Research Article

Genome-regulated Assembly of a ssRNA Virus May Also Prepare It for Infection

Rebecca Chandler-Bostock [1†], Richard J. Bingham [2†], Sam Clark [2†], Andrew J. P. Scott [1†], Emma Wroblewski [1], Amy Barker [1], Simon J. White [1‡], Eric C. Dykeman [2], Carlos P. Mata [1§], Jen Bohon [3‖], Erik Farquhar [3], Reidun Twarock [*2] and Peter G. Stockley [*1]

*1 - **Astbury Centre for Structural Molecular Biology,** University of Leeds, Leeds LS2 9JT, UK*
*2 - **Departments of Mathematics and Biology & York Cross-Disciplinary Centre for Systems Analysis,** University of York, York, UK*
*3 - **CWRU Center for Synchrotron Biosciences,** NSLS-II, Brookhaven National Laboratory, Upton, NY 11973, USA*

***Correspondence to Reidun Twarock and Peter G. Stockley:** reidun.twarock@york.ac.uk (R. Twarock), p.g.stockley@leeds.ac.uk (P.G. Stockley)*
https://doi.org/10.1016/j.jmb.2022.167797
Edited by M.F. Summers

Abstract

Many single-stranded, positive-sense RNA viruses regulate assembly of their infectious virions by forming multiple, cognate coat protein (CP)-genome contacts at sites termed Packaging Signals (PSs). We have determined the secondary structures of the bacteriophage MS2 ssRNA genome (gRNA) frozen in defined states using constraints from X-ray synchrotron footprinting (XRF). Comparison of the footprints from phage and transcript confirms the presence of multiple PSs in contact with CP dimers in the former. This is also true for a virus-like particle (VLP) assembled around the gRNA *in vitro* in the absence of the single-copy Maturation Protein (MP) found in phage. Since PS folds are present at many sites across gRNA transcripts, it appears that this genome has evolved to facilitate this mechanism of assembly regulation. There are striking differences between the gRNA-CP contacts seen in phage and the VLP, suggesting that the latter are inappropriate surrogates for aspects of phage structure/function. Roughly 50% of potential PS sites in the gRNA are not in contact with the protein shell of phage. However, many of these sit adjacent to, albeit not in contact with, PS-binding sites on CP dimers. We hypothesize that these act as PSs transiently during assembly but subsequently dissociate. Combining the XRF data with PS locations from an asymmetric cryo-EM reconstruction suggests that the genome positions of such dissociations are non-random and may facilitate infection. The loss of many PS-CP interactions towards the 3′ end of the gRNA would allow this part of the genome to transit more easily through the narrow basal body of the pilus extruding machinery. This is the known first step in phage infection. In addition, each PS-CP dissociation event leaves the protein partner trapped in a non-lowest free-energy conformation. This destabilizes the protein shell which must disassemble during infection, further facilitating this stage of the life-cycle.

Introduction

Formation of an infectious virion is the critical event in any viral life-cycle. Virion architecture must therefore be carefully controlled to provide sufficient protection for the genetic cargo during transfer between host cells, whilst enabling its timely release at the start of a new round of infection. The molecular details regulating this life-cycle directionality and timing are still largely obscure for virtually all viruses.

We,[1–13] and now others[14–17] propose that for many families of ssRNA virions assembly is regulated by multiple dispersed RNA sequences/motifs (Packaging Signals, PSs) within the viral genome (gRNA). This is also the case for the para-retrovirus Hepatitis B Virus.[10,13] PSs in many of these genomes consist of RNA stem-loops presenting a cognate coat protein (CP) recognition motif in their loops. Their CP affinities vary because of sequence variation of the often sparse recognition motif presented in secondary structure elements of differing RNA folding propensities.[11] This variation defines a preferred assembly path,[6,11,18] the Hamiltonian path, helping to ensure faithful and efficient assembly in complex environments. The inferred packaging specificity of this mechanism is consistent with experimental outcomes of natural infections,[19] and it also evolves spontaneously in an artificial self-assembling enzyme.[20] A simple analogy is to consider the PS-CP contacts formed as a type of molecular Velcro, directing the formation of, and stabilizing, the virion formed. Too many gRNA-CP contacts, would result in stable virions unable to respond to host cues by gRNA release and/or uncoating. Here we use the model virion, bacteriophage MS2, to address the question of how Nature prevents this outcome. In particular, we investigate the roles that contacts between gRNA and viral proteins play throughout assembly and at the start of infection.

PS-mediated assembly regulation was first identified for bacteriophage MS2.[1,5,21–23] Translational repression of the replicase cistron within its gRNA occurs at the start of phage assembly, switching gRNA function from that of a polycistronic mRNA to an assembly substrate.[24] The molecular basis of this repression involves the sequence-specific binding of a CP dimer (CP_2) to an RNA stem-loop,[25–28] the translational operator (TR), encompassing the start codon of replicase. We showed using mass spectrometry (MS)[1–3] and single-molecule fluorescence correlation spectroscopy (smFCS),[5] that TR-CP_2 interaction induces a protein conformational change. This converts a symmetric (C/C subunits) to an asymmetric (A/B subunits) CP_2 conformation by altering the orientations of the FG-loops of polypeptide joining the F & G β-strands in each subunit. TR-like oligonucleotides act as allosteric effectors.[21,27] Both symmetric and asymmetric protein conformers are required to build a $T = 3$ capsid (Figure 1). Consistent with this observation, *in vitro* both types of dimer need to be present for assembly to occur, suggesting a simple rules-based model for virus assembly. Multiple, dispersed PSs[22–23] in the gRNA each induce conformational changes favouring asymmetry in the incoming CP_2. These proteins form the fivefold axes of the capsid shell, implying that there are up to 60 PS sites across the gRNA. *In vitro* reassembly of a virus-like particle (VLP) around the 3569 nt long gRNA, in the absence of the single copy Maturation Protein (MP), is known to be driven by the formation of both RNA-CP_2 and CP_2-CP_2 contacts,[5] as would be expected from this model. The VLP product of that reaction is indistinguishable from the icosahedrally-averaged X-ray structure of phage.[29–30]

Asymmetric cryo-electron microscopy (cryo-EM)[31–33] reconstructions (Figure 1(b)) allow visualisation of both the MP and gRNA structures in the infectious phage, which are both absent in crystal structures. These include multiple gRNA PS-CP contacts[18] principally at A/B dimers, as expected (Figure 1(b)). Since the TR sequence in the gRNA is unique, these multiple contacts are consistent with the binding of sequence-related stem-loops distributed across the gRNA[18,23] consistent with the assumed assembly mechanism. In the absence of TR, or other gRNA PSs, CP dimers are stable in the C/C conformation.[1] This is consistent with the need for cognate gRNA to stimulate phage capsid assembly. In the presence of gRNA or PSs, assembly is rapid (seconds) both *in vivo* and *in vitro*,[5] compared with CP-alone assembly (days).

The highest-resolution asymmetric cryo-EM reconstruction of MS2,[33] which is mostly at atomic resolution, is unfortunately incomplete across regions of the gRNA, which is described as "flexible". We therefore used the minimally invasive technique of X-ray synchrotron footprinting (XRF)[34–36] to determine the secondary structure of bacteriophage MS2 gRNA, and its molecular contacts *in situ,* i.e. within phage particles. We compared these results with those for a full-length, protein-free gRNA transcript, and for the latter following its *in vitro* reassembly into a VLP lacking MP. The results show that many stem-loops encompassing features of PS sites are present in the gRNA transcript. Many of these PSs (31) remain in contact with the protein shell in infectious phage. This is more than expected on the basis of cryo-EM reconstruction[33] (16). CP-gRNA contacts in the VLP are distinct from those in the phage suggesting that MP and the gRNA PS-CP interactions play multiple roles in both defining assembly initiation, and specifying a protein shell of the correct size and curvature. It appears that programmed PS dissociation events may facilitate genome release from the phage particle during infection.

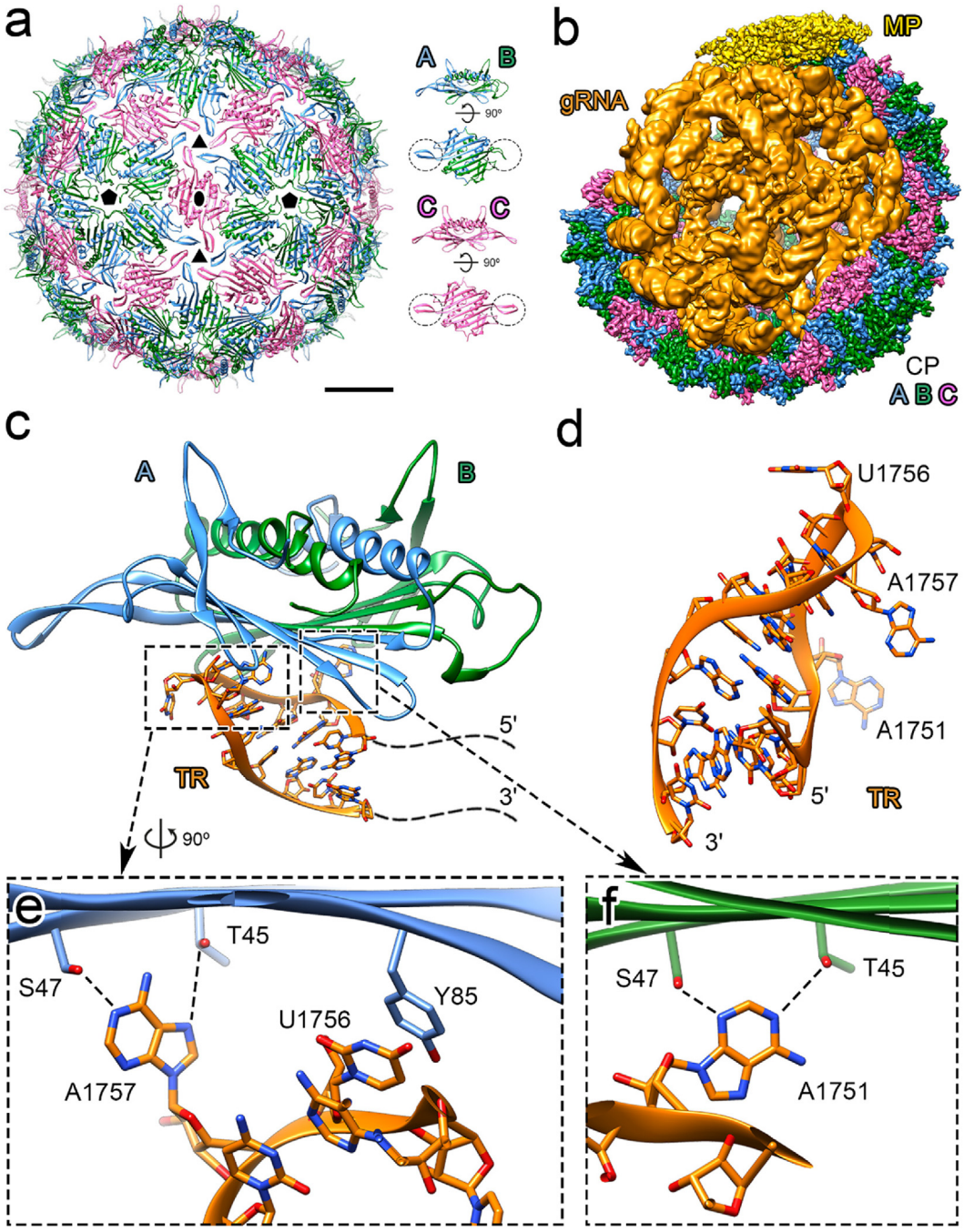

Figure 1. Structures of bacteriophage MS2 & the TR PS complex with a CP_2. (a) Atomic model of the front-half capsid of the crystal structure (PDB: 1ZDH) of MS2 modelled with 180 CP subunits as ribbon diagrams coloured in blue (A subunit), green (B subunit) or pink (C subunit) to indicate the differing quasi-conformers, viewed along a twofold axis.[29–30] To the right, A/B and C/C dimers are shown, as ribbon views from the side and the capsid interior. The dashed circles highlight the positions and conformations of their FG-loops; (b) The cryo-EM density map obtained without imposing icosahedral symmetry (EMD-8397) reveals the asymmetric organization of phage MS2.[31–33] The capsid colour-coding is as in (a), but with one C/C-dimer replaced by the MP (grey). The internal orange structure, viewed with a section of the capsid computationally removed, is a model of the ordered parts of the gRNA in that reconstruction; (c) The crystal structure of the TR oligonucleotide within a VLP in complex with an A/B CP dimer, (PDB:1ZDI). The boxed areas show enlargements of the sequence-specific contacts made between TR and the CP. Extruded nucleotide bases from TR make sequence-specific contacts with the CP_2, as indicated. U1756 (−5) forms a stacking interaction with the side-chain of Tyr58 in A subunits, adenines 1751/1757(−10 and −4) hydrogen bond, in distinct orientations to the side-chains of Thr45 and Ser47 in both A and B subunits. Nucleotides are numbered relative to the gRNA, whilst those in brackets are relative to the first A of the replicase start codon in TR, which is also commonly used.

Results

XRF workflow and choice of viral footprinting targets

MS2 bacteriophage (Figure 1) has one of the first genomes sequenced. Its secondary structure has been inferred previously from combinations of enzymatic and chemical footprinting.[37–39] It is therefore an ideal choice with which to benchmark the XRF of its gRNA. Infectious particles contain single copies of the genome, the Maturation Protein (MP),[31–33] and 89 CP_2, each in one of two distinct, quasi-equivalent conformations. There are 60 asymmetric A/B and 29 symmetric C/C CP_2 positions in its $T = 3$ capsid, with the additional C/C lattice position occupied by MP.[40–42] The B conformation requires that an essential conserved Pro68 residue adopt a *cis* peptide bond, which is *trans* in A & C conformers.[43–44] PS-binding is distal from the side-chain of Pro68[21] and the allosteric effect of PS binding is propagated via its impact on the dynamics of the β-strands attached to the loop.[27]

XRF is very versatile with unparalleled time-resolution (<100 millisecond exposures), allowing its use in analysis of RNA folding and ribosomal assembly processes.[45] It can be used on samples in solution, in the frozen state, or even in living cells.

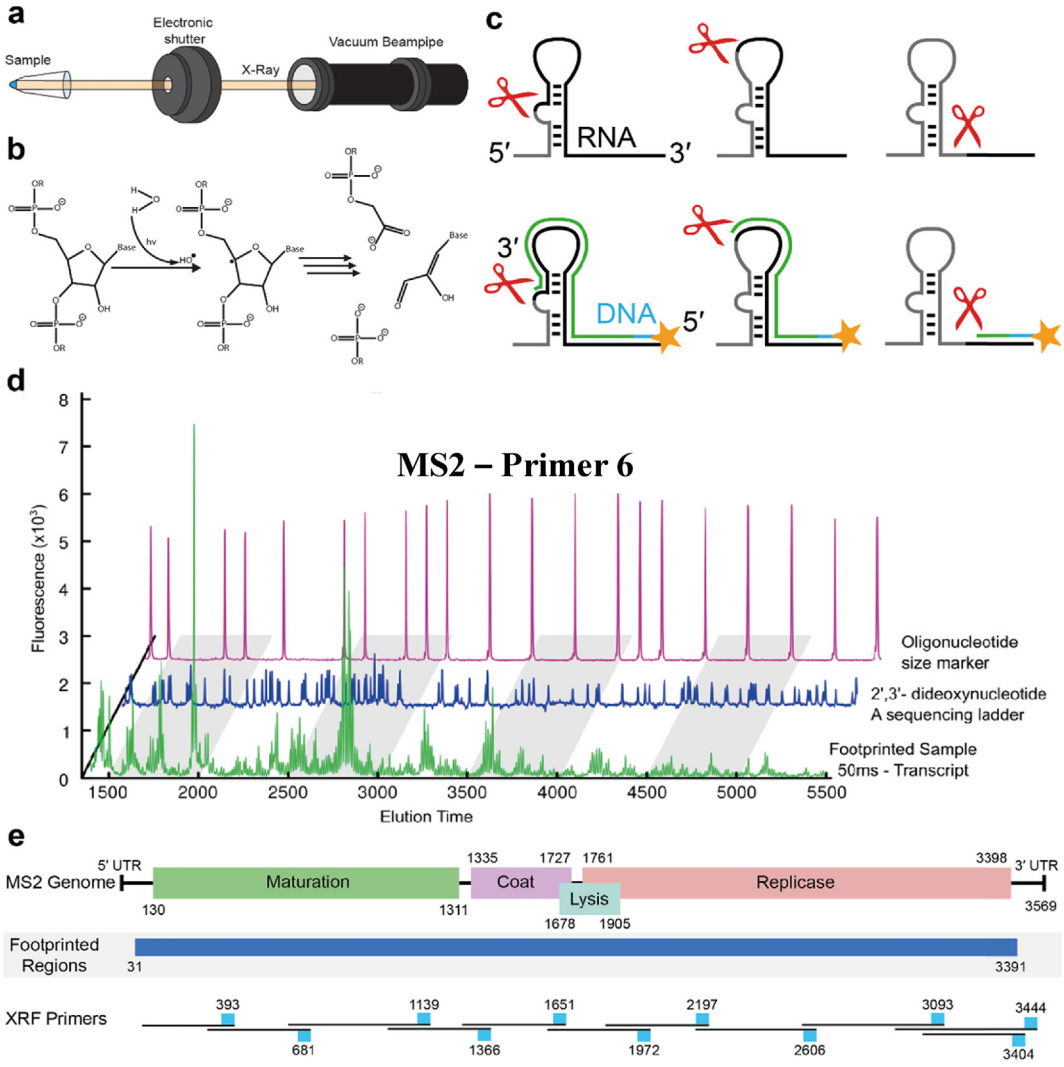

Figure 2. Details of X-ray footprinting analysis of the MS2 gRNA. (a) Cartoon showing the arrangement of XRF samples with respect to the X-ray beam. Samples are spun down in Eppendorf tubes, flash frozen and then placed in matched sample holders for exposure; (b) The assumed principal chemical reaction leading to polynucleotide chain cleavage; (c) A cartoon illustrating the readout of reactivity at each nucleotide by primer extension; (d) Example capillary electrophoresis extension traces for a footprinted sample (green), together with a sequencing ladder (blue) and oligonucleotide size markers (pink); (e) Gene map of the MS2 gRNA (top) together with the region footprinted (middle) that integrates the data from the primer extensions shown (bottom).

Table 1 XRF Details on Primers - reproducibility and normalization factors.

P-ID	Primer annealing position	XRF nucleotide data used	Average PCC			Normalisation Factor		
			T	V	R	T	V	R
1	393–372	31–340	0.95 ± 0.018	0.86 ± 0.067	–	1181 ± 20	1691 ± 87	–
2	681–657	279–628	0.98 ± 0.003	0.90 ± 0.031	0.51 ± 0.204	952 ± 23	737 ± 38	78 ± 7
3	1139–1119	584–1083	0.91 ± 0.034	0.95 ± 0.006	–	713 ± 107	681 ± 65	–
4	1366–1342	966–1315	0.94 ± 0.035	0.84 ± 0.041	–	378 ± 102	979 ± 47	–
5	1651–1627	1251–1600	0.90 ± 0.008	0.95 ± 0.010	–	99 ± 7	1536 ± 130	–
6	1972–1953	1573–1922	0.97 ± 0.012	0.74	0.57	8335 ± 347	185 ± 7	108 ± 7
7	2197–2177	1796–2145	0.80 ± 0.079	0.91 ± 0.036	–	808 ± 93	218 ± 30	–
8	2606–2583	2137–2556	0.83 ± 0.043	0.94 ± 0.027	–	130 ± 2	124 ± 5	–
9	3093–3074	2542–3041	0.94 ± 0.010	0.50 ± 0.046	–	2021 ± 93	189 ± 8	–
10	3404–3383	3003–3352	0.90 ± 0.032	0.97 ± 0.009	–	252 ± 26	109 ± 3	–
11	3444–3426	2892–3391	0.93 ± 0.023	0.51 ± 0.252	–	377 ± 39	53 ± 1	–

Table shows from left to right: primer IDs (P-ID), positions, read areas covered by primer extension, average Pearson correlation coefficients (PCCs) over replicates for different states (T = transcript, V = virion, and R = reassembled/VLP), and normalization factors with standard deviation. The non-normalized per nucleotide reactivities in the region between the first and last size marker of each primer extension product were computed as peak height multiplied by peak width for peaks in the footprinted trace; values shown are averaged over triplicate determinations.

It works by photolysing solvent water molecules, a reaction that accounts for virtually all the absorbed photon energy, creating hydroxyl radicals throughout the sample. RNA ribose moieties are differentially modified on the basis of their flexibility (Figure 2). This leads to phosphodiester backbone cleavages at modified sites, making it similar to other footprinting technologies, such as SHAPE.[46–49] Footprinting frozen samples allows direct comparison of the results with cryo-EM studies. In order to understand the roles of gRNA in CP_2 shell assembly, and the impact of MP on this process, we determined XRF per nucleotide reactivities of the MS2 gRNA in an infectious virion, in a VLP formed only with CP_2, or as a protein-free transcript. Samples for XRF were flash-frozen, shipped on dry-ice to the Brookhaven National Synchrotron Light Source II, Beamline 17-BM,[36] and multiple frozen replicates exposed to X-rays at −30 °C for 10–100 milliseconds. Following exposure, the samples were returned on dry-ice to the host laboratory for further processing. Test exposures and quantitative primer extensions show that avoiding multiple cleavages within the full-length gRNA, which is a common approach in the field, is difficult to achieve because its different regions cleave at very different rates. We therefore identified conditions that on average yield single cleavages across each region probed by an individual primer. This assumes that within a frozen sample phosphodiester backbone cleavages outside this region do not cause significant conformational changes within the footprinting period, which seems reasonable.

Eleven primers, 19–24 nts long, were identified that produce overlapping reverse transcripts across the region 32–3352 nts for the MS2 gRNA from exposed samples (Figure 2(e), Table 1). Absorption at 260 nm was used to estimate RNA recoveries and transcript concentrations. Triplicate samples at X-ray exposures of 0, 25, 50 & 100 milliseconds were analysed using a commercial capillary electrophoresis service (DNASeq, University of Dundee), alongside ROX 400HD size-markers (Applied Biosystems™) and reference 2′,3′-dideoxyA sequence ladders (Figure 2(d)). Data only become reliable after an "entry peak" in the electropherograms, creating blank areas adjacent to, and including, the regions where some primers anneal. Genome sequences in these regions were analysed using the primer extension products from sites located further 3' (Table 1). We were unable to obtain extension products with primers designed to anneal at the 3' end of the gRNA (nt 3569) or for which extension would terminate at the 5' nucleotide. Both terminal gRNA regions are predicted to form base-paired stem-loops which may explain this outcome. The reactivity profiles obtained cover ∼93% of the gRNA sequence, including all but the final 7 nts of the replicase gene, and ∼1000/1048 nts (∼95%) defined as "flexible" in the highest resolution cryo-EM structure due to weak or uninterpretable density.[33] Open-source software, QuShape,[48] was used to correct the primary electropherograms by applying signal smoothing, as well as baseline, decay and mobility shift corrections. Accurate alignment of peaks and RNA sequence is vital for subsequent analysis. Protection by the phage protein shell typically results in lower footprinted signals than for free RNAs, creating problems with sequence matching. We also detected peak alignment issues due to differential mobility of oligonucleotides of different lengths through the capillary. Software algorithms to correct these problems were developed (Sup. Figure 1), and the details of data processing are described in Materials & Methods.

Determination of gRNA secondary structures using X-ray reactivities

Experimental verification of RNA secondary structure is essential for oligonucleotides > 50 nts in length. Its prediction, e.g. by S-fold alone,[50] on an RNA as long as the MS2 genome results in multiple alternative structures of comparable folding free energies due to the conformational complexity of the gRNA. Adding footprinting reactivities or accessibilities as constraints dramatically improves the accuracy of such predictions.[51–52] We employed a well-established approach to incorporate the corrected XRF nucleotide reactivities into secondary structure predictions here[46] (Figure 3). Base-pairing probabilities are weighted by Boltzmann factors, i.e. exponential functions that include the reactivities via a scaling factor, *m*. This determines how strongly the experimental data influence the S-fold outcomes, for sensible values of *m*. In addition, an offset value, *b,* is applied to ensure that data points with negative values, which occur occasionally in the data, do not perturb the calculations. It is not *a priori* clear which values of *m* and *b* are appropriate for the free energy calculations. In previous studies, a single value for each of these parameters was chosen based on the accuracy with which they reproduced a known element of secondary structure within the footprinted region. The secondary structure for the entire sequence was then calculated using the same parameter values.

The TR stem-loop (nucleotides 1748–1762) is readily visible in the asymmetric cryo-EM map and would be an obvious target for this *m* & *b* parameter selection. To explore this possibility, we examined the gRNA footprints (Figure 3(a)) from transcript (yellow line) and virion (green line) generated by Primer #6 (Table 1). These data for the gRNA around the TR site are shown in front of the respective reactivity plots. The per nucleotide reactivities of these gRNA states are shown. Error bars from triplicate determinations imply that the results are highly reproducible. The gRNA is more protected in the virion than in the transcript, as expected. S-fold predictions for 1000 sample folds across this region, constrained by XRF reactivities, were generated at each of the different *m* = values are 0 – 7; then for *b* = they are -6 – 0 values. This plot suggests that TR and TR-1 are stem-loops in both virion and transcript gRNAs at all *m* & *b* values tested (see Sup. Figure 2). This outcome shows that these PSs are too stable to be sensitive to changes over the parameter ranges used. They are therefore unable to distinguish virion and transcript folds. In contrast, the difference map of the raw reactivity data for TR+1 shows distinct behaviour, a PS-like fold appearing only in the virion (Figure 3).

S-fold calculations were therefore used to identify areas in *m* & *b* parameter space where TR and TR +1a (green), or TR and TR+1b (blue), are the dominant folds (Figure 3(c) & (d)). The *m* & *b* parameter combinations that best represent this trend are the ones for which these folds occur most frequently (i.e. 998 times for the gRNA in the virion and 983 times for same region in the transcript) over 1000 sample folds of the primer read region. These occur at the following unique parameter values: virion *m* = 0.8, *b* = -2.0, & transcript *m* = 6.0, *b* = -3.8. These positions are indicted by black dots in Figure 3(d) & Sup. Figure 2. Using these values for the secondary structure prediction of the entire 3569 nt long gRNA, as virion or transcript, yields the images shown in Sup. Figure 3. Where primer reads overlap, the average reactivity value for each nucleotide was used for the computations. A well characterised, long-range intramolecular interaction (the “Min Jou kissing loop”) is known to occur in the virion gRNA and is visible in the cryo-EM map[33,53]. This kissing loop contact emerges naturally in the calculation of the secondary structure of the full virion gRNA based on the XRF data. However, in initial computations for Primer #6 only, used to identify the *m* & *b* combination for global structure prediction (Figure 3), the nucleotides encompassing one half of this interaction were forced to remain single-stranded, as their partner nucleotides are not within this read. The transcript and *in virio* gRNA secondary structures reveal significant local conformational differences, including folding of PS sites in the virion that are not present in the transcript, as predicted.[54] Overall, however, there is minimal variation around a dominant conformation (the 1000 sample folds agree in 97% of the nucleotides). There is no evidence for significant conformational variants, as proposed previously.[55]

gRNA-protein interactions within a virus-like particle (VLP)

Multiple RNA PS-CP interactions are also expected to occur in the VLP.[3,56] We do not have an asymmetric cryo-EM reconstruction to confirm this assumption, but footprinting of the TR PS is very clear in the XRF data for the VLP. In order to compare the XRF data between phage and VLP, a 50 nt long frame was slid across the region in increments of 1 nt, and Pearson Correlation Coefficients (PCCs) computed for each frame position and ascribed to the central nucleotide position (Figure 3(e)). Only frames centred on nucleotides 1650 to ~1780 show a positive correlation, implying that similar secondary structures occur only in this region, which encompasses TR-1 and TR, implying that both act as PSs in the VLP. In contrast, the region encompassing TR+1 is negatively correlated, indicative of variation in the gRNA conformation/interactions between the two particles.

This has consequences for nucleation of assembly in each case. Assembly initiation in phage most likely occurs at PSs TR-1, TR, and TR+1. The MP bound at a stem-loop directly adjacent to the 3′-end of the gRNA also makes

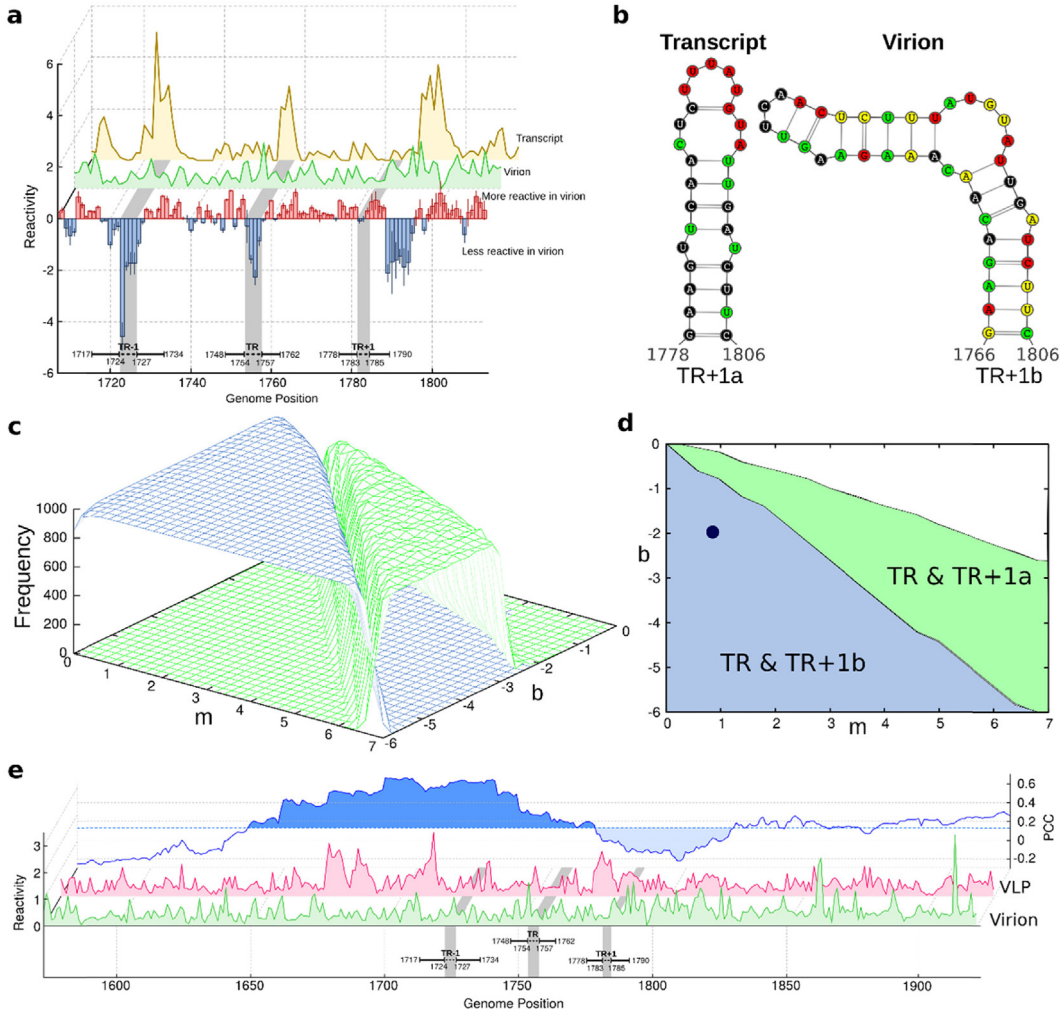

Figure 3. Defining Secondary Structures Using X-ray Footprinting Data. (a) Shows a comparison of the smoothed averages of triplicate nucleotide XRF reactivities between nucleotides 1700–1820 for primer extension reactions with Primer #6: for transcript (yellow) and infectious MS2 phage (green). The relative reactivity differences between these states are shown as histograms with error bars. Red indicates a site that is more reactive in the virion than in the transcript, whilst blue indicates sites that are less reactive; (b) The calculated secondary structures of TR +1 in the gRNA based on S-fold/XRF analysis in the transcript (TR+1a) and in the phage (TR+1b). (c) A folding landscape created by scanning *m* & *b* values showing the number of times TR and its neighbouring PS, TR+1, occur with a PS fold (blue) or as a distinct stem-loop (green) in 1000 structure folds generated for each combination of *m* & *b* parameters; (d) Top view of the folding landscape in (c). The black dot on the blue surface identifies the *m* & *b* parameters that result in the maximum number of predicted PS folds (998/1000). These values were then used for secondary structure prediction of the entire gRNA in the virion (Sup Figure 3). Similar analysis yielded parameter values for calculation of the transcript secondary structure (Sup Figures 2 & 3). (e) A comparison of the XRF reactivities across the region probed by Primer #6 for both VLP (pink) and virion (green). Differences and similarities are highlighted by calculation of PCC values, shown at the rear. The blue dotted line shows the average PCC value across this read. The data are well correlated between nucleotides 1650–1750 but drop significantly either side of this.

protein–protein contacts with the CP dimers bound at TR and TR+1 (Figure 1(d)), implying that these contacts are made early in assembly. These intermolecular contacts create a large loop in the 3′ half (nts 1748–3569) of the gRNA, contributing to the overall organization of the packaged genome. In the VLP, in contrast, assembly initiation likely still occurs at TR-1 and TR, but the 3′ half of the gRNA is less organized. The low PCC values for the noisy replicates of the XRF reactivities (Table 1) are consistent with this interpretation. These noisy signals prevent calculation of a unique gRNA secondary structure in the VLP. The implication is that the VLP, which is often studied in this system and others, is an inadequate surrogate for the infectious virion.

gRNA-protein interactions within the virion

XRF reactivity-guided S-fold returns a dominant secondary structure for the gRNA in phage (Figure 3; Sup. Figure 3) with a 97% nucleotide pairing identity across an ensemble of 1000 sample folds. Stem-loops with loop nucleotides that are more protected in phage than in the transcript reveal the positions of PSs in contact with the CP layer. A prominent example is TR at nucleotide positions 1748–1762[26,28,57] (Figure 1). This site is predicted to be an identical stem-loop in both transcript and virion, but the low reactivity values in the tetraloop in phage are consistent with its known contacts with CP. In particular, the two 3′ loop nucleotides, that are known to be important for CP recognition (Figure 1(c)), are more protected in the virion, whilst the nucleotide 5′ to these remains unaffected. In contrast, nucleotides both 5′ and 3′ to the loop, and its most 5′ residue, become more reactive. These changes can be rationalised by reference to both solution[58] and crystallographic structures[28] for the TR-CP_2 complex.

Using these results as a guide, we identified a further 31 potential PSs, in addition to TR, from the 65 SLs in the XRF-constrained S-fold secondary structure of gRNA in phage (Sup. Figures 3 & 4). In each of these SLs, at least two loop nucleotides are in the more protected category of reactivity values (green or black, Sup. Figure 4). Thus at least 31 PS-like sites, more than double the 16 tightly-bound PSs seen by cryo-EM reconstruction,[33] are in contact with the CP shell at some point. Many of these SLs mimic some of the key features in TR that result in sequence-specific CP binding (Figure 1), consistent with their making CP contacts post-assembly. Since we do not have an absolute scale for the degree of footprinting to expect, we are constrained in our interpretation of the XRF data. For example, we are unable to interpret less protected PSs as being of intrinsically lower affinity for a CP dimer. In any event, we would expect some variation across the ensemble of particles probed, with TR being the most strongly footprinted since it is the highest affinity site.[57] The footprinted SLs identified here therefore represent the minimum number of PS sites within the gRNA.

A further 34 RNA SLs are present in the footprinted gRNA secondary structure in addition to those identified as PSs in phage. The cryo-EM map (Figure 4 & Sup. Movie) of the gRNA suggests that many of these are proximal to the inner CP surfaces in phage, i.e. they are orientated such that they could have acted transiently as PSs during assembly, as has been proposed.[5,31] At least 3 of these (corresponding to nucleotides 1452–1476; 1966–1989 & 2062–2086) are also likely PSs since they were identified as being in contact with the capsid shell via previous CLIP-Seq analysis.[7] CLIP-Seq also identifies TR-1, TR and TR+1 as PSs, in agreement with the XRF result. Such sites may be dynamic within the protein shell and only bind CP_2 transiently. Note, while the XRF criteria used here (i.e. the *m* and *b* values) identify all the PS SLs seen by cryo-EM, several of these do not have protected loops by our XRF criteria. Therefore, they may or may not have acted as transient PSs.

By both XRF or cryo-EM analysis, it appears that a significant fraction (~50%) of gRNA PSs may only function during phage assembly but then dissociate from their CP_2 contacts. Capsid-internal rearrangements of the gRNA tertiary structure due to gRNA condensation during assembly could easily dislocate detached PSs from their CP contacts, preventing rebinding. The loss of each of these RNA PS-CP contacts makes gRNA release, and thus infection, easier.

Discussion

Virion assembly transforms relatively harmless molecular components with great efficiency into a precise molecular machine with the ability to detect, invade and ultimately subvert a host cell. RNA PS-mediated assembly[11] highlights the roles that gRNA-viral structural protein interactions play in this process, focusing attention on the gRNA conformation which previously has been largely neglected. The lack of such information in many other viruses reflects the fact that before the advent of asymmetric cryo-EM reconstructions, the visualisation of ordered RNA fragments within viral structures was the exception rather than the rule.[8,59] Even in most of these exceptions, identifying the nucleotide sequences involved in the RNA-CP contacts is not possible.[59] XRF allows us to do this, detecting differences in conformations and intermolecular contacts between encapsidated and unencapsidated states of the same gRNA with minimal perturbation. Its extension to other aspects of viral life-cycles, including dynamic changes during cell entry and assembly, should therefore be straightforward and informative.

Implications of the encapsidated gRNA conformation for phage assembly

Dai *et al.* built a three-dimensional model of the MS2 gRNA structure (Figure 4) based on their cryo-EM density, but it is incomplete.[33] We used their model to confirm that the footprinted PSs identified above (Sup. Figure 4), as well as the potential transient PS SLs *in virio*, are located externally in the gRNA, i.e. are facing the protein shell. Since this is the case, XRF footprints at these PS positions are likely due to gRNA-CP interactions rather than RNA-RNA tertiary contacts. The RNA phages are one of the most studied groups of viral pathogens. The XRF data, together with multiple additional structural and functional studies, allow us to describe a pathway for MS2 assembly, and infer a

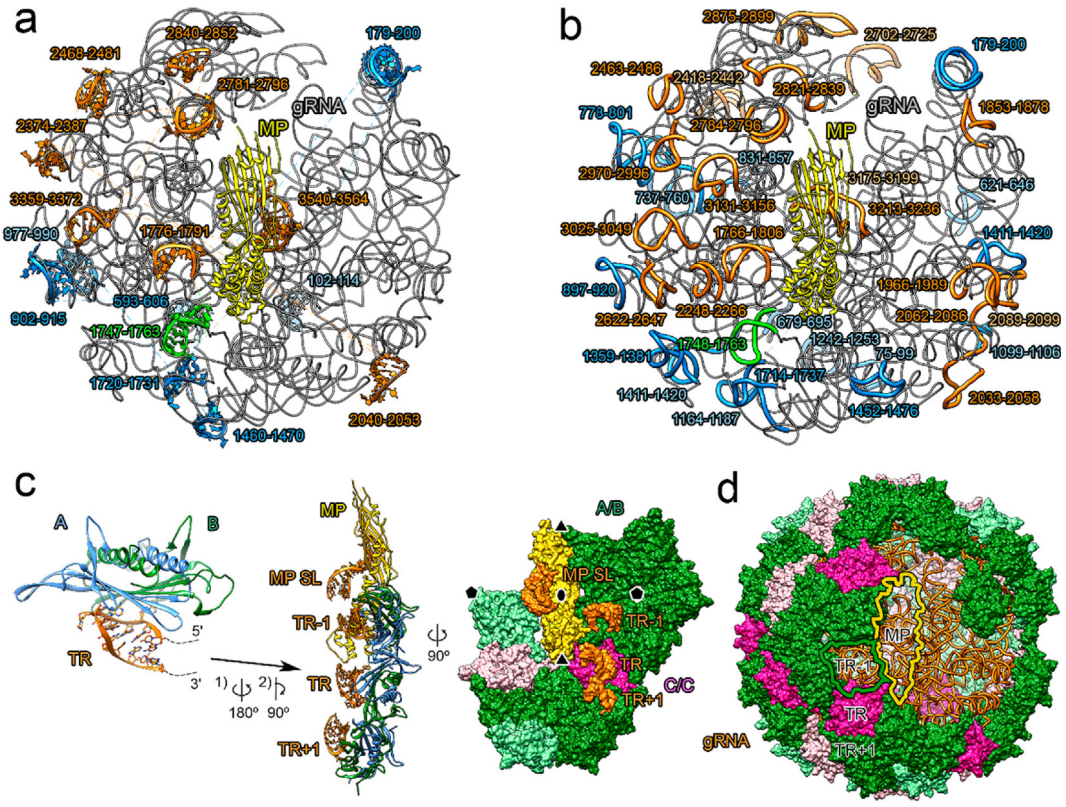

Figure 4. gRNA PSs Identified by XRF & the Inferred Impact on Phage Assembly. (a) Grey ribbon representation of the MS2 gRNA model based on Dai et al[33]. The image is centred on the MP shown as a yellow ribbon. Tightly bound SLs from the cryo-EM reconstruction are shown as ribbon models with TR in green, with the PSs 5′ and 3′ to it, i.e. TR-1 and TR+1, in blue and orange, respectively. (b) As in (a) but with the additional PSs based on the XRF data (Sup Figure 4) included. See Supplementary Movie for more information. (c) The implications of viral protein-gRNA contacts for assembly. A CP dimer bound at TR, and MP bound to the 3′ SL form part of an extensive assembly initiation complex with the CP dimers bound at TR-1 and TR+1, viewed edge on (left) and then from the interior of the phage particle (right). CP dimer conformers are colour-coded as green (A/B) and pink (C/C). Saturated colours indicate that RNA PSs (orange) remain bound to those CP subunits post-assembly. Together this complex determines the size and symmetry of the capsid that can be assembled by binding further CP dimers. It also blocks access for replicase to its 3′ binding site, and commits the gRNA molecule to assembly. (d) The fully assembled phage, viewed centred above the MP, but with the MP itself, as well as neighbouring CP_2 lacking bound PSs, removed.

mechanism for its gRNA uncoating, in unprecedented detail. The molecular switch repurposing phage gRNAs from multi-cistronic mRNAs into assembly substrates[24] was believed to require only the sequence-specific interaction between a CP dimer and the translational operator stem-loop (TR). This would block ribosome access to the replicase start codon.[25] Asymmetric cryo-EM phage reconstructions, as well as many previous studies of the roles of the gRNA in assembly,[12,2–6] suggest that this is only the first of many defined gRNA-structural protein contacts required to make it infectious.[31–33]

For instance, the MP must bind to its stem-loop located at the 3´-end of the gRNA[33] more tightly than any of the gRNA PS-CP contacts. The latter all break whilst the gRNA:MP complex enters the host cell with the gRNA at the start of infection. MP also contacts the CP dimers bound at TR and TR+1 both of which together with TR-1, from the footprinting data, are in contact with the gRNA. Cryo-EM reconstructions allow us to place these contacts in the three-dimensional architecture of the phage. Although these three PS sites are adjacent along the gRNA (~40 nts apart), and some of their bound CP subunits form neighbouring CP-CP contacts, in three dimensions these CPs form parts of distinct symmetry axes in the surface lattice. Interaction with the MP pushes the CP dimer bound at TR into a C/C-position, even though the CP dimer adopts the A/B-like conformation.[33] Note, crystal structures reveal that MS2 VLPs soaked with TR-like oligonucleotides are bound in a unique orientation at A/B dimers, but in both possible orien-

tations at C/C dimers,[28] consistent with this observation. This dimer is a good candidate to sit on the first two-fold site formed during phage assembly. Its neighbouring CP dimers, bound at PSs TR-1 and TR+1, occupy A/B positions but in separate neighbouring five-fold capsomers. The complex consisting of MP and three CP dimers bound to both the gRNA region around TR and the RNA 3′-end, is a likely assembly intermediate. It can act as a nucleation complex for $T = 3$ particle assembly, via defined PS-CP_2 contacts that define the size and symmetry of the phage particle. Its formation is a direct consequence of the presence of the MP. It may also act as a stable switch for the genome becoming an assembly substrate, since the gRNA-bound MP blocks access by the replicase complex to its binding site at the 3′-end of the viral RNA, whilst the replicase start codon is simultaneously sequestered by the CP_2 bound at TR.

VLP versus infectious phage assembly

VLP assembly around the gRNA also depends on the formation of multiple PS-CP contacts.[5,54] However, our data suggest that the conformation of the gRNA in a VLP differs markedly from that in phage (Figure 3(e)), most likely because the gRNA loop between TR and the MP binding site is missing. In asymmetric cryo-EM reconstructions of phage, and symmetry-averaged VLP structures, the capsid forms because of the ordered interdigitation of CP dimers in either A/B or C/C conformations. For both phage and VLP, assembly beyond the initiation stage is most likely to occur due to formation of additional gRNA-CP dimer interactions, i.e. CP_2-PS, which will induce formation of A/B dimers at appropriate sites. Beyond the assembly initiation complex described above, XRF confirms that there are a minimum of 31 PS sites in contact with the CP shell, whilst yet more are adjacent to this protein layer and might have become dissociated post-assembly. Most PS-like folds are also present in the gRNA transcript, implying that the genome has evolved to support PS-mediated assembly (Sup. Figure 5). XRF reads are less reproducible in the VLP compared with phage (Figure 3(e) & Table 1). This confirms the importance of the MP for assembly initiation, and potentially for orchestrating the order of PS-CP_2 contacts. Virion assembly studies in the absence of MP cannot therefore capture essential aspects of the PS-regulated assembly mechanism and VLPs are therefore inappropriate surrogates for phage assembly *in vivo*.

Do the roles of the gRNA-CP contacts extend beyond virus assembly?

XRF data confirm that not all potential PSs within the gRNA are bound in the infectious particle. Indeed, if a particular PS is bound in only a fraction of particles during the footprinting period (100 msec), then it is likely to appear less modified than its equivalent in the transcript. This suggests differences in reactivities at the PS binding sites could reflect differences in the strength of the PS-CP_2 contacts. Genome condensation and molecular crowding resulting from encapsidation could provide the driving force for an internal conformational rearrangement resulting in the loss of some of these contacts.[5] Whatever the molecular origins of these dissociation events, we speculate that they may play a significant role(s) in phage infectivity.

If this were the case, then dissociation events would be expected to be non-random in the context of the phage structure. Mapping the PSs identified by XRF into the cryo-EM density, we see precisely that situation. All the CP dimers around one of the five-fold capsomers adjacent to the MP have PSs bound (Figure 5(a)). In contrast, the PSs in its neighbouring capsomer retain only one PS contact. That five-fold capsomer sits above the 3′ domain of the gRNA which, with MP bound at its 3′-end, has to exit the particle first during transit of the basal body in the pilus-extruding machinery,[60] the first step of infection. Dissociation of multiple PS-CP contacts would make this extrusion easier (Figure 5(b)). It also traps the FG-loops from the corresponding CP dimers in a non-preferred conformational state (Sup. Figure 6). They can only convert to their preferred C/C conformations by disrupting the protein shell. This outcome also implies that the five-fold capsomer lacking PSs is under structural strain, and thus more likely to rupture. Thus, dissociation of bound PSs contributes directly to making infection more likely.

Could this mechanism be more widespread in virology?

The presence of different conformational states that cannot all be simultaneously fulfilled, i.e. the trade-off between gRNA condensation and formation of PS-CP contacts, is an example of "molecular frustration".[61–62] This phenomenon, initially described by Wolynes and colleagues with respect to protein folding, is responsible for the gradual loss of PS-CP contacts, ultimately preparing the phage for gRNA release. It thus provides a driving force for infection. These insights imply a previously unsuspected aspect of natural virus assembly, i.e. that it simultaneously sets up the conditions of disassembly and gRNA release.

It will be interesting to see if the molecular mechanisms visible in a simple phage underlie infections by more sophisticated viral pathogens. Variation of this theme is obvious in Qβ particles since they encompass additional stabilisation due to their disulphide-linked CP subunits at the particle five-folds,[63] which their MS2 equivalents lack. These might be released by the redox conditions upon cell entry. PS-mediated assembly regulation appears to occur very widely across many viral families,[11] yet the asymmetric structures of rel-

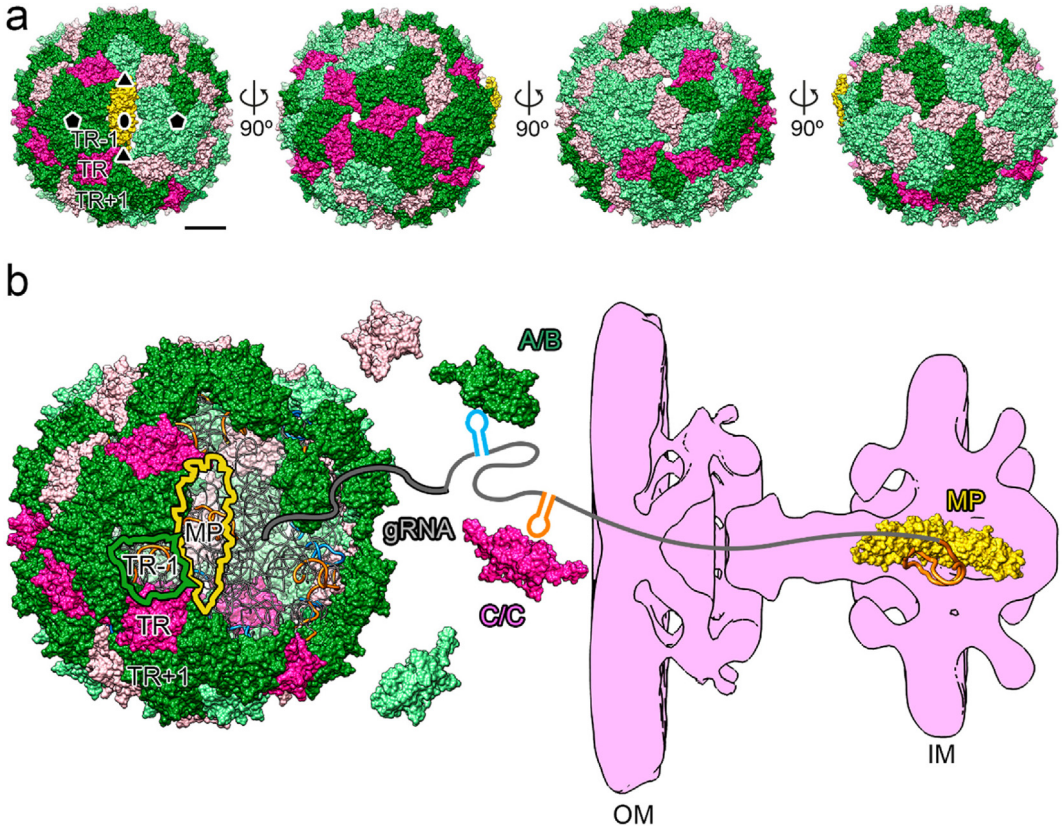

Figure 5. Potential Roles of PS Dissociation for Phage Infection. (a) An image of the infectious phage based on the asymmetric cryo-EM reconstruction, viewed in a series of images rotated by 90°. Lightly coloured subunits are CP dimers lacking an underlying bound PS, whilst those shown in saturated darker colours are bound to gRNA PSs. The subunits are colour-coded as in earlier Figures to indicate the conformations of their FG-loops. (b) The image in (a) with MP and the CP_2 pentamer lacking most of its bound PSs removed, viewed along the two-fold axis. This likely allows the MP, together with the bound 3′ end of the gRNA, to dissociate following association with a bacterial pilus. Pilus retraction would introduce the proximal phage particle to the pilin-extruding cellular machinery (pink outline based on a cryo-EM reconstruction[60] which makes a channel from the outer membrane (OM) to the cytoplasm via the inner membrane (IM). The MP occupies roughly two thirds of the channel volume and can easily "tow" the attached gRNA through, as indicated, because of the relative paucity of CP_2-PS contacts that need to be broken. This reaction is favoured because it relieves the local geometric strain in the particle surface lattice generated by PS dissociation. Only the MP:gRNA complex enters the bacterium with the CP subunits being left on the outside.

atively few of these virions have been determined. This prevents direct visualisation of the presumed intermolecular interactions between gRNA and CP, especially as many of these virions, unlike MS2, have CPs that encompass extended polypeptide chains at their N- or C-termini. These "tails" add additional flexibility in CP-PS contacts creating significantly disordered regions which are unlikely to be captured fully by current structural techniques. Molecular frustration could thus be a more general phenomenon preparing virions for the next steps of their life-cycles. This feat has been difficult to rationalise previously. Designing ways to test this hypothesis directly with MS2 is beyond the scope of this manuscript. As results with the VLP demonstrate, such studies will need to include any protein components required for genome release, such as MP. Their presence is literally vital for infectivity. The molecular details of both virion assembly and disassembly are important for understanding both natural infections, and the potential use of VLPs in gene therapy applications. The insights derived here provide unique perspectives on previously poorly understood aspects of the viral life-cycles.

Materials and Methods

Source of infectious MS2 phage & reassembly of MS2 VLPs

Aliquots of infectious MS2 bacteriophage (ATCC 15597-B1) were propagated in C3000 (ATCC) cells at an MOI of 10:1. Under these conditions lysis throughout the culture occurs within a tight

window of ~10 min. Following concentration by precipitation, phage particles were purified by size-exclusion chromatography (NAP10, GE). Given the regulation of the phage life-cycle, we assumed throughout that all particles are both identical and infectious.

MS2 gRNA was transcribed with a T7 kit (NEB) from a pSMART HCAmp plasmid containing a full-length clone (Genbank V00642.1), linearised with *Hpa* I directly 3′ of the genome, and subsequently purified using an RNeasy kit (Qiagen), as per the manufacturer's instructions. MS2 gRNA was extracted from virions using the same kit. Both RNAs were eluted into nuclease-free water (Severn Biotech Ltd) and their integrity assessed on denaturing agarose gels (not shown). Sample concentrations were adjusted to a minimum of 200 ng/μL (with respect to RNA concentration in virus samples). All virus and RNA samples, including unexposed controls, were stored at −80 °C until shipment for XRF. Virus-like particles were reassembled *in vitro* using a procedure known to yield $T = 3$ shells,[5] starting with a mixture of CPs purified from glacial acetic acid dissociated phage and a gRNA transcript.

Determination of XRF reactivities

Samples were footprinted on beamline 17-BM XFP at Brookhaven National Laboratory,[36] NY, USA. Each beamline session was preceded by a measurement of the relative beam strength. This was achieved via a calibration curve of the photo-bleaching of Alexafluor-488, diluted in 10 mM sodium-phosphate buffer (pH 7.4) exposed to the beam for 10, 15, 20 or 30 milliseconds with 762 μm aluminium attenuation. The fluorescence intensity was measured by fluorimeter for all samples and normalised relative to a no exposure sample. The fluorescence was plotted against exposure time and a rate constant, *k*, calculated. This was used as a measure of relative beam strength which is comparable between runs and allows sample exposure times to be adjusted to produce similar levels of RNA hydroxyl radical modification between sessions. Samples were mounted in the beamline in a temperature-controlled (-30 °C, ensuring samples remained frozen), 96-well motorised holder, which accommodates individual or strips of 8–12 PCR tubes. Samples were exposed to the beam via a shutter mechanism for 10–100 msec. The holder moves to align each sample tube to the beam for the programmed exposure time. Post-exposure, samples were stored at −80 °C for return shipping to the host laboratory for further processing.

RNA was extracted from exposed virions using Buffer AVL (Viral RNA extraction buffer; Qiagen) for 10 min at room temperature. RNA was bound to magnetic beads (RNAClean XP; Beckman Coulter) and incubated for 10 min. The RNA-bound beads were washed three times with 70% (v/v) ethanol and allowed to air dry for 5 min, before elution with 12 μL nuclease-free water. The extracted RNAs and beamline-exposed free RNAs were reverse transcribed using Superscript IV (Invitrogen) and a sequence specific 5′ 6-carboxyfluorescein (FAM)-labelled primer that hybridised 3′ of the region of interest. Sequencing ladders were synthesised from *in vitro* transcribed RNA (2 μg RNA template per ladder) using hexachloro-fluorescein (HEX) labelled primers (same sequence as experimental sample primers) and the addition of a 1:1 molar concentration of ddATP:dNTP. RNA was degraded with RNase H (NEB, 5 units per sample) and the cDNA purified by precipitation (3x volumes of ethanol, 0.3 M sodium acetate, 1% volume of glycogen). Experimental samples were resuspended in 20 μL formamide and sequencing ladders were resuspended in 12 μL H_2O, then 1.6 μL of each ladder was spiked into each experimental sample. The samples were heated to 65 °C for 10 min then transferred to a 96-well plate and frozen for shipping to DNASeq (Dundee, UK) for capillary electrophoresis.

Determination of PSs in contact with CPs in virio by analysis of the XRF data

The strongest PS in the MS2 gRNA is TR,[57] and its modification pattern in XRF was used to benchmark other potential PSs. There are differences in reactivity depending on which nucleotide is examined. Figure 1 shows the TR structure present in crystal structures of the TR in complex with the CP shell (Sup. info).[28] Note this conformation differs from the structure derived from NMR[58] of free TR oligonucleotides in which the base at A1757 intercalates between neighbouring base pairs. The CP complex shows that the adenines at positions A1757(A-10) and A1751(A-4) hydrogen bond to amino acids Thr45 & Ser47 in the A and B subunits, respectively. The designations in brackets are for ease of comparison with previous structural data in which bases were numbered relative to the A (+1) of the AUC replicase start codon. In a TR encompassing oligonucleotide in solution, NMR shows that the A1751 base intercalates between its neighbouring base pairs. The conformational change needed to allow it to make the contacts to a CP subunit are consistent with the XRF reactivities of its neighbouring nucleotides.

Given the reactivity within the TR-CP_2 complex, we examined the XRF data associated with every stem-loop present in the calculated encapsidated gRNA secondary structure. There are 74 such stem-loops, 53 of which have an equivalent in the transcript (Sup. Figure 4). Of these, loop sizes vary from 8(1); 7(2); 6(3); 5(8); 4(27) to 3(12) nucleotides in length. The most frequent, tetraloops, are potential PSs. In TR the 3′ loop nucleotides, U.A, are less reactive in the virion. This is consistent with U1756(U-5) forming a

stacking interaction with the side-chain of Tyr85 in the A subunit. We therefore identified all tetraloops in which the two most 3′ loop nucleotides show XRF protection in the virion relative to the transcript, provided there is an equivalent stem-loop in the transcript. This comparison suggests that all 11 such stem-loops are gRNA PSs. There is one example of a C.A dinucleotide as the 3′ residues of the stem-loop (UUCA, nts 788–791). In isolated oligonucleotides, a cytidine at the 1756(C-5) position makes an additional intramolecular hydrogen bond relative to the wild-type TR sequence, stabilising the bound conformation and explaining its higher CP affinity.[28,64–65] It is therefore not surprising that the XRF reactivity suggests this site is a PS. We then examined the other tetraloops (3) in which a YR (pyrimidine/purine) motif shows low reactivity, i.e. shown as green or black in Sup. Figure 4. For stem-loops presenting tetraloops but without an equivalent secondary structure in the transcript to compare with, we identified those in which the 3′ two nucleotides of the loop (7), or a YR motif (1), have low reactivity (i.e. on average are smaller or equal to the mean (0.6)).

In total, this analysis identifies 22 of the stem-loops with tetraloops as PSs. An equivalent analysis for triloops and pentaloop, identifies a further 6 PSs for each, making a total of 32 potential PS sites. To confirm these tentative assignments, we interrogated the cryo-EM reconstruction rejecting stem-loops involved in RNA-RNA contacts, or a combination of RNA-RNA and CP contacts. This reduced the number of potential PS sites by 2, leaving 32 PSs that could have acted during assembly. This is more than the 15 sites seen to be tightly bound in the well-resolved regions of the EM reconstruction,[33] and it does not include all of those sites. Note, XRF reactivities are averages so if there are positions where the structure of the phage is heterogeneous due to conformational flexibility, e.g. as a result of repeated transient PS binding, the outcome could be misleading. Despite these caveats, it is clear that multiple PSs within the gRNA remain bound to the CP shell, as expected from PS-mediated assembly.[1,11] It is also clear that many other stem-loops could have acted as PSs and subsequently dissociated from the protein layer.

Structural analysis of the MS2 capsid

In order to identify which CP dimers in phage are in contact with the gRNA, the deposited model of the asymmetric structure (PDB 5TC1, https://doi.org/10.1038/nature20589) was interrogated to determine if the inner surface of each A/B or C/C dimer was in contact (saturated colours) or not (unsaturated colours) with the viral gRNA SLs (Figure 5(a)). The impact of PS dissociation on phage structure was assessed by determining the degree of displacement from an idealised $T = 3$ lattice structure (PDB 1ZDH, https://doi.org/10.1006/jmbi.1997.1144) for each CP_2 in the asymmetric structure. For this purpose, the icosahedrally symmetric structure model was fitted globally as a rigid body into the asymmetric cryo-EM map with the Fit in Map routine in UCSF Chimera (https://doi.org/10.1002/jcc.20084). A copy of each CP_2 was then generated and fitted individually into the cryo-EM map. Finally, the RMSD values between each individual CP_2 in the $T = 3$ averaged model and their equivalents in the asymmetric map were calculated. CP_2 displacements are mostly modest (RMSD < 0.6 Å) but can be significantly larger (RMSD > 3.4 Å), implying heterogeneity across the entire structure.

X-ray footprinting data analysis

QuShape[48] has previously been used successfully for the analysis of XRF data from ribosomal RNAs. Due to the significantly longer length of the MS2 gRNA, and the additional protection levels arising from encapsidation within the viral capsid, this software is insufficient in our context. We therefore developed a suite of additional algorithms (Sup. Figure 1), that we implemented in combination with QuShape functions as described below.

Initial Data Processing. The CE samples were extracted using the 'ABIF reader' tool within the QuShape package, and the following pre-processing steps performed on the raw electropherogram traces (For more information on these steps, see QuShape[48]):

- *triangular smoothing* of all traces (footprinted sample, ddA ladder and size markers) for noise removal;
- *baseline correction* – a frame of 60 elution time points was slid in increments of 5 along all traces, and the minimum intensity was subtracted from each data point in every frame position;
- *signal decay correction* was applied to all but the size marker trace;
- *mobility shift corrections*, mitigating differences in peak positions in the ddA ladder and footprinted sample due to different retention times of the dyes.

These preprocessing steps were performed over 20 frames of the raw electropherograms. The start of the first frame was chosen at least 50 elution time points after the entry peak of the electropherogram, and its end point at least 50 elution points before the exit peak, and each subsequent window was obtained via reduction by 5 elution points at either end. The analysis protocol below was carried out for each frame individually, and then the average taken over all frames, in order to mitigate against bias due to the position of the exit peak.

In each case, size marker positioning was performed as an additional preprocessing step. For this, the size marker positions in the

electropherogram were determined using the *peak finder* function of QuShape on the size marker trace, and the 21 highest peaks (corresponding to the 21 size markers used) were extracted into the peak list $\boldsymbol{Peaks}^{\boldsymbol{SM}} = \left\{X_1^{SM} \cdots X_{21}^{SM}\right\}$.

Peak Identification. The *peak finder* function of QuShape was also used to identify peak positions (X_i^{FS}) and intensities $(I_i^{FS} = I\left(X_i^{FS}\right))$ in the footprinted sample trace that were retained in the peak list, $\boldsymbol{Peaks}^{\boldsymbol{FS}} = \left\{\boldsymbol{P}_1^{FS} \cdots \boldsymbol{P}_n^{FS}\right\}$, where $\boldsymbol{P}_i^{FS} = \left(X_i^{FS}, I_i^{FS}\right)$. As this method often misses peaks, especially at saddle points and in wide and shallow minima, we developed additional algorithms to mitigate this problem (see Sup. Figure 1; I-III):

Step I: Saddles were identified as peaks in the negative modulus of the first derivative of the footprinted sample $(-\left|I^{FS'}\right|)$. To exclude minima, the positions corresponding to minima in between peaks were disregarded. The intensity values in the footprinted sample at the obtained shoulder positions were added to $\boldsymbol{Peaks}^{\boldsymbol{FS}}$. To identify any missing peaks, the distances between adjacent peaks $\boldsymbol{P}_i^{FS}$ and $\boldsymbol{P}_{i+1}^{FS}$ in $\boldsymbol{Peaks}^{\boldsymbol{FS}}$, together with their mean (W_m) and standard deviation (W_{sd}), were then calculated. For any distances greater than $2W_m - 4W_{sd}$, a new peak was inserted at position $X_i^{FS} + W_m$ between $\boldsymbol{P}_i^{FS}$ and $\boldsymbol{P}_{i+1}^{FS}$ with intensity $I\left(X_i^{FS} + W_m\right)$.

Step II: Any peaks in the footprinted samples $\left(\boldsymbol{Peaks}^{\boldsymbol{FS}}\right)$ located between peaks in the size marker peak list $(\boldsymbol{Peaks}^{\boldsymbol{SM}})$ were exported into a separate list $(\boldsymbol{Peaks}^{\boldsymbol{FS(sub)}})$. The interval between any two adjacent size marker peaks was then partitioned into bins, each representing a nucleotide in the genome, and peaks in $\boldsymbol{Peaks}^{\boldsymbol{FS(sub)}}$ assigned to these bins. Due to random variation in peak positions along the chromatograms, some peaks will fall into the same bin, and were therefore reassigned to other bins. To achieve this, for any bins containing no peaks, the closest bin with more than one peak was identified. Their peaks were then redistributed such that a single peak was assigned to both the empty and the overpopulated bin. In case $\boldsymbol{Peaks}^{\boldsymbol{FS(sub)}}$ exceeded the number of bins, the following procedure was applied: The average over peak positions and intensities associated with each bin was taken; if two distinct bins containing more than one peak were both equidistant from an empty bin, then the bin containing a peak position closest to the edge of the empty bin was selected for reassignment into the empty bin.

Step III: If $\boldsymbol{Peaks}^{\boldsymbol{FS(sub)}}$ contains fewer peaks than bins, peaks may be associated with the incorrect bin. This was mitigated by comparison of the different replicates. For this, the peak distribution of each replicate was represented as a sequence B_j encoding peak height with reference to the maximal intensity in the ensemble $(\max(I^{FS(sub)}))$ as low (L), medium (M), or high (H) as follows:

$$B_j = \begin{cases} L & \text{if} \quad I_j^{FS(sub)} < 0.33 * \max\left(I^{FS(sub)}\right) \\ M & \text{if} \quad 0.33 * \max\left(I^{FS(sub)}\right) < I_j^{FS(sub)} < 0.66 * \max\left(I^{FS(sub)}\right) \\ H & \text{if} \quad I_j^{FS(sub)} > 0.66 * \max\left(I^{FS(sub)}\right) \end{cases}$$

If $\boldsymbol{Peaks}^{\boldsymbol{FS(sub)}}$ contains fewer peaks than size marker bins, assignment of some size marker bins as unoccupied was achieved using an alignment of the sequences B_j of different replicates using the Needleman-Wunsch algorithm. The following protocol was used:

- The intensity profiles $(I^{FS(sub)})$ of the peaks in $\boldsymbol{Peaks}^{\boldsymbol{FS(sub)}}$ were extracted as a sequence;
- sequences corresponding to different replicates were aligned pairwise using the Needleman-Wunsch algorithm, and a zero value was inserted into the sequences in any position where the algorithm identified an insertion, creating extended intensity profiles;
- for any two such extended intensity profile, the Pearson correlation coefficient (PCC) was computed, and for each replicate all its possible extended profiles and PCC values with other replicates retained;
- the sum over all pairwise PCCs was computed for any possible combination of extended replicate profiles;
- the combination with the greatest sum was chosen as the final alignment, uniquely assigning a specific extended sequence to each replicate; intensity profiles $(I^{FS(sub)})$ were updated accordingly by inserting a peak with zero peak height in the middle of the bin.

The following *final reassignment check* was performed:

- pairwise PCCs were computed based on $\boldsymbol{Peaks}^{\boldsymbol{FS(sub)}}$ for different replicates;
- if any of these pairwise PCCs was above a threshold, C_{PCC}, the $\boldsymbol{Peaks}^{\boldsymbol{FS(sub)}}$ of the replicate that did not contribute to this computation was reduced down further by removing bins with a peak amplitude of less than 33% of the maximal peak amplitude in $\boldsymbol{Peaks}^{\boldsymbol{FS(sub)}}$;
- the above process was repeated for different values of C_{PCC} from 0.95 down to 0.7 in steps of 0.05;
- in order to avoid overfitting, alignments resulting in any value migrating more than three bins in $\boldsymbol{Peaks}^{\boldsymbol{FS(sub)}}$ were discarded;
- a sequence representation was then generated for each replicate in $\boldsymbol{Peaks}^{\boldsymbol{FS(sub)}}$ by inserting a peak with zero peak height in the middle of the bin as described above.

The individual aligned $\boldsymbol{Peaks}^{\boldsymbol{FS(sub)}}$ sub lists were then combined into an aligned peak list for the full primer region covered by the size marker trace $\boldsymbol{Peaks}^{\boldsymbol{FS(part)}}$.

Reactivity Calculations

Reactivities were calculated for each of the bins in $\boldsymbol{Peaks}^{\boldsymbol{FS(part)}}$ as the areas under the peaks, $A_i^{FS(part)} = I_i^{FS(part)} * w_i^{FS(part)}$, where $w_i^{FS(part)}$ denotes the width of the ith peak. Average and standard errors over areas and intensities for different replicate datasets were calculated. To achieve a meaningful comparison between reactivities from different footprinted samples, the footprinted peak areas were counter corrected against background using QuShape methods. In particular, *counter corrected reactivities* were obtained as $R = \bar{A}_{tr} - sf * \bar{A}_b$, where $\bar{A}_{tr}$ and $\bar{A}_b$ denote the average peak areas for the footprinted and background samples respectively, and *sf* denotes the scaling factor derived for background samples (i.e. the values at 0 milliseconds exposure). The standard error was then calculated using trigonometric error propagation as:

$$SE_R = \sqrt{(SE_{tr})^2 + (sf * SE_b)^2}.$$

The average over all replicates (R values; excluding outliers) and the normalization factor (*nf*) were calculated using QuShape methods. All R and SE_R values were then divided by *nf* to obtain the normalized area difference, R^N, and the normalized standard error SE_R^N.

Difference plots were calculated for different states of the gRNA (transcript; *in virio*; & VLP) at the optimal exposure time (50 milliseconds, see main text) as follows: The counter corrected reactivities R of both sets were combined, and a normalization factor *nf* determined for this combined set. R and SE_R were then renormalized individually using this *nf* from the combined set. The difference values were then calculated as $R_{V-T}^N = R_{V*}^N - R_{T*}^N$, where R_{V*}^N and R_{T*}^N are the renormalized area differences for the individual sets, respectively. Standard errors for the difference maps were again calculated based on trigonometric error propagation applied to renormalized standard error values as above.

We noted that the outcome of the procedure described above may produce anomalies if the frame used for preprocessing includes the exit peak. To mitigate against such errors, we computed pairwise PCCs for the normalised reactivity profiles based on different preprocessing windows. For each frame, the average PCC was computed, and any frames with average PCCs of less than 0.8 were discarded. The mean of the normalised reactivity profiles in the remaining frames was used as the final normalized reactivity profile, and the standard error was calculated using trigonometric propagation:

$$SE_{Final} = \frac{1}{\sqrt{n}}\sqrt{\overline{(SE_R)^2}},$$

where n denotes the number of windows used to calculate the average, and SE_R represent the standard errors of the normalized data for each window.

Positioning of the ddA ladder. The positions and intensities of the peaks in the ddA ladder trace were determined using QuShape methods and filtered into the target peak list $\boldsymbol{Peaks}^{\boldsymbol{SL}}$ following the same procedure as outlined for the intensity peaks of the footprinted samples above. In particular, Steps I and II were repeated in order to identify missing peaks in the ddA ladder trace, and for bin reassignment of peaks in $\boldsymbol{Peaks}^{\boldsymbol{SL}}$ with positions between pairs of adjacent positions in $\boldsymbol{Peaks}^{\boldsymbol{SM}}$. Association of ddA ladder peaks with size marker bins was achieved as follows:

- If a single peak was located in a bin, the intensity of that peak was assigned to that bin;
- if more than one peak was assigned to a bin, the average intensity of these peaks was assigned as a single value to this bin;
- if no peaks were assigned to a bin, then the intensity value of that bin was set to 0.

The individually processed lists ($\boldsymbol{Peaks}^{\boldsymbol{SL(sub)}}$) were then combined into a single aligned peak list for the full primer region ($\boldsymbol{Peaks}^{\boldsymbol{SL(part)}}$), encompassing nucleotides between the start and end of the size marker trace.

In order to correct for errors in peak identification in individual ddA ladders, results were benchmarked against the ladders associated with different replicates (Figure S1 IV). In particular, the intensity profiles of ddA ladders in $\boldsymbol{Peaks}^{\boldsymbol{SL(part)}}$ were translated into binary sequences B_j:

$$B_j = \begin{cases} 0 & if \quad I_j^{SL(part)} < c_1 * \max\left(I^{SL(part)}\right) \\ 1 & if \quad I_j^{SL(part)} > c_1 * \max\left(I^{SL(part)}\right) \end{cases}$$

where the constant c_1 was chosen such that the mean PCC over all pairwise combinations of the binary sequences in the ensemble was maximised.

The entries B_j of the vectors $\boldsymbol{B}$ were then added element-wise into a 'ballot-box' array, $\boldsymbol{Bbox} = \{Bb_1, \cdots, Bb_n\}$, facilitating comparison between replicate ladders. This array was then converted into a binary consensus sequence with elements B_j^{final,c_2}:

$$B_j^{final,c_2} = \begin{cases} 0 & if \quad Bb_j < c_2 * N \\ 1 & if \quad Bb_j > c_2 * N \end{cases}$$

where N denotes the total number of datasets in the ensemble, that is, the number of ddA ladders per plate sequenced, and c_2 a constant that we have varied in

the range 0.5–0.95 in increments of 0.05. For each fixed value of c_2, this consensus sequence was compared with the genomic sequence to determine the optimal position of the ddA ladder. For this, the viral genome sequence was converted into a binary sequence B_j^{gen}, associating a 1 with each occurrence of U and a zero with every other nucleotide:

$$B_j^{gen} = \begin{cases} 1 & if \quad nucleotide = U \\ 0 & else \end{cases}$$

The consensus sequence B_j^{final,c_2} was then slid along B_j^{gen} in increments of 1 nucleotide, and for each frame position, the number of correctly matched U's (1s) calculated. For each fixed value of c_2, the maximal number of matches and corresponding frame position(s) were recorded. The position of the ddA ladder with respect to the genomic sequence was then taken to be the frame position with the maximal number of matches across all values of c_2.

CRediT authorship contribution statement

Rebecca Chandler-Bostock: Methodology, Investigation, Writing – original draft. **Richard J. Bingham:** Methodology, Investigation, Visualization, Writing – original draft. **Sam Clark:** Software, Data curation, Methodology, Investigation, Writing – original draft. **Andrew J.P. Scott:** Methodology, Investigation, Writing – original draft. **Emma Wroblewski:** Methodology, Investigation. **Amy Barker:** Methodology, Investigation. **Simon J. White:** Methodology, Investigation. **Eric C. Dykeman:** Methodology, Investigation. **Carlos P. Mata:** Methodology, Investigation, Writing – original draft. **Jen Bohon:** Methodology. **Erik Farquhar:** Methodology. **Reidun Twarock:** Conceptualization, Methodology, Visualization, Formal analysis, Supervision, Writing – original draft, Project administration, Funding acquisition. **Peter G. Stockley:** Conceptualization, Visualization, Supervision, Project administration, Funding acquisition, Methodology, Writing – original draft.

DATA AVAILABILITY

The links to websites containing the original data are within the manuscript.

Acknowledgements

We are grateful to Professor Sarah Woodson, Johns Hopkins University, for her encouragement and support in the use of XRF.

Portions of this work used the XFP (17-BM) beamline at NSLS-II. Development of XFP was made possible by the US National Science Foundation, Division of Biological Infrastructure (grant No. 1228549), while operations support of XFP was provided by the US National Institutes of Health (grant No. P30-EB-009998). NSLS-II, a US Department of Energy (DOE) Office of Science User Facility and is operated for the DOE Office of Science by Brookhaven National Laboratory under Contract No. DE-SC0012704. We also thank DNA Sequencing & Services (MRC I PPU, School of Life Sciences, University of Dundee, Scotland, www.dnaseq.co.uk) for DNA sequencing.

Funding

PGS & RT thank The Wellcome Trust (Joint Investigator Award Nos. 110145 & 110146 to PGS & RT, respectively) for funding. We also acknowledge the financial support of The Trust of infrastructure and equipment in the Astbury Centre, University of Leeds (089311/Z/09/Z; 090932/Z/09/Z & 106692), and for their additional support, together with The University of Leeds, of the Astbury Biostructure Facility. RT acknowledges additional funding via an EPSRC Established Career Fellowship (EP/R023204/1) and a Royal Society Wolfson Fellowship (RSWF \R1\180009).

Competing interests

Authors declare that they have no competing interests.

Data and materials availability

The processed data from the capillary electrophoresis data analysis is available as a collection on Figshare (https://doi.org/10.6084/m9.figshare.c.5395302). All other data are available from the corresponding author on reasonable request.

Code and additional materials availability

The software package used to analyse the capillary electrophoresis (BoXFP), as well as the Supplementary Movie, and larger scale images of the MS2 secondary structures *in virio* as well as a transcript are available to download from GitHub at https://github.com/MathematicalComputational-Virology/XRFanalysis.

Appendix A. Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2022.167797.

Received 20 May 2022;
Accepted 15 August 2022;
Available online 20 August 2022

Keywords:
RNA PS-mediated virion assembly;
bacteriophage MS2;
RNA X-ray footprinting;
molecular frustration;
phage infection

† These authors contributed equally to this manuscript.
‡ Present addresses: Room 208, Biology/Physics Building, Department of Molecular and Cell Biology, University of Connecticut, 91 N. Eagleville Rd., Unit-3125, Storrs, CT, USA.
§ Present addresses: Biocomputing Unit, Department of Macromolecular Structures, National Centre for Biotechnology (CSIC) & Electron and Confocal Microscopy Unit (UCCTs), National Centre for Microbiology (ISCIII). Majadahonda, Madrid, Spain.
|| Present addresses: Los Alamos National Laboratory, Los Alamos, NM 87545, USA.

References

1. Stockley, P.G. 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.
2. Morton, V.L. et al, (2010). The impact of viral RNA on assembly pathway selection. *J. Mol. Biol.* **401**, 298–308.
3. Rolfsson, Ó., 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.
4. Knapman, T.W., Morton, V.L., Stonehouse, N.J., Stockley, P.G., Ashcroft, A.E., (2010). Determining the topology of virus assembly intermediates using ion mobility spectrometry-mass spectrometry. *Rapid Commun. Mass Spectrom.* **24**, 3033–3042.
5. Borodavka, A., Tuma, R., Stockley, P.G., (2012). Evidence that viral RNAs have evolved for efficient, two-stage packaging. *Proc. Natl. Acad. Sci. U S A* **109**, 15769–15774.
6. Dykeman, E.C., Stockley, P.G., Twarock, R., (2014). Solving a Levinthal's paradox for virus assembly identifies a unique antiviral strategy. *Proc. Natl. Acad. Sci. U S A* **111**, 5361–5366.
7. Rolfsson, O. et al, (2016). Direct Evidence for Packaging Signal-Mediated Assembly of Bacteriophage MS2. *J. Mol. Biol.* **428**, 431–448.
8. Shakeel, S. et al, (2017). Genomic RNA folding mediates assembly of human parechovirus. *Nat. Commun.* **8**, 5.
9. Patel, N. et al, (2017). Rewriting nature's assembly manual for a ssRNA virus. *Proc. Natl. Acad. Sci. U S A* **114**, 12255–12260.
10. Patel, N. et al, (2017). HBV RNA pre-genome encodes specific motifs that mediate interactions with the viral core protein that promote nucleocapsid assembly. *Nat. Microbiol.* **2**, 17098.
11. Twarock, R., Stockley, P.G., (2019). RNA-Mediated Virus Assembly: Mechanisms and Consequences for Viral Evolution and Therapy. *Annu. Rev. Biophys.* **48**, 495–514.
12. Chandler-Bostock, R. et al, (2020). Assembly of infectious enteroviruses depends on multiple, conserved genomic RNA-coat protein contacts. *PLoS Pathog.* **16**, e1009146.
13. Patel, N. et al, (2021). In vitro functional analysis of gRNA sites regulating assembly of hepatitis B virus. *Commun. Biol.* **4**, 1407.
14. Sokoloski, K.J. et al, (2017). Identification of Interactions between Sindbis Virus Capsid Protein and Cytoplasmic vRNA as Novel Virulence Determinants. *PLoS Pathog.* **13**, e1006473.
15. Brown, R.S., Anastasakis, D.G., Hafner, M., Kielian, M., (2020). Multiple capsid protein binding sites mediate selective packaging of the alphavirus genomic RNA. *Nat. Commun.* **11**, 4693.
16. Brown, R.S., Kim, L., Kielian, M., (2021). Specific Recognition of a Stem-Loop RNA Structure by the Alphavirus Capsid Protein. *Viruses* **13**
17. Kiser, L.M., Sokoloski, K.J., Hardy, R.W., (2021). Interactions between capsid and viral RNA regulate Chikungunya virus translation in a host-specific manner. *Virology* **560**, 34–42.
18. Twarock, R., Leonov, G., Stockley, P.G., (2018). Hamiltonian path analysis of viral genomes. *Nat. Commun.* **9**, 2021.
19. 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. U S A* **109**, 1907–1912.
20. Tetter, S. et al, (2021). Evolution of a virus-like architecture and packaging mechanism in a repurposed bacterial protein. *Science* **372**, 1220–1224.
21. 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.
22. Dykeman, E.C., Stockley, P.G., Twarock, R., (2013). Packaging signals in two single-stranded RNA viruses imply a conserved assembly mechanism and geometry of the packaged genome. *J. Mol. Biol.* **425**, 3235–3249.
23. Dykeman, E.C. et al, (2011). Simple rules for efficient assembly predict the layout of a packaged viral RNA. *J. Mol. Biol.* **408**, 399–407.
24. Eigen, M., Biebricher, C.K., Gebinoga, M., Gardiner, W.C., (1991). The hypercycle. Coupling of RNA and protein biosynthesis in the infection cycle of an RNA bacteriophage. *Biochemistry* **30**, 11005–11018.
25. Witherell, G.W., Gott, J.M., Uhlenbeck, O.C., (1991). Specific interaction between RNA phage coat proteins and RNA. *Prog. Nucleic Acid Res. Mol. Biol.* **40**, 185–220.
26. Carey, J., Uhlenbeck, O.C., (1983). Kinetic and thermodynamic characterization of the R17 coat protein-ribonucleic acid interaction. *Biochemistry* **22**, 2610–2615.
27. Dykeman, E.C., Twarock, R., (2010). All-atom normal-mode analysis reveals an RNA-induced allostery in a bacteriophage coat protein. *Phys. Rev. E: Stat. Nonlinear Soft Matter Phys.* **81**, 031908.
28. Valegård, K., Murray, J.B., Stockley, P.G., Stonehouse, N. J., Liljas, L., (1994). Crystal structure of an RNA bacteriophage coat protein-operator complex. *Nature* **371**, 623–626.
29. Valegård, K., Liljas, L., Fridborg, K., Unge, T., (1990). The three-dimensional structure of the bacterial virus MS2. *Nature* **345**, 36–41.

30. Golmohammadi, R., Valegård, K., Fridborg, K., Liljas, L., (1993). The refined structure of bacteriophage MS2 at 2.8 A resolution. *J. Mol. Biol.* **234**, 620–639.
31. Koning, R.I. et al, (2016). Asymmetric cryo-EM reconstruction of phage MS2 reveals genome structure in situ. *Nat. Commun.* **7**, 12524.
32. Zhong, Q. et al, (2016). Genetic, Structural, and Phenotypic Properties of MS2 Coliphage with Resistance to ClO(2) Disinfection. *Environ. Sci. Technol.* **50**, 13520–13528.
33. Dai, X. et al, (2017). In situ structures of the genome and genome-delivery apparatus in a single-stranded RNA virus. *Nature* **541**, 112–116.
34. Sclavi, B., Sullivan, M., Chance, M.R., Brenowitz, M., Woodson, S.A., (1998). RNA folding at millisecond intervals by synchrotron hydroxyl radical footprinting. *Science* **279**, 1940–1943.
35. Adilakshmi, T., Soper, S.F.C. & Woodson, S.A. (2009) In Biophysical, Chemical, and Functional Probes of RNA Structure, Interactions and Folding: Part A, pp. 239–258.
36. Asuru, A. et al, (2019). The XFP (17-BM) beamline for X-ray footprinting at NSLS-II. *J. Synchrotron Radiat.* **26**, 1388–1399.
37. Beekwilder, M.J., Nieuwenhuizen, R., van Duin, J., (1995). Secondary structure model for the last two domains of single-stranded RNA phage Q beta. *J. Mol. Biol.* **247**, 903–917.
38. Groeneveld, H., Thimon, K., van Duin, J., (1995). Translational control of maturation-protein synthesis in phage MS2: a role for the kinetics of RNA folding? *RNA* **1**, 79–88.
39. Olsthoorn, R.C., van Duin, J., (1996). Evolutionary reconstruction of a hairpin deleted from the genome of an RNA virus. *Proc. Natl. Acad. Sci. U S A* **93**, 12256–12261.
40. Dent, K.C. et al, (2013). The asymmetric structure of an icosahedral virus bound to its receptor suggests a mechanism for genome release. *Structure* **21**, 1225–1234.
41. Meng, R. et al, (2019). Structural basis for the adsorption of a single-stranded RNA bacteriophage. *Nat. Commun.* **10**, 3130.
42. Gorzelnik, K.V., Zhang, J., (2021). Cryo-EM reveals infection steps of single-stranded RNA bacteriophages. *Prog. Biophys. Mol. Biol.* **160**, 79–86.
43. Hill, H.R., Stonehouse, N.J., Fonseca, S.A., Stockley, P.G., (1997). Analysis of phage MS2 coat protein mutants expressed from a reconstituted phagemid reveals that proline 78 is essential for viral infectivity. *J. Mol. Biol.* **266**, 1–7.
44. Stonehouse, N.J. et al, (1996). Crystal structures of MS2 capsids with mutations in the subunit FG loop. *J. Mol. Biol.* **256**, 330–339.
45. Adilakshmi, T., Bellur, D.L., Woodson, S.A., (2008). Concurrent nucleation of 16S folding and induced fit in 30S ribosome assembly. *Nature* **455**, 1268–1272.
46. Deigan, K.E., Li, T.W., Mathews, D.H., Weeks, K.M., (2009). Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U S A* **106**, 97–102.
47. Low, J.T., Weeks, K.M., (2010). SHAPE-directed RNA secondary structure prediction. *Methods* **52**, 150–158.
48. Karabiber, F., McGinnis, J.L., Favorov, O.V., Weeks, K.M., (2013). QuShape: rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *RNA* **19**, 63–73.
49. Boerneke, M.A., Ehrhardt, J.E., Weeks, K.M., (2019). Physical and Functional Analysis of Viral RNA Genomes by SHAPE. *Annu. Rev. Virol.* **6**, 93–117.
50. Ding, Y., Chan, C.Y., Lawrence, C.E., (2004). Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res.* **32**, W135–W141.
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. Zarringhalam, K., Meyer, M.M., Dotu, I., Chuang, J.H., Clote, P., (2012). Integrating chemical footprinting data into RNA secondary structure prediction. *PLoS ONE* **7**, e45160.
53. Fiers, W. et al, (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**, 500–507.
54. Borodavka, A., Tuma, R., Stockley, P.G., (2013). A two-stage mechanism of viral RNA compaction revealed by single molecule fluorescence. *RNA Biol.* **10**, 481–489.
55. Chang, J.Y. et al, (2020). Hierarchical natural move Monte Carlo refines flexible RNA structures into cryo-EM densities. *RNA* **26**, 1755–1766.
56. Basnak, G. et al, (2010). Viral genomic single-stranded RNA directs the pathway toward a T=3 capsid. *J. Mol. Biol.* **395**, 924–936.
57. Hirao, I., Spingola, M., Peabody, D., Ellington, A.D., (1998). The limits of specificity: an experimental analysis with RNA aptamers to MS2 coat protein variants. *Mol Divers* **4**, 75–89.
58. Borer, P.N. et al, (1995). Proton NMR and structural features of a 24-nucleotide RNA hairpin. *Biochemistry* **34**, 6488–6503.
59. Chen, Z.G. et al, (1989). Protein-RNA interactions in an icosahedral virus at 3.0 A resolution. *Science* **245**, 154–159.
60. Harb, L. et al, (2020). ssRNA phage penetration triggers detachment of the F-pilus. *Proc. Natl. Acad. Sci.* **117**, 25751–25758.
61. Chen, M. et al, (2020). Surveying biomolecular frustration at atomic resolution. *Nat. Commun.* **11**, 5944.
62. Zhang, B., Wolynes, P.G., (2017). Genomic Energy Landscapes. *Biophys. J.* **112**, 427–433.
63. Ashcroft, A.E. et al, (2005). Engineering thermal stability in RNA phage capsids via disulphide bonds. *J. Nanosci. Nanotechnol.* **5**, 2034–2041.
64. Gell, C. et al, (2008). Single-molecule fluorescence resonance energy transfer assays reveal heterogeneous folding ensembles in a simple RNA stem-loop. *J. Mol. Biol.* **384**, 264–278.
65. Valegård, K. et al, (1997). The three-dimensional structures of two complexes between recombinant MS2 capsids and RNA operator fragments reveal sequence-specific protein-RNA interactions. *J. Mol. Biol.* **270**, 724–738.