the HAV genome. RNA 3: 291-302.

La Scola B, et al. 2008. The virophage as a unique parasite of the giant mimivirus. Nature 455: 100-104.

Lu HH, Wimmer E. 1996. Poliovirus chimeras replicating under the translational control of genetic elements of hepatitis C virus reveal unusual properties of the internal ribosomal entry site of hepatitis C virus. Proc Natl Acad Sci U S A 93: 1412-1417.

Lukashev AN. 2010. Recombination among picornaviruses. Rev Med Virol 20: 327-337.

Markham NR, Zuker M. 2008. UNAFold: software for nucleic acid folding and hybridization. Methods Mol Biol 453: 3-31.

Martin DP, van der Walt E, Posada D, Rybicki EP. 2005. The evolutionary value of recombination is constrained by genome modularity. PLoS Genet 1: e51.

Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26: 2462-2463.

Martinez-Salas E, Fernandez-Miragall O. 2004. Picornavirus IRES: structure function relationship. Curr Pharm Des 10: 3757-3767.

McKnight KL, Lemon SM. 1998. The rhinovirus type 14 genome contains an internally located RNA structure that is required for viral replication. RNA 4: 1569-1584.

Mironov AA, Dyakonova LP, Kister AE. 1985. A kinetic approach to the prediction of RNA secondary structures. J Biomol Struct Dyn 2: 953-962.

Moratorio G, Costa-Mattioli M, Piovani R, Romero H, Musto H, Cristina J. 2007. Bayesian coalescent inference of hepatitis A virus populations: evolutionary rates and patterns. J Gen Virol 88: 3039-3042.

Moya A, Elena SF, Bracho A, Miralles R, Barrio E. 2000. The evolution of RNA viruses: A population genetics view. Proc Natl Acad Sci U S A 97: 6967-6973.

Muse SV. 1995. Evolutionary analyses of DNA sequences subject to constraints of secondary structure. Genetics 139: 1429-1439.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11: 715-724.

Ngandu NK, 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.

Ooms M, Abbink TE, Pham C, Berkhout B. 2007. Circularization of the HIV-1 RNA genome. Nucleic Acids Res 35: 5253-5261.

Pagan I, Holmes EC. 2010. Long-term evolution of the Luteoviridae: time scale and mode of virus speciation. J Virol 84: 6177-6187.

Paul AV, Rieder E, Kim DW, van Boom JH, 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: 10359-10370.

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

Pilipenko EV, Blinov VM, Chernov BK, Dmitrieva TM, Agol VI. 1989. Conservation of the secondary structure elements of the 5'-untranslated region of cardio- and aphthovirus RNAs. Nucleic Acids Res 17: 5701-5711.

Pollard VW, Malim MH. 1998. The HIV-1 Rev protein. Annu Rev Microbiol 52: 491-532.

Poon AF, Lewis FI, Frost SD, Kosakovsky Pond SL. 2008. Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models. Bioinformatics 24: 1949-1950.

Posada D, Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci U S A 98: 13757-13762.

Ravetch JV, Horiuchi K, Zinder ND. 1977. Nucleotide sequences near the origin of replication of bacteriophage f1. Proc Natl Acad Sci U S A 74: 4219-4222.

Reynolds JE, Kaminski A, Kettinen HJ, Grace K, Clarke BE, Carroll AR, Rowlands DJ, Jackson RJ. 1995. Unique features of internal initiation of hepatitis C virus RNA translation. EMBO J 14: 6010-6020.

Richmond TJ, Davey CA. 2003. The structure of DNA in the nucleosome core. Nature 423: 145-150.

Rivera R, Nollens HH, Venn-Watson S, Gulland FM, Wellehan JF, Jr. 2010. Characterization of phylogenetically diverse astroviruses of marine mammals. J Gen Virol 91: 166-173.

Scheffler K, Martin DP, Seoighe C. 2006. Robust inference of positive selection from recombining coding sequences. Bioinformatics 22: 2493-2499.

Semegni JY, Wamalwa M, Gaujoux R, Harkins GW, Gray A, Martin DP. 2011. NASP: a parallel program for identifying evolutionarily conserved nucleic acid secondary structures from nucleotide sequence alignments. Bioinformatics 27: 2443-2445.

Shang L, Xu W, Ozer S, Gutell RR. 2012. Structural constraints identified with covariation analysis in ribosomal RNA. PLoS One 7: e39383.

Shao Y, Chan CY, Maliyekkel A, Lawrence CE, Roninson IB, Ding Y. 2007. Effect of target secondary structure on RNAi efficiency. RNA 13: 1631-1640.

Shepherd DN, Martin DP, Varsani A, Thomson JA, Rybicki EP, Klump HH. 2006. Restoration of native folding of single-stranded DNA sequences through reverse mutations: an indication of a new epigenetic mechanism. Arch Biochem Biophys 453: 108-122.

Simmonds P. 2006. Recombination and selection in the evolution of picornaviruses and other Mammalian positive-stranded RNA viruses. J Virol 80: 11124-11140.

Simmonds P, Smith DB. 1999. Structural constraints on RNA virus evolution. J Virol 73: 5787-5794.

Simmonds P, Welch J. 2006. Frequency and dynamics of recombination within different species of human enteroviruses. J Virol 80: 483-493.

Simmonds P, Tuplin A, Evans DJ. 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: 1337-1351.

Simmonds P, Karakasiliotis I, Bailey D, Chaudhry Y, Evans DJ, Goodfellow IG. 2008. Bioinformatic and functional analysis of RNA secondary structure elements among different genera of human and animal caliciviruses. Nucleic Acids Res 36: 2530-2546.

Simon-Loriere E, Martin DP, Weeks KM, Negroni M. 2010. RNA structures facilitate recombination-mediated gene swapping in HIV-1. J Virol 84: 12675-12682.

Smith AW, Akers TG, Latham AB, Skilling DE, Bray HL. 1979. A new calicivirus isolated from a marine mammal. Arch Virol 61: 255-259.

Steil BP, Barton DJ. 2009. Cis-active RNA elements (CREs) and picornavirus RNA replication. Virus Res 139: 240-252.

Tafer H, Ameres SL, Obernosterer G, Gebeshuber CA, Schroeder R, Martinez J, Hofacker IL. 2008. The impact of target site accessibility on the design of effective siRNAs. Nat Biotechnol 26: 578-583.

Teterina NL, Gorbalenya AE, Egger D, Bienz K, Rinaudo MS, Ehrenfeld E. 2006. Testing the modularity of the N-terminal amphipathic helix conserved in picornavirus 2C proteins and hepatitis C NS5A protein. Virology 344: 453-467.

Thiviyanathan V, Yang Y, Kaluarachchi K, Rijnbrand R, Gorenstein DG, Lemon SM. 2004. High-resolution structure of a picornaviral internal cis-acting RNA replication element (cre). Proc Natl Acad Sci U S A 101: 12688-12693.

Tian D, Wang Q, Zhang P, Araki H, Yang S, Kreitman M, Nagylaki T, Hudson R, Bergelson J, Chen JQ. 2008. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature 455: 105-108.

Tuplin A, Evans DJ, 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.

van Rensburg HG, Henry TM, Mason PW. 2004. Studies of genetically defined chimeras of a European type A virus and a South African Territories type 2 virus reveal growth determinants for foot-and-mouth disease virus. J Gen Virol 85: 61-68.

Voigt CA, Martinez C, Wang ZG, Mayo SL, Arnold FH. 2002. Protein building blocks preserved by recombination. Nat Struct Biol 9: 553-558.

Watson JD, Crick FH. 1953. The structure of DNA. Cold Spring Harb Symp Quant Biol 18: 123-131.

Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW, Jr., Swanstrom R, Burch CL, Weeks KM. 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. Nature 460: 711-716.

Wilkinson KA, Merino EJ, Weeks KM. 2006. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. Nat Protoc 1: 1610-1616.

Willis IM. 1993. RNA polymerase III. Genes, factors and transcriptional specificity. Eur J Biochem 212: 1-11.

Witwer C, Rauscher S, Hofacker IL, Stadler PF. 2001. Conserved RNA secondary structures in Picornaviridae genomes. Nucleic Acids Res 29: 5079-5089.

Wuchty S, Fontana W, Hofacker IL, Schuster P. 1999. Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers 49: 145-165.

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

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

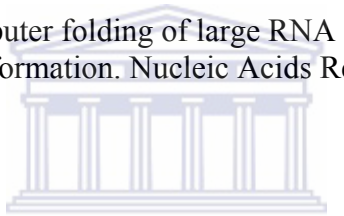
Yang Y, Yi M, Evans DJ, Simmonds P, Lemon SM. 2008. Identification of a conserved RNA replication element (cre) within the 3Dpol-coding sequence of hepatoviruses. J Virol 82: 10118-10128.

Ye K, Malinina L, Patel DJ. 2003. Recognition of small interfering RNA by a viral suppressor of RNA silencing. Nature 426: 874-878.

Zhang B, Dong H, Stein DA, Iversen PL, Shi PY. 2008. West Nile virus genome cyclization and RNA replication require two pairs of long-distance RNA interactions. Virology 373: 1-13.

Zuker M. 1989. Computer prediction of RNA structure. Methods Enzymol 180: 262-288.

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



6. SUPPLEMENTARY DATA

Supplementary Table 1. Association between base-pairing and constraints on synonymous substitution rates across coding region

Please Note:

All of the supplementary data and materials are available at: www.sanbi.ac.za/~emil/msc_supp.zip

A. FMDV

Observed:

	Constrained	Unconstrained	Total
0	214.0	129.0	343.0
1	211.0	132.0	343.0
2	347.0	189.0	536.0
3	765.0	359.0	1124.0
Total	1537.0	809.0	2346.0

Expected:

	Constrained	Unconstrained	Total
0	224.7	118.3	343.0
1	224.7	118.3	343.0
2	351.1	184.4	536.0
3	736.4	387.6	1124.0
Total	1537.0	809.0	2346.0

cutoff = 1.0
chi = 7.28
pvalue = 0.05

B. HEV-A

Observed:

	Constrained	Unconstrained	Total
0	256.0	135.0	391.0
1	204.0	115.0	319.0
2	352.0	150.0	502.0
3	657.0	324.0	981.0
Total	1469.0	724.0	2193.0

Expected:

	Constrained	Unconstrained	Total
0	261.9	129.1	391.0
1	213.6	105.2	319.0
2	336.2	165.7	502.0
3	657.1	323.9	981.0
Total	1469.0	724.0	2193.0

cutoff = 1.0
chi = 3.92
pvalue = 0.03

C. HEV-B

Observed:

	Constrained	Unconstrained	Total
0	200.0	222.0	422.0
1	130.0	202.0	332.0
2	200.0	219.0	419.0
3	468.0	569.0	1037.0
Total	998.0	1212.0	2210.0

Expected:

	Constrained	Unconstrained	Total
0	190.5	231.4	422.0
1	149.9	182.1	332.0
2	189.2	229.7	419.0
3	468.3	568.7	1037.0
Total	998.0	1212.0	2210.0

cutoff = 1.0
chi = 6.80
pvalue = 0.06

D. HEV-C

Observed:

	Constrained	Unconstrained	Total
0	207.0	228.0	435.0
1	173.0	169.0	342.0
2	263.0	201.0	464.0
3	527.0	460.0	987.0
Total	1170.0	1058.0	2228.0

Expected:

	Constrained	Unconstrained	Total
0	228.4	206.6	435.0
1	179.5	162.4	342.0
2	243.6	220.3	464.0
3	518.3	468.7	987.0
Total	1170.0	1058.0	2228.0

cutoff = 1.0
chi = 8.32
pvalue = 0.04

E. HRV-A

Observed:			
	Constrained	Unconstrained	Total
0	157.0	302.0	459.0
1	124.0	196.0	320.0
2	150.0	262.0	412.0
3	344.0	662.0	1006.0
Total	775.0	1422.0	2197.0

Expected:			
	Constrained	Unconstrained	Total
0	161.9	297.1	459.0
1	112.9	207.1	320.0
2	145.3	266.6	412.0
3	354.9	651.1	1006.0
Total	775.0	1422.0	2197.0

cutoff = 1.0
chi = 2.66
pvalue = 0.32

F. HRV-B

Observed:			
	Constrained	Unconstrained	Total
0	159.0	305.0	459.0
1	120.0	200.0	320.0
2	152.0	280.0	412.0
3	3662.0	1006.0	
Total	775.0	1422.0	2197.0

Expected:			
	Constrained	Unconstrained	Total
0	161.9	297.1	459.0
1	112.9	207.1	320.0
2	145.3	266.6	412.0
3	354.8	651.1	1006.0
Total	775.0	1422.0	2197.0

cutoff = 1.0
chi = 2.66
pvalue = 0.44

G. HAV

Observed:			
	Constrained	Unconstrained	Total
0	198.0	213.0	411.0
1	133.0	156.0	289.0
2	244.0	247.0	491.0
3	467.0	571.0	1038.0
Total	1042.0	1187.0	2229.0

Expected:			
	Constrained	Unconstrained	Total
0	192.1	218.8	411.0
1	135.1	153.9	289.0
2	229.5	261.4	491.0
3	485.3	552.7	1038.0
Total	1042.0	1187.0	2229.0

cutoff = 1.0
chi = 3.39
pvalue = 0.33

H. HPeV

Observed:			
	Constrained	Unconstrained	Total
0	195.0	248.0	443.0
1	145.0	202.0	347.0
2	192.0	214.0	406.0
3	459.0	535.0	994.0
Total	991.0	1199.0	2190.0

Expected:			
	Constrained	Unconstrained	Total
0	200.5	242.5	443.0
1	157.1	189.9	347.0
2	183.7	222.3	406.0
3	449.8	544.2	994.0
Total	991.0	1199.0	2190.0

cutoff = 1.0
chi = 2.97
pvalue = 0.39

I. DuHV

Observed:			
	Constrained	Unconstrained	Total
0	248.0	171.0	419.0
1	200.0	123.0	323.0
2	335.0	174.0	509.0
3	639.0	363.0	1002.0
Total	1422.0	831.0	2253.0

Expected:			
	Constrained	Unconstrained	Total
0	264.4	154.5	419.0
1	203.8	119.1	323.0
2	321.2	187.7	509.0
3	632.4	369.6	1002.0
Total	1422.0	831.0	2253.0

cutoff = 1.0
chi = 4.75
pvalue = 0.19

J. EMCV

Observed:			
	Constrained	Unconstrained	Total
0	128.0	305.0	433.0
1	103.0	217.0	320.0
2	155.0	302.0	457.0
3	372.0	735.0	1107.0
Total	758.0	1559.0	2317.0

Expected:			
	Constrained	Unconstrained	Total
0	141.6	291.3	433.0
1	104.6	215.3	320.0
2	149.5	307.4	457.0
3	362.1	744.8	1107.0
Total	758.0	1559.0	2317.0

cutoff = 1.0
chi = 2.69
pvalue = 0.44

Supplementary table 2. Consensus ranking results

A. FMDV

Consensus Rank	NASP ID	Location	Length	Reference
1	81	7956-8076	120	
2	79	7801-7850	49	
3	36	4313-4445	132	
4	2	618-649	31	Bassili et al., 2004
5	176	7808-7843	35	
6	6	1006-1035	29	Fernandez et al., 2003
7	261	7982-8011	29	
8	111	2905-2917	12	
9	4	788-867	79	Martinez et al., 2003
10	283	4208-4225	17	
11	125	4379-4416	37	
12	8	1119-1161	42	Mason et al., 2002
13	122	4198-4239	41	
14	35	4254-4311	57	
15	124	4322-4371	49	
16	313	7986-8007	21	
17	218	4384-4411	27	
18	215	4204-4230	26	
19	333	6077-6087	10	
20	326	4211-4222	11	
21	286	4388-4408	20	
22	93	1125-1155	30	
23	78	7675-7797	122	
24	216	4275-4296	21	
25	9	1232-1319	87	
26	343	7989-8002	13	
27	198	2033-2045	12	
28	213	4062-4076	14	
29	189	1129-1150	21	
30	296	6074-6090	16	

B. HEV-A

Consensus Rank	NASP ID	Location	Length	Reference
1	5	4415-4456	41	Steil and Barton, 2008
2	12	4423-4447	24	
3	308	4410-4459	49	
4	138	4386-4549	163	
5	203	4406-4463	57	
6	368	5299-5369	70	
7	78	4393-4542	149	Goodfellow et al., 2000
8	369	5303-5366	63	
9	371	4880-4994	114	
10	285	5293-5375	82	
11	64	4929-4945	16	
12	80	4463-4496	33	
13	161	2583-2607	24	Paul et al., 2000
14	129	4887-4990	103	
15	127	5349-5363	14	
16	303	7244-7338	94	Paul et al., 2000
17	162	4894-4984	90	
18	197	1440-1490	50	
19	141	4474-4485	11	
20	216	996-1124	128	
21	33	4707-4723	16	
22	118	987-1270	283	
23	100	3412-3562	150	
24	234	1805-1834	29	
25	69	2775-2914	139	
26	97	4284-4309	25	
27	288	5127-5138	11	
28	377	2785-2903	118	
29	136	4430-4441	11	
30	37	1077-1109	32	

C. HEV-B

Consensus Rank	NASP ID	Location	Length	Reference
1	217	4476-4517	41	Goodfellow et al., 2000
2	342	4482-4511	29	
3	437	4486-4506	20	
4	78	4369-4518	149	Goodfellow et al., 2000
5	63	3541-3561	20	
6	408	1466-1483	17	
7	428	3657-3687	30	
8	62	3420-3540	120	
9	66	3612-3731	119	
10	446	5409-5426	17	
11	127	7304-7399	95	Paul et al., 2000
12	24	1567-1631	64	
13	292	1278-1331	53	
14	391	7358-7382	24	
15	295	1462-1487	25	
16	202	3544-3558	14	
17	21	1253-1350	97	
18	160	1576-1623	47	
19	207	3793-3804	11	
20	205	3633-3720	87	
21	532	3664-3677	13	
22	332	3645-3710	65	
23	255	6478-6488	10	Paul et al., 2000
24	407	1285-1324	39	
25	509	5412-5423	11	
26	484	1291-1320	29	
27	68	3790-3807	17	
28	155	1267-1339	72	
29	361	5586-5596	10	
30	522	7271-7282	11	

D. HEV-C

Consensus Rank	NASP ID	Location	Length	
1	66	7244-7346	102	
2	115	4678-4721	43	Steil and Barton, 2008
3	71	7540-7630	90	
4	147	7552-7619	67	
5	141	7251-7278	27	Paul et al., 2000
6	179	4684-4715	31	
7	41	4668-4729	61	Goodfellow et., 200
8	194	7255-7274	19	Paul et al., 2000
9	198	7560-7610	50	
10	179	4688-4710	22	
11	236	4691-4707	16	
12	224	7258-7272	14	
13	241	7260-7270	10	
14	226	7568-7582	14	
15	146	7490-7511	21	
16	191	7091-7111	20	
17	63	7081-7120	39	
18	115	4694-4705	11	
19	238	6026-6038	12	
20	65	7182-7217	35	
21	138	7086-7115	29	
22	142	7288-7303	15	
23	195	7290-7301	11	
24	67	7347-7360	13	
25	197	7493-7508	15	
26	242	7570-7580	10	
27	186	6020-6045	25	
28	225	7495-7506	11	
29	227	7587-7603	16	
30	223	7094-7108	14	

E. HRV-A

Consensus Rank	NASP ID	Location	Length	Reference
1	129	3392-3419	27	
2	237	6528-6551	23	
3	8	652-671	48	Kistler et al., 2007
4	209	3396-3415	19	Gerber et al., 2001
5	34	2760-2785	25	
6	75	6517-6562	45	
7	38	3312-3432	120	
8	185	833-843	10	Kistler et al., 2007
9	71	6277-6308	31	
10	41	3559-3600	41	
11	168	6523-6556	33	
12	256	2989-3009	20	
13	164	6281-6304	23	
14	122	2764-2781	17	
15	40	3482-3506	24	
16	205	2985-3013	28	
17	117	2387-2421	34	
18	10	827-849	22	
19	204	2767-2778	11	
20	35	2941-3091	150	
21	157	5701-5726	25	
22	141	4377-4394	17	
23	235	6284-6301	17	
24	95	830-846	16	
25	216	4380-4391	11	
26	155	5553-5574	21	
27	9	727-752	25	
28	175	7041-7113	72	
29	123	2980-3018	38	
30	101	1249-1266	17	

F. HRV-B

Consensus Rank	NASP ID	Location	Length	
1	128	2374-2425	51	McKnight and Lemon, 1998
2	231	2380-2418	38	
3	300	2385-2413	28	
4	341	2389-2409	20	
5	39	2974-3116	142	
6	33	2355-2503	148	
7	37	2824-2882	58	
8	24	1706-1728	22	
9	136	2981-3073	92	
10	35	2650-2753	103	
11	104	805-816	11	Tapparel et., 2007
12	118	1709-1725	16	
13	10	802-819	17	
14	143	3440-3471	31	
15	125	2102-2198	96	
16	134	2832-2874	42	
17	245	3444-3467	23	
18	103	746-782	36	
19	228	2111-2145	34	
20	147	3633-3656	23	
21	30	2072-2221	149	
22	21	1468-1482	14	
23	339	2121-2133	12	
24	239	2993-3061	68	
25	344	3291-3307	16	
26	44	3402-3552	150	
27	249	3636-3653	17	
28	45	3575-3672	97	
29	242	3284-3314	30	
30	350	6658-6668	10	

G. HAV

Consensus Rank	NASP ID	Location	Length	Reference
1	86	6006-6092	86	
2	206	6018-6082	64	
3	11	746-857	111	Brown et al., 1995
4	317	6027-6072	45	
5	105	7406-7464	58	
6	121	786-825	39	Kolupaeva et al., 2000
7	398	6033-6066	33	
8	340	796-814	18	Brown et al., 1995
9	230	7346-7393	47	
10	242	791-819	28	
11	445	6038-6062	24	
12	231	7414-7456	42	Kusov et al., 1997
13	332	7352-7387	35	
14	408	7357-7382	25	
15	415	799-811	12	
16	104	7248-7398	150	
17	333	7420-7450	30	Kusov et al., 1997
18	451	7361-7378	17	
19	27	2428-2443	15	
20	122	826-846	20	
21	429	3231-3244	13	
22	243	830-843	13	
23	344	1149-1165	16	
24	417	1152-1162	10	
25	409	7424-7446	22	
26	147	2430-2441	11	
27	452	7427-7443	16	
28	247	1134-1180	46	
29	218	6703-6717	14	Kusov et al., 1997
30	192	5102-5112	10	

H. PTeV

Consensus Rank	NASP ID	Location	Length	Reference
1	159	6807-6842	35	
2	181	6812-6837	25	
3	45	5890-5936	46	
4	116	6799-6851	52	
5	196	6817-6832	15	
6	56	6789-6866	77	
7	63	4456-4512	56	
8	76	1590-1616	26	
9	65	500-521	21	Witwer et al., 2001
10	96	4840-4858	18	
11	32	4560-4581	21	
12	24	3588-3604	16	
13	131	1594-1612	18	
14	126	503-518	15	Witwer et al., 2001
15	85	3591-3601	10	
16	127	554-593	39	
17	134	2838-2852	14	
18	143	4843-4855	12	
19	4	481-618	137	Witwer et al., 2001
20	48	6165-6197	32	
21	117	6890-6952	62	
22	8	1116-1132	16	
23	160	6899-6944	45	
24	70	1119-1129	10	
25	14	1585-1621	36	
26	40	5541-5570	29	
27	109	6221-6250	29	
28	115	6733-6744	11	
29	106	5972-5998	26	
30	176	5904-5921	17	