JVI Journals.ASM.org

Evidence of Pervasive Biologically Functional Secondary Structures within the Genomes of Eukaryotic Single-Stranded DNA Viruses

Brejnev Muhizi Muhire,[a] Michael Golden,[a] Ben Murrell,[b] Pierre Lefeuvre,[c] Jean-Michel Lett,[c] Alistair Gray,[d] Art Y. F. Poon,[e,f] Nobubelo Kwanele Ngandu,[g] Yves Semegni,[h] Emil Pavlov Tanov,[i] Adérito Luis Monjane,[a,d] Gordon William Harkins,[i] Arvind Varsani,[j,k,l] Dionne Natalie Shepherd,[d] Darren Patrick Martin[a]

Institute of Infectious Diseases and Molecular Medicine, Computational Biology Group, University of Cape Town, Cape Town, South Africa[a]; Department of Medicine, University of California, San Diego, San Diego, California, USA[b]; CIRAD, UMR PVBMT CIRAD—Université de la Réunion, Pôle de Protection des Plantes, Saint-Pierre, La Réunion, France[c]; Department of Molecular and Cell Biology, University of Cape Town, Rondebosch, Cape Town, South Africa[d]; BC Centre for Excellence in HIV/AIDA, Vancouver, Canada[e]; Department of Medicine, University of British Columbia, Vancouver, Canada[f]; Institute of Infectious Diseases and Molecular Medicine, Division of Medical Virology, University of Cape Town, Cape Town, South Africa[g]; Department of Mathematics and Physics, Cape Peninsula University of Technology, Cape Town, South Africa[h]; South African National Bioinformatics Institute, University of the Western Cape, Cape Town, South Africa[i]; School of Biological Sciences, University of Canterbury, Christchurch, New Zealand[j]; Biomolecular Interaction Centre, University of Canterbury, Christchurch, New Zealand[k]; Electron Microscope Unit, Division of Medical Biochemistry, Department of Clinical Laboratory Sciences, University of Cape Town, Rondebosch, Cape Town, South Africa[l]

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 analyzed these for evidence of natural selection favoring 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 analyzed here.

Besides encoding structural, regulatory, and enzymatic proteins, the nucleotide sequences of viral genomes encode a wide range of regulatory motifs associated with, among other things, transcription (1, 2), translation (3), replication (4), and genome packaging (5). 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 tRNA, rRNA, mRNA, 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 (6–12), the regulation of gene expression (13), the control of transcription (14), translation (15–17), and gene splicing (18), and the modulation of host antiviral responses (19–21).

Within viral genomes, biologically important structural elements tend to be highly conserved across even distantly related species (22). In ssRNA viruses, for example, they include the Rev response element (RRE) in human and simian immunodeficiency viruses (23–25), the *cis*-acting replication elements (CREs) of flaviviruses (26), luteoviruses (27), carmoviruses (11), coronaviruses (28), alphaflexiviruses (29), reoviruses (30), and picornaviruses (31, 32), the internal ribosomal entry site (IRES) elements of flaviviruses (33, 34), picornaviruses (35), pestiviruses (36) and dicistroviruses (37), and the cap-independent translation elements (CITEs) found in many plant-infecting ssRNA viruses (38, 39).

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 (9, 40, 41) and optimal gene expression (42–46), in geminiviruses, nanoviruses, and circoviruses highly conserved stem-loop structures at their replication origins are essential for the initiation and termination of replication (6, 47–49). Besides these few examples, however, it

Received 23 October 2013 Accepted 25 November 2013
Published ahead of print 27 November 2013
Address correspondence to Darren Patrick Martin, darrenpatrickmartin@gmail.com.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/JVI.03031-13.

doi:10.1128/JVI.03031-13

is currently unknown how pervasive biologically important secondary-structural elements are within the genomes of these viruses (50–52).

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 (53), 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 favoring the maintenance of these structures. Whereas in coding regions these signals might include codon usage biases (54, 55) and decreased rates of synonymous substitution (56), throughout the genome these signals could also include high rates of reversion substitution (51, 57) and increased frequencies of complementarily coevolving nucleotide pairs, particularly among those nucleotides predicted to be base paired within secondary-structural elements (58–60).

Accordingly, experimental investigations of individual structural elements within some ssDNA virus genomes have clearly demonstrated the existence of strong natural selection favoring 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 (61, 62). 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 restabilize the structural element (51).

Here, we examine the biological relevance of pervasive computationally predicted secondary structures within diverse eukaryote-infecting ssDNA virus genomes. After using a free-energy minimization 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 eukaryotic ssDNA virus genomes, we further identify what are likely some of the most functionally important uncharacterized structural elements within these genomes.

MATERIALS AND METHODS

Data set preparation. All available circovirus, anellovirus, parvovirus, nanovirus, and geminivirus full-genome sequences were obtained from GenBank (http://www.ncbi.nlm.nih.gov/genbank/) between April 2011 and August 2012. Full-genome sequences for each of the five families were preliminarily aligned separately using MUSCLE (63) implemented in MEGA5 (64) and subdivided into data sets of sequences sharing at least 75% sequence identity. This was done to ensure reasonable alignment accuracy during subsequent sequence analyses (65) while at the same time providing enough sequence diversity to enable the accurate characterization of evolutionary processes acting to maintain predicted secondary structures.

A set of 23 data sets was obtained, each containing between 21 and 519 full-genome sequences. Each of these full-genome sequence data sets was realigned using MUSCLE (with default settings) and, where necessary, manually edited. The resulting alignments will here be referred to as large data sets (Table 1 identifies the data sets and explains the naming system). This distinction is important because many of the analyses performed

TABLE 1 List of the 23 large data sets obtained

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

[a] The name of the data set 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 data set.

could be carried out only on subsets of these data sets. Specifically, from each of the large data sets we first extracted an intermediate data set. In all but four cases each contained one representative sequence from each of the 30 most divergent viral sequence lineages in the large data sets. The exceptional cases were the AnelloTTSuV1, AnelloTTV, ParvoHBoV, and ParvoMPV data sets that, respectively, contained only 21, 22, 21, and 26 sequences, all of which were included in the intermediate data set. From each of the intermediate data sets we further extracted a small data set containing one representative sequence of each of the 10 most divergent lineages.

Detection of conserved secondary-structural elements within 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) (66) to identify the conserved secondary-structural elements within the set of representative full genomic sequences in each of the small data sets.

NASP takes as input a set of aligned nucleic acid sequences and uses Gibbs free energy (67) and Boltzmann probability (68) techniques implemented in the hybrid-ss component of the UNAFold software package (69) 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 overall thermodynamic stability.

Precisely, NASP produces sets of pairing matrices for each of the input sequences in each small data set, 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 (66). The use by NASP of weighted sums in the calculation of its M Matrix is intended to counteract unavoidable sampling biases in sequence data sets 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 data sets and indicated the subsets of structural elements referred to as high-confidence structure sets (HCSSs) that accounted for the analyzed sequences having significantly lower MFE scores than those expected in randomized sequences with identical base compositions (66). 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 data sets) or circular (for all 20 of the other small data sets) ssDNA at either 37°C (for animal-infecting circoviruses, anelloviruses, and parvoviruses) or 25°C (for plant-infecting geminiviruses and nanoviruses) under 0 M magnesium and 1 M 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 data set 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.

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 disfavoring 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 (70) and Fu and Li (71). We calculated Tajima's *D* and Fu and Li's *F* statistics for the paired and unpaired sites in each of the 23 large data sets.

Since in all 23 of the analyzed large data sets there were invariably fewer paired sites (i.e., those paired within the HCSS) than unpaired sites (i.e., the remainder of the sites in the various data sets), 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 data sets.

From each large data set we produced 100 data sets, each consisting of sites (i.e., entire large data set alignment columns) randomly sampled with replacement from the pool of unpaired sites. These permutation data sets contained the same numbers of sites as their corresponding paired-site data sets. Tajima's *D* and Fu and Li's *F* statistics were then calculated for all of the paired-site and permutation data sets. For each of the 23 data sets, 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 data set were lower than those calculated for the 100 permuted data sets.

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 favors the preservation of amino acid sequences (i.e., selection disfavoring nonsynonymous substitutions), and selection at the nucleotide level, which favors 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 data set alignments. Within each of the resulting gene data sets, all sites encoding amino acids in two or more different frames were removed. Following this, 43 gene data sets containing 200 or more nucleotide sites were retained for further analysis (see Table S1 in the supplemental material for details of these data sets). Gene data sets excluded from this set because they retained too few sites included the following: the *ren*, *mp*, and *trap* genes of the GeminiTYLCV, GeminiEACMV, and GeminiMYVYV data sets; ORF2 and ORF3 of the AnelloTTSuV1, AnelloTTSuV2, and AnelloTTV data sets; the *vp1* and *vp2* of the ParvoHBoV data set; and the *vp1* of the ParvoMPV.

Two methods were used to estimate synonymous substitution rates at individual codon sites within the 43 gene data sets: partitioning approach for inference of selection (PARRIS) (72) and fast unconstrained Bayesian approximation (FUBAR) (73). Both of these methods apply the time-reversible MG94 codon substitution model which utilizes a 61-by-61 codon substitution matrix (74), and both allow independent distributions for synonymous and nonsynonymous 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 analyzed. 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) (75), it was likely that many of the sequences being analyzed here were recombinant (76–80). It was therefore necessary to take steps to explicitly account for recombination within these analyses. Accordingly, prior to selection anal-

yses the genetic algorithm for recombination detection (GARD) (81) method was used to detect recombination breakpoint sites. These sites were then used to partition the alignment into "mostly recombination-free" subalignments (it is unlikely that every recombination event was detected and accounted for). For each of these subalignments a phylogenetic tree was inferred, and the 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 (72, 73).

Within each gene data set, each codon was categorized as being a paired codon if its third-position nucleotide was paired within the relevant HCSS and as 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 data sets 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.

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 (51). We detected evidence of complementary coevolution between pairs of sites within the large data sets using a customized version of the SpiderMonkey coevolution script written in HYPHY (82). For any chosen pair of sites within a large data set, the script compares the standard independent sites of a 4-by-4 HKY85 nucleotide substitution model (83) to a 16-by-16 Muse-modified HKY85 coevolution model (called M95) (84) to determine which of these best describes the evolution of individual site pairs. In our case, entries in the M95 16-by-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/λ. A maximum-likelihood (ML) ratio test enabled us to determine whether nucleotide pairs were coevolving. Whereas a λ of >1 for particular coevolving site pairs indicated that they displayed a tendency toward complementary coevolution, a λ of 1 indicated a tendency toward site pairs evolving independently, and λ of <1 indicated a tendency toward their coevolving noncomplementarily. We identified site pairs displaying strong evidence of complementary coevolution as those with both associated maximum-likelihood estimates of a λ of >1 and Muse 95 versus HKY85 likelihood ratio test *P* values of <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 (nt) of one another and (ii) pairs of nucleotides that were polymorphic in the input data set. Also, since recombination can undermine the accuracy with which the phylogenetic trees used to detect coevolution reflect the actual evolutionary relationships of the analyzed sequences, we took steps to account for recombination in our analyses. Each large data set was analyzed for recombination using Recombination Detection Program (RDP), version 4.13 (85), 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 a 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 (86). Each of these alignments and their corresponding phylogenetic trees were used as inputs for our complementary coevolution analysis.

Customized computational tools. All computer scripts used in all the analyses conducted can be downloaded from the University of Cape Town Computational Biology website (http://web.cbio.uct.ac.za/~brejnev/downloads/Scripts.zip) and a customized computer program and data sets for visualization 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).

RESULTS AND DISCUSSION

Numerous evolutionarily conserved secondary structures are evident within eukaryotic ssDNA virus genomes. We assembled 23 full-genome sequence data sets representing the families *Circoviridae*, *Anelloviridae*, *Parvoviridae*, *Nanoviridae*, and *Geminiviridae* (Table 1). In each of these sets, between 69 and 316 conserved secondary-structural elements were identified using the minimum free-energy (MFE) approach implemented in the computer program NASP (66). From these lists of conserved structural elements, NASP identified subsets of between 5 and 132 high-confidence structural elements. These lists, here referred to as high-confidence structure set (HCSS) lists, contained those structures primarily responsible for the analyzed genomes having greater degrees of predicted structural stability than those of randomized 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 data sets. 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 (Fig. 1 and 2; see also Table S2 in the supplemental material).

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 uncharacterized conserved structures may have been preserved during evolution due to their biological importance.

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 (87, 88), 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 the HCSSs.

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

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)

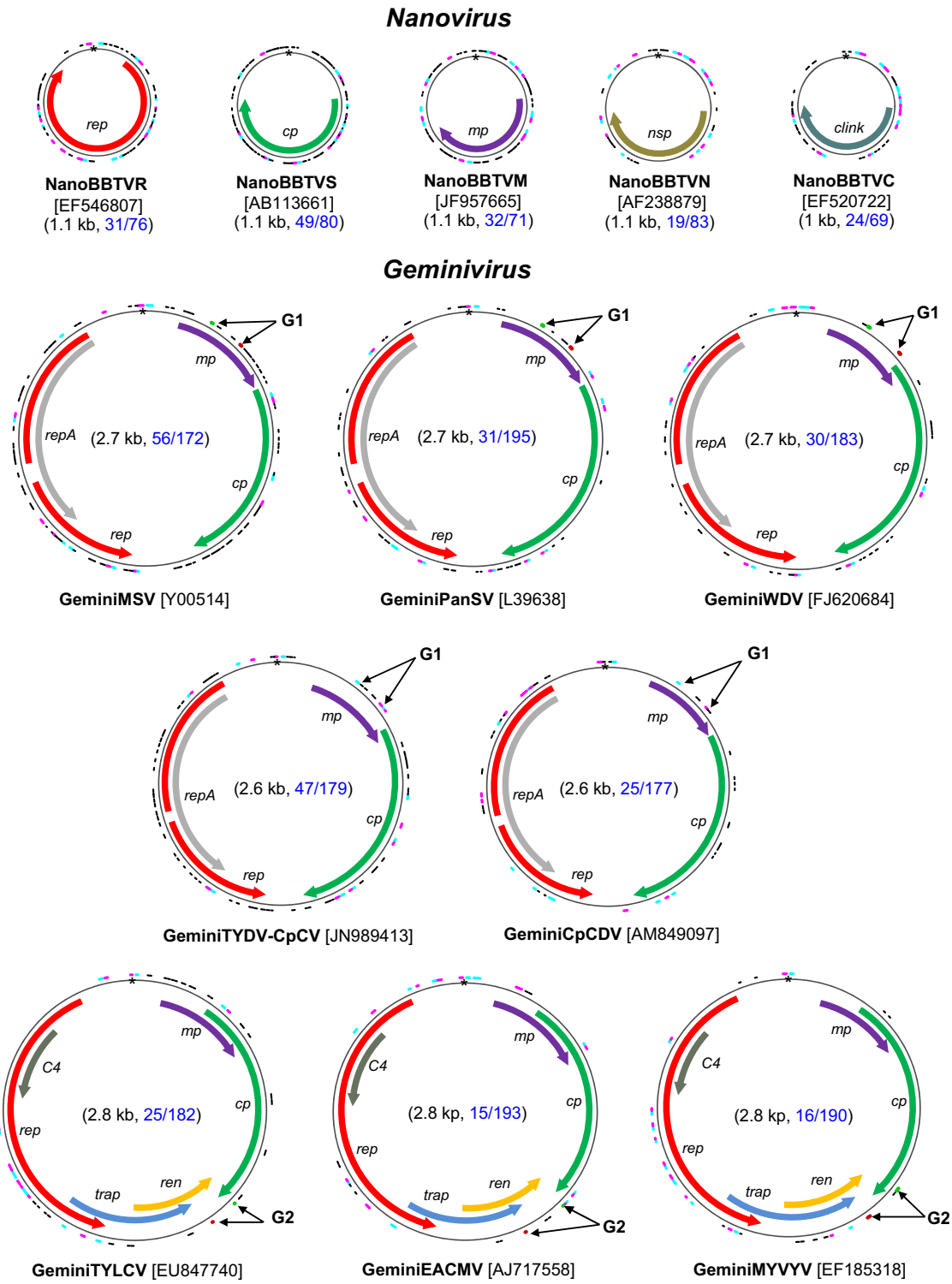

FIG 1 Secondary-structure map of plant-infecting ssDNA viruses; genome organization maps of plant-infecting ssDNA viruses. In each map, the arcs represent the coordinates of identified structural elements within high-confidence structure sets (HCSSs). These highly conserved structural elements are those primarily responsible for the estimated structural stability of the analyzed genomes being greater than that of randomized sequences with identical nucleotide compositions. The 10 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 uncharacterized but likely biologically functional structures that are apparently conserved across multiple data sets (shown in green and brown when these were not ranked among the top 10 within their respective HCSSs). 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 are as follows: *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.

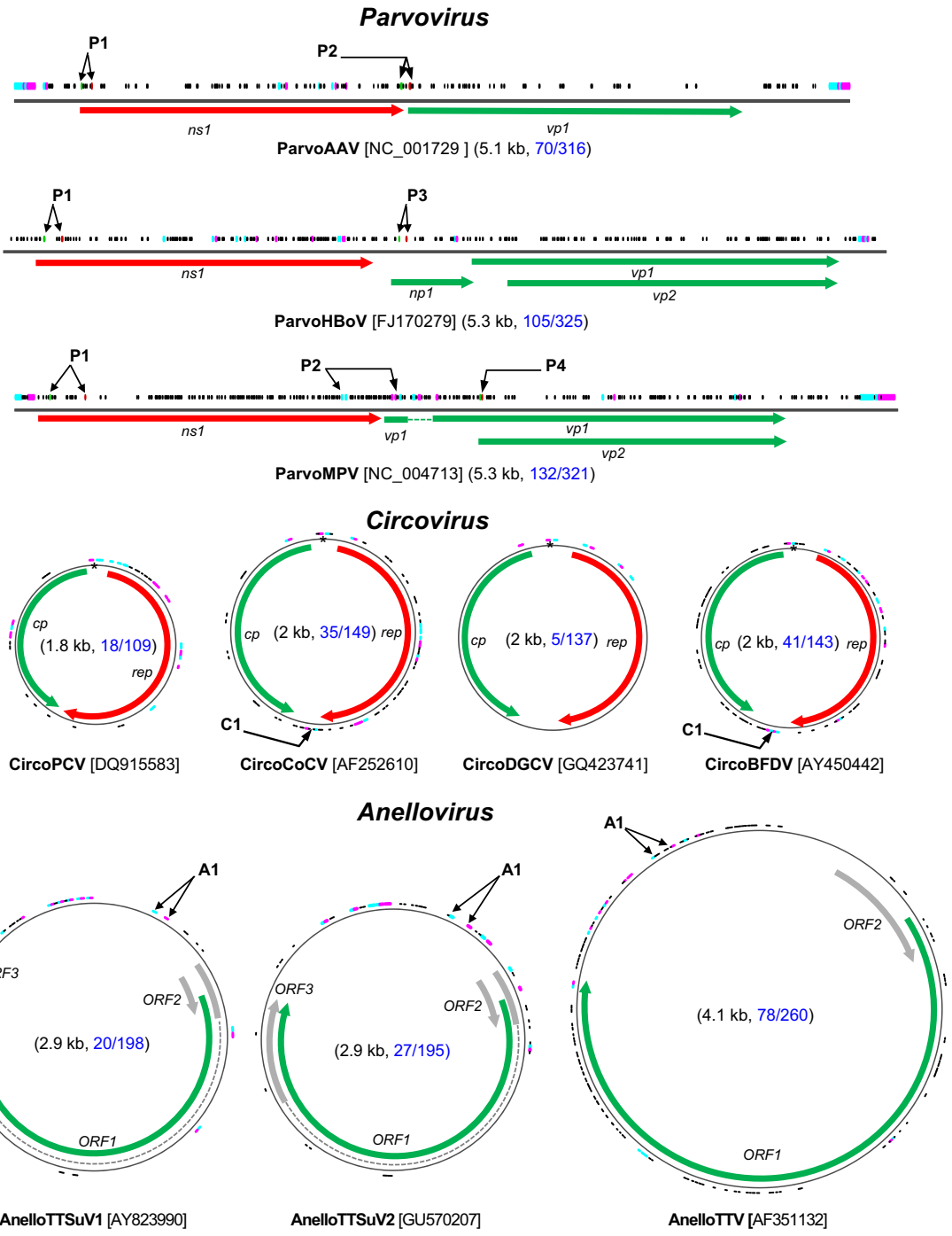

FIG 2 Secondary-structure map of animal-infecting ssDNA viruses; genome organization maps of animal-infecting ssDNA viruses. In each map, the arcs (for circular genomes) and vertical lines (for linear genomes) represent the coordinates of identified structural elements within high-confidence structure sets (HCSSs). The 10 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 uncharacterized but likely biologically functional structures that are apparently conserved across multiple data sets (shown in green and brown when these were not ranked among the top 10 structures within their respective HCSSs). 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 nonstructural protein; *np1*, small nonstructural protein; *vp1*, major virion/viral protein; *vp2*, minor virion/viral protein, and *ORF*, unnamed open reading frame.

TABLE 2 Tajima's *D* and Fu and Li *F* statistics for paired and unpaired genomic site alignments

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

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

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 (70, 71). 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 data sets containing only base-paired sites (constructed from the large data set alignments by removing all unpaired nucleotide sites) than for data sets containing only unpaired sites (constructed from the large data set alignments by removing all paired nucleotide sites).

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

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 analyzed genomes.

However, to test whether values of these statistics were significantly lower at paired sites than at unpaired sites in the 21/23 large data sets displaying the expected trend, for each data set we applied a permutation test involving resampling of identical numbers of sites from the unpaired-site data set as were present within the paired-site data set (in each data set unpaired sites were invariably more numerous than the paired sites). In each case a *P* value was computed as the proportion of the 100 permutated unpaired-site data sets that yielded lower *D* or *F* values than the corresponding paired-site data set. In this test, if the *P* value is <0.05, a *D* or *F* value for an unpaired-site data set that was lower than that of its corresponding paired-site data set would be expected less than 5% of the time if the null model of neutral evolution was true.

In 11/23 data sets both the *D* and *F* statistic permutation tests yielded evidence that paired sites within the HCSS lists experience significantly stronger (*P* values of <0.05) purifying selection than the remainder of the genomic sites. In a further 6/23 cases, either the *D* or *F* statistic test yielded at least marginal evidence (*P* values of <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 analyzed anellovirus data sets were among the six data sets with no evidence of purifying selection acting on paired sites. It is perhaps also noteworthy that of the 11 data sets displaying strong evidence of base pairing associated with purifying selection, only two (both of them circoviruses, CircoCoCV and CircoBFDV) were from the 10 mammal- and bird-infecting virus data sets. While it is not possible to directly compare the plant- and animal-infecting virus data sets 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 data sets were, in general, substantially smaller than our plant virus data sets and that our analysis therefore simply lacked sufficient power to differentiate between the numbers of low-frequency polymorphisms within the paired- and unpaired-data set 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.

Synonymous substitution rates are unusually low at paired genomic sites. We hypothesized that selection favoring 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 data sets (see Table S1 in the supplemental material) were inferred using the random-effects-likelihood selection analysis methods PARRIS (75) and FUBAR (73). These methods indicated that in 27/43 of these data sets, the median substitution rates of codons with paired third-position nucleotides were signif-

TABLE 3 Comparison of synonymous substitution rates at paired- and unpaired-codon sites

Data set	Gene(s) studied	No. of sequences[c]	PARRIS[a]	FUBAR[b]
CircoPCV	*rep*, *cp*	30, 29	*rep*	*rep*
CircoCoCV	*rep*, *cp*	30, 30	*rep*	*rep*
CircoDGCV	*rep*, *cp*	30, 30	*rep*	*rep*
CircoBFDV	*rep*, *cp*	30, 29	*rep*, *cp*	*rep*, *cp*
AnelloTTSuV1	ORF1	17	ORF1	ORF1
AnelloTTSuV2	ORF1	30	ORF1	ORF1
AnelloTTV	ORF1	21	ORF1	ORF1
ParvoAAV	*ns1*, *vp1*	23, 30	*ns1*, *vp1*	*ns1*, *vp1*
ParvoHBoV	*ns1*, *np1*	21, 21	*ns1*	*ns1*
ParvoMPV	*ns1*, *vp2*	25, 18	*ns1*	*ns1*, *vp2*
NanoBBTV-R	*rep*	28	*rep*	*rep*
NanoBBTV-S	*cp*	29	*cp*	*cp*
NanoBBTV-M	*mp*	27	*mp*	*mp*
NanoBBTV-N	*nsp*	27		
NanoBBTV-C	*clink*	30		
GeminiMSV	*rep*, *cp*, *mp*	30, 30, 30	*rep*, *cp*	*rep*, *cp*, *mp*
GeminiWDV	*rep*, *cp*, *mp*	30, 30, 30	*cp*, *mp*	*cp*, *mp*
GeminiPanSV	*rep*, *cp*, *mp*	30, 30, 30		
GeminiTYDV-CpCV	*rep*, *cp*, *mp*	30, 30, 30	*cp*	*cp*
GeminiCpCDV	*rep*, *cp*, *mp*	30, 30, 30	*rep*, *cp*, *mp*	*rep*, *cp*, *mp*
GeminiTYLCV	*rep*, *cp*	27, 30		
GeminiEACMV	*rep*, *cp*	30, 30	*rep*	*rep*
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, $P < 0.05$) 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, $P < 0.05$) lower than those at unpaired-codon sites.
[c] Values are given in respective order for the genes studied.

icantly lower than those of codons with unpaired third-position nucleotides (multiple comparison-corrected Mann-Whitney U-test, P value of <0.05). An additional five data sets yielded similar evidence but only with one of the two selection analysis methods (Table 3).

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 also that these structures likely 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 (89) simply because base-paired nucleotides might be predisposed to lower mutation rates than their unpaired counterparts (90, 91).

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 data set were used to infect maize plants (92). 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 Gemini-TYLCV data set) were used to infect tomato plants (52).

While over 101 days postinfection 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 data sets, 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 exact test (93), 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 (939/2,641, or ~36% of considered sites) within the HCSS, yielding significant evidence (P value of 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 of 0.021) for mutations to accumulate more frequently at unpaired nucleotide sites.

Although no analogous experimental data are 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

TABLE 4 Association between paired sites and complementarily coevolving sites

Data set	Chi-square value	*P* value
CircoPCV	190.9307	4.20×10^{-14}
CircoCoCV	0.2272	0.14
CircoDGCV	143.2324	3.15×10^{-14}
CircoBFDV	62.5998	1.59×10^{-13}
ParvoAAV	185.5472	4.08×10^{-14}
ParvoHBoV	96.656	2.13×10^{-14}
ParvoMPV	137.077	3.02×10^{-14}
AnelloTTSuV1	117.9971	2.60×10^{-14}
AnelloTTSuV2	38.2243	2.41×10^{-08}
AnelloTTV	70.6212	1.55×10^{-14}
NanoBBTV-R	107.8986	2.37×10^{-14}
NanoBBTV-S	20.398	1.28×10^{-04}
NanoBBTV-M	49.9491	7.88×10^{-11}
NanoBBTV-N	48.2752	1.79×10^{-10}
NanoBBTV-C	21.1911	8.81×10^{-05}
GeminiMSV	212.2187	4.67×10^{-14}
GeminiWDV	89.9702	1.98×10^{-14}
GeminiPanSV	82.3437	1.81×10^{-14}
GeminiTYDV-CpCV	28.6975	2.43×10^{-06}
GeminiCpCDV	98.1122	2.16×10^{-14}
GeminiTYLCV	159.2665	3.50×10^{-14}
GeminiEACMV	175.6639	3.86×10^{-14}
GeminiMYVYV	364.9167	8.03×10^{-14}

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 favoring the maintenance of their biological functionality.

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 data sets. For each large data set, 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 data set, CircoCoCV, we found strong significant associations (multiple testing corrected *P* values of <0.0001) between paired sites within the HCSSs and sites for which complementary coevolution was detected (Table 4). It is noteworthy that the CircoCoCV was one of the two animal-infecting virus data sets displaying both strong evidence of base-pairing-associated negative selection and evidence of strong selection disfavoring 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 disfavoring any substitutions at these sites.

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 favoring 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.

Potentially important structural elements within eukaryotic 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 analyzed, 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 (94). Specifically, these rankings were based on the following: (i) the degree to which structural elements were conserved across the analyzed genomes, (ii) the degree to which synonymous substitution rates were constrained at codon sites containing nucleotides that are predicted to be base paired, and (iii) the degree to which nucleotides predicted to be base paired coevolve with one another. Rankings based collectively on these three criteria are here referred to as consensus rankings (see Table S2 in the supplemental material).

The 10 highest ranked structural elements based on these criteria within each of the 23 analyzed HCSSs are plotted in magenta and cyan in Fig. 1 and Fig. 2, and are listed in Table S2 in the supplemental material. 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 (95), 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 analyzed data sets. 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 among an ensemble (potentially very large) 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-characterized 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) (see Table S2 in the supplemental material).

Additional well-characterized structures detected include the replication-associated protein gene (*rep*) intron-associated structure (GeminiMSV; rank 16) (51), the parvovirus transcription attenuation stem-loop structures (ParvoMPV; ranks 17 and 34)

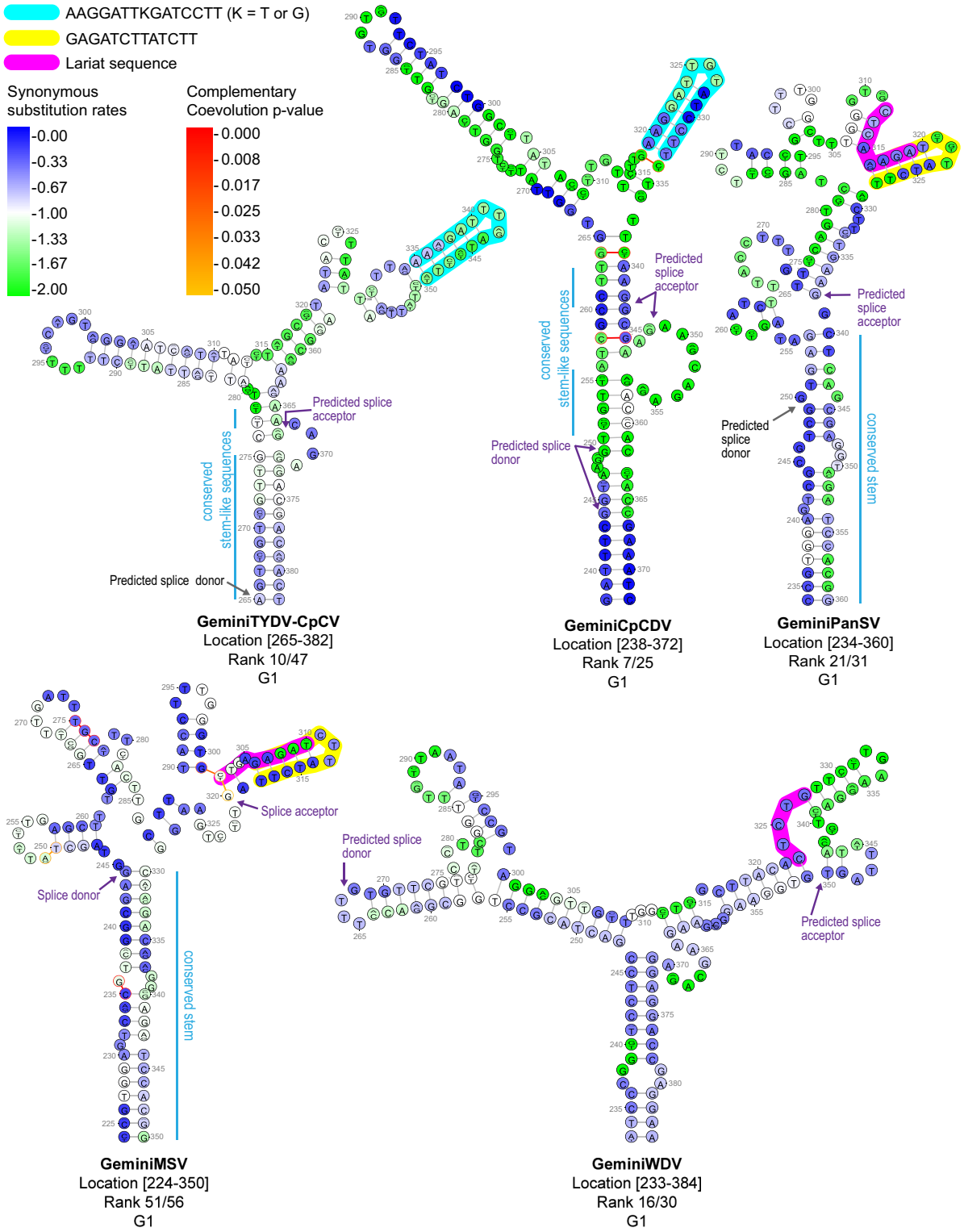

FIG 3 Secondary structure associated with the intron of the mastrevirus movement protein gene. A secondary-structure associated with the movement protein gene intron was predicted in all five mastrevirus data sets. This structure is highly conserved and contains splice donor and acceptor sites (indicated by arrows), as well as, in the case of the GeminiMSV, GeminiPanSV, and GeminiWDV, 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 the case of GeminiCpCDV, base-pairing interactions displaying significant associated complementary coevolution (*P* value of <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 color 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 data sets, 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 data sets, it is plausible that this structure has some function during splicing of the movement protein intron.

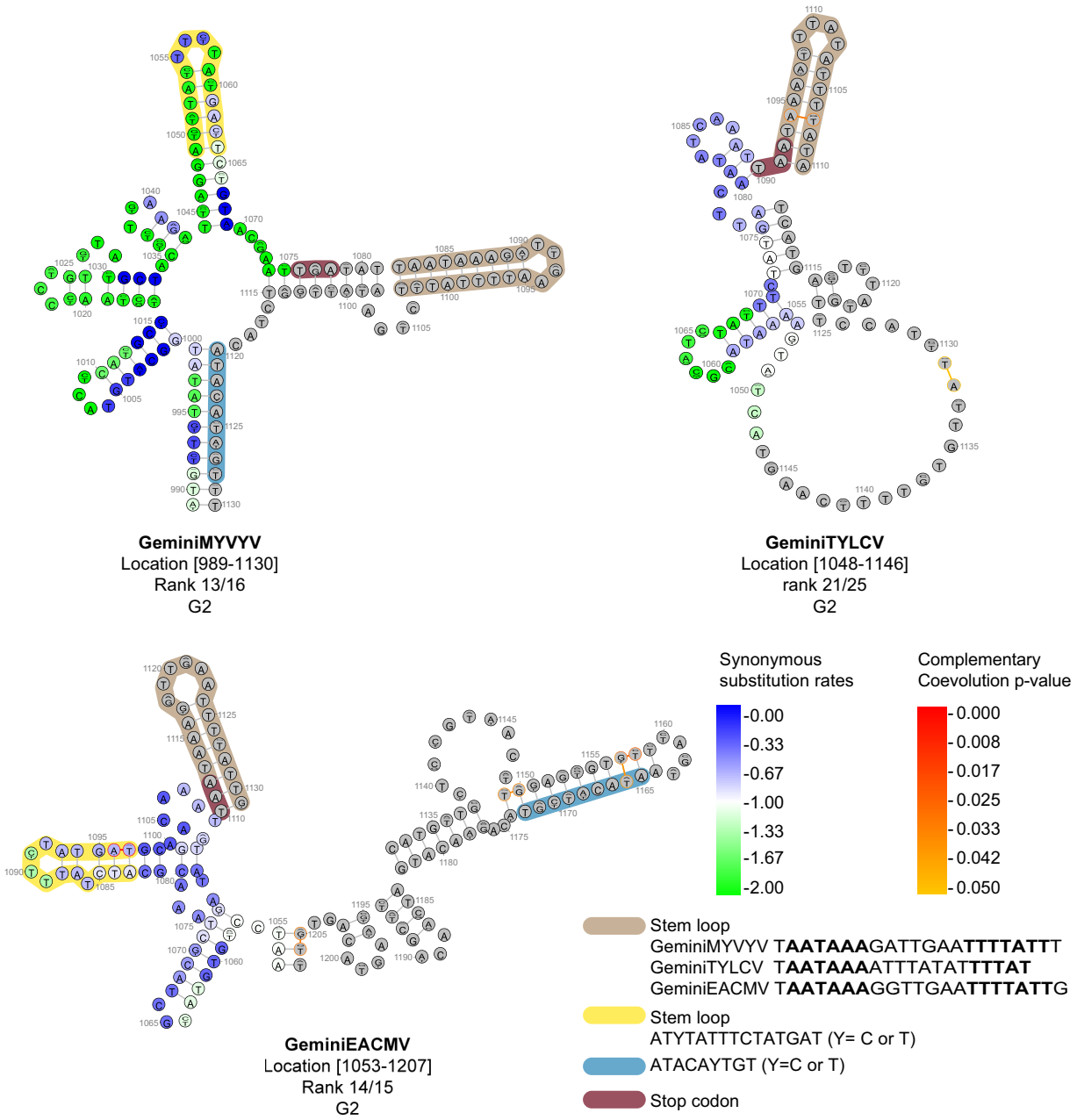

FIG 4 Secondary structure associated with the 3′ end of the begomovirus coat protein gene. A secondary structure with a potential role in transcriptional termination was predicted at the end of the coat protein gene of the begomovirus data sets, GeminiTYLCV, GeminiEACMV, and GeminiMYVYV. In all of these, the structure has a stop codon and a stem-loop containing a polyadenylation signal (the complementary polyadenylation signaling sequences within the stem-loops are in bold text). A common stem-loop structure between the GeminiEACMV and GeminiMYVYV data set is highlighted in yellow. Nucleotide logos and colors, respectively, indicate degrees of sequence variability and associated synonymous nucleotide substitution rates as outlined on Fig. 3. Nucleotides falling outside genes are shaded gray. Base-pairing interactions displaying significant associated complementary coevolution (P value of <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.

(Table S2) (42), and the 3′ complementary strand T-shaped structure that binds to the viral capsid in some parvoviruses (ParvoMPV; rank 3) (Table S2) (96).

Besides these well-known structures, we sought to identify other uncharacterized, 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 analyzed.

Geminivirus. We identified a particularly conserved 126- to 157-nt secondary structure within the movement protein (*mp*) gene of all five analyzed mastrevirus data sets (GeminiMSV,

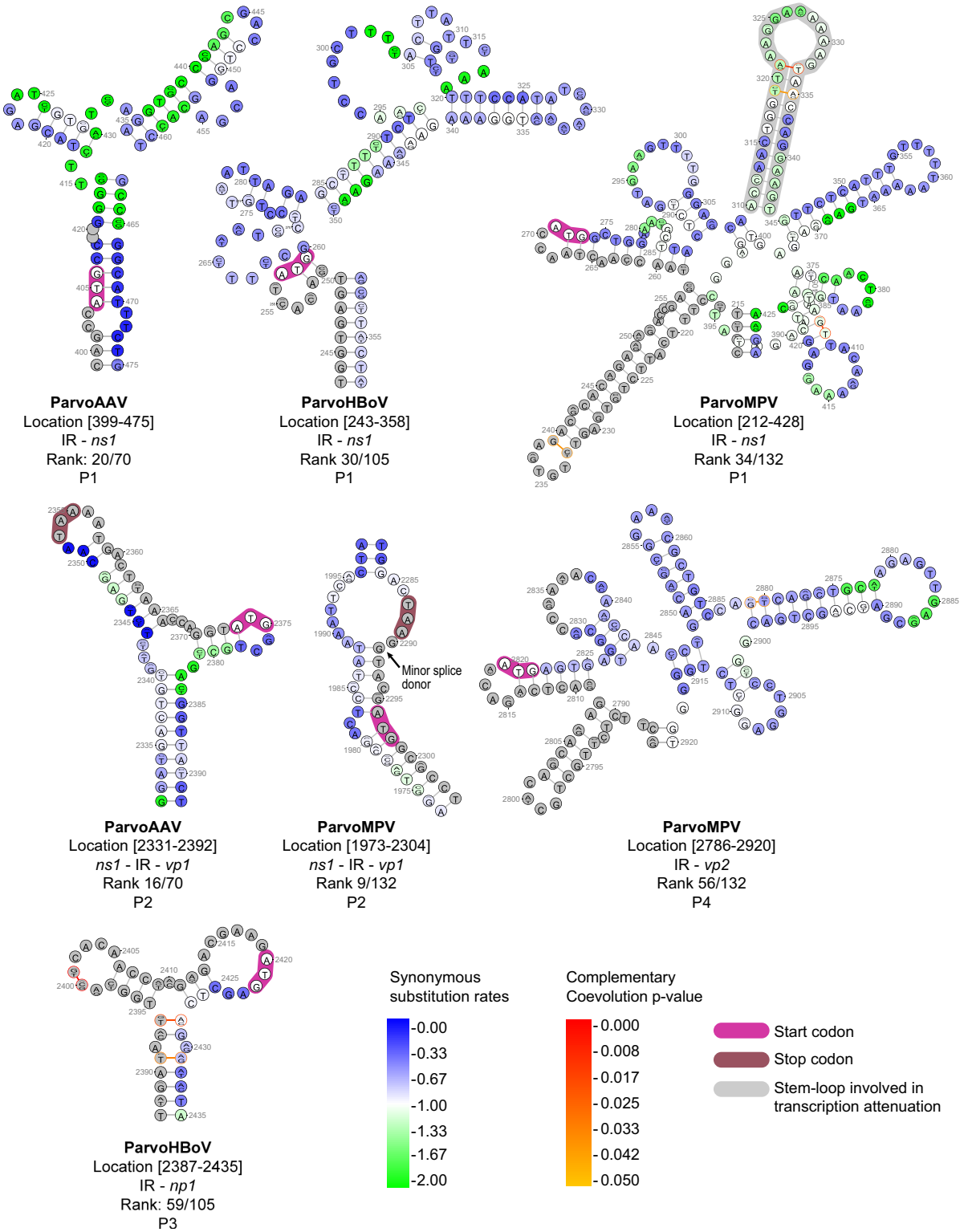

FIG 5 Parvovirus secondary structures predicted at the start of genes. Secondary structures predicted at the start of genes represented in the parvovirus data sets ParvoAAV, ParvoHBoV, and ParvoMPV are shown. These include those spanning the start of the large nonstructural proteins (*ns1*; P1), the major viral/virion proteins (*vp1*; P2), the small nonstructural protein (*np1*; P3) and the minor viral/virion proteins (*vp2*, P4). Nucleotide logos and colors, respectively, indicate degrees of sequence variability and associated synonymous nucleotide substitution rates as outlined on Fig. 3. Base-pairing interactions displaying significant associated complementary coevolution (*P* value of <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 gray. 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.

GeminiPanSV, GeminiWDV, GeminiTYDV-CpCV, and GeminiCpCDV) (structure G1 in Fig. 1 and 3). In all of these data sets 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 GeminiMSV). The structure in the GeminiMSV data set displayed a particularly high degree of conformational similarity with that in the GeminiPanSV data set, with the two structures sharing a nearly identical 21-nucleotide-long stem sequence (Fig. 3), indicating that they are almost certainly homologous. Although the sequences within this structure differ substantially between the other mastrevirus data sets, they all contain the splice donor, acceptor, and branch sites previously identified (or predicted) in mastrevirus *mp* introns (97) (Fig. 3), 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 (18, 98, 99).

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 data sets (structure G2 in Fig. 1 and 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 (Fig. 4). It is likely, therefore, that this structure may be functional either within ssDNA as a transcriptional terminator or within transcribed mRNA during polyadenylation.

Parvovirus. We identified a variety of uncharacterized parvovirus genomic and/or mRNA structural elements with potential functionality at the start of the large nonstructural (*ns1*) gene (structure P1 in Fig. 2 and 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 Fig. 2 and Fig. 5) (16th out of 70 HCSS structures in ParvoAAV and 9th out of 132 HCSS structures in ParvoMPV), the start of the small nonstructural (*np1*) gene (structure P3 in Fig. 2 and Fig. 5) (59th out of 105 HCSS structure in ParvoHBoV), and the start of the minor virion protein (*vp2*) gene (structure P4 in Fig. 2 and Fig. 5) (56th out of 132 HCSS structures in ParvoMPV). Although there were no sequence similarities shared between positionally analogous structures in the different parvovirus data sets, this was not unexpected, given that these data sets represent species within different genera (with sequences in different data sets sharing, on average, only 57.8% sequence identity). The ParvoMPV intergenic region (IR)-*ns1* structure contains a stem-loop identified to play role in transcription attenuation of parvovirus minute virus of mice (42) (structure P1 in Fig. 5). 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 (Fig. 5). This tendency was also noted in other data sets analyzed, and it is plausible that structures spanning the start codons of genes in these different families are functional either within partially single-stranded DNA during the initiation of transcription or in transcribed mRNA during the initiation of translation.

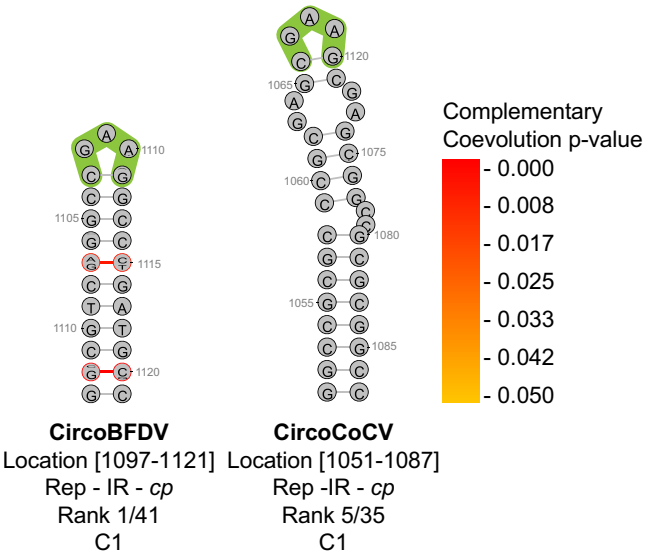

FIG 6 Conserved circovirus stem-loop structure within the intergenic region. A stem-loop structure that is highly conserved within the intergenic regions of the two circovirus data sets, CircoCoCV and CircoBFDV, is shown. Nucleotide logos reflect degrees of sequence variability. The stems of these structures 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.

Circovirus. While we were unable to identify any secondary structures that were clearly conserved across all five circovirus data sets analyzed, within the moderately divergent CircoCoCV and CircoBFDV data sets (these two data sets share, on average, 64% pairwise sequence identity), we identified an intergenic region (IR) stem-loop structure (structure C1 in Fig. 2 and 6), which is highly conserved in each of the respective data sets (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 data sets, in both data sets 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 posttranscriptional processing of mRNA transcripts.

Anellovirus. A conserved T-shaped structure was identified in the IRs of the two anellovirus data sets, AnelloTTSuV1 and AnelloTTuSV2 (structure A1 in Fig. 2 and 7) (7th out of 20 structures in the AnelloTTSuV1 HCSS and 1st out of 27 structures in the AnelloTTSuV2 HCSS). Even though these two data sets are moderately divergent (sequences within them share, on average, 60.7% pairwise identity), the structure is strikingly conserved between the two data sets. In both data sets 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 Fig. 7).

Given the high degree to which this structure has been conserved between these two moderately divergent anellovirus data sets, we attempted to identify a homologous structure within our

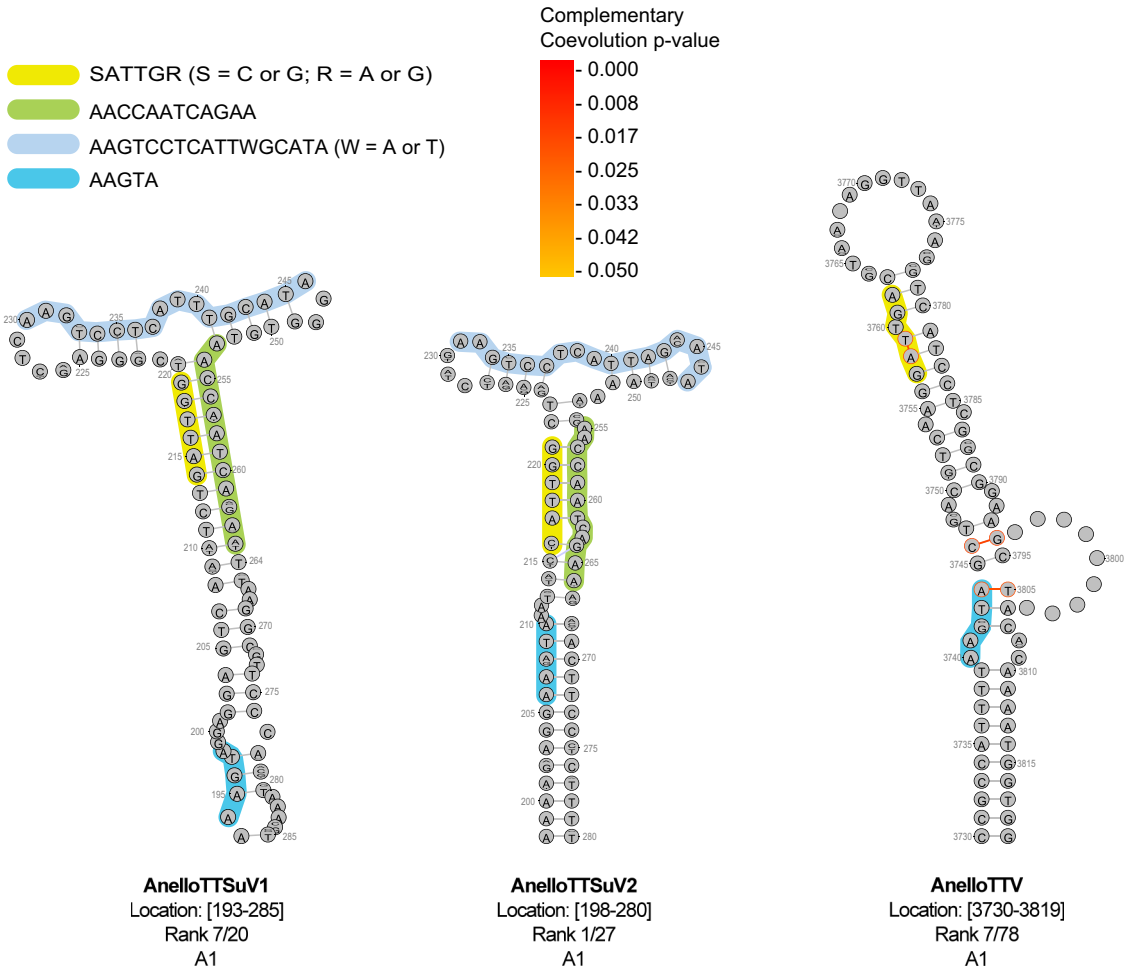

FIG 7 Anellovirus highly conserved intergenic T-shaped structures. A T-shaped structure predicted within the intergenic region (IR) of two divergent anellovirus data sets, AnelloTTSuV1 and AnelloTTSuV2, is shown. These structures have homologous 17-nucleotide-long sequences on top of the T (highlighted 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 data set 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 data sets. In the AnelloTTV structure, base-pairing interactions displaying significant associated complementary coevolution (*P* value of <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.

third highly divergent anellovirus data set (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 Fig. 2 and Fig. 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 nucleotides long) at similar positions within their stems (outlined in blue and yellow in Fig. 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 origin of replication (*ori*) sites reside. Given that the virion strand *ori* sites 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, characterization of replication-competent sub-full-length *Torque teno virus* (TTV) genomes (which are closely related to those represented in our AnelloTTV data set) 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, bellow the HCSS in our AnelloTTV data set [data not shown]) (100). 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.

Conclusion. Using computational methods we have identified numerous secondary structures that probably form at least transiently within eukaryotic ssDNA virus genomes and shown that a significant proportion of these predicted structures are likely bio-

TABLE 5 Summary of results

Data set	>5 NASP structures[a]	*dS* paired-codon sites < *dS* unpaired-codon sites[b]	Selection at paired sites[c]	Complementary coevolution[d]
CircoPCV	+	+	+	+
CircoCoCV	+	+	+	−
CircoDGCV	−	+	−	+
CircoBFDV	+	+	+	+
AnelloTTSuV1	+	+	−	+
AnelloTTSuV2	+	+	−	+
AnelloTTV	+	+	−	+
ParvoAAV	+	+	−	+
ParvoHBoV	+	+	+	+
ParvoMPV	+	+	+	+
NanoBBTV-R	+	+	+	+
NanoBBTV-S	+	+	+	+
NanoBBTV-M	+	+	+	+
NanoBBTV-N	+	−	+	+
NanoBBTV-C	+	−	+	+
GeminiMSV	+	+	+	+
GeminiWDV	+	+	+	+
GeminiPanSV	+	−	+	+
GeminiTYDV-CpCV	+	+	+	+
GeminiCpCDV	+	+	+	+
GeminiTYLCV	+	−	−	+
GeminiEACMV	+	+	+	+
GeminiMYVYV	+	+	+	+

[a] Data sets that had more than five structures significantly conserved in all lineages are indicated with a plus sign.
[b] Data sets in which at least for one gene alignment the synonymous substitution (*dS*) rates at paired-codon sites were significantly lower than those at the unpaired-codon sites are indicated with a plus sign.
[c] Data sets 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 indicated with a plus sign.
[d] Data sets in which a statistically significant association between paired sites and complementarily coevolving sites is detected are indicated with plus sign.

logically functional (Table 5). We have further provided a few examples of currently uncharacterized 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 disfavoring 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 analyzed 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 toward complementary coevolution with their predicted pairing partners, presumably to maintain the biological function of their parent structures.

Even though we have provided 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 favoring the overall maintenance of pervasive biologically functional nucleic acid secondary structures has likely been a major theme in the evolutionary history of eukaryotic ssDNA viruses.

ACKNOWLEDGMENTS

B.M.M. is funded by the University of Cape Town, South Africa. P.L. and J.-M.L. are funded by the Conseil Régional de La Réunion, European Union (FEDER) and Centre de Coopération Internationale en Recherche Agronomique pour le Développement. A.L.M. is funded by the National Research Foundation (South Africa) and the Carnegie Corporation of New York. B.M. is funded by the CFAR Translational Virology Core (P30 AI036214), Molecular Epidemiology (Avant Garde grant; DP1 DA034978). A.F.Y.P. is supported by a New Investigator Award from the Canadian Institutes of Health Research (Canadian HIV Vaccine Initiative) and by a Scholar Award from the Michael Smith Foundation for Health Research/St. Paul's Hospital Foundation—Providence Health Care Research Institute. D.N.S. is funded by Pannar (Pty.), Ltd. D.P.M. and G.W.H. are funded by the South African National Research Foundation.

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

REFERENCES

1. **Yuen L, Moss B.** 1987. Oligonucleotide sequence signaling transcriptional termination of vaccinia virus early genes. Proc. Natl. Acad. Sci. U. S. A. **84:**6417–6421. http://dx.doi.org/10.1073/pnas.84.18.6417.
2. **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:**4482–4494. http://dx.doi.org/10.1111/j.1742-4658.2006.05454.x.
3. **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:**4655–4664. http://dx.doi.org/10.1128/JVI.78.9.4655-4664.2004.
4. **Song SI, Miller WA.** 2004. *cis* and *trans* Requirements for rolling circle replication of a satellite RNA. J. Virol. **78:**3072–3082. http://dx.doi.org/10.1128/JVI.78.6.3072-3082.2004.
5. **Stockley PG, Twarock R, Bakker SE, Barker AM, Borodavka A, Dykeman E, Ford RJ, Pearson AR, Phillips SEV, Ranson NA, Tuma R.** 2013. Packaging signals in single-stranded RNA viruses: nature's alternative to a purely electrostatic assembly mechanism. J. Biol. Phys. **39:**277–287. http://dx.doi.org/10.1007/s10867-013-9313-0.
6. **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:**152–160. http://dx.doi.org/10.1006/viro.2001.1203.
7. **Berns KI.** 1990. Parvovirus replication. Microbiol. Rev. **54:**316–329.
8. **Gronenborn B.** 2004. Nanoviruses: genome organisation and protein function. Vet. Microbiol. **98:**103–109. http://dx.doi.org/10.1016/j.vetmic.2003.10.015.

9. **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:**3034–3039.
10. **Faurez F, Dory D, Grasland B, Jestin A.** 2009. Replication of porcine circoviruses. Virol. J. **6:**60. http://dx.doi.org/10.1186/1743-422X-6-60.
11. **Sun X, Simon AE.** 2006. A cis-replication element functions in both orientations to enhance replication of Turnip crinkle virus. Virology **352:**39–51. http://dx.doi.org/10.1016/j.virol.2006.03.051.
12. **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:**186–195. http://dx.doi.org/10.1006/viro.1995.1467.
13. **Ilyinskii PO, Schmidt T, Lukashev D, Meriin AB, Thoidis G, Frishman D, Shneider AM.** 2009. Importance of mRNA secondary structural elements for the expression of influenza virus genes. OMICS **13:**421–430. http://dx.doi.org/10.1089/omi.2009.0036.
14. **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:**2876–2885.
15. **Guo L, Allen EM, Miller WA.** 2001. Base-pairing between untranslated regions facilitates translation of uncapped, nonpolyadenylated viral RNA. Mol. Cell **7:**1103–1109. http://dx.doi.org/10.1016/S1097-2765(01)00252-0.
16. **Zuo X, Wang J, Yu P, Eyler D, Xu H, Starich MR, Tiede DM, Simon AE, Kasprzak W, Schwieters CD, Shapiro BA, Wang YX.** 2009. Solution structure of the cap-independent translational enhancer and ribosome-binding element in the 3′ UTR of turnip crinkle virus. Proc. Natl. Acad. Sci. U. S. A. **107:**1385–1390. http://dx.doi.org/10.1073/pnas.0908140107.
17. **Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW, Swanstrom R, Burch CL, Weeks KM.** 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. Nature **460:**711–716. http://dx.doi.org/10.1038/nature08237.
18. **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:**1305–1310. http://dx.doi.org/10.4161/rna.22343.
19. **Wikström FH, Meehan BM, Berg M, Timmusk S, Elving J, Fuxler L, Magnusson M, Allan GM, McNeilly F, Fossum C.** 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:**4919–4927. http://dx.doi.org/10.1128/JVI.02797-06.
20. **Wikström FH, Fossum C, Fuxler L, Kruse R, Lövgren T.** 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:**156–166. http://dx.doi.org/10.1016/j.vetimm.2010.09.010.
21. **Simmonds P, Tuplin A, Evans DJ.** 2004. Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: implications for virus evolution and host persistence. RNA. **10:**1337–1351. http://dx.doi.org/10.1261/rna.7640104.
22. **Simon AE, Gehrke L.** 2009. RNA conformational changes in the life cycles of RNA viruses, viroids, and virus-associated RNAs. Biochim. Biophys. Acta **1789:**571–583. http://dx.doi.org/10.1016/j.bbagrm.2009.05.005.
23. **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. http://dx.doi.org/10.4161/rna.9.1.18178.
24. **Powell DM, Amaral MC, Wu JY, Maniatis T, Greene WC.** 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:**973–978. http://dx.doi.org/10.1073/pnas.94.3.973.
25. **Le SY, Malim MH, Cullen BR, Maizel JV.** 1990. A highly conserved RNA folding region coincident with the Rev response element of primate immunodeficiency viruses. Nucleic Acids Res. **18:**1613–1623. http://dx.doi.org/10.1093/nar/18.6.1613.
26. **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. http://dx.doi.org/10.1128/JVI.78.3.1352-1366.2004.
27. **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. http://dx.doi.org/10.1006/viro.2001.1268.
28. **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:**6720–6730. http://dx.doi.org/10.1128/JVI.77.12.6720-6730.2003.
29. **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. http://dx.doi.org/10.1016/S0022-2836(02)01369-4.
30. **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:**221–229. http://dx.doi.org/10.1006/viro.2001.0825.
31. **Paul AV, Rieder E, Kim DW, van Boom JH, Wimmer E.** 2000. Identification of an RNA hairpin in poliovirus RNA that serves as the primary template in the in vitro uridylylation of VPg. J. Virol. **74:**10359–10370. http://dx.doi.org/10.1128/JVI.74.22.10359-10370.2000.
32. **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. http://dx.doi.org/10.1016/S0042-6822(03)00346-5.
33. **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:**2233–2250. http://dx.doi.org/10.3390/v4102233.
34. **Thurner C.** 2004. Conserved RNA secondary structures in *Flaviviridae* genomes. J. Gen. Virol. **85:**1113–1124. http://dx.doi.org/10.1099/vir.0.19462-0.
35. **Pelletier J, Sonenberg N.** 1988. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. Nature **334:**320–325. http://dx.doi.org/10.1038/334320a0.
36. **Kolupaeva VG, Pestova TV, Hellen CU.** 2000. Ribosomal binding to the internal ribosomal entry site of classical swine fever virus. RNA. **6:**1791–1807. http://dx.doi.org/10.1017/S1355838200000662.
37. **Kanamori Y, Nakashima N.** 2001. A tertiary structure model of the internal ribosome entry site (IRES) for methionine-independent initiation of translation. RNA **7:**266–274. http://dx.doi.org/10.1017/S1355838201001741.
38. **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:**1629–1633. http://dx.doi.org/10.1042/BST0351629.
39. **Simon AE, Miller WA.** 2013. 3′ Cap-independent translation enhancers of plant viruses. Annu. Rev. Microbiol. **67:**21–42. http://dx.doi.org/10.1146/annurev-micro-092412-155609.
40. **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:**258–264. http://dx.doi.org/10.1006/viro.1996.0058.
41. **Sun Y, Chen AY, Cheng F, Guan W, Johnson FB, Qiu J.** 2009. Molecular characterization of infectious clones of the minute virus of canines reveals unique features of bocaviruses. J. Virol. **83:**3956–3967. http://dx.doi.org/10.1128/JVI.02569-08.
42. **Perros M, Spegelaere P, Dupont F, Vanacker JM, Rommelaere J.** 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:**2645–2653. http://dx.doi.org/10.1099/0022-1317-75-10-2645.
43. **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:**3515–3521.
44. **Ben-Asher E, Aloni Y.** 1984. Transcription of minute virus of mice, an autonomous parvovirus, may be regulated by attenuation. J. Virol. **52:**266–276.
45. **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:**12–16. http://dx.doi.org/10.1073/pnas.86.1.12.
46. **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:**316–327. http://dx.doi.org/10.1016/0042-6822(88)90502-8.
47. **Orozco BM, Hanley-Bowdoin L.** 1996. A DNA structure is required for geminivirus replication origin function. J. Virol. **70:**148–158.
48. **Hafner GJ, Stafford MR, Wolter LC, Harding RM, Dale JL.** 1997. Nicking and joining activity of banana bunchy top virus replication protein in vitro. J. Gen. Virol. **78:**1795–1799.
49. **Cheung AK.** 2006. Rolling-circle replication of an animal circovirus

genome in a theta-replicating bacterial plasmid in *Escherichia coli*. J. Virol. **80:**8686–8694. http://dx.doi.org/10.1128/JVI.00655-06.
50. **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. http://dx.doi.org/10.1080/07391102.1994.10508036.
51. **Shepherd DN, Martin DP, Varsani A, Thomson JA, Rybicki EP, Klump HH.** 2006. Restoration of native folding of single-stranded DNA sequences through reverse mutations: an indication of a new epigenetic mechanism. Arch. Biochem. Biophys. **453:**108–122. http://dx.doi.org/10.1016/j.abb.2005.12.009.
52. **Martin DP, Lefeuvre P, Varsani A, Hoareau M, Semegni J, Dijoux B, Vincent C, Reynaud B, Lett J.** 2011. Complex recombination patterns arising during geminivirus coinfections preserve and demarcate biologically important intra-genome interaction networks. PLoS Pathog. **7:**e1002203. http://dx.doi.org/10.1371/journal.ppat.1002203.
53. **Schultes EA, Spasic A, Mohanty U, Bartel DP.** 2005. Compact and ordered collapse of randomly generated RNA sequences. Nat. Struct. Mol. Biol. **12:**1130–1136. http://dx.doi.org/10.1038/nsmb1014.
54. **Hasegawa M, Yasunaga T, Miyata T.** 1979. Secondary structure of MS2 phage RNA and bias in code word usage. Nucleic Acids Res. **7:**2073–2079. http://dx.doi.org/10.1093/nar/7.7.2073.
55. **Cardinale DJ, DeRosa K, Duffy S.** 2013. Base composition and translational selection are insufficient to explain codon usage bias in plant viruses. Viruses **5:**162–181. http://dx.doi.org/10.3390/v5010162.
56. **Ngandu NK, Scheffler K, Moore P, Woodman Z, Martin D, Seoighe C.** 2008. Extensive purifying selection acting on synonymous sites in HIV-1 group M sequences. Virol. J. **5:**160. http://dx.doi.org/10.1186/1743-422X-5-160.
57. **Cheung AK.** 2005. Detection of rampant nucleotide reversion at the origin of DNA replication of porcine circovirus type 1. Virology **333:**22–30. http://dx.doi.org/10.1016/j.virol.2004.12.016.
58. **Cheng N, Mao Y, Shi Y, Tao S.** 2012. Coevolution in RNA molecules driven by selective constraints: evidence from 5S rRNA. PLoS One **7:**e44376. http://dx.doi.org/10.1371/journal.pone.0044376.
59. **Fernández N, Fernandez-Miragall O, Ramajo J, García-Sacristán A, Bellora N, Eyras E, Briones C, Martínez-Salas E.** 2011. Structural basis for the biological relevance of the invariant apical stem in IRES-mediated translation. Nucleic Acids Res. **39:**8572–8585. http://dx.doi.org/10.1093/nar/gkr560.
60. **Hofacker I, Fekete M, Flamm C, Huyenen M, Rauscher S, Stolorz P, Stadler PF.** 1998. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. Nucleic Acids Res. **26:**3825–3836. http://dx.doi.org/10.1093/nar/26.16.3825.
61. **Cheung AK.** 2004. Detection of template strand switching during initiation and termination of DNA replication of porcine circovirus. J. Virol. **78:**4268–4277. http://dx.doi.org/10.1128/JVI.78.8.4268-4277.2004.
62. **Cheung AK.** 2004. 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:**9016–9029. http://dx.doi.org/10.1128/JVI.78.17.9016-9029.2004.
63. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32:**1792–1797. http://dx.doi.org/10.1093/nar/gkh340.
64. **Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S.** 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. **28:**2731–2739. http://dx.doi.org/10.1093/molbev/msr121.
65. **Wilm A, Mainz I, Steger G.** 2006. An enhanced RNA alignment benchmark for sequence alignment programs. Algorithms Mol. Biol. **1:**19. http://dx.doi.org/10.1186/1748-7188-1-19.
66. **Semegni JY, Wamalwa M, Gaujoux R, Harkins GW, Gray A, Martin DP.** 2011. NASP: a parallel program for identifying evolutionarily conserved nucleic acid secondary structures from nucleotide sequence alignments. Bioinformatics **27:**2443–2445. http://dx.doi.org/10.1093/bioinformatics/btr417.
67. **Greiner W, Neise L, Stöcker H.** 1995. Thermodynamics and statistical mechanics. Springer-Verlag, New York, NY.
68. **Ding Y.** 2003. A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res. **31:**7280–7301. http://dx.doi.org/10.1093/nar/gkg938.
69. **Markham NR, Zuker M.** 2008. UNAFold: software for nucleic acid folding and hybridization. Methods Mol. Biol. **453:**3–31. http://dx.doi.org/10.1007/978-1-60327-429-6_1.
70. **Tajima F.** 1989. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics **123:**585–595.
71. **Fu YX, Li WH.** 1993. Statistical tests of neutrality of mutations. Genetics **133:**693–709.
72. **Scheffler K, Martin DP, Seoighe C.** 2006. Robust inference of positive selection from recombining coding sequences. Bioinformatics **22:**2493–2499. http://dx.doi.org/10.1093/bioinformatics/btl427.
73. **Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K.** 2013. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol. Biol. Evol. **30:**1196–1205. http://dx.doi.org/10.1093/molbev/mst030.
74. **Muse SV, Gaut BS.** 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. **11:**715–724.
75. **Scheffler K, Martin DP, Seoighe C.** 2006. Robust inference of positive selection from recombining coding sequences. Bioinformatics **22:**2493–2499. http://dx.doi.org/10.1093/bioinformatics/btl427.
76. **Lefeuvre P, Lett Varsani J-MA, Martin DP.** 2009. Widely conserved recombination patterns among single-stranded DNA viruses. J. Virol. **83:**2697–2707. http://dx.doi.org/10.1128/JVI.02152-08.
77. **Julian L, Piasecki T, Chrzastek K, Walters M, Muhire B, Harkins GW, Martin DP, Varsani A.** 2013. Extensive recombination detected among beak and feather disease virus isolates from breeding facilities in Poland. J. Gen. Virol. **94:**1086–1095. http://dx.doi.org/10.1099/vir.0.050179-0.
78. **Padidam M, Sawyer S, Fauquet CM.** 1999. Possible emergence of new geminiviruses by frequent recombination. Virology **265:**218–225. http://dx.doi.org/10.1006/viro.1999.0056.
79. **Shackelton LA, Hoelzer K, Parrish CR, Holmes EC.** 2007. Comparative analysis reveals frequent recombination in the parvoviruses. J. Gen. Virol. **88:**3294–3301. http://dx.doi.org/10.1099/vir.0.83255-0.
80. **Navas-Castillo J, Sánchez-Campos S, Noris E, Louro D, Accotto GP, Moriones E.** 2000. Natural recombination between tomato yellow leaf curl virus-is and tomato leaf curl virus. J. Gen. Virol. **81:**2797–2801.
81. **Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW.** 2006. GARD: a genetic algorithm for recombination detection. Bioinformatics **22:**3096–3098. http://dx.doi.org/10.1093/bioinformatics/btl474.
82. **Pond SLK, Frost SDW, Muse SV.** 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics **21:**676–679. http://dx.doi.org/10.1093/bioinformatics/bti079.
83. **Hasegawa M, Kishino H, Yano T.** 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22:**160–174. http://dx.doi.org/10.1007/BF02101694.
84. **Muse SV.** 1995. Evolutionary analyses of DNA sequences subject to constraints of secondary structure. Genetics **139:**1429–1439.
85. **Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P.** 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics **26:**2462–2463. http://dx.doi.org/10.1093/bioinformatics/btq467.
86. **Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O.** 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. **59:**307–321. http://dx.doi.org/10.1093/sysbio/syq010.
87. **McCormack JC, Simon AE.** 2004. Biased hypermutagenesis associated with mutations in an untranslated hairpin of an RNA virus. J. Virol. **78:**7813–7817. http://dx.doi.org/10.1128/JVI.78.14.7813-7817.2004.
88. **Staplin WR, Miller WA.** 2008. In vivo analyses of viral RNA translation. Methods Mol. Biol. **451:**99–112.
89. **Simmonds P, Smith DB.** 1999. Structural constraints on RNA virus evolution. J. Virol. **73:**5787–5794.
90. **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. http://dx.doi.org/10.1021/bi00462a015.
91. **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. http://dx.doi.org/10.1186/1471-2156-6-20.
92. **Monjane AL, Pande D, Lakay F, Shepherd DN, van der Walt E, Lefeuvre P, Lett Varsani J-MA, Rybicki EP, Martin DP.** 2012. Adaptive evolution by recombination is not associated with increased mutation rates in maize streak virus. BMC Evol. Biol. **12:**252. http://dx.doi.org/10.1186/1471-2148-12-252.
93. **Fisher RA.** 1922. On the Interpretation of χ^2 from contingency tables,

and the calculation of P. J. R. Stat. Soc. **85:**87–94. http://dx.doi.org/10.2307/2340521.

94. **Golden M, Martin D.** 2013. DOOSS: a tool for visual analysis of data overlaid on secondary structures. Bioinformatics **29:**271–272. http://dx.doi.org/10.1093/bioinformatics/bts667.
95. **Ray SS, Pal SK.** 2013. RNA secondary structure prediction using soft computing. IEEE/ACM Trans. Comput. Biol. Bioinform. **10:**2–17. http://dx.doi.org/10.1109/TCBB.2012.159.
96. **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:**4629–4635.
97. **Wright E, Heckel T, Groenendijk J, Davies J, Boulton M.** 1997. Splicing features in maize streak virus virion- and complementary-sense gene expression. Plant J. **12:**1285–1297. http://dx.doi.org/10.1046/j.1365-313x.1997.12061285.x.
98. **Warf MB, Berglund JA.** 2010. Role of RNA structure in regulating pre-mRNA splicing. Trends Biochem. Sci. **35:**169–178. http://dx.doi.org/10.1016/j.tibs.2009.10.004.
99. **Munroe SH.** 1984. Secondary structure of splice sites in adenovirus mRNA precursors. Nucleic Acids Res. **12:**8437–8456. http://dx.doi.org/10.1093/nar/12.22.8437.
100. **de Villiers E-M, Borkosky SS, Kimmel R, Gunst K, Fei J-W.** 2011. The diversity of torque teno viruses: *in vitro* replication leads to the formation of additional replication-competent subviral molecules. J. Virol. **85:**7284–7295. http://dx.doi.org/10.1128/JVI.02472-10.