jmb Article

Packaging Signals in Two Single-Stranded RNA Viruses Imply a Conserved Assembly Mechanism and Geometry of the Packaged Genome

Eric C. Dykeman[1], Peter G. Stockley[2] and Reidun Twarock[1]

[1] - **Departments of Mathematics and Biology and York Centre for Complex Systems Analysis,** University of York, York YO10 5DD, UK
[2] - **Astbury Centre for Structural Molecular Biology,** University of Leeds, Leeds LS2 9JT, UK

Correspondence to Reidun Twarock: *Departments of Mathematics and Biology and York Centre for Complex Systems Analysis, University of York, York YO10 5DD, UK.* reidun.twarock@york.ac.uk
http://dx.doi.org/10.1016/j.jmb.2013.06.005
Edited by J. Johnson

Abstract

The current paradigm for assembly of single-stranded RNA viruses is based on a mechanism involving non-sequence-specific packaging of genomic RNA driven by electrostatic interactions. Recent experiments, however, provide compelling evidence for sequence specificity in this process both *in vitro* and *in vivo*. The existence of multiple RNA packaging signals (PSs) within viral genomes has been proposed, which facilitates assembly by binding coat proteins in such a way that they promote the protein–protein contacts needed to build the capsid. The binding energy from these interactions enables the confinement or compaction of the genomic RNAs. Identifying the nature of such PSs is crucial for a full understanding of assembly, which is an as yet untapped potential drug target for this important class of pathogens. Here, for two related bacterial viruses, we determine the sequences and locations of their PSs using Hamiltonian paths, a concept from graph theory, in combination with bioinformatics and structural studies. Their PSs have a common secondary structure motif but distinct consensus sequences and positions within the respective genomes. Despite these differences, the distributions of PSs in both viruses imply defined conformations for the packaged RNA genomes in contact with the protein shell in the capsid, consistent with a recent asymmetric structure determination of the MS2 virion. The PS distributions identified moreover imply a preferred, evolutionarily conserved assembly pathway with respect to the RNA sequence with potentially profound implications for other single-stranded RNA viruses known to have RNA PSs, including many animal and human pathogens.

Introduction

Single-stranded RNA (ssRNA) viruses are one of the largest groups of viral pathogens, yet their assembly mechanisms are still poorly understood. In particular, the potential roles in this process of defined interactions between RNA segments within their genomes and the coat proteins (CPs) that form a protective capsid layer have largely been neglected. This is mostly due to the difficulty in identifying such CP–RNA contacts, called packaging signals (PSs), in viral genomes. We introduce here a new method for the identification of PSs in ssRNA genomes and demonstrate its consequences for two related RNA phages, MS2 and GA, from the Leviviridae family. We show that geometric constraints on the positions of PSs in the viral genomes can be formulated via a Hamiltonian path, a geometrical concept that, in general, provides information about connectivity between different vertices of a graph and, here in particular, about the graph representing the RNA density in proximity to capsid. We demonstrate that the Hamiltonian path concept, in combination with bioinformatics and structural data, reveals astonishing insights into the biology of these phages. For example, it shows that the PS distribution in both phages must have the same geometric organization in contact with the proteins of the capsid, implying that their assembly pathways and the conformations of their packaged genomes in the capsid must be identical.

The current paradigm of assembly assumes that RNA packaging is not sequence-specific and driven by electrostatic interactions.[1–4] This has been reinforced by the fact that capsid-like particles can form *in vitro* and, in simulations, *in silico* in the presence of non-cognate RNAs, polyanions or even nanoparticles.[5–8] This mechanism of assembly, however, fails to account for the observed encapsidation specificity in most of these viruses recovered from natural hosts.[9] This discrepancy has led to proposals that there must be specialized cellular compartments to sequester viral RNAs and CPs, or that only nascent RNA chains are packaged. The assumption that packaging is driven by non-specific interactions also contradicts recent single-molecule fluorescence correlation spectroscopy (smFCS) assays of reassembly of two model viruses, the plant virus, satellite tobacco necrosis virus (STNV) and bacteriophage MS2.[10–12] These assays used ~100- to 1000-fold lower concentration of CP than in most *in vitro* ensemble assembly reactions, reducing the dominance of CPs on the reaction and perhaps giving a more accurate reflection of *in vivo* conditions. These assays monitored the co-assembly of CPs onto protein-free genomic RNA under defined conditions. For both viruses, the RNAs are initially bound by a subset of the proteins required to form their capsids, resulting in a sudden collapse in their hydrodynamic radii from an ensemble of conformations that are mostly too large to fit within the confines of the respective virions. In the collapsed state, the RNA–CP complexes are small enough to fit and they have the size and symmetry of partially formed capsids, that is, they are the result of capsid assembly rather than formation of a CP–RNA aggregate that subsequently rearranges into a capsid. Following the initial collapse, additional CPs are recruited to the growing shells to complete capsids in a second, slower stage of assembly. The yields and fidelity of correctly assembled species from this process are extremely high. Non-viral and non-cognate viral RNAs also promote capsid assembly under these conditions, but with dramatically reduced efficiency and with a majority of malformed or aggregated species and without the compaction stage. This principal difference in assembly behavior for cognate and non-cognate RNAs suggests that specific RNA–CP interactions are important for assembly efficiency and that, therefore, the non-specific interactions in an electrostatic model of assembly cannot fully account for the behavior observed in these experiments. An alternative assembly model that explicitly takes the impact of such dispersed PS in the viral genomes into account[13] demonstrates that their distribution is crucial for assembly efficiency. This suggests that identification of PS in ssRNA genomes is important in order to fully understand their assembly behavior.

We have previously made progress on this task with STNV.[14,15] We used *in vitro* RNA SELEX to identify preferred RNA binding sequences with affinity for the STNV CP subunit. This yielded a series of aptamers, most of which had statistically significant matches to regions of the STNV genome and encompassed sequences that were able to form stem–loops, with a loop sequence motif of -A.X.X.A-, where X is any nucleotide. The three known STNV strains (1, 2 and C) have genomes that could in principle present up to 30 copies of such motifs. It is therefore tempting to speculate that this aptamer motif represents a consensus PS for STNV. To examine the potential functional significance of these PSs, we carried out *in vitro* reassembly assays with the best aptamer and a variant with a -U.U.U.U- loop motif.[15] STNV CP subunits will not assemble *in vitro* in the absence of RNA under these conditions. The preferred sequence triggered formation of $T = 1$ VLPs more efficiently than the sequence variant, and longer genomic fragments were more efficient than long, non-cognate RNAs. These results support the idea that the STNV genome contains multiple PSs that are absent from a non-cognate RNA. The STNV CP-aptamer VLP formed crystals, and we solved its structure by X-ray diffraction. In the presence of repeated copies of the preferred RNA stem–loop fragment, the CP has undergone a conformational change, compared to the virion and a VLP containing a synthetic CP mRNA.[16,17] With the aptamer present, a region of polypeptide (residues 8–11) is visible in the electron density map that is disordered and invisible when the genomic RNA is packaged. The ordered region extends an N-terminal helix that is positioned around the 3-fold axes of the $T = 1$ particle. This extension contains a number of positively charged amino acids (RRKS) that could potentially prevent formation of virion 3-fold because of electrostatic repulsion.[15] Binding of the aptamer appears to have screened this effect, identifying a possible functional role for PSs in this case.

The similarity in behavior between STNV and MS2 in the smFCS assay suggests that there are conserved features in the assembly mechanisms for these viruses. To identify the PSs in MS2, we examined its genomic sequence and that of the evolutionarily related RNA phage GA. There is extensive evidence that PSs exist in the RNA phages. Firstly, there is a well-known, single-copy, high-affinity CP binding site (TR) that acts as a translational operator, putting the expression of the replicase gene under the control of the CP dimer concentration.[18,19] The MS2 19-nt TR site and its GA equivalent are also thought to be the assembly initiation triggers for capsid formation both *in vitro* and *in vivo*.[18,20–22] A number of studies support the idea that additional sites within the MS2 genome also promote capsid assembly.[23–25] We have also shown that stem–loop binding to the CP plays a critical role in assembly. It acts as an allosteric regulator of CP conformation. TR binding promotes a shift in the conformation of the CP dimer from a symmetric (C/C-like) form (pink in Fig. 1a) to an asymmetric (or A/B-like) form.[22,24,26–28] This "Dimer Switching Model" (DSM)[28] is one of the functional consequences of PS–CP interactions in these viruses. Both types of dimer are required to construct a $T = 3$ capsid of the correct size and symmetry. Individually, each form, RNA-free and TR-bound, is kinetically trapped, failing to produce capsids over many hours or days. If both forms are mixed in solution or if CP is in the presence of weaker binding stem–loops, $T = 3$ capsids rapidly self-assemble. Crucially, the allosteric consequences of RNA stem–loop binding are not highly sequence-specific,[26] in principle allowing many similar but non-identical sites within the genome to play the same role. Assembly reactions with longer RNAs show that assembly with such molecules is highly cooperative, consistent with this idea and the presence of multiple PSs.[24] Ideally, there would be 60 PSs in such a system able to promote switching at the 60 required A/B dimers, although this requirement is not absolute since shorter RNAs can also be packaged. Presumably the kinetics of capsid formation would be slower in these cases, which would depend significantly on CP–CP interactions alone to reach completion. Note that, in MS2 and GA, we expect 60 PSs due to the functional roles of the PSs in dimer switching at the

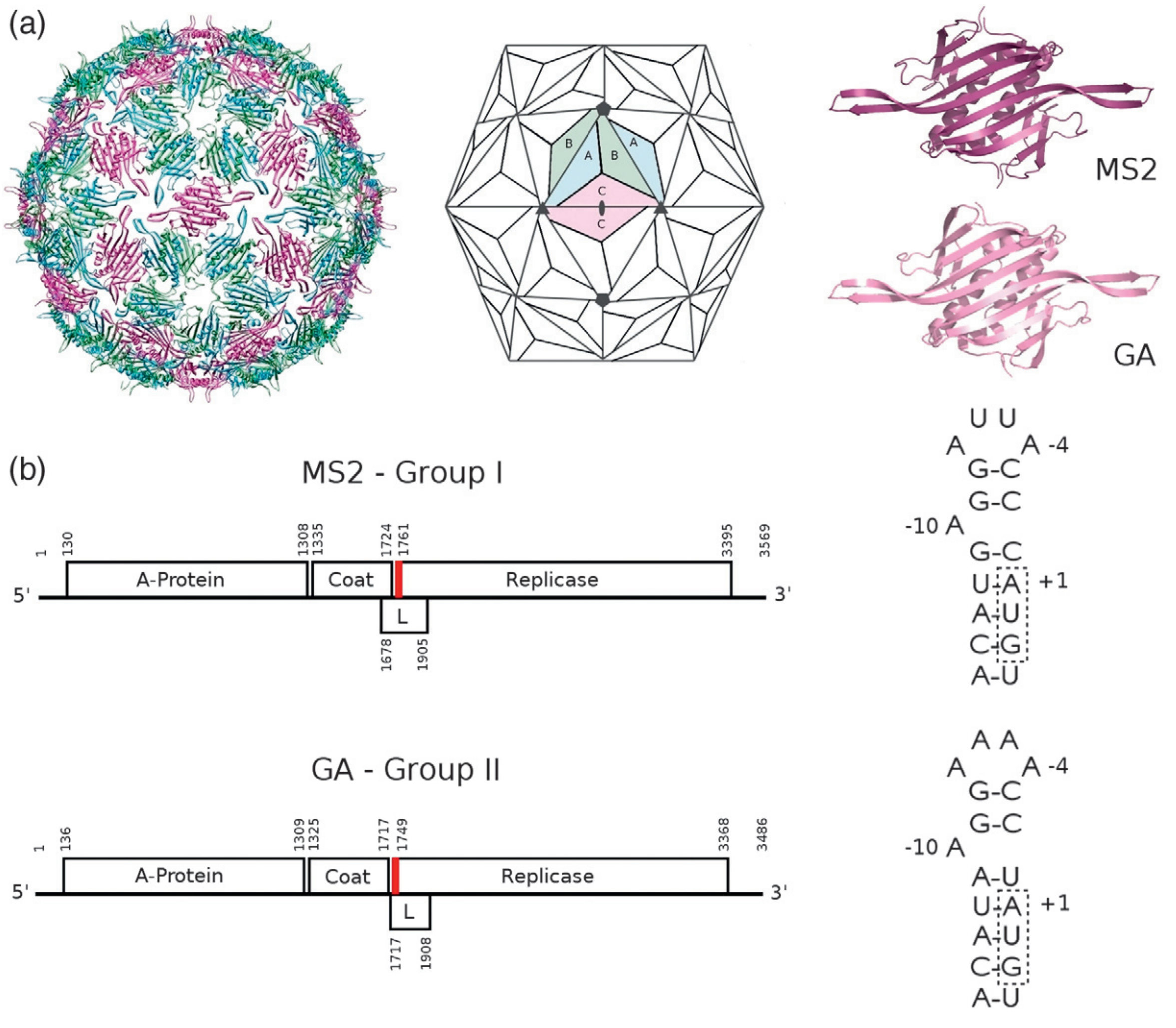

Fig. 1. The components of MS2 and GA. (a) The MS2 capsid ($T = 3$; Protein Data Bank ID 1ZDH) viewed along a 2-fold axis, shown next to a schematic indicating the locations of the 60 A/B (blue/green) and 30 C/C (pink) dimers. The structural similarity of MS2 and GA (Protein Data Bank ID 1GAV) CPs is revealed by cartoons of the C^{α} chains of their C/C dimers (RMSD = 3.7 Å). (b) Genetic maps of the phage genomes and alongside the secondary structures of their high-affinity translational operators whose locations are marked in the maps by the red stripes.

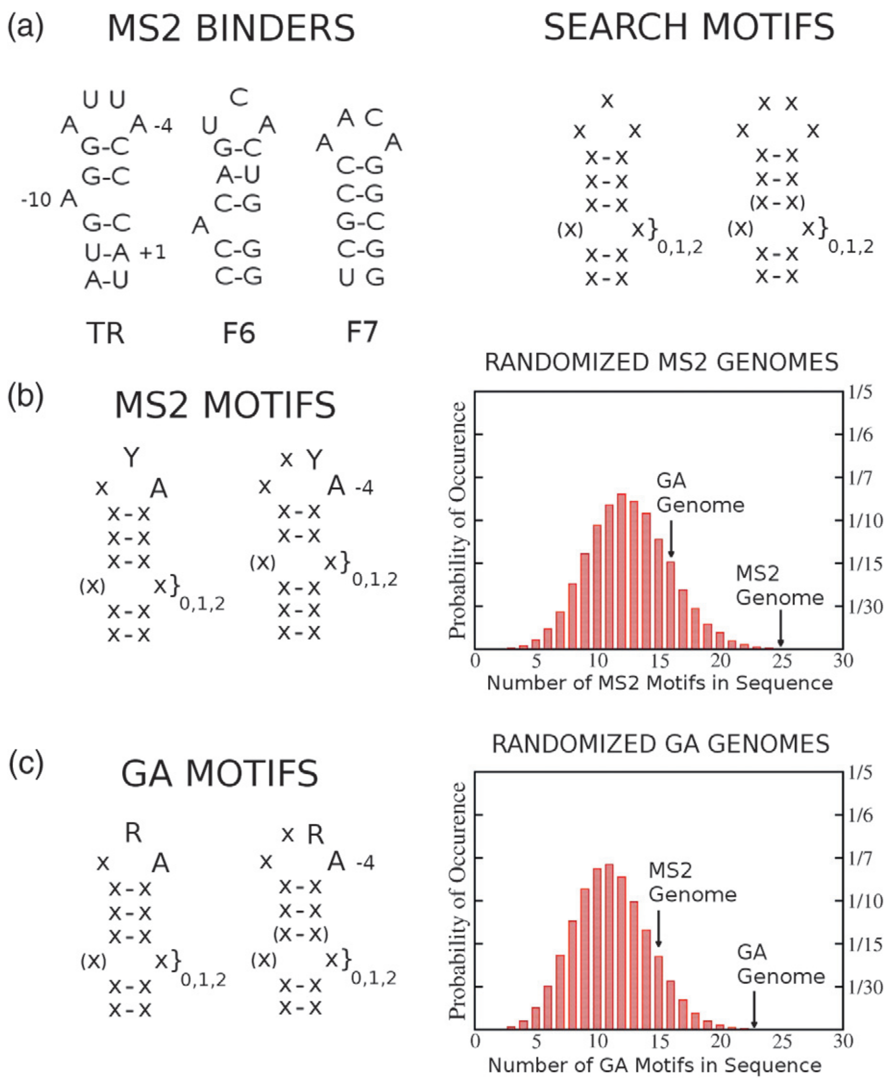

Fig. 2. Search motifs for putative PSs. (a) Sequences of high-affinity stem–loops that bind to the MS2 CP (left) and the search motifs that capture their essential features (right). The latter also encompasses the equivalent sites for GA.[18,31] (x) and (x-x) denote a nucleotide or base pair, respectively, that can be omitted, while $x\}_{0,1,2}$ indicates a location for 0, 1 or 2 additional nucleotides. (b) and (c) show the derived recognition motifs that discriminate between MS2 and GA CP binding. The adjacent histograms illustrate the number of occurrences of such recognition motifs in ensembles of 30,000 randomized versions of each genome. The positions of the MS2 and GA genome in this distribution are indicated explicitly.

A/B sites. In other viral systems, this number could vary depending on the functional roles of the PSs, but the general analysis principles adopted here could also be applied to this more general setting.

In contrast to STNV, the PSs within the MS2 and GA genomes cannot be identified by sequence analysis alone. We have therefore developed a new approach that uses Hamiltonian paths to formulate geometric constraints on genome organization in addition to biochemical (RNA SELEX) information about the RNA sequence preferences of both MS2 and GA phage CP subunits.[29–34] We show that there are characteristic stem–loop motifs with little sequence identity but that occur with statistically significant frequencies in the respective genomes, consistent with them having a functional role. These predictions were further validated by comparison with prior secondary structure probing of the protein-free RNA suggesting that ~60% of these potential PS sequences should be present in the form of stem–loops.[35–37] By combining our PS predictions with structural information from cryo-electron microscopy (cryo-EM) studies on the location of the genome inside the MS2 capsid and other *Leviviridae*,[38,39] we predict their locations within both virions using our new graph theoretical approach. We show that the packaged genomes must have a defined conformation inside each viral particle. This is entirely consistent with the first asymmetric structure of MS2 determined using cryo-electron tomography and sub-tomographic averaging and explains the observations reported in that study. Such defined PS–CP interactions add a previously unsuspected constraint on the evolution of viral genomes and are potential targets for novel anti-viral therapy.

Results

Identification of phage-specific recognition motifs important for selective genome packaging

Understanding the functional roles of PSs in MS2 and GA requires identification of both cognate stem–loop (SL) motifs and their locations within the two genomes. To identify putative PSs other than the known single-copy, high-affinity ones, we applied two minimalistic criteria. Firstly, we assumed that all PSs are in the form of RNA stem–loops and, secondly, that these exhibit at least some of the recognition determinants identified previously for the respective high-affinity sites.[18,20,29–34] Since all known high-affinity stem–loops for MS2 and GA conform to the general search pattern shown in Fig. 2a, we used a sliding window approach (see Materials and Methods) to locate all such sites. This identified the maximal number of lowest-energy, non-overlapping stem–loops with a stem length >5 bp. The MS2 and GA genomes contain 118 and 112 such sites, respectively. Note that the smFCS data suggest that, in general, PSs are already present in the secondary structure of the RNA and are readily available for CP binding. Note that in the case of overlapping stem-loops stability rather than affinity has been the criterion of choice.

We then used information on the structures of RNA aptamers selected *in vitro* (using RNA SELEX)[31–34] as preferred binding sites for MS2 CPs, as well as affinity measurements on sequence variants of the equivalent GA sites, to identify a characteristic recognition motif common to all high-affinity stem–loops for each phage (Fig. 2). For MS2, these are stem–loops with a loop of the form (x)xYA, and for GA, these are stem–loops with a loop of the form (x)xRA, where Y and R denote pyrimidines and purines, respectively; x and (x) denote any nucleotide and a further nucleotide that could be omitted, respectively. If such motifs are important for selective packaging, they must occur non-randomly in their respective genomes. To test this, we created 30,000 randomized sequences with the same base composition as each of the phage genomes and determined the frequency of occurrence of matches to the minimal sequence motifs in each, using the same sliding window approach (Materials and Methods). Consensus motifs with loops of the form (x)xYA and (x)xRA occur with much higher frequency than would be expected by chance alone in the MS2 and GA genomes (cf. Fig. 2b and c), respectively, while they occur with frequencies close to that expected by chance for the non-cognate genomes. This strongly supports their putative roles in assembly and selective packaging. There are 25 and 23 such putative PS sites within MS2 and GA, respectively.

Identification of additional PSs

In an idealized DSM, both phage genomes should contain 60 PSs, each one promoting the conformational switching to an A/B dimer during assembly, and the putative PSs identified above only account for about a third of these. To determine further potential PSs, we created a scoring matrix (see Materials and Methods) based on previous structural studies, RNA SELEX, and sequence/functional group variation experiments.[29–34] We used it to rank all 118 {112} of the possible stable stem–loops for their potential CP affinity and, hence, suitability as PSs (for ease of notation, here and in the following, figures for GA are indicated in curly brackets behind those for MS2). They were then sorted into classes based on their predicted binding affinity relative to TR (Fig. 1b), assuming a dissociation constant for TR of K_D = 1.5 nM.[40] The first class contains the 25 {23} highest-affinity sites with the minimal common recognition motif determined above (with a predicted score of ≥10% of the affinity of TR, i.e., K_D up to 15 nM). Two further classes of putative PSs retain at least partial features of this common recognition motif. Those with the loop sub-motif (x)xxA (same motif for both MS2 and GA) form a class predicted to have binding affinities ≥1% of TR, that is, K_D up to 150 nM and those with (x)xYx {(x)xRx}, ≥0.1% of TR, that is, K_D up to 1.5 μM. In total there are 53 {54} stem–loops in these three classes (Supplementary Figs. 1 and 2). All other potential stem–loops have lower predicted CP affinities. We reason that, since the 53 {54} stem–loops retain some features of the recognition motif, their lower affinities for CP could be compensated by their effectively high concentrations within phage RNA genomes, especially if they are suitably positioned in three dimensions to interact productively with CPs during assembly.[10] Indeed, other SLs with lower affinity can potentially also occur but can only be identified using further constraints, which emerge from the Hamiltonian path approach (see below). Before we carried out this analysis, we compared the assigned 53 {54} SLs with the published solution structures of the phage genomes.

Locating the putative PSs in the solution structures of the genomes

The secondary structures of the MS2 and GA genomes have been mapped using a combination of phylogenetic analysis and experimental structure probing.[35–37] Of the putative PSs that we have identified, 35 {39}, that is, 58% {65%}, are reported to be present in the solution conformations of the genomic RNAs, ready to interact with phage CPs during assembly. Seven of these have been previously identified as potential CP binding sites.[41] This is good evidence that we have identified PSs. The PSs present in the solution structure include 14 {15} of the 25 {23} highest-affinity classes of stem–loop motifs identified above, as well as 21 {24} of the additional PS that were identified using the scoring matrix. There is evidence from assembly assays with MS2 that RNA conformational changes occur during encapsidation,[10,24] opening the possibility that the remaining 18 {15} of the total 53 {54} putative PSs could also occur after minor local refolding. We estimated from Mfold (cf. Ref. 42) the free-energy difference between the reported solution structure and the refolded structure exhibiting these remaining PSs. For all these additional stem–loops, the cost of local refolding was <10 kcal/mol, comparable to the binding free energy of isolated TR stem–loops to MS2 CP (~12 kcal/mol).[40] This suggests that these additional PSs are likely to occur if the additional favorable contributions resulting from CP–CP interactions within the protein shell facilitated by PS binding are taken into account. The smFCS data[10] suggest that MS2 RNA exists as an ensemble of structures in solution, a result observed for other long RNAs.[43,44] These ensembles could represent altered tertiary structures with similar secondary structures or molecules differing at both tertiary and secondary structure levels. The solution structure probing data for several RNA phages favor the former explanation, but the sensitivity and time resolution of such techniques cannot exclude the presence of minor conformers presenting the additional PSs.

Assuming that localized RNA refolding occurs in response to CP binding, our analysis has identified 53 {54}, that is, ~90%, of the potential PSs in both phages needed by the DSM. The locations of these PSs are shown as dots, color-coded according to predicted CP affinity, in cartoon drawings of the secondary structures of MS2 (Fig. 3) and GA (Supplementary Fig. 3). The majority of these putative PSs (49 {46}) lie within coding regions (Fig. 3 and Supplementary Fig. 3). The presence of such functional structures within open reading frames contrasts with the results of structure probing of the human immunodeficiency virus genome, which suggested that coding regions were relatively unstructured.[45,46] Regulatory RNA structures within the RNA phage coding regions are, however, well known.[35–37] Our analysis suggests that, for the RNA phages, each gene contains putative PSs.

Predicting the three-dimensional layout of the RNA PSs in the virion

In order to do this, we introduce here a new method for the localization of individual PSs in the tertiary structure of packaged genomes. It is applicable to any ssRNA virus with a monopartite genome for which the icosahedrally averaged cryo-EM density shows a polyhedral organization

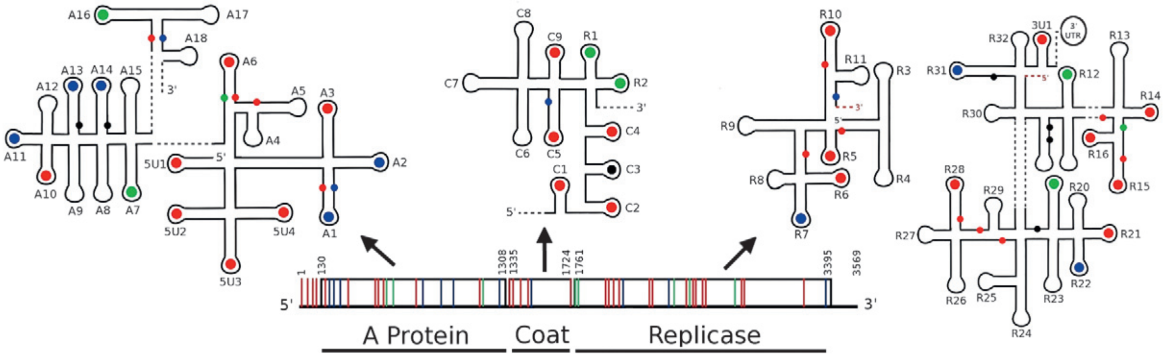

Fig. 3. Locations of the putative MS2 PSs. The 53 putative MS2 PSs are shown as green, blue and red bars on the genetic map, color-coded according to their predicted affinities for CP (green, highest affinity; blue, intermediate affinity; red, lowest affinity), alongside cartoons of regions of the solution structure.[36,37] Colored dots in loops indicate the 35 putative PSs that are present in the absence of CP, while the smaller ones on stems indicate the positions of the 18 PSs that would require local refolding. Black dots represent 7 further potential stem–loops that have positions that are consistent with the identified Hamiltonian path and that could be used to obtain the 60 PSs needed in an idealized DSM. PS locations are numbered increasing in the 5′-to-3′ direction, preceded by the genetic locus, that is, 5′UTR (*un*translated *r*egion) (5U#), A protein (A#), CP (C#), replicase (R#) and 3′UTR (3U#).

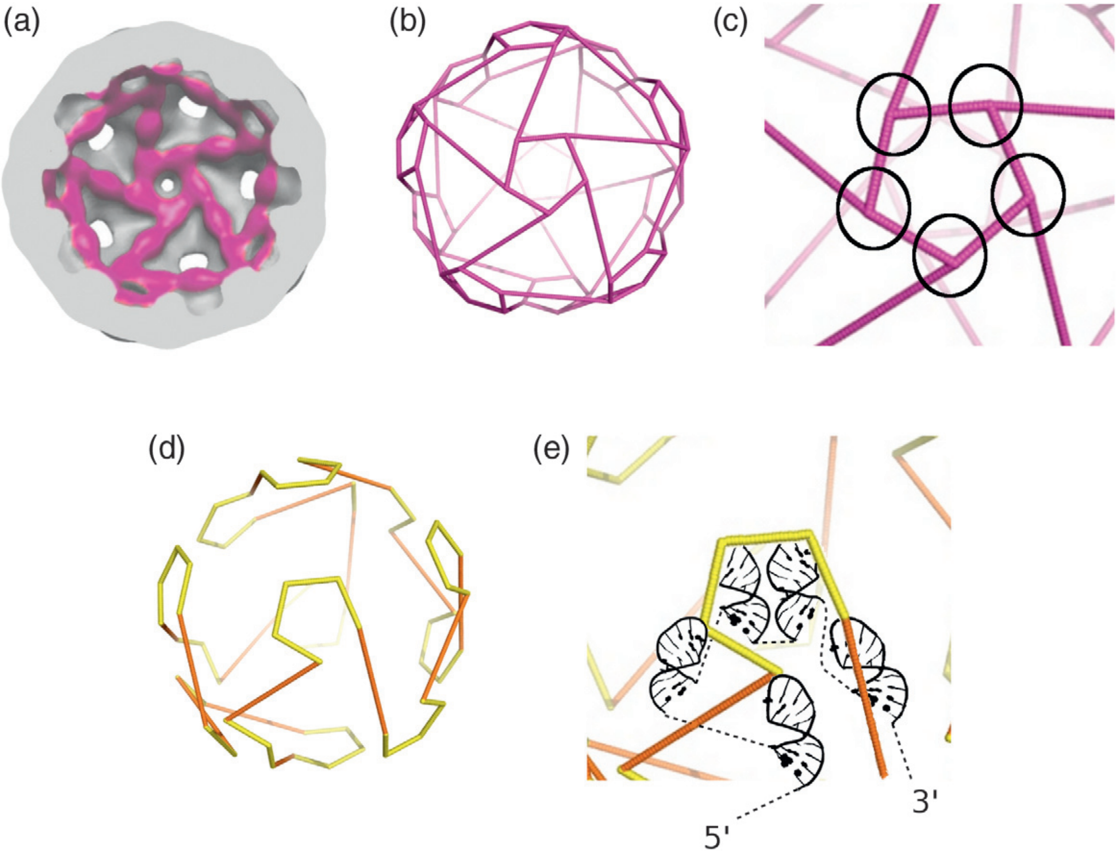

Fig. 4. Localizing PSs in the tertiary structures of packaged RNA genomes using the Hamiltonian path method. (a) The icosahedrally averaged cryo-EM electron density for genomic MS2 RNA (purple) in the virion in contact with capsid protein, viewed from the center of a half-capsid along the central 5-fold axis.[28,39] (b) The density from (a) represented as a polyhedron with 60 vertices. (c) A close-up view of the vertices (circled) around a 5-fold axis that would contact A/B dimers in the DSM. (d) A Hamiltonian path is a path on the polyhedron in (b) that contacts each of its 60 vertices precisely once. Each of the >40,000 possible such paths is characterized by a series of short (yellow) and long (orange) segments. In the RNA phages, single-copy maturation proteins are known to bind close to the ends of the genomic RNA effectively circularizing it.[28] Only 66 Hamiltonian paths have the property of being circular. One such path is shown in (d). Assembly of phages with RNAs following all 66 possible paths or icosahedral averaging of phages encompassing just one of these paths will generate the polyhedron of density seen in (a). (e) A Hamiltonian path encodes the relative positions of PSs in contact with the 60 A/B dimers. Such interactions are shown schematically here for a section of the path in (d).

of the RNA in proximity to the capsid and for which the functional roles of the PS suggest the contact points between PS and CP. Previous intermediate resolution cryo-EM structures of both MS2 and GA[39] suggest that the density in proximity to the capsid (Fig. 4a) is organized as a polyhedron (Fig. 4b). Since the vertices of the polyhedron in Fig. 4b are located underneath the A/B dimers, this is consistent with the contact sites between PSs and CP being at those vertices, indicated by the circles shown in Fig. 4c for 5 of the 60 vertices in a magnified version of the polyhedron. In the idealized assembly case, the RNA genome would contact each A/B dimer within the capsid, and the RNA would visit each of these vertices by extending a stem–loop to the CP layer. However, there is no requirement that the C/C positions be visited, and they would only be occupied when the RNA transits from an A/B dimer around one 5-fold vertex to an A/B dimer around a different vertex. Thus, the binding of RNA at A/B and C/C dimer positions is functionally distinct, and this should be reflected in the density for bound RNA at these positions in cryo-EM structures. Indeed, in a previous sub-nanometer cryo-EM structure of MS2, the density does appear slightly different underneath A/B and C/C dimers, consistent with this expectation from the DSM.

Since the vertices of the polyhedron in Fig. 4b are located underneath the A/B dimers, this is consistent with the contact sites between PSs and CP being at those vertices, indicated by the circles shown in Fig. 4c. In the idealized assembly case, the RNA genome would contact each A/B dimer within the capsid, and the RNA would visit each of these vertices by extending a stem–loop to the CP layer. Since only one contact per A/B dimer is possible due to steric constraints, the RNA must be positioned asymmetrically in the polyhedron in such a way that it contacts each vertex precisely once. Mathematically, this implies a scenario describable as a Hamiltonian path on the polyhedron representing the averaged density, that is, a connected path that visits every vertex precisely once.[47] There are 40,678[28] such paths for this polyhedron, and an example is shown in Fig. 4d, but only 66 of these are circular and hence consistent with the fact that both 5′ and 3′ ends of the genomic RNA are in contact with the single copy of the phage maturation protein. From a topological point of view, this circularizes the RNA, and it is reasonable to assume that the maturation protein in GA plays a similar role (amino acid sequence identity between the MS2 and GA maturation proteins is 46.7%). The Hamiltonian paths for this polyhedron are characterized by a unique sequence of long (orange) and short (yellow) edges (see Fig. 4d). Each such scenario encodes a possible organization of the RNA, consistent with both the cryo-EM density and the functional roles of the PSs in dimer switching at the A/B sites. The importance of this result is that it provides a finite number of options for the ways in which the putative PSs can be mapped into the cryo-EM density. In particular, the problem of mapping PSs into the cryo-EM density is now translated into the problem of checking which (if any) of these Hamiltonian path organizations is consistent with (appropriate subsets of) the ensemble of 53 {54} PS candidates determined earlier.

If the putative PSs identified above do describe a Hamiltonian path for the genome, then their distribution must reflect this characteristic pattern of short and long edges. Short edges represent distances between the PSs in contact with A/B dimers around the same 5-fold vertex and long edges between those at adjacent 5-fold (which traverse the C/C dimer at a 2-fold position). Using the crystal structure of an MS2 capsid containing multiple copies of TR bound at every CP dimer (Protein Data Bank ID 1ZDH), we estimate that a spacer of at least 15 nt between two PSs is required for them to be able to contact A/B dimers at different 5-fold axes. This structural constraint allows us to identify clusters of adjacent PSs that must be located at the same 5-fold axis, that is, those separated in the sequence by fewer nucleotides than this. A direct comparison of all these clusters with all the possible 66 circular Hamiltonian paths is not possible for the following reasons: (i) PSs separated by more than 15 nt could potentially merge into larger clusters, that is, be bound at the same 5-fold, if the RNA sequences between them extend into the interior of the capsid, contributing to the inner shell of RNA observed by cryo-EM that encompasses ~35% of the genome[38]; (ii) other stem–loops from the set of 118 {112} could potentially occupy the remaining 7 {6} unidentified PS sites in contact with the vertices.

As a necessary condition for a match between Hamiltonian path organization and SL distribution, we tested the alignment of a specific connected genome fragment (nucleotides 156–1746 in the MS2 genome) containing few unassigned PSs (i.e., PS from the set of 118 possible, excluding the 53 previously assigned SLs) to fragments of all 66 Hamiltonian paths. For this comparison, it was necessary to group the PSs into clusters of 2, 3 and 5 SLs. This is because all the circular Hamiltonian paths have two, three or five consecutive vertices separated by short edges and hence only such groupings of PS clusters can match to a Hamiltonian path, see Fig. 5a. For the genome fragment under consideration, we determined all possible groupings of PSs such that all assigned PSs form part of a cluster, while any number of the unassigned ones may be contained in a cluster. As illustrated in Supplementary Fig. 4 (Supplementary Fig. 5 for GA), there are 126 {63} options for the clustering of these 25 predicted PSs and the 3 potential additional stem–loops in this region. Two examples are illustrated schematically in Fig. 5b, with broken lines indicating the positions of the three possible additional stem–loops. Stem–loops indicated in the black box in the middle represent the data before clustering, with red boxes identifying stem–loops that must belong to a cluster due to their proximity. Two examples (out of the 126 possible ones) of grouping these further into clusters of 2, 3 and 5 are shown above and below the black box in Fig. 5b. We compared all 126 options numerically with all subsets of all 66 Hamiltonian paths. Strikingly, only the solution shown above the box in Fig. 5 showed a match, implying that there is a single Hamiltonian path that is compatible with the locations of these putative PSs. We checked that this solution is also consistent with the PS clustering across the remainder of the genome and found that it is (see Fig. 3 and Supplementary Fig. 3). A similar analysis for GA PSs reveals exactly the same Hamiltonian path as the unique solution consistent with their distribution, even though their locations within the primary sequence of the GA genome differ significantly from their MS2 equivalents.

The identification of a single Hamiltonian path as a match for the PS distributions is consistent with our previous modeling of the kinetics of assembly of MS2 with RNA fragments encompassing PSs.[28] We examined the statistical significance of this result by assessing the likelihood that 28 PSs randomly clustered in groups of 2, 3 and 5 match a single Hamiltonian path. In order to be consistent with the above analysis, we randomly picked 126 different such combinations of clusters (out of the total 7713 different ways in which 28 PSs could potentially be arranged in terms of clusters of 2, 3 and 5) and assessed the fit of these 126 options with the possible MS2 Hamiltonian paths. This procedure was repeated a million times, that is, for a million randomly chosen combinations of 126 cluster sequences, to obtain the likelihood that a random configuration of 28 PS clusters would fit to a unique

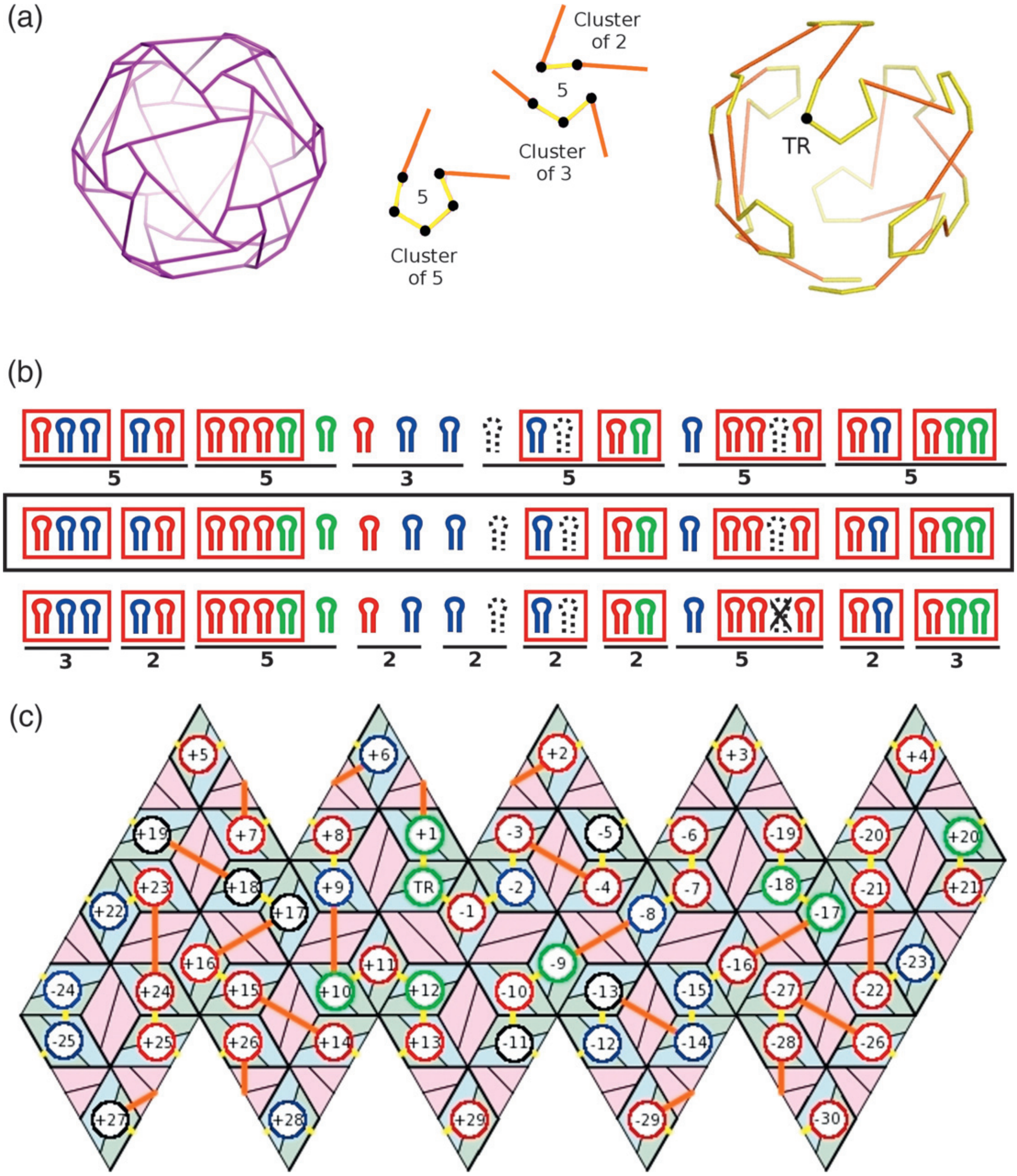

Fig. 5. Prediction of genome organization. Combining the locations of the putative PSs with all possible Hamiltonian path organizations for the RNA in contact with capsid proteins predict the layout of the genome within the virion. (a) The Hamiltonian paths encode the different ways in which the polyhedral outer shell of RNA density seen in icosahedrally averaged cryo-EM reconstructions (purple polyhedron, left) can be realized asymmetrically by the RNA molecule (1 of the 40,678 possibilities is shown, together with the location of the TR PS). In a circular Hamiltonian path arrangement, the 60 PS are located in groups of 2, 3 or 5 around the particle 5-fold axes, and an example of each is shown in (a). (b) Combinatorial analysis of 25 out of 53 MS2 PSs, color-coded by their predicted CP affinity (black box, middle): PSs separated by ≤ 15 nt must be restricted to the same 5-fold vertex (red boxes) and can be grouped into clusters of 2, 3 and 5 stem–loops (black underlines and numbers) allowing 126 different combinations. Two of these are shown above and below the black box; the top one is the only combination consistent with a circular Hamiltonian path. (c) The location of this unique path with respect to a net representing the capsid organization (shown as one of the two possible equivalent representations). The Hamiltonian path predicts the locations of individual PSs in contact with A/B CP dimers. Numbering (0) starts at the assembly initiation site TR, with positive and negative numbers indicating PSs toward the 3′ and 5′ ends, respectively.

Hamiltonian path. A match to a unique Hamiltonian path occurs only 0.8% of the time, with the overwhelming majority of clustering sequences fitting to more than one path. Given that both MS2 and GA exhibit the same Hamiltonian path organization, we moreover computed the likelihood that two sets of 126 cluster sequences, representing two different PS configurations, would both match to a single identical path. Remarkably, out of the million PS configurations sampled, none were found that matched only a single path, implying that the match of the predicted PSs for two phage genomes is highly significant. This suggests that the identified stem–loops are functionally important and that this analysis has provided novel insights into the assembly of these virions.

Discussion

PSs can play important cooperative roles in ssRNA virus assembly, and it is therefore important to develop methods that allow identification of them. The Hamiltonian path method presented here is applicable to all ssRNA viruses for which information on the structure of the RNA density in proximity to capsid is known, for example, from cryo-EM, and for which information regarding RNA–CP affinities is available, for example, via SELEX. We have demonstrated this method here explicitly for two viruses, MS2 and GA, and have illustrated that the nature of the PS distribution provides important insights into the assembly of these particles.

The nature and roles of genomic RNA PSs

Many ssRNA viral capsids and nucleocapsids have been shown to assemble spontaneously *in vitro* into structures resembling the virions recovered from host cells. In the presence of RNA, these reactions occur despite the fact that the RNA requires confinement to a high packing density.[48] Favorable interactions between viral CPs and between CPs and the RNA must enable this entropically costly confinement. In principle, this could arise via a purely electrostatic mechanism, but this cannot account for the two-stage assembly mechanism seen for two viruses from bacterial and plant hosts in smFCS assays.[10] One of those viruses, STNV, has a CP that encompasses a positively charged N-terminal arm, a common feature in many viruses, which has been assumed to interact non-sequence-specifically with RNAs. STNV behaves similarly to MS2, where it is known that, for the interaction with the highest-affinity PS to form a TR: CP_2 complex, over 80% of the binding energy is non-electrostatic.[49] PSs composed of multiple sequence/structure motifs throughout a viral genome making favorable contacts to the RNA are an alternative mechanism to regulating assembly. Modeling the consequences of CP–PS interactions[13] shows that RNAs presenting PSs of differing affinity for CP do better in capsid assembly reactions than do polymers that contain PSs of a uniform affinity. Such a difference would be seized upon by evolution to improve the efficiency of viral life cycles by stabilizing sequences that act as PSs. Appropriate positioning of PSs throughout a genome could then be used to facilitate the correct protein–protein interactions required to form capsids at the low concentrations of viral CP found *in vivo* during the early stages of the formation of progeny viruses.[10]

One outcome of the evolutionary constraint implied by PSs is that viral genomes might have simple repeated sequence motifs. This is obviously not the case, but the efficiency of assembly is just one of many selective pressures on a viral genome. The gene products must also be encoded, and regulatory regions for ribosome and replicase binding must also be formed. As a result, the PS motifs are likely to be quite short, reducing their uniqueness. This would not matter if the PSs act collectively; for example, many cellular RNAs could present single copies of a particular PS, but CPs would not bind stably to such sites unlike on cognate viral RNAs where they would be rapidly incorporated into larger complexes via additional contacts between CPs. The presence of multiple CP–PS complexes in virions might also be seen in their structures. Satellite tobacco mosaic virus is a virus in the same class as STNV. It is notable because X-ray structures reveal that large fractions of its packaged genome exist as stem–loops positioned along the $T = 1$ capsid's 2-fold axes.[50] This has lead to attempts to identify which genomic sequences form these stem–loops (=PSs) using structure probing,[51] and a molecular model of the virus has been built based on these predictions.[52] That is, the assumption has been made that the encapsidated RNA has a defined structure. The X-ray structure of the comovirus bean pod mottle virus also reveals an icosahedrally ordered short segment of ssRNA in the electron density map.[53,54] This again implies that a significant fraction of the viral genome is in contact with the protein shell in every viral particle. Similarly, in Pariacoto virus, substantial sections of the genomic RNA are icosahedrally ordered, forming a dodecahedral cage within the icosahedral protein shell.[55,56] Such highly ordered RNA conformations within virions are relatively rare[54] since most X-ray structures do not show density for the genome, but a large number of cryo-EM reconstructions of ssRNA viruses that do show layers of RNA adjacent to the inner surfaces of protein capsids are now available.[57,58] (Note that the different results from X-ray and cryo-EM reflect the differing amounts of data at different resolutions that each technique records and uses for structure determination.[59])

These results can all be interpreted in terms of the PS hypothesis. It has been explicitly suggested for satellite tobacco mosaic virus that the sections of ordered RNA visible in the electron density map play roles as PSs, although in that case, it has not been possible to demonstrate this directly.[60] For STNV and the RNA phages, *in vitro* reassembly assays with RNA fragments encompassing the respective putative PSs have shown that they will preferentially trigger assembly. The molecular mechanisms of these effects differ in the different viral types. For STNV, the PSs appear to overcome electrostatic repulsion between CP monomers,[15] whereas in MS2 and presumably GA, the PSs determine the locations of the A/B dimers within capsids.[22–24,26] PSs might therefore be a common solution that viruses have found to the generic problem of ensuring that they preferentially package their genomes in the presence of competing cellular RNAs.[61] Animal viruses that package only nascent genomes, such as poliovirus[62] or assemble adjacent to the RNA-dependent RNA polymerase, could also use PSs, but ones that only form through kinetic folding of the RNA transcript. The precise molecular effects of CP–PS interaction may differ in each case, dependent on the CP architecture and the sequence/structure of the PS, but have the common outcome of promoting faithful capsid assembly. For the large number of viruses, including viruses infecting plants, animals and humans,[63–65] with known high-affinity CP binding sites (PSs), lower-affinity variants of such sites are also likely to exist within their genomes. Such high-affinity sites would not be required in a purely electrostatic mechanism of assembly, but once they are present, the selective advantages in favor of multiple sites may render their functioning in this way unavoidable.

Even though many of the PSs identified for MS2 and GA are predicted to have low intrinsic affinities for their respective CPs, they may still be significant in context of the genomes displaying suitably disposed high-affinity sites due to their locally high concentration and the further binding energy from formation of CP–CP contacts. For MS2 at least, there is good evidence that sites other than TR can influence/promote assembly. Mutation of the stem–loop binding site within the phage CP leads to non-assembling phenotypes *in vivo*, but mutants of the TR site yield normal levels of phage.[25] The effects on reassembly efficiency of RNAs carrying multiple copies of TR show that additional sites increase assembly rates and yields,[20,21] as does the use of fragments encompassing natural extensions to the TR site itself.[23] For reassembly with longer RNAs, it has been shown that TR is important to ensure assembly around shorter fragments but becomes progressively less vital as the size of the RNA to be packaged increases.[20,21] This is consistent with the idea that many stem–loops can behave as PSs and is the only conclusion possible from the smFCS assays showing that genomic fragments and the MS2 genome undergo hydrodynamic collapse in the first stage of assembly. RNA collapse can also be induced by addition of multivalent cations, but such RNAs are poor substrates for assembly, implying that cation- and CP-induced collapses are not the same process, the latter being dependent on the RNA sequence, that is, on PSs.[66]

PSs and the conformation of genomes in virions

For MS2 and GA, PSs fall into two groups: those predicted to have relatively high affinities for their respective CPs, and those whose affinities are predicted to be extremely low in isolation but would likely be bound in the context of the packaged genome. Many of the cognate high-affinity class of PSs have been reported to be present in the solution structures of MS2 and GA RNA. This is consistent with the results of smFCS experiments,[10] in particular, the speed of the first stage collapse is in line with CP binding to preexisting PSs. For PS not present in solution, we have checked that the cost of local refolding is small compared with the binding energy from PS–CP contact, making PS-mediated compaction a distinct possibility.

For MS2, formation of a capsid of the correct size and symmetry requires the ordered switching from symmetric to asymmetric protein dimers, triggered via binding of the RNA PSs. The clustering of PSs detected for both MS2 and GA enable us to show that only 1 of the possible 66 circular Hamiltonian paths is compatible with the data despite the fact that the sequences and locations of the PSs are different for these two viruses. This result implies evolutionary conservation of the overall assembly mechanism because a Hamiltonian path defines the nature of the assembly pathway and of the assembly intermediates and kinetics.[28] The fact that a unique pathway is selected in both cases must mean that this pathway is advantageous to these viruses. Inspection of the geometry of the assembly intermediates on this pathway shows that, after assembly initiation at TR, which is located roughly in the middle of the genome, assembly proceeds so that the 5′ and 3′ ends of the RNA build up two hemispheres of the particle independently,[28] hence avoiding steric clashes of both ends during assembly. Indeed, both ends of the RNA bind to the single copy of maturation protein and so never come into direct contact with each other. We have shown by modeling that PS–CP interactions contribute favorably to efficient build-up of a viral capsids.[13] Viruses face the problem of having a plethora of options to assemble their capsid proteins and exploring all of these would slow down the assembly process dramatically, akin to Levinthal's paradox in protein folding.[67] PS–CP interactions are one way of avoiding this problem and biasing assembly to a specific pathway, hence gaining a vital advantage over the host's anti-viral defenses by speeding up assembly.

A specific assembly pathway, however, in turn implies that the conformation of the genomic RNA within every virus particle must be the same. This is at first glance surprising, given that a unique conformation of the RNA in contact with the CP layer is entropically costly, and must therefore be compensated for in another way. However, the advantage of overcoming the virus assembly analog of Levinthal's paradox as explained above is a clear advantage that would certainly outweigh this cost. Moreover, a PS-mediated mechanism provides an added benefit in the initiation phase by defining the starting point of assembly at the PS with highest affinity to capsid protein. The PS with the highest affinity for CP in MS2 (TR) has about double the affinity of the next highest site, which is bound by the adjacent dimer in the preferred assembly pathway. This is also true for GA although the affinity difference has not been determined. In both cases, this combination of PSs marks the starting point of the assembly process. Such precise control of assembly initiation is not achievable in an electrostatic assembly mechanism because it relies on stochastic initiation. In contrast, PS–CP contacts define the pathways of assembly, ensuring the most efficient build-up of the container via cooperative RNA–CP contacts, akin to crystallization. The great fidelity of *in vivo* capsid assembly[9] would emerge naturally from such a mechanism.

These conclusions are corroborated by the recently determined asymmetric structure of MS2 bound to its normal cellular receptor, the bacterial pilus.[68] In this study, pilus fragments were used to bind MS2 virions via their unique maturation proteins. This creates an asymmetric complex, a pilus decorated along its sides with phage that can be used for a tomographic reconstruction without symmetry averaging. Individual tomograms have too small a signal-to-noise ratio to interpret directly; thus, several thousand particles were averaged via sub-tomographic averaging. Note that this is an averaging of many asymmetric structures and not an imposition of an assumed symmetry. Although the resulting asymmetric structure is at moderate resolution (~39 Å), the capsid component is clearly mostly icosahedral, allowing higher-resolution data to be used in its interpretation. There is defined density within the protein shell that must belong to the phage RNA and would not be visible without its conformation being identical, at this resolution, in every particle in the data set. Icosahedral averaging of the tomogram RNA density reproduces the concentric double shell of density seen in averaged cryo-EM reconstructions, suggesting that the asymmetric structure is a true image of the phage particle. The work described above on the identification of PSs and their implications for RNA structure within virions via the DSM and Hamiltonian path neatly explains this independently derived structure. None of these observations means that RNA-free CP assembly, assembly with sub-genomic or non-cognate RNA fragments will not occur. At artificially high CP concentrations, all these reactions will occur, but without the accuracy and speed of the pathway based on PSs.

The conserved assembly mechanisms described here for two RNA phages and their implications for the structure of their virions open up radically new insights into ssRNA virus biology. PS-dependent assembly mechanisms may be much more common than widely realized as the work described here suggests. Revealing these mechanisms has only been possible via a unique interaction of theory with experimental observation, and similar approaches to other viral systems may yet uncover even more similarities, making the roles of putative PSs in assisting virion assembly *in vivo* tangible novel anti-viral drug targets.[22]

Materials and Methods

Computational identification of stem–loops potentially able to bind CP

We determined the locations of all nucleotide sequences in the phage genomes consistent with formation of stem–loops with a 3- or 4-nt loop motif corresponding to the search pattern shown in Fig. 2a. To do this, a window of 3 or 4 nt is slid across the genomic sequence in increments of 1 nt, testing in each case whether the flanking nucleotides can (Watson–Crick) base pair. From the list of 2087 stem–loops determined in this way, we have retained those that are predicted to be stable using UNAFold,[42] resulting in a pruned list of 623 stem–loops. As these contain overlapping stem–loops, this list was further pruned to identify the most stable, non-overlapping stem–loops with a minimum of 5 bp and a maximum of 7 bp. The 7-bp cutoff reflects the length of the 19-nt-long fragment TR, the known high-affinity site in the MS2 genome. As experiments with extensions of TR by 2 bp (cf. the TR-clamp in Ref. 27) show, longer fragments may lead to steric clashes. This resulted in a total of 118 stem–loops for MS2 and 112 stem–loops for GA.

Construction of the scoring matrix

The scoring matrix in Table 1 was constructed using measurements of the binding affinities of a variety of TR stem–loops to MS2[20,21,29–34] and GA CP[18] based on single-nucleotide variations. Binding affinities were measured in terms of the association constant K_a using filter-binding assays and stopped-flow kinetics. We normalized these K_a values to that of the wild-type TR and rounded to nearest powers of 10. Values of 1 in the table hence correspond to mutants with affinities similar to TR.

Table 1. The scoring matrix

	A	G	C	U
−7	1	0.1	0.1	0.1
−6	1	1	1	1
−5	0.01	0.01	6	1
−4	1	0.001	0.001	0.001

The scoring matrix provides a qualitative ranking of the binding affinities of the stem–loops to capsid protein relative to the highest-affinity stem–loop in the genomic sequences, for example, in MS2, the PS TR (here normalized to 1). The values in this matrix are based on the effects of single-nucleotide variations in the TR sequence on the CP affinity.[18,29–34] These incorporate the effects of individual nucleotides at positions −4 to −7 in the stem–loops. Note that the effects of many multiple nucleotide changes within TR have not been determined experimentally. Here, the effects of single substitutions are assumed to be multiplicative on the overall affinity. A similar matrix for GA[18] can be derived from the MS2 matrix above by swapping the scores between A and C and between G and U in the −5 row.

We have checked that the values in this table qualitatively reproduce the critical recognition features (the adenosine in position −4 and pyrimidine in position −5; cf. Fig. 2b) and hence provide a suitable representation of the affinity of a general stem–loop relative to wild-type TR.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jmb.2013.06.005

Acknowledgements

We thank Prof. Olsthoorn (University of Leiden), Prof. Groeneveld (University of Utrecht) and Dr. Beekwilder for sending us copies of their Ph.D. theses detailing aspects of their work on the RNA phages and Prof. Neil Ranson (University of Leeds) for careful reading of the manuscript and helpful comments. R.T. and E.C.D. also thank the Leverhulme Trust for funding via a Research Leadership Award.

Received 15 April 2013;
Received in revised form 22 May 2013;
Accepted 3 June 2013
Available online 11 June 2013

Keywords:
packaging signal;
viral genome organization;
RNA SELEX;
Hamiltonian path;
virus assembly

Abbreviations used:
ssRNA, single-stranded RNA; PS, packaging signal; CP, coat protein; smFCS, single-molecule fluorescence correlation spectroscopy; STNV, satellite tobacco necrosis virus; DSM, Dimer Switching Model; cryo-EM, cryo-electron microscopy.

References

1. van der Schoot, P. & Bruinsma, R. (2005). Electrostatics and the assembly of an RNA virus. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **71**, 061928–061939.
2. Belyi, V. A. & Muthukumar, M. (2006). Electrostatic origin of the genome packing in viruses. *Proc. Natl Acad. Sci. USA*, **103**, 7174–17178.
3. Comas-Garcia, M., Cadena-Nava, R. D., Rao, A. L., Knobler, C. M. & Gelbart, W. M. (2012). *In vitro* quantification of the relative packaging efficiencies of single-stranded RNA molecules by viral capsid protein. *J. Virol.* **86**, 12271–12282.
4. Cadena-Nava, R. D., Comas-Garcia, M., Garmann, R. F., Rao, A. L., Knobler, C. M. & Gelbart, W. M. (2012). Self-assembly of viral capsid protein and RNA molecules of different sizes: requirement for a specific high protein/RNA mass ratio. *J. Virol.* **86**, 3318–3326.
5. Bancroft, J. B., Hills, G. J. & Markham, R. (1967). A study of the self-assembly process in a small spherical virus. Formation of organised structures from protein subunits *in vitro*. *Virology*, **31**, 354–379.
6. Bancroft, J. B., Hiebert, E. & Bracker, C. E. (1969). The effects of various polyions on shell formation of spherical viruses. *Virology*, **39**, 924–930.
7. Kivenson, A. & Hagan, M. F. (2010). Mechanisms of capsid assembly around a polymer. *Biophys. J.* **99**, 619.
8. Elrad, O. M. & Hagan, M. F. (2010). Encapsulation of a polymer by an icosahedral virus. *Phys. Biol.* **7**, 045003.
9. Routh, A., Domitrovic, T. & Johnson, J. E. (2012). Host RNAs, including transposons, are encapsidated by a eukaryotic single-stranded RNA virus. *Proc. Natl Acad. Sci. USA*, **109**, 1907–1912.
10. Borodavka, O., Tuma, R. & Stockley, P. G. (2012). Evidence that viral RNA has evolved for efficient, two-stage packaging. *Proc. Natl Acad. Sci. USA*, **109**, 15769–15774.
11. Golmohammadi, R., Valegård, K., Fridborg, K. & Liljas, L. (1993). The refined structure of bacteriophage MS2 at 2.8 Å resolution. *J. Mol. Biol.* **234**, 620–639.
12. Jones, T. A. & Liljas, L. (1984). Structure of satellite tobacco necrosis virus after crystallographic refinement at 2.5 Å resolution. *J. Mol. Biol.* **177**, 735–767.
13. Dykeman, E. C., Stockley, P. G. & Twarock, R. (2013). How to build a viral capsid in the presence of genomic RNA. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **87**, 022717.
14. Bunka, D. H., Lane, S. W., Lane, C. L., Dykeman, E. C., Ford, R. J., Barker, A. M. *et al.* (2011). Degenerate RNA packaging signals in the genome of satellite tobacco necrosis virus: implication for the assembly of a T = 1 capsid. *J. Mol. Biol.* **413**, 51–65.
15. Ford, R. J., Barker, A. M., Bakker, S. E., Coutts, R. H., Ranson, N. A., Phillips, S. E. V. *et al.* (2013). Sequence-specific, RNA–protein interactions overcome electrostatic barriers preventing assembly of Satellite Tobacco Necrosis Virus coat protein. *J. Mol. Biol.* **425**, 1050–1064.
16. Liljas, L., Unge, T., Jones, T. A., Fridborg, K., Lövgren, S., Skoglund, U. & Strandberg, B. (1982). Structure of

satellite tobacco necrosis virus at 3.0 Å resolution. *J. Mol. Biol.* **159**, 93–108.

17. Lane, S. W., Dennis, C. A., Lane, C. L., Trinh, C. H., Rizkallah, P. J., Stockley, P. G. & Phillips, S. E. V. (2011). Construction and crystal structure of recombinant STNV capsids. *J. Mol. Biol.* **413**, 41–50.
18. Gott, J. M., Wilhelm, L. J. & Uhlenbeck, O. C. (1991). RNA binding properties of the coat protein from bacteriophage GA. *Nucleic Acids Res.* **19**, 6499–6503.
19. Lodish, H. F. & Zinder, N. D. (1966). Mutants of the bacteriophage f2. VIII. Control mechanisms for phage-specific syntheses. *J. Mol. Biol.* **19**, 333–348.
20. Beckett, D., Wu, H. N. & Uhlenbeck, O. C. (1988). Roles of operator and non-operator RNA sequences in bacteriophage R17 capsid assembly. *J. Mol. Biol.* **204**, 939–947.
21. Beckett, D. & Uhlenbeck, O. C. (1988). Ribonucleoprotein complexes of R17 coat protein and a translational operator analog. *J. Mol. Biol.* **204**, 927–938.
22. Stockley, P. G., Rolfsson, O., Thompson, G. S., Basnak, G., Francese, S., Stonehouse, N. J. *et al.* (2007). A simple, RNA-mediated allosteric switch controls the pathway to formation of a T = 3 viral capsid. *J. Mol. Biol.* **369**, 541–552.
23. Basnak, G., Morton, V. L., Rolfsson, O., Stonehouse, N. J., Ashcroft, A. E. & Stockley, P. G. (2010). Viral genomic single-stranded RNA directs the pathway toward a T = 3 capsid. *J. Mol. Biol.* **395**, 924–936.
24. Rolfsson, O., Toropova, K., Ranson, N. A. & Stockley, P. G. (2010). Mutually-induced conformational switching of RNA and coat protein underpins efficient assembly of a viral capsid. *J. Mol. Biol.* **401**, 309–322.
25. Peabody, D. S. (1997). Role of the coat protein–RNA interaction in the life cycle of bacteriophage MS2. *Mol. Gen. Genet.* **254**, 358–364.
26. Dykeman, E. C., Stockley, P. G. & Twarock, R. (2010). Dynamic allostery controls coat protein conformer switching during MS2 phage assembly. *J. Mol. Biol.* **395**, 916–923.
27. Morton, V. L., Dykeman, E. C., Stonehouse, N. J., Ashcroft, A. E., Twarock, R. & Stockley, P. G. (2010). The impact of viral RNA on assembly pathway selection. *J. Mol. Biol.* **401**, 298–308.
28. Dykeman, E. C., Grayson, N. E., Toropova, K., Ranson, N., Stockley, P. G. & Twarock, R. (2011). Simple rules for efficient assembly predict the layout of a packaged viral RNA. *J. Mol. Biol.* **408**, 399–407.
29. Grahn, E., Stonehouse, N. J., Murray, J. B., van den Worm, S., Valegård, K., Fridborg, K. *et al.* (1999). Crystallographic studies of RNA hairpins in complexes with recombinant MS2 capsids: implications for binding requirements. *RNA*, **5**, 131–138.
30. Grahn, E., Moss, T., Helgstrand, C., Fridborg, K., Sundaram, M., Tars, K. *et al.* (2001). Structural basis of pyrimidine specificity in the MS2 RNA hairpin–coat-protein complex. *RNA*, **7**, 1616–1627.
31. Hirao, I., Spingola, M., Peabody, D. & Ellington, A. D. (1999). The limits of specificity: an experimental analysis with RNA aptamers to MS2 coat protein variants. *Mol. Diversity*, **4**, 75–89.
32. Lago, H., Fonseca, S. A., Murray, J. B., Stonehouse, N. J. & Stockley, P. G. (1998). Dissecting the key recognition features of the MS2 bacteriophage translational repression complex. *Nucleic Acids Res.* **26**, 1337–1344.
33. Convery, M. A., Rowsell, S., Stonehouse, N. J., Ellington, A. D., Hirao, I., Murray, J. B. *et al.* (1998). Crystal structure of an RNA aptamer–protein complex at 2.8 Å resolution. *Nat. Struct. Biol.* **5**, 133–139.
34. Rowsell, S., Stonehouse, N. J., Convery, M. A., Adams, C. J., Ellington, A. D., Hirao, I. *et al.* (1998). Crystal structures of a series of RNA aptamers complexed to the same protein target. *Nat. Struct. Biol.* **5**, 970–975.
35. Van Duin, J. & Tsareva, N. (2006). Single-stranded RNA phages. In *The Bacteriophages* (Calendar, R., ed.), pp. 175–196, 2nd edit. Oxford University Press, New York, NY.
36. Groeneveld, H. (1997). Secondary structure of bacteriophage MS2 RNA: translational control by kinetics of RNA folding. Ph.D. Thesis, University of Leidun.
37. Olsthoorn, R. C. L. (1996). Structure and evolution of RNA phages. Ph.D. Thesis, University of Leiden.
38. Toropova, K., Basnak, G., Twarock, R., Stockley, P. G. & Ranson, N. A. (2008). The three-dimensional structure of genomic RNA in bacteriophage MS2: implications for assembly. *J. Mol. Biol.* **375**, 824–836.
39. Van den Worm, S. H., Koning, R. I., Warmenhoven, H. J., Koerten, H. K. & van Duin, J. (2006). Cryo electron microscopy reconstructions of the Leviviridae unveil the densest icosahedral RNA packing possible. *J. Mol. Biol.* **363**, 858–865.
40. Lago, H., Parrott, A. M., Moss, T., Stonehouse, N. J. & Stockley, P. G. (2001). Probing the kinetics of formation of the bacteriophage MS2 translational operator complex: identification of a protein conformer unable to bind RNA. *J. Mol. Biol.* **305**, 1131–1144.
41. Beekwilder, J. (1997). Secondary structure of the RNA genome of bacteriophage Qβ. Ph.D. Thesis, University of Leiden.
42. Markham, N. R. & Zuker, M. (2005). DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.* **33**, W577–W581.
43. Gopal, A., Zhou, Z. H., Knobler, C. M. & Gelbart, W. M. (2012). Visualizing large RNA molecules in solution. *RNA*, **18**, 284–299.
44. Yoffe, A. M., Prinsen, P., Gopal, A., Knobler, C. M., Gelbart, W. M. & Ben-Shaul, A. (2008). Predicting the sizes of large RNA molecules. *Proc. Natl Acad. Sci. USA*, **105**, 16153–16158.
45. Wilkinson, K. A., Gorelick, R. J., Vasa, S. M., Guex, N., Rein, A., Mathews, D. H. *et al.* (2008). High--throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biol.* **6**, e96.
46. Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., Bess, J. W., Swanstrom, R. *et al.* (2009). Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, **460**, 711–716.
47. Hamilton, W. R. (1858). An account of the Icosian calculus. *Proc. R. Ir. Acad.* **6**, 415–416.
48. Speir, J. A. & Johnson, J. E. (2012). Nucleic acid packaging in viruses. *Curr. Opin. Struct. Biol.* **22**, 65–71.
49. Carey, J. & Uhlenbeck, O. C. (1983). Kinetic and thermodynamic characterization of the R17 coat

protein–ribonucleic acid interaction. *Biochemistry*, **22**, 2610–2615.

50. Larson, S. B., Koszelak, S., Day, J., Greenwood, A., Dodds, J. A. & McPherson, A. (1993). Double-helical RNA in satellite tobacco mosaic virus. *Nature*, **361**, 179.
51. Schroeder, S. J., Stone, J. W., Bleckley, S., Gibbons, T. & Mathews, D. M. (2011). Ensemble of secondary structures for encapsidated satellite tobacco mosaic virus RNA consistent with chemical probing and crystallography constraints. *Biophys. J.* **101**, 167–175.
52. Zeng, Y., Larson, S. B., Heitsch, C. E., McPherson, A. & Harvey, S. C. (2012). A model for the structure of satellite tobacco mosaic virus. *J. Struct. Biol.* **180**, 110–116.
53. Lin, T., Cavarelli, J. & Johnson, J. E. (2003). Evidence for assembly-dependent folding of protein and RNA in an icosahedral virus. *Virology*, **314**, 26–33.
54. Schneemann, A. (2006). The structural and functional role of RNA in icosahedral virus assembly. *Annu. Rev. Microbiol.* **60**, 51–67.
55. Tang, L., Johnson, K. N., Ball, L. A., Lin, T., Yeager, M. & Johnson, J. E. (2001). The structure of Pariacoto virus reveals a dodecahedral cage of duplex RNA. *Nat. Struct. Biol.* **8**, 77–83.
56. Rudnick, J. & Bruinsma, R. (2005). Icosahedral packing of RNA viral genomes. *Phys. Rev. Lett.* **94**, 038101.
57. Toropova, K., Stockley, P. G. & Ranson, N. A. (2011). Visualising a viral RNA genome poised for release from its receptor complex. *J. Mol. Biol.* **408**, 408–419.
58. Bakker, S. E., Ford, R. J., Barker, A. M., Robottom, J., Saunders, K., Pearson, A. R. *et al.* (2012). Isolation of an asymmetric RNA uncoating intermediate for a single-stranded RNA plant virus. *J. Mol. Biol.* **417**, 65–78.
59. Tsuruta, H., Reddy, V. S., Wikoff, W. R. & Johnson, J. E. (1998). Imaging RNA and dynamic protein segments with low-resolution virus crystallography: experimental design, data processing and implications of electron density maps. *J. Mol. Biol.* **284**, 1439–1452.
60. Larson, S. B. & McPherson, A. (2001). Satellite tobacco mosaic virus RNA: structure and implications for assembly. *Curr. Opin. Struct. Biol.* **11**, 59–65.
61. Harrison, S. C., Olson, A. J., Schutt, C. E., Winkler, F. K. & Bricogne, G. (1978). Tomato bushy stunt virus at 2.9 Å resolution. *Nature*, **276**, 368–373.
62. Nugent, C. I., Johnson, K. L., Sarnow, P. & Kirkegaard, K. (1999). Functional coupling between replication and packaging of poliovirus replicon RNA. *J. Virol.* **73**, 427–435.
63. Qu, F. & Morris, T. J. (1997). Encapsidation of turnip crinkle virus is defined by a specific packaging signal and RNA size. *J. Virol.* **71**, 1428–1435.
64. Kim, D. Y., Firth, A. E., Atasheva, S., Frolova, E. I. & Frolov, I. (2011). Conservation of a packaging signal and the viral genome RNA packaging mechanism in alphavirus evolution. *J. Virol.* **85**, 8022–8036.
65. Sasaki, J. & Taniguchi, K. (2003). The 5′-end sequence of the genome of Aichi virus, a picornavirus, contains an element critical for viral RNA encapsidation. *J. Virol.* **77**, 3542–3548.
66. Borodavka, A., Tuma, R. & Stockley, P. G. (2013). A two-stage mechanism of viral RNA compaction revealed by single molecule fluorescence. *RNA Biol.* **10**, 1–9.
67. Levinthal, C. (1969). How to fold graciously. pp. 22–24, University of Illinois Press, Urbana, IL.
68. Dent, K. C., Thompson, R., Barker, A. M., Barr, J. N., Hiscox, J. A., Stockley, P. G. & Ranson, N. A. (2013). The asymmetric structure of an icosahedral virus bound to its receptor suggests a mechanism for genome release. *Structure*. In press. http://dx.doi.org/10.1016/j.str.2013.05.012.