Evolutionary Impacts of Secondary Structures within the Genomes of Eukaryote-Infecting Single-Stranded DNA Viruses

Thesis presented for the degree of

DOCTOR OF PHILOSOPHY

in the

Department of Integrative BioMedical Sciences

at the

UNIVERSITY OF CAPE TOWN

August 2015

Author:

Brejnev Muhizi Muhire

Supervisor:

A/Prof. Darren Martin

Abstract

Secondary structures forming through base-pairing in virus genomes have been proven to regulate several processes during viral replication cycles, including genome replication, transcription, post-transcriptional activities, protein synthesis, genome packaging, generation of viral sub-genomes and evasion of host-cell immune responses. Although computational DNA/RNA folding methods based-on free energy minimisation approaches are capable of predicting structures that form within virus genomes, these methods are not entirely accurate. Notably, many of structures that are accurately predicted will likely have no biological importance within the genomes in which they reside because even randomly generated single-stranded RNA/DNA sequences will form stable secondary structures. Nevertheless, with additional genome evolution analyses involving the detection of natural selection, sequence co-evolution, and genetic recombination, it is possible to both validate the existence of, and infer the biological importance of, computationally predicted structures. Here I implement and deploy free bioinformatics tools to (1) automate nucleotide and protein sequences classification into datasets useful for downstream molecular evolution analyses; (2) improve the accuracy of computational virus-genome-scale secondary structure prediction; (3) enable the identification of biologically relevant secondary structures using signals of purifying selection, coevolution and recombination within aligned sequence datasets; and (4) enable efficient visualisation of structural and selection data for better characterisation of individual secondary structural elements. Using these tools I carried-out large scale studies that predicted and characterised novel functional secondary structures, that potentially regulate transcription, translation, gene splicing, and replication, within the genomes of eukaryote-infecting ssDNA viruses (*Circoviridae*, *Anelloviridae*, *Parvoviridae*, *Nanoviridae*, and *Geminiviridae*). I show that purifying selection tends to be stronger at base-paired sites than it is at unpaired sites and, wherever mutations are tolerable within paired regions, I demonstrate that there exist strong associations between base-pairing and complementary coevolution. Finally, I show that the recombinant genomes of some, but not all, eukaryote-infecting ssDNA virus groups display weak evidence of both homologous

and non-homologous recombination break-points preferentially occurring at genome sites that minimally disrupt secondary structures. Altogether, these results suggest that natural selection acting to maintain important biologically functional secondary structural elements has been a major process during the evolution of eukaryote-infecting ssDNA viruses.

Acknowledgements

I wish to thank Associate Professor Darren Martin who supervised this research and ensured my intellectual progress and financial stability over the course of my PhD degree. Under Darren's supervision, I quickly learned a large amount of biology (a challenge to any mathematician wanting to transform into a bioinformatician) and I was fortunate to spend a lot of time working closely with him, which significantly enhanced my skills in software development and scientific writing: skills crucial to my future bioinformatics career.

I thank the University of Cape Town (UCT) for the International & Refugee Scholarship awarded to me. This was my first award without which I wouldn't have managed to join the Bioinformatics programme. I acknowledge the Carnegie Corporation and Poliomyelitis Research Foundation for funding this research.

Special thanks to the Head of our Research Group, Associate Professor Nicky Mulder. Her guidance, courage and kindness have been an inspiration throughout my postgraduate studies and will continue to be, in my future research endeavours. Special thanks to Michael Golden for making tremendous contributions toward this work, his close collaboration played an influential role in improving my programming skills and aptitude. A special thank you to Arvind Varsani, Emil Tanov, Fredrick Nindo, Ben Murrell, Gordon Harkins, Philippe Roumagnac, Pierre Lefeuvre, Simona Kraberger, Daisy Stainton, Adérito Monjane, Rebone Meraba, Penelope Hartnady and the UCT's Computational Biology research group for your intellectual and moral contribution which always kept me moving forward. This thesis is dedicated to my daughter Joanne.

Acronyms

BLAST	:	Basic Local Alignment Search Tool
BMV	:	Brome mosaic virus
cp	:	Capsid protein gene
CP	:	Capsid protein
DEmARC	:	DivErsity pArtitioning by hieRarchical Clustering
DENV	:	Dengue virus
DOOSS	:	Data Overlaid On Secondary Structure
dN	:	Non-synonymous substitution rates
DNA	:	Deoxyribonucleic acid
dS	:	Synonymous substitution rates
dsDNA	:	Double-stranded DNA
dsRNA	:	Double-stranded RNA
FUBAR	:	Fast Unconstrained Bayesian Approximation
EMF	:	Enhance Metafile Format
HCSS	:	High Confidence Structure Set
HCV	:	Hepatitis C virus
HIV	:	Human immunodeficiency virus
HyPhy	:	Hypothesis testing using Phylogenies
GARD	:	Genetic Algorithm for Recombination Detection
GUI	:	Graphical User Interface
GDI	:	Graphical Device Interface
kb	:	Kilobase
MFE	:	Minimum Free Energy
ML	:	Maximum Likelihood

mp	:	Movement protein gene
MP	:	Movement protein
MSV	:	Maize streak virus
NASP	:	Nucleic Acid Secondary-Structure Predictor
NAVA	:	Nucleic Acid Visualisation and Analysis
nt	:	Nucleotide
NW	:	Needleman-Wunsch
OTUs	:	Operational Taxonomic Units
PARRIS	:	Partitioning Approach for Inference of Selection
PASC	:	PAirwise Sequence Comparison
PCV	:	Porcine circovirus
PNG	:	Portable Network Graphic
RNA	:	Ribonucleic acid
RDP	:	Recombination Detection Program
rep	:	Replication associated-protein gene
Rep	:	Replication associated protein
SDT	:	Sequence Demarcation Tool
SHAPE	:	Selective 2‘-Hydroxyl Acylation by Primer Extension
SIV	:	Simian immunodeficiency virus
ssDNA	:	Single-stranded DNA
ssRNA	:	Single-stranded RNA

Table of contents

List of tables

List of figures

Chapter 1 : Introduction

Despite the increasing accuracy of current computational tools for predicting secondary structures within nucleic acid molecules such as those found in virus genomes, these tools do not have the capacity to identify biologically functional elements among the predicted stable structures. For virologists wishing to understand the exact biological function and impact of genomic secondary structure on virus evolution, the identification of biologically functional structural elements against a background of non-functional structures, is therefore a daunting task. Here I develop and implement a range of molecular evolution computational tools that greatly improve the prediction, visualisation and characterisation of functional genomic secondary structures within virus genomes. These tools essentially rely on biological sequence data to computationally predict secondary structural elements and assess their biological relevance by applying different molecular evolution analyses to test for evidence of natural selection acting to maintain these structures. These tools are to play pivotal role in guiding lab experiments that ideally need to focus exclusively on the most likely functional elements when studying the biological functions of genomic secondary structures. Adding to the implementation, I use these tools in large scale studies that identify and characterise biologically functional structures conserved within the genomes of selected eukaryote-infecting single-stranded DNA (ssDNA) virus families (*Circoviridae*, *Anelloviridae*, *Parvoviridae*, *Nanoviridae* and *Geminiviridae*). This introductory chapter reviews the current advances in computational prediction of nucleic acid secondary structures pointing out major limitations and caveats and also describes the diverse group of eukaryote-infecting ssDNA viruses that form the basis of this research.

1.1 Nucleic acid secondary structures within virus genomes

Besides the characteristic double-helix formed via hydrogen bond interactions between complementary nucleotides (i.e. Watson-Click A-T and C-G, and Wobble base-paring T-G) by deoxyribonucleic acid (DNA), as well as the more recently characterized ribonucleic acid (RNA) double-helix (Safaee et al. 2013), DNA/RNA molecules that are single-stranded can form complex structures via base-pairing

interactions between nucleotides on the same strand. These structures are usually referred to as “secondary structures” when considering only pairing interactions between nucleotides or “tertiary structures” when accounting for their configuration in three dimensional space.

In living organisms these structures have various regulatory functions. In the RNA interference pathway for example, well defined hairpins facilitate nucleic-acid—nucleic-acid interactions of short interfering RNAs (Harborth et al. 2003) and during peptide synthesis they play a role in the nucleic-acid—amino-acid interactions of transfer RNAs (Cramer 1971; Ishitani et al. 2003), messenger RNAs (Iserentant and Fiers 1980; Krieg and Melton 1984) and ribosomal RNAs (Noller et al. 1981; Noller 1984). The functional complexity of DNA/RNA secondary structures is determined by their conformational dynamics during the exploration of a free-energy landscape that has the potential to lead to different stable configurations that are suited to different biological functions. Apart from change in physiological conditions such as temperature, pH, sodium and magnesium concentration, the key triggers of DNA/RNA conformational transitions are proteins known as RNA chaperones and helicases that drive structural-transitions over a large energy barrier by destabilising RNA helices and allowing the metastable RNA to take on conformations that are thermodynamically stable (Pyle and Green 1995; Treiber and Williamson 2001). Furthermore, tertiary conformational changes that do not necessary alter the secondary structure but modify the three-dimension configuration of helices allow molecules to bind and interact with disparate targets (Dethoff et al. 2012).

Besides functional elements encoded within viral genomes, such as regulatory, structural, and enzymatic proteins, and various functional motifs for transcription (Yuen and Moss 1987; Hefferon et al. 2006), translation (Shen and Miller 2004) replication (Song and Miller 2004) and genome packaging (Stockley et al. 2013), the genomes of viruses frequently encode regulatory information within their abundant stable secondary structures. Some of the structures have catalytic and regulatory functions at various stages of the viral life cycle including replication (Le et al. 1990; Powell et al. 1997; Fernandes et al. 2012), transcription (Koev et al. 1999), translation (You et al. 2004; Miller et al. 2007; Simon and Miller 2013) and post-transcriptional processing (Moss et al. 2012). Although the nature and function of

most stable genomic structures are still largely unknown, their existence in large numbers throughout all virus families, and the ranges of alternative configurations that these structures can form, constitute an important but under-explored layer of biological information encoded within virus genomes: a factor implying that the maintenance and modification of genomic secondary structures has likely been a major theme during the evolution of many viruses lineages.

1.1.1 Impact of secondary structure on virus evolution

Virus genome evolution is the process by which viral genomes change in sequence, structure or size over generations under the influence of natural biological and environmental processes that drive nucleotide substitutions (Domingo and Holland 1997), nucleotide insertions and deletions (McCullers et al. 1999), recombination (Worobey and Holmes 1999), and genome component reassortment (McCullers et al. 1999). The key characteristic of evolution is that genes and other biologically functional sequences (such as protein binding motifs and secondary structures) that provide some competitive advantage are either maintained or evolve to become more useful over time whereas non-functional or maladaptive sequences that incur a competitive disadvantage are either not maintained or are driven to extinction. As a consequence of millions of years of adaptive evolution, virus genomes that are sampled in nature today contain protein encoding genes and functional motifs that are presently mostly evolving under strong selection favouring the maintenance of their functionality: a type of selection called purifying or negative selection that disfavours genetic changes in genomic sequences that are already functionally optimal (Holmes 2003; Edwards et al. 2006; Wertheim and Kosakovsky Pond 2011). Like other functionally optimal sequence elements, the biologically important secondary structures that are present today within virus genomes have likely been selectively maintained in the face of persistently ongoing genomic mutation and recombination.

1.1.1.1 Mutation

Although mutations occurring within the non-coding regions of genomes are often assumed to be silent in that they do not code for protein amino acid sequences, they can impact regulatory elements such as functional secondary structures or protein-

binding motifs such as those located in gene promoter regions. Similarly, within protein-coding regions nucleotide substitutions that are synonymous (i.e. they do not alter an encoded amino acid) are not necessarily silent in that they might still disrupt or alter other functional sequence elements such as amino acids encoded in an overlapping reading frame (Mizokami et al. 1997), protein binding motifs (for example, genome packaging signals; Kim et al. 2011) or secondary structures (Zanini and Neher 2013), and will therefore often be selectively disfavoured. In many virus families genomic secondary structures have been associated with strong purifying selection (Simmonds and Smith 1999; Cloete et al. 2014) which has motivated the development of powerful tools aimed at identifying coding regions with excessively low synonymous substitution rates so as to pinpoint locations within genes of functional nucleic acid sequence elements such as biologically important secondary structures (Sealfon et al. 2015).

It is, however, important to point out that low nucleotide substitution rates (and hence low synonymous substitution rates) at nucleotide sites that are base-paired within a secondary structural element may occur even if the structural element in question is not biologically functional. This is because nucleotides are less prone to the types of chemical alterations that cause mutations (Lindahl and Nyberg 1974) and strand-breakage (Parthasarathi et al. 1995) when they are base-paired than when they are not base-paired. It is therefore not always entirely convincing to invoke the action of selection pressures acting against deleterious mutations that destabilise functional secondary structures as the cause of excessively low synonymous substitution rates at nucleotide sites that form base-pairs within such structures.

To further complicate matters, either the evolutionary constraints on nucleotide substitutions that are imposed by functional secondary structures, or an innate feature of structured nucleic acids (i.e. a feature that is not necessarily associated with the functional properties of the structured nucleic acids) have in some cases been associated with what appear to be mismatch repair mechanisms that result in the extremely rapid reversion of mutations that destabilised secondary structures (Cheung 2005; Shepherd et al. 2006). These mechanisms may differ from simple compensatory complementary mutations at base-pairing nucleotide sites that restore

pairing interactions (Hofacker et al. 1998; Fernández et al. 2011; Cheng et al. 2012a) in that they result in the precise reversion of the primary mutations.

The possibility that the presence of secondary structures, irrespective of their functionality, might induce low rates of nucleotide substitution means that efforts to identify functional secondary structures using nucleotide sequence data should ideally also examine sequences for other signals of natural selection favouring the preservation of these structures. For example, signal of complementary coevolution between base-paired nucleotides favouring maintenance of pairings that is detectable either computationally (Hofacker et al. 1998) or experimentally (Fernández et al. 2011; Cheng et al. 2012b), and signal of lower than expected frequency of minor allele polymorphisms at paired-sites detectable within aligned sequences using computational approaches (Tajima 1989; Fu and Li 1993), allow determining whether maintenance of structural elements has been essential during the evolutionary history of the genome and can therefore be used to investigate the biological relevance of particular structures.

1.1.1.2 Genetic recombination and genome reassortment

Besides mutation, two other important evolutionary mechanisms that are often important during virus evolution are genetic recombination (in which homologous or non-homologous fragments of genome sequence are exchanged between virus genomes), and genome reassortment (in which homologous genome components are exchanged between two multi-component virus genomes: i.e. virus genomes with multiple chromosomes or segments). These two processes allow faster evolution than could be achieved through mutation alone (Stemmer 1994; Crameri et al. 1998) and contribute to the ongoing diversification of many virus species.

It is noteworthy that, following genetic recombination and genome reassortment the survival and viability of recombinant and reassortant genomes greatly depends on conservation of important coevolved intra- and inter-molecular interactions, including those occurring between functional nucleic acid domains and the proteins that they bind to, amino acids interacting within proteins to ensure that they fold correctly, amino acids that ensure the specificity of inter-protein binding and base-paired nucleotides that ensure the stability of secondary structures. Accordingly,

recombinant viral genomes observed in nature have been noted to have predicted degrees of intra-protein amino acid interaction disruption (Lefeuvre et al. 2009a; Woo et al. 2014; Golden et al. 2014a) and functional nucleic acid interaction disruptions (Golden et al. 2014a) that are far lower than those that would be expected in the absence of negative selection disfavouring the disruption of these interactions. Such efforts to find evidence of selection disfavouring the survival of recombinants with disrupted secondary structures used sequences sampled from nature (Golden et al. 2014a) and those obtained from evolution experiments (Martin et al. 2011c) and also relied on computational approaches that detect recombination within genomes, fold individual genomes and produce *"in silico"* recombinants genomes used for a permutation test measuring the degree of structure disruption.

Furthermore, presence of secondary structures within the genome can have impact on recombination patterns. For instance, specific stable stem-loop structures are known to induce recombination in certain regions within the genome by triggering template switching of RNA-dependent DNA polymerase in HIV (Galetto et al. 2004; Galetto et al. 2006) and RNA-dependent RNA polymerase in Brome mosaic virus (BMV; Figlerowicz 2000). In addition, analysis of HIV has demonstrated that recombination frequently occurs at regions corresponding to experimentally detected secondary structures (Simon-Loriere et al. 2010).

Finally, it is worth stressing that the conservation of pervasive computationally detectable secondary structures in both the coding and non-coding genomic regions of viruses in diverse families strongly suggests, but is not definitive proof of, the fitness advantages that such structures provide. Such proof requires multiple lines of evidence that selection has actively disfavoured the disruption or removal of such structures within genetically altered virus variants (whether these alterations are caused by mutation, recombination, or re-assortment) from a wide range of different virus species.

1.1.2 Computational prediction of secondary structure

Prediction of the secondary structures of large single-stranded DNA/RNA molecules remains a complex problem in Bioinformatics due its computational complexity and the low level degree of accuracy of available prediction algorithms. Many of the most

popular computational tools that are currently widely used such as UNAFold (Markham and Zuker 2008) and RNAalifold (Bernhart et al. 2008) implement a minimum free-energy method, and predict the structural configuration of DNA/RNA molecules as the configuration with the lowest estimated free-energy.

However, due to the metastable nature of DNA/RNA structures, the free-energy landscape of various conformations of one large DNA/RNA molecule are punctuated with deep conformational wells or local minima, such that structural conformations associated with different wells that have similar local free-energy minima can vary quite substantially from one another (Dethoff et al. 2012). The flexibility of single-stranded DNA/RNA molecules under dynamic physiological conditions (where metabolite concentration, pH and temperature can vary) and the binding of these molecules to proteins such as chaperones can cause them to undergo conformational transitions (i.e. “movements” between different low-free energy wells; Herschlag et al. 1994). Since there are no automated methods that fully model the dynamics of DNA/RNA structure, existing methods of RNA/DNA structure prediction are unable to generate structural models that fully reflect a reality. By failing to predict all sets of favourable secondary structural configurations that might have functional relevance under varied environmental conditions, these methods all have room for improvement (Dethoff et al. 2012).

Another major limitation in this field is the inability to model tertiary structures of long RNA/DNA molecules. RNA/DNA 3D-folding approaches are at a much earlier stage of development, than protein folding approaches. Most methods for predicting DNA/RNA 3-D structures are limited to analysing sequences of less than 100 nucleotides (Zhao et al. 2012) and most also require manual manipulations (Laing and Schlick 2011). Although tertiary conformational changes may not alter the underlying secondary structures, they are often crucial for intermolecular interactions between the structure and other molecules or structures (Giegé 2008). For instance, mRNA riboswitches are structural regulatory elements capable of switching protein production on or off based on conformational changes that alter their tertiary structure (reviewed in Serganov and Patel 2012). A full understanding of the biological functions of many other secondary structural elements will therefore also likely require adequate knowledge of their tertiary structures.

Furthermore, most MFE-based computational methods are unable to predict complex structures known as “pseudoknots”. Pseudoknots are structures minimally composed of two helical-segments connected by base-pairing between the single-stranded regions of otherwise single-stranded “loops” or “bubbles” embedded within these helices (Staple and Butcher 2005). Many pseudoknots are known to be biologically functional and some RNA molecules that contain pseudoknots are even known to be catalytically active (Wadkins et al. 1999; Staple and Butcher 2005).

Besides these limitations, in recent years improvements in secondary structure prediction have been achieved in methods such as CoFold which accounts for co-transcriptional folding of RNA molecules (Proctor and Meyer 2013), PPfold which uses stochastic context-free grammar and phylogenetic analysis to improve predictions (Sükösd et al. 2011), RNAalifold which applies thermodynamic energy minimization approach but also accounts for covariation of nucleotides and evolutionary conservation of structures (Bernhart et al. 2008), and NASP which identifies the subset of evolutionarily conserved base-pairs that are responsible for real DNA/RNA molecules having lower MFEs than randomly generated DNA/RNA molecules (Semegni et al. 2011). While most DNA/RNA folding methods use one sequence to predict secondary structures, some are capable of predicting the consensus structure of a group of homologous sequences. Besides enabling the detection of structure conservation across many lineages, such methods also enable the use of other biological information, such as nucleotide covariation between potentially base-paired sites. Among these, NASP (the structure prediction method that is widely employed in this project), uses an alignment of related sequences as input and predicts their consensus secondary structure using hybrid-ss-min a component of UNAFold (Markham and Zuker 2008). NASP also statistically determines the subset of individual elements that contributes to the actual analysed sequences displaying a degree of structural stability that is significantly greater than that seen in randomly shuffled sets of sequences with the same nucleotide (or dinucleotide) composition (Semegni et al., 2011).

However, it must be stressed that even though most of the evolutionary conserved structural elements that are computationally predicted by methods such as NASP most likely really do exist (and might therefore have some biological relevance),

these methods still cannot adequately describe the entire ensemble of (potentially very biologically important) alternative structural configurations that RNA/DNA molecules likely form under physiological conditions. Even for those structural features evident within the relatively static snapshots (some of the methods do provide a few potential alternative structures) of RNA/DNA molecules that these methods provide, they are unable to provide definitive proof of functional relevance. However, complementary to the analytical barriers of comparative folding tools such as NASP, are experimental structural identification and analysis approaches (discussed in the next section) which can verify the actual existence and/or the biological relevance of particular structures within small numbers of genomes, and selection-based computational approaches for analysis of homologous sequences (discussed in the previous section) that can verify the functional significance (and hence the likely existence) of large numbers of structural elements in large numbers of related genomes.

1.1.3 Experimental prediction of secondary structure

Given the difficulty of accurately predicting secondary structures, data from experimental methods play an indispensable role in the verification of computationally predicted structure models. One popular experimental method is termed “Selective 2’-hydroxyl acylation analysed by primer extension” (SHAPE; Wilkinson et al. 2006). SHAPE determines whether nucleotides within a RNA molecule are base-paired or not based on the chemical modification of unpaired nucleotides within RNA molecules under approximately physiological conditions. Modifications are identified as stops during a primer extension reaction with reverse transcriptase and are 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 base-paired, display low SHAPE reactivity, whereas unpaired sites show higher degrees of chemical modification and thus, high SHAPE reactivity (Wilkinson et al. 2006). SHAPE data are then used to correct and substantially improve secondary structure predictions made by MFE-based folding methods (Deigan et al. 2009). This method has been used to predict the full genome secondary structures of several RNA viruses including *Human immunodeficiency*

virus (HIV; Watts et al. 2009), *Simian immunodeficiency virus* (SIV; Pollom et al. 2013) and *Hepatitis C virus* (HCV; Mauger et al. 2015).

It is noteworthy that although the SHAPE method is very accurate in determining paired bases, it does not determine their pairing partners and the results can therefore be subjective, based on the performance of the structure model used. Thus in most circumstances, it is worthwhile to carryout additional analyses that validate the structure obtained with aid of SHAPE. For example, methods detecting selection and coevolution have been used to obtain independent support for SHAPE derived structured regions within HCV (Mauger et al. 2015) and *Dengue virus 2* (DENV2; manuscript in preparation) genomes. Besides confirming the likely existence of particular SHAPE predicted structural elements, these methods give evidence that such structures are evolving in a manner consistent with selection favouring their preservation and hence are likely biologically relevant.

Furthermore, despite the complexity of tertiary structure prediction, the accuracy of most 3D RNA structure prediction methods can be greatly improved if they are guided (i.e. the solution-space is constrained) by evolutionary, biochemical or biophysical data (Magnus et al. 2014). For example, existing 3D structure determination methods that utilise information derived from known structures of other RNA molecules largely outperform methods that are solely based on the physical laws governing the folding process (Magnus et al. 2014). Currently, purely theoretical computational methods of 3D structure including DMD (Ding et al. 2008), FARNA (Das and Baker 2007), SimRNA (Sripakdeevong et al. 2012), and V-Fold (Cao and Chen 2009) can only be used with high confidence for molecules up to 40 nt and with moderate confidence for molecule up to 80 nt (Magnus et al. 2014). Increasing the lengths of analysed sequences complicates the problem further and the accuracy of these methods decreases dramatically. In this case reasonable prediction of long-range interactions is achieved with aid of existing experimentally derived structural data. Although RNA 3D structure prediction is expensive and complex, the field is continually moving forward with new experimentally determined/verified 3D structures consistently being added to public databases each year (Xin et al. 2008; Chojnowski et al. 2014).

Besides improving structure prediction or 3D structure models, experimental methods are also valuable for examining the biological impact of individual structural elements during the viral infectious cycle. Many of these methods consist of mutagenesis assays that engineer mutations within the genome to partially or completely destabilise potential biologically functional structural elements. The biological relevance of the structure is evident if mutant and wild-type genomes display significant differences in infectivity, replication, gene expression, transmission or any other measurable biological characteristic. Examples of the experimentally detected relevance of particular structural elements include that of a stem-loop at the virion strand origin of replication in *Porcine circovirus 1* (Cheung 2004a; Cheung 2005), the small structural element near the replication associated protein gene intron of *Maize streak virus* (Shepherd et al. 2006), and structure involved in the trans-activation of genes in simian retroviruses (Tabernero et al. 1996).

1.2 Eukaryote-infecting single-stranded DNA viruses

Single-stranded DNA (ssDNA) viruses are among the smallest viruses on earth in terms of their genome lengths, particle sizes and numbers of genes. Their genomes range in size from ~2 kb to ~8 kb and have either linear or circular configurations (Fauquet 2006). Currently, the International Committee on Virus Taxonomy (ICTV; http://www.ictvonline.org), an international scientific community involved in virus classification, recognises seven ssDNA virus families including five infecting eukaryotes (*Circoviridae*, *Anelloviridae*, *Parvoviridae*, *Nanoviridae* and *Geminiviridae*) and two infecting prokaryotes (*Microviridae* and *Inoviridae*). Besides these families, a diversity of unclassified replicons and plasmids resembling ssDNA viruses has been sampled around the world from terrestrial and aquatic environments (Rosario et al. 2009; Blinkova et al. 2009; Blinkova et al. 2010; Kraberger et al. 2014), and their enormous degree of diversity suggests that the host ranges of these viruses/virus-like elements could span the tree of life and have major impacts on global ecosystems (Martin et al. 2011a).

Although many ssDNA viruses infecting humans and other animals do not generally cause major diseases, *Beak and feather disease virus* and *Porcine circovirus* cause

deadly diseases to parrots and swine respectively, and they are major problems in parrot and porcine industries (Alarcon et al. 2013; Harkins et al. 2014). Similarly, plant-infecting ssDNA viruses transmitted by insects cause economically and socially important diseases to plants in various parts of the world. For example geminiviruses and nanoviruses infect and cause serious yield losses in a wide range of food crops including maize potato, bananas, beans and cassava (McKenzie et al. 2002; Boulton 2003; Kumar et al. 2011).

It is noteworthy that despite their immense diversity ssDNA viruses generally share strikingly similar characteristics including genome replication mechanisms and genome configurations (linear or circular). However, examples of some other distinctive characteristics which differentiate them include the number of genome components (one, two, or multiple components), numbers of genes, gene organisation, tissue tropisms and host-ranges.

1.2.1 Diversity of eukaryote-infecting ssDNA viruses

Currently, the ICTV classification of eukaryote-infecting ssDNA viruses consists of five virus families containing 24 genera and 425 species (Figure 1-1).

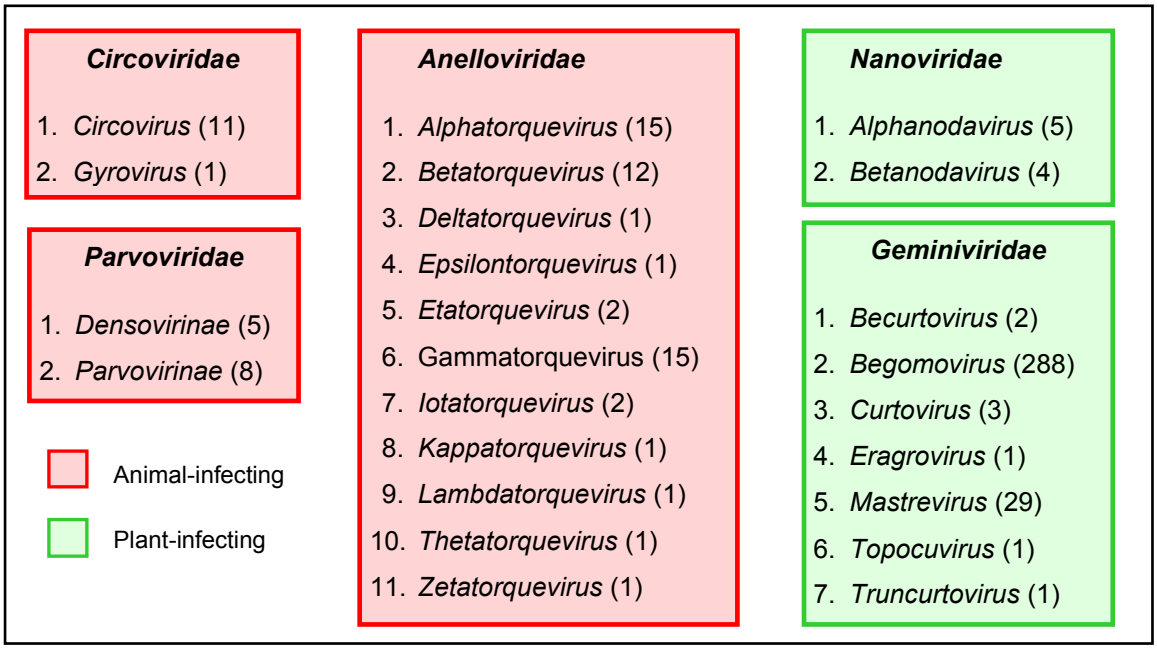


Figure 1-1. Virus genera within eukaryote-infecting ssDNA virus families

Animal and plant infecting viral genera (subfamilies in the case of *Parvoviridae*) are enclosed in red and green shaded boxes, respectively. Family names are written in bold at the top of each box, while the number of species within each genus is shown in brackets. There are 24 taxonomically

recognised genera/subfamilies and 425 named species within the group of eukaryote-infecting ssDNA viruses. (Source: http://www.ictvonline.org).

Besides that all these viruses consist of small single-stranded DNA molecules taking a circular or linear form, they have some common characteristics that are unique to ssDNA viruses. For instance the rolling-circle replication (RCR) mechanism is shared by most families of viruses with circular genomes including the *Circoviridae* (Cheung 2006), *Geminiviridae* (Gutierrez 1999) and *Nanoviridae* (Gronenborn 2004), while a variant of RCR, called rolling-hairpin replication (RHR), is common to those with linear genomes in the family *Parvoviridae* (Berns 1990). Also, the functional motifs involved during replication are particularly conserved within and across families of circular and linear genomes. In addition, there are similarities in genome organisation and layout amongst viruses sharing the same particle structure and host type. Specifically, while plant-infecting ssDNA viruses with circular genomes all encode at least three proteins (a replication associated protein, a capsid protein and a movement protein), those infecting animals generally encode at least two proteins (a replication associated protein and a capsid protein).

1.2.2 Characterised genomic secondary structures

Only a few structures have been well-characterised in ssDNA viruses and these are mainly involved in the modulation of replication and transcription.

A conserved structural element at the origin of replication of circular ssDNA viruses is a common characteristic in most families (Orozco and Hanley-Bowdoin 1996; Hafner et al. 1997; Steinfeldt et al. 2001; Cheung 2006). Although, the sequence of this structure varies from family to family, a “AxTAxT↓AC” motif in the loop sequence that flanks the virion strand origin of replication nicking site (indicated by an arrow) is strikingly conserved among geminiviruses (Heyraud et al. 1993; Fauquet 2006), circoviruses (Cheung 2006), nanoviruses (Timchenko et al. 1999) and an enormous diversity of uncharacterised ssDNA-like viruses that are found in the environment. Similarly, parvoviruses, which have linear genomes, contain a conserved T-shaped structural element at their origins of replication (Ashktorab and Srivastava 1989; Cossons et al. 1996; Sun et al. 2009). Although this structure might have divergent sequences in different species, its stability and conformation is strikingly conserved.

Another important structure located approximately 140-380 nt downstream the promoter region P4, regulates translation in parvoviruses (Ben-Asher and Aloni 1984; Resnekov and Aloni 1989; Krauskopf et al. 1991; Perros et al. 1994). However, there are no reports yet that have investigated the existence of such regulatory elements in other circular ssDNA viruses.

Apart from these few characterised structures, it is still unknown how pervasive and biologically relevant secondary structures are within the genomes of eukaryote-infecting ssDNA viruses and whether the numerous predicted uncharacterised structures have any regulatory functions during the viral replication cycle in, for example, complementary sense replication, genome packaging, gene splicing, genome movement (for plant viruses).

Historically, structural biologists and bioinformatics researchers have focused primarily on characterising human infecting RNA viruses such as HIV, DENV, HCV and Poliovirus. Very little work has been done on the prediction and characterisation of structural elements in ssDNA viruses. The identification and enumeration of biologically functional secondary structural elements within the genomes of viruses in this group would provide a foundation for future studies seeking to investigate the actual biological functions of specific secondary structural elements within ssDNA virus genomes.

1.2.3 Evolution of eukaryote-infecting ssDNA viruses

DNA viruses are known to have lower mutation rates than RNA viruses, mainly because DNA polymerases tend to be less error prone than either RNA-depended RNA polymerases (which are used during the replication of plus stranded, negative stranded and double-stranded RNA viruses) or RNA-depended DNA polymerases (also called reverse transcriptases; which are used during the replication of retroviruses; Duffy et al. 2008). This is because DNA polymerases generally have error correcting mechanisms whereas RNA polymerases generally do not (Garcia-Diaz and Bebenek 2007). Even though ssDNA virus generally mutate much slower (~ 10^{-6} – 10^{-5} mutations/site/replication) than ssRNA viruses (~10^{-5}-10^{-3} mutations/site/replication), the substitution rates for both groups (i.e. the rate at which arising mutations become part of the pool of genetic diversity within virus

populations) tends to be similar for both groups (~ 10^{-5} - 10^{-3} substitution/site/year; Shackelton et al. 2005; Shackelton and Holmes 2006; Garcia-Diaz and Bebenek 2007).

Other major evolutionary mechanisms, through which genetic diversity arises in ssDNA viruses, are genetic recombination and genome component reassortment. High frequencies of recombination have been detected within most families of eukaryote-infecting ssDNA viruses: for example in geminiviruses (Martin et al. 2011c), circoviruses (Ma et al. 2007), anelloviruses (Leppik et al. 2007), parvoviruses (Shackelton et al. 2007). Similarly, high frequencies of genome component reassortment have been detected within both the nanoviruses (Grigoras et al. 2014) and begomoviruses (the only two ssDNA virus groups that are known to have genomes with multiple components; Idris et al. 2008).

A comparative analysis of recombination breakpoints distribution within the genomes of ssDNA viruses has demonstrated that recombination events are non-random, as breakpoint hotspots tend to occur outsides or toward the ends of genes (Lefeuvre et al. 2009a). Furthermore, analysis on begomovirus genomes revealed that recombination events that occur within genes tend to be less disruptive of protein folding interactions than could be achieved if recombination breakpoints were randomly distributed (Lefeuvre et al. 2007). These findings suggest that selection acting against virus genomes expressing improperly folded chimeric proteins is one of the major factors driving the evolution of ssDNA viruses.

Furthermore, nucleic acid secondary and tertiary structures that form within the genomes of ssDNA viruses may be crucial for their survival and may therefore place strong constraints on the evolution of these viruses. While several lines of inquiry have extensively investigated and characterised various genomic secondary structures in RNA viruses, in ssDNA viruses genomes secondary structures have not been very well studied and their pervasiveness and overall impacts on ssDNA virus evolution remains a mystery. Specifically, the presence of biologically functional secondary structures within ssDNA genomes could have a range of potential effects including: (1) influencing the survival of mutants such as is seen for RNA viruses (Simmonds and Smith 1999; Garcia-Diaz and Bebenek 2007; Cloete et al. 2014), (2) influencing where mutation events are most likely to occur as has been seen in RNA

viruses (Simmonds and Smith 1999; Cloete et al. 2014) and bacteria and eukaryotes (Gu et al. 2014), (3) influencing where recombination events are most likely to occur, as has been noted in HIV (Galetto et al. 2006; Simon-Loriere et al. 2010) and BMV (Figlerowicz 2000), and (4) influencing which recombinants are most likely to survive as has also been noted in HIV (Golden et al. 2014a).

Therefore, a major hypothesis of this study is that the ssDNA virus genomes that have been able to replicate and survive in nature are those that have maintained important base-pairing interactions within biologically functional genomic DNA secondary structural elements. As such, the main objective of this research is to implement and apply an array of comparative sequence analysis techniques to determine whether the patterns of natural selection that are evident within ssDNA virus genome sequences are consistent with the presence of pervasive biologically functional secondary structural elements within these genomes.

1.3 Thesis structure

The remainder of this thesis consists of five chapters. Chapter 2 describes a computational tool that I developed both for the classification of virus genome/protein sequences into operational taxonomic units (OTUs) and for the creation of sequence datasets suitable for the types of molecular evolution analyses that I perform in the remainder of the thesis. Chapter 3 presents a variety of computational tools I developed to predict the likely biological functionality of secondary structural elements that are evident within virus genomes. Chapter 4 is a study that applies the computational tools discussed in Chapter 3 to the prediction and characterisation of biologically functional secondary structures within the genomes of eukaryote-infecting ssDNA viruses. Chapter 5 is a study that uses computational methods discussed in Chapter 3 to investigate the impact of genomic secondary structure on recombination in ssDNA viruses and to detect evidence of selection favouring the maintenance within these viruses of functional secondary structures following genetic recombination. Chapter 6 serves as a conclusion chapter that ties together the various themes covered in the thesis.

All chapters in this thesis other than Chapters 1 and 6, are presented as research papers with each consisting of abstract, introduction, materials and methods, results

and discussion, and conclusion sections. At the end of each chapter, my contributions and those of my co-authors are outlined and appropriate acknowledgements are made.

Chapter 2 : Sequence Demarcation Tool (SDT): a tool for objective classification of virus genomes

2.1 Abstract

The perpetually increasing rate at which viral full-genome sequences are being determined is creating a pressing demand for computational tools that will aid the objective classification of these genome sequences. Taxonomic classification approaches that are based on pairwise genetic identity measures are potentially highly automatable and are progressively gaining favour with the International Committee on Taxonomy of Viruses (ICTV). There are, however, various issues with the calculation of such measures that could potentially undermine the accuracy and consistency with which they can be applied to virus classification. Firstly, pairwise sequence identities computed based on multiple sequence alignments rather than on multiple independent pairwise alignments can lead to the deflation of identity scores with increasing dataset sizes. Also, when gap-characters need to be introduced during sequence alignments to account for insertions and deletions, methodological variations in the way that these characters are introduced and handled during pairwise genetic identity calculations can cause high degrees of inconsistency in the way that different methods classify the same sets of sequences. Here we present Sequence Demarcation Tool (SDT), a free user-friendly computer program that aims to provide a robust and highly reproducible means of objectively using pairwise genetic identity calculations to classify any set of nucleotide or amino acid sequences. SDT can produce publication quality pairwise identity plots and colour-coded distance matrices to further aid the classification of sequences according to ICTV approved taxonomic demarcation criteria. Besides a graphical interface version of the program for Windows computers, command-line versions of the program are available for a variety of different operating systems (including a parallel version for cluster computing platforms).

2.2 Introduction

The ever advancing rate at which novel viral genomes are being determined is creating a serious challenge both for taxonomists seeking to ensure the consistent and accurate classification of these genomes, and for laboratory virologists attempting to accurately name newly determined genome sequences prior to deposition into public sequence databases. Given that in many cases the only taxonomically useful information that is available for a particular genome sequence is the sequence data itself, the use of pairwise nucleotide sequence identity measures is becoming increasingly popular as a means of objectively classifying bacteria (Kim et al. 2014) and viruses (Bao et al. 2012; Muhire et al. 2013) into consistent and practically useful operational taxonomic units (OTUs) such as variants, strains, species or genera. In the case of many viruses which have small genomes (< 30 kb long), whole genome sequences can be efficiently aligned, and genome-wide pairwise sequence identity scores are therefore used routinely for their functional classification. Accordingly, for over 50% of currently known virus families, the International Committee on Taxonomy of Viruses (ICTV) has, amongst other phylogenetic and biological criteria, endorsed the use of genome-wide nucleotide or amino acid sequence identity thresholds for the classification of novel virus isolates (according to ICTV proposals published since the 8th ICTV Report; http://ictvonline.org/).

Despite the obvious appeal of using genetic identity scores between pairs of sequences to objectively classify these sequences, there is frequently a lack of clarity on exactly how such scores should be calculated. For example, given a new virus sequence and the desire to classify it based on an established ICTV approved species demarcation threshold, there are many different ways in which a researcher might determine whether or not it should be included within an already established species. Computer programs such as MUSCLE (Edgar 2004), CLUSTALW (Larkin et al. 2007), MAFFT (Katoh and Standley 2013) or Basic Local Alignment Search Tool (BLAST; Altschul et al. 1990) could be used to make either multiple individual pairwise sequence alignments or a single multiple sequence alignment and other programs such as MEGA5 (Tamura et al. 2011), PHYLIP (Felsenstein 1989), PAUP (Swofford 2002) or GENEIOUS (http://www.geneious.com/) could be used to

calculate a variety of different pairwise identity scores. Unsurprisingly, for a given pair of sequences, different combinations of alignment and pairwise identity calculation approaches will in many cases yield a fairly broad range of possible sequence identity scores.

Whereas different alignment methodologies will very frequently infer different patterns of insertions and deletions (indels) during the evolutionary histories of any particular pair of sequences (Thompson et al. 1999; Katoh et al. 2005; Wilm et al. 2006), independent pairwise alignments of sequences will tend to yield higher pairwise identity scores than those calculated for the same pairs of sequences within the context of multiple sequence alignments (Katoh et al. 2005; Sievers et al. 2013). Also, when calculating pairwise identity scores between any particular pair of sequences, the way in which indels are treated can have a very substantial impact on the identity scores that are calculated. Specifically, indel characters (usually "-") that were inserted during multiple or pairwise sequence alignment might either be ignored or treated as a fifth character state. If indels are treated as a fifth character state then sites where both of the sequences being compared have indel characters might either be ignored or treated as matches (in which case they will inflate identity scores). Conversely, if sites where one but not the other sequence has an indel character are treated as mismatches the calculated identity scores will be lower than if such sites were ignored.

Particularly pertinent in the context of ever-increasing sequence database sizes is the fact that for any given pair of sequences, the differences between all these various alignment and identity score calculation approaches are expected to increase as the number of sequences that are being compared increases. This is because the computational complexity of accurately aligning multiple sequences increases exponentially with the number of sequences being aligned (Elias 2006). Put simply what this means is that as sequence numbers get larger multiple sequence alignments will tend to become more inaccurate. Although correction of alignments by eye is generally recommended for small datasets, it is not a practical option for datasets containing hundreds of sequences drawn from multiple different virus species. Alignment by eye is particularly undesirable in the context of taxonomic classification as it is both time-consuming and has the potential to

seriously undermine the objectivity and consistency with which sequences are classified.

The pairwise identity calculation approaches that will be least impacted by these problems are those relying exclusively on independent pairwise alignments. Besides being unaffected by dataset sizes, pairwise alignment is computationally tractable: i.e. given a specified set of rules for penalising mismatches and inserting gap characters, the optimal pairwise alignment can always be found in a reasonable time (Needleman and Wunsch 1970). Pairwise alignments also lack sites where both sequences have indel characters and are therefore far less affected by how indel characters are treated during identity score calculations. When calculating the identity scores of pairwise aligned sequences, the issue of gap character handling can be even further minimised by simply ignoring all sites at which a gap character is present in either one of the sequences being compared: an approach commonly adopted when calculating evolutionary distances in the context of phylogenetic tree construction (Felsenstein 2004; Lemey et al. 2009).

The demand for computational tools that will expedite the consistent and accurate classification of the increasing numbers of complete virus genomes deposited in public databases each year has prompted the development of computer programs such as PAirwise Sequence Comparison (PASC; Bao et al. 2012), and DivErsity pArtitioning by hieRarchical Clustering (DEmARC; Lauber and Gorbalenya 2012). Besides providing a means for virologists to accurately classify novel virus genome sequences at the species level prior to their publication, these tools have been especially useful both in the refinement of taxonomic classification criteria and for updating the classifications of hundreds of virus genome sequences that have been deposed in publically accessible sequence databases over the past three decades (Matthijnssens et al. 2008; Bao et al. 2012; Muhire et al. 2013; Fiallo-Olivé et al. 2014).

PASC, the most widely used of these programs, is a web-based tool developed by the National Centre for Biotechnology Information (NCBI; Bao et al. 2012). Given a novel virus genome sequence, PASC compares this to a defined set of publicly available sequences and then uses either BLAST (Altschul et al. 1990) similarity scores or Needleman-Wunsch (NW; Needleman and Wunsch 1970) pairwise-

alignment based genetic identity scores to generate frequency distributions of pairwise genetic identity scores (based on both the input and database sequences). The output can then be used to either classify the input sequence or manually identify taxonomically optimal pairwise identity-based species or genus demarcation thresholds.

Rather than focusing on pairwise identity scores determined from multiple sequence alignments, DEmARC utilises multiple sequence alignments and model-based pairwise evolutionary distance calculations that ignore indel sites. In this regard, DEmARC is perhaps better suited to the objective identification of phylogenetically supported taxonomic demarcation criteria than for use by general virologists for the classification of new sequences based on pairwise identity-based classification criteria. It is also worth noting that while applicable to the analysis of nucleotide sequence data, DEmARC was specifically designed for the analysis of conserved amino acid sequence domains: an intended application that would substantially diminish alignment accuracy issues.

While both PASC and DEmARC are potentially very useful for the establishment of objective classification criteria and the refinement of existing virus classifications, in our opinion neither of the approaches is ideally suited for use by general virologists seeking to accurately and consistently classify the novel virus genomes that they sequence into either established ICTV approved species or strains or other functionally useful OTUs. Whereas DEmARC demands the analysis of carefully constructed and edited multiple sequence alignments, PASC forces users to scan a novel sequence against a representative selection of related sequences that is generally tailored specifically to classify genomes only down to the species level (i.e., the list of sequences in many cases excludes sequences that might be of interest for making strain, variant or other higher resolution OTU classifications). PASC also relies entirely on analysing sequences in the configuration in which they were submitted to the public sequence databases. This is particularly problematic because the NW pairwise alignment method implemented in PASC encounters difficulties when circular genome sequences have been deposited with inconsistent starting and ending coordinates. The developers of PASC have therefore recommended the use of a BLAST-based alignment comparison approach that is much less affected by this

issue (Bao et al. 2012). However, from a viral taxonomic classification perspective, there remains a potentially serious consistency issue when it comes to using BLAST scores instead of NW alignment-based pairwise identity scores. Specifically, in a given dataset containing both closely related and distantly related genome sequences, whereas BLAST similarity scores between the closely related sequences might be calculated across the entire genome length, the BLAST similarity scores for the more distantly related sequences may only be calculated across the portions of the sequences that are most conserved. Besides this consistency issue, there is also no obvious way to translate BLAST scores into genome-wide pairwise identity scores: i.e. the intuitively obvious measure of genome-wide similarity that is generally used by the ICTV in their classification guidelines and is overwhelmingly preferred by general virologists when describing the genetic relatedness of virus isolates.

Here we present Sequence Demarcation Tool for Windows (SDT version 1.2; www.cbio.uct.ac.za/SDT), a stand-alone program with a simple user friendly graphical interface. Rather than being targeted at hard-core virus taxonomists, SDT is specifically targeted at laboratory and field virologists wanting to rapidly and consistently use the pairwise identity-based ICTV taxonomic guidelines to tentatively classify new viral genome sequences. The program has, however, also been recently used for the reclassification of viruses in the family *Geminiviridae* (Razavinejad et al. 2013; Gharouni Kardani et al. 2013; Muhire et al. 2013; Kanakala et al. 2013; Manzoor et al. 2013; Oluwafemi et al. 2014; Fiallo-Olivé et al. 2014; Varsani et al. 2014b; Du et al. 2014; Paz-Carrasco et al. 2014; Varsani et al. 2014c), in the classification of viruses in the families *Circoviridae* (Stenzel et al. 2014) and *Nanoviridae* (Grigoras et al. 2014), the characterisation of novel highly divergent viral genomes sampled during metagenomic surveys (Dayaram et al. 2013; Sikorski et al. 2013), and the comparison of protein sequence similarities in already characterised viruses species (Phelps et al. 2014) and novel viruses (Piasecki et al. 2013; Dayaram et al. 2013; Bernardo et al. 2013; Ge et al. 2013; Sikorski et al. 2013; Du et al. 2013; Dayaram et al. 2014; Varsani et al. 2014a). SDT is functionally similar to PASC in that it objectively applies a robust NW-based pairwise alignment approach with a pairwise identity calculation that ignores alignment positions containing indel characters. The primary differences between SDT and PASC are that: (1) it is not

restricted to using predefined sets of sequences, (2) it is geared specifically to the objective taxonomic classification of new virus sequences within the context of ICTV endorsed pairwise identity based strain, species and genus demarcation thresholds, and (3) it can produce publication quality colour coded pairwise-identity matrices with sequences ordered according to their degrees of phylogenetic relatedness. We also provide both command-line versions of SDT for Linux (SDT_Linux) and MacOS (SDT_MacOS), and a parallel Message Passing Interface based version for Linux (SDTMPI_Linux) that can be used on high performance computing clusters.

2.3 Materials and methods

2.3.1 Implementation of SDT

A graphical user interface for SDT (available at www.cbio.uct.ac.za/SDT), is implemented in Visual Basic 6.0 and runs on Windows XP, 7 and 8. Command-line versions of SDT, SDT_Linux and SDT_MacOS and a parallel version, SDTMPI_Linux are provided for both 32 and 64 bit operating systems and are all written in Python. While SDT has a graphical user interface that is complete with data visualisation tools, the command-line versions only produce numerical data. However, all these versions apply the same sequence identity calculation procedures.

2.3.2 Sequence identity calculation

Given an input FASTA file, SDT aligns every unique pair of sequences (S sequences yield [S x (S-1)]/2 alignments) using the NW algorithms implemented in MUSCLE (Edgar 2004), ClustalW (Larkin et al. 2007) or MAFFT (Katoh and Standley 2013) (the user can choose whichever program he/she prefers), and computes the identity score for each pair of sequences as 1-M/N, where M is the number of mismatched nucleotides and N is the total number of columns along the alignment where neither sequence has a gap character. The program then uses the NEIGHBOR component of PHYLIP (Felsenstein 1989) to generate a rooted neighbour-joining phylogenetic tree of sequences according to which computed scores are rearranged so as to order sequences according to their likely degrees of evolutionary relatedness. Finally, SDT generates a frequency distribution of pairwise-identities. The graphical

program interface (Figure 2-1) provides both publication quality visualisations of results and enables results to be saved in a variety of graphical and numerical data file formats.

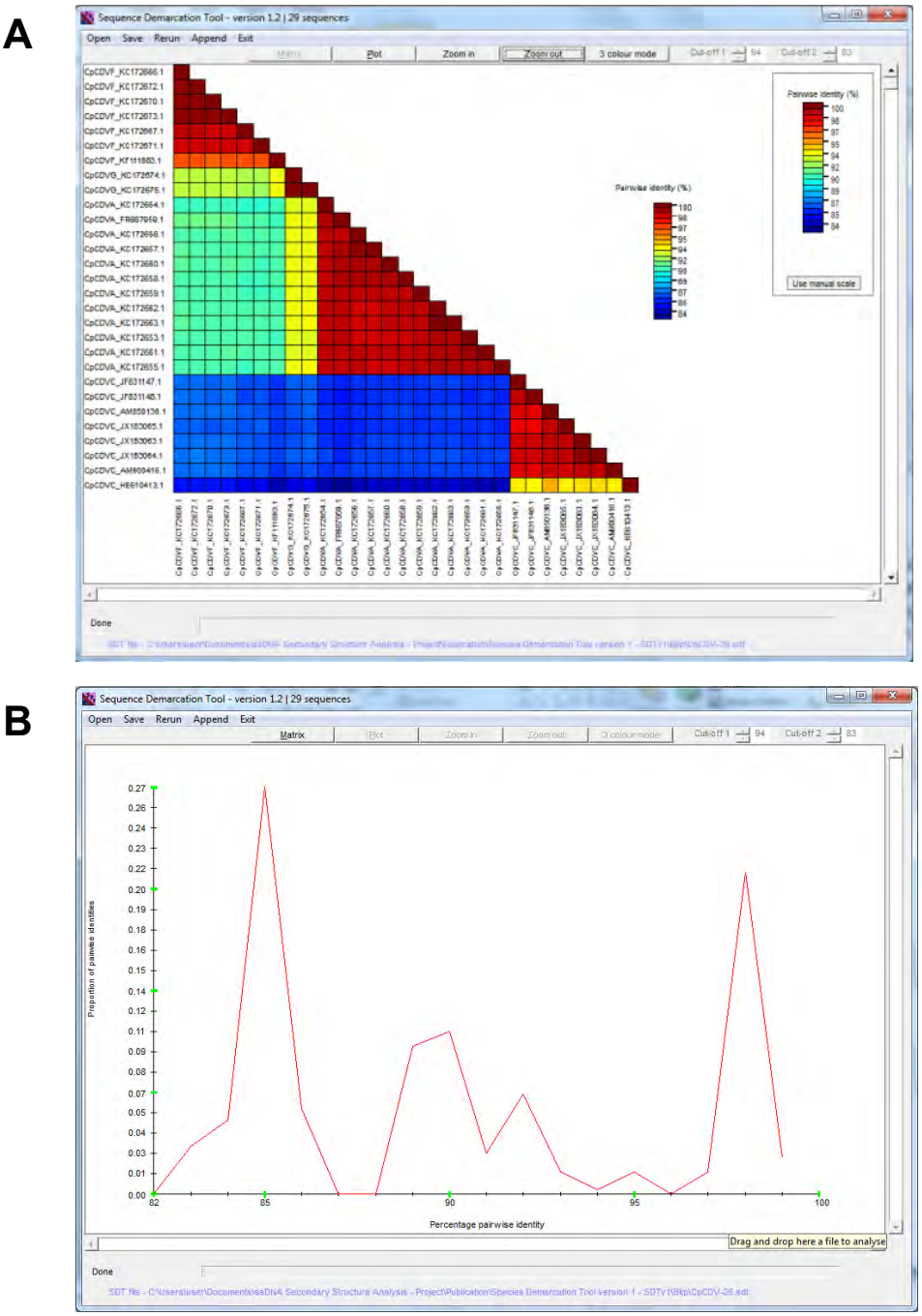


Figure 2-1. The SDT interface

(A) Colour-coded pairwise identity matrix generated from 29 *Chickpea chlorotic dwarf virus* genomes. Each coloured cell represents a percentage identity score between two sequences (one indicated horizontally to the left and the other vertically at the bottom). A coloured key indicates the correspondence between pairwise identities and the colours displayed in the matrix. **(B)** Pairwise identity frequency distribution plot. The horizontal axis indicates percentage pairwise identities, and

the vertical axis indicates proportions of these identities within the distribution. While peaks on the graph indicate pairwise sequence identity thresholds that would yield the most ambiguous classifications, troughs indicate thresholds that would yield the least ambiguous classifications and could therefore be tentatively used as relatively conflict free operational taxonomic unit demarcation cut-offs.

2.3.3 Pairwise identity matrix and pairwise identity distribution plots

SDT displays pairwise identity scores using a colour-coded matrix (Figure 2-1 A) which provides more intuitively accessible insights into the overall relationships between sequences in a dataset than the tables of pairwise sequence identity scores that are widely used for this purpose. The colours in this matrix can be adjusted to reflect, for example, an ICTV species demarcation criterion such that identities between sequences that are over the threshold are represented in a shade of one colour whereas those that fall below the threshold are represented in a shade of a different colour. The ordering of sequences along the axes of the matrix reflects the ordering of the sequences as they would appear in a neighbour joining phylogenetic tree: i.e. the pairwise identities between sequences are clustered within the matrix in an evolutionarily meaningful way. This makes it very easy to check exactly which groups of sequences a novel sequence is most closely related to and, depending on the colours of the cells in the matrix, immediately indicates which genus, species, strain or other OTU it could most appropriately be assigned to. For a detailed example of how SDT pairwise identity matrices can be applied to the classification of novel virus genomes please refer to (Muhire et al. 2013)and (Varsani et al. 2014b).

2.3.4 Usage of pre-computed identity scores

When the computations are finalised, all versions of SDT allow a completed analysis session to be saved to a SDT readable file (with file extension ".sdt") which subsequently can be reloaded. Upon reloading such a file in SDT, the program allows the addition of new sequences and then only computes scores for those sequence pairs that include the newly added sequences. Doing this vastly speeds up the analysis of new sequences and allows a user to very efficiently grow the size of project specific datasets.

2.3.5 Creation of datasets based on sequence identities

Given a set of input sequences and their corresponding pairwise sequence identity scores it is possible for SDT to objectively generate datasets comprising sequences of a desired level of diversity/identity that are tailored to further genome evolution analyses such as inference of patterns of natural selection or the identification of conserved genomic secondary structures (Muhire et al. 2014a; Stenzel et al. 2014). Once sequence identity scores are computed, SDT provides an efficient way to generate such datasets. All that is required of the user is to indicate a minimum and a maximum identity percentage and the program will then partition the input sequence dataset into sets of non-overlapping sequence files, with each file containing only sequence pairs with identities that are within the user specified range.

2.3.6 The SDT_Linux, SDT_MacOS and SDTMPI_Linux command line versions

The Python coded command-line versions of SDT for Linux, MacOS and high performance computing clusters are ideal for inclusion within automated sequence classification pipelines. These versions apply precisely the same sequence identity calculation approach as SDT but only generate pairwise identity scores in various comma separated value (CSV) text formats. Although there is no graphical output from these versions, the CSV files that are generated are formatted such that a colour coded pairwise identity matrix and distribution plot can easily be constructed using the R programming language (www.r-project.org) or MATLAB (http://www.mathworks.com/products/matlab/). Also, the .sdt formatted files that are generated by these versions of the program can be opened in the graphical interface version of the program to produce colour-coded distance matrices and pairwise identity plots. Whereas the SDT_Linux, SDT_MacOS and SDTMPI_Linux versions all require that Python (available from https://www.python.org) be installed on the machines on which they are run, the SDTMPI_Linux version additionally requires the installation of the Python Message Passing Interface library (MPI4PY; available at http://mpi4py.scipy.org/docs/usrman/install.html).

2.3.7 Comparison of SDT performance with alternative sequence comparison methods

For an objective comparison of SDT's consistency with that of alternative pairwise sequence comparison methods, we used SDT and DEmARC to analyse the same set of 25 mastrevirus full genome sequences within the context of progressively increasing dataset sizes. Although it was not possible to run this test with PASC (due to the stringent sequence input requirements of this program), it is anticipated that PASC would have exactly the same degree of consistency as SDT (it too relies on pairwise sequence alignments). A dataset of 400 mastrevirus full genome sequences (Supplementary Dataset 1), was progressively subdivided to generate five sub-alignments of 200, 100, 50 and 25 sequences, all containing the same set of 25 sequences. These were all analysed unaligned by SDT which produced pairwise identity scores for each of the 300 pairwise comparisons between the 25 sequences common to all five datasets. The identity scores once produced were converted to Hamming distances by subtracting them from one (so as to enable a more direct comparison with DEmARC). After aligning each individual dataset using MUSCLE (with default settings), Hamming genetic distances and DEmARC evolutionary distances were calculated for each of the same 300 pairwise sequence comparisons in each of the five alignments.

2.3.8 Comparison of parallel and serial versions of SDT

We analysed 1000 publically available begomovirus sequences (Supplementary Dataset 2); requiring 499,500 pairwise sequence alignments of ~2800nts; Table 2-1) with 32 and 64 bit versions of SDTMPI_Linux and SDT_Linux using MUSCLE to perform pairwise alignments. The 32 and 64 bit versions of SDT_Linux were run on a 2.8 GHz computer with 24 GB of RAM (with this 32 bit version by definition being restricted to using < 2 GB of RAM), and the 32 bit and 64 bit versions of SDTMPI_Linux were run on 20 or 40 cores each running at 2.8 GHz with 24 GB of RAM (again with the 32 bit version being restricted to using < 2 GB of RAM).

2.4 Results and discussion

2.4.1 The consistency of SDT relative to alternative virus classification tools

Although all of the pairwise comparison methods produced very similar results for sequences sharing between 90 and 100% pairwise identity, distinct differences between the methods were clearly observable in all datasets for sequence pairs sharing less than 80% identity (Figure 2-2). This observation is expected since sequence alignment only becomes non-trivial (and hence more error prone) when some of the sequences being aligned have accumulated multiple nucleotide substitution, insertion and deletion mutations since their most recent common ancestors.

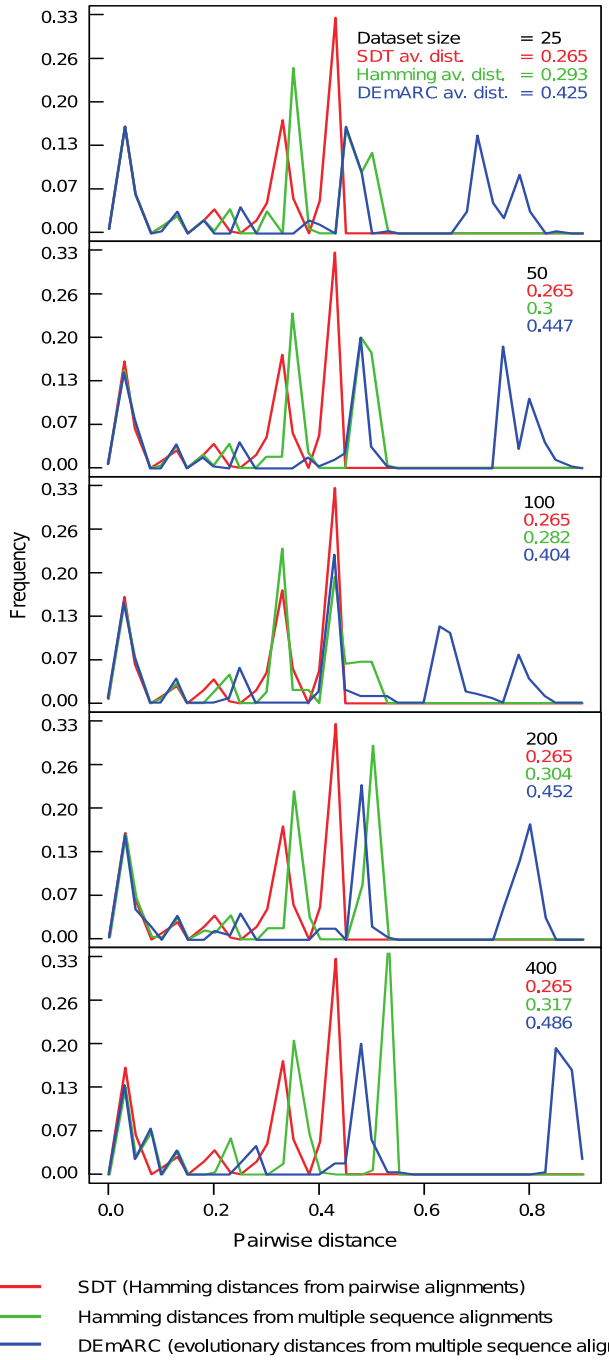


Figure 2-2. Distribution of pairwise genetic/evolutionary distances of the same set of 25 mastrevirus full genome sequences in the context of progressively larger sequence datasets

The constant frequency distribution (represented by red graph) illustrates the consistency of pairwise distance calculation based on pairwise alignments while the changing frequency distributions (represented by blue and green graphs) indicate how pairwise distance scores based on multiple sequence alignment tend to become inflated as dataset sizes get larger.

For all datasets SDT yielded identical pairwise identity score distributions whereas the distributions yielded by the multiple sequence alignment-based methods differed substantially between the different datasets. This indicates that SDT is absolutely consistent whereas the other methods are not. It should be pointed out here that the absolute consistency of SDT is an obvious consequence of it using pairwise sequence alignments rather than multiple sequence alignments. In this regard it is absolutely certain that PASC too would have been found to be similarly consistent had it been flexible enough to allow the analysis of the various datasets.

Other points that should be noted in Figure 2-2 are that, firstly, the multiple alignment-based comparison methods always yielded higher average distance estimates than SDT, and secondly, that the magnitudes of these differences tend to increase with increasing dataset size (with the 100 sequence alignment being a notable exception). These observations simply confirm that pairs of sequences in the context of multiple sequence alignments tend to appear less similar to one another than they do in the context of pairwise sequence alignments.

It is important to point out here that the higher degrees of identity inferred by SDT do not necessarily imply that SDT identity estimates are more accurate than those inferred from the multiple sequence alignments. It is entirely plausible that, relative to the gap characters inserted during the pairwise alignment of two sequences, the positions of gap characters within pairs of sequences that are drawn from a multiple sequence alignment might better reflect the patterns of insertion and deletion that actually occurred during the evolution of the sequences. It is in fact expected that identity estimates based on pairwise alignments could at least slightly overestimate the relatedness of sequences: for example, even two completely random sequences can yield pairwise identity scores of > 40% following pairwise alignment. In the context of virus classification, however, this is not necessarily a bad quality: particularly in a publication environment that incentivises the discovery of novel virus genera, species and strains. If anything, slightly overestimating pairwise identity estimates will force a degree of conservatism when proposing that new taxonomic groupings be created to accommodate novel virus isolates.

2.4.2 Speed gains of SDT with parallelisation

In addition to the graphical version of SDT being extremely easy for non-specialists to use (it is very difficult to even purposefully manipulate the program to yield inflated or deflated identity scores), the software is also flexible enough to be of interest to more specialist users. For example, the command line versions can be directly slotted into analysis pipelines to automatically identify rational operational taxonomic unit demarcation thresholds and then automatically apply these to the subdivision of large datasets for downstream analyses. In this regard the SDTMPI_Linux version of SDT was specifically designed for the analysis of large datasets (containing more than 1000 sequences) and is well suited for inclusion in high throughput viral metagenome sequencing pipelines. The improvement in analysis speed afforded by SDTMPI_Linux over SDT_Linux was illustrated by an analysis of 1000 begomovirus sequences (requiring 499,500 pairwise sequence alignments of ~2800nts; Table1). The 32 bit version of SDT_Linux took 3740.37 min (~62.34 h) whereas SDTMPI_Linux running on 20 cores (each with similar specifications to that used with the serial version) took 188.56 min (~3.14 h). SDTMPI_Linux running on 40 cores took only 96.63 min (1.61 h). The speed-up improvements with 20 and 40 cores were therefore 19.8 and 38.7 fold, respectively. Overall the 64 bit version of SDT yielded a further 13% increase in speed which is likely due to more efficient memory use. The 64 bit version of SDT would likely yield even better performance gains over the 32 bit version when analysing longer sequences (Table 2-1).

Table 2-1. Speed-ups achieved with the parallelised versions of SDT

System	Program	Number of Sequences	Number of cores	Processor speed (GHz)	RAM (GB)	Time (min.)	Time (hrs.)	Speed up
32 bit	SDT_Linux	1000	1	2.8	24	3740.37	62.34	
	SDTMPI_Linux	1000	20	2.8	24	188.56	3.14	19.8 fold
	SDTMPI_Linux	1000	40	2.8	24	96.63	1.61	38.7 fold
64 bit	SDT_Linux	1000	1	2.8	24	3343.37	55.72	
	SDTMPI_Linux	1000	20	2.8	24	173.02	2.76	19.3 fold
	SDTMPI_Linux	1000	40	2.8	24	85.43	1.42	39.1 fold

2.5 Conclusions

We present a free open-source cross-platform computer program that has been specifically designed to enable general virologists to consistently classify newly determined virus full genome sequences according to ICTV endorsed pairwise genetic identity based genus, species and strain demarcation recommendations. Besides providing the means to minimise inconsistencies in virus taxonomic classifications, the program is suitable for use both by biologists with limited computational skills and more computationally capable biologists that require the rapid automated analysis of very large datasets. Unlike the similar sequence classification tool, PASC, SDT is not exclusively designed for full virus genome based pairwise identity calculations but is also usable as an amino acid sequence classifier – a fact which could make it very useful for the characterisation of novel highly divergent viruses.

Although we have primarily focused here on the merits of SDT relative to PASC and DEmARC, it should be stressed that SDT is not a competitor of PASC and DEmARC – it is instead a complementary tool that could be used in conjunction with these other methods to establish and effectively implement pairwise identity based virus classification systems. For example DEmARC might be used by the ICTV to establish a solid evolutionary rationale for defining a particular set of species, PASC might then be used by individual ICTV working groups to establish easy to apply pairwise identity thresholds that optimally conform with the DEmARC classifications, and SDT (or equivalent software) might be used by individual virologists to consistently apply these thresholds during the tentative classification of novel virus isolates that they submit to public sequence databases. Finally, even if SDT is not deemed suitable as a classification tool by particular ICTV working groups, it will still have widespread utility as a tool for graphically visualising colour coded pairwise genetic similarities of large numbers of sequences – a niche that is currently unfilled by any other sequence analysis software.

The various versions of SDT that have been described here, along with instructions for their installation and use, are freely available at www.cbio.uct.ac.za/SDT.

2.6 Authors' contributions and acknowledgements

Main author's contribution

I designed and implemented the Windows GUI, Linux/MacOS command line and the parallel versions of SDT, tested its performance in comparison to other existing similar computer programs and wrote ~80% of the manuscript.

Co-authors' contributions

Darren Martin and Arvind Varsani played a major role in conceiving the SDT project and supervising its implementation. They collectively contributed to writing about 20% of the manuscript and Arvind edited the figures.

Acknowledgements

I thank the Centre for High Performance Computing (CHPC; www.chpc.ac.za) in Cape Town and the University of Cape Town's Information and Communication Technology Services (ICTS; http://hpc.uct.ac.za) for granting access to their high performance computer clusters. Particularly, I thank the CHPC for a Python parallel programming course that I attended during the 2012 Summer School at the Free State University. The course became a source of motivation that lead to successful development of the parallel version of SDT.

Chapter 3 : Computational tools for the identification within virus genomes of secondary structures with likely biological functionality

3.1 Abstract

Besides biological information encoded within viral genomes by genes, functional domains and conserved motifs, secondary structures formed through base-paring of nucleotides across the genome have been found to play a role in many of the biological processes essential for the survival of these pathogens. Recent advances in molecular evolutionary modelling have prompted development and application of various computational tools for the prediction and characterisation of genomic nucleic acid structures for several virus species. However, current computational methods are not entirely reliable as some of the predicted structures may not really exist. In fact even structures that actually exist and are apparently conserved across multiple related genomes may not have any biological function. It is possible to identify potential biologically important structural elements by validating predictions based on other biological evidence such as natural selection, prior to conducting extensive and expensive lab experiments to determine the exact function of identified structures. Here I present a range of novel computational tools that have been developed for the prediction and characterisation of likely functional secondary structures that are evolutionarily conserved amongst related virus genomes, including: (1) a tool that objectively organises virus genomic sequences into analysable datasets, (2) a minimum free energy-based computer program that improves the prediction of evolutionarily conserved structures within virus genomes, (3) computer scripts for natural selection inferences and data analyses aimed at determining degrees of evolutionary support for predicted structures, and (4) computer applications introducing a novel approach to the visualisation of secondary structure and natural selection data. The performance of these tools was assessed using some ssDNA and HIV-1M viral datasets. All developed tools are freely available at http://web.cbio.uct.ac.za/~brejnev/ComputationalTools.html.

3.2 Introduction

Across all known domains of life on Earth, Watson-Crick base-pairing within and between nucleic acid molecules determine their catalytic activities and or functional roles within living cells. Besides the obvious role of helical double-stranded DNA in storing the genetic information of all known life, the internal base-pairing of single-stranded nucleic acid molecules such as microRNA (miRNA; Vermeulen et al. 2007), ribosomal RNA (rRNA; Mears et al. 2002) and transfer RNA (tRNA; Wende et al. 2014) is also central component of all known life. Notably, in the world of viruses, these structures also form within all single-stranded RNA (ssRNA) and single-stranded DNA (ssDNA) viral genomes and they have been found to play regulatory functions in many virus families during important biological processes such as genome replication (Cheung 2004a), transcription (Koev et al. 1999), translation (Koev et al. 1999), gene-splicing (Moss et al. 2012), generation of viral subgenomes (Koev et al. 1999; Roby et al. 2014) and evasion of host immune responses (Roby et al. 2014).

While there are structures with known functional roles, many potentially biologically important structural elements that are present within viral genomes remain uncharacterised. This has prompted structural biologists and viral evolutionists to invest their efforts in designing and utilising tools for the prediction and characterisation of these elements. Such tools include a wide range of computational methods mainly relying on free energy minimisation approaches (Hofacker and Stadler 2006; Markham and Zuker 2008; Semegni et al. 2011; Dela-Moss et al. 2014). In addition to these purely computational methods, hybrid computational-experimental methods (Wilkinson et al. 2006) have proven extremely useful for the accurate prediction of secondary structural elements that form transiently within the ssRNA genomes of several viruses (Watts et al. 2009; Pollom et al. 2013; Dela-Moss et al. 2014).

Despite the predictive improvements achieved by these methods, currently there are no computer programs for identifying biologically functional secondary structures that account for structural conservation while also explicitly accounting for other relevant evolutionary processes such as mutational dynamics, natural selection and genetic

recombination. As has been demonstrated by several studies, it is possible to validate the existence and biological importance of structural elements within single-stranded nucleic acid molecules by studying signals of molecular evolution that are consistent with the evolutionary preservation of structural elements. This is because: (1) a structural element that is biologically functional is likely to be conserved within genomes of closely related species or families (Hofacker et al. 1998; Collier et al. 2002; Pedersen et al. 2006; Dela-Moss et al. 2014; Muhire et al. 2014a), and (2) natural selection patterns should manifest a tendency towards the maintenance of these elements through strong purifying selection at base-paired sites, for example such sites will have low synonymous substitution rates (Cloete et al. 2014; Muhire et al. 2014a) and/or complementary coevolution between paired nucleotides (Eddy and Durbin 1994; Tuplin et al. 2002; Cheng et al. 2012a; Cloete et al. 2014; Muhire et al. 2014a), and (3) there should be evidence that selection acts to maintain structural configurations following genetic recombination (Golden et al. 2014a). However, no existing computer programs can predict biologically functional nucleic acid structures using all of these various layers of biological information. A computational pipeline comprising novel algorithms and tools capable of augmenting currently available functional structure predictions with evolutionary evidence would be extremely useful in efforts to pinpoint the subsets of structural elements within single-stranded nucleic acid molecules that warrant further biological analyses.

It is noteworthy that bioinformatics-based visualisation tools are helping to intuitively present and understand molecular biology data and are gaining popularity as they become more sophisticated (Agapito et al. 2013). The most successful of these tools have focused on the visualisation of genomics/transcriptomics data (Sturn et al. 2002; Conesa et al. 2005; Karolchik et al. 2011) nucleic acid structures (De Rijk et al. 2003; Martinez et al. 2008; Darty et al. 2009), protein structures (Herráez 2006; Porollo and Meller 2007) and protein-protein interactions (Shannon et al. 2003; Brown et al. 2009; Salazar et al. 2014). There are, however, very few available tools for visualising how nucleotide or codon sites evolve and coevolve within the context of protein or nucleic acid structures.

Here, I present a computational framework comprising a range of computer programs and data analysis scripts for the identification of biologically functional

secondary structures, the detection of selection and coevolution, and the visualisation and analysis of genome evolution within the context of genomic secondary structure. The pipeline comprises: (1) Sequence Demarcation Tool (SDT), a program that uses pairwise sequence identities to objectively organise a group sequences into analysable datasets (Muhire et al. 2014b); (2) Nucleic Acid Structure Predictor (NASP), a program that identifies evolutionary conserved nucleic acid secondary structures (Semegni et al. 2011); (3) scripts to analyse patterns of natural selection and coevolution using Hypothesis testing using Phylogenies (HyPhy), a program designed to test complex evolutionary models within a phylogenetic context; (4) Nucleic Acid Visualisation and Analysis (NAVA), a nucleic acid secondary structure and evolutionary data visualisation program, that also ranks individual inferred structural elements in order of their potential biological functionality based on (i) their degrees of conservation, (ii) the degree to which structures influence the amino acid encoding potential of their composite and (iii) the degree to which their base-paired nucleotides coevolve; (5) StructureMap, a visualization program that indicates the genomic locations of secondary structural elements; (6) SelectionMap, a program for the quantification and visualisation of similarities and differences between natural selection signals evident within homologous genes sampled from different species; and (7) Fold disruption test: a module that tests for avoidance of nucleic acid fold disruption following genetic recombination. Finally I assess the usability and performance of these tools, all of which are freely available along with documentation for their effective use from:

http://web.cbio.uct.ac.za/~brejnev/ComputationalTools.html.

3.3 Material and methods

3.3.1 Sequence Demarcation Tool (SDT): objective creation of datasets using pairwise sequence identities

The accuracy of molecular evolution analyses such as those aimed at the prediction of evolutionarily conserved nucleic acid secondary structures, the analysis of natural selection and the detection of genetic recombination largely depends on the quality of the sequence datasets used. Highly divergent sequences (i.e. those sharing <75% pairwise identities) can often not be accurately aligned (Wilm et al. 2006). Alignment errors can seriously impact estimates of secondary structural conservation, inferences of selection acting on coding sequences (it can cause false positive signals of positive selection), and the detection of genetic recombination (misaligned genome regions can be detected as recombination events). Conversely, analyses of small numbers of very similar sequences (i.e. those sharing > 95% pairwise identities) can lack sufficient power both to differentiate between the degrees to which different structural elements are conserved and to detect signals of natural selection and genetic recombination.

In this regard, I implement a method within SDT (described in Chapter 2 Muhire et al. 2014b) which, given a dataset of biological sequences (DNA, RNA or Amino-acid) computes pairwise identities and uses these to objectively partition the input dataset into smaller non overlapping datasets, each consisting exclusively of sequences that share a user-defined degree of sequence identity (Muhire et al. 2014a).

Once the identity scores are calculated, the user clicks on the save menu on the program interface (Figure 3-1 A) and choses the “Create dataset [based on p-identities]” option. This displays a data creation window (Figure 3-1 B). In the data creation window the user must enter the maximum and the minimum identities and the program then creates the appropriate datasets.

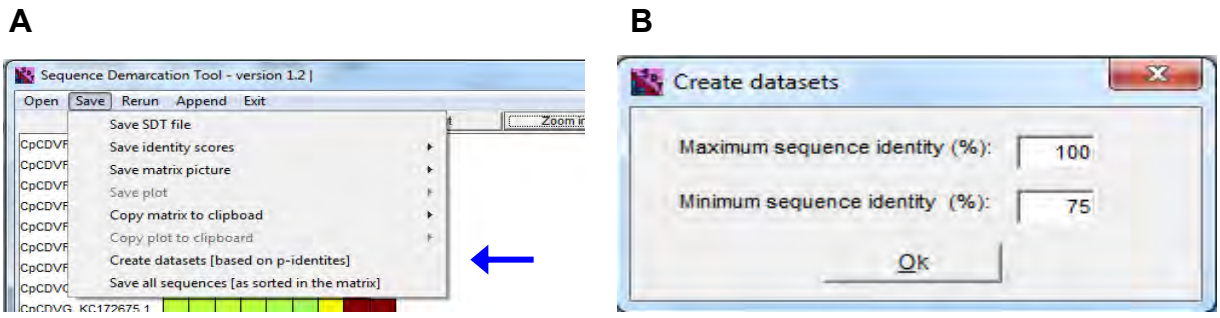


Figure 3-1. SDT's dataset creation window

(A) The **Create datasets [based on p-identities]** submenu (pointed by the blue arrow) is enabled once pairwise identities have been computed. If it is clicked the Create datasets window is displayed. **(B)** The **Created datasets window** allows the user to enter the maximum and the minimum sequence identities, which are used to partition the input alignment into files that each contains sequence pairs sharing identities between the entered minimum and maximum identities.

The SDT graphical user interface (GUI) is written in Visual Basic 6.0. Command line versions of SDT for Linux and MacOS and a parallel version of SDT written in Python produce a ".SDT" output file that can be loaded into the SDT GUI to directly generate datasets without rerunning an analysis. Lastly, datasets that are created following an SDT analysis need to be aligned (preferably using MUSCLE) prior to proceeding to downstream analyses.

3.3.2 Nucleic Acid Structure Predictor (NASP): prediction of evolutionary conserved structures

The identification of evolutionary conserved nucleic acid secondary structures within virus genomes is carried out by NASP (available on http://web.cbio.uct.ac.za/~yves/nasp/; Semegni et al. 2011). NASP takes an alignment of evolutionary related genomes, uses the hybrid-ss-min component of the UNAFOLD package (Markham and Zuker 2008) to fold sequences, determines the consensus structure of all the sequences in the input alignment, and then applies a recursive permutation test to determine the subset of conserved structural elements that is responsible for the input sequences having a greater degree of structural stability than 95% of the sequences in randomly generated alignments with the same base-composition. The subset of conserved structural elements thus identified by NASP is referred to as the high confidence structure set (HCSS). For any given input alignment the HCSS will generally consists of 0% to 20% of all the predicted structural elements. The genomic coordinates of the HCSS elements are written to a

concatenate (CT) file that is processed with custom Python scripts for further analysis (Muhire et al. 2014a).

3.3.3 Test for degrees of natural selection acting at paired and unpaired sites

Mutations inducing disruption of base-pairing within biologically functional secondary structural elements within a viral genome can result in partial or complete loss of fitness, in which case the genomes will likely be purged by selection. Such deleterious mutations will either revert to wild-type (Cheung 2005; Shepherd et al. 2006) or will become fixed and be compensated for by complementary mutations that restore base-pairing (Hofacker et al. 1998; Fernández et al. 2011; Cheng et al. 2012a). There should therefore be evidence of negative selection acting against nucleotide substitutions at base-paired sites within biologically functional secondary structural elements.

Here are computer scripts I wrote in Python and R programing that statistically test whether paired sites within HCSSs display higher degrees of negative selection at both the nucleotide and codon levels compared to sites which are not included within the HCSSs.

3.3.3.1 Purifying selection analysis - Tajima and Fu-Li-based permutation test

This test estimates the strength of purifying selection at the nucleotide level and determines whether there is significantly greater purifying selection at paired sites than there is at unpaired sites within the HCSSs of an alignment. It consists of the following two sets of scripts:

(a) Splitting alignments into paired and unpaired site alignments

The Python script *"Split_Paired-Unpaired.py"* performs a profile alignment of a small alignment of selected sequences (in FASTA format) that was used by NASP for predicting paired sites within a HCSS, and a large alignment file consisting of all available sequences under investigation (structure predictions are generally only carried out by NASP on a subset of up to 10 sequences in any dataset because the

program is extremely computationally intensive). The script then maps coordinates of paired sites produced by NASP to the full large alignment, and separates the sites within the full alignment into those corresponding to paired sites in the HCSS (called the paired site alignment) and those from the remainder of the alignment (called the unpaired site alignment).

(b) Tajima - Fu & Li permutation test

After the production of paired and unpaired site alignments, the *"Tajima_Fu-Li.py"* script allows one to compare degrees of negative selection between the two separated alignments based on Tajima's D and Fu & Li's F statistics (Tajima 1989; Fu and Li 1993) using implementations of these tests in the Evolutionary genetics and genomic library (Egglib; http://seqlib.sourceforge.net/). Given the two alignments, the scripts compute Tajima's D and Fu & Li F statistics for the paired alignment, and 100 simulated alignments each consisting of the same number of sites as the paired alignment with each site being randomly sampled with replacement from the unpaired alignment. A p-value is computed as the fraction of the simulated alignments that have Tajima or Fu and Li statistics lower or equal to that of the paired alignment.

3.3.3.2 Comparison of synonymous substitution rates between paired and unpaired nucleotides within codons

This test estimates degrees purifying selection within gene alignments that exclude regions with overlapping open reading frames. It determines whether rates of synonymous substitutions (dS; i.e. nucleotide substitutions that do not change the encoded amino acid) at codon sites containing paired nucleotides are significantly lower than those at codon sites containing unpaired nucleotides. In this test the only nucleotides that are considered are those at third codon positions. The third codon position is used because it is the site where synonymous mutations can most freely occur (i.e. the vast majority of mutations at first codon sites and all mutations at second codon sites are non-synonymous). The test involves the use of Python and R scripts that (1) infer dS rates using selection analysis tools implemented in the computer program Hypothesis testing using Phylogenies (HyPhy; Pond et al. 2005)

which runs on a computer cluster, and (2) perform statistical tests on the output files obtained from HyPhy.

(a) Recombination detection and the inference of synonymous substitution rates

Since unaccounted for recombination could undermine the accuracy with which synonymous and non-synonymous substitution rates are estimated (Scheffler et al. 2006), recombination breakpoint positions within the input alignment are first detected using the Genetic Algorithm for Recombination Detection (GARD; Kosakovsky Pond et al. 2006) method in HyPhy and then dS rates are estimated using both the Partitioning Approach for Inference of Selection (PARRIS; Scheffler et al. 2006) and the Fast Unconstrained Bayesian Approximation (FUBAR; Murrell et al. 2013) methods implemented in HyPhy. These HyPhy methods are computationally intensive and require high performance computing resources. For this reason, I created a Python script that automatically generates shell scripts that are required to run all of the tools for all of the input files.

(b) Test for significantly lower substitution rates at paired codon sites than at unpaired codon sites

Custom Python scripts are used to map the rates inferred by PARRIS and FUBAR to the proper alignment coordinates of HCSS and other potentially base-paired site coordinates that were inferred by NASP. The mapping allows the determination of synonymous substitution rates at paired codon sites (i.e. those codons where the site at the third position is base-paired) and unpaired codon sites (i.e. those codons where the third position site is unpaired). An R script is used to perform a Mann-Whitney U test to determine whether synonymous substitution rates are indeed significantly lower at paired than unpaired nucleotide sites.

3.3.4 Complementary coevolution test for paired nucleotide sites

Preservation of biologically functional secondary structures is likely to be crucial during the evolution of viral genomes. While there should be selection against any nucleotide substitutions at all at base-paired sites that are absolutely essential for the

integrity of important secondary structural elements (Simmonds and Smith 1999), substitutions that are more tolerable could potentially be compensated for by complementary substitutions at their pairing partners such that the pairing interactions are maintained (Cheung 2004a). Accordingly, several studies have used evidence of complementary coevolution to infer the existence of biologically functional secondary structures within virus genomes (Hofacker et al. 1998; Fernández et al. 2011; Cheng et al. 2012b).

Here I present a complementary coevolution detection method that initially uses Recombination Detection Program (RDP4; Martin et al. 2010) for detecting recombination within an alignment of nucleotide sequences and then uses a HyPhy script (hereafter referred to as "Coevolution script") that implements a coevolution model developed by Muse (Muse 1995), to detect sites which may be complimentarily coevolving. Lastly, Python and R scripts were written to test for associations between sites that are predicted to be base-paired and site-pairs that appear to be complementarily coevolving.

3.3.4.1 Recombination detection

Prior to the analysis of coevolution RDP is used to detect recombination. Given an alignment of *N* sequences RDP detects recombination and produces a distributed alignment in which recombinant regions are extracted individual sequences and readded to the alignment as unique sequences, thus creating a recombination-free alignment containing >*N* sequences. Using a sliding window I create sub-alignments containing N sequences (containing maximum sequence content). All the recombination-free sub-alignments are converted to PHYLIP format, and for each a maximum likelihood phylogenetic tree is inferred using PhyML3.0 with the HKY85 nucleotide substitution model (Guindon et al. 2010). The sub-alignments in PHYLIP format and their corresponding phylogenetic trees in NEWICK format are used as input for the coevolution script.

3.3.4.2 The coevolution script

The coevolution script written in HyPhy implements a coevolution model that fits the 4X4 HKY85 nucleotide substitution model (Hasegawa et al. 1985) and a modified version of HKY85 model (hereafter referred to as M95; Muse 1995) that accounts for

base-pairing, to every individual pair of sites within an alignment and applies a likelihood ratio test to determine which model better describes the substitution patterns observed at the two sites (Muse 1995).

The M95 coevolution model is represented in a 16X16 nucleotide substitution matrix (Figure 3-2) whose entries represent the rates of changes from one pair of nucleotides to another. To account for both complementary and non-complementary coevolution, the rates of change from unpaired states to paired states (Watson-Crick and Wobble base-pairs) are multiplied by a factor, ***λ*** (complementary base-pairing factor) and rates of changes from paired states to unpaired states are multiplied by ***1/λ***. Coevolution is inferred when the M95 versus HKY85 likelihood ratio test has a p-value less than 0.05. However, complementary coevolution is detected when the estimate of ***λ*** is greater than one while non-complementary coevolution is detected when ***λ*** is less than one.

	AT	TA	CG	GC	AA	AC	…
AT	*	0	0	0	$\beta\pi_A/\lambda$	$\alpha\pi_C/\lambda$	…
TA	0	*	0	0	$\beta\pi_A/\lambda$	0	…
CG	0	0	*	0	0	0	…
GC	0	0	0	*	0	$\mu/4\lambda$	…
AA	$\beta\pi_T\lambda$	$\beta\pi_T\lambda$	0	0	*	$\beta\pi_C$	…
AC	$\alpha\pi_T\lambda$	0	0	$\alpha\pi_G\lambda$	$\beta\pi_A\lambda$	*	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Figure 3-2. Muse 95 nucleotide substitution model (M95)

The matrix represents the Muse 95 (M95) coevolution model which is a 16-by-16 version of the HKY85 nucleotide substitution model modified to account for constraints on base-paring. Each cell represents the rate of change from a pair of nucleotides to another pair with one nucleotide substituted. To account for base-pairing the rates of changes from unpaired to paired nucleotides, Watson-Crick base-pairing (e.g. AC → GC) and Wobble base-pairing (AT → GT) are multiplied by a base-pairing factor λ while those of changes from the paired to unpaired nucleotides are multiplied by $1/\lambda$. The rates of changes involving two nucleotide substitutions are ignored and are represented by zero. π_i represents the base frequencies (where i=A, C, G or T) while * is used to equate the sum of rates in every row or column to 1. (Adapted from Muse 1995)

3.3.4.3 Running the coevolution script and performing statistical test

Since coevolution needs to be detected for every site against every other site within *n* nucleotides (usually *n*=100 or larger) and due to the large numbers of input recombination-free sub-alignments and trees obtained for every unique distributed alignment, the problem becomes complex and computationally intense such that high performance computing resources are absolutely necessary. To address this problem, a Python script was written which detects the input sub-alignments and corresponding trees based on their names (which are appropriately indexed) and generates the multitude of Python scripts that are needed to run the HyPhy coevolution script. Each script runs for a small section of alignment and skips all sites that are invariant. The output data are filtered using a Python script and chi-squared tests for associations between complementarily coevolving sites and base-pared sites are carried out using an R script.

3.3.5 Nucleic Acid Visualisation and Analysis (NAVA): secondary structure ranking and visualisation

3.3.5.1 Ranking

The approach used to rank secondary structures in order of their likely biological functionality is implemented in the computer program, NAVA (available on https://sites.google.com/site/cbiomichael/software); an updated version of the computer program Data Overlaid On Secondary Structures (DOOSS; Golden and Martin 2013). NAVA is specifically designed to enable investigators to visualise and rank secondary structures according to their likely biological relevance. NAVA ranks structures using a Mann Whitney U test that compares the distribution of data values (either synonymous substitution rates or complementary coevolution p-values) for an individual structure to the distribution of data values of all other structures

3.3.5.2 Visualisation

For visualisation, NAVA uses the full genome sequences and a CT file containing evolutionary conserved structures (the file indicates coordinates of sites which are paired with one another), and employs the computer program VARNA (Darty et al.

2009) to draw the secondary structures. It further maps the gene codon-alignment and corresponding synonymous substitution rates to the structures so that the rates can be visualised on the structural elements of interest. In the structure each nucleotide is coloured using a green-blue scale corresponding to the value of the rate. Similarly the alignment used for coevolution is mapped to the structure and complementary coevolving sites are indicated using lines that have an orange-red colour scale that is based on the value of their corresponding complementary coevolution p-value.

For detailed instructions on how to use NAVA please visit the following webpage: https://sites.google.com/site/cbiomichael/software

3.3.6 StructureMap: visualisation of genome-wide secondary structure map

I present a computer program StructureMap that draws a full genome secondary structure map, displaying locations of structures that are ranked based on their likely biological functionality. StructureMap takes as input a text file containing the length of the full genome and the coordinates of structural elements ranked in order of their likely biological importance and displays a genome map in either linear or circular configuration (as determined by the user) where locations of structures are marked by arcs (for circular genomes) or horizontal lines (for linear genomes) with the top ranking structures highlighted in different colours. Structure maps drawn for different virus species are scaled according to the length of the genomes such that structure maps of related species can be accurately compared for identification of functionally homologous structures within distantly related species (Muhire et al. 2014a; Stenzel et al. 2014).

StructureMap (Fig 3-3) is written in Visual Basic 6.0, and uses the Microsoft Windows graphics device interface “GDI32” library to produces publication quality figures in both Enhanced Metafile Format (EMF) and Portable Network Graphics (PNG) formats.

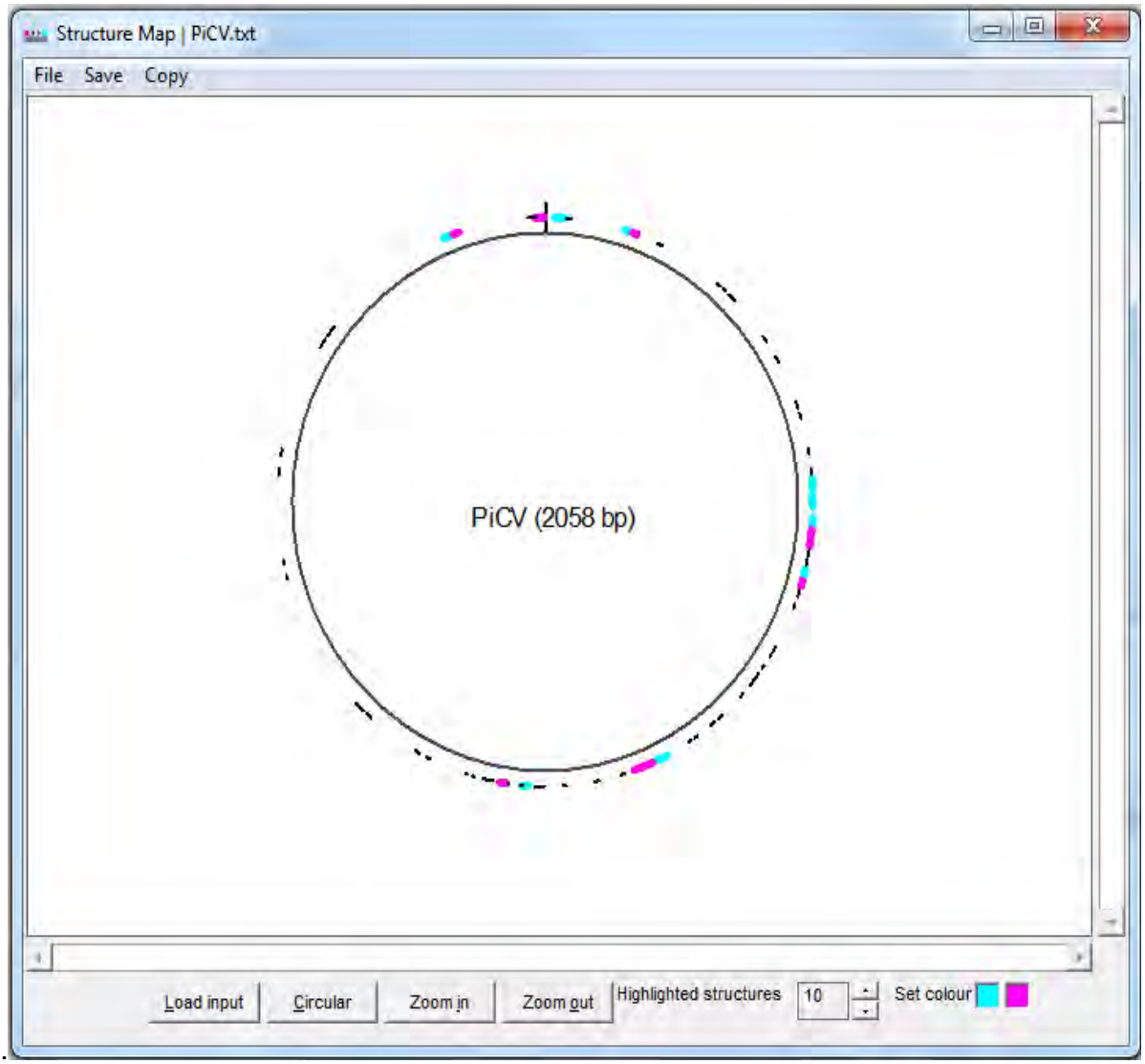


Figure 3-3. StructureMap interface

The StructureMap interface displaying the structure map of the *Pigeon circovirus* (PiCV) genome. The origin of replication is located at the 12 o-clock position while at the centre of the map the name of the input file "PiCV" is displayed along with the length of the genome (enclosed by brackets). The ten highest ranked structures are presented in arcs coloured in cyan (indicating the 5' stem of the structure) and magenta (indicating the 3' stem of the structure) while the remaining predicted structures are represented by black arcs. The menus allow input files to be loaded, and the copying or saving of maps in either EMF or PNG format. Command buttons allow input files to be loaded, the setting of genome types to linear or circular and zooming-in or -out of the map. The spin control allows a user to set the number of structures that will be highlighted. By clinking on the coloured boxes, the user can set a colour of their preference to be used for highlighting the top-ranked structures.

StructureMap was used to compare full genome structures of circoviruses in the species *Pigeon circovirus* (PiCV) and *Beak and feather disease virus* (Stenzel et al. 2014), and several other species within the virus families *Geminiviridae*, *Anelloviridae*, *Nanoviridae* and *Parvoviridae* (see Chapter 4).

3.3.7 SelectionMap: visualisation of natural selection patterns within genes

The presence of functional secondary structures, regulatory motifs and other functional domains within genes has an impact on the degree of selection detectable within gene encoding nucleotide sequences (Ngandu et al. 2008; Muhire et al. 2014a). Inference and comparison of natural selection patterns within the gene alignments of related virus species allows the identification of homologous codon sites that are maintained under the same or similar selective pressures in these species. Such comparisons also highlight differences in selection patterns between related virus species that are likely indicative of differences in the selective environments of the species: differences that might be attributable to their specific preferred hosts or transmission vectors.

To facilitate such analyses I wrote a computer program, SelectionMap (Figure 3-4), that graphically illustrates the types and degrees of natural selection acting at each codon site along the gene alignments of related virus species. The program takes as input the selection data from the selection detection tools, Fast Unconstrained Bayesian Approximation (FUBAR; Murrell et al. 2013) and Mixed Effects Model of Evolution (MEME; Murrell et al. 2012) both of which are implemented in HyPhy. The codon alignments of homologous genes from two or three species are aligned using a profile codon alignment option of MUSCLE (Edgar 2004), and then are separated before being used for inference of selection. FUBAR is used to infer degrees of negative/positive selection at each individual codon site while MEME is used to infer evidence of episodic positive selection. Different colours are used to differentiate sites in the two or three datasets that are concurrently under negative selection for the same amino acid (blue), under negative selection but for different amino acids in two or three different datasets (green), sites under negative selection within only one of the datasets (grey), sites under positive selection in individual datasets (red) and sites under episodic selection in individual datasets (orange).

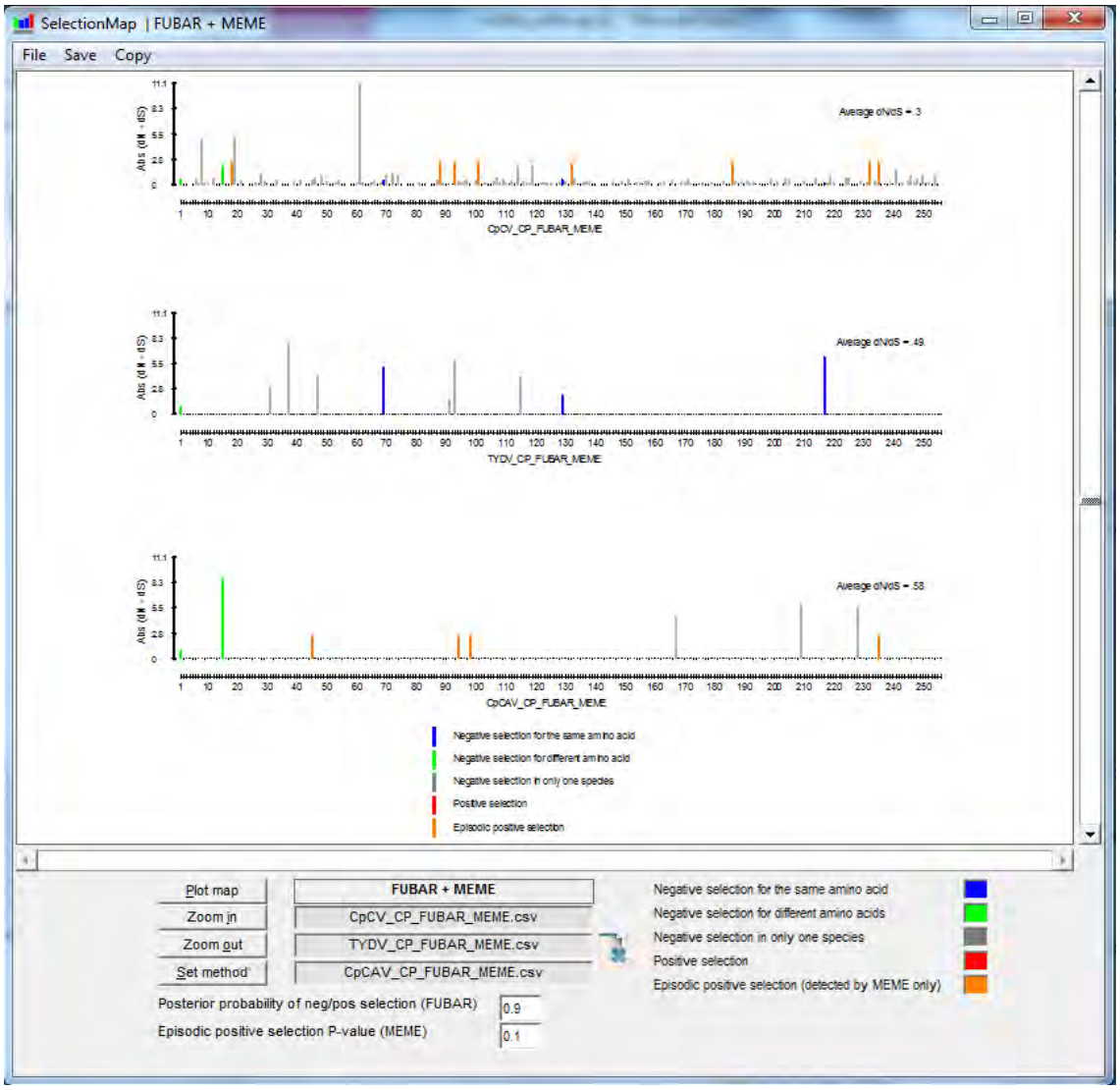


Figure 3-4. SelectionMap interface

The SelectionMap interface displaying a selection map comparing signals of natural selection detected in (**a**) *Chickpea chlorotic dwarf virus* (CpCDV), (**b**) *Tobacco yellow dwarf virus* (TYDV) and (**c**) *Chickpea chlorosis Australia virus* (CpCAV) capsid proteins. On the selection map, the degree of selection is represented by the height of the bar (corresponding to the absolute value of the synonymous substitution rate minus the non-synonymous substitution rates; Abs(dN - dS). Blue bars represents sites under selection for the same amino acid in multiple species, green bars represent sites under negative selection in multiple species but for different amino acids, grey bars represent sites under negative selection in only one of the species, red bars represent sites under positive selection and orange bars represent sites under episodic positive selection (sites where MEME inferred positive selection while FUBAR did not). The average dN/dS ratio is shown at the top-right of each gene. On the program interface the menus allows one to copy or save the generated selection map in EMF and PNG formats. Command buttons allow one to plot the map, zoom-in and –out and set the analysis method(s) to use (three options are supported FUBAR + MEME, FUBAR and MEME). Buttons below the "FUBAR+MEME" label are used to load the input files. Alternatively input files can be dragged and dropped into the picture box within the program interface. The text boxes labelled "Posterior probability of negative/positive selection" and "Episodic positive selection p-value" allow one to set the statistical significance thresholds at which selection is inferred by FUBAR and MEME, respectively. The delete button allows one to remove input files and reset the program. The colour boxes allow one to change the colours used to represent different types of selection.

Additional information such as gene annotations and secondary structures can be added manually to the map to observe whether the presence of some functional motifs/domains or structural elements might have an influence on the selection patterns observed.

The SelectionMap interface (Figure 3-4) was developed using Visual Basic 6.0 and uses the GDI32 library to produce publication quality image in either EMF or PNG format. The program has been used for characterisation of selection patterns within the genes of ssDNA virus species (Stenzel et al. 2014; Kraberger et al. 2015).

3.3.8 Nucleic acid fold disruption test

Genetic recombination has been reported in many different virus families (Navas-Castillo et al. 2000; Simon-Loriere and Holmes 2011; Martin et al. 2011a; Bujarski 2013) In many of these families recombination breakpoint distributions are apparently non-random with distinct and often quite highly conserved recombination hot- and cold-spots being detectable (Lefeuvre et al. 2009a; Martin et al. 2011b; Martin et al. 2011a; Stenzel et al. 2014). The potential evolutionary advantages of recombination include enabling genomes to explore sequence space more effectively than can be achieved by mutation alone (Stemmer 1994; Crameri et al. 1998) and defending genomes against the accumulation of deleterious mutations (Muller 1964; Felsenstein 1974). However, recombination can also be deleterious when it induces disruption of important intramolecular interactions within nucleic acid or protein molecules (Martin et al. 2005b; Lefeuvre et al. 2007; Rokyta and Wichman 2009). Therefore, it is plausible that the breakpoint distributions apparent within natural recombinants might display some evidence of selection acting to preserve important nucleic acid and protein interactions (nucleic acid / protein folds; Lefeuvre et al. 2007; Martin et al. 2011c).

I therefore implement a module under RDP4 (described by Martin et al. 2015) and similar to that used in (Golden et al. 2014a), which performs a nucleic acid folding disruption test to investigate whether natural recombinants have significantly less disrupted nucleic acid foldings than that which would occur in the absence of selection against recombination induced nucleic acid folding disruption. The test uses methods already implemented in RDP4 to detect recombinant sequences, their

parental sequences and their recombination breakpoints. This information is used to construct (1) "mimic" recombinant sequences (using identified recombination breakpoints and the two identified parental sequences) and (2) simulated recombinants (also using the two identified parents but with randomised breakpoint locations). All the recombinant sequences are folded and a permutation test is used to help determine the degree of disruption of the "real" recombinants compared to randomised recombinants.

3.3.8.1 Simulation of recombinant genomes

RDP4 provides a detailed list of recombination events that are evident within an alignment of virus genomes. Each event consists of (1) a pair of breakpoints, (2) a major parent (a sequence contributing a longer fragment to the recombinant) and minor parent (a sequence contributing a shorter section to the recombinant), and (3) the "event length" or "number of informative sites" (which is the number of polymorphic sites from the 5' to the 3' breakpoint varying between the minor and major parental sequences). It is noteworthy that the real parental sequences are unlikely to have been sampled in our alignment; thus the major and minor parental sequences refer to sequences in our alignment that are simply most similar to the actual parental sequences.

For each detected recombination event the program reconstructs the recombinant sequence (hereafter referred to as "Mimic recombinant" or M-recombinant) by joining fragments from the major and minor parents such as the fragment delineated with breakpoints comes from the minor parent while the two fragments at the two ends come from the major parent (see Figure 3-5). For every mimic recombinant, I simulate a 100 or more recombinants (hereafter referred to as S-recombinants) using the same major and minor parental sequences as for the M-recombinant, and randomly choosing the location of the 5' breakpoint and finding the closest 3' breakpoint that keeps the same event length as for the M-recombinant. This ensures that M- and S- recombinants all have exactly the same genetic distance from both the major and minor parental sequences – this is important for our permutation-based test for nucleic acid structural disruption. In constructing recombinants the circularity of genomes is accounted for when genomes are circular. If the 5' breakpoint falls toward the end of the genome then the 3' breakpoint is put after the

number sites equal to the event length wrapping around the end to the start of the genome.

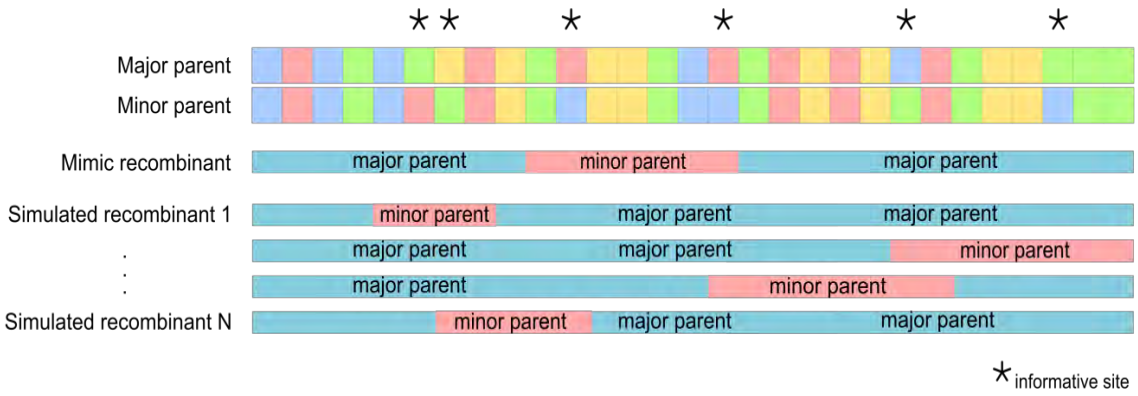


Figure 3-5. Diagrammatic representation of simulation of recombinants

For a particular recombination event indicating a major parent, a minor parent, and locations of pairs of recombination breakpoints delineating a fragment of sequence derived from the minor parent an *in silico* mimic of the real recombinant sequence is created using the minor and the major parental sequences. Following that, a set of N simulated recombinants is generated in a similar way as the mimic recombinant, but using random starting and ending positions, whilst maintaining the same number of variable nucleotides between the breakpoints sites as they are in the mimic recombinant. In this example the event length = 2 therefore both the mimic and simulated recombinants all have two such ‘‘informative’’ sites between breakpoints (source of the figure: Golden et al. 2014).

3.3.8.2 Folding parental and recombinant sequences

For each recombination event, the major and minor parents, the mimic recombinant and all simulated recombinants are folded using hybrid-ss-min component of UNAFold (Markham and Zuker 2008) run under RDP. The user can set folding parameters such as the genome type (which can be either circular or linear, DNA or RNA), and also the temperature of hybridisation.

3.3.8.3 Permutation based fold disruption test

In our fold disruption test I account for both broken and formed base-pairing within the recombinant sequence. The “Base-pairing disruption score” for a recombinant refers to the number of base-pairings which are present in both major and minor parental sequence folds, but which are absent from the recombinant fold. Conversely, the “Aberrant base-pairing score” refers to the number of base-pairings that are present within the recombinant fold but which are absent from both parental folds.

To perform the fold disruption test, I first calculate the "Total Mimic disruption score" (Tot M-score; Table 3-1) by summing up all Mimic recombinant disruption scores for all N recombination events detected, and, secondly, at each event I randomly choose one score from the disruption scores of the 100 simulated recombinants, and add these chosen scores of all N events to obtain the "Total simulated disruption score" (Tot S-score; Table 3-1). This is repeated 100 000 times and at each time the Tot M-score and Tot S-score are compared. The fold disruption p-value is equivalent to the proportion of times the Tot S-score is greater than Tot M-Score in the 100 000 repetitions.

Table 3-1. Permutation based fold disruption test

Event	M-recomb[a]	S-recomb-1[b]	S-recomb-2	S-recomb-3	. . .	S-recomb-100	
1							Rnd(1-100)[d]
2							Rnd(1-100)
3							Rnd(1-100)
.							.
.							.
.							.
N							Rnd(1-100)
	Tot M-score[c]						Tot S-score[e]

[a]Column containing the fold disruption scores of the Mimic recombinant for all recombination events under consideration.

[b]Columns from S-recomb-1 to S-recomb-100 contain the fold disruption scores for 100 simulated recombinants for all recombination events.

[c]The real disruption score: the sum of disruption scores of all Mimic recombinants constructed from all N recombination events.

[d]Column contains disruption scores for one S-recombinant randomly chosen from the 100 simulated recombinants for each recombination event.

[e]The simulated disruption score: the sum of disruption scores from all S-recombinants randomly chosen for all recombination events.

The aberrant base-pairing disruption score is similar to the disruption score except that it involves base-pairing formation rather than disruption.

We have applied our test to investigate the impact of natural selection favouring maintenance of functional secondary structures on genetic recombination patterns within the genome of HIV-1M (Golden et al. 2014a).

3.4 Results and discussions

3.4.1 Objectivity of dataset creation based on pairwise sequence identities

Although the pairwise identity calculation used in SDT was initially implemented as a multiple sequence alignment-free means of classifying viruses into operational taxonomic units (examples of its use in this role can be found in Muhire et al. 2013; Fiallo-Olivé et al. 2014; Grigoras et al. 2014; Stenzel et al. 2014), the tool was subsequently augmented to enable the automated creation of nucleotide sequence datasets optimised for molecular evolution studies.

Here I illustrate the automated generation of optimised datasets using 30 unaligned mastrevirus genome sequences: *Chickpea chlorotic dwarf virus* (CpCDV; 5 sequences), *Wheat dwarf virus* (WDV; 5 sequences), *Panicum streak virus* (PanSV; 9 sequences) and *Maize streak virus* (MSV; 11 sequences). For purposes of this example I opted to partition the dataset so that each sub-dataset would contain sequences that were all between 78 and 100% identical to one another. Given that the mastrevirus species demarcation threshold is 78% (Muhire et al. 2013), these analysis settings would be expected to split the dataset into sub-datasets that each contain only sequences from individual virus species. Accordingly, with these settings the program yields four non-overlapping datasets: a five sequence CpCDV dataset, a five sequence WDV dataset, a nine sequence PanSV dataset and an eleven sequence MSV dataset (Supplementary Figure 1).

If the pairwise identity scores are pre-computed (as would be the case if the input was a standard SDT file), this method is very efficient even on very large datasets containing more than 2000 sequences.

3.4.2 Performance in prediction of functional secondary structures

I analysed well-known functional secondary structural elements within eukaryote-infecting ssDNA and single-stranded RNA (ssRNA) virus genomes that have been experimentally validated to assess the accuracy with which my approach predicts the existence of biologically important secondary structural elements.

3.4.2.1 Previously characterised structural elements are highly ranked within HCSSs

Secondary structure predictions made using NASP generated high confidence structure sets (HCSSs) for 23 datasets from ssDNA virus families of *Circoviridae, Anelloviridae, Parvoviridae, Nanoviridae* and *Geminiviridae* (discussed in details in Chapter 4; Muhire et al. 2014a), with each HCSS consisting of only an average of 20% of all of the predicted structures. These HCSSs are expected to be significantly enriched for structures that really do form *in vivo*. However, even if we were absolutely certain that all of the HCSS structures actually do form *in vivo*, conservation alone is not compelling evidence for the biological functionality of these structures. It would be more compelling to specifically test these structures for evidence of natural selection acting to preserve them.

3.4.2.2 Natural selection based evidence for structural conservation

For the ~20% of predicted structural elements that fall within the HCSSs of all the various virus datasets that were examined, the median synonymous substitution rates of paired nucleotides were consistently lower than those of unpaired nucleotides (Figure 3-6). This general trend indicates firstly, that, besides the unavoidable inaccuracies associated with computational-based structure predictions, many of the HCSS elements predicted within coding regions do really exist and, secondly, that strong selection pressures favour the maintenance of a large proportion of these structural elements. Furthermore, these substitution rates enable the ranking of structures in order of their likely biological importance since the structures with the lowest associated synonymous substitution rates at paired sites are likely evolving under the strongest evolutionary pressures.

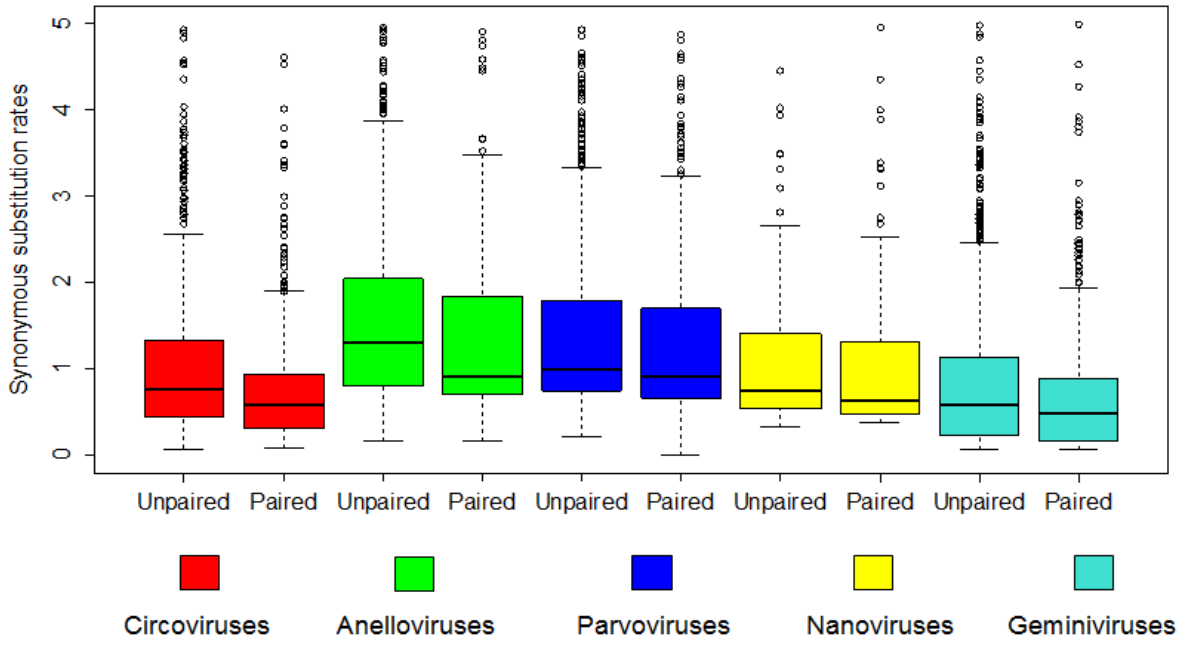


Figure 3-6. Paired vs unpaired sites synonymous substitution rates for eukaryote-infecting ssDNA virus *rep* genes

The box plots compare the synonymous substitution rates between paired and unpaired sites within eukaryotic ssDNA virus *rep* gene codon-alignments. Rates were considered only for the synonymous sites (i.e. the 3rd nucleotide position within each codon). For five eukaryote-infecting ssDNA virus families (described in Table 4-1, Chapter 4; Muhire et al. 2014a) the median synonymous substitution rates of paired sites are consistently lower than those of unpaired sites. This provides evidence that a large proportion of the structures predicted within the HCSS elements that fall within coding regions both really exist, and are likely biologically functional.

3.4.2.3 Complementary coevolution between base-paired sites

Our complementary coevolution test was applied with a view to determine the biological relevance of predicted structural elements within the HCSSs. This test yielded overwhelming evidence of an association of between paired and complementary coevolving sites in full genome datasets drawn from five different ssDNA families (*Geminiviridae, Nanoviridae, Annelloviridae, Parvoviridae, and Circoviridae*; Muhire et al. 2014a) and two ssRNA virus families, *Togaviridae* (Cloete et al. 2014) *and Flaviviridae* (Mauger et al. 2015). Importantly, the complementary coevolution p-values associated with these HCSS structural elements provided an additional means of ranking them in order of their biological importance.

Furthermore the output of our complementary coevolution test bears a striking resemblance to the experimentally determined SHAPE-reactivities of both HCV and

DENV2. Specifically, I have found that there are strong associations between genome regions with high degrees of complementary coevolution and genome structured regions with low degrees of SHAPE reactivity (i.e. parts of the genome that are likely involved in base-pairing interactions). For HCV, I computed the number of significant complementary coevolution interactions per site (i.e. those with an associated p-value <= 0.05) and used a 200nt sliding window to average these values and those of estimated nucleotide-by-nucleotide Shannon Entropies (SE) derived from SHAPE reactivity measurements. In Figure 3-7 it is strikingly clear that peaks of numbers of coevolving nucleotide partners per nucleotide (blue graph) correspond to troughs of average SE (red graph), and peaks of average SE correspond with troughs of the coevolution plot. This strongly supports our hypothesis that complementary coevolution hotspots occur within structured regions while coldspots occur within unstructured regions.

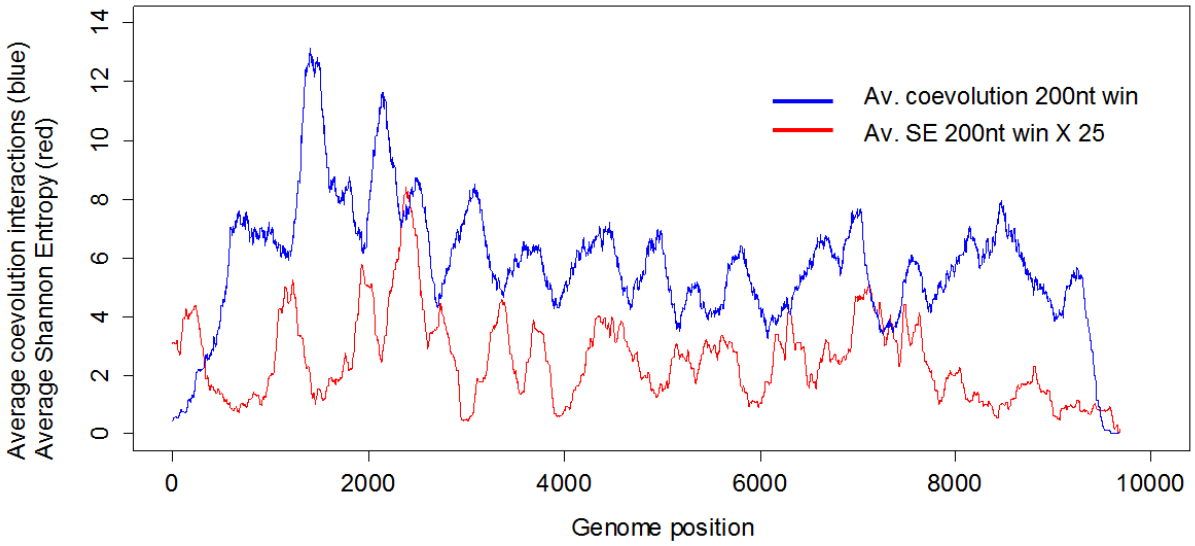


Figure 3-7. Comparison of Hepatitis C virus (HCV) genomic regions with high rates of complementary coevolution and regions of low Shannon Entropy (SE).

A 200nt sliding window is used along the HCV genome to average the number of coevolution interactions (p-value 0.05) and the SE values from HCV SHAPE experiment by (Mauger et al. 2015). For clarity of the SE plot, each average SE value is multiplied by 25 (SE values are generally small compared to the average number of coevolution interactions). Regions of high rates of complementary coevolution consistently correspond with regions of low Shannon Entropy.

As with HCV, for DENV2 I obtained strong evidence for 6/24 highly structured regions determined by SHAPE correspond to regions of highly coevolving regions (manuscript in preparation). However, there was no overall tendency of coevolution

between particular base-paired nucleotides predicted by the SHAPE derived structure model and those identified by our complementary coevolution test. Also, while SHAPE is very accurate in determining whether a nucleotide is paired or unpaired, it does not determine its base-pairing partner.

Importantly, both the agreements and disagreements between the results of our complementary coevolution test and the SHAPE experimental data and structure model are consistent with the hypothesis that RNA structure can be highly dynamic: a hypothesis proposing that a single ssRNA molecule can take on multiple similarly stable configurations. Static structure models such as those that are presently produced by the SHAPE method are unable to capture this dynamism. It is therefore unsurprising that, despite the clear associations across the DENV2 genome between regions with decreased SHAPE reactivity and those with increased numbers of complementarily coevolving base-pairs, there was so little correspondence between base-paired sites in the SHAPE structure model and those which I identified were coevolving.

3.4.2.4 Ranking of structures in order of their biological importance

While the NASP structure conservation score used to rank HCSS placed all structures with know-biological importance within the top 30%, an augmented ranking based on the consensus of this NASP score and the rankings yielded by the synonymous substitution and complementary coevolution analyses placed these structures among the top 22% (Muhire et al. 2014a). The well-described stem-loop that is essential for replication in several groups of eukaryote-infecting ssDNA viruses were among the top ranking structures in many species (2/4 circoviruses had this structure ranked first, and it was ranked second within 3/4 parvovirus datasets and 4/9 geminivirus datasets). Overall this structure ranked within top 14% of structures in geminiviruses, the top 12% in circoviruses, the top 13% in parvoviruses, and the top 50% within nanoviruses (Muhire et al. 2014a).

3.4.3 StructureMap showed the location of homologous highly ranked structures within circoviruses, geminiviruses, parvoviruses and anelloviruses

I used our approach to predict and rank conserved secondary structures within the genomes of two different circovirus species, PiCV (HCSS n = 35) and BFDV (HCSS n = 41), and compared their respective genome maps (Figure 3-8 A). An additional test was performed to determine structural conservation by aligning structures within the context of an alignment (Figure 3-8 B). Three important positionally analogous structures that are highly ranked (among the top six out 35 and 41 HCSS structures in PiCV and BFDV respectively) were identified: (1) the *ori* structure important for genome replication (position 0-66; p-value = 0.035; Figure 3-8 A), (2) a structural element highly conserved in both species near the 5' end of *rep* (S2; position 232 – 305; median p-value = 0.038; Figure 3-8 B) and (3) a structural element also highly conserved in both species within the intergenic region between the *rep* and *cp* stop codons, (S3; position 1029 – 1104; p-value = 0.016; Figure 3-8 B).

The *rep* structures in these species have clear conformational similarities even though the sequences within the structures are moderately diverged. The structure within the intergenic region between the *rep* and *cp* stop codons is a relatively simple stem-loop structure, described in Chapter 4 (Muhire et al. 2014a). The loop sequence contains a conserved pentanucleotide "**CGAAG**" and a G-C rich stem sequence 10 to 15 nucleotides long that displays strong evidence of complementary coevolution between base-paired nucleotides (Muhire et al. 2014a; Stenzel et al. 2014).

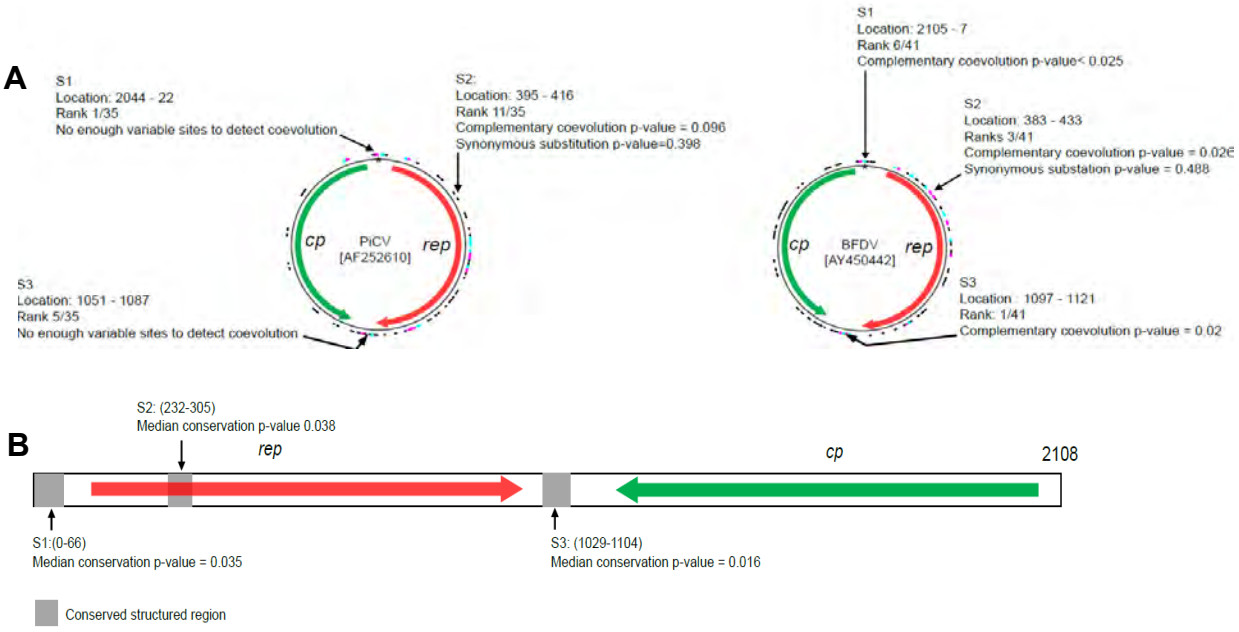


Figure 3-8. Comparison of secondary structure maps of PiCV and BFDV genomes

(A) The genome maps show the locations of the most conserved structural elements (i.e. structural elements within the HCSSs) inferred for *Pigeon circovirus* (PiCV) and *Beak and feather disease virus* (BFDV) genomes, structures have been ranked based on base-pairing conservation scores, synonymous substitution rates and evidence of complementary coevolution favouring the maintenance of base-pairing. Structures are shown using arcs coloured in black or cyan and magenta (to distinguish the two complementary parts of the stem sequences for the ten highest ranked structures). In the PiCV and BFDV datasets, S1 at the origin of replication respectively ranks first and sixth, S2 within the *rep* gene respectively ranks eleventh and third, and S3 within the intergenic region between the *rep* and *cp* stop codons ranks fifth and first. **(B)** Displays a secondary structure conservation map of aligned PiCV and BFDV genomes, indicating three regions containing significantly conserved genomic secondary structural features (associated conservation p-values are given for each region; from Stenzel et al. 2014).

Using the StructureMap visualisation approach and individual nucleotide resolution visualisation of particular structural elements with NAVA, I was able to identify and characterise new structures, across many related species within the geminiviruses, anelloviruses, and parvoviruses, that could be involved in the regulation of gene splicing, transcription, translation and virion or complementary sense genome replication (these are discussed in detail in Chapter 4).

3.4.4 SelectionMap visualisation and the characterisation of selection signals within some ssDNA virus genes

Pigeon circovirus (PiCV) and *Beak and feather disease virus* (BFDV) were analysed using SelectionMap to characterise and compare selection signals between the

movement protein (*mp*), capsid protein (*cp*) *and replication associated protein* (*rep*) genes along with overlaid gene annotation information specifying the locations of functional motifs/domains (Stenzel et al. 2014).

As expected, all genes were found to be evolving predominantly under selection disfavouring amino acid substitutions (i.e. negative selection: average dN/dS < 1). However the *cp* contained about twice the number of sites under positive selection as were observed in the *rep* gene, possibly due to host adaptive immunity more strongly influencing the evolution of the capsid protein (CP).The CP is more exposed to the host immune system than the replication associated protein (Rep; Pogranichnyy et al. 2000; Blanchard et al. 2003). Also, the large proportion of sites (8% in *rep* and 9% in *cp*) detected to be under negative selection but for different amino acids in the different species could represent sites important for host adaptation which, since the last common ancestor of PiCV and BFDV might have diverged under adaptive directional positive selection (i.e. selection favouring host adaptive amino acid substitutions within these proteins). These sites could possibly include those determining differences in the epidemiology, host *ranges,* cell tropisms, transmission modes, replication levels and disease phenotypes of BFDV and PiCV.

On the selection map, it is additionally possible to incorporate information on the locations of conserved secondary structures and functional motifs/domains so as to illustrate the extent to which these influence patterns of selection. Using SelectionMap we compared selection signals between *rep* genes of two monocot-infecting mastreviruses, MSV (n = 88) and PanSV (n = 43), overlaid NASP predicted HCSS elements on these genes and then overlaid additional information about the location of functional motifs/domains (Figure 3-9).

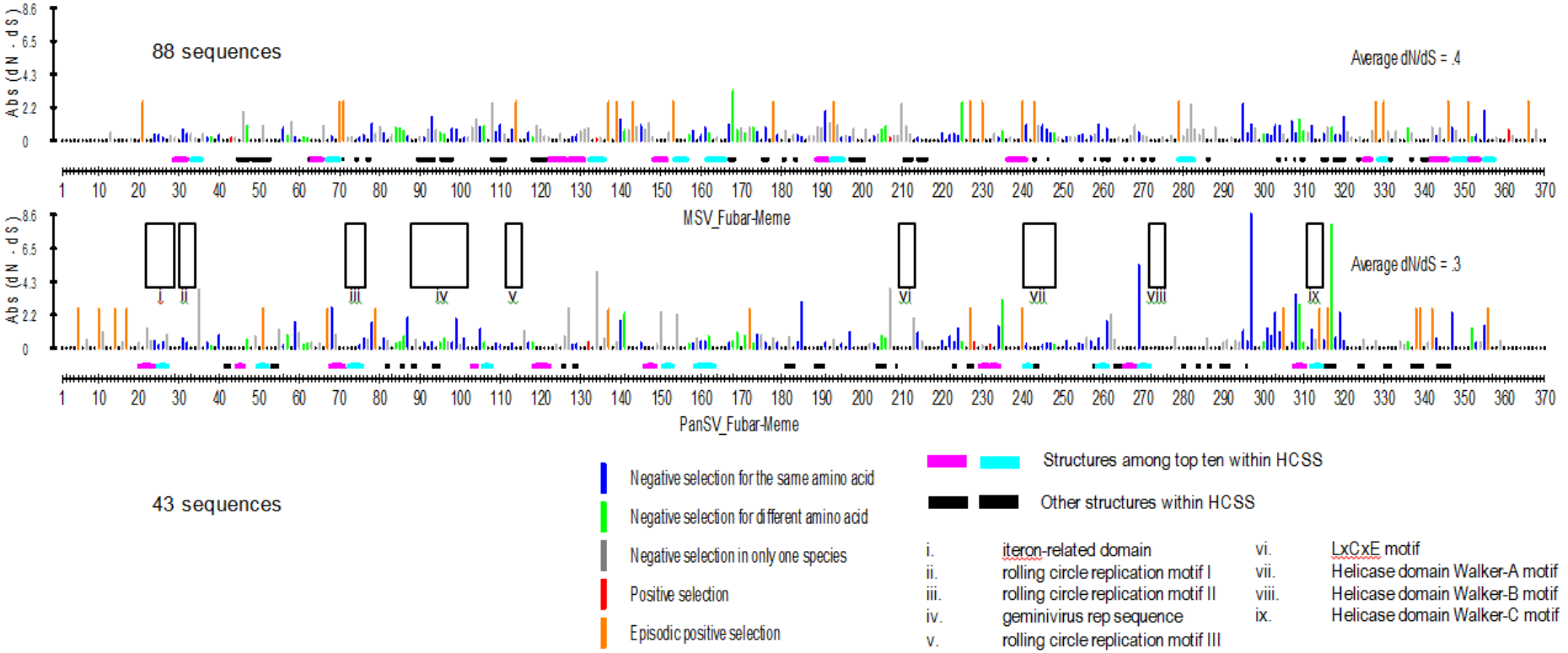


Figure 3-9. Comparison MSV and PanSV *rep* gene selection maps

Signals of natural selection were detected using FUBAR and MEME from MSV and PanSV *rep* sequences (n = 88 and n = 43 respectively). dN=non-synonymous substitution rates and dS=synonymous substitution rates. The height of each bar corresponds to the absolute value of dN-dS that represents the degree of positive or negative selection detected. While sites under negative selection (blue, green and grey) and sites under positive selection (red) were detected by FUBAR with a 0.9 posterior probability cutoff, sites under episodic selection (orange) are sites detected to evolve under positive selection by MEME only (not detected by FUBAR) with a p-value cutoff of 0.1 .The average dN/dS (0.3) indicates that overall the PanSV *rep* gene is evolving under a greater degree of purifying selection than MSV (0.4). The HCSS sizes were 29/69 and 24/79 total NASP predicted structures in MSV and PanSV respectively. The top ten ranking structures are coloured in cyan and magenta. Boxes labelled with roman numbers (from i to ix) represent the locations of conserved domains/motifs found within the Rep protein that is expressed from this gene in both MSV and PanSV.

As mentioned above, both the MSV and PanSV *rep* genes evolve predominantly under strong negative selection (average dN/dS < 1). It is noteworthy that the highly structured regions tend to contain a high proportion of sites evolving under negative selection, particularly at regions associated with motifs/domains i-v, vii and ix (more than 71% structured sites are evolving under negative selection. As a formal statistical test is required, I propose to implement a 2x2 chi-squared test within SelectionMap to enable automatic testing of whether there is an association between paired sites and sites evolving under negative selection. This is important because such an association would imply that, along with biological imperatives to preserve particular functional domains within the proteins expressed from these genes, secondary structures may have an influence on the selection patterns observed, which could help maintain the integrity of nucleotides within these important motifs/domains.

Other eukaryote-infecting ssDNA virus species that have been analysed using this approach include a newly discovered mastrevirus, *Chickpea chlorotic dwarf virus* (CpCDV; n=205), from Sudan whose selection signals were compared to those of three other well sampled/characterised dicot-infecting mastreviruses, *Chickpea chlorosis Australia virus* (CpCAV; n=13), *Chickpea chlorosis virus* (CpCV n=28) and *Tobacco yellow dwarf virus* (TYDV; n=9) in order to gain new and useful biological knowledge (Kraberger et al. 2015).

3.4.5 Nucleic acid folding disruption test applied to HIV-1M genomes

We have applied the nucleic acid fold disruption test to investigate the impact of selection favouring the maintenance of functional RNA secondary structures on genetic recombination within HIV1-M genomes (Golden et al. 2014b). The results indicate that (1) Mimic recombinants (M-recombinants) tend to have significantly fewer aberrant base-pairs than simulated recombinant (S-recombinants; i.e. medians of 259 and 308 aberrant base-pairs with an associated p-value = 0.019) and (2) the number of disrupted base-pairings in M-recombinants was significantly lower than those in the S-recombinants (medians of 52 and 70 disrupted base-pairs, respectively; p-value =0.005).

These results collectively imply that relative to parental sequences, the M-recombinants have better maintained base-pairing configurations compared to S-recombinants. This is consistent with the hypothesis that natural selection has acted both in the short term to purge recombinants with disrupted RNA folds, and possibly in the longer term to modify the genome architecture of HIV to ensure that recombination prone sites correspond with those where recombination will have a minimal impact on genomic RNA folding (Golden et al. 2014b).

A similar large scale analysis to that carried out on HIV-1M sequences was also conducted on a large number of eukaryote-infecting ssDNA virus datasets and is discussed in Chapter 5.

3.5 Conclusion

Adding to existing computational tools, I have introduced a range of novel evolutionary analysis tools for detecting biologically functional secondary structures within virus genomes and analysing their impacts on mutational dynamics and recombination during the evolution of these genomes. The novelty of our tools lies within the approaches that I have implemented which include: (1) methods to predict evolutionarily conserved genomic secondary structures (2) a sophisticated evolutionary model-based approach to computationally validate and rank predicted structural elements in order of their likely biological importance, and (3) an efficient means for visualising structural and selection data so as to allow comparative analyses of biological features between even distantly related virus genomes (i.e. the approaches implemented in NAVA, StructureMap and SelectionMap).

I have shown that, in addition to using computational approaches that minimise false positive rates of structural element prediction (Semegni et al. 2011), the incorporation of natural selection and complementary coevolution information can substantially increase the accuracy with which functional structural elements are predicted. Furthermore, besides various statistical tests I have developed, I have found that the simple visualisation of genome evolution data can greatly enhance our understanding of mechanisms underlying the evolution viruses. Although visualization tools have been implemented and used widely in other areas of

molecular evolution, the introduction of secondary structure related evolutionary data visualisation approaches with NAVA, StructureMap and SelectionMap provides a new perspective in our overarching efforts to understand virus genome evolution.

The main limitations of our methods are: (1) that our structural predictions and our analyses of these predictions are limited to focusing on only one of several metastable structural configurations that any particular nucleotide sequence might in reality form under physiological conditions, (2) due to its computational intensity, the nucleic acid fold disruption test is based on a simple nucleic acid folding method (hybrid-ss-min implemented in UNAFold) without any further tests for the validation of predicted structures such that this test includes a high proportion of structures that probably do not actually form in nature: this could seriously diminish the power of the test, (3) the nucleic acid folding disruption test only considers two recombination breakpoints per analysed genome at a time, which is sometimes unrealistic in that many natural recombinants frequently have more than two detectable break points; (4) another limitation of the folding disruption test is that it does not examine actual parental sequences and it is likely that subtle difference in structure between the sequences identified as recombinants in the recombination analysis and the actual parents could further diminish the accuracy with which the test is able to infer recombination-induced nucleic acid folding disruption.

Feasible future improvements on the tools presented include: (1) full automation of SelectionMap based analysing by incorporating the HyPhy executable programs for seamless selection analysis from within the program, (2) automatic simultaneous overlaying in SelectionMap of gene annotation data from GeneBank files and secondary structure information in a similar approach to that applied in StructureMap, (3) implementing in SelectionMap a 2x2 chi-squared test for associations between genomic regions that are structured and those that are under significant purifying selection (4) increasing the accuracy and speed of folding for the recombination-induced fold disruption test and (5) developing web-based versions of the SelectionMap, StructureMap and SDT computer programs as to improve their accessibility.

All computational tools are freely accessible at:
http://web.cbio.uct.ac.za/~brejnev/ComputationalTools.html.

3.6 Authors' contributions and acknowledgements

Main author's contribution

I developed the following tools: (1) the SDT GUI and its command line and parallel version, (2) Python scripts used to compare degrees of negative selection at the nucleotide level between paired and unpaired nucleotide site alignments, (3) Python and R scripts comparing synonymous substitution rates between codons containing paired and unpaired nucleotide sites, (4) Python and R scripts testing for evidence of complementary coevolution between paired sites, (5) StructureMap, (6) SelectionMap, and (7) the RDP implementation of the recombination-induced nucleic acid folding disruption test. I was also involved in all the analyses performed using these tools.

Co-authors' contributions

1. Darren Martin and Arvind Varsani inspired and supervised the design and development of SDT, StructureMap, SelectionMap and Fold Disruption Test.
2. Yves Semegni and colleagues developed NASP.
3. Art Poon wrote the HyPhy coevolution script.
4. Michael Golden developed NAVA, contributed to the implementation of the fold disruption test and analysed the HIV-1M genomes.

Acknowledgements

I thank the Centre for High Performance Computing in Cape Town and the Information Communication Technology Services Department at the University of Cape Town for providing access to their high-performance computing clusters.

Chapter 4 : Biologically functional secondary structures within eukaryote-infecting ssDNA virus genomes

4.1 Abstract

Single-stranded DNA (ssDNA) viruses have genomes that are potentially capable of forming complex secondary structures through Watson-Crick base-pairing between their constituent nucleotides. A few of the structural elements formed by such base-pairings are, in fact, known to have important functions during the replication of many ssDNA viruses. Unknown, however, are (i) whether numerous additional ssDNA virus genomic structural elements predicted to exist by computational DNA folding methods actually exist and (ii) whether those structures that do exist have any biological relevance. We therefore computationally inferred lists of the most evolutionarily conserved structures within a diverse selection of animal- and plant-infecting ssDNA viruses drawn from the families *Circoviridae*, *Anelloviridae*, *Parvoviridae*, *Nanoviridae*, and *Geminiviridae* and analysed these for evidence of natural selection favouring the maintenance of these structures. While we find evidence that is consistent with purifying selection being stronger at nucleotide sites that are predicted to be base-paired than at sites predicted to be unpaired, we also find strong associations between sites that are predicted to pair with one another and site pairs that are apparently coevolving in a complementary fashion. Collectively, these results indicate that natural selection actively preserves much of the pervasive secondary structure that is evident within eukaryote-infecting ssDNA virus genomes and, therefore, that much of this structure is biologically functional. Lastly, we provide examples of various highly conserved but completely uncharacterized structural elements that likely have important functions within some of the ssDNA virus genomes analysed here.

4.2 Introduction

Besides encoding structural, regulatory and enzymatic proteins, the nucleotide sequences of viral genomes encode a wide range of regulatory motifs associated with, amongst other things, transcription (Yuen and Moss 1987; Hefferon et al. 2006), translation (Shen and Miller 2004), replication (Song and Miller 2004) and genome packaging (Stockley et al. 2013). Other types of biologically relevant information encoded by many nucleotide sequences, including those of viruses, are the thermodynamically stable secondary and tertiary structures that these sequences form under physiological conditions.

While the capacity of single-stranded nucleic acid molecules to fold into higher order structures is crucial in all living organisms for the correct functioning of transfer RNA, ribosomal RNA, messenger RNA and small regulatory RNA molecules, such structures are also particularly important in the biology of many viruses with single-stranded DNA (ssDNA) and single-stranded RNA (ssRNA) genomes. Such structures can play vital roles during the entire viral reproductive cycle including the initiation of genome replication (Ashktorab and Srivastava 1989; Berns 1990; Mohan et al. 1995; Steinfeldt et al. 2001; Gronenborn 2004; Sun and Simon 2006; Faurez et al. 2009), the regulation of gene expression (Ilyinskii et al. 2009), the control of transcription (Koev et al. 1999), translation (Guo et al. 2001; Zuo et al. 2009; Watts et al. 2009) and gene splicing (Moss et al. 2012), and the modulation of host anti-viral responses (Simmonds et al. 2004; Wikström et al. 2007; Wikström et al. 2011).

Within viral genomes, biologically important structural elements tend to be highly conserved across even distantly related species (Simon and Gehrke 2009). In ssRNA viruses, for example, they include the rev response element (RRE) in Human and Simian immunodeficiency viruses (Le et al. 1990; Powell et al. 1997; Fernandes et al. 2012), the *cis*-acting replication elements (CREs) of flaviviruses (You et al. 2004), luteoviruses (Koev et al. 2002), carmoviruses (Sun and Simon 2006), coronaviruses (Raman et al. 2003), alphaflexiviruses (Pillai-Nair et al. 2003), reoviruses (Chen et al. 2001), and picornaviruses (Paul et al. 2000; Nagashima et al. 2003), the internal ribosomal entry site (IRES) elements of flaviviruses (Thurner 2004; Piñeiro and Martinez-Salas 2012), picornaviruses (Pelletier and Sonenberg

1988), pestiviruses (Kolupaeva et al. 2000) and dicistroviruses (Kanamori and Nakashima 2001), and the cap-independent translation elements (CITEs) found in many plant-infecting ssRNA viruses (Miller et al. 2007; Simon and Miller 2013).

Similarly, in many ssDNA virus genomes DNA secondary structural elements have been identified that have crucial biological functions. While in parvoviruses these include structural elements that are essential for genome replication (Ashktorab and Srivastava 1989; Cossons et al. 1996; Sun et al. 2009) and optimal gene expression (Ben-Asher and Aloni 1984; Bohenzky et al. 1988; Resnekov and Aloni 1989; Krauskopf et al. 1991; Perros et al. 1994), in geminiviruses, nanoviruses and circoviruses highly conserved stem-loop structures at their replication origins are essential for the initiation and termination of replication (Orozco and Hanley-Bowdoin 1996; Hafner et al. 1997; Steinfeldt et al. 2001; Cheung 2006). Besides these few examples, however, it is currently unknown how pervasive biologically important secondary structural elements are within the genomes of these viruses (Morozov et al. 1994; Shepherd et al. 2006; Martin et al. 2011c).

It is important to stress the distinction between the simple existence within viral genomes of pervasive stable secondary structures and the biological importance of these structures. Above a certain length even randomly generated single-stranded oligonucleotides will form stable secondary structures (Schultes et al. 2005), and it is therefore plausible that many essentially functionless secondary structural elements might exist within ssDNA and ssRNA viral genomes.

It is, however, theoretically possible to computationally determine the functional importance within viral genomes of secondary structural elements (detected either by computational prediction or by more rigorous laboratory analyses) by simply examining patterns of evolution that are evident within groups of related genome sequences. Specifically, although biologically functional secondary structural elements should be evolutionarily conserved across diverse viral lineages, the nucleotide sequences from which these elements are composed should display distinctive signals of natural selection favouring the maintenance of these structures. Whereas in coding regions these signals might include codon usage biases (Hasegawa et al. 1979; Cardinale et al. 2013), and decreased rates of synonymous substitution (Ngandu et al. 2008), throughout the genome these signals could also

include high rates of reversion substitution (Cheung 2005; Shepherd et al. 2006) and increased frequencies of complementarily coevolving nucleotide pairs – particularly amongst those nucleotides predicted to be base-paired within secondary structural elements (Hofacker et al. 1998; Fernández et al. 2011; Cheng et al. 2012b).

Accordingly, experimental investigations of individual structural elements within some ssDNA virus genomes have clearly demonstrated the existence of strong natural selection favouring the maintenance of these elements. For example, when mutations were experimentally introduced that disrupted particular base-pairings within a stem-loop structure at the origin of replication of the circovirus, *Porcine circovirus* 1 (PCV-1), the disrupted base-pairings were rapidly restored during replication through a DNA polymerase mediated template switching mechanism (Cheung 2004b; Cheung 2004a). Similarly, in the geminivirus, *Maize streak virus* (MSV), mutations that potentially disrupted base-pairings within a complex computationally predicted structural element were found to very predictably revert to the original nucleotide so as to re-stabilise the structural element (Shepherd et al. 2006).

Here we examine the biological relevance of pervasive computationally predicted secondary structures within diverse eukaryote-infecting ssDNA virus genomes. After using a free energy minimisation approach to identify conserved secondary structural elements within groups of closely related full genome sequences, we applied various tests to determine whether mutational processes differed between structured and unstructured genome regions in ways consistent with the evolutionary conservation of the identified structural elements. While we provide strong evidence of extensive biologically relevant secondary structures within eukaryote-infecting ssDNA virus genomes, we further identify what are likely some of the most functionally important uncharacterised structural elements within these genomes.

4.3 Materials and methods

4.3.1 Dataset preparation

All available circovirus, anellovirus, parvovirus, nanovirus and geminivirus full genome sequences were obtained from GenBank (http://www.ncbi.nlm.nih.gov/pubmed/) between April 2011 and August 2012. Full genome sequences for each of the five families were preliminarily aligned separately using MUSCLE (Edgar 2004) implemented in MEGA5 (Tamura et al. 2011) and subdivided into datasets of sequences sharing at least 75% sequence identity. This was done to ensure reasonable alignment accuracy during subsequent sequence analyses (Wilm et al. 2006) while at the same time, providing enough sequence diversity to enable the accurate characterisation of evolutionary processes acting to maintain predicted secondary structures.

A set of 23 datasets was obtained, each containing between 21 and 519 full genome sequences. Each of these full genome sequence datasets was realigned using MUSCLE (with default settings) and, where necessary, manually edited. The resulting alignments will hereafter be referred to as “large datasets” (Table 3-1). This distinction is important because many of the analyses performed could only be carried out on subsets of these datasets. Specifically, from each of the large datasets we first extracted an “intermediate dataset”. In all but four cases each contained one representative sequence from each of the 30 most divergent viral sequence lineages in the large datasets. The exceptional cases were the AnelloTTSuV1, AnelloTTV, ParvoHBoV, and ParvoMPV datasets that respectively contained only 21, 22, 21 and 26 sequences, all of which were included in the intermediate dataset. From each of the intermediate datasets we further extracted a “small dataset” containing one representative sequence of each of the 10 most divergent lineages.

Table 4-1. List of the 23 large datasets obtained

	Name[a]	Families	Constituent virus species	Size[b]
1	CircoPCV	*Circoviridae*	*Porcine circovirus 2*	519
2	CircoCoCV	*Circoviridae*	*Columbid circovirus*	36
3	CircoDGCV	*Circoviridae*	*Duck circovirus, Goose circovirus, Muscovy duck circovirus, Cygnus olor circovirus*	49
4	CircoBFDV	*Circoviridae*	*Beak and feather disease virus*	184
5	AnelloTTSuV1	*Anelloviridae*	*Torque teno sus virus 1*	21
6	AnelloTTSuV2	*Anelloviridae*	*Torque teno sus virus 2, Porcine torque teno virus 2*	44
7	AnelloTTV	*Anelloviridae*	*Torque teno virus*	22
8	ParvoAAV	*Parvoviridae*	*Adeno-associated virus,*	34
9	ParvoHBoV	*Parvoviridae*	*Human boccavirus - 2, 3,4; Porcine boccavirus 1,2; Gorilla boccavirus, Bovine parvovirus 1, Canine minute virus*	21
10	ParvoMPV	*Parvoviridae*	*Mouse parvovirus 4, Rat minute virus, Mouse parvovirus, Minute virus, LuIII virus, Hamster parvovirus*	26
11	NanoBBTV-R	*Nanoviridae*	*Banana bunchy top virus component R*	221
12	NanoBBTV-S	*Nanoviridae*	*Banana bunchy top virus component S*	189
13	NanoBBTV-M	*Nanoviridae*	*Banana bunchy top virus component M*	150
14	NanoBBTV-N	*Nanoviridae*	*Banana bunchy top virus component N*	148
15	NanoBBTV-C	*Nanoviridae*	*Banana bunchy top virus component C*	122
16	GeminiMSV	*Geminiviridae*	*Maize streak virus*	759
17	GeminiWDV	*Geminiviridae*	*Wheat dwarf virus*	138
18	GeminiPanSV	*Geminiviridae*	*Panicum streak virus*	41
19	GeminiTYDV-CpCV	*Geminiviridae*	*Tobacco yellow dwarf virus, Chickpea chlorosis virus, Chickpea yellows virus*	41
20	GeminiCpCDV	*Geminiviridae*	*Chickpea chlorotic dwarf virus*	43
21	GeminiTYLCV	*Geminiviridae*	*Tomato yellow leaf curl virus*	228
22	GeminiEACMV	*Geminiviridae*	*East African cassava mosaic virus, South African cassava mosaic virus*	146
23	GeminiMYVYV	*Geminiviridae*	*Malvastrum yellow vein Yunnan virus, Cotton leaf curl Multan virus isolate, Bhendi yellow vein India virus*	254

[a] The name of the dataset is made up of the prefix of its family and the abbreviation of the main virus species it contains.

[b] The number of full genome sequences in the dataset.

4.3.2 Detection of conserved secondary structural elements within eukaryote-infecting ssDNA virus genomes

Since biologically relevant secondary structural elements are likely to be at least partially conserved during evolution, we used the computer program Nucleic Acid Secondary Structure Predictor (NASP; Semegni et al. 2011), to identify the conserved secondary structural elements within the set of representative full genomic sequences in each of the small datasets.

NASP takes as input a set of aligned nucleic acid sequences and uses Gibbs free-energy (Greiner et al. 1995) and Boltzmann probability (Ding 2003) techniques implemented in the hybrid-ss component of the UNAFold software package (Markham and Zuker 2008) to determine an ensemble of nearly minimum free energy (MFE) folds for each of the input sequences. It then uses a nucleotide-shuffling based permutation test to statistically determine the sets of conserved structural elements within the folded sequences that contribute most to their over-all thermodynamic stability.

Precisely, NASP produces sets of “pairing matrices” for each of the input sequences in each small dataset which are then compressed into a “consensus base-pairing matrix,” called the M matrix, using a weighted sum of the pairing matrices obtained for each of the input sequences (Semegni et al. 2011). The use by NASP of weighted sums in the calculation of its M Matrix is intended to counteract unavoidable sampling biases in sequence datasets so as to ensure that similar structures within very closely related sequences do not make unfair contributions to the “conservation scores” that NASP calculates for the individual structural elements that it identifies.

Importantly, in our study NASP provided a conservation score for each discrete structural element identified within the M-matrices calculated for each of the small datasets, and indicated the subsets of structural elements referred to as “high confidence structure sets” (HCSSs), that accounted for the analysed sequences having significantly lower MFE scores than those expected in randomised sequences with identical base compositions (Semegni et al. 2011). Whereas the

conservation scores for the individual structural elements provided an obvious way of ranking these in order of their likely biological relevance, the demarcation of HCSSs provided an objective means of focusing further analyses into the biological relevance of secondary structures on just the structural elements that are most likely to really occur.

In our NASP analysis, sequences were folded as either linear (for the three parvovirus small datasets) or circular (for all 20 other small datasets) ssDNA at either 37oC (for animal-infecting circoviruses, anelloviruses and parvoviruses) or 25oC (for plant-infecting geminiviruses and nanoviruses) under 0M magnesium and 1M sodium ionic conditions. The HCSS was identified using 100 nucleotide shuffling permutations with a permutation p-value threshold of 0.05. In all subsequent analyses, the only nucleotides considered as being paired within secondary structures were those occurring within columns of the large and intermediate dataset nucleotide sequence alignments that corresponded with nucleotides identified by NASP as being paired within the HCSS. Whereas these paired nucleotides were referred to as occurring at “paired-sites”, all other nucleotides were referred to as occurring at “unpaired-sites”.

4.3.3 Neutrality tests for elevated negative selection at paired sites

Structural elements that increase the fitness of virus genomes are expected to be selectively preserved such that selection disfavouring nucleotide substitutions should be stronger at paired-sites than at unpaired-sites. Specifically, paired-sites might display stronger evidence of purifying selection than unpaired-sites in neutrality tests such as those proposed by Tajima (Tajima 1989) and Fu and Li (Fu and Li 1993). We calculated Tajima’s D and Fu and Li’s F statistics for the paired- and unpaired-sites in each of the 23 large datasets.

Since in all 23 of the analysed large datasets there were invariably fewer paired-sites (i.e. those paired within the HCSS) than there were unpaired-sites (i.e. the remainder of sites in the various datasets) we devised a permutation test involving the random selection of identical numbers of paired and unpaired-sites and the comparison of summary selection statistics between these paired- and unpaired-site datasets.

From each large dataset we produced 100 datasets each consisting of sites (i.e. entire large dataset alignment columns) randomly sampled with replacement from the pool of unpaired-sites. These permutation datasets contained the same numbers of sites as their corresponding paired-site datasets. Tajima's D and Fu and Li's F statistics were then calculated for all of the paired-site and permutation datasets. For each of the 23 datasets, the probability that purifying selection was operating more strongly on paired-sites than on unpaired-sites was calculated as being approximately equivalent to the proportion of times the D and F statistics calculated for the paired-site dataset were lower than those calculated for the 100 permuted datasets.

4.3.4 Codon-based tests of synonymous substitution rates at paired versus unpaired genomic sites

Biologically important structural elements that occur within protein-coding sequences are expected to display both selection at the codon level which favours the preservation of amino acid sequences (i.e. selection disfavouring non-synonymous substitutions), and selection at the nucleotide level favouring the maintenance of base-pairing interactions within the structural elements. This "double selection" at codons that contain constituent nucleotides which form base-pairs within biologically important secondary structures should result in such codons displaying synonymous substitution rates that are lower than those occurring in codons consisting of unpaired nucleotides.

To determine whether codons corresponding to paired genomic sites displayed significantly lower synonymous substitution rates than those occurring at unpaired genomic sites, nucleotide sequences corresponding to known/suspected genes were extracted from each of the 23 intermediate dataset alignments. Within each of the resulting "gene datasets" all sites encoding amino acids in two or more different frames were removed. Following this, 43 gene datasets containing 200 or more nucleotide sites were retained for further analysis (see Supplementary Table 1 for details of these datasets). Gene datasets excluded from this set because they retained too few sites included the *ren*, *mp* and *trap* genes of the GeminiTYLCV, GeminiEACMV and GeminiMYVYV datasets; *ORF2* and *ORF3* of the AnelloTTSuV1,

AnelloTTSuV2 and AnelloTTV datasets; the *vp1* and *vp2* of the ParvoHBoV dataset; and the *vp*1 of the ParvoMPV.

Two methods were used to estimate synonymous substitution rates at individual codon sites within the 43 gene datasets: Partitioning Approach for Inference of Selection (PARRIS; Scheffler et al. 2006) and Fast Unconstrained Bayesian Approximation (FUBAR; Murrell et al. 2013). Both of these methods apply the time-reversible MG94 codon substitution model which utilises a 61 X 61 codon substitution matrix (Muse and Gaut 1994) and both allow independent distributions for synonymous and non-synonymous rates. PARRIS is a random effects likelihood (REL) method permitting the use of only three discrete categories for each rate. FUBAR on the other hand is an approximate Bayesian method which permits the use of many more rate classes (20 in our case) so as to increase the resolution with which, for example, subtle differences in selection pressures operating on individual codons can be distinguished.

Both FUBAR and PARRIS rely on the use of phylogenetic trees to describe the evolutionary relationships of the sequences being analysed. While it is well established that genetic recombination undermines the accuracy of phylogenetic inference (and by extension many phylogenetics-oriented codon-based selection analysis methods; Scheffler et al. 2006), it was likely that many of the sequences being analysed here were recombinant (Padidam et al. 1999a; Navas-Castillo et al. 2000; Shackelton et al. 2007; Lefeuvre et al. 2009b; Julian et al. 2013). It was therefore necessary to take steps to explicitly account for recombination within these analyses. Accordingly, prior to selection analyses the Genetic Algorithm for Recombination Detection (GARD; Kosakovsky Pond et al. 2006) method was used to detect recombination breakpoint sites. These sites were then used to partition the alignment into “mostly recombination free” (it is unlikely that every recombination event was detected and accounted for) sub-alignments. For each of these sub-alignments a phylogenetic tree was inferred and these trees were collectively used as inputs for the PARRIS and FUBAR analyses – both of which allow phylogenetic tree topologies and branch lengths to vary across different alignment partitions so as to facilitate accurate inference of natural selection in the presence of recombination (Scheffler et al. 2006; Murrell et al. 2013).

Within each gene dataset, each codon was categorised as being a “paired-codon” if its third position nucleotide was paired within the relevant HCSS, and an “unpaired-codon” if its third position nucleotide was not a paired nucleotide within the relevant HCSS. Using a Mann-Whitney U test, we determined whether in each of the 43 gene datasets paired-codons had significantly lower synonymous substitution rates than unpaired-codons. All p-values thus calculated were step-down corrected to account for multiple testing.

4.3.5 Testing whether paired sites complementarily coevolve

Mutations at paired-sites may be tolerable within biologically important structural elements if they are followed by compensatory mutations that restore base-pairing (Shepherd et al. 2006). We detected evidence of complementary coevolution between pairs of sites within the large datasets using a customised version of the SPIDERMONKEY coevolution script written in HyPhy(Pond et al. 2005). For any chosen pair of sites within a large dataset, the script compares the standard independent sites 4 X 4 HKY85 nucleotide substitution model (Hasegawa et al. 1985) to a 16 X 16 Muse-Modified HKY85 coevolution model (called M95; Muse 1995) to determine which of these best describes the evolution of individual site pairs. In our case entries in the M95 16 X 16 substitution matrix representing changes that potentially maintain base-pairing (including both Watson-Crick pairings such as A-T and G-C and the wobble pair T-G) are multiplied by a pairing factor, λ, and those involving changes between paired- and unpaired-states are multiplied by $1/\lambda$. A maximum likelihood ratio test enabled us to determine whether nucleotide pairs were coevolving. Whereas $\lambda > 1$ for particular coevolving site pairs indicated that they displayed a tendency towards complementary coevolution, $\lambda = 1$ indicated a tendency towards site pairs evolving independently and $\lambda < 1$ indicated a tendency towards them coevolving non-complementarily. We identified site pairs displaying strong evidence of complementary coevolution as those with both associated maximum likelihood estimates of $\lambda > 1$, and Muse 95 versus HKY85 likelihood ratio test p-values <0.05.

Importantly, due to computational intensity considerations, we restricted our analyses to testing for coevolution only between (i) pairs of sites that were within 100 nucleotides of one another and (ii) pairs of nucleotides that were polymorphic in the

input dataset. Also, since recombination can undermine the accuracy with which the phylogenetic trees used to detect coevolution reflect the actual evolutionary relationships of the analysed sequences, we took steps to account for recombination in our analyses. Each large dataset was analysed for recombination using Recombination Detection Program (RDP) version 4.13 (Martin et al. 2010) which produced a "distributed alignment" in which fragments of recombinant sequences derived from different parental viruses were split up into different sequences. For each of the 23 distributed alignments obtained, a 100 nucleotide sliding window was moved 1 nucleotide step at time across the alignment. At every step the N longest nucleotide sequences were selected (where N is the number of sequences in the original alignment), and saved to an alignment file. Sequences in consecutive windows containing exactly the same N longest sequences were merged into one file (ensuring that no sites from the original alignment were duplicated in the merged alignment). Maximum-likelihood (ML) phylogenetic trees were inferred for each of the resulting alignments under the HKY85 nucleotide substitution model using PhyML3.0 (Guindon et al. 2010). Each of these alignments and their corresponding phylogenetic tree was used as inputs for our complementary coevolution analysis.

4.3.6 Customized computational tools

All computer scripts used in all the analyses conducted can be downloaded from http://web.cbio.uct.ac.za/~brejnev/downloads/ComputationalTools/ and a customised computer program and datasets for visualisation of predicted structural elements can be downloaded from http://web.cbio.uct.ac.za/~brejnev/downloads/DOOSS.zip (unzip these files and please see the README file for instructions).

4.4 Results and discussion

4.4.1 Numerous evolutionarily conserved secondary structures are evident within eukaryote-infecting ssDNA virus genomes

We assembled 23 full-genome sequence datasets representing the families *Circoviridae*, *Anelloviridae*, *Parvoviridae*, *Nanoviridae and Geminiviridae* (Table 4-1). In each of these between 69 and 316 conserved secondary structural elements were identified using the minimum free energy (MFE) approach implemented in the

computer program NASP (Semegni et al. 2011). From these lists of conserved structural elements, NASP identified subsets of between five and 132 “high confidence” structural elements. These lists, hereafter referred to as “high confidence structure set” (HCSS) lists, contained those structures primarily responsible for the analysed genomes having greater degrees of predicted structural stability than those of randomised sequences with identical nucleotide compositions. Notably, most of the previously described biologically important structures in these viral genomes were present within the top 30% of structures in the HCSS lists of their respective datasets. These included hairpin structures at the virion strand origins of replication in circoviruses, nanoviruses, and geminiviruses and T-shaped structures required for replication in parvoviruses. The genomic coordinates of structures within all 23 of the HCSSs were mapped onto their respective genomes (Figure 4-1; Figure 4-2; Supplementary Table 2).

Clearly our computational approach for predicting secondary structure suggests that there exist more conserved genomic secondary structures within many of these ssDNA virus genomes than is currently appreciated. It is plausible that, as is the case with currently known secondary structures within these genomes, many of the uncharacterised conserved structures may have been preserved during evolution due to their biological importance.

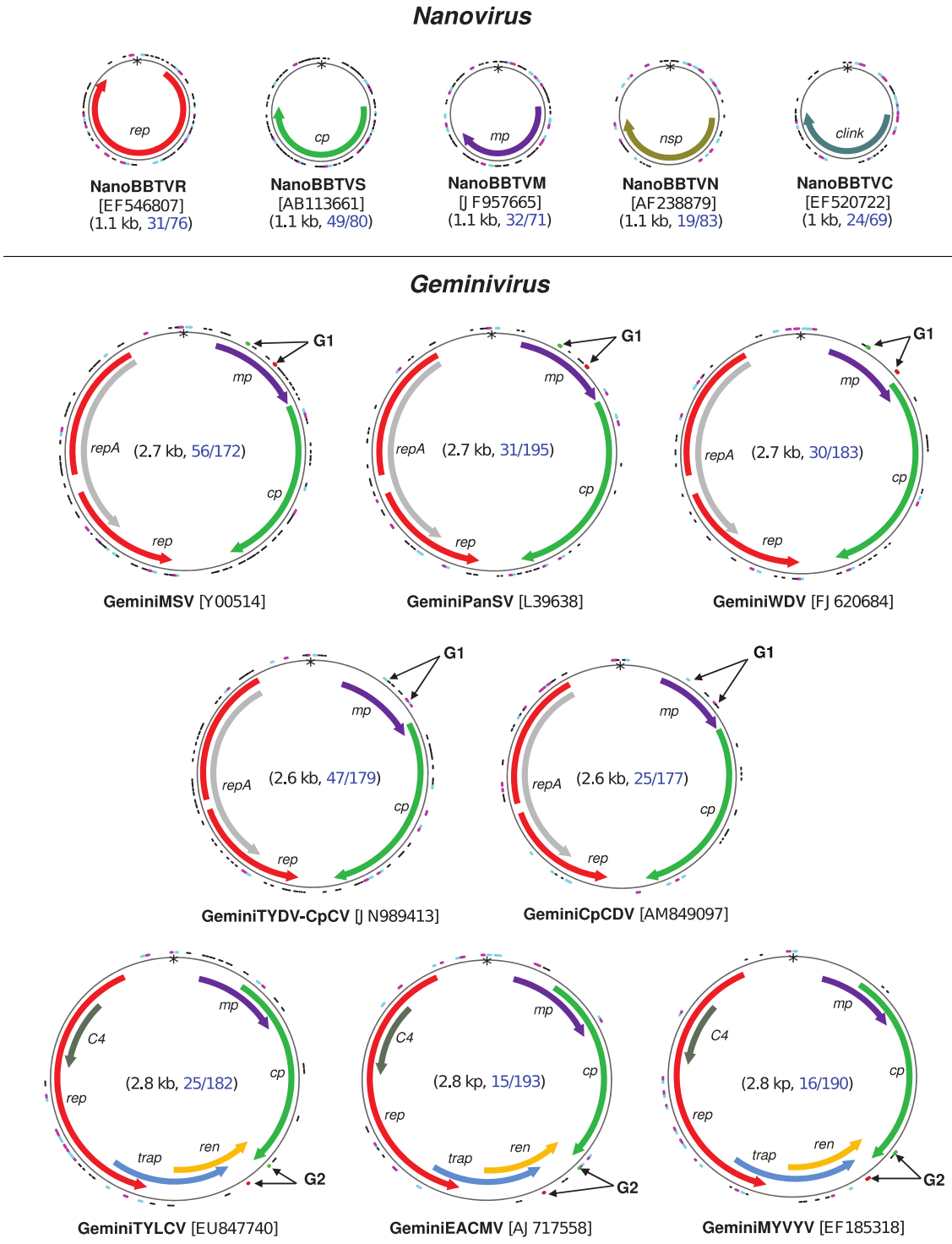


Figure 4-1. Secondary structure map of plant infecting ssDNA viruses

Genome organisation maps of plant-infecting ssDNA viruses. In each map, the arcs represent the coordinates of identified structural elements within "high confidence structure sets" (HCSS). These highly conserved structural elements are those primarily responsible for the estimated structural stability of the analysed genomes being greater than that of randomised sequences with identical nucleotide compositions. The ten structures collectively displaying the greatest degrees of base-pairing conservation, lowest associated synonymous substitution rates and greatest degrees of complementary coevolution between paired nucleotides are shown using arcs in cyan and magenta (to distinguish the two complementary parts of the stem sequences). All remaining structures are

shown using black arcs. Black arrows indicate examples of currently uncharacterised but likely biologically functional structures that are apparently conserved across multiple datasets (coloured in green and brown when these were not ranked among the top ten within their respective HCSS). The known secondary structural elements at the virion strand origins of replication are indicated by a star symbol at the 12 o'clock position of the genomes. Numbers in brackets indicate the lengths of the genomes in kilobases (kb) and the ratio of the numbers of high confidence structures over the total numbers of predicted secondary structures. Italicized abbreviations of gene names: *rep*=replication associated protein; *cp* = coat protein; *mp* = movement protein; *clink* = cell cycle link protein; *nsp* = nuclear shuttle protein; *ren* = replication enhancer protein; *trap* = transcription activator protein.

Although directly testing the biological relevance of any one of the identified potential structures would require detailed mutational analyses of their constituent nucleotides within the context of infectious cloned genomes or analysis of recombinant viral constructs, followed by extensive quantitative fitness assays (McCormack and Simon 2004; Staplin and Miller 2008), there are less cumbersome computational approaches for testing whether the identified structures collectively (as opposed to individually) are likely to have any biological relevance. In this regard the biological relevance of the structures in our HCSSs could be tested by comparing how their constituent nucleotides evolve relative to those at positions located outside of the HCSSs.

Therefore, in our subsequent analyses we focused on testing whether, relative to the remainder of nucleotides in the genomes (hereafter referred to as unpaired nucleotides), nucleotides predicted to be base-paired within the HCSS (hereafter referred to as paired nucleotides), are evolving in ways suggestive of their parental structural elements possessing some biological function.

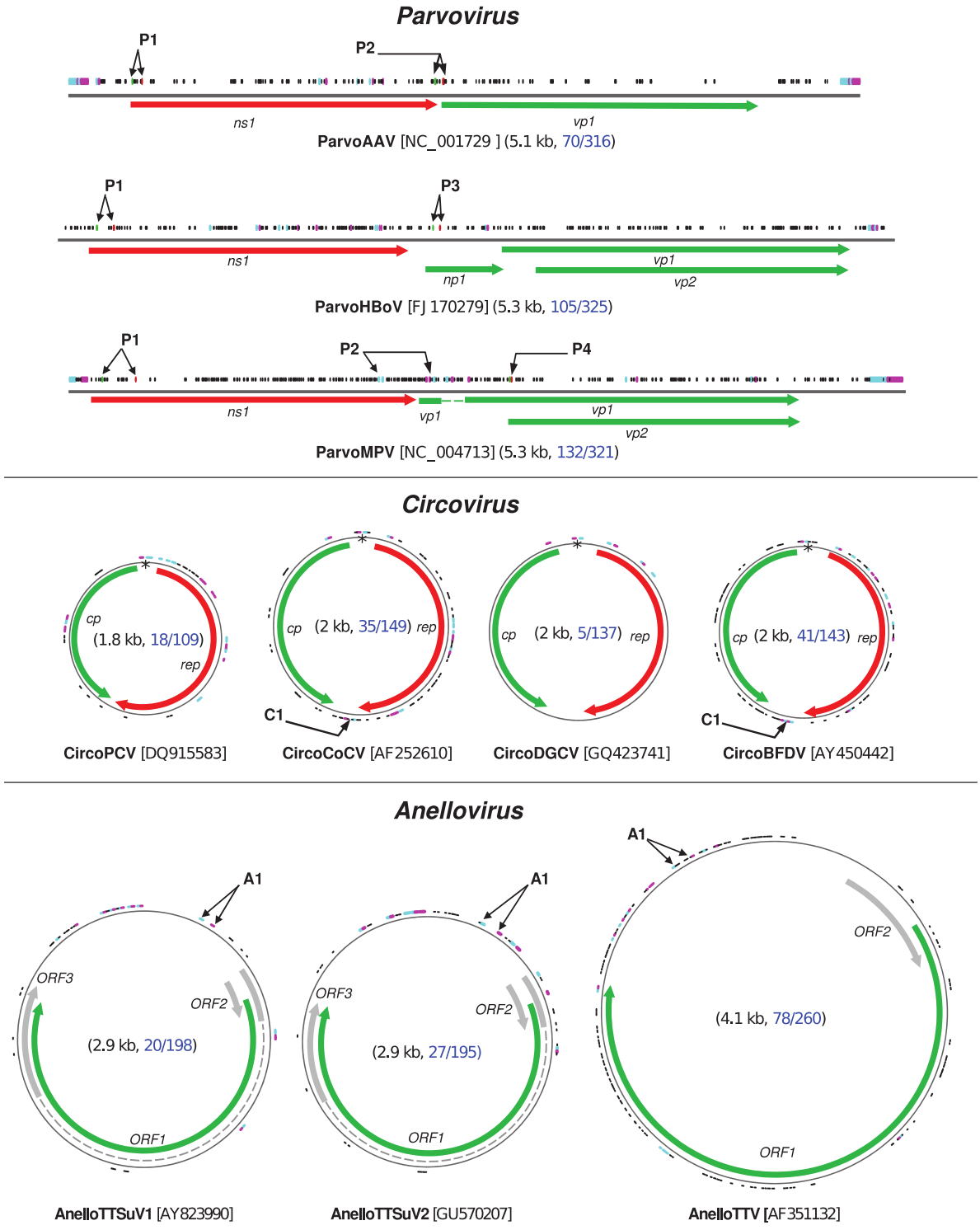


Figure 4-2. Secondary structure map of animal infecting ssDNA viruses

Genome organisation maps of animal-infecting ssDNA viruses. In each map, the arcs (for circular genomes) and vertical lines (for linear genomes) represent the coordinates of identified structural elements within "high confidence structure sets" (HCSS). The ten structures collectively displaying the greatest degrees of base-pairing conservation, lowest associated synonymous substitution rates and greatest degrees of complementary coevolution between paired nucleotides are shown using arcs in cyan and magenta (to distinguish the two complementary parts of the stem sequences). All remaining structures are shown using black arcs/vertical lines. Black arrows indicate examples of currently

uncharacterised but likely biologically functional structures that are apparently conserved across multiple datasets (coloured in green and brown when these were not ranked among the top ten structures within their respective HCSS). The known secondary structural elements at the virion strand origins of replication are indicated by a star symbol at the 12 o'clock position of the genomes. Numbers in brackets indicate the lengths of the genomes in kilobases (kb) and the ratio of the numbers of high confidence structures over the total numbers of predicted secondary structures. Italicized abbreviations represent the gene names encoding the following proteins: *rep*=replication associated protein; *cp* = coat protein; *ns1* = large non-structural protein; *np1* = small non-structural protein; *vp1* = major virion/viral protein; *vp2* = minor virion/viral protein and *ORF* = unnamed open reading frame.

4.4.2 Purifying selection is apparently strongest at paired nucleotide sites

Nucleotide sites involved in biologically important base-pairing interactions might be expected to evolve under a greater degree of purifying selection (selection against change) than sites that are not base-paired. Also, sequences evolving under purifying selection are expected to have lower frequencies of minor allele polymorphisms than those evolving under neutral selection and are, therefore, expected to yield negative values for Tajima's D and Fu and Li's F statistics (Tajima 1989; Fu and Li 1993). If purifying selection was stronger at base-paired-sites than at unpaired-sites we would expect to see lower values of the D and F statistics for datasets containing only base-paired-sites (constructed from the large dataset alignments by removing all unpaired nucleotide sites) than for datasets containing only unpaired-sites (constructed from the large dataset alignments by removing all paired nucleotide sites).

In all but one of the 23 large datasets (the exception being the circovirus dataset, CircoDGCV; Table 4-1), both the paired and unpaired-site alignments consistently yielded negative D and F test static values (see Table 4-2). In 16/23 of the datasets both the D and F statistics were lower for the paired-site than for the unpaired-site datasets. In 5/23 datasets (the anellovirus datasets AnelloTTSuV2 and AnelloTTV, the parvovirus datasets, ParvoAAV, ParvoHBoV and the geminivirus dataset GeminiTYLCV) either the D or F statistics were lower for the paired-site datasets than for unpaired-site datasets. In only 2/23 cases (CircoDGCV and AnelloTTSuV1) did the unpaired-site dataset yield both values of D and F statistics lower than those yielded by the paired-site datasets.

This observation is consistent with our hypothesis that, if paired-sites within the 23 HCSS lists really do reside within biologically important secondary structures, they should display higher degrees of purifying selection than other sites within the analysed genomes.

Table 4-2. Tajima's D – Fu and Li statistics for paired and unpaired genome site alignments

		Tajima's D			Fu & Li's F		
	Large dataset	**Paired[a]**	**Permuted unpaired[b]**	**p-value**	**Paired[c]**	**Permuted unpaired[d]**	**p-value**
1	CircoPCV	-2.08	-1.90	0.06	-5.33	-4.91	0.13
2	CircoCoCV	-2.48	-1.75	< 0.01	-4.98	-3.17	< 0.01
3	CircoDGCV	1.44	0.71	0.99	1.40	0.58	0.96
4	CircoBFDV	-1.53	-1.33	< 0.01	-2.54	-1.77	< 0.01
5	AnelloTTSuV1	-0.72	-1.09	1	-0.19	-1.09	1
6	AnelloTTSuV2	-0.96	-0.95	0.54	-1.09	-1.38	0.92
7	AnelloTTV	-0.10	-0.09	0.5	-0.95	-0.95	0.49
8	ParvoAAV	-0.56	-0.55	0.53	0.80	0.67	0.85
9	ParvoHBoV	-1.09	-1.02	0.07	0.25	0.18	0.78
10	ParvoMPV	-0.22	-0.12	0.08	-0.31	-0.01	0.01
11	NanoBBTVR	-1.31	-1.21	0.28	-3.08	-2.27	0.03
12	NanoBBTVS	-0.96	-0.65	< 0.01	-2.79	-2.00	0.02
13	NanoBBTVM	-0.77	-0.38	0.05	-2.32	-1.74	0.14
14	NanoBBTVN	-1.61	-1.45	0.13	-4.26	-3.36	0.02
15	NanoBBTVC	-1.44	-0.80	< 0.01	-5.03	-3.28	< 0.01
16	GeminiMSV	-2.02	-1.72	< 0.01	-5.34	-3.26	< 0.01
17	GeminiWDV	-1.26	-0.61	< 0.01	-4.03	-2.99	0.04
18	GeminiPanSV	-0.91	-0.61	< 0.01	-0.53	-0.05	< 0.01
19	GeminiTYDV-CpCV	-1.28	-0.55	< 0.01	-2.30	-0.29	< 0.01
20	GeminiCpCDV	-0.88	-0.10	< 0.01	-0.71	0.19	< 0.01
21	GeminiTYLCV	-1.67	-1.71	0.61	-3.44	-3.23	0.26
22	GeminiEACMV	-1.33	-0.83	< 0.01	-2.58	-1.31	< 0.01
23	GeminiMYVYV	-1.23	-0.78	< 0.01	-3.25	-1.28	< 0.01

[a]Tajima's D for paired-sites alignments corresponding to the HCSS.

[b]Average Tajima's D for 100 permuted alignments sampled for the unpaired-sites.

[c]Fu and Li's F for paired-sites alignments corresponding to the HCSS.

[d]Average Fu and Li's F for 100 permuted alignments sampled for the unpaired-sites

However, to test whether values of these statistics were significantly lower at paired-sites than at unpaired-sites in the 21/23 large datasets displaying the expected trend, for each dataset we applied a permutation test involving resampling of identical numbers of sites from the unpaired-site dataset as were present within the paired-site dataset (in each dataset unpaired-sites were invariably more numerous the paired-sites). In each case a p-value was computed as the proportion of the 100 permutated unpaired-site datasets that yielded lower D or F values than the corresponding paired-site dataset. In this test a p-value <0.05 indicates that you would expect to see a D or F value for an unpaired-site dataset that was lower than

that of its corresponding paired-site dataset less than 5% of the time if the null model of neutral evolution was true.

In 11/23 datasets both the D and F statistic permutation tests yielded evidence that paired-sites within the HCSS lists experience significantly stronger (p-values < 0.05) purifying selection than the remainder of genomic sites. In a further 6/23 cases, either the D, or F statistic test yielded at least marginal evidence (p-values < 0.08) of paired-sites experiencing stronger purifying selection than unpaired-sites. Therefore, in only 6/23 cases, was there absolutely no evidence of paired-sites experiencing significantly stronger purifying selection than unpaired-sites.

Interestingly, all three of the analysed anellovirus datasets were among the six datasets with no evidence of purifying selection acting on paired-sites. It is perhaps also noteworthy that of the eleven datasets displaying strong evidence of base-pairing associated with purifying selection, only two (both of them circoviruses, CircoCoCV and CircoBFDV) were from the ten mammal- and bird-infecting virus datasets. While it is not possible to directly compare the plant- and animal-infecting virus datasets to one another, it is plausible that increased structural stability afforded by the lower physiological temperatures of plants relative to animals might contribute to the genomic structures of the plant viruses being more evolutionarily stable than those of their warm-blooded animal counterparts. A more mundane explanation, however, could simply be that our animal virus datasets were, in general, substantially smaller than our plant virus datasets and that our analysis therefore, simply lacked sufficient power to differentiate between the numbers of low frequency polymorphisms within the paired and unpaired dataset fractions.

Regardless of possible differences between animal and plant viruses, collectively these results indicate that a substantial proportion of paired-sites within at least 17/23 of the HCSSs are evolving in a manner that is consistent with many of these structures being evolutionarily preserved.

4.4.3 Synonymous substitution rates are unusually low at paired genomic sites

We hypothesised that selection favouring the maintenance of base-pairing within secondary structures might be particularly evident when these structures occurred within protein-coding regions of the genome. Essentially, we investigated whether codons in which third codon position nucleotides were predicted to be base-paired within the HCSSs had significantly lower synonymous substitution rates than those with unpaired nucleotides in the third codon position.

Synonymous substitution rates at individual codon sites within 43 gene datasets (Supplementary Table 1) were inferred using the random effects likelihood selection analysis methods, PARRIS (Scheffler et al. 2006) and FUBAR (Murrell et al. 2013). These methods indicated that in 27/43 of these datasets, the median substitution rates of codons with paired third position nucleotides were significantly lower than those of codons with unpaired third position nucleotides (multiple comparison corrected Mann-Witney U-test p-value < 0.05). An additional five datasets yielded similar evidence but only with one of the two selection analysis methods (Table 4-3).

Table 4-3. Comparison of synonymous substitution rates at paired and unpaired codon-sites

	Datasets	Genes studied	Number of sequences	PARRIS[a]	FUBAR[b]
1	CircoPCV	*rep*, *cp*	30, 29	*rep*	*rep*
2	CircoCoCV	*rep*, *cp*	30, 30	*rep*	*rep*
3	CircoDGCV	*rep*, *cp*	30, 30	*rep*	*rep*
4	CircoBFDV	*rep*, *cp*	30, 29	*rep*, *cp*	*rep*, *cp*
5	AnelloTTSuV1	*ORF1*	17	*ORF1*	*ORF1*
6	AnelloTTSuV2	*ORF1*	30	*ORF1*	*ORF1*
7	AnelloTTV	*ORF1*	21	*ORF1*	*ORF1*
8	ParvoAAV	*ns1, vp1*	23, 30	*ns1, vp1*	*ns1, vp1*
9	ParvoHBoV	*ns1, np1*	21, 21	*ns1*	*ns1*
10	ParvoMPV	*ns1, vp2*	25, 18	*ns1*	*ns1, vp2*
11	NanoBBTV-R	*Rep*	28	*rep*	*rep*
12	NanoBBTV-S	*Cp*	29	*cp*	*cp*
13	NanoBBTV-M	*Mp*	27	*mp*	*mp*
14	NanoBBTV-N	*Nsp*	27	-	-
15	NanoBBTV-C	*Clink*	30	-	-
16	GeminiMSV	*rep*, *cp*, *mp*	30, 30, 30	*rep*, *cp*	*rep*, *cp*, *mp*
17	GeminiWDV	*rep*, *cp*, *mp*	30, 30, 30	*cp*, *mp*	*cp*, *mp*
18	GeminiPanSV	*rep*, *cp*, *mp*	30, 30, 30	-	-
19	GeminiTYDV-CpCV	*rep*, *cp*, *mp*	30, 30, 30	*cp*	*cp*
20	GeminiCpCDV	*rep*, *cp*, *mp*	30, 30, 30	*rep*, *cp*, *mp*	*rep*, *cp*, *mp*
21	GeminiTYLCV	*rep*, *cp*	27, 30	-	-
22	GeminiEACMV	*rep*, *cp*	30, 30	*rep*	*rep*
23	GeminiMYVYV	*rep*, *cp*	29, 28	*rep*	*rep*

[a]Gene alignments in which the synonymous substitution rates (computed using PARRIS) at paired codon sites are significantly (Mann Whitney U test p-value < 0.05) lower than those at unpaired codon sites.

[b]Gene alignments in which the synonymous substitution rates (computed using FUBAR) at paired codon sites are significantly (Mann Whitney U test p-value < 0.05) lower than those at unpaired codon sites.

The results of these analyses therefore strongly support our hypothesis that two layers of selection – one operating at the amino acid sequence level and the other at the nucleotide sequence level – are likely acting on nucleotide sites within the HCSSs that fall within coding regions. This suggests not only that many of the predicted secondary structures represented within the HCSSs really do exist (either within single-stranded genomic DNAs themselves, or within the RNA transcripts that are produced from them), but that these structures likely also make a substantial contribution to the fitness of the genomes within which they reside.

While evidence of lower degrees of nucleotide polymorphism and decreased synonymous substitution rates at paired-sites than at unpaired-sites provides strong support for the existence of many of the predicted secondary structural elements within the HCSSs, it must be stressed that this result does not necessarily imply that these elements are biologically functional. The reason for this is that besides

influencing which arising mutations are deleterious and which are neutral (and, therefore, which mutations are likely to be purged from populations by natural selection), the presence of secondary structures within ssDNA genomes could potentially also influence the basal rates at which sites within these genomes become mutated (Simmonds and Smith 1999), simply because base-paired nucleotides might be predisposed to lower mutation rates than their unpaired counterparts (Frederico et al. 1990; Xia and Yuen 2005).

4.4.4 In short-term evolution experiments mutations tend to preferentially accumulate at unpaired sites

If paired-sites within the HCSSs really do form base-pairs within genomic secondary structures, we hypothesized that these sites might accumulate fewer mutations than unpaired-sites. We tested this hypothesis using mutation data from a series of previously published short-term evolution experiments. In one experiment infectious cloned genomes of two Maize streak virus isolates (called MSV-MatA and MSV-VW) closely related to those in the GeminiMSV dataset, were used to infect maize plants (Monjane et al. 2012). In another experiment infectious cloned genomes of a Tomato yellow leaf curl virus isolate (called TYX) and a Tomato leaf curl Comoros virus isolate (called TOX; both closely related to sequences included in the GeminiTYLCV dataset) were used to infect tomato plants (Martin et al. 2011c).

While over 101 days post-infection the MSV-MatA and MSV-VW genomes were noted to have accumulated 41 and 33 mutations, respectively, at 52 distinct nucleotide sites, over 120 days the TYX and TOX genomes had respectively accumulated 31 and 105 mutations at 135 distinct nucleotide sites. As described previously for our small datasets, we predicted the secondary structures of each genome pair using NASP in order to obtain, for each pair, its own specific HCSS. We used these HCSSs to construct two-by-two contingency tables for paired-sites (sites predicted to be paired within the HCSS) and unpaired-sites (all sites in the genome other than the HCSS paired-sites) versus variable sites (those where mutations occurred) and invariable sites (those where mutations did not occur) and used these in a Fisher's exact test (Fisher 1922), to assess whether variable sites were significantly clustered outside rather than inside paired-sites.

For MSV-MatA and MSV-VW 11/52 variable sites (~21%) were located at paired nucleotide sites (939/2641 or ~36% of considered sites) within the HCSS, yielding significant evidence (p-value = 0.019) that mutations tended to occur more frequently at unpaired nucleotides. Similarly, for TYX and TOX only 5/135 variable sites (~4%) were located at paired nucleotide sites (237/2724 or ~9% of considered sites) within the HCSS regions, indicating a significant tendency (p-value = 0.021) for mutations to accumulate more frequently at unpaired nucleotide sites.

Although no analogous experimental data is currently available for any of the other plant- and animal-infecting ssDNA viruses investigated here, it is nevertheless important that even in short term geminivirus evolution experiments such as these, where selection has not had prolonged periods to purge slightly deleterious mutations, there remains such an obvious trend for mutations to preferentially occur at unpaired-sites.

Unfortunately, even though these experiments were short-term (lasting between 101 and 120 days), it remains possible that selection, in addition to a decreased biochemical predisposition to mutation, was responsible for the relatively lower mutation frequencies at paired-sites within these genomes. While still consistent with our hypothesis that selection is acting on secondary structures to maintain their biological functionality, these results suggest that the alternative hypothesis – that base-paired-sites within secondary structures are simply biochemically predisposed to mutate more slowly than unpaired-sites – is also entirely plausible.

Therefore, although we had established up to this point that secondary structures are likely quite pervasive within ssDNA virus genomes, we were unable to definitively attribute the apparent evolutionary conservation of these structures to natural selection favouring the maintenance of their biological functionality.

4.4.5 Base-paired sites tend to complementarily coevolve

It is expected that, independent of different basal mutation rates at paired- and unpaired-sites, nucleotide substitutions that occur at paired-sites within biologically functional secondary structures might only be tolerable if coupled with complementary substitutions that reconstitute base-pairing. Therefore, in order to

test for natural selection acting to maintain secondary structures without the confounding effects of base-pairing dependent basal mutation rate variation, we directly tested for evidence of paired-sites within the HCSSs coevolving with one another in a manner consistent with the maintenance of their base-pairing. Specifically, we tested for associations between sites predicted to be base-paired within the HCSSs and sites detectably coevolving in a complementary fashion within the 23 large datasets. For each large dataset we performed a two-by-two contingency test of site pairs predicted to be paired versus unpaired on the one hand, and sites predicted to be coevolving versus not coevolving on the other.

In all but one circovirus dataset, CircoCoCV, we found strong significant associations (multiple testing corrected p-values <0.0001) between paired-sites within the HCSSs and sites for which complementary coevolution was detected (Table 4-4). It is noteworthy that the CircoCoCV was one of the two animal-infecting virus datasets displaying both strong evidence of base-pairing associated negative selection, and evidence of strong selection disfavouring synonymous substitutions at paired codon sites within coding regions. Therefore, the lack of significant evidence of coevolution between nucleotides predicted to be paired within the CircoCoCV HCSS may simply be due to strong selection disfavouring any substitutions at these sites.

Table 4-4. Association between paired sites and complementarily coevolving sites

	Dataset	Chi-squared value	*p*-values
1	CircoPCV	190.9307	4.20×10^{-14}
2	CircoCoCV	0.2272	0.14
3	CircoDGCV	143.2324	3.15×10^{-14}
4	CircoBFDV	62.5998	1.59×10^{-13}
5	ParvoAAV	185.5472	4.08×10^{-14}
6	ParvoHBoV	96.656	2.13×10^{-14}
7	ParvoMPV	137.077	3.02×10^{-14}
8	AnelloTTSuV1	117.9971	2.60×10^{-14}
9	AnelloTTSuV2	38.2243	2.41×10^{-08}
10	AnelloTTV	70.6212	1.55×10^{-14}
11	NanoBBTV-R	107.8986	2.37×10^{-14}
12	NanoBBTV-S	20.398	1.28×10^{-04}
13	NanoBBTV-M	49.9491	7.88×10^{-11}
14	NanoBBTV-N	48.2752	1.79×10^{-10}
15	NanoBBTV-C	21.1911	8.81×10^{-05}
16	GeminiMSV	212.2187	4.67×10^{-14}
17	GeminiWDV	89.9702	1.98×10^{-14}
18	GeminiPanSV	82.3437	1.81×10^{-14}
19	GeminiTYDV-CpCV	28.6975	2.43×10^{-06}
20	GeminiCpCDV	98.1122	2.16×10^{-14}
21	GeminiTYLCV	159.2665	3.50×10^{-14}
22	GeminiEACMV	175.6639	3.86×10^{-14}
23	GeminiMYVYV	364.9167	8.03×10^{-14}

Besides providing additional evidence that many of the structures represented within the HCSSs really do form either within the genomes of these ssDNA viruses, or within their RNA transcripts, this result provides the most compelling evidence yet that natural selection is favouring the maintenance of a substantial proportion of these structures. The simple fact that many of the structures represented within the HCSSs likely provide significant fitness advantages to the genomes in which they occur, in turn, suggests that many of these structures have as yet, undetermined biological functions.

4.4.6 Potentially important structural elements within eukaryote-infecting ssDNA virus genomes

Whereas we provided evidence of pervasive evolutionarily conserved (and therefore, likely biologically functional) secondary structures within the various ssDNA virus genomes that we have analysed, we have not up to this point examined any of the individual computationally inferred structural elements in any significant detail. Fortunately, some of the analyses that we performed provide a straightforward means of ranking the identified structures within the HCSSs in order of their likely

biological functionality (Golden and Martin 2013). Specifically, these rankings were based on: (1) the degree to which structural elements were conserved across the analysed genomes; (2) the degree to which synonymous substitution rates were constrained at codon sites containing nucleotides that are predicted to be base-paired and; (3) the degree to which nucleotides predicted to be base-paired coevolve with one another. Rankings based collectively on these three criteria are hereafter, referred to as "consensus rankings" (Supplementary Table 2).

The ten highest ranked structural elements based on these criteria within each of the 23 analysed HCSSs are plotted in magenta and cyan in Figure 4-1 and Figure 4-2, and are listed in Supplementary 2. It is important to point out that although these top ranked structures contributed most to the signals detected in our earlier association tests, it is possible that many of them do not actually exist in the exact form that we have inferred either in the ssDNA genomes themselves, or in the RNA molecules transcribed from these genomes. Besides expected inaccuracies in the computational inference of DNA and RNA secondary structures (Ray and Pal 2013), it is likely that even if these structural elements have been accurately inferred, the exact base-pairing configurations within the presented consensus structures will likely vary between the different genomes within each of the analysed datasets. Also, it is very likely that, even within an individual genome, many of these structures will not be static but will instead represent a single reasonably stable base-pairing configuration amongst a (potentially very large) ensemble of similarly stable alternative configurations. It should therefore, be borne in mind that the actual base-pairing interactions within the tertiary structures represented by many of these structural elements, might vary as the structural elements continually transition between their alternative forms.

Among the individual structural elements that achieved the highest consensus rankings were all of the well-characterised secondary structures found at the origins of replication of circoviruses (ranks 1 to 6), nanoviruses (ranks 8 to 28), geminiviruses (ranks 1 to 12) and parvoviruses (ranks 1 to 35; Supplementary Table 2).

Additional well-characterised structures detected include the replication associated protein gene (rep) intron associated structure (GeminiMSV; rank 16; Shepherd et al.

2006), the parvovirus transcription attenuation stem-loop structures (ParvoMPV; ranks 17 and 34; Supplementary Table 2; Perros et al. 1994), the 3' complementary strand T-shaped structure that binds to the viral capsid in some parvoviruses (ParvoMPV; rank 3; Supplementary Table 2; Willwand and Hirt 1991).

Besides these well-known structures, we sought to identify other uncharacterised, but likely biologically functional, structural elements within some of these genomes. Rather than exhaustively enumerating every predicted secondary structural element that might have some biological relevance, we instead focus here on a few examples of the elements that have apparently been conserved across multiple, highly divergent viral lineages in the various viral families that we analysed.

4.4.6.1 Geminivirus

We identified a particularly conserved 126 to 157 nt secondary structure within the movement protein (mp) gene of all five analysed mastrevirus datasets (GeminiMSV, GeminiPanSV, GeminiWDV, GeminiTYDV-CpCV and GeminiCpCDV; structure G1 in Figure 4-1 and Figure 4-3). In all of these datasets other than GeminiMSV, the entire structure was within the HCSS (7th out of 25 in GeminiCpCDV, 10th out of 47 in GeminiTYDV-CpCV, 21st out of 31 in GeminiPanSV, 16th out of 30 in GeminiWDV, and 51st in Geminins). The structure in the GeminiMSV dataset displayed a particularly high degree of conformational similarity with that in the GeminiPanSV dataset with the two structures sharing a nearly identical 21-nucleotide long stem sequence (Figure 4-3) indicating that they are almost certainly homologous. Although the sequences within this structure differ substantially between the other mastrevirus datasets, they all contain the splice donor, acceptor and branch sites previously identified (or predicted) in mastrevirus *mp* introns (Figure 4-3; Wright et al. 1997), suggesting that the structure is possibly functional within the *mp* mRNA transcript where it might facilitate *mp* intron splicing. Also, likely acceptor and donor sites identified within these various sequences tend to occur at junctions between paired and unpaired nucleotides – a factor which might enhance the accessibility of these sites during splicing (Munroe 1984; Warf and Berglund 2010; Moss et al. 2012).

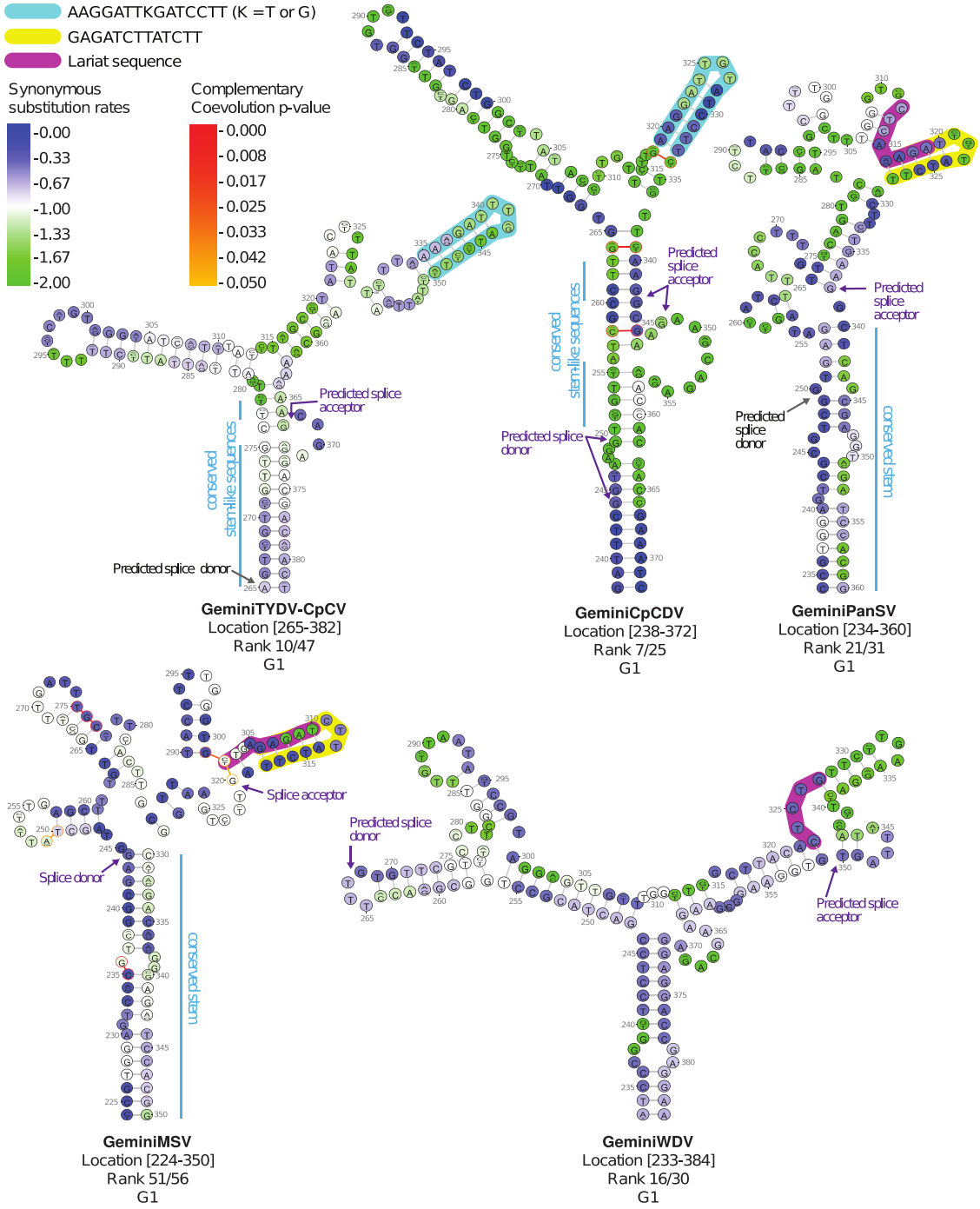


Figure 4-3. Secondary structure associated with the intron of the mastrevirus movement protein gene

A secondary structure associated with the movement protein gene intron was predicted in all five mastrevirus datasets. This structure is highly conserved and contains splice donor and acceptor sites (indicated by arrows), as well as, in the case of the GeminiMSV, GeminiPanSV and GeminiWDV, a likely lariat sequences (outlined in pink). The similarities between these structures include homologous stem-loop structures conserved in all but GeminiWDV (highlighted in blue and yellow), a highly conserved stem-structure found in both GeminiMSV and GeminiPanSV, and conserved sequences in the stems of GeminiTYDV-CpCV and GeminiCpCDV. The rank ratio shows the actual rank of a structure over the total number of structures predicted in the high confidence structure set (HCSS). This structure is highly ranked in GeminiCpCDV and GeminiTYDV-CpCV (ranked 7th out of

25 structures in HCSS and 10th out of 47 structures in the HCSS set respectively). In case of GeminiCpCDV, base-pairing interactions displaying significant associated complementary coevolution (p-value <0.05) are represented by a red line where the degree of redness reflects the p-value. Whereas nucleotide sequence variability is reflected by a sequence logo at each position, each position is also associated with a colour ranging from blue to green depicting the rate of synonymous substitutions of the codon site at which the nucleotide is located. Low synonymous substitution rates are observable in the stem region in all datasets, indicating that there is a high degree of conservation at these particular sites. Although the sequence of this structure is divergent in all five mastrevirus datasets, it is plausible that this structure has some function during splicing of the movement protein intron.

Another highly conserved secondary structure that is most likely functional within geminivirus genomes was identified near the 3' end of the coat protein (cp) genes of begomoviruses in the GeminiTYLCV, GeminiEACMV and GeminiMYVYV datasets (structure G2 in Figure 4-1 and Figure 4-4). This structure contains a conserved stem-loop sequence immediately 3' of the cp stop codon that contains the likely polyadenylation signals of both virion and complementary strand RNA transcripts (Figure 4-4). It is likely therefore, that this structure may be functional either within ssDNA as a transcriptional terminator or within transcribed mRNA during polyadenylation.

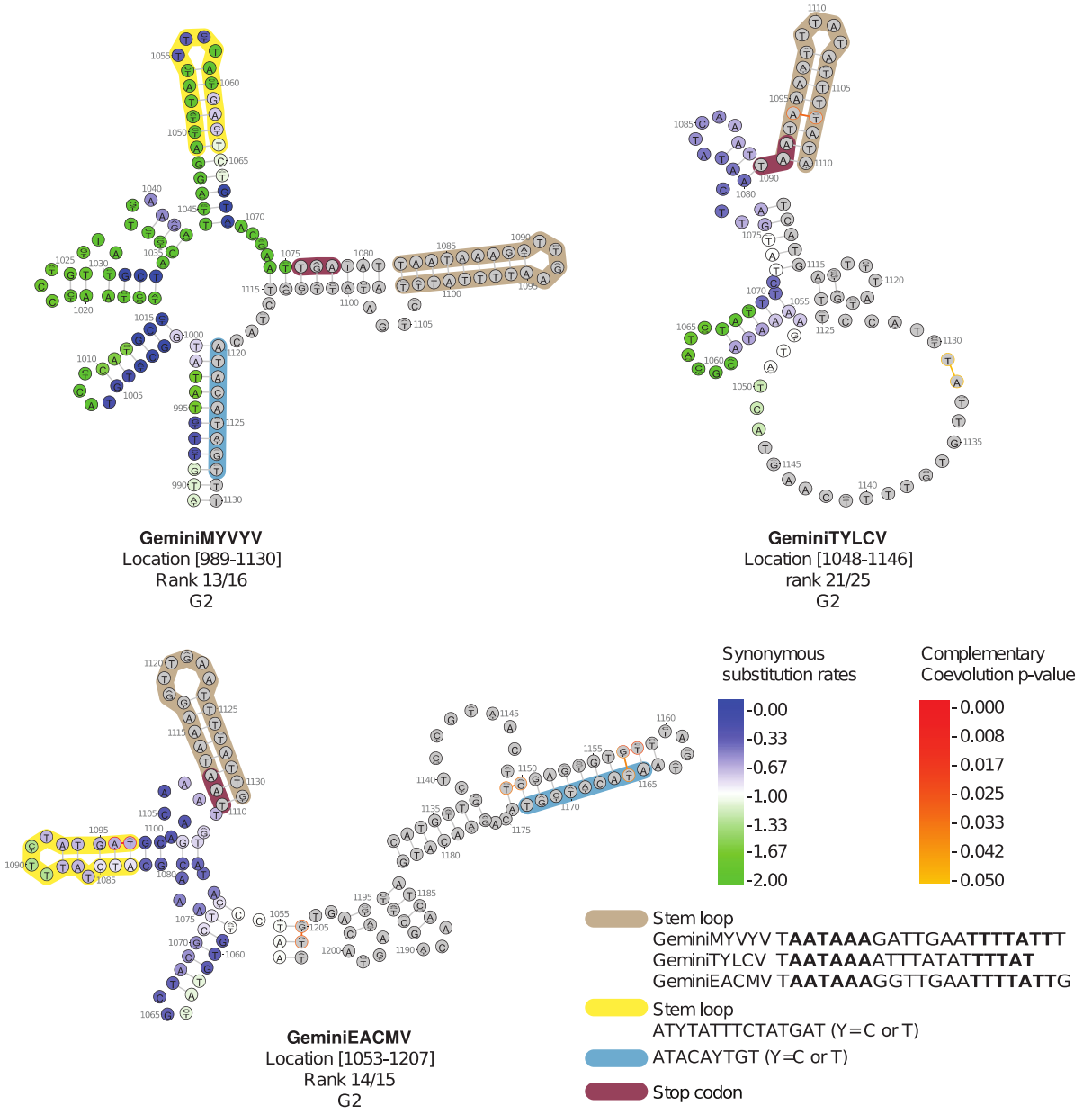


Figure 4-4. Secondary structure associated with the 3' end of the begomovirus coat protein gene

A secondary structure with a potential role in transcriptional termination was predicted at the end of the coat protein gene of the begomovirus datasets, GeminiTYLCV, GeminiEACMV and GeminiMYVYV. In all these, the structure has a stop codon and a stem-loop containing a polyadenylation signal (the complementary polyadenylation signalling sequences within the stem-loops are in bold text). A common stem-loop structure between the GeminiEACMV and GeminiMYVYV dataset is highlighted in yellow. Nucleotide logos and colours respectively indicate degrees of sequence variability and associated synonymous nucleotide substitution rates as outlined in Figure 4-3. Nucleotides falling outside genes are shaded grey. Base-pairing interactions displaying significant associated complementary coevolution (*p*-value <0.05) are represented by a red line where the degree of redness reflects the *p*-value. The rank ratio shows the actual rank of a structure over the total number of structures predicted in the high confidence structure set.

4.4.6.2 Parvovirus

We identified a variety of uncharacterised parvovirus genomic and/or mRNA structural elements with potential functionality at the start of the large non-structural (*ns1*) gene (Structure P1 in Figure. 3-2 and Figure 4-5; 20th out of 70 HCSS structures in ParvoAAV, 30th out of 105 HCSS structures in ParvoHBoV and 34th out of 132 HCSS structures in ParvoMPV), the start of the major virion/viral protein (*vp1*) gene (Structure P2 in Figure 4-2 and Figure 4-5; 16th out of 70 HCSS structures in ParvoAAV and 9th out of 132 HCSS structures in ParvoMPV), the start of the small non-structural (*np1*) gene (structure P3 in Figure 4-2 and Figure 4-5, 59th out of 105 HCSS structure in ParvoHBoV) and the start of the minor virion protein (*vp2*) gene (structure P4 in Figure 4-2 and Figure 4-5; 56th out of 132 HCSS structures in ParvoMPV). Although there were no sequence similarities shared between positionally analogous structures in the different parvovirus datasets, this was not unexpected given that these datasets represent species within different genera (with sequences in different datasets sharing on average only 57.8% sequence identity). The ParvoMPV IR-*ns1* structure contains a stem-loop identified to play role in transcription attenuation of Parvovirus minute virus of mice (structure P1; Figure 4-5; Perros et al. 1994). In this regard, it is noteworthy that start codons within the structures that we have identified are consistently located either within, or immediately adjacent to unpaired loop or bulge regions (Figure 4-5). This tendency was also noted in other datasets analysed (see Supplementary Figure 2), and it is plausible that structures spanning the start codons of genes in these different families are functional within either partially single-stranded DNA during the initiation of transcription, or in transcribed mRNA during the initiation of translation.

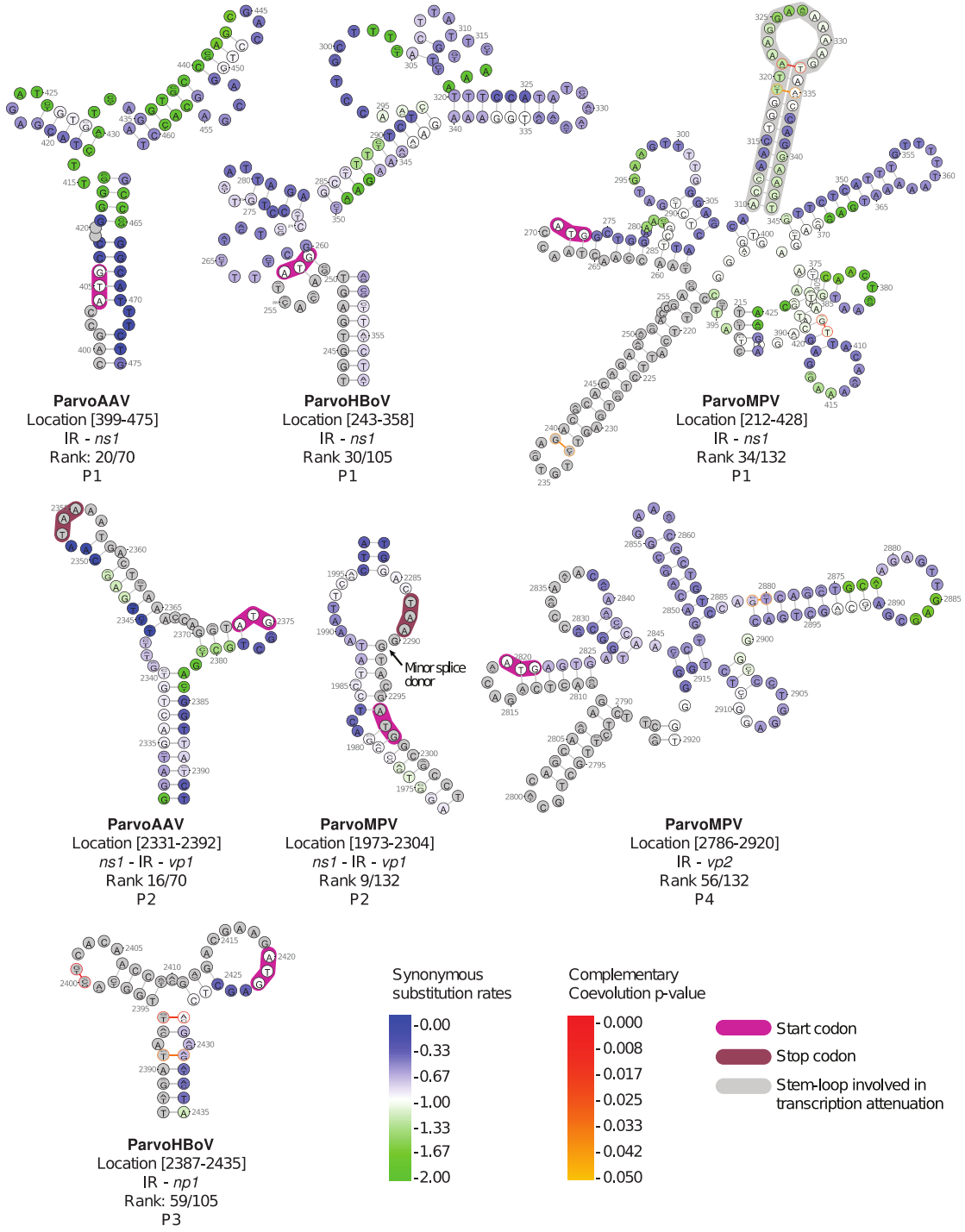


Figure 4-5. Parvovirus secondary structures predicted at the start of genes

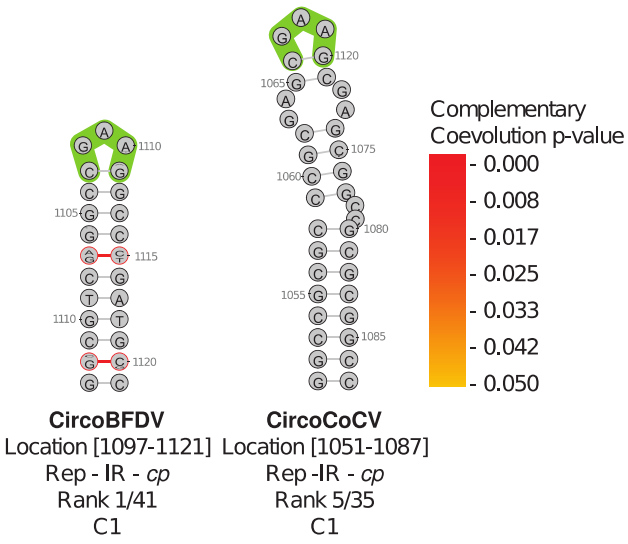
Secondary structures predicted at the start of genes represented in the parvovirus datasets ParvoAAV, ParvoHBoV and ParvoMPV are shown. These include those spanning the start of the large non-structural proteins (*ns1;* P1), the major viral/virion proteins (*vp1;* P2), the small non-structural protein (*np1;* P3) and the minor viral/virion proteins (*vp2,* P4). Nucleotide logos and colours respectively indicate degrees of sequence variability and associated synonymous nucleotide substitution rates as outlined in Figure 4-3. Base-pairing interactions displaying significant associated complementary coevolution (*p*-value <0.05) are represented by a line where the degree of redness reflects the p-value. The rank ratio shows the actual rank of a structure over the total number of structures predicted in the high confidence structure set. The ParvoMPV IR-*ns1* stem-loop involved in

transcription attenuation is highlighted in grey. In the depicted structures start codons are consistently located either within or immediately adjacent to an unpaired loop or bulge, which might enhance the accessibility of these codons during transcription or translation.

4.4.6.3 Circovirus

While we were unable to identify any secondary structures that were clearly conserved across all five circovirus datasets analysed, within the moderately divergent CircoCoCV and CircoBFDV datasets (these two datasets share on average 64% pairwise sequence identity), we identified an intergenic region (IR) stem-loop structure (structure C1 in Figure 4-2 and Figure 4-6), which is highly conserved in each of the respective datasets (ranked 5th out of 35 HCSS structures in CircoCoCV and 1st out 41 HCSS structures in CircoBFDV). Despite the sequences of this structure sharing no obvious similarity between the two datasets, in both datasets the stem is GC rich (and therefore, predicted to be very stable) with a loop sequence containing a conserved pentanucleotide (CGAAG). This structural element could potentially contain the complementary strand replication origin or it might be functional either during the termination of transcription, or in the post-transcriptional processing of mRNA transcripts.

Figure 4-6. Conserved circovirus stem-loop structure within the intergenic region



Conserved circovirus stem-loop structure within the intergenic region. A stem-loop structure that is highly conserved within the intergenic regions of the two circovirus data sets, CircoCoCV and CircoBFDV, is shown. Nucleotide logos reflect degrees of sequence variability. The stems of these structures have high GC-contents and display clear evidence of complementary coevolution between base-paired nucleotides within the CircoBFDV data set (base-pairing interactions displaying significant associated complementary coevolution with P value of <0.05 are represented by red lines, where the degree of redness reflects the P value). Additionally, a pentanucleotide loop (highlighted in green) is highly conserved at the top of the stem in both data sets.

4.4.6.4 Anellovirus

A conserved T-shaped structure was identified in the IRs of the two anellovirus datasets; AnelloTTSuV1 and AnelloTTuSV2 (structure A1 in Figure 4-2 and Figure 4-7; 7th out of 20 structures in the AnelloTTSuV1 HCSS and 1st out of 27 structures in the AnelloTTSuV2 HCSS). Even though these two datasets are moderately divergent (sequences within them share on average 60.7% pairwise identity), the structure is strikingly conserved between the two datasets. In both datasets it has a nearly identical predicted T-shaped conformation with a highly conserved 17-nucleotide long sequence at the top of the "T" (highlighted in sky blue in Figure 4-7).

Given the high degree to which this structure has been conserved between these two moderately divergent anellovirus datasets, we attempted to identify a homologous structure within our third highly divergent anellovirus dataset (AnelloTTV). The most likely homologue of this structure also resides within the IR and is ranked 7th out of 78 structures in the AnelloTTV HCSS (structure A1 in Figure 4-2 and Figure 4-7). However, the AnelloTTV structure has a stem-loop rather than a T-shaped configuration and lacks the 17-nucleotide sequence that is conserved in the AnelloTTSuV1 and AnelloTTSuV2 structures. All three Anellovirus structures nevertheless, contain two similar sequences (five and six nucleotide long) at similar positions within their stems (outlined in blue and yellow in Figure 4-7), which strongly suggests that these structures are indeed homologous.

Unlike with many other circular ssDNA viruses that replicate by rolling circle replication, it is presently unknown where the Anellovirus virion and complementary strand origins of replication reside. Given that the virion strand origins of replication of other ssDNA viruses generally occur within IRs and have a characteristic stem-loop structure with an A-T rich loop sequence, it is plausible that this highly conserved Anellovirus structural element might contain the Anellovirus virion strand ori. However, characterisation of replication competent sub-full-length Torque teno virus (TTV) genomes (which are closely related to those represented in our AnelloTTV dataset) has suggested that the TTV virion strand ori is approximately 470 nucleotides 3' of the highly conserved TTV stem-loop structure that we have identified here (in the region of a small stem-loop structure ranked 83rd, below the HCSS in our AnelloTTV dataset; data not shown; de Villiers et al. 2011).

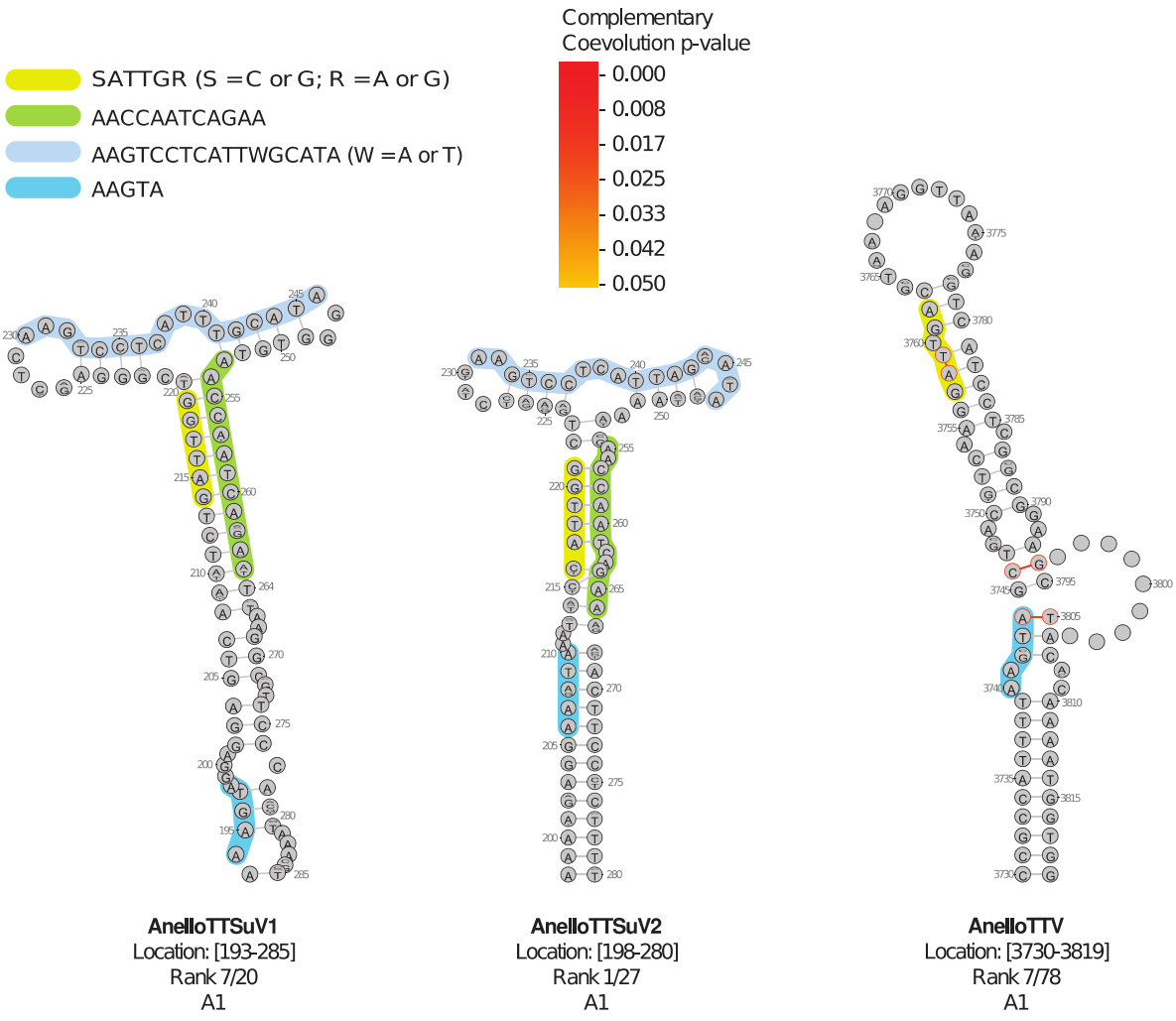


Figure 4-7. Anellovirus highly conserve intergenic T-shaped structures

A T-shaped structure predicted within the intergenic region (IR) of two divergent anellovirus datasets, AnelloTTSuV1 and AnelloTTSuV2 is shown. These structures have homologous 17-nucleotide long sequences on top of the "T" (highlighted in sky blue) and similar sequences in the stem (highlighted in green, yellow and blue). The homologue to these structures in the even more divergent AnelloTTV dataset has a stem-loop rather than a "T" configuration. It shares similar sequences (highlighted using yellow and blue) with the ones found in the other anellovirus datasets. In the AnelloTTV structure, base-pairing interactions displaying significant associated complementary coevolution (*p*-value <0.05) are represented by a red line where the degree of redness reflects the p-value. Nucleotide logos reflect the degree of sequence diversity at individual sites.

Importantly, the structure we have identified falls outside the genomic region that is conserved within these defective genomes and, in the TTV genome at least, is therefore, unlikely to be the virion strand ori. Apart from possibly containing the virion strand ori, this highly conserved structural element could alternatively be involved in either complementary strand replication or transcriptional regulation, both of which are also carried out by IR sequences in all other known ssDNA viruses.

4.5 Conclusion

Using computational methods we have identified numerous secondary structures that probably form at least transiently, within eukaryote-infecting ssDNA virus genomes, and shown that a significant proportion of these predicted structures are likely biologically functional (Table 4-5). We have further provided a few examples of currently uncharacterised genomic secondary structures which, due to high degrees of evolutionary conservation across multiple highly divergent viral lineages, likely play a central role in the biology of the various ssDNA viruses examined here.

Although we found evidence consistent with natural selection strongly disfavouring the accumulation of substitutions at paired-sites, we also found that paired-sites tended to display lower nucleotide variability than unpaired-sites. Using data from published evolution experiments, we showed that, in at least one of the analysed virus families (the geminiviruses), it is possible that this discrepancy may simply be due to mutation frequencies at paired-sites being lower than those at unpaired-sites (possibly due to base-paired nucleotides being less mutable than unpaired nucleotides). We were nevertheless able to clearly demonstrate the action of selection by showing that those base-paired-sites which do accumulate mutations display a significant tendency towards complementary coevolution with their predicted pairing partners – presumably to maintain the biological function of their parent structures.

Table 4-5. Summary of results

	Dataset	>5 NASP structures[a]	dS paired codon sites< dS unpaired codon sites [b]	Selection at paired sites[c]	Complementary coevolution[d]
1	CircoPCV	+	+	-	+
2	CircoCoCV	+	+	+	-
3	CircoDGCV	-	+	-	+
4	CircoBFDV	+	+	+	+
5	AnelloTTSuV1	+	+	-	+
6	AnelloTTSuV2	+	+	-	+
7	AnelloTTV	+	+	-	+
8	ParvoAAV	+	+	-	+
9	ParvoHBoV	+	+	-	+
10	ParvoMPV	+	+	+	+
11	NanoBBTV-R	+	+	+	+
12	NanoBBTV-S	+	+	+	+
13	NanoBBTV-M	+	+	+	+
14	NanoBBTV-N	+	-	+	+
15	NanoBBTV-C	+	-	+	+
16	GeminiMSV	+	+	+	+
17	GeminiWDV	+	+	+	+
18	GeminiPanSV	+	-	+	+
19	GeminiTYDV-CpCV	+	+	+	+
20	GeminiCpCDV	+	+	+	+
21	GeminiTYLCV	+	-	-	+
22	GeminiEACMV	+	+	+	+
23	GeminiMYVYV	+	+	+	+

[a] datasets that had more than 5 structures significantly conserved in all lineages are given a "+" sign

[b] datasets in which at least for one gene alignment the synonymous substitution rates at paired codon sites were significantly lower than those at the unpaired codon sites are given a "+" sign

[c] datasets in which purifying selection detected within paired nucleotide sites was significantly stronger than that at unpaired nucleotide sites based on F and D statistics are given a "+" sign

[d] datasets in which a statistically significant association between paired sites and complementarily coevolving sites is detected are given a "+" sign

Despite providing compelling evidence of pervasive biologically functional secondary structures within eukaryote-infecting ssDNA viruses, it is important to reiterate that our study has certain limitations. It is very likely that the complex genomic structures of these viruses are not entirely static. The secondary and tertiary structures of these entire genomes are, in fact, very likely to shift continually between large numbers of different thermodynamically stable states. We cannot therefore, be absolutely certain if the computationally predicted structures identified here are a good reflection of those which form most commonly within these ssDNA virus genomes. Also, although examples of individual genomic structural elements that are highly conserved across divergent virus lineages are likely to have some biological functionality, we cannot know without further laboratory experimentation either what the precise functions of

these structures might be, or whether they function within the context of ssDNA or transcribed RNA.

Regardless of whether specific individual structures form, or are functional within ssDNA or transcribed RNA molecules, it is absolutely clear from our study that, at the whole-genome scale, selection favouring the overall maintenance of pervasive biologically functional nucleic acid secondary structures has likely been a major theme in the evolutionary history of eukaryote-infecting ssDNA viruses.

4.6 Authors' contributions and acknowledgements

Main author's contribution

I downloaded all ssDNA virus sequences available from NCBI public database and prepared all the 23 full genome and 43 gene datasets used in this study. I ran all the tools for secondary structure prediction, inference of natural selection and complementary coevolution (NASP, PARRIS & FUBAR and the coevolution script), and wrote Python and R programming scripts used to analyse the data generated. I performed statistical tests; ranked the structural elements based on their biological relevance and produced the figures of the biologically important structures. Lastly, I wrote ~80% of the manuscript.

Co-authors' contribution

1. Darren Martin inspired and supervised the development of the computational tools applied to identify evolutionary conserved genomic structures that are potentially biologically important. He also wrote 20% of the manuscript and made critical contributions in editing both the manuscript and the figures.
2. Arvind Varsani edited the manuscript and all figures. He also provided many newly determined sequences of *Beak and feather disease virus, Columbid circovirus* and *Banana bunchy top virus* which were not publically available.
3. Michael Golden assisted in writing computer scripts used for synonymous substitution rate and complementary coevolution based tests. He developed a statistical test used to rank structures, and contributed in visualising structures and producing the figures. He also edited the manuscript.

4. Yves Semegni was involved in the prediction of evolutionary conserved secondary structures using NASP. Besides his major contribution to the development of NASP, he helped to install and run NASP. He also helped during the interpretation of the output and contributed to editing the manuscript.
5. Ben Murrell contributed to the selection analyses using FUBAR and in devising the statistical tests based on synonymous substitution rates. He additionally made critical edits on the manuscripts.
6. Art Poon was involved in developing and implementing the coevolution model, and contributed to editing the manuscript.
7. Nobubelo Ngandu helped to install HyPhy based programmes for detection of recombination (GARD) and natural selection(PARRIS), contributed to writing scripts used to ran both GARD and PARRIS and was involved in the analysis of natural selection data. She also participated in editing the manuscript.
8. Emil Tanov helped in testing and running several scripts used in all the analyses and contributed to data analysis.
9. Adérito Monjane and Pierre Lefeuvre provided data from mutational experiments of *Maize streak virus* and *Tomato yellow leaf curl virus* isolates, respectively. In addition they also helped edit the manuscript.
10. Gordon Harkins, Dionne Shepherd, Jean-Michel Lett and Alistair Gray contributed in conceiving the approaches applied in this study. Gordon, Michel and Dionne also made critical editing to the manuscript.

Acknowledgements

I thank the Centre for High Performance Computing (CHPC) in Cape Town and the Information Communication Technology Services (ICTS) Department at the University of Cape Town for providing access to their high performance computer clusters.

Chapter 5 : Impact of secondary structures on patterns of recombination within the genomes of eukaryote-infecting ssDNA viruses

5.1 Abstract

While recombination is a major evolutionary mechanism whereby many viruses can explore sequence space, it can also potentially result in the disruption of coevolved intra-genome interaction networks. Both *in vitro* evolution experiments and analyses of natural RNA and DNA virus recombinants have shown that viruses expressing chimeric proteins with less disrupted structures relative to parental genomes are more likely to replicate and survive than those that don't: i.e. natural selection apparently disfavours recombinants that express improperly folded chimeric proteins. It is plausible that selection could similarly favour recombinant viral genomes in which biologically important secondary structural elements have remained intact. I implemented computational tools to test this hypothesis in eukaryote-infecting ssDNA viruses. The tools include: (1) recombination-induced DNA folding disruption tests for natural recombinants and viral subgenomics, and (2) statistical tests for associations between the locations of nucleotide sites that are base-paired within predicted secondary structural elements and the locations of homologous and non-homologous recombination breakpoints. I observed only weak evidence that, relative to their parental genomes, both natural ssDNA virus recombinants and subgenomics in some virus groups have less disrupted secondary structures than do randomly generated recombinants/subgenomes. I also found no evidence amongst both ssDNA virus natural recombinants and subgenomics that recombination breakpoints occur more frequently at base-paired sites than at unpaired sites. Collectively, these results suggest that natural selection acting against recombinants with disrupted nucleic acid secondary structures has not obviously been a general constraint during the evolution of ssDNA viruses.

5.2 Introduction

Genetic recombination is a molecular evolutionary process by which an offspring RNA/DNA molecule referred to as “recombinant” is produced by joining fragments of two or more other RNA/DNA molecules referred to as “parentals”. Homologous recombination occurs when a recombinant is generated by the replacement of a segment of sequence in one of the parentals by a homologous segment from another parental. Non-homologous recombination occurs when the recombinant either inherits a segment of sequence from one of its parentals that has no homologue in its other parental (i.e. there is a sequence insertion) or when recombination occurs between different sites in the same molecule and the recombinant loses a segment of sequence (i.e. the recombinant has only one parental and a deletion occurs). Homologous recombination is a particularly important evolutionary mechanism as it permits both more extensive exploration of sequence space and more rapid removal of harmful mutations than could be achieved by mutation alone (Muller 1964; Felsenstein 1974).

Despite its evolutionary advantages, however, recombination between distantly related genomes that unites divergent genome fragments of independent evolutionary histories can disrupt co-adapted intra-genome interaction networks such as those occurring between different genome components (Escriu et al. 2007; Galli et al. 2010), between different proteins, between nucleic acids and proteins, between amino acids within individual proteins, and between nucleotides within structured DNA or RNA.

Analysis of chimeric proteins that are expressed by both recombinants sampled in nature (Lefeuvre et al. 2007; Simon-Loriere et al. 2009; Woo et al. 2014; Golden et al. 2014a) and those arising during evolution experiments (Martin et al. 2005b; Galli et al. 2010), have demonstrated that, relative to their parental genomes, recombinants tend to have lower degrees of intra-protein amino acid interaction disruption than do randomly generated recombinants. Recombinant genomes that preserve these interactions are expected to have a greater chance of replicating and surviving while those with disrupted interactions are expected to be eliminated by natural selection. Also, evidence is emerging showing that in viruses with extensive genomic secondary structure, selection might also disfavour the survival of

recombinants in which base-pairing interactions within functional nucleic acid structures have been disrupted (Martin et al. 2011c; Golden et al. 2014a).

It is noteworthy that in some cases maintenance of structural integrity following non-homologous recombination can impact the biological activity and/or functionality of subgenomic ssRNA and ssDNA molecules. Subgenomic molecules are generally generated via discontinuous transcription in ssRNA viruses (van Dinten et al. 1997), defective replication in ssDNA (Marriott and Dimmock 2010; Bach and Jeske 2014) or incomplete degradation of genomes by cellular enzymes in both ssDNA and ssRNA viruses (Roby et al. 2014). These molecules are apparently crucial for the biological success of certain ssRNA (Grdzelishvili et al. 2005) and ssDNA viruses (Ndunguru et al. 2006) in that their presence can modulate various cellular pathways so as to regulate viral replication and transcription rates: effects which can ultimately result in the modulation of pathogenicity (Fan et al. 2011). Importantly, experimental analysis of some ssRNA viruses have shown that the integrity of secondary structures within some subgenomics is required for their functionality (Wang et al. 1999) and it is therefore plausible that particular structural elements within these molecules might be selectively preserved within subgenomics.

As demonstrated in Chapter 4, mutations preferentially occur in the single-stranded parts of eukaryote-infecting ssDNA virus genomes rather than at sites which are base-paired within secondary structures (Muhire et al. 2014a): sites at which such mutations might be most deleterious. It is conversely plausible that recombination breakpoints, whether homologous or non-homologous, might be mechanistically predisposed to preferentially occur at base-paired sites within the genomes of ssDNA and ssRNA viruses. It is, for example, well established that secondary structures within some RNA virus genomes can have a strong influence on recombination patterns, with stable stem-loop structures specifically inducing a type of “replicational” recombination (as opposed to a strand breakage and rejoining type of recombination) called “copy-choice” recombination. These stem-loop structures apparently trigger template switching of reverse transcriptases (in viruses like HIV; Galetto et al. 2004; Galetto et al. 2006) and RNA-depend RNA polymerases (in viruses like *Brome mosaic virus*; BMV; Figlerowicz 2000). Crucially, in viruses like HIV, highly structured genome regions have far higher recombination breakpoint

densities than do unstructured genome regions in viruses sampled from nature (Simon-Loriere et al. 2010). This suggests that the impacts of secondary structure on recombination breakpoint patterns might also be readily detectable in the publically available genome sequence data of other species.

Here I use a novel homologous recombination-induced nucleic acid folding disruption test (previously described in Chapter 3) to test for evidence of natural selection acting against the survival of natural eukaryote-infecting ssDNA virus recombinants with disrupted secondary structures and also use a permutation-based test to determine whether the secondary structures within actual subgenomics are less disrupted than if subgenomic deletions occurred at random. I apply additional tests to determine whether distributions of both homologous recombination breakpoints that are detectable in natural ssDNA virus recombinants and non-homologous recombination breakpoints observed in natural eukaryote-infecting ssDNA virus subgenomics are associated with the locations of secondary structural elements.

5.3 Material and methods

5.3.1 Dataset preparation

The assembly of full genome sequence datasets for different groups of circoviruses, anelloviruses, parvoviruses, nanoviruses and geminiviruses (Table 5-1) has been described in Chapter 4. These datasets consisted of between 21 and 519 sequences each and were the focus of homologous recombination detection and recombination-induced folding disruption tests. For the detection of non-homologous recombination in ssDNA virus genomes, unpublished defective genome sequences were obtained for *Chickpea chlorotic dwarf virus* (courtesy of Arvind Varsani at the University of Canterbury), and a variety of published studies involving ssDNA viruses in the families *Anelloviridae* (de Villiers et al. 2011), *Geminiviridae* (Patil et al. 2007; Bach and Jeske 2014) and *Parvoviridae* (Hoelzer et al. 2008). Each subgenome was aligned to the parental virus full genome sequence to allow the identification of breakpoint locations. Datasets that included genomes belonging to the same genus (i.e. they were fairly closely evolutionary related) were grouped, and in total five datasets (Begomovirus, Curtovirus, Mastrevirus, Alphatorquevirus, Protoparvovirus) were assembled for these tests (Table 5-2).

5.3.2 Homologous recombination detection

Recombination Detection Program (RDP4; Martin et al. 2010) was used to identify recombination breakpoints, recombinant sequences and parental sequences. Specifically, evidence of recombination was detected using six different recombination detection methods: RDP (Martin and Rybicki 2000), GENECONV (Padidam et al. 1999b), MaxChi (Smith 1992), BootScan (Martin et al. 2005a) and SiScan (Gibbs et al. 2000). Recombination events detected by at least three methods were considered as credible. Pairs of genomes sharing more than 95% sequence identity were excluded from the preliminary recombination detection scans so as to minimise the numbers of tests performed to detect recombination (and subsequently reduce the magnitude of the Bonferroni multiple testing correction used to avoid false positive inferences of recombination). After this preliminary scan, the detected recombination events were manually checked to ensure that recombinant sequences, breakpoints locations and parental sequences were correctly identified. Five out of the 23 datasets that contained evidence of ten or more recombination events were selected for the folding disruption tests and nine of the remaining datasets (in which less than ten recombination events were detected) containing closely evolutionary related genomes were merged into four datasets in an attempt to improve the power of the tests. Thus nine datasets in total were obtained for these tests (Table 5-1).

5.3.3 Recombination-induced DNA fold disruption test

The recombination-induced DNA secondary structure disruption test, previously described in Chapter 3, was implemented in RDP4 (Martin et al. 2015) and is essentially the same to that which was recently applied to an analysis of HIV-1M genomes (Golden et al. 2014a). For every detected recombination event, one Mimic recombinant (M-recombinant) was constructed along with 100 Simulated recombinants (S-recombinants). Each M-recombinant was constructed taking minor parent segments (each bounded by a 5’ and 3’ breakpoint site) and joining these to their corresponding major parental segments at the detected 5’ and 3’ breakpoint sites. For each recombinant the major parent was considered as the genome contributing the larger segment whereas the minor parent was the genome contributing the smaller segment. Each S-recombinant was constructed in a similar

way except that the 5' breakpoint location was randomly chosen and the 3' breakpoint was fixed downstream such that the number of variable sites differentiating the parental sequences between the 5' and 3' breakpoints was exactly the same as in the corresponding M-recombinant. Maintaining the number of variable sites between breakpoints insured that under conditions of random recombination the S-recombinants and M-recombinants would be expected to have similar degrees of fold disruption: a factor crucial for the permutation test which I used to infer the probability that S-recombinants did not have significantly higher degrees of folding disruption than the M-recombinants.

For each recombination event, an M-recombinant genome was generated along with 100 S-recombinant genomes. Each of these sequences was computationally folded along with the major and minor parental sequences using the hybrid-ss-min component of UNAFold (Markham and Zuker 2008), to determine whether the predicted base-paired nucleotides within the M-recombinant sequences were collectively significantly less different to those of their parental sequences than were those of the S-recombinants. After folding I quantified two measures:

(1) The base-pairing disruption score: the number of predicted base-paired nucleotides that were present in both parental sequences, but were not present in the M/S recombinant.
(2) The aberrant base-pairing score: the number of predicted base-paired nucleotides that were present in the M/S-recombinant, but were not present in either of the parental sequences.

Using the scores that were obtained for the M- and S-recombinants, I performed a permutation test that determines whether M-recombinants tend to have lower degree of fold disruption/aberrant base-pairing than S-recombinants (i.e. whether natural ssDNA virus recombinants have lower degree of disruption/aberrant base-pairing than expected under random recombination). For more details for this test, please see Chapter 3, Section 3.3.8.

5.3.4 Test for association between genomic secondary structures and homologous recombination breakpoints

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). I tested for evidence of such influences in eukaryote-infecting ssDNA virus genomes. For each dataset base-paired sites were determined using hybrid-ss-min and recombination breakpoints were identified using RDP4. A Fisher's exact was applied to test whether recombination breakpoint positions detected within the various ssDNA virus groups were significantly more or less clustered at base-paired sites than they were at unpaired sites.

5.3.5 Test for association between genomic secondary structure and non-homologous recombination breakpoints

I assessed the influence of secondary structures on non-homologous recombination by testing whether breakpoints of natural subgenomes preferentially occur within or closer to secondary structures than would be expected under completely random recombination. From each of the five alignments comprising virus genomes and their associated subgenomes, all breakpoints were identified as the locations of the 5' and 3'prime ends of subgenome molecules relative to their parental full genome. Each viral full genome was folded using NASP to obtain a HCSS as described in Chapter 4. Breakpoints were mapped to the full genome HCSS and a Fisher's exact test was applied to determine whether observed breakpoints were clustered more or less frequently within secondary structures than was expected by chance.

5.3.6 Non-homologous recombination-induced subgenomic fold disruption test

This test is very similar to the recombination-induced DNA fold disruption test described above in section 5.3.3. Using each subgenome's parental sequence, I simulated 100 subgenomes of the same length as the real subgenome. All parental genomes and corresponding real and simulated subgenomic sequences were folded

using hybrid-ss-min, and then every obtained subgenomic structure was compared to the homologous region of its parental structure. The fold disruption score was computed as the number of base pairs not in the subgenomic structure but present in corresponding parental structure and the aberrant score computed as the number of base pairs appearing in subgenomic structure that were absent in the corresponding parental structure. Then a permutation test was applied to determine the extent to which the real subgenomics had lower degrees of fold disruption/aberrant base-pairing formation than the simulated subgenomes.

5.4 Results and discussion

5.4.1 Weak evidence of selection against recombinants with altered secondary structures is evident in some eukaryote-infecting ssDNA viruses

Only two of the nine datasets included in this test, ParvoHBoV, and GeminiMSV + GeminiPanSV, yielded some evidence that natural recombinants had lower degrees of predicted base-pairing disruption than was expected under random recombination in the absence of any selection (Table 5-1). Similarly, only one of the nine datasets, GeminiMYVYV yielded evidence that natural recombinants had lower numbers of aberrant base-pairs than would have been expected under random recombination in the absence of any selection. Collectively, these results suggest that natural selection has either acted over the short-term to purge recombinants with altered/destabilised secondary structures, or acted over the longer-term to insure that recombination-prone sites within these genomes (Lefeuvre et al. 2007; Varsani et al. 2008; Tyumentsev et al. 2014) are arranged so as to minimise the impacts of recombination on genomic secondary structure.

Although I found absolutely no evidence of selection acting against recombination-induced secondary structure disruption in six of the nine analysed datasets, it must be emphasised that the tests I used are not particularly powerful.

Specifically the power of the tests I used was potentially undermined by the following factors: (1) the actual parental sequences were not used (they never were sampled) and it is likely that if the actual parents had been used subtle structural differences

between parentals and recombinants could have been much clearer; (2) each individual recombination event was represented by one pair of 5' and one 3' breakpoints which is not entirely realistic because natural recombinants often have more than two breakpoints; (3) the secondary structure predictions I used were not entirely accurate; and (4) it is possible that non-viable viruses were included among analysed sequences, which could have violated the implicit assumption that all of the analysed ssDNA virus genomes were reasonably fit and free of recombination-induced DNA structure disruption. In addition to these four factors, there were too few available sequences in some of the analysed datasets for us to detect sufficient numbers recombination events to even apply the test. Even when >10 recombination events were detected I only expected to achieve a significant p-value with the test if selection had been extremely strong. When this same test was applied to an analysis of 434 HIV M-recombinants (a dataset >20 times larger than all but one of those examined here), it detected selection against both aberrant ($p = 0.005$) and disrupted ($p= 0.019$) base-pairings.

Table 5-1. Homologous recombination-based tests for fold disruption and breakpoint co-localisation with secondary structures

Family	Dataset name	Number of sequences	Recombination Events[a]	Fold disruption test		Breakpoint co-localisation with secondary structures		
				Bp. disrupt.[b]	Aberr. bp[c]	Major parent[d]	Minor parent[e]	Recombinant[f]
Parvoviridae	ParvoAAV	34	16/35	0.974	0.581	0.320	0.518	0.382
	ParvoHBoV	21	12/42	0.021	0.189	0.314	0.126	0.779
	ParvoMPV	26	13/35	0.504	0.430	0.679	0.679	0.304
Anelloviridae	AnelloTTSuV2	44	18/37	0.363	0.305	0.683	0.524	0.184
Circoviridae	CircoCoCV + CircoDGCV + CircoBFDV	269	24/65	0.584	0.114	0.395	0.013	0.062
Geminiviridae	GeminiMYVYV	254	61/133	0.848	0.036	0.460	0.685	0.883
	GeminiMSV + GeminiPanSV	800	16/49	0.050	0.866	0.812	0.974	0.861
	GeminiTYDV-CpCV +GeminiCpCDV	84	15/37	0.120	0.337	0.636	0.861	0.874
	GeminiTYLCV + GeminiEACMV	374	31/67	0.231	0.489	0.348	0.238	0.892

[a]The number of recombination events for which the major and minor parents, and breakpoint positions were well-identified over the number of all recombination events detected.

[b]The base-pairing disruption p-value indicates whether natural recombinant genomes have lower degrees of fold disruption than that expected under random recombination.

[c]The aberrant base-pairing p-value indicates whether natural recombinant genomes have lower degrees of aberrant base-pairing formation than that expected under random recombination.

[d]The p-value indicating whether observed breakpoints tend to co-localise with secondary structures within the major parental genomes.

[e]The p-value indicating whether observed breakpoints tend to co-localise with secondary structures within the minor parental genomes.

[f]The p-value indicating whether observed breakpoints tend to co-localise with secondary structures within the recombinant genomes.

5.4.2 No evidence that homologous recombination breakpoints preferentially occur within genomic secondary structures

Amongst the ssDNA virus datasets analysed here, I found no evidence of recombination breakpoints clustering at base-paired sites within predicted secondary structures (Table 5-1). Similarly, when considering each parental genome separately I found no evidence of recombination breakpoints clustering within secondary structures in all but one analysed dataset. The exceptional dataset, comprised of circovirus sequences, yielded a p-value = 0.013 indicating that recombination breakpoints in this family tend to co-localise with secondary structures within the minor parental genome.

5.4.3 No association between genomic secondary structures and non-homologous recombination breakpoints

In all but one of five datasets, I found no evidence that breakpoints within subgenomic molecules preferentially occur within sites that are base-paired (Table 5-2). Conversely, a subsequent permutation test that was applied to all these datasets showed a general trend for breakpoints in these molecules to occur more frequently at non-base paired sites that were well separated from structured elements, than would be expected under completely random recombination: this was particularly apparent for the Parvovirus and Curtovirus datasets (p-value 0.021 and 0.041 respectively). This tendency indicates that my failure to detect clustering of breakpoints at or near to base-paired sites was not simply due to insufficient power in the test that I used i.e. more data would likely not increase one's ability to detect co-localisation of non-homologous recombination breakpoints and base-paired sites.

While this trend is different to that observed in some RNA viruses where breakpoints within subgenomic molecules tend to fall at base-paired sites (Figlerowicz 2000), it is possible that in ssDNA viruses the modulation of subgenome generation occurs via different pathways. It is entirely plausible that this could be the case, particularly if subgenomics are primarily generated either from molecules that are in their double-stranded transcriptionally active configurations (Jeske et al. 2001) or due to strand-cleavage which is more likely to occur in single-stranded nucleic acid molecules

(Parthasarathi et al. 1995). This trend is also similar to that observed for mutations in Chapter 4 and is consistent with the hypothesis that recombinants/subgenomes that are able to accumulate in nature are those where recombination events have had a minimally disruptive impact on secondary structure.

Table 5-2. Subgenomics-based tests for fold disruption and breakpoint co-localisation with secondary structures

Family	Genus / dataset name	Species	Number of datasets	Number of subgenomes	Number of events[a]	Association p-value[b]	Fold disruption test.	
							Bp. Disrupt.[c]	Abber. bp[d]
Geminiviridae	*Begomovirus*	*Indian cassava mosaic virus, Sri Lankan cassava mosaic virus and East African cassava mosaic virus*	8	10	42	0.640	0.195	0.978
	Curtovirus	*Beet curly top virus*	1	19	35	0.738	0.386	1
	Mastrevirus	*Chickpea chlorotic dwarf virus and Maize streak virus*	2	26	316	0.994	1	1
Anelloviridae	*Alphatorquevirus*	*Torque teno virus*	6	6	40	0.031	0.018	0.120
Parvoviridae	*Protoparvovirus*	*Canine parvovirus*	1	5	5	0.720	0.988	0.996

[a]Number of sequence fragments obtained after mapping subgenomic sequences to the corresponding viral full genome sequences.

[b]P-value indicating whether subgenomic breakpoints tend to co-localise with structural elements within HCSS predicted using NASP.

[c]P-value indicating whether subgenomics have lower degrees of fold disruption than that expected under random non-homologous recombination.

[d]P-value indicating whether subgenomics have lower degrees of aberrant base-pairing than that expected under random non-homologous recombination.

5.4.4 Weak evidence of selection acting against subgenomics with altered secondary structure

After observing that non-homologous recombination breakpoints preferentially occur outside secondary structures, I assessed whether this might be due to selection acting against the disruption of base-paring interactions within functional secondary structures. The permutation test I applied only yielded evidence that secondary structural interactions might be more preserved within anellovirus subgenomics than is expected if all randomly generated subgenomics were equally viable (p-value = 0.018; Table 5-2).

Although I managed to obtain a signal for anelloviruses, this test was also not particularly powerful due to the inability to simulate subgenomics while maintaining the same genetic distance as that between the real subgenomics and their parental genomes. In the Parvovirus, Mastrevirus and Curtovirus datasets, some subgenomic molecules contained fragments of sequence that were highly divergent from (and possibly not homologous with) the parental genomes, which inflated the fold disruption and aberrant base-pairing scores, whereas all simulated subgenomic sequences were entirely identical to regions of the parental sequence from which they were simulated and therefore yielded significantly lower scores. Additionally insufficient numbers of subgenomes was also a major limitation, and it is entirely plausible that analyses of larger datasets could yield a clearer trend.

5.5 Conclusion

Using computational techniques I have investigated the impact of genomic secondary structures on patterns of homologous and non-homologous recombination in eukaryote-infecting ssDNA viruses. I found only marginal evidence that some highly recombining ssDNA viruses such as geminiviruses might display lower degrees of recombinationally induced folding disruption than is expected under random recombination in the absence of selection. While consistent with the hypothesis that natural selection acts against some ssDNA virus recombinants containing disrupted secondary structures, it remains unknown either how general this phenomenon is during ssDNA virus evolution or how much of a constraint it is on the evolution of these viruses.

It is very likely that the folding disruption tests were particularly underpowered during the analyses of the eukaryote-infecting ssDNA virus datasets owing both to the small numbers of sequences in some of the datasets and the low numbers of recombination events detected within them.

Contrary to what has been observed in some RNA viruses, I found no evidence of homologous recombination breakpoints co-localising with secondary structures within ssDNA virus genomes. Similarly I showed that in eukaryote-infecting ssDNA virus subgenomes while there is no clear evidence that non-homologous recombination breakpoints are preferentially clustered within secondary structures, these breakpoints display a tendency to fall at unpaired sites. Although subgenomic molecules that were partially divergent from the parental genomes affected the subgenomic-based fold disruption tests, the anellovirus dataset (in which subgenomes were all uniformly similar to their parental genomes) yielded significant evidence that selection has likely acted against subgenomes with disrupted secondary structures.

Despites the technical limitations of the analyses that I performed, overall I found marginal evidence that during the evolutionary history of some eukaryote-infecting ssDNA viruses, natural selection has possibly acted against recombinants with disrupted genomic secondary structures. It remains unclear, however, to what degree the disruption of base-pairing interactions within genomic secondary structures has been a general constraint on the evolution of ssDNA viruses.

5.6 Authors' contributions and acknowledgements

Main author's contribution

I wrote the Fold disruption module within RDP4 and all Python scripts that were used to test for association between genomic secondary structure and homologous/non-homologous recombination breakpoints. I also, performed all the analyses.

Co-authors' contribution

1. Darren Martin supervised development of the Fold disruption module.
2. Michael Golden participated during implementation and testing of the Fold disruption test.

Acknowledgements

I would like to thank Dr Arvind Varsani, University of Canterbury for providing *Chickpea chlorotic dwarf virus* subgenomes. Also I thank the Centre for High Performance Computing (CHPC) in Cape Town and the Information Communication Technology Services (ICTS) Department at the University of Cape Town for providing access to their high performance computer clusters.

Chapter 6 : Concluding remarks

6.1 Summary of findings

The rapid advance of DNA sequencing technologies and increasing numbers of terrestrial and aquatic metagenomics surveys focused on studying virus diversity in the environment have dramatically increased the quantity of novel viral genome sequence data that is being determined. This has produced a pressing demand for bioinformatics tools dedicated to processing, analysing and storing sequence data (Roossinck et al. 2015). My PhD research has contributed to the field of virus evolution, through the achievement of two major objectives: (1) the development and deployment of free bioinformatics tools for both identifying and testing the biological functionality of, nucleic acid secondary structural elements within viral genome, and (2) using these tools to assess the evolutionary impacts of secondary structure on the evolution of eukaryote-infecting ssDNA virus genomes.

6.1.1. Bioinformatics Tools

SDT is gaining popularity (it has gained over 90 citations since it was first published in Muhire et al. 2013), and serves as the starting point for many molecular evolution studies that are currently been carried out throughout the world. It is also being used for the taxonomic classification of viral genome sequences based on ICTV protocols (which are crucial for the organisation and structuring of the publically stored sequence data that is used by the scientific community). SDT is user-friendly and besides being specifically designed to apply ICTV species demarcation guidelines, it additionally offers an automated means of objectively creating, from large unaligned sequence files, sequence datasets with user-specified degrees of sequence diversity. It is noteworthy that SDT is useful for the analysis of both publically available and newly determined DNA and protein sequences, and provides publication quality sequence identity heatmaps that have become very popular, particularly amongst virologists.

The innovative tools for the computational prediction of biologically relevant secondary structural elements and the approaches implemented for the visualisation of these elements (illustrating biological evidence including degree of selection and complementary coevolutionary interactions), are set to make a large positive impact on the ease with which genomic secondary structures can be studied and characterised. Using selection-detection based analyses I demonstrated the likely biological functionality of many of conserved secondary structural elements detected by the NASP method (another tool developed in our research group).

Although methods for detecting associations between base-pairing within secondary structural elements and decreased substitution-rates at the individual nucleotide- and codon-levels provides evidence that computationally inferred structural elements really do exist, they are incapable of proving that selection is acting to maintain these structures due to their biological functionality: besides the fact that decreased substitution frequencies could arise through selection acting to maintain overlapping functional motifs, base-paired nucleotides might simply be more protected from mutation than non-base-paired nucleotides. Fortunately, the detection of complementary coevolution at base-paired nucleotide sites is capable of providing direct unambiguous evidence of selection acting to maintain secondary structural elements: a process that can only be explained if the associated structural elements are biologically functional. In Chapter 4 I showed that, for eukaryote-infecting ssDNA viruses at least there exists stronger evidence for complementary coevolution than for selection acting against mutations at base paired sites during the evolutionary maintenance of secondary structures. This does not imply that the coevolution detection methods which I have used are more accurate or robust than the negative selection detection methods that I have used, but it certainly indicates that complementary coevolution is a major mechanism whereby important genomic secondary structural interactions are maintained during the evolution of ssDNA viruses. Our research group is in the process of testing whether this is also true for ssRNA viruses.

In this regard, it is interesting that the ssRNA viruses, DENV2 and HCV (discussed in Chapter 3), display substantially less evidence than ssDNA viruses of complementary coevolution between nucleotides that are base-paired within the

experimentally-determined (i.e. SHAPE-derived) genome secondary structure models of these viruses. However, within these datasets I detected high degrees of genome-wide complementary coevolution when I allowed individual nucleotides to potentially pair with multiple other nucleotides (i.e. within alternative structural conformations and/or the tertiary structure rather than just those indicated by the static SHAPE models): a finding that revealed a strong correlation between complementary-coevolution hotspots and predicted structured regions, albeit not particular base-pairings, within the static SHAPE models. This surprising result might be attributable to the inability of the SHAPE structure models to capture the entire ensemble of stable structural conformations that are likely formed by large RNA molecules.

The strong correlation between HCV and DENV2 genome regions where nucleotides are complementarily coevolving with one another and the apparent SHAPE reactivity of nucleotides suggests that it may be possible to use the coevolution detection approaches that I have applied here to improve the accuracy with which secondary structures are predicted. Specifically, it will be interesting in the future to test whether the degrees to which particular nucleotides are co-evolving with nearby nucleotides could be used to constrain MFE-based RNA folding methods so as obtain more accurate secondary structure models: i.e. using co-evolutionary propensity in the same way as SHAPE reactivities are used in conjunction with computational MFE-based folding to produce structure models that are more accurate than those produced by unconstrained MFE folding methods alone.

Another strength of the tools that I have presented here lies in their utility for visualising secondary structure information in that they: (1) allow high resolution visualisation of individual secondary structures with the possibility of overlaying additional biological data i.e. synonymous substitution, complementary coevolution and SHAPE reactivities (see NAVA described in Chapter 3); and (2) allow genome-wide visualisation of secondary structure and selection data. The visualizations that are enabled by these tools facilitate, for example, the comparative identification of regulatory elements that are conserved between related viruses (see StructureMap and SelectionMap in Chapter 3).

Importantly, these tools have all been put to productive use. Apart from being used for sequence characterisation and the identification of functional secondary structures within the genomes of various eukaryote-infecting ssDNA (Candresse et al. 2014; Muhire et al. 2014a; Stenzel et al. 2014; Kraberger et al. 2014; Kraberger et al. 2015) and some ssRNA viruses (Cloete et al. 2014; Golden et al. 2014a; Mauger et al. 2015) these tools have paved the way for other on-going large-scale studies focusing on identifying and characterising the numerous biologically functional secondary structural elements that are evident within the genomes of 45 virus species within the mastrevirus genus of the family *Geminiviridae* and 56 virus species drawn from twelve different families of ssRNA viruses, including important human-infecting pathogens such influenza virus, poliovirus, coronavirus, DENV and HCV. Although, for most of these virus families some functional RNA structural elements have already been characterised, these large-scale studies will both help test hypotheses pertaining to the nature and functionality of these structural elements, and uncover and characterise additional hitherto unknown elements.

6.1.2. Pervasive secondary-structures in eukaryote-infecting ssDNA virus genomes

Using the analysis tools that I developed, I identified and characterised the genomic secondary structures evident within five families of eukaryote-infecting ssDNA viruses. Besides some previously characterised structures with known roles, including the regulation of rolling-circle and rolling-hairpin replication, I identified hundreds of other uncharacterised likely biologically functional structures that potentially have function during movement, transcription, translation, gene splicing, and replication.

However, it is not yet possible to use computational tools to determine whether many of these elements form and are functional within the genomic ssDNA of these viruses or within the mRNA transcripts that are produced from these ssDNA genomes. My analyses were also unable to determine what the precise biological functions of these structural elements are. Further experimental work is therefore required to understand the regulatory pathways/mechanisms within which these structural elements function.

It is noteworthy that my analyses were likely impacted by unavoidable sampling biases arising due to the fact that some virus species have more publicly available full genome sequence data within the sequence diversity ranges that I focused on than did others (*Maize streak virus* had 759 full genomes available, *Porcine circovirus* had 519 but *Torque teno virus* had only 22 and *Human bocavirus* only 21). There were also very few or no publically available subgenome sequences for certain virus groups which precluded these groups from analyses seeking to determine whether secondary structure was associated with the formation of subgenome molecules. However, the approach (implemented in SDT) for sequence selection adopted when I assembled the datasets so as to maintain a degree of sequence diversity between 75% and 90% between individual examined genomes, ensured that the analyses were, in general, sufficiently powered to test the various hypotheses that were the focus of my thesis.

The selection analyses that I performed indicated that purifying selection tends to be stronger at predicted base-paired sites than at unpaired sites, whereas evidence of complementary coevolution was significantly stronger between nucleotides predicted to be base-paired than between nucleotides predicted to not pair. Both these lines of evidence strongly imply that numerous secondary structural elements have likely been evolutionarily conserved within ssDNA virus genomes and that complementary coevolution is potentially a major mechanism whereby secondary structures are maintained within genomes.

Furthermore, I detected weak evidence for the most recombination prone of the ssDNA viruses (particularly those in the family *Geminiviridae*), that secondary structures tend to be slightly more conserved within natural recombinants than would be expected under completely random recombination. Additionally, I observed that both homologous recombination breakpoints and non-homologous recombination break-points that occur during the generation of subgenomic ssDNA molecules preferentially fall at un-paired sites (i.e. where they are likely to be less disruptive of secondary structural interactions).

Overall these results provide evidence that natural selection acting to maintain the integrity of important interactions within biologically functional genomic secondary

structural elements has at least weakly constrained the evolution of some eukaryote-infecting ssDNA viruses.

6.2 Major challenges

ssDNA/ssRNA molecules likely transit between different stable conformations with favourable base-pairing and/or stacking energy configurations, giving particular molecules the ability to interact with different targets and have different structure-dependent functions (Dethoff et al. 2012). The inability to very accurately model the tertiary structures of large ssRNA/ssDNA molecules such viral genomes (usually > 5kb) is still a major bottleneck for all applications in this field. Currently, all computational tools for ssDNA/ssRNA secondary structure prediction (including those used here), provide only one or a few versions of the many possible structural conformations that particular molecules can adopt, and only a very small number of methods can predict interactions that only form within tertiary structures. Thus, it is obvious that many important intramolecular interactions were likely overlooked in the analyses that I have performed.

Computational power and time has been a major limiting factor in many of the analyses that I have performed: particularly with respect to the tests that I used to test whether selection against recombinants with disrupted secondary structures has had a detectable impact on recombination patterns in eukaryote-infecting ssDNA viruses. The bottle-neck in this test was the large number *in-silico* generated recombinant sequences that required folding. As a result, I had to forgo the use of NASP (which is accurate but enormously computationally intensive) in favour of using hybrid-ss-min. For every recombination event that I detected, the folding disruption test that I devised requires the folding of 100 full genome sequences. What this means is that if anybody intends using this test in the future either for larger genomes, or for species that are more recombination-prone, then it will likely be necessary to implement a fully parallelised version of the test.

Another future challenge of mine will be to implement more user-friendly tests in SelectionMap and StructureMap to determine the influence of genome secondary structures on selection patterns at the nucleotide and codon levels.

6.3 Future prospects

Despite the current progress in predicting and studying the functions of secondary structural elements within virus genomes, improvements in tertiary structure prediction could greatly improve our knowledge of the nature and functions of these regulatory elements. However, the complexity of tertiary structure prediction limits the best available tools to analysing sequences that are smaller than 100 nt long: a size exceeded by many of the biologically and medically important structural elements within virus genomes: for example most such elements are over 200 nt long in HIV (Watts et al. 2009) and HCV (Mauger et al. 2015). Tertiary structure prediction tools are limited because they need to consider a very large number of possible 3D conformations. To overcome this problem, these tools are generally overly reliant on existing experimentally determined 3D structures (Beauchamp et al. 2011). However, too few experimentally determined 3D nucleic acid structures are presently available in public databases (such as RNA BRICKS; Chojnowski et al. 2014), to enable the prediction of 3D structure models for any but a few small RNA molecules.

Given their effectiveness in predicting the structures of large ssDNA and ssRNA genome molecules, current secondary structure prediction approaches have made a major contribution to the study of functional structures within several ssRNA and ssDNA viruses. As has been shown in this thesis, the incorporation of untapped evolutionary information that is ubiquitously present within nucleotide sequence alignments can both complement the prediction of structures and identify subsets of structural elements that are biologically functional. Future research should therefore focus on using evolutionary information to refine computational secondary structure prediction tools to the point where they yield structure models that are more accurate, and potentially more biologically meaningful, than models that can be produced using laborious experimental approaches such as SHAPE.

References

Agapito G, Guzzi PH, Cannataro M (2013) Visualization of protein interaction networks: problems and solutions. BMC Bioinformatics 14 Suppl 1:S1. doi: 10.1186/1471-2105-14-S1-S1

Alarcon P, Rushton J, Nathues H, Wieland B (2013) Economic efficiency analysis of different strategies to control post-weaning multi-systemic wasting syndrome and porcine circovirus type 2 subclinical infection in 3-weekly batch system farms. Prev Vet Med 110:103–118. doi: 10.1016/j.prevetmed.2012.12.006

Altschul SF, Gish W, Miller W, et al. (1990) Basic local alignment search tool. J Mol Biol 215:403–10. doi: 10.1016/S0022-2836(05)80360-2

Ashktorab H, Srivastava A (1989) Identification of nuclear proteins that specifically interact with adeno-associated virus type 2 inverted terminal repeat hairpin DNA. J Virol 63:3034–9.

Bach J, Jeske H (2014) Defective DNAs of beet curly top virus from long-term survivor sugar beet plants. Virus Res 183:89–94. doi: 10.1016/j.virusres.2014.01.028

Bao Y, Chetvernin V, Tatusova T (2012) PAirwise Sequence Comparison (PASC) and its application in the classification of filoviruses. Viruses 4:1318–27. doi: 10.3390/v4081318

Beauchamp K, Sripakdeevong P, Das R (2011) Why Can't We Predict RNA Structure At Atomic Resolution? doi: 10.1007/978-3-642-25740-7

Ben-Asher E, Aloni Y (1984) Transcription of minute virus of mice, an autonomous parvovirus, may be regulated by attenuation. J Virol 52:266–76.

Bernardo P, Golden M, Akram M, et al. (2013) Identification and characterisation of a highly divergent geminivirus: evolutionary and taxonomic implications. Virus Res 177:35–45. doi: 10.1016/j.virusres.2013.07.006

Bernhart SH, Hofacker IL, Will S, et al. (2008) RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinformatics 9:474. doi: 10.1186/1471-2105-9-474

Berns KI (1990) Parvovirus replication. Microbiol Rev 54:316–29.

Blanchard P, Mahé D, Cariolet R, et al. (2003) Protection of swine against post-weaning multisystemic wasting syndrome (PMWS) by porcine circovirus type 2 (PCV2) proteins. Vaccine 21:4565–75.

Blinkova O, Rosario K, Li L, et al. (2009) Frequent detection of highly diverse variants of cardiovirus, cosavirus, bocavirus, and circovirus in sewage samples collected in the United States. J Clin Microbiol 47:3507–13. doi: 10.1128/JCM.01062-09

Blinkova O, Victoria J, Li Y, et al. (2010) Novel circular DNA viruses in stool samples of wild-living chimpanzees. J Gen Virol 91:74–86. doi: 10.1099/vir.0.015446-0

Bohenzky RA, LeFebvre RB, Berns KI (1988) Sequence and symmetry requirements within the internal palindromic sequences of the adeno-associated virus terminal repeat. Virology 166:316–27.

Boulton MI (2003) Geminiviruses: major threats to world agriculture. Ann Appl Biol 142:143–143. doi: 10.1111/j.1744-7348.2003.tb00239.x

Brown KR, Otasek D, Ali M, et al. (2009) NAViGaTOR: Network Analysis, Visualization and Graphing Toronto. Bioinformatics 25:3327–9. doi: 10.1093/bioinformatics/btp595

Bujarski JJ (2013) Genetic recombination in plant-infecting messenger-sense RNA viruses: overview and research perspectives. Front Plant Sci 4:68. doi: 10.3389/fpls.2013.00068

Candresse T, Filloux D, Muhire B, et al. (2014) Appearances can be deceptive: revealing a hidden viral infection with deep sequencing in a plant quarantine context. PLoS One 9:e102945. doi: 10.1371/journal.pone.0102945

Cao S, Chen SJ (2009) A new computational approach for mechanical folding kinetics of RNA hairpins. Biophys J 96:4024–4034. doi: 10.1016/j.bpj.2009.02.044

Cardinale DJ, DeRosa K, Duffy S (2013) Base composition and translational selection are insufficient to explain codon usage bias in plant viruses. Viruses 5:162–81. doi: 10.3390/v5010162

Chen D, Barros M, Spencer E, Patton JT (2001) Features of the 3'-consensus sequence of rotavirus mRNAs critical to minus strand synthesis. Virology 282:221–9. doi: 10.1006/viro.2001.0825

Cheng N, Mao Y, Shi Y, Tao S (2012a) Coevolution in RNA molecules driven by selective constraints: evidence from 5S rRNA. PLoS One 7:e44376. doi: 10.1371/journal.pone.0044376

Cheng N, Mao Y, Shi Y, Tao S (2012b) Coevolution in RNA molecules driven by selective constraints: evidence from 5S rRNA. PLoS One 7:e44376. doi: 10.1371/journal.pone.0044376

Cheung AK (2005) Detection of rampant nucleotide reversion at the origin of DNA replication of porcine circovirus type 1. Virology 333:22–30. doi: 10.1016/j.virol.2004.12.016

Cheung AK (2004a) Palindrome regeneration by template strand-switching mechanism at the origin of DNA replication of porcine circovirus via the rolling-circle melting-pot replication model. J Virol 78:9016–29. doi: 10.1128/JVI.78.17.9016-9029.2004

Cheung AK (2006) Rolling-circle replication of an animal circovirus genome in a theta-replicating bacterial plasmid in Escherichia coli. J Virol 80:8686–94. doi: 10.1128/JVI.00655-06

Cheung AK (2004b) Detection of template strand switching during initiation and termination of DNA replication of porcine circovirus. J Virol 78:4268–77.

Chojnowski G, Waleń T, Bujnicki JM (2014) RNA Bricks - A database of RNA 3D motifs and their interactions. Nucleic Acids Res. doi: 10.1093/nar/gkt1084

Cloete LJ, Tanov EP, Muhire BM, et al. (2014) The influence of secondary structure, selection and recombination on rubella virus nucleotide substitution rate estimates. Virol J 11:166. doi: 10.1186/1743-422X-11-166

Collier AJ, Gallego J, Klinck R, et al. (2002) A conserved RNA structure within the HCV IRES eIF3-binding site. Nat Struct Biol 9:375–80. doi: 10.1038/nsb785

Conesa A, Götz S, García-Gómez JM, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–6. doi: 10.1093/bioinformatics/bti610

Cossons N, Faust EA, Zannis-Hadjopoulos M (1996) DNA polymerase delta-dependent formation of a hairpin structure at the 5’ terminal palindrome of the minute virus of mice genome. Virology 216:258–64. doi: 10.1006/viro.1996.0058

Cramer F (1971) Three-dimensional structure of tRNA. Prog Nucleic Acid Res Mol Biol 11:391–421. doi: 10.1016/S0079-6603(08)60333-5

Crameri A, Raillard S, Bermudez E, Stemmer W (1998) DNA shuffling of a family of genes from diverse species accelerates directed evolution. Nature 288–291.

Darty K, Denise A, Ponty Y (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics 25:1974–5. doi: 10.1093/bioinformatics/btp250

Das R, Baker D (2007) Automated de novo prediction of native-like RNA tertiary structures. Proc Natl Acad Sci U S A 104:14664–14669. doi: 10.1073/pnas.0703836104

Dayaram A, Galatowitsch M, Harding JS, et al. (2014) Novel circular DNA viruses identified in Procordulia grayi and Xanthocnemis zealandica larvae using metagenomic approaches. Infect Genet Evol 22:134–41. doi: 10.1016/j.meegid.2014.01.013

Dayaram A, Potter KA, Moline AB, et al. (2013) High global diversity of cycloviruses amongst dragonflies. J Gen Virol 94:1827–40. doi: 10.1099/vir.0.052654-0

De Rijk P, Wuyts J, De Wachter R (2003) RnaViz 2: an improved representation of RNA secondary structure. Bioinformatics 19:299–300. doi: 10.1093/bioinformatics/19.2.299

De Villiers E-M, Borkosky SS, Kimmel R, et al. (2011) The diversity of torque teno viruses: in vitro replication leads to the formation of additional replication-competent subviral molecules. J Virol 85:7284–95. doi: 10.1128/JVI.02472-10

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. doi: 10.1073/pnas.0806929106

Dela-Moss LI, Moss WN, Turner DH (2014) Identification of conserved RNA secondary structures at influenza B and C splice sites reveals similarities and differences between influenza A, B, and C. BMC Res Notes 7:22. doi: 10.1186/1756-0500-7-22

Dethoff EA, Chugh J, Mustoe AM, Al-Hashimi HM (2012) Functional complexity and regulation through RNA dynamics. Nature 482:322–330. doi: 10.1038/nature10885

Ding F, Sharma S, Chalasani P, et al. (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. RNA 14:1164–1173. doi: 10.1261/rna.894608

Ding Y (2003) A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res 31:7280–7301. doi: 10.1093/nar/gkg938

Domingo E, Holland JJ (1997) RNA virus mutations and fitness for survival. Annu Rev Microbiol 51:151–178. doi: 10.1146/annurev.micro.51.1.151

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. doi: 10.1099/vir.0.014712-0

Du Z, Chen M, Wang Z, et al. (2014) Isolation and molecular characterization of a distinct begomovirus and its associated betasatellite infecting Hedyotis uncinella (Hook. et Arn.) in Vietnam. Virus Genes 80:246–50. doi: 10.1007/s11262-014-1043-2

Du Z, Tang Y, Zhang S, et al. (2013) Identification and molecular characterization of a single-stranded circular DNA virus with similarities to Sclerotinia sclerotiorum hypovirulence-associated DNA virus 1. Arch Virol. doi: 10.1007/s00705-013-1890-5

Duffy S, Shackelton L a, Holmes EC (2008) Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet 9:267–276. doi: 10.1038/nrg2323

Eddy SR, Durbin R (1994) RNA sequence analysis using covariance models. Nucleic Acids Res 22:2079–88.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–7. doi: 10.1093/nar/gkh340

Edwards CTT, Holmes EC, Pybus OG, et al. (2006) Evolution of the human immunodeficiency virus envelope gene is dominated by purifying selection. Genetics 174:1441–1453. doi: 10.1534/genetics.105.052019

Elias I (2006) Settling the intractability of multiple alignment. J Comput Biol 13:1323–1339. doi: 10.1089/cmb.2006.13.1323

Escriu F, Fraile A, García-Arenal F (2007) Constraints to genetic exchange support gene coadaptation in a tripartite RNA virus. PLoS Pathog 3:0067–0074. doi: 10.1371/journal.ppat.0030008

Fan Y-H, Nadar M, Chen C-C, et al. (2011) Small noncoding RNA modulates japanese encephalitis virus replication and translation in trans. Virol J 8:492. doi: 10.1186/1743-422X-8-492

Fauquet CM (2006) The diversity of single stranded DNA viruses. Biodiversity 7:38–44. doi: 10.1080/14888386.2006.9712793

Faurez F, Dory D, Grasland B, Jestin A (2009) Replication of porcine circoviruses. Virol J 6:60. doi: 10.1186/1743-422X-6-60

Felsenstein J (2004) Inferring Phylogenies. Sinauer Associates, Inc., Sunderland, Massachusetts

Felsenstein J (1974) The evolutionary advantage of recombination. Genetics 78:737–756.

Felsenstein J (1989) PHYLIP – Phylogeny Inference Package (Version 3.2). Cladistics 5:163–166. doi: 10.1111/j.1096-0031.1989.tb00562.x

Fernandes J, Jayaraman B, Frankel A (2012) The HIV-1 Rev response element: an RNA scaffold that directs the cooperative assembly of a homo-oligomeric ribonucleoprotein complex. RNA Biol 9:6–11. doi: 10.4161/rna.9.1.18178

Fernández N, Fernandez-Miragall O, Ramajo J, et al. (2011) Structural basis for the biological relevance of the invariant apical stem in IRES-mediated translation. Nucleic Acids Res 39:8572–85. doi: 10.1093/nar/gkr560

Fiallo-Olivé E, Chirinos DT, Geraud-Pouey F, et al. (2014) Complete genome sequence of Jacquemontia yellow mosaic virus, a novel begomovirus from Venezuela related to other New World bipartite begomoviruses infecting Convolvulaceae. Arch Virol. doi: 10.1007/s00705-014-1996-4

Figlerowicz M (2000) Role of RNA structure in non-homologous recombination between genomic molecules of brome mosaic virus. Nucleic Acids Res 28:1714–1723.

Fisher RA (1922) On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. J R Stat Soc 85:87. doi: 10.2307/2340521

Frederico L, Kunkel T, Shaw B (1990) A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry 29:2532–2537.

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133:693–709.

Galetto R, Giacomoni V, Véron M, Negroni M (2006) Dissection of a circumscribed recombination hot spot in HIV-1 after a single infectious cycle. J Biol Chem 281:2711–2720. doi: 10.1074/jbc.M505457200

Galetto R, Moumen A, Giacomoni V, et al. (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. doi: 10.1074/jbc.M405476200

Galli A, Kearney M, Nikolaitchik OA, et al. (2010) Patterns of Human Immunodeficiency Virus type 1 recombination ex vivo provide evidence for coadaptation of distant sites, resulting in purifying selection for intersubtype recombinants during replication. J Virol 84:7651–7661. doi: 10.1128/JVI.00276-10

Garcia-Diaz M, Bebenek K (2007) Multiple Functions of DNA Polymerases. CRC Crit Rev Plant Sci 26:105–122. doi: 10.1080/07352680701252817

Ge X, Wu Y, Wang M, et al. (2013) Viral metagenomics analysis of planktonic viruses in East Lake, Wuhan, China. Virol Sin 28:280–90. doi: 10.1007/s12250-013-3365-y

Gharouni Kardani S, Heydarnejad J, Zakiaghl M, et al. (2013) Diversity of beet curly top Iran virus isolated from different hosts in Iran. Virus Genes 46:571–5. doi: 10.1007/s11262-013-0875-5

Gibbs MJ, Armstrong JS, Gibbs AJ (2000) Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics 16:573–82.

Giegé R (2008) Toward a more complete view of tRNA biology. Nat Struct Mol Biol 15:1007–1014. doi: 10.1038/nsmb.1498

Golden M, Martin D (2013) DOOSS: a tool for visual analysis of data overlaid on secondary structures. Bioinformatics 29:271–2. doi: 10.1093/bioinformatics/bts667

Golden M, Muhire BM, Semegni Y, Martin DP (2014a) Patterns of recombination in HIV-1M are influenced by selection disfavouring the survival of recombinants with disrupted genomic RNA and protein structures. PLoS One 9:e100400. doi: 10.1371/journal.pone.0100400

Golden M, Muhire BM, Semegni Y, Martin DP (2014b) Patterns of recombination in HIV-1M are influenced by selection disfavouring the survival of recombinants with disrupted genomic RNA and protein structures. PLoS One 9:e100400. doi: 10.1371/journal.pone.0100400

Grdzelishvili VZ, Garcia-Ruiz H, Watanabe T, Ahlquist P (2005) Mutual interference between genomic RNA replication and subgenomic mRNA transcription in brome mosaic virus. J Virol 79:1438–1451. doi: 10.1128/JVI.79.3.1438-1451.2005

Greiner W, Neise L, Stöcker H (1995) Thermodynamics and statistical mechanics. Springer-Verlag, New York

Grigoras I, Del Cueto Ginzo AI, Martin DP, et al. (2014) Genome Diversity and Evidence of Recombination and Reassortment in Nanoviruses from Europe. J Gen Virol. doi: 10.1099/vir.0.063115-0

Gronenborn B (2004) Nanoviruses: genome organisation and protein function. Vet Microbiol 98:103–109. doi: 10.1016/j.vetmic.2003.10.015

Gu W, Li M, Xu Y, et al. (2014) The impact of RNA structure on coding sequence evolution in both bacteria and eukaryotes. BMC Evol Biol 14:87. doi: 10.1186/1471-2148-14-87

Guindon S, Dufayard J-F, Lefort V, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–21. doi: 10.1093/sysbio/syq010

Guo L, Allen EM, Miller WA (2001) Base-pairing between untranslated regions facilitates translation of uncapped, nonpolyadenylated viral RNA. Mol Cell 7:1103–9.

Gutierrez C (1999) Geminivirus DNA replication. Cell Mol Life Sci 56:313–29.

Hafner GJ, Stafford MR, Wolter LC, et al. (1997) Nicking and joining activity of banana bunchy top virus replication protein in vitro. J Gen Virol 78 (Pt 7):1795–9.

Harborth J, Elbashir SM, Vandenburgh K, et al. (2003) Sequence, chemical, and structural variation of small interfering RNAs and short hairpin RNAs and the effect on mammalian gene silencing. Antisense Nucleic Acid Drug Dev 13:83–105. doi: 10.1089/108729003321629638

Harkins GW, Martin DP, Christoffels A, Varsani A (2014) Towards inferring the global movement of beak and feather disease virus. Virology 450-451:24–33. doi: 10.1016/j.virol.2013.11.033

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:169–174.

Hasegawa M, Yasunaga T, Miyata T (1979) Secondary structure of MS2 phage RNA and bias in code word usage. Nucleic Acids Res 7:2073–9.

Hefferon KL, Moon Y-S, Fan Y (2006) Multi-tasking of nonstructural gene products is required for bean yellow dwarf geminivirus transcriptional regulation. FEBS J 273:4482–94. doi: 10.1111/j.1742-4658.2006.05454.x

Herráez A (2006) Biomolecules in the computer: Jmol to the rescue. Biochem Mol Biol Educ 34:255–61. doi: 10.1002/bmb.2006.494034042644

Herschlag D, Khosla M, Tsuchihashi Z, Karpel RL (1994) An RNA chaperone activity of non-specific RNA binding proteins in hammerhead ribozyme catalysis. EMBO J 13:2913–2924.

Heyraud F, Matzeit V, Schaefer S, et al. (1993) The conserved nonanucleotide motif of the geminivirus stem-loop sequence promotes replicational release of virus molecules from redundant copies. Biochimie 75:605–615. doi: 10.1016/0300-9084(93)90067-3

Hoelzer K, Shackelton L a, Holmes EC, Parrish CR (2008) Within-host genetic diversity of endemic and emerging parvoviruses of dogs and cats. J Virol 82:11096–105. doi: 10.1128/JVI.01003-08

Hofacker IL, Fekete M, Flamm C, et al. (1998) Automatic detection of conserved RNA structure elements in complete RNA virus genomes. Nucleic Acids Res 26:3825–3836. doi: 10.1093/nar/26.16.3825

Hofacker IL, Stadler PF (2006) Memory efficient folding algorithms for circular RNA secondary structures. Bioinformatics 22:1172–6. doi: 10.1093/bioinformatics/btl023

Holmes EC (2003) Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. J Virol 77:11296–11298. doi: 10.1128/JVI.77.20.11296-11298.2003

Idris AM, Mills-Lujan K, Martin K, Brown JK (2008) Melon chlorotic leaf curl virus: characterization and differential reassortment with closest relatives reveal adaptive virulence in the squash leaf curl virus clade and host shifting by the host-restricted bean calico mosaic virus. J Virol 82:1959–1967. doi: 10.1128/JVI.01992-07

Ilyinskii PO, Schmidt T, Lukashev D, et al. (2009) Importance of mRNA secondary structural elements for the expression of influenza virus genes. OMICS 13:421–30. doi: 10.1089/omi.2009.0036

Iserentant D, Fiers W (1980) Secondary structure of mRNA and efficiency of translation initiation. Gene 9:1–12. doi: 10.1016/0378-1119(80)90163-8

Ishitani R, Nureki O, Nameki N, et al. (2003) Alternative Tertiary Structure of tRNA for Recognition by a Posttranscriptional Modification Enzyme. Cell 113:383–394. doi: 10.1016/S0092-8674(03)00280-0

Jeske H, Lütgemeier M, Preiß W (2001) DNA forms indicate rolling circle and recombination-dependent replication of Abutilon mosaic virus. EMBO J 20:6158–6167. doi: 10.1093/emboj/20.21.6158

Julian L, Piasecki T, Chrzastek K, et al. (2013) Extensive recombination detected among beak and feather disease virus isolates from breeding facilities in Poland. J Gen Virol 94:1086–95. doi: 10.1099/vir.0.050179-0

Kanakala S, Verma HN, Vijay P, et al. (2013) Response of chickpea genotypes to Agrobacterium-mediated delivery of Chickpea chlorotic dwarf virus (CpCDV) genome and identification of resistance source. Appl Microbiol Biotechnol 97:9491–501. doi: 10.1007/s00253-013-5162-9

Kanamori Y, Nakashima N (2001) A tertiary structure model of the internal ribosome entry site (IRES) for methionine-independent initiation of translation. RNA 7:266–74.

Karolchik D, Hinrichs AS, Kent WJ (2011) The UCSC Genome Browser. Curr Protoc Hum Genet Chapter 18:Unit18.6. doi: 10.1002/0471142905.hg1806s71

Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33:511–8. doi: 10.1093/nar/gki198

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–80. doi: 10.1093/molbev/mst010

Kim DY, Firth AE, Atasheva S, et al. (2011) Conservation of a packaging signal and the viral genome RNA packaging mechanism in alphavirus evolution. J Virol 85:8022–8036. doi: 10.1128/JVI.00644-11

Kim M, Oh H-S, Park S-C, Chun J (2014) Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol 64:346–351. doi: 10.1099/ijs.0.059774-0

Koev G, Liu S, Beckett R, Miller WA (2002) The 3'-Terminal Structure Required for Replication of Barley Yellow Dwarf Virus RNA Contains an Embedded 3' End. Virology 292:114–126.

Koev G, Mohan BR, Miller WA (1999) Primary and secondary structural elements required for synthesis of barley yellow dwarf virus subgenomic RNA1. J Virol 73:2876–85.

Kolupaeva VG, Pestova T V, Hellen CU (2000) Ribosomal binding to the internal ribosomal entry site of classical swine fever virus. RNA 6:1791–807.

Kosakovsky Pond SL, Posada D, Gravenor MB, et al. (2006) GARD: a genetic algorithm for recombination detection. Bioinformatics 22:3096–8. doi: 10.1093/bioinformatics/btl474

Kraberger S, Kumari SG, Hamed AA, et al. (2015) Molecular diversity of Chickpea chlorotic dwarf virus in Sudan : High rates of intra-species recombination – a driving force in the emergence of new strains. Infect Genet Evol 29:203–215. doi: 10.1016/j.meegid.2014.11.024

Kraberger S, Mumtaz H, Claverie S, et al. (2014) Identification of an Australian-like dicot-infecting mastrevirus in Pakistan. Arch Virol. doi: 10.1007/s00705-014-2299-5

Krauskopf A, Bengal E, Aloni Y (1991) The block to transcription elongation at the minute virus of mice attenuator is regulated by cellular elongation factors. Mol Cell Biol 11:3515–21. doi: 10.1128/MCB.11.7.3515.Updated

Krieg PA, Melton DA (1984) Functional messenger RNAs are produced by SP6 in vitro transcription of cloned cDNAs. Nucleic Acids Res 12:7057–7070. doi: 10.1093/nar/12.18.7057

Kumar PL, Hanna R, Alabi OJ, et al. (2011) Banana bunchy top virus in sub-Saharan Africa: Investigations on virus distribution and diversity. Virus Res 159:171–182. doi: 10.1016/j.virusres.2011.04.021

Laing C, Schlick T (2011) Computational approaches to RNA structure prediction, analysis, and design. Curr Opin Struct Biol 21:306–318. doi: 10.1016/j.sbi.2011.03.015

Larkin M a, Blackshields G, Brown NP, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–8. doi: 10.1093/bioinformatics/btm404

Lauber C, Gorbalenya AE (2012) Partitioning the genetic diversity of a virus family: approach and evaluation through a case study of picornaviruses. J Virol 86:3890–904. doi: 10.1128/JVI.07173-11

Le SY, Malim MH, Cullen BR, Maizel J V (1990) A highly conserved RNA folding region coincident with the Rev response element of primate immunodeficiency viruses. Nucleic Acids Res 18:1613–23.

Lefeuvre P, Lett J, Reynaud B, Martin DP (2007) Avoidance of protein fold disruption in natural virus recombinants. PLoS Pathog 3:e181. doi: 10.1371/journal.ppat.0030181

Lefeuvre P, Lett J-M, Varsani A, Martin DP (2009a) Widely conserved recombination patterns among single-stranded DNA viruses. J Virol 83:2697–707. doi: 10.1128/JVI.02152-08

Lefeuvre P, Lett J-M, Varsani A, Martin DP (2009b) Widely conserved recombination patterns among single-stranded DNA viruses. J Virol 83:2697–707. doi: 10.1128/JVI.02152-08

Lemey P, Salemi M, Vandamme A (2009) The Phylogenetics Handbook, A Practical Approach to Phylogenetic Analysis and Hypothesis Testing, Second Edi. Cambridge University Press, New York

Leppik L, Gunst K, Lehtinen M, et al. (2007) In vivo and in vitro intragenomic rearrangement of TT viruses. J Virol 81:9346–9356. doi: 10.1128/JVI.00781-07

Lindahl T, Nyberg B (1974) Heat-induced deamination of cytosine residues in deoxyribonucleic acid. Biochemistry 13:3405–3410. doi: 10.1021/bi00713a035

Ma C-M, Hon C-C, Lam T-Y, et al. (2007) Evidence for recombination in natural populations of porcine circovirus type 2 in Hong Kong and mainland China. J Gen Virol 88:1733–1737. doi: 10.1099/vir.0.82629-0

Magnus M, Matelska D, Lach G, et al. (2014) Computational modeling of RNA 3D structures, with the aid of experimental restraints. RNA Biol 11:1–15. doi: 10.4161/rna.28826

Manzoor MT, Ilyas M, Shafiq M, et al. (2013) A distinct strain of chickpea chlorotic dwarf virus (genus Mastrevirus, family Geminiviridae) identified in cotton plants affected by leaf curl disease. Arch Virol. doi: 10.1007/s00705-013-1911-4

Markham NR, Zuker M (2008) UNAFold: software for nucleic acid folding and hybridization. In: Keith JM (ed) Methods Mol. Biol. Humana Press, Totowa, NJ, pp 3–31

Marriott AC, Dimmock NJ (2010) Defective interfering viruses and their potential as antiviral agents. Rev Med Virol 20:51–62. doi: 10.1002/rmv.641

Martin D, Rybicki E (2000) RDP: detection of recombination amongst aligned sequences. Bioinformatics 16:562–563. doi: 10.1093/bioinformatics/16.6.562

Martin DP, Biagini P, Lefeuvre P, et al. (2011a) Recombination in eukaryotic single stranded DNA viruses. Viruses 3:1699–738. doi: 10.3390/v3091699

Martin DP, Briddon RW, Varsani A (2011b) Recombination patterns in dicot-infecting mastreviruses mirror those found in monocot-infecting mastreviruses. Arch Virol 156:1463–9. doi: 10.1007/s00705-011-0994-z

Martin DP, Lefeuvre P, Varsani A, et al. (2011c) Complex recombination patterns arising during geminivirus coinfections preserve and demarcate biologically important intra-genome interaction networks. PLoS Pathog 7:e1002203. doi: 10.1371/journal.ppat.1002203

Martin DP, Lemey P, Lott M, et al. (2010) RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26:2462–3. doi: 10.1093/bioinformatics/btq467

Martin DP, Murrell B, Golden M, et al. (2015) RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol 1:vev003–vev003. doi: 10.1093/ve/vev003

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

Martin DP, van der Walt E, Posada D, Rybicki EP (2005b) The evolutionary value of recombination is constrained by genome modularity. PLoS Genet 1:e51. doi: 10.1371/journal.pgen.0010051

Martinez HM, Maizel J V, Shapiro BA (2008) RNA2D3D: a program for generating, viewing, and comparing 3-dimensional models of RNA. J Biomol Struct Dyn 25:669–83. doi: 10.1080/07391102.2008.10531240

Matthijnssens J, Ciarlet M, Heiman E, et al. (2008) Full genome-based classification of rotaviruses reveals a common origin between human Wa-Like and porcine rotavirus strains and human DS-1-like and bovine rotavirus strains. J Virol 82:3204–19. doi: 10.1128/JVI.02257-07

Mauger DM, Golden M, Yamane D, et al. (2015) Functionally conserved architecture of hepatitis C virus RNA genomes. doi: 10.1073/pnas.1416266112

McCormack JC, Simon AE (2004) Biased hypermutagenesis associated with mutations in an untranslated hairpin of an RNA virus. J Virol 78:7813–7. doi: 10.1128/JVI.78.14.7813-7817.2004

McCullers JA, Wang GC, He S, Webster RG (1999) Reassortment and insertion-deletion are strategies for the evolution of influenza B viruses in nature. J Virol 73:7343–7348.

McKenzie CL, Shatters RG, Doostdar H, et al. (2002) Effect of geminivirus infection and Bemisia infestation on accumulation of pathogenesis-related proteins in tomato. Arch Insect Biochem Physiol 49:203–214. doi: 10.1002/arch.10020

Mears JA, Cannone JJ, Stagg SM, et al. (2002) Modeling a minimal ribosome based on comparative sequence analysis. J Mol Biol 321:215–34.

Miller WA, Wang Z, Treder K (2007) The amazing diversity of cap-independent translation elements in the 3'-untranslated regions of plant viral RNAs. Biochem Soc Trans 35:1629–33. doi: 10.1042/BST0351629

Mizokami M, Orito E, Ohba K, et al. (1997) Constrained evolution with respect to gene overlap of hepatitis B virus. J Mol Evol 44 Suppl 1:S83–S90. doi: 10.1007/PL00000061

Mohan BR, Dinesh-Kumar SP, Miller WA (1995) Genes and cis-acting sequences involved in replication of barley yellow dwarf virus-PAV RNA. Virology 212:186–95. doi: 10.1006/viro.1995.1467

Monjane AL, Pande D, Lakay F, et al. (2012) Adaptive evolution by recombination is not associated with increased mutation rates in Maize streak virus. BMC Evol Biol 12:252. doi: 10.1186/1471-2148-12-252

Morozov Sy, Chernov B, Merits A, Blinov V (1994) Computer-assisted predictions of the secondary structure in the plant virus single-stranded DNA genome. J Biomol Struct Dyn 11:837–847.

Moss WN, Dela-Moss LI, Priore SF, Turner DH (2012) The influenza A segment 7 mRNA 3' splice site pseudoknot/hairpin family. RNA Biol 9:1305–10. doi: 10.4161/rna.22343

Muhire B, Martin DP, Brown JK, et al. (2013) A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). Arch Virol 158:1411–24. doi: 10.1007/s00705-012-1601-7

Muhire BM, Golden M, Murrell B, et al. (2014a) Evidence of pervasive biologically functional secondary structures within the genomes of eukaryotic single-stranded DNA viruses. J Virol 88:1972–89. doi: 10.1128/JVI.03031-13

Muhire BM, Varsani A, Martin DP (2014b) SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. PLoS One 9:e108277. doi: 10.1371/journal.pone.0108277

Muller H (1964) The relation of recombination to mutational advance. Mutat Res 2–9.

Munroe SH (1984) Secondary structure of splice sites in adenovirus mRNA precursors. Nucleic Acids Res 12:8437–56.

Murrell B, Moola S, Mabona A, et al. (2013) FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol Biol Evol 30:1196–205. doi: 10.1093/molbev/mst030

Murrell B, Wertheim JO, Moola S, et al. (2012) Detecting individual sites subject to episodic diversifying selection. PLoS Genet 8:e1002764. doi: 10.1371/journal.pgen.1002764

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

Muse S V, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11:715–24.

Nagashima S, Sasaki J, Taniguchi K (2003) Functional analysis of the stem-loop structures at the 5′ end of the Aichi virus genome. Virology 313:56–65. doi: 10.1016/S0042-6822(03)00346-5

Navas-Castillo J, Sánchez-Campos S, Noris E, et al. (2000) Natural recombination between Tomato yellow leaf curl virus-is and Tomato leaf curl virus. J Gen Virol 81:2797–801.

Ndunguru J, Legg JP, Fofana IBF, et al. (2006) Identification of a defective molecule derived from DNA-A of the bipartite begomovirus of East African cassava mosaic virus. Plant Pathol 55:2–10. doi: 10.1111/j.1365-3059.2005.01289.x

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–53.

Ngandu NK, Scheffler K, Moore P, et al. (2008) Extensive purifying selection acting on synonymous sites in HIV-1 Group M sequences. Virol J 5:160. doi: 10.1186/1743-422X-5-160

Noller HF (1984) Structure of ribosomal RNA. Annu Rev Biochem 53:119–62. doi: 10.1146/annurev.bi.53.070184.001003

Noller HF, Kop J, Wheaton V, et al. (1981) Secondary structure model for 23S ribosomal RNA. Nucleic Acids Res 9:6167–6189. doi: 10.1093/nar/9.22.6167

Oluwafemi S, Kraberger S, Shepherd D, et al. (2014) A high degree of African streak virus diversity within Nigerian maize fields includes a new mastrevirus species from Axonopus compressus. Arch Virol. doi: 10.1007/s00705-014-2090-7

Orozco BM, Hanley-Bowdoin L (1996) A DNA structure is required for geminivirus replication origin function. J Virol 70:148–58.

Padidam M, Sawyer S, Fauquet CM (1999a) Possible emergence of new geminiviruses by frequent recombination. Virology 265:218–25. doi: 10.1006/viro.1999.0056

Padidam M, Sawyer S, Fauquet CM (1999b) Possible emergence of new geminiviruses by frequent recombination. Virology 265:218–25. doi: 10.1006/viro.1999.0056

Parthasarathi S, Varela-Echavarría A, Ron Y, et al. (1995) Genetic rearrangements occurring during a single cycle of murine leukemia virus vector replication: characterization and implications. J Virol 69:7991–8000.

Patil BL, Dutt N, Briddon RW, et al. (2007) Deletion and recombination events between the DNA-A and DNA-B components of Indian cassava-infecting geminiviruses generate defective molecules in Nicotiana benthamiana. Virus Res 124:59–67. doi: 10.1016/j.virusres.2006.10.003

Paul A V, Rieder E, Kim DW, et al. (2000) Identification of an RNA hairpin in poliovirus RNA that serves as the primary template in the in vitro uridylylation of VPg. J Virol 74:10359–70.

Paz-Carrasco LC, Castillo-Urquiza GP, Lima ATM, et al. (2014) Begomovirus diversity in tomato crops and weeds in Ecuador and the detection of a recombinant isolate of rhynchosia golden mosaic Yucatan virus infecting tomato. Arch Virol. doi: 10.1007/s00705-014-2046-y

Pedersen JS, Bejerano G, Siepel A, et al. (2006) Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol 2:e33. doi: 10.1371/journal.pcbi.0020033

Pelletier J, Sonenberg N (1988) Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. Nature 334:320–5. doi: 10.1038/334320a0

Perros M, Spegelaere P, Dupont F, et al. (1994) Cruciform structure of a DNA motif of parvovirus minute virus of mice (prototype strain) involved in the attenuation of gene expression. J Gen Virol 75 (Pt 10:2645–53.

Phelps NBD, Mor SK, Armien AG, et al. (2014) Isolation and Molecular Characterization of a Novel Picornavirus from Baitfish in the USA. PLoS One 9:e87593. doi: 10.1371/journal.pone.0087593

Piasecki T, Harkins GW, Chrząstek K, et al. (2013) Avihepadnavirus diversity in parrots is comparable to that found amongst all other avian species. Virology 438:98–105. doi: 10.1016/j.virol.2013.01.009

Pillai-Nair N, Kim K-H, Hemenway C (2003) Cis-acting Regulatory Elements in the Potato Virus X 3′ Non-translated Region Differentially Affect Minus-strand and Plus-strand RNA Accumulation. J Mol Biol 326:701–720. doi: 10.1016/S0022-2836(02)01369-4

Piñeiro D, Martinez-Salas E (2012) RNA structural elements of hepatitis C virus controlling viral RNA translation and the implications for viral pathogenesis. Viruses 4:2233–50. doi: 10.3390/v4102233

Pogranichnyy RM, Yoon KJ, Harms PA, et al. (2000) Characterization of immune response of young pigs to porcine circovirus type 2 infection. Viral Immunol 13:143–53.

Pollom E, Dang KK, Potter EL, et al. (2013) Comparison of SIV and HIV-1 genomic RNA structures reveals impact of sequence evolution on conserved and non-

conserved structural motifs. PLoS Pathog 9:e1003294. doi: 10.1371/journal.ppat.1003294

Pond SLK, Frost SDW, Muse S V (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676–9. doi: 10.1093/bioinformatics/bti079

Porollo A, Meller J (2007) Versatile annotation and publication quality visualization of protein complexes using POLYVIEW-3D. BMC Bioinformatics 8:316. doi: 10.1186/1471-2105-8-316

Powell DM, Amaral MC, Wu JY, et al. (1997) HIV Rev-dependent binding of SF2/ASF to the Rev response element: possible role in Rev-mediated inhibition of HIV RNA splicing. Proc Natl Acad Sci U S A 94:973–8.

Proctor JR, Meyer IM (2013) CoFold: An RNA secondary structure prediction method that takes co-transcriptional folding into account. Nucleic Acids Res. doi: 10.1093/nar/gkt174

Pyle AM, Green JB (1995) RNA folding. Curr Opin Struct Biol 5:303–310. doi: 10.1016/0959-440X(95)80091-3

Raman S, Bouma P, Williams GD, Brian DA (2003) Stem-loop III in the 5' untranslated region is a cis-acting element in bovine coronavirus defective interfering RNA replication. J Virol 77:6720–30. doi: 10.1128/JVI.79.19.12434-12446.2005

Ray SS, Pal SK (2013) RNA secondary structure prediction using soft computing. IEEE/ACM Trans Comput Biol Bioinform 10:2–17. doi: 10.1109/TCBB.2012.159

Razavinejad S, Heydarnejad J, Kamali M, et al. (2013) Genetic diversity and host range studies of turnip curly top virus. Virus Genes 46:345–53. doi: 10.1007/s11262-012-0858-y

Resnekov O, Aloni Y (1989) RNA polymerase II is capable of pausing and prematurely terminating transcription at a precise location in vivo and in vitro. Proc Natl Acad Sci U S A 86:12–6.

Roby JA, Pijlman GP, Wilusz J, Khromykh AA (2014) Noncoding subgenomic flavivirus RNA: Multiple functions in west Nile virus pathogenesis and modulation of host responses. Viruses 6:404–427. doi: 10.3390/v6020404

Rokyta DR, Wichman HA (2009) Genic incompatibilities in two hybrid bacteriophages. Mol Biol Evol 26:2831–9. doi: 10.1093/molbev/msp199

Roossinck MJ, Martin DP, Roumagnac P (2015) Plant virus metagenomics: Advances in virus discovery. Phytopathol PHYTO12140356RVW. doi: 10.1094/PHYTO-12-14-0356-RVW

Rosario K, Duffy S, Breitbart M (2009) Diverse circovirus-like genome architectures revealed by environmental metagenomics. J Gen Virol 90:2418–24. doi: 10.1099/vir.0.012955-0

Safaee N, Noronha AM, Rodionov D, et al. (2013) Structure of the parallel duplex of poly(A) RNA: Evaluation of a 50 year-old prediction. Angew Chemie - Int Ed 52:10370–10373. doi: 10.1002/anie.201303461

Salazar GA, Meintjes A, Mazandu GK, et al. (2014) A web-based protein interaction network visualizer. BMC Bioinformatics 15:129. doi: 10.1186/1471-2105-15-129

Scheffler K, Martin DP, Seoighe C (2006) Robust inference of positive selection from recombining coding sequences. Bioinformatics 22:2493–9. doi: 10.1093/bioinformatics/btl427

Schultes EA, Spasic A, Mohanty U, Bartel DP (2005) Compact and ordered collapse of randomly generated RNA sequences. Nat Struct Mol Biol 12:1130–6. doi: 10.1038/nsmb1014

Sealfon RS, Lin MF, Jungreis I, et al. (2015) FRESCo: Finding regions of excess synonymous constraint in diverse viruses. Genome Biol. doi: 10.1186/s13059-015-0603-7

Semegni JY, Wamalwa M, Gaujoux R, et al. (2011) NASP: a parallel program for identifying evolutionarily conserved nucleic acid secondary structures from nucleotide sequence alignments. Bioinformatics 27:2443–5. doi: 10.1093/bioinformatics/btr417

Serganov A, Patel DJ (2012) Molecular recognition and function of riboswitches. Curr Opin Struct Biol 22:279–286. doi: 10.1016/j.sbi.2012.04.005

Shackelton L a, Hoelzer K, Parrish CR, Holmes EC (2007) Comparative analysis reveals frequent recombination in the parvoviruses. J Gen Virol 88:3294–301. doi: 10.1099/vir.0.83255-0

Shackelton L a, Holmes EC (2006) Phylogenetic evidence for the rapid evolution of human B19 erythrovirus. J Virol 80:3666–9. doi: 10.1128/JVI.80.7.3666-3669.2006

Shackelton L a, Parrish CR, Truyen U, Holmes EC (2005) High rate of viral evolution associated with the emergence of carnivore parvovirus. Proc Natl Acad Sci U S A 102:379–384. doi: 10.1073/pnas.0406765102

Shannon P, Markiel A, Ozier O, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–504. doi: 10.1101/gr.1239303

Shen R, Miller WA (2004) The 3' untranslated region of tobacco necrosis virus RNA contains a barley yellow dwarf virus-like cap-independent translation element. J Virol 78:4655–64. doi: 10.1128/JVI.78.9.4655

Shepherd DN, Martin DP, Varsani A, et al. (2006) Restoration of native folding of single-stranded DNA sequences through reverse mutations: an indication of a new epigenetic mechanism. Arch Biochem Biophys 453:108–22. doi: 10.1016/j.abb.2005.12.009

Sievers F, Dineen D, Wilm A, Higgins DG (2013) Making automated multiple alignments of very large numbers of protein sequences. Bioinformatics 29:989–95. doi: 10.1093/bioinformatics/btt093

Sikorski A, Massaro M, Kraberger S, et al. (2013) Novel myco-like DNA viruses discovered in the faecal matter of various animals. Virus Res 177:209–16. doi: 10.1016/j.virusres.2013.08.008

Simmonds P, Smith DB (1999) Structural constraints on RNA virus evolution. J Virol 73:5787–94.

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 10:1337–51. doi: 10.1261/rna.7640104

Simon AE, Gehrke L (2009) RNA conformational changes in the life cycles of RNA viruses, viroids, and virus-associated RNAs. Biochim Biophys Acta 1789:571–83. doi: 10.1016/j.bbagrm.2009.05.005

Simon AE, Miller WA (2013) 3’ cap-independent translation enhancers of plant viruses. Annu Rev Microbiol 67:21–42. doi: 10.1146/annurev-micro-092412-155609

Simon-Loriere E, Galetto R, Hamoudi M, et al. (2009) Molecular mechanisms of recombination restriction in the envelope gene of the human immunodeficiency virus. PLoS Pathog. doi: 10.1371/journal.ppat.1000418

Simon-Loriere E, Holmes EC (2011) Why do RNA viruses recombine? Nat Rev Microbiol 9:617–26. doi: 10.1038/nrmicro2614

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. doi: 10.1128/JVI.01302-10

Smith JM (1992) Analyzing the mosaic structure of genes. J Mol Evol 34:126–9.

Song SI, Miller WA (2004) cis and trans Requirements for Rolling Circle Replication of a Satellite RNA. J Virol 78:3072–3082. doi: 10.1128/JVI.78.6.3072-3082.2004

Sripakdeevong P, Beauchamp K, Das R (2012) RNA 3D Structure Analysis and Prediction. RNA 3D Struct. Anal. Predict. Nucleic Acids Mol. Biol. 27. pp 43–65

Staple DW, Butcher SE (2005) Pseudoknots: RNA structures with diverse functions. PLoS Biol 3:0956–0959. doi: 10.1371/journal.pbio.0030213

Staplin WR, Miller WA (2008) In vivo analyses of viral RNA translation. Methods Mol Biol 451:99–112. doi: 10.1007/978-1-59745-102-4_7

Steinfeldt T, Finsterbusch T, Mankertz A (2001) Rep and Rep' protein of porcine circovirus type 1 bind to the origin of replication in vitro. Virology 291:152–60. doi: 10.1006/viro.2001.1203

Stemmer W (1994) Rapid evolution of a protein in vitro by DNA shuffling. Nature 389–391.

Stenzel T, Piasecki T, Chrząstek K, et al. (2014) Pigeon circoviruses display patterns of recombination, genomic secondary structure and selection similar to those of Beak and feather disease viruses. J Gen Virol 1338–1351. doi: 10.1099/vir.0.063917-0

Stockley PG, Twarock R, Bakker SE, et al. (2013) Packaging signals in single-stranded RNA viruses: nature's alternative to a purely electrostatic assembly mechanism. J Biol Phys 39:277–87. doi: 10.1007/s10867-013-9313-0

Sturn A, Quackenbush J, Trajanoski Z (2002) Genesis: cluster analysis of microarray data. Bioinformatics 18:207–208. doi: 10.1093/bioinformatics/18.1.207

Sükösd Z, Knudsen B, Vaerum M, et al. (2011) Multithreaded comparative RNA secondary structure prediction using stochastic context-free grammars. BMC Bioinformatics 12:103. doi: 10.1186/1471-2105-12-103

Sun X, Simon AE (2006) A cis-replication element functions in both orientations to enhance replication of Turnip crinkle virus. Virology 352:39–51. doi: 10.1016/j.virol.2006.03.051

Sun Y, Chen AY, Cheng F, et al. (2009) Molecular characterization of infectious clones of the minute virus of canines reveals unique features of bocaviruses. J Virol 83:3956–67. doi: 10.1128/JVI.02569-08

Swofford DL (2002) PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Inc., Sunderland, Massachusetts

Tabernero C, Zolotukhin AS, Valentin A, et al. (1996) The posttranscriptional control element of the simian retrovirus type 1 forms an extensive RNA secondary structure necessary for its function. J Virol 70:5998–6011.

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

Tamura K, Peterson D, Peterson N, et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–9. doi: 10.1093/molbev/msr121

Thompson JD, Plewniak F, Poch O (1999) A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res 27:2682–90.

Thurner C (2004) Conserved RNA secondary structures in Flaviviridae genomes. J Gen Virol 85:1113–1124. doi: 10.1099/vir.0.19462-0

Timchenko T, de Kouchkovsky F, Katul L, et al. (1999) A single rep protein initiates replication of multiple genome components of faba bean necrotic yellows virus, a single-stranded DNA virus of plants. J Virol 73:10173–10182.

Treiber DK, Williamson JR (2001) Beyond kinetic traps in RNA folding. Curr Opin Struct Biol 11:309–314. doi: 10.1016/S0959-440X(00)00206-2

Tuplin A, Wood J, Evans DJ, et al. (2002) Thermodynamic and phylogenetic prediction of RNA secondary structures in the coding region of hepatitis C virus. RNA 8:824–41. doi: 10.1017.S1355838202554066

Tyumentsev AI, Tikunova N V, Tikunov AY, Babkin I V (2014) Recombination in the evolution of human bocavirus. Infect Genet Evol 28:11–4. doi: 10.1016/j.meegid.2014.08.026

Van Dinten LC, den Boon JA, Wassenaar AL, et al. (1997) An infectious arterivirus cDNA clone: identification of a replicase point mutation that abolishes discontinuous mRNA transcription. Proc Natl Acad Sci U S A 94:991–996. doi: 10.1073/pnas.94.3.991

Varsani A, Kraberger S, Jennings S, et al. (2014a) A novel papillomavirus in Adelie penguin (Pygoscelis adeliae) faeces sampled at the Cape Crozier colony, Antarctica. J Gen Virol. doi: 10.1099/vir.0.064436-0

Varsani A, Martin DP, Navas-Castillo J, et al. (2014b) Revisiting the classification of curtoviruses based on genome-wide pairwise identity. Arch Virol. doi: 10.1007/s00705-014-1982-x

Varsani A, Navas-Castillo J, Moriones E, et al. (2014c) Establishment of three new genera in the family Geminiviridae: Becurtovirus, Eragrovirus and Turncurtovirus. Arch Virol. doi: 10.1007/s00705-014-2050-2

Varsani A, Shepherd DN, Monjane AL, et al. (2008) Recombination, decreased host specificity and increased mobility may have driven the emergence of maize streak virus as an agricultural pathogen. J Gen Virol 89:2063–2074. doi: 10.1099/vir.0.2008/003590-0

Vermeulen A, Robertson B, Dalby AB, et al. (2007) Double-stranded regions are essential design components of potent inhibitors of RISC function. RNA 13:723–30. doi: 10.1261/rna.448107

Wadkins TS, Perrotta AT, Ferré-D'Amaré AR, et al. (1999) A nested double pseudoknot is required for self-cleavage activity of both the genomic and

antigenomic hepatitis delta virus ribozymes. RNA 5:720–727. doi: 10.1017/S1355838299990209

Wang J, Carpenter CD, Simon AE (1999) Minimal sequence and structural requirements of a subgenomic RNA promoter for turnip crinkle virus. Virology 253:327–336. doi: 10.1006/viro.1998.9538

Warf MB, Berglund JA (2010) Role of RNA structure in regulating pre-mRNA splicing. Trends Biochem Sci 35:169–78. doi: 10.1016/j.tibs.2009.10.004

Watts JM, Dang KK, Gorelick RJ, et al. (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. Nature 460:711–6. doi: 10.1038/nature08237

Wende S, Platzer EG, Jühling F, et al. (2014) Biological evidence for the world's smallest tRNAs. Biochimie 100:151–8. doi: 10.1016/j.biochi.2013.07.034

Wertheim JO, Kosakovsky Pond SL (2011) Purifying selection can obscure the ancient age of viral lineages. Mol Biol Evol 28:3355–3365. doi: 10.1093/molbev/msr170

Wikström FH, Fossum C, Fuxler L, et al. (2011) Cytokine induction by immunostimulatory DNA in porcine PBMC is impaired by a hairpin forming sequence motif from the genome of Porcine Circovirus type 2 (PCV2). Vet Immunol Immunopathol 139:156–66. doi: 10.1016/j.vetimm.2010.09.010

Wikström FH, Meehan BM, Berg M, et al. (2007) Structure-dependent modulation of alpha interferon production by porcine circovirus 2 oligodeoxyribonucleotide and CpG DNAs in porcine peripheral blood mononuclear cells. J Virol 81:4919–27. doi: 10.1128/JVI.02797-06

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 1:1610–6. doi: 10.1038/nprot.2006.249

Willwand K, Hirt B (1991) The minute virus of mice capsid specifically recognizes the 3' hairpin structure of the viral replicative-form DNA: mapping of the binding site by hydroxyl radical footprinting. J Virol 65:4629–35.

Wilm A, Mainz I, Steger G (2006) An enhanced RNA alignment benchmark for sequence alignment programs. Algorithms Mol Biol 1:19. doi: 10.1186/1748-7188-1-19

Woo J, Robertson DL, Lovell SC (2014) Constraints from protein structure and intra-molecular coevolution influence the fitness of HIV-1 recombinants. Virology 454-455:34–39. doi: 10.1016/j.virol.2014.01.029

Worobey M, Holmes EC (1999) Evolutionary aspects of recombination in RNA viruses. J Gen Virol 80:2535–2543.

Wright E, Heckel T, Groenendijk J, et al. (1997) Splicing features in maize streak virus virion- and complementary-sence gene expression. Plant J 12:1285–1297.

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. doi: 10.1186/1471-2156-6-20

Xin Y, Laing C, Leontis NB, Schlick T (2008) Annotation of tertiary interactions in RNA structures reveals variations and correlations. RNA 14:2465–2477. doi: 10.1261/rna.1249208

You S, Stump DD, Branch AD, Rice CM (2004) A cis-Acting Replication Element in the Sequence Encoding the NS5B RNA-Dependent RNA Polymerase Is Required for Hepatitis C Virus RNA Replication. J Virol 78:1352–1366. doi: 10.1128/JVI.78.3.1352-1366.2004

Yuen L, Moss B (1987) Oligonucleotide sequence signaling transcriptional termination of vaccinia virus early genes. Proc Natl Acad Sci U S A 84:6417–21.

Zanini F, Neher R a (2013) Quantifying selection against synonymous mutations in HIV-1 env evolution. J Virol 87:11843–50. doi: 10.1128/JVI.01529-13

Zhao Y, Huang Y, Gong Z, et al. (2012) Automated and fast building of three-dimensional RNA structures. Sci Rep. doi: 10.1038/srep00734

Zuo X, Wang J, Yu P, et al. (2009) Solution structure of the cap-independent translational enhancer and ribosome-binding element in the 3 0 UTR of turnip crinkle virus. Biophys Comput Biol 107:1385–1390. doi: 10.1073/pnas.0908140107

Supplementary information

Supplementary Table 1: 43 Gene alignments obtained from the 23 virus groups

	Family	Intermediate dataset name	Number of sequences	Gene full name	Gene short name
1	*Circoviridae*	CircoPCV	30	Replication associated protein	*Rep*
2		CircoPCV	29	Capsid protein	*Cp*
3		CircoCoCV	30	Replication associated protein	*Rep*
4		CircoCoCV	30	Capsid protein	*Cp*
5		CircoDGCV	30	Replication associated protein	*Rep*
6		CircoDGCV	30	Capsid protein	*Cp*
7		CircoBFDV	30	Replication associated protein	*Rep*
8		CircoBFDV	29	Capsid protein	*Cp*
9	*Anelloviridae*	AnelloTTSV1	17	Open reading frame 1	*ORF1*
10		AnelloTTSV2	30	Open reading frame 1	*ORF1*
11		AnelloTTV	21	Open reading frame 1	*ORF1*
12	*Parvoviridae*	ParvoAAV	23	Large non-structural protein	*NS1*
13		ParvoAAV	30	Major structural protein	*VP1*
14		ParvoHBoV	21	Large non-structural protein	*NS1*
15		ParvoHBoV	21	Small non-structural protein	*NP1*
16		ParvoMPV	25	Large non-structural protein	*NS1*
17		ParvoMPV	18	Minor structural protein	*VP2*
18	*Nanoviridae*	NanoBBTV-R	28	Replication associated protein	R
19		NanoBBTV-S	29	Capsid protein	S
20		NanoBBTV-M	27	Movement protein	M
21		NanoBBTV-N	27	Nuclear shuttle protein	N
22		NanoBBTV-C	30	Cell cycle link protein	*Clink*
23	*Geminiviridae*	GeminiMSV	30	Replication associated protein	*Rep*
24		GeminiMSV	30	Capsid protein	*Cp*
25		GeminiMSV	30	Movement protein	MP
26		GeminiWDV	30	Replication associated protein	*rep*
27		GeminiWDV	30	Capsid protein	*Cp*
28		GeminiWDV	30	Movement protein	MP
29		GeminiPanSV	30	Replication associated protein	*Rep*
30		GeminiPanSV	30	Capsid protein	*Cp*
31		GeminiPanSV	29	Movement protein	*Mp*
32		GeminiTYDV-CpCV	30	Replication associated protein	*Rep*
33		GeminiTYDV-CpCV	30	Capsid protein	*Mp*
34		GeminiTYDV-CpCV	30	Movement protein	*Rep*
35		GeminiCpCDV	30	Replication associated protein	*Mp*
36		GeminiCpCDV	30	Capsid protein	*Rep*
37		GeminiCpCDV	30	Movement protein	*Mp*
38		GeminiTYLCV	27	Replication associated protein	*Rep*
39		GeminiTYLCV	30	Capsid protein	*Cp*
40		GeminiEACMV	30	Replication associated protein	*Rep*
41		GeminiEACMV	30	Capsid protein	*Cp*
42		GeminiMYVYV	29	Replication associated protein	*Rep*
43		GeminiMYVYV	28	Capsid protein	*Cp*

Supplementary Table 2: Consensus rankings of high confidence secondary structure sets (HCSS)

Consensus ranking of structures in HCSS is based on base-pairing conservation scores, associated synonymous substitution rates and degrees of complementary coevolution (ranks are given in the column labelled "Cons. rank"). Each structure inherited the lower of the ranks of its daughter or parent structures. Structures highlighted in yellow are those that are currently well-characterised in eukaryote-infecting ssDNA virus genomes. Structures that are individually discussed in the text are highlighted in green.

CircoPCV (genome length =1796)										
Cons. Rank[a]	**Structure ID / NASP Rank[b]**	**Coordinates Including Gaps**				**Syn. Sub. Rank[c]**	**Coev. Rank[d]**	**Min. Rank[e]**	**Biological function**	**Reference**
		Left hand side		**Right hand side**						
1	1	443	459	469	485	24	64	1		
2	16	488	494	520	526	1	55	1		
3	2	3	13	1776	1786	NA	NA	2	Initiation of genome replication	Steinfeldt, T., et al. (2001)
4	6	101	107	211	217	40	2	2		
5	3	1402	1414	1418	1430	91	44	3		
6	7	48	56	273	281	83	3	3		
7	4	14	21	287	294	67	19	4		
8	9	75	84	227	236	47	4	4		
9	5	681	698	1373	1390	25	34	5		
10	11	95	100	219	224	41	5	5		
11	14	115	124	191	200	20	6	6		
12	8	803	809	816	822	7	NA	7		
13	15	137	146	174	183	15	7	7		
14	10	1605	1618	1622	1635	69	NA	10		
15	12	967	974	1062	1069	38	NA	12		
16	13	1335	1343	1351	1359	72	56	13		
17	17	148	152	161	165	31	23	17		
18	18	1119	1127	1161	1169	86	NA	18		
CircoCoCV (genome length =2058)										
1	1	10	22	2044	2056	NA	NA	1	Initiation of genome replication	Mankertz, A., et al. (2000)
2	2	488	498	910	920	55	NA	2		
3	3	506	518	553	565	78	NA	3		
4	10	97	103	107	113	3	NA	3		
5	4	1051	1058	1080	1087	NA	NA	4		Figure 4-2., C1 and Figure 4-6., CircoCoCV
6	5	881	887	896	902	74	NA	5		
7	6	591	598	605	612	32	NA	6		
8	7	875	880	904	909	68	NA	7		
9	8	532	538	546	552	47	NA	8		
10	30	1930	1938	1942	1950	8	NA	8		
11	9	395	403	408	416	67	NA	9		
12	26	1177	1182	1192	1197	10	23	10		
13	11	698	703	715	720	98	NA	11		
14	12	1575	1582	1598	1605	11	NA	11		
15	13	451	458	964	971	77	NA	13		
16	14	568	575	580	587	35	NA	14		

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

Cons. Rank[a]	Structure ID / NASP Rank[b]	Coordinates Including Gaps				Syn. Sub. Rank[c]	Coev. Rank[d]	Min. Rank[e]	Biological function	Reference
		Left hand side		Right hand side						
17	15	24	31	2036	2043	NA	NA	15		
18	16	618	625	805	812	106	NA	16		
19	31	1004	1009	1284	1289	16	24	16		
20	17	727	733	737	743	87	NA	17		
21	18	1264	1271	1276	1283	24	NA	18		
22	19	1031	1036	1101	1106	NA	NA	19		
23	20	1039	1043	1096	1100	NA	NA	20		
24	21	688	694	757	763	123	NA	21		
25	22	831	836	840	845	49	NA	22		
26	23	1738	1744	1748	1754	82	NA	23		
27	24	635	640	793	798	114	NA	24		
28	25	1454	1459	1473	1478	97	NA	25		
29	27	139	147	246	254	134	NA	27		
30	28	1059	1062	1074	1077	NA	NA	28		
31	29	1119	1122	1127	1130	NA	NA	29		
32	32	310	317	337	344	85	NA	32		
33	33	1732	1736	1756	1760	109	NA	33		
34	34	476	481	930	935	57	NA	34		
35	35	227	231	235	239	65	NA	35		
CircoDGCV (genome length =2046)										
1	1	3	13	2025	2035	NA	NA	1	Initiation of genome replication	Todd, D., et al. (2003) and Hattermann, K. et al. (2003)
2	2	102	107	131	136	34	NA	2		
3	3	234	239	244	249	14	NA	3		
4	4	289	302	1933	1946	5	13	4		
5	5	108	111	126	129	51	NA	5		
CircoBFDV (genome length =2049)										
1	1	1097	1107	1111	1121	NA	6	1		Figure 4-2., C1 and Figure 4-6., CircoBFDV
2	35	1127	1130	1134	1137	NA	1	1		
3	2	383	398	418	433	51	21	2		
4	7	192	201	309	318	2	37	2		
5	27	1122	1125	1138	1141	NA	2	2		
6	3	7	16	2095	2104	NA	18	3	Initiation of genome replication	Bassami, M.R., et al. (1998)
7	10	490	499	547	556	3	48	3		
8	11	123	128	137	142	NA	3	3		
9	4	863	871	877	885	104	NA	4		
10	12	261	271	287	297	4	NA	4		
11	21	1146	1151	1171	1176	NA	4	4		
12	5	1473	1484	1494	1505	24	NA	5		
13	6	1618	1632	1636	1650	103	13	6		
14	29	177	183	322	328	6	39	6		
15	8	31	37	41	47	NA	10	8		
16	9	607	619	640	652	113	NA	9		
17	28	509	514	533	538	11	NA	11		
18	13	784	794	799	809	97	NA	13		
19	31	501	505	541	545	13	49	13		
20	14	17	23	2085	2091	54	19	14		
21	39	1529	1539	1660	1670	88	14	14		

[a]Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b]NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID.
[c] Syn. Sub ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d]Coev. ranking is a ranking based on complementarily coevolution. NA indicates structures which had invariable sites.
[e]Minimum rank from NASP, Syn. Sub. and Coev. ranks.

Cons. Rank[a]	Structure ID / NASP Rank[b]	Coordinates Including Gaps				Syn. Sub. Rank[c]	Coev. Rank[d]	Min. Rank[e]	Biological function	Reference
		Left hand side		Right hand side						
22	15	215	228	247	260	62	38	15		
23	16	1214	1220	1233	1239	NA	NA	16		
24	17	1301	1307	1321	1327	NA	NA	17		
25	18	856	861	886	891	80	NA	18		
26	19	821	829	921	929	48	NA	19		
27	20	1155	1159	1164	1168	NA	NA	20		
28	22	1741	1749	1753	1761	107	NA	22		
29	25	1698	1703	1719	1724	22	NA	22		
30	32	358	367	447	456	60	22	22		
31	23	655	662	667	674	116	NA	23		
32	24	1262	1266	1270	1274	NA	NA	24		
33	26	1836	1841	1859	1864	52	44	26		
34	30	1705	1709	1714	1718	35	NA	30		
35	36	1964	1977	2015	2028	30	50	30		
36	33	1981	1987	2008	2014	47	NA	33		
37	34	1296	1299	1329	1332	NA	NA	34		
38	37	971	978	1020	1027	91	NA	37		
39	38	1731	1739	1763	1771	114	NA	38		
40	40	1222	1225	1229	1232	NA	NA	40		
41	41	1200	1203	1242	1245	NA	NA	41		
AnelloTTSuV1 (genome length = 2984)										
1	1	2926	2935	2939	2948	NA	NA	1		
2	2	718	725	731	738	9	NA	2		
3	3	2824	2832	2836	2844	NA	NA	3		
4	4	2961	2967	2971	2977	NA	NA	4		
5	5	2638	2646	2724	2732	NA	NA	5		
6	6	2858	2865	2870	2877	NA	NA	6		
7	157 / 7	193	220	254	285	NA	NA	7		Figure 4-2., A1 and Figure 4-7., AnelloTTSuV1
8	8	2890	2897	2902	2909	NA	NA	8		
9	12	712	717	741	746	8	NA	8		
10	9	1079	1087	1093	1101	43	NA	9		
11	10	2649	2655	2689	2695	NA	NA	10		
12	11	336	341	346	351	NA	NA	11		
13	13	2674	2678	2683	2687	NA	NA	13		
14	14	2175	2181	2221	2227	NA	NA	14		
15	15	1550	1565	1597	1612	54	15	15		
16	16	2616	2624	2847	2855	NA	NA	16		
17	17	2400	2407	2420	2427	NA	NA	17		
18	18	394	399	424	429	NA	NA	18		
19	19	2705	2709	2715	2719	NA	NA	19		
20	20	2606	2612	2883	2889	NA	NA	20		
AnelloTTSuV2 (genome length = 2931)										
1	1	198	210	268	280	NA	11	1		Figure 4-2., A1 and Figure 4-7., AnelloTTSuV2
2	12	663	670	772	779	1	NA	1		
3	2	2875	2883	2887	2895	NA	NA	2		
4	3	333	339	353	359	NA	NA	3		
5	4	2784	2792	2797	2805	NA	NA	4		
6	5	2676	2683	2689	2696	NA	NA	5		
7	9	752	757	765	770	5	72	5		
8	6	2858	2866	2904	2912	NA	NA	6		

[a]Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b]NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID.
[c] Syn. Sub ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d]Coev. ranking is a ranking based on complementarily coevolution. NA indicates structures which had invariable sites.
[e]Minimum rank from NASP, Syn. Sub. and Coev. ranks.

Cons. Rank[a]	Structure ID / NASP Rank[b]	Coordinates Including Gaps				Syn. Sub. Rank[c]	Coev. Rank[d]	Min. Rank[e]	Biological function	Reference
		Left hand side		Right hand side						
9	7	472	480	547	555	NA	NA	7		
10	8	326	332	361	367	NA	NA	8		
11	17	1470	1483	1487	1500	NA	NA	17		
12	19	2480	2484	2489	2493	8	77	8		
13	10	2849	2855	2914	2920	NA	NA	10		
14	11	2817	2823	2828	2834	NA	61	11		
15	13	613	618	623	628	NA	27	13		
16	14	2020	2027	2216	2223	22	NA	14		
17	15	62	67	316	321	NA	8	8		
18	24	190	194	300	304	NA	59	24		
19	16	23	27	37	41	80	NA	16		
20	18	94	100	105	111	NA	49	18		
21	20	687	692	740	745	NA	66	20		
22	21	2461	2466	2550	2555	NA	NA	21		
23	22	80	82	87	89	NA	NA	22		
24	23	2708	2713	2717	2722	NA	15	15		
25	25	491	494	499	502	NA	NA	25		
26	26	1766	1772	1783	1789	42	NA	26		
27	27	76	78	91	93	NA	NA	27		
AnelloTTV (Genome length = 4129)										
1	1	1540	1548	1552	1560	40	NA	1		
2	12	2499	2509	3619	3629	1	26	1		
3	2	3434	3442	3458	3466	NA	20	2		
4	37	2488	2495	3631	3638	2	27	2		
5	3	3487	3495	3500	3508	NA	38	3		
6	4	3855	3867	3905	3917	NA	6	4		
7	5	3730	3739	3810	3819	NA	17	5		Figure 4-2., A1 and Figure 4-7., AnelloTTV
8	44	2519	2524	3602	3607	5	28	5		
9	6	3175	3183	3188	3196	NA	NA	6		
10	7	3376	3382	3395	3401	NA	NA	7		
11	55	3850	3853	3919	3922	194	7	7		
12	8	2909	2915	2926	2932	64	NA	8		
13	14	906	914	937	945	61	8	8		
14	9	3473	3481	3523	3531	NA	39	9		
15	78	921	923	928	930	100	9	9		
16	10	3088	3097	3102	3111	16	NA	10		
17	11	3941	3949	3955	3963	NA	NA	11		
18	59	2710	2719	3592	3601	11	29	11		
19	68	3745	3747	3793	3795	NA	12	12		
20	13	2589	2593	2598	2602	42	NA	13		
21	15	3268	3274	3415	3421	NA	NA	15		
22	16	3843	3849	3924	3930	NA	41	16		
23	17	2319	2325	2330	2336	63	NA	17		
24	57	2651	2655	2689	2693	17	NA	17		
25	18	4043	4051	4055	4063	NA	NA	18		
26	73	3742	3744	3805	3807	NA	18	18		
27	19	1205	1212	1265	1272	148	52	19		
28	46	723	728	752	757	19	NA	19		
29	20	1915	1922	2104	2111	108	NA	20		
30	21	4099	4107	4111	4119	NA	NA	21		
31	22	2273	2280	2286	2293	145	NA	22		
32	40	1536	1538	1561	1563	22	NA	22		

[a]Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b]NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID.
[c] Syn. Sub ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d]Coev. ranking is a ranking based on complementarily coevolution. NA indicates structures which had invariable sites.
[e]Minimum rank from NASP, Syn. Sub. and Coev. ranks.

Cons. Rank[a]	Structure ID / NASP Rank[b]	Coordinates Including Gaps				Syn. Sub. Rank[c]	Coev. Rank[d]	Min. Rank[e]	Biological function	Reference
		Left hand side		Right hand side						
33	23	3278	3282	3405	3409	NA	NA	23		
34	70	507	512	777	782	23	NA	23		
35	24	3308	3313	3326	3331	NA	NA	24		
36	26	2989	2994	3003	3008	24	NA	24		
37	25	2969	2974	2980	2985	35	NA	25		
38	27	3291	3295	3350	3354	NA	NA	27		
39	52	2741	2748	2816	2823	27	NA	27		
40	28	4026	4033	4067	4074	NA	NA	28		
41	29	4018	4025	4076	4083	NA	NA	29		
42	77	514	518	762	766	29	NA	29		
43	30	3024	3031	3039	3046	57	NA	30		
44	76	2656	2659	2684	2687	30	NA	30		
45	31	2858	2863	2882	2887	58	NA	31		
46	42	2749	2753	2810	2814	31	NA	31		
47	32	1233	1238	1247	1252	144	NA	32		
48	33	2793	2797	2803	2807	43	NA	33		
49	34	2179	2186	2222	2229	136	NA	34		
50	35	4011	4017	4089	4095	NA	NA	35		
51	36	1219	1223	1258	1262	147	NA	36		
52	38	3333	3336	3342	3345	NA	NA	38		
53	67	1135	1139	1146	1150	38	NA	38		
54	39	23	35	65	77	NA	NA	39		
55	41	1196	1201	1273	1278	152	53	41		
56	43	2762	2765	2769	2772	103	NA	43		
57	65	2612	2615	2628	2631	44	NA	44		
58	45	3755	3759	3781	3785	NA	NA	45		
59	56	1347	1356	1583	1592	82	46	46		
60	47	2832	2835	2840	2843	124	NA	47		
61	48	2263	2268	2301	2306	128	NA	48		
62	49	575	579	583	587	NA	NA	49		
63	50	3870	3873	3900	3903	NA	NA	50		
64	60	2899	2904	2937	2942	50	NA	50		
65	51	1228	1231	1253	1256	156	NA	51		
66	53	2864	2867	2877	2880	76	NA	53		
67	54	2017	2021	2026	2030	137	NA	54		
68	58	1924	1928	2095	2099	116	NA	58		
69	61	3245	3250	3258	3263	NA	NA	61		
70	62	3750	3753	3787	3790	NA	NA	62		
71	63	1990	1994	2003	2007	62	NA	62		
72	64	1469	1473	1508	1512	142	NA	64		
73	66	3315	3317	3322	3324	NA	NA	66		
74	69	2134	2139	2147	2152	75	NA	69		
75	71	2462	2467	2472	2477	166	NA	71		
76	72	3288	3290	3356	3358	NA	NA	72		
77	74	2066	2074	2081	2089	111	NA	74		
78	75	523	527	694	698	80	NA	75		
ParvoAAV (genome length = 5040)										
1	1	1	41	85	125	NA	3	1	Genome replication	Lusby, E., et al. (1980)
2	6	42	50	54	62	NA	1	1		
3	2	4916	4956	5000	5040	NA	NA	2	Genome replication	Lusby, E., et al. (1980)
4	5	4957	4965	4969	4977	NA	2	2		

[a]Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b]NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID.
[c] Syn. Sub ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d]Coev. ranking is a ranking based on complementarily coevolution. NA indicates structures which had invariable sites.
[e]Minimum rank from NASP, Syn. Sub. and Coev. ranks.

Cons. Rank[a]	Structure ID / NASP Rank[b]	Coordinates Including Gaps				Syn. Sub. Rank[c]	Coev. Rank[d]	Min. Rank[e]	Biological function	Reference
		Left hand side		Right hand side						
5	3	4979	4987	4991	4999	NA	6	3		
6	69	1836	1840	2001	2005	3	NA	3		
7	4	64	72	76	84	NA	NA	4		
8	7	1916	1921	1936	1941	114	NA	7		
9	8	1593	1602	1636	1645	147	NA	8		
10	9	175	184	190	199	NA	30	9		
11	10	521	530	537	546	11	NA	10		
12	11	1842	1852	1990	2000	20	NA	11		
13	12	1698	1708	2077	2087	211	NA	12		
14	13	1869	1873	1946	1950	229	NA	13		
15	28	220	233	320	333	13	57	13		
16	17	2331	2340	2383	2392	14	NA	14		Figure 4-2., P2 and Figure 4-5., ParvoAAV *ns1*-IR-*vp1*
17	14	434	439	456	461	142	NA	14		
18	15	2739	2749	2766	2776	176	NA	15		
19	47	411	414	462	465	179	15	15		
20	16	399	408	466	475	47	NA	16		Figure 4-2., P1 and Figure 4-5., ParvoAAV IR-*ns1*
21	35	1358	1365	1386	1393	107	16	16		
22	42	3299	3302	3306	3309	16	NA	16		
23	30	670	676	687	693	17	NA	17		
24	18	1462	1467	1474	1479	55	NA	18		
25	19	2216	2224	2245	2253	38	NA	19		
26	20	2680	2689	2797	2806	104	NA	20		
27	21	2712	2716	2721	2725	72	NA	21		
28	62	2440	2451	4714	4725	290	21	21		
29	22	3289	3294	3311	3316	110	NA	22		
30	23	2937	2943	2952	2958	257	43	23		
31	24	1610	1614	1626	1630	NA	NA	24		
32	63	4634	4641	4647	4654	24	NA	24		
33	25	729	738	798	807	289	54	25		
34	26	4531	4540	4547	4556	243	NA	26		
35	48	2514	2521	2527	2534	260	26	26		
36	27	3068	3078	4693	4703	112	NA	27		
37	29	301	306	311	316	288	NA	29		
38	31	1398	1405	1529	1536	183	NA	31		
39	32	2226	2229	2238	2241	108	NA	32		
40	33	1973	1976	1981	1984	274	NA	33		
41	34	3079	3087	3164	3172	141	NA	34		
42	56	2594	2597	2613	2616	261	35	35		
43	36	1223	1228	1283	1288	NA	NA	36		
44	37	417	421	426	430	NA	NA	37		
45	39	1749	1753	1758	1762	37	NA	37		
46	38	2310	2319	2398	2407	81	NA	38		
47	40	2635	2640	2645	2650	262	NA	40		
48	41	1236	1240	1251	1255	126	NA	41		
49	43	206	212	362	368	90	NA	43		
50	44	3478	3489	3704	3715	74	NA	44		
51	45	215	218	358	361	195	NA	45		
52	46	2705	2708	2726	2729	122	NA	46		
53	60	1219	1222	1290	1293	236	47	47		
54	49	1865	1867	1952	1954	239	NA	49		

[a]Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b]NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID. [c] Syn. Sub ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d]Coev. ranking is a ranking based on complementarily coevolution. NA indicates structures which had invariable sites.
[e]Minimum rank from NASP, Syn. Sub. and Coev. ranks.

Cons. Rank[a]	Structure ID / NASP Rank[b]	Coordinates Including Gaps				Syn. Sub. Rank[c]	Coev. Rank[d]	Min. Rank[e]	Biological function	Reference
		Left hand side		Right hand side						
55	50	1052	1064	1098	1110	200	NA	50		
56	51	3616	3624	3634	3642	113	NA	51		
57	52	1031	1038	1043	1050	215	NA	52		
58	53	4601	4607	4613	4619	57	NA	53		
59	54	2166	2170	2193	2197	282	62	54		
60	55	4521	4529	4557	4565	61	NA	55		
61	57	3644	3648	3653	3657	227	NA	57		
62	58	2344	2349	2361	2366	254	63	58		
63	59	4052	4061	4109	4118	246	64	59		
64	61	1212	1217	1294	1299	NA	NA	61		
65	64	4818	4821	4830	4833	NA	NA	64		
66	65	1782	1788	1807	1813	118	NA	65		
67	66	2321	2324	2393	2396	106	NA	66		
68	67	2866	2870	2892	2896	265	NA	67		
69	68	1191	1196	1206	1211	185	NA	68		
70	70	1558	1563	1569	1574	163	NA	70		
ParvoHBV (genome length = 5329)										
1	1	5162	5171	5190	5199	NA	83	1		
2	2	964	973	1860	1869	107	29	2		
3	41	2718	2726	2733	2741	119	2	2		
4	3	5201	5208	5214	5221	NA	NA	3		
5	37	1403	1411	1643	1651	137	3	3		
6	99	2020	2025	2056	2061	3	NA	3		
7	4	5173	5178	5183	5188	NA	84	4		
8	53	2031	2035	2046	2050	4	NA	4		
9	54	1455	1463	1524	1532	147	4	4		
10	5	1260	1268	1280	1288	89	NA	5		
11	12	2603	2609	2642	2648	5	NA	5		
12	87	1427	1432	1533	1538	142	5	5		
13	6	1102	1108	1776	1782	157	30	6		
14	7	1015	1021	1827	1833	134	31	7		
15	8	1949	1954	1959	1964	19	NA	8		
16	102	2169	2173	2235	2239	8	NA	8		
17	9	522	532	556	566	98	NA	9		
18	10	4755	4767	4771	4783	NA	NA	10		
19	14	421	427	437	443	38	10	10		
20	11	1893	1902	1907	1916	76	54	11		
21	52	403	408	456	461	23	11	11		
22	13	3378	3387	3452	3461	NA	NA	13		
23	65	3037	3042	3050	3055	14	NA	14		
24	15	1022	1028	1816	1822	141	32	15		
25	88	2848	2857	2940	2949	159	15	15		
26	16	3254	3260	3264	3270	NA	NA	16		
27	59	2611	2615	2634	2638	16	NA	16		
28	17	4828	4836	4931	4939	NA	NA	17		
29	18	1067	1071	1081	1085	22	NA	18		
30	ID 182 / rank 25	243	250	351	358	18	NA	18		Figure 4-2., P1 and Figure 4-5., ParvoHBoV IR-*ns1*
31	19	1094	1099	1785	1790	154	33	19		
32	27	1373	1381	1679	1687	155	19	19		
33	20	1985	1992	1998	2005	86	NA	20		
34	95	1383	1388	1673	1678	153	20	20		
35	21	5126	5135	5254	5263	NA	85	21	Genome replication	Kapoor, A., et al. (2011)

[a]Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b]NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID.
[c] Syn. Sub ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d]Coev. ranking is a ranking based on complementarily coevolution. NA indicates structures which had invariable sites.
[e]Minimum rank from NASP, Syn. Sub. and Coev. ranks.

Cons. Rank[a]	Structure ID / NASP Rank[b]	Coordinates Including Gaps				Syn. Sub. Rank[c]	Coev. Rank[d]	Min. Rank[e]	Biological function	Reference
		Left hand side		Right hand side						
36	105	1133	1139	1703	1709	166	21	21		
37	22	1919	1926	1930	1937	103	NA	22		
38	23	1047	1052	1792	1797	152	34	23		
39	24	3563	3568	3573	3578	NA	NA	24		
40	43	1060	1066	1087	1093	24	NA	24		
41	26	2176	2184	2224	2232	57	NA	26		
42	89	681	689	2698	2706	26	62	26		
43	28	1467	1475	1488	1496	110	NA	28		
44	29	4566	4572	4577	4583	NA	NA	29		
45	30	4672	4679	4683	4690	NA	NA	30		
46	31	4368	4376	4387	4395	NA	NA	31		
47	69	671	679	2708	2716	31	61	31		
48	32	373	381	387	395	118	96	32		
49	33	3831	3842	3943	3954	NA	NA	33		
50	34	1032	1038	1807	1813	150	35	34		
51	35	1122	1131	1747	1756	167	36	35		
52	36	3242	3247	3276	3281	NA	NA	36		
53	42	1040	1045	1801	1806	151	37	37		
54	38	2993	3000	3029	3036	77	94	38		
55	48	988	994	1852	1858	124	38	38		
56	39	44	50	120	126	NA	NA	39		
57	57	1006	1010	1835	1839	128	39	39		
58	40	3314	3319	3324	3329	NA	NA	40		
59	ID 125 / rank 44	2387	2391	2431	2435	48	66	44		Figure 4-2., P3 and Figure 4-5., ParvoHBoV IR-*np1*
60	45	4088	4095	4102	4109	NA	NA	45		
61	46	1497	1504	1509	1516	123	NA	46		
62	47	3348	3352	3362	3366	NA	NA	47		
63	50	3766	3772	3789	3795	NA	48	48		
64	49	1293	1300	1326	1333	94	NA	49		
65	81	217	225	2772	2780	49	52	49		
66	51	141	146	167	172	NA	NA	51		
67	55	2872	2878	2890	2896	133	NA	55		
68	86	1875	1881	1940	1946	106	55	55		
69	56	3998	4004	4025	4031	NA	NA	56		
70	92	1594	1601	1622	1629	56	NA	56		
71	58	3661	3668	3737	3744	NA	NA	58		
72	61	833	842	867	876	59	NA	59		
73	60	4415	4426	4595	4606	NA	NA	60		
74	98	3001	3004	3024	3027	61	NA	61		
75	62	2297	2303	2308	2314	NA	NA	62		
76	63	2245	2253	2360	2368	NA	NA	63		
77	64	2800	2805	2815	2820	117	NA	64		
78	74	1660	1663	1669	1672	121	64	64		
79	66	3806	3810	3818	3822	NA	NA	66		
80	83	2112	2117	2129	2134	66	NA	66		
81	67	4480	4489	4584	4593	NA	NA	67		
82	68	2483	2488	2511	2516	126	NA	68		
83	70	4493	4502	4550	4559	NA	NA	70		
84	71	4219	4224	4242	4247	NA	NA	71		
85	72	4005	4010	4018	4023	NA	NA	72		
86	73	3341	3345	3367	3371	NA	NA	73		
87	77	3086	3098	4790	4802	NA	73	73		

[a]Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b]NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID.
[c] Syn. Sub ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d]Coev. ranking is a ranking based on complementarily coevolution. NA indicates structures which had invariable sites.
[e]Minimum rank from NASP, Syn. Sub. and Coev. ranks.

Cons. Rank[a]	Structure ID / NASP Rank[b]	Coordinates Including Gaps				Syn. Sub. Rank[c]	Coev. Rank[d]	Min. Rank[e]	Biological function	Reference
		Left hand side		Right hand side						
88	80	3069	3077	4973	4981	NA	74	74		
89	75	3485	3491	3500	3506	NA	80	75		
90	76	4703	4709	4720	4726	NA	NA	76		
91	78	5136	5140	5248	5252	NA	86	78		
92	93	639	643	650	654	78	NA	78		
93	79	5142	5146	5242	5246	NA	87	79		
94	82	4607	4612	4619	4624	NA	NA	82		
95	84	1302	1306	1320	1324	85	NA	84		
96	85	708	716	734	742	95	NA	85		
97	90	3691	3698	3704	3711	NA	NA	90		
98	91	3860	3866	3871	3877	NA	NA	91		
99	94	5078	5085	5106	5113	NA	NA	94		
100	96	3633	3638	3644	3649	NA	NA	96		
101	97	189	194	199	204	NA	NA	97		
102	100	4665	4671	4694	4700	NA	NA	100		
103	101	4911	4914	4924	4927	NA	NA	101		
104	103	78	87	104	113	NA	NA	103		
105	104	3407	3413	3444	3450	NA	NA	104		
ParvoMPV (genome length = 5334)										
1	1	5106	5187	5235	5316	NA	27	1	Genome replication	Sun, Y., et al. (2009)
2	119	3547	3555	3617	3625	207	1	1		
3	2	3	27	96	120	NA	28	2	Genome replication / binding the viral capsid	Sun, Y., et al. (2009); Willwand, K. and Hirt, B. (2009)
4	3	2404	2416	2439	2451	NA	NA	3		
5	20	1996	2006	2274	2284	3	36	3		
6	4	30	43	79	92	NA	29	4		
7	5	5195	5211	5215	5231	NA	NA	5		
8	39	4346	4356	4374	4384	170	5	5		
9	42	1973	1980	2297	2304	5	37	5		Figure 4-2, P2 and Figure 4-5., ParvoMLP *ns1*-IR-*vp1*
10	6	2325	2336	2545	2556	NA	NA	6		
11	116	4338	4342	4387	4391	173	6	6		
12	7	848	858	871	881	177	NA	7		
13	25	2031	2038	2230	2237	7	NA	7		
14	8	2708	2718	2742	2752	NA	NA	8		
15	68	2024	2028	2238	2242	8	NA	8		
16	112	4513	4522	4992	5001	82	8	8		
17	9	310	321	333	344	231	20	9	Transcription attenuation	Perros, M., et al. (1994)
18	10	1718	1724	1741	1747	182	NA	10		
19	11	2249	2256	2262	2269	35	34	11		
20	34	2040	2045	2224	2229	11	NA	11		
21	12	913	921	949	957	115	NA	12		
22	13	2869	2878	2890	2899	119	NA	13		
23	103	3210	3216	3260	3266	203	13	13		
24	14	2124	2131	2135	2142	27	NA	14		
25	64	2098	2102	2164	2168	14	NA	14		
26	15	3582	3590	3599	3607	87	NA	15		
27	35	2077	2082	2209	2214	15	NA	15		
28	16	1248	1254	1268	1274	139	NA	16		

[a]Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b]NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID.
[c] Syn. Sub ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d]Coev. ranking is a ranking based on complementarily coevolution. NA indicates structures which had invariable sites.
[e]Minimum rank from NASP, Syn. Sub. and Coev. ranks.

Cons. Rank[a]	Structure ID / NASP Rank[b]	Coordinates Including Gaps				Syn. Sub. Rank[c]	Coev. Rank[d]	Min. Rank[e]	Biological function	Reference
		Left hand side		Right hand side						
29	50	2091	2096	2170	2175	16	NA	16		
30	17	1836	1847	2589	2600	20	35	17		
31	76	5019	5021	5027	5029	NA	17	17		
32	18	4402	4412	4433	4443	226	NA	18		
33	91	1830	1834	2602	2606	18	38	18		
34	ID 239 / rank 19	212	215	425	428	NA	51	19	Transcription attenuation	Figure 4-2., P1 and Figure 4-5., ParvoMLP IR-*ns1*; Perros, M., et al. (1994)
35	21	2352	2358	2473	2479	NA	NA	21		
36	22	1070	1075	1086	1091	32	NA	22		
37	126	1766	1771	2625	2630	22	39	22		
38	23	3432	3441	3446	3455	137	NA	23		
39	24	979	991	1133	1145	33	NA	24		
40	127	2184	2186	2197	2199	24	NA	24		
41	63	2105	2108	2120	2123	25	NA	25		
42	26	2849	2855	2860	2866	40	NA	26		
43	27	3647	3659	3916	3928	126	56	27		
44	28	4790	4795	4801	4806	NA	NA	28		
45	106	967	974	1186	1193	28	NA	28		
46	29	1601	1608	1628	1635	110	NA	29		
47	37	2144	2148	2153	2157	29	NA	29		
48	30	1312	1315	1320	1323	38	NA	30		
49	51	2177	2182	2201	2206	30	NA	30		
50	31	1796	1800	1806	1810	134	NA	31		
51	32	2340	2345	2483	2488	NA	NA	32		
52	33	1380	1386	1468	1474	194	NA	33		
53	36	1445	1450	1457	1462	180	NA	36		
54	38	4623	4630	4899	4906	NA	NA	38		
55	40	49	54	58	63	NA	NA	40		
56	ID 282 / rank 41	2786	2787	2919	2920	128	NA	41		Figure 4-2., P4 and Figure 4-5., ParvoMLP IR-*vp2*
57	109	1076	1077	1082	1083	41	NA	41		
58	60	515	522	546	553	42	NA	42		
59	43	3770	3782	3836	3848	88	57	43		
60	44	1165	1168	1173	1176	90	NA	44		
61	45	1400	1405	1438	1443	214	NA	45		
62	46	3869	3874	3881	3886	167	NA	46		
63	67	2054	2058	2070	2074	46	NA	46		
64	47	345	354	363	372	51	NA	47		
65	48	1555	1561	1594	1600	190	NA	48		
66	49	45	48	75	78	NA	NA	49		
67	93	1051	1059	1113	1121	49	NA	49		
68	52	2422	2424	2429	2431	NA	NA	52		
69	117	231	233	239	241	NA	52	52		
70	53	929	932	939	942	120	NA	53		
71	54	808	817	822	831	249	NA	54		
72	55	3334	3340	3350	3356	54	NA	54		
73	56	2649	2657	2666	2674	NA	NA	56		
74	121	1301	1306	1352	1357	56	NA	56		
75	57	4317	4322	4327	4332	102	NA	57		

[a]Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b]NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID.
[c] Syn. Sub ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d]Coev. ranking is a ranking based on complementarily coevolution. NA indicates structures which had invariable sites.
[e]Minimum rank from NASP, Syn. Sub. and Coev. ranks.

Cons. Rank[a]	Structure ID / NASP Rank[b]	Coordinates Including Gaps				Syn. Sub. Rank[c]	Coev. Rank[d]	Min. Rank[e]	Biological function	Reference
		Left hand side		Right hand side						
76	58	1369	1375	1479	1485	143	NA	58		
77	59	1924	1928	1947	1951	61	NA	59		
78	85	997	999	1004	1006	59	NA	59		
79	61	2306	2309	2570	2573	NA	NA	61		
80	62	2314	2317	2563	2566	NA	NA	62		
81	81	3717	3724	3739	3746	63	NA	63		
82	65	1539	1545	1684	1690	148	NA	65		
83	129	3665	3670	3680	3685	65	NA	65		
84	66	3861	3868	3892	3899	155	NA	66		
85	86	1881	1888	1896	1903	66	NA	66		
86	72	1931	1934	1941	1944	67	NA	67		
87	69	2497	2500	2505	2508	NA	NA	69		
88	70	5191	5193	5232	5234	NA	NA	70		
89	71	4810	4814	4821	4825	NA	NA	71		
90	73	64	67	71	74	NA	NA	73		
91	74	1876	1880	1913	1917	123	NA	74		
92	75	1863	1867	1960	1964	144	NA	75		
93	122	993	995	1008	1010	75	NA	75		
94	77	860	863	867	870	204	NA	77		
95	125	3312	3321	3461	3470	77	NA	77		
96	78	1416	1419	1427	1430	218	NA	78		
97	79	1412	1415	1432	1435	227	NA	79		
98	80	2721	2726	2731	2736	NA	NA	80		
99	82	2420	2421	2433	2434	NA	NA	82		
100	83	1546	1551	1663	1668	156	NA	83		
101	84	3850	3858	3901	3909	236	NA	84		
102	110	3371	3376	3383	3388	86	NA	86		
103	87	2960	2968	2979	2987	232	NA	87		
104	88	1612	1615	1621	1624	165	NA	88		
105	89	2680	2687	2767	2774	NA	NA	89		
106	90	4157	4166	4201	4210	253	NA	90		
107	115	1220	1224	1232	1236	91	NA	91		
108	92	4178	4185	4189	4196	252	NA	92		
109	94	2793	2797	2802	2806	NA	NA	94		
110	120	4065	4075	4102	4112	94	NA	94		
111	95	1524	1531	1700	1707	150	NA	95		
112	96	3001	3008	3015	3022	174	NA	96		
113	97	1327	1333	1343	1349	114	NA	97		
114	98	2391	2394	2457	2460	NA	NA	98		
115	99	142	147	170	175	NA	NA	99		
116	100	4236	4244	4259	4267	247	NA	100		
117	101	4848	4859	4863	4874	NA	NA	101		
118	102	1920	1922	1952	1954	118	NA	102		
119	104	4358	4362	4369	4373	111	NA	104		
120	105	890	896	905	911	157	NA	105		
121	107	1726	1729	1734	1737	217	NA	107		
122	108	4707	4712	4717	4722	NA	NA	108		
123	123	923	924	947	948	109	NA	109		
124	111	4692	4698	4750	4756	NA	NA	111		
125	113	5087	5094	5098	5105	NA	NA	113		
126	114	3952	3960	3999	4007	213	NA	114		
127	118	1360	1364	1493	1497	200	161	118		
128	124	2396	2397	2455	2456	NA	NA	124		

[a]Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b]NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID.
[c] Syn. Sub ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d]Coev. ranking is a ranking based on complementarily coevolution. NA indicates structures which had invariable sites.
[e]Minimum rank from NASP, Syn. Sub. and Coev. ranks.

Cons. Rank[a]	Structure ID / NASP Rank[b]	Coordinates Including Gaps				Syn. Sub. Rank[c]	Coev. Rank[d]	Min. Rank[e]	Biological function	Reference
		Left hand side		Right hand side						
129	128	5074	5077	5082	5085	NA	NA	128		
130	130	2494	2495	2509	2510	NA	NA	130		
131	131	189	193	201	205	NA	NA	131		
132	132	737	741	746	750	192	NA	132		
NanoBBTV-R (genome length =1158)										
1	1	379	389	393	403	34	8	1		
2	22	789	794	810	815	49	1	1		
3	26	345	348	368	371	1	NA	1		
4	2	131	143	147	159	26	NA	2		
5	3	614	625	629	640	40	NA	3		
6	27	276	279	306	309	3	NA	3		
7	4	519	531	688	700	60	NA	4		
8	5	7	18	1127	1138	NA	NA	5	Initiation of genome replication	Hafner, G.J., et al. (1997)
9	6	828	838	848	858	5	13	5		
10	11	496	504	731	739	61	5	5		
11	7	896	903	907	914	12	NA	7		
12	23	39	45	452	458	7	10	7		
13	8	55	67	182	194	22	NA	8		
14	9	351	356	361	366	9	NA	9		
15	10	328	335	411	418	15	9	9		
16	12	589	594	598	603	28	NA	12		
17	13	266	274	310	318	13	NA	13		
18	14	761	768	781	788	45	NA	14		
19	31	823	826	859	862	21	14	14		
20	15	22	29	1104	1111	NA	NA	15		
21	16	236	241	246	251	18	NA	16		
22	17	114	117	123	126	30	NA	17		
23	18	1051	1054	1058	1061	NA	NA	18		
24	19	421	427	431	437	32	NA	19		
25	29	941	949	953	961	19	NA	19		
26	20	915	919	923	927	58	NA	20		
27	21	475	478	483	486	36	NA	21		
28	25	78	82	96	100	23	NA	23		
29	24	877	882	887	892	25	NA	24		
30	28	660	663	667	670	50	NA	28		
31	30	463	469	489	495	37	NA	30		
NanoBBTV-S (genome length = 1128)										
1	1	940	949	1035	1044	5	NA	1		
2	5	397	403	425	431	78	1	1		
3	2	337	346	381	390	2	2	2		
4	3	531	539	545	553	52	12	3		
5	12	178	183	187	192	3	NA	3		
6	4	741	748	753	760	64	24	4		
7	24	167	174	198	205	4	NA	4		
8	6	630	636	641	647	36	19	6		
9	19	359	365	371	377	6	21	6		
10	36	281	285	313	317	68	6	6		
11	7	1054	1067	1077	1090	61	NA	7		
12	8	991	997	1001	1007	9	NA	8		
13	9	406	412	416	422	30	8	8		
14	30	461	463	468	470	15	9	9		
15	10	706	712	728	734	46	26	10		

[a]Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b]NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID.
[c] Syn. Sub ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d]Coev. ranking is a ranking based on complementarily coevolution. NA indicates structures which had invariable sites.
[e]Minimum rank from NASP, Syn. Sub. and Coev. ranks.

Cons. Rank[a]	Structure ID / NASP Rank[b]	Coordinates Including Gaps				Syn. Sub. Rank[c]	Coev. Rank[d]	Min. Rank[e]	Biological function	Reference
		Left hand side		Right hand side						
16	26	7	10	1120	1123	10	NA	10		
17	46	288	291	308	311	NA	10	10		
18	11	454	460	473	479	58	29	11		
19	29	663	668	676	681	11	37	11		
20	13	957	966	975	984	41	NA	13		
21	41	250	255	260	265	NA	13	13		
22	14	687	693	795	801	56	27	14		
23	27	104	107	137	140	14	NA	14		
24	15	648	656	807	815	35	30	15		
25	35	354	356	378	380	55	15	15		
26	16	207	217	572	582	69	17	16		
27	17	238	249	557	568	16	18	16		
28	18	11	15	1114	1118	NA	NA	18	Initiation of genome replication	Hafner, G.J., et al. (1997)
29	21	761	765	770	774	19	NA	19		
30	20	322	326	332	336	67	23	20		
31	23	713	717	722	726	20	31	20		
32	22	112	116	129	133	39	NA	22		
33	39	618	621	625	628	54	22	22		
34	31	24	26	1091	1093	23	NA	23		
35	25	481	485	500	504	57	38	25		
36	38	80	82	86	88	25	NA	25		
37	28	144	148	153	157	37	NA	28		
38	37	694	701	783	790	29	28	28		
39	32	73	78	89	94	50	NA	32		
40	33	512	515	519	522	72	34	33		
41	34	892	894	903	905	33	NA	33		
42	40	118	121	125	128	NA	NA	40		
43	42	895	896	900	901	NA	NA	42		
44	43	1068	1069	1074	1075	NA	NA	43		
45	44	109	110	135	136	NA	NA	44		
46	45	832	837	932	937	NA	NA	45		
47	47	175	176	194	195	NA	NA	47		
48	48	1100	1102	1106	1108	NA	NA	48		
49	49	868	871	878	881	NA	NA	49		
NanoBBTV-M (genome length = 1080)										
1	1	29	40	44	55	NA	43	1		
2	12	346	354	360	368	1	NA	1		
3	2	753	763	769	779	NA	46	2		
4	21	799	809	879	889	NA	2	2		
5	3	110	121	133	144	NA	NA	3		
6	27	286	293	371	378	7	3	3		
7	4	892	901	988	997	NA	33	4		
8	5	557	568	583	594	12	NA	5		
9	6	146	154	164	172	NA	55	6		
10	7	295	304	308	317	23	17	7		
11	29	222	231	654	663	13	7	7		
12	8	405	418	423	436	19	28	8		
13	17	505	510	543	548	8	26	8		
14	9	7	17	1065	1075	NA	52	9	Initiation of genome replication	Hafner, G.J., et al. (1997)
15	32	519	521	531	533	20	9	9		
16	10	201	208	213	220	NA	22	10		

[a]Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b]NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID.
[c] Syn. Sub ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d]Coev. ranking is a ranking based on complementarily coevolution. NA indicates structures which had invariable sites.
[e]Minimum rank from NASP, Syn. Sub. and Coev. ranks.

Cons. Rank[a]	Structure ID / NASP Rank[b]	Coordinates Including Gaps				Syn. Sub. Rank[c]	Coev. Rank[d]	Min. Rank[e]	Biological function	Reference
		Left hand side		Right hand side						
17	11	943	950	954	961	NA	47	11		
18	16	511	517	535	541	11	38	11		
19	28	245	252	262	269	NA	11	11		
20	13	178	186	190	198	NA	35	13		
21	25	627	632	636	641	NA	13	13		
22	14	723	729	734	740	NA	29	14		
23	22	462	469	473	480	14	NA	14		
24	15	695	704	708	717	NA	36	15		
25	26	392	399	642	649	16	41	16		
26	18	911	917	928	934	NA	20	18		
27	19	1018	1022	1027	1031	NA	NA	19		
28	20	844	848	853	857	NA	NA	20		
29	23	22	28	1041	1047	NA	23	23		
30	24	67	79	83	95	NA	45	24		
31	30	608	613	621	626	25	34	25		
NanoBBTV-N (genome length = 1124)										
1	1	152	163	167	178	NA	15	1		
2	18	974	977	987	990	NA	1	1		
3	2	82	96	103	117	NA	3	2		
4	9	678	683	755	760	2	NA	2		
5	3	185	194	198	207	NA	8	3		
6	4	942	952	1044	1054	NA	6	4		
7	5	363	372	402	411	21	NA	5		
8	6	48	55	63	70	NA	NA	6		
9	7	800	806	812	818	NA	NA	7		
10	8	424	431	450	457	32	17	8		
11	14	626	630	657	661	9	NA	9		
12	10	7	18	1108	1119	NA	NA	10	Initiation of genome replication	Hafner, G.J., et al. (1997)
13	11	723	730	734	741	16	NA	11		
14	12	892	896	900	904	NA	NA	12		
15	13	713	718	745	750	17	NA	13		
16	17	335	340	345	350	13	NA	13		
17	15	231	236	281	286	40	NA	15		
18	16	22	29	1094	1101	NA	NA	16		
19	19	635	638	646	649	19	NA	19		
NanoBBTV-C (genome length = 1030)										
1	1	110	124	128	142	NA	1	1		
2	21	434	437	442	445	1	NA	1		
3	2	28	40	44	56	NA	NA	2		
4	3	724	731	742	749	NA	NA	3		
5	9	205	212	268	275	30	3	3		
6	4	582	592	597	607	20	NA	4		
7	13	197	202	278	283	23	4	4		
8	5	304	311	318	325	7	NA	5		
9	22	220	224	231	235	32	5	5		
10	6	239	246	253	260	16	NA	6		
11	7	648	654	658	664	9	NA	7		
12	24	370	377	387	394	10	7	7		
13	8	684	692	698	706	25	NA	8		
14	10	920	924	928	932	NA	NA	10		

[a]Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b]NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID.
[c] Syn. Sub ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d]Coev. ranking is a ranking based on complementarily coevolution. NA indicates structures which had invariable sites.
[e]Minimum rank from NASP, Syn. Sub. and Coev. ranks.

Cons. Rank[a]	Structure ID / NASP Rank[b]	Coordinates Including Gaps				Syn. Sub. Rank[c]	Coev. Rank[d]	Min. Rank[e]	Biological function	Reference
		Left hand side		Right hand side						
15	11	7	13	1019	1025	NA	NA	11	Genome replication	Hafner, G.J., et al. (1997)
16	12	893	897	901	905	NA	NA	12		
17	14	407	412	419	424	18	NA	14		
18	15	750	755	815	820	NA	NA	15		
19	23	621	625	711	715	15	NA	15		
20	16	179	183	188	192	NA	NA	16		
21	17	164	169	173	178	NA	NA	17		
22	18	761	765	774	778	NA	NA	18		
23	19	756	760	781	785	NA	NA	19		
24	20	460	465	470	475	35	NA	20		
GeminiMSV (genome length = 2745)										
1	1	3	20	2718	2735	NA	NA	1	Initiation of genome replication	Orozco, B.M. and Hanley-Bowdoin, L., (1996)
2	48	2440	2449	2600	2609	1	5	1		
3	2	1628	1641	1760	1773	30	8	2		
4	13	1397	1404	1408	1415	25	2	2		
5	3	535	547	551	563	102	23	3		
6	16	1481	1487	1493	1499	20	3	3		
7	4	2116	2127	2131	2142	29	NA	4		
8	43	1476	1480	1501	1505	10	4	4		
9	5	808	824	901	917	141	NA	5		
10	25	1663	1671	1703	1711	5	10	5		
11	6	1512	1526	1975	1989	98	52	6		
12	7	2185	2195	2201	2211	35	19	7		
13	9	1418	1428	1435	1445	22	7	7		
14	8	429	439	448	458	110	44	8		
15	15	1680	1687	1691	1698	90	9	9		
16	ID 172 / rank10	2050	2051	2276	2277	77	45	10		Shepherd, D.N., et al. (2006)
17	11	2391	2400	2404	2413	34	77	11		
18	44	1675	1678	1699	1702	39	11	11		
19	12	1124	1133	1137	1146	93	14	12		
20	42	1932	1936	1940	1944	13	NA	13		
21	14	2498	2506	2511	2519	112	NA	14		
22	23	1619	1627	1870	1878	14	53	14		
23	24	1264	1269	1274	1279	NA	15	15		
24	39	920	936	940	956	100	16	16		
25	17	1815	1824	1828	1837	23	NA	17		
26	18	379	386	391	398	49	NA	18		
27	19	65	72	84	91	41	33	19		
28	20	1028	1035	1066	1073	58	NA	20		
29	30	1080	1087	1172	1179	129	20	20		
30	21	1997	2003	2007	2013	65	NA	21		
31	22	2687	2695	2709	2717	NA	NA	22		
32	50	700	707	800	807	92	22	22		
33	26	1208	1214	1222	1228	NA	NA	26		
34	31	569	575	580	586	64	26	26		
35	27	1100	1108	1114	1122	147	NA	27		
36	28	2029	2034	2038	2043	105	63	28		

[a]Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b]NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID.
[c] Syn. Sub ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d]Coev. ranking is a ranking based on complementarily coevolution. NA indicates structures which had invariable sites.
[e]Minimum rank from NASP, Syn. Sub. and Coev. ranks.

Cons. Rank[a]	Structure ID / NASP Rank[b]	Coordinates Including Gaps				Syn. Sub. Rank[c]	Coev. Rank[d]	Min. Rank[e]	Biological function	Reference
		Left hand side		Right hand side						
37	29	1156	1160	1165	1169	66	NA	29		
38	32	1545	1554	1558	1567	133	NA	32		
39	33	656	664	682	690	82	NA	33		
40	34	363	374	405	416	69	NA	34		
41	35	2365	2370	2375	2380	95	NA	35		
42	36	1990	1996	2015	2021	113	NA	36		
43	47	2386	2389	2414	2417	36	78	36		
44	37	143	148	153	158	76	NA	37		
45	38	2228	2235	2266	2273	128	49	38		
46	55	126	131	135	140	NA	39	39		
47	40	1884	1890	1895	1901	99	NA	40		
48	41	1715	1719	1741	1745	59	NA	41		
49	46	1234	1238	1254	1258	NA	42	42		
50	54	1229	1232	1259	1262	NA	43	43		
51	ID 65 / rank 45	224	230	344	350	46	NA	45		Figure 4-1., G1 and Figure 4-3., GeminiMSV
52	49	474	479	492	497	67	70	49		
53	53	711	718	726	733	88	50	50		
54	51	1021	1025	1074	1078	68	62	51		
55	52	247	250	257	260	73	NA	52		
56	56	988	995	1000	1007	108	NA	56		
GeminiPanSV (genome length = 2763)										
1	1	1652	1665	1784	1797	NA	8	1		
2	2	5	16	2740	2751	NA	NA	2	Initiation of genome replication	Orozco, B.M. and Hanley-Bowdoin, L., (1996)
3	4	1543	1551	1556	1564	2	NA	2		
4	11	1312	1318	1323	1329	NA	2	2		
5	3	555	568	580	593	136	15	3		
6	26	1193	1198	1203	1208	3	NA	3		
7	14	457	463	471	477	114	4	4		
8	5	2380	2388	2392	2400	12	NA	5		
9	20	1127	1132	1142	1147	5	NA	5		
10	31	1645	1651	1891	1897	22	5	5		
11	6	1225	1232	1236	1243	34	11	6		
12	7	732	739	832	839	91	NA	7		
13	8	1915	1928	1953	1966	86	47	8		
14	9	2138	2145	2152	2159	101	NA	9		
15	21	1839	1847	1851	1859	9	12	9		
16	10	1677	1683	1688	1694	NA	NA	10		
17	12	1705	1711	1715	1721	NA	NA	12		
18	13	2107	2118	2231	2242	89	33	13		
19	15	2530	2538	2542	2550	73	NA	15		
20	16	1253	1259	1265	1271	NA	NA	16		
21	17	234	240	354	360	59	NA	17		Figure 4-1., G1 and Figure 4-3., GeminiPanSV
22	18	1060	1066	1099	1105	30	20	18		
23	19	2682	2688	2715	2721	NA	NA	19		
24	22	1969	1978	2018	2027	119	NA	22		
25	23	2722	2727	2732	2737	93	NA	23		

[a]Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b]NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID.
[c] Syn. Sub ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d]Coev. ranking is a ranking based on complementarily coevolution. NA indicates structures which had invariable sites.
[e]Minimum rank from NASP, Syn. Sub. and Coev. ranks.

Cons. Rank[a]	Structure ID / NASP Rank[b]	Coordinates Including Gaps				Syn. Sub. Rank[c]	Coev. Rank[d]	Min. Rank[e]	Biological function	Reference
		Left hand side		Right hand side						
26	24	1440	1453	1464	1477	149	35	24		
27	25	1354	1358	1367	1371	NA	NA	25		
28	27	80	86	97	103	NA	NA	27		
29	28	316	320	324	328	56	NA	28		
30	29	1288	1292	1296	1300	NA	NA	29		
31	30	2439	2445	2454	2460	78	NA	30		
GeminiWDV (genome length = 2762)										
1	1	3	19	2735	2751	NA	NA	1	Initiation of genome replication	Orozco, B.M. and Hanley-Bowdoin, L., (1996)
2	19	2313	2319	2358	2364	1	NA	1		
3	2	22	31	40	49	NA	NA	2		
4	3	1195	1205	1209	1219	71	NA	3		
5	4	1378	1385	1389	1396	NA	NA	4		
6	5	2325	2332	2337	2344	28	NA	5		
7	16	844	854	865	875	42	5	5		
8	6	1529	1539	1555	1565	NA	NA	6		
9	7	2664	2678	2708	2722	NA	6	6		
10	8	2195	2202	2206	2213	15	NA	8		
11	9	625	631	674	680	22	NA	9		
12	10	653	658	663	668	18	NA	10		
13	11	1740	1745	1750	1755	NA	NA	11		
14	12	1921	1928	2029	2036	108	NA	12		
15	13	1400	1405	1409	1414	NA	NA	13		
16	14	233	246	371	384	65	NA	14		Figure 4-1., G1 and Figure 4-3., GeminiWDV
17	20	1103	1112	1126	1135	87	14	14		
18	15	1241	1246	1251	1256	NA	NA	15		
19	22	1058	1064	1153	1159	62	15	15		
20	24	1040	1045	1167	1172	35	16	16		
21	17	220	232	485	497	21	NA	17		
22	18	1821	1828	1853	1860	36	NA	18		
23	21	2117	2124	2130	2137	55	NA	21		
24	23	1443	1450	1470	1477	NA	NA	23		
25	25	632	638	643	649	64	NA	25		
26	26	2071	2076	2081	2086	70	NA	26		
27	27	2367	2373	2409	2415	90	NA	27		
28	28	2599	2605	2613	2619	NA	NA	28		
29	29	1626	1631	1636	1641	NA	NA	29		
30	30	817	820	826	829	NA	NA	30		
GeminiTYDV-CpCV (genome length = 2642)										
1	1	697	710	795	808	131	24	1		
2	4	1552	1565	1570	1583	65	1	1		
3	2	3	12	2621	2630	NA	NA	2	Initiation of genome replication	Orozco, B.M. and Hanley-Bowdoin, L., (1996)
4	21	2494	2502	2544	2552	NA	2	2		
5	3	1103	1114	1123	1134	45	20	3		
6	40	389	392	401	404	4	NA	4		
7	5	850	858	862	870	46	NA	5		
8	9	2301	2307	2314	2320	5	NA	5		
9	6	1062	1069	1074	1081	60	NA	6		

[a]Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b]NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID.
[c] Syn. Sub ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d]Coev. ranking is a ranking based on complementarily coevolution. NA indicates structures which had invariable sites.
[e]Minimum rank from NASP, Syn. Sub. and Coev. ranks.

Cons. Rank[a]	Structure ID / NASP Rank[b]	Coordinates Including Gaps				Syn. Sub. Rank[c]	Coev. Rank[d]	Min. Rank[e]	Biological function	Reference
		Left hand side		Right hand side						
10	7	265	275	372	382	116	NA	7		Figure 4-1., G1 and Figure 4-3., GeminiTYDV-CpCV
11	8	552	560	573	581	53	NA	8		
12	10	2379	2388	2392	2401	50	NA	10		
13	16	2250	2259	2270	2279	10	NA	10		
14	11	1711	1719	1729	1737	NA	40	11		
15	12	1798	1809	1840	1851	96	14	12		
16	13	1907	1916	1923	1932	98	52	13		
17	14	2124	2136	2144	2156	125	49	14		
18	15	592	597	601	606	63	NA	15		
19	17	1401	1409	1462	1470	25	NA	17		
20	18	34	37	42	45	NA	NA	18		
21	19	27	32	2608	2613	NA	NA	19		
22	20	1223	1229	1234	1240	NA	30	20		
23	27	924	936	1027	1039	122	21	21		
24	22	961	971	975	985	114	46	22		
25	23	1143	1147	1153	1157	58	48	23		
26	24	2057	2065	2086	2094	78	NA	24		
27	38	1381	1387	1489	1495	24	NA	24		
28	25	1862	1873	1879	1890	100	36	25		
29	26	562	565	569	572	101	NA	26		
30	28	1626	1633	1647	1654	NA	NA	28		
31	29	283	292	301	310	30	NA	29		
32	30	1414	1419	1423	1428	66	NA	30		
33	31	334	340	344	350	132	NA	31		
34	43	1161	1171	1275	1285	47	31	31		
35	32	2369	2376	2419	2426	33	NA	32		
36	33	1938	1946	2027	2035	113	NA	33		
37	34	584	589	691	696	82	NA	34		
38	35	1608	1614	1679	1685	NA	NA	35		
39	36	1367	1369	1375	1377	75	NA	36		
40	37	2108	2117	2157	2166	81	50	37		
41	39	636	640	676	680	92	NA	39		
42	46	2402	2405	2409	2412	40	NA	40		
43	41	20	22	2618	2620	NA	NA	41		
44	42	520	524	529	533	57	NA	42		
45	44	1349	1353	1359	1363	76	NA	44		
46	45	1947	1953	1958	1964	102	NA	45		
47	47	630	634	682	686	94	NA	47		
GeminiCpCDV (genome length = 2600)										
1	1	1071	1079	1096	1104	25	NA	1		
2	12	1650	1659	1899	1908	31	1	1		
3	2	2312	2321	2325	2334	14	NA	2		
4	20	1552	1561	1919	1928	5	2	2		
5	3	1490	1501	2289	2300	9	18	3		
6	4	10	24	2566	2580	NA	NA	4		
7	5	238	247	363	372	123	8	5		Figure 4-1., G1 and Figure 4-3., GeminiCpCDV
8	6	2167	2175	2207	2215	29	NA	6		
9	7	1127	1138	1237	1248	68	NA	7		
10	8	1482	1489	2304	2311	7	19	7		
11	9	1863	1873	1878	1888	85	NA	9		

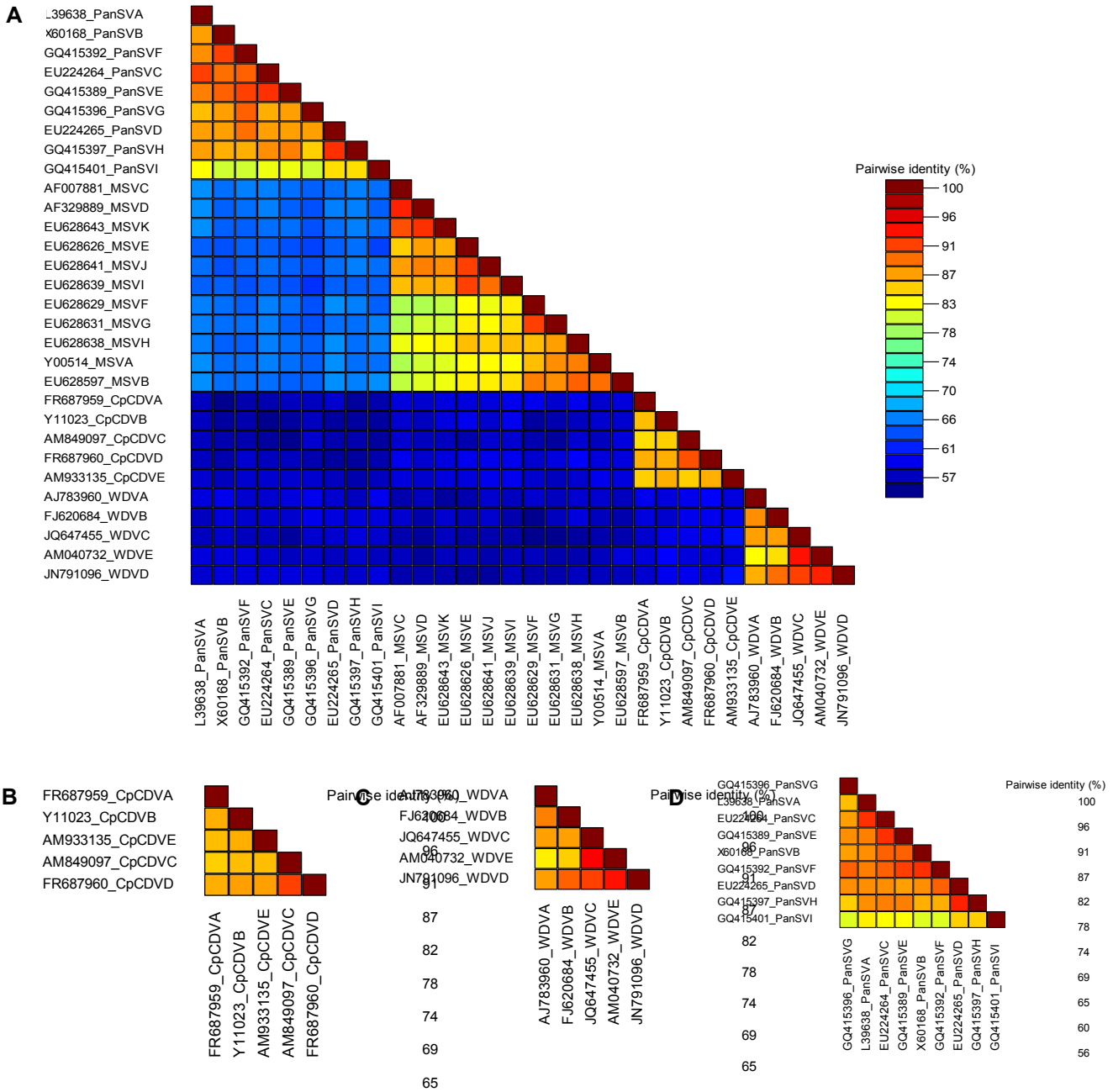
[a]Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b]NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID.
[c] Syn. Sub ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d]Coev. ranking is a ranking based on complementarily coevolution. NA indicates structures which had invariable sites.
[e]Minimum rank from NASP, Syn. Sub. and Coev. ranks.

Cons. Rank[a]	Structure ID / NASP Rank[b]	Coordinates Including Gaps				Syn. Sub. Rank[c]	Coev. Rank[d]	Min. Rank[e]	Biological function	Reference
		Left hand side		Right hand side						
12	10	3	9	2583	2589	NA	NA	10	Initiation of genome replication	Orozco, B.M. and Hanley-Bowdoin, L., (1996)
13	11	318	324	328	334	59	20	11		
14	13	2341	2347	2351	2357	38	NA	13		
15	14	880	885	895	900	82	NA	14		
16	15	844	854	858	868	70	NA	15		
17	16	1514	1519	1538	1543	NA	NA	16		
18	17	2118	2125	2134	2141	55	NA	17		
19	18	1994	2000	2004	2010	84	32	18		
20	19	2181	2186	2200	2205	43	NA	19		
21	21	616	624	635	643	61	NA	21		
22	22	2385	2391	2396	2402	52	NA	22		
23	23	531	540	778	787	86	NA	23		
24	24	373	378	383	388	32	NA	24		
25	25	654	660	770	776	104	NA	25		
GeminiTYLCV (genome length = 2832)										
1	1	1532	1544	1586	1598	NA	30	1		
2	2	309	320	332	343	NA	NA	2		
3	12	1892	1899	1932	1939	91	2	2		
4	3	1904	1913	1918	1927	61	NA	3		
5	4	2048	2056	2060	2068	48	NA	4		
6	24	1658	1667	1812	1821	67	4	4		
7	5	2713	2723	2731	2741	NA	23	5		
8	6	3	9	2817	2823	NA	NA	6	Initiation of genome replication	Orozco, B.M. and Hanley-Bowdoin, L., (1996)
9	9	1843	1848	1853	1858	6	NA	6		
10	7	1825	1837	1866	1878	20	19	7		
11	8	394	401	405	412	NA	NA	8		
12	10	41	54	156	169	NA	17	10		
13	11	253	258	263	268	NA	NA	11		
14	13	1735	1741	1765	1771	17	NA	13		
15	14	2101	2107	2111	2117	62	NA	14		
16	15	122	128	133	139	NA	NA	15		
17	16	653	667	672	686	88	15	15		
18	17	1722	1727	1783	1788	47	NA	17		
19	18	97	104	109	116	NA	NA	18		
20	19	1013	1018	1023	1028	27	NA	19		
21	ID 138 / rank 20	1048	1050	1144	1146	50	49	20		Figure 4-1., G2 and Figure 4-4., GeminiTYLCV
22	21	1421	1429	1433	1441	NA	NA	21		
23	22	1717	1721	1790	1794	31	NA	22		
24	23	183	191	196	204	NA	NA	23		
25	25	437	450	1489	1502	76	31	25		
GeminiEACMV (genome length =2820)										
1	1	3	13	2801	2811	NA	NA	1	Initiation of genome replication	Orozco, B.M. and Hanley-Bowdoin, L., (1996)
2	2	2714	2725	2730	2741	NA	41	2		
3	3	1860	1867	1872	1879	8	NA	3		
4	4	464	472	477	485	NA	NA	4		

[a]Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b]NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID.
[c] Syn. Sub ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d]Coev. ranking is a ranking based on complementarily coevolution. NA indicates structures which had invariable sites.
[e]Minimum rank from NASP, Syn. Sub. and Coev. ranks.

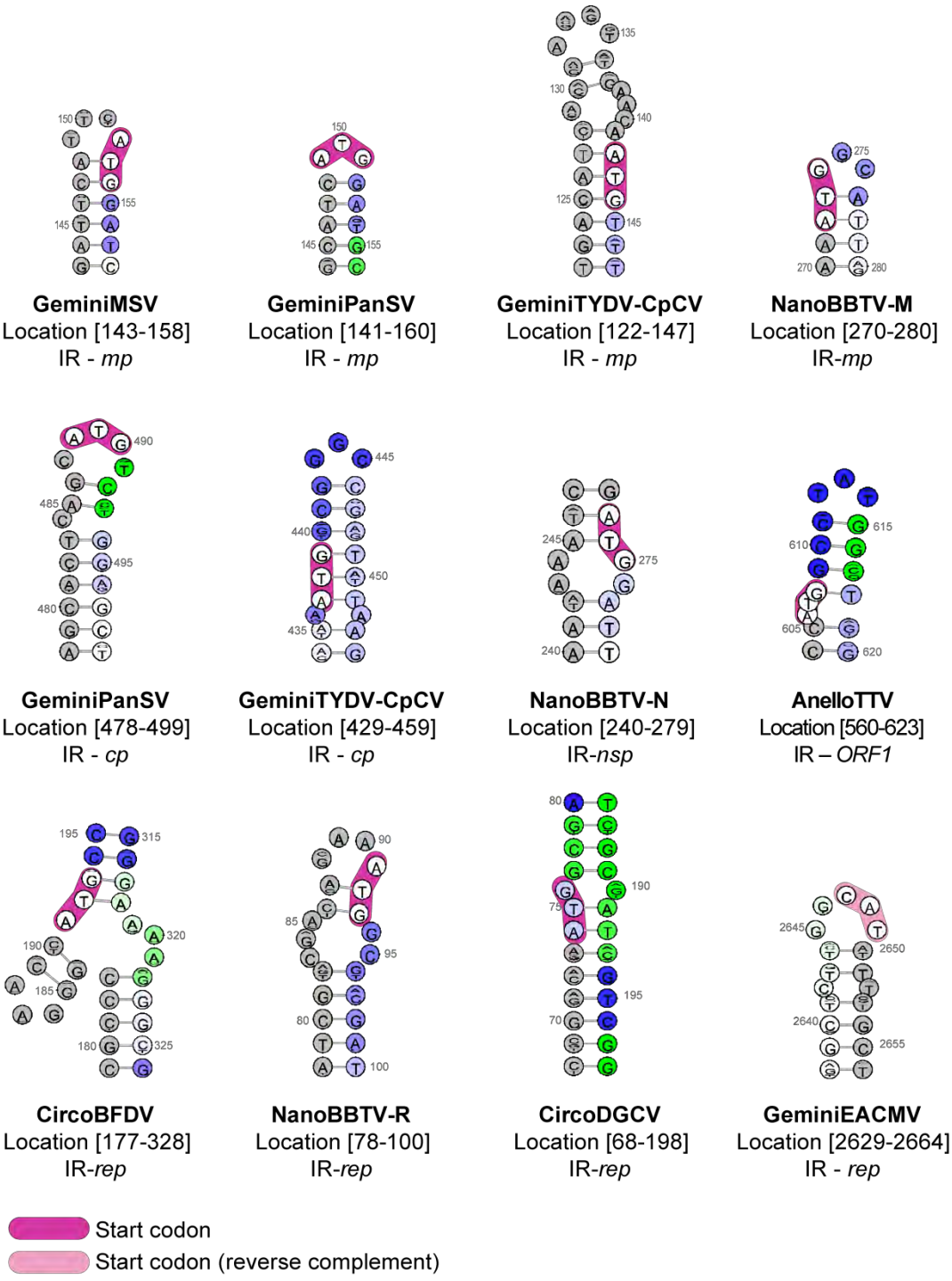
Cons. Rank[a]	Structure ID / NASP Rank[b]	Coordinates Including Gaps				Syn. Sub. Rank[c]	Coev. Rank[d]	Min. Rank[e]	Biological function	Reference
		Left hand side		Right hand side						
5	5	2400	2412	2486	2498	86	53	5		
6	6	28	47	169	188	NA	NA	6		
7	7	1650	1665	1756	1771	61	30	7		
8	11	1034	1039	1044	1049	7	NA	7		
9	8	2629	2636	2657	2664	47	NA	8		
10	9	994	1000	1004	1010	75	NA	9		
11	10	1820	1827	1852	1859	102	NA	10		
12	12	1925	1932	1937	1944	69	NA	12		
13	13	1289	1295	1301	1307	NA	65	13		
14	ID 157 / rank 14	1053	1055	1205	1207	NA	NA	14		Figure 4-1., G2 and Figure 4-4., GeminiEACMV
15	15	202	211	216	225	NA	NA	15		
GeminiMVYVY (genome length = 2806)										
1	1	3	14	2786	2797	NA	NA	1	Initiation of genome replication	Orozco, B.M. and Hanley-Bowdoin, L., (1996)
2	10	1992	1996	2000	2004	1	NA	1		
3	2	2033	2041	2045	2053	4	NA	2		
4	3	1896	1905	1910	1919	37	31	3		
5	4	2084	2092	2096	2104	20	19	4		
6	5	1190	1197	1258	1265	NA	36	5		
7	6	1576	1585	1598	1607	NA	NA	6		
8	7	1751	1756	1779	1784	28	45	7		
9	8	1838	1851	1858	1871	31	NA	8		
10	9	2334	2339	2344	2349	NA	NA	9		
11	11	1355	1364	1370	1379	NA	53	11		
12	12	1276	1286	1312	1322	NA	NA	12		
13	ID 96 / rank 13	989	999	1120	1130	18	NA	13		Figure 4-1., G2 and Figure 4-4., GeminiMYVYV
14	14	2640	2646	2694	2700	NA	14	14		
15	15	1562	1567	1623	1628	NA	NA	15		
16	16	341	347	365	371	NA	NA	16		

[a]Consensus ranking in which each structure gets the minimum rank from the NASP, Syn. Sub. and Coev. rankings.
[b]NASP ranking is based on base-pairing conservation scores. This rank is also used as structure ID.
[c] Syn. Sub ranking is a ranking based on synonymous substitution rates. NA indicates structures falling outside coding regions.
[d]Coev. ranking is a ranking based on complementarily coevolution. NA indicates structures which had invariable sites.
[e]Minimum rank from NASP, Syn. Sub. and Coev. ranks.



Supplementary Figure 1: Pairwise identity-based partitioning datasets

(**A**) A pairwise identity matrix from the input dataset of 30 reference mastrevirus genomes. The input dataset was portioned using the minimum of 78 and maximum of 100 identity thresholds. (**B**), (**C**), (**D**), and (**E**) are pairwise identity matrices for the resulting small datasets of *Chickpea chlorotic dwarf virus* (CpCDV; 5 sequences), *Wheat dwarf virus* (WDV; 5 sequences), *Panicum streak virus* (PanSV; 9 sequences) and *Maize streak virus* (MSV; 11 sequences) respectively.



Supplementary Figure 2: Other ssDNA virus genomic secondary structures spanning the start of genes

Some examples of structures spanning start codons in geminiviruses, nanoviruses anelloviruses and circoviruses are shown. In all these structures the start codon is consistently positioned within or immediately adjacent to a loop or a bulge, which potentially enhances the accessibility or recognition of these sites during translation initiation.

Supplementary Dataset 1: Full genome sequences used to compare SDT to other methods

Available at:

http://web.cbio.uct.ac.za/~brejnev/downloads/Supplementary_Dataset1.fas

Supplementary Dataset 2: Full genome sequences used to assess the speed gained with parallelisation of SDT

Available at:

http://web.cbio.uct.ac.za/~brejnev/downloads/ Supplementary_Dataset2.fas

Appendix

Author's publications associated with the thesis

1. Muhire, B. M., Golden, M., Murrell, B., Lefeuvre, P., Lett, J.-M., Gray, A., Poon, A.Y.F., Ngandu, N.K., Semegni, Y., Tanov, E.P., Monjane, A.L., Harkins, W.H., Varsani, A., Shepherd, D.N. & Martin, D. P. (2014). Evidence of pervasive biologically functional secondary-structures within the genomes of eukaryotic single-stranded DNA viruses. Journal of virology, 88 (4), 1972–1989.
2. Muhire, B.M., Varsani, A., Martin, D.P., (2014). SDT: a virus classification tool based on pairwise sequence alignment and identity calculation (PLoS One).
3. Muhire, B.M., Martin, D. P., Brown, J. K., Navas-Castillo, J., Moriones, E., Zerbini, F. M., Rivera-Bustamante, R.,Malathi, V.G., Briddon, R.W. & Varsani, A. (2013). A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). Archives of virology, 158(6), 1411–24.
4. Golden, M., Muhire, B.M., Semegni, Y., Martin, D. (2014). Patterns of recombination in HIV-1M are influenced by selection disfavouring the survival of recombinants with disrupted genomic RNA and protein structures (PLoS One).
5. Cloete, L., Tanov, E., Muhire, B.M., Martin, D.P., Harkins, G.W. (2014). Bayesian Coalescent Inference of Rubella Virus Nucleotide Substitution Rates. (Journal of General Virology).
6. Martin, D.P., Murrell, B., Golden, M., Khoosal, A., Muhire, B. (2015). RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evolution 1 (1), vev003
7. Kraberger, S., Kumari, SG., Hamed, AA., Gronenborn, B., Thomas, JE., Sharman, M., Harkins, GW., Muhire, BM., Martin, DP., Varsani, A. (2015). Molecular diversity of *Chickpea chlorotic dwarf virus* in Sudan: High rates of intra-species recombination a driving force in the emergence of new strains. Infect Genet Evol. doi: 10.1016/j.meegid.2014.11.024
8. Stenzel, T., Piasecki, T., Chrząstek, K., Julian, L., Muhire, B.M., Golden, M., Martin, D., Varsani, A. (2014). Genomes of Pigeon circoviruses display patterns

of recombination, genomic secondary structure and selection similar to those of Beak and feather disease viruses. Journal of General Virology, (95), 1338–1351.

9. Julian, L., Piasecki, T., Chrz, K., Walters, M., Muhire, B.M., Harkins, G.W., Martin, D.P. & Varsani, A. (2013). Extensive recombination detected amongst Beak and feather disease virus isolates from breeding facilities in Poland. Journal of General Virology, (94), 1086–1095.
10. Monjane, A., Martin, D.P., Lakay, F., Muhire, B.M., Pande, D., Varsani, A., Harkins, G.W., Shepherd, D.N., Rybicki, E. (2014). Adaptive Recombination Between Synthetic Chimeric Viruses Yields Large Fitness Gains. Journal of Virology, doi:10.1128/JVI.00709-14.
11. Candresse, T., Filloux, D., Muhire, B.M., Julian, C., Galzi, S., Fort, G., Bernardo, P., Daugrois, JH, Fernandez, E., Martin, D.P., Varsani, A., Roumagnac, P. (2014). Appearances can be deceptive: Revealing a hidden viral infection with deep sequencing in a plant quarantine context. (Plos One).