doi:10.1038/nature20589

In situ structures of the genome and genome-delivery apparatus in a single-stranded RNA virus

Xinghong Dai[1,2], Zhihai Li[2,3,4], Mason Lai[3], Sara Shu[1], Yushen Du[1], Z. Hong Zhou[2,3] & Ren Sun[1,2]

Packaging of the genome into a protein capsid and its subsequent delivery into a host cell are two fundamental processes in the life cycle of a virus. Unlike double-stranded DNA viruses, which pump their genome into a preformed capsid[1–3], single-stranded RNA (ssRNA) viruses, such as bacteriophage MS2, co-assemble their capsid with the genome[4–7]; however, the structural basis of this co-assembly is poorly understood. MS2 infects *Escherichia coli* via the host 'sex pilus' (F-pilus)[8]; it was the first fully sequenced organism[9] and is a model system for studies of translational gene regulation[10,11], RNA–protein interactions[12–14], and RNA virus assembly[15–17]. Its positive-sense ssRNA genome of 3,569 bases is enclosed in a capsid with one maturation protein monomer and 89 coat protein dimers arranged in a $T=3$ icosahedral lattice[18,19]. The maturation protein is responsible for attaching the virus to an F-pilus and delivering the viral genome into the host during infection[8], but how the genome is organized and delivered is not known. Here we describe the MS2 structure at 3.6 Å resolution, determined by electron-counting cryo-electron microscopy (cryoEM) and asymmetric reconstruction. We traced approximately 80% of the backbone of the viral genome, built atomic models for 16 RNA stem–loops, and identified three conserved motifs of RNA–coat protein interactions among 15 of these stem–loops with diverse sequences. The stem–loop at the 3′ end of the genome interacts extensively with the maturation protein, which, with just a six-helix bundle and a six-stranded β-sheet, forms a genome-delivery apparatus and joins 89 coat protein dimers to form a capsid. This atomic description of genome–capsid interactions in a spherical ssRNA virus provides insight into genome delivery via the host sex pilus and mechanisms underlying ssRNA–capsid co-assembly, and inspires speculation about the links between nucleoprotein complexes and the origins of viruses.

We imaged MS2 particles embedded in vitreous ice with a K2 direct electron detector attached to the end of an energy filter in a Titan Krios electron microscope and averaged more than 330,000 particles to calculate a 3.6 Å resolution reconstruction without imposing any symmetry (Extended Data Fig. 1). The external structure of the asymmetric reconstruction (Fig. 1a) appears to be similar to the icosahedrally averaged crystallographic structure of MS2, which contains 90 coat protein dimers[18,20]. However, a small density bulge emerges at one of the two-fold axes and its structure differs from all of the other 89 coat protein dimers (Fig. 1b, Supplementary Video 1). Initial analysis of secondary structures of this asymmetric feature and more detailed amino-acid tracing indicated it to be a maturation protein monomer. High-resolution structural features in the maturation protein, such as side chains, are consistent with the estimated 3.6 Å resolution of the map (Fig. 1d).

The reconstruction also reveals the ssRNA genome density inside the capsid. This density is a bit noisy and somewhat broken in the 3.6 Å resolution map, but structural features typical of RNA are apparent when the map is low-pass filtered to 6 Å resolution (Fig. 1c, Supplementary Video 1), indicating that the RNA chain shows some flexibility relative to the capsid shell. To identify possible structural heterogeneity due to the flexible RNA, we carried out further 3D classification of the dataset and obtained ten structures. These structures were almost identical to each other except for several short, possibly flexible, segments (Extended Data Fig. 2, Supplementary Video 2). The RNA density was not uniformly distributed within the capsid, with the maturation protein-proximal side of the space more densely packed than the distal side (Fig. 1c). The majority of the density shows prominent major and minor grooves (Fig. 1c), hallmarks of double helices, indicating that most of the ssRNA has folded into stem–loops. We identified more than 50 stem–loops, most of which contact the capsid at the tip (that is, the loop region; Fig. 1c).

Among the stem–loops identified, 16 showed clearly resolved individual nucleotides and even features that distinguish purines from pyrimidines (Fig. 1d). Such features enabled us to derive possible sequences for each stem–loop, which were used to search against the viral genome to identify its genuine sequence. Using the identified sequences of these 16 stem–loops as landmarks across the genome, we were able to trace the backbone for more than 80% of the genome, build atomic models for the 16 stem–loops, and identify long-range base-pairing and kissing-loop interactions (Fig. 2, Extended Data Figs 3–9, and Supplementary Data 1–3).

Among the 16 stem–loops, the one at the 3′ end of the genome interacts with the maturation protein and each of the remaining 15 binds to a coat protein dimer (Extended Data Fig. 9). The 15 stem–loops vary greatly in their lengths and sequences (Fig. 3a). Comparison of the 15 stem–loop–coat protein dimer structures reveals how these different stem–loops can be specifically recognized by the same coat protein during genome packaging. The non-sequence-specific, negatively charged phosphates of the stem–loop RNA backbone interact with the coat protein dimer through a patch of positively charged/polar residues on the two coat protein monomers (Lys43, Arg49, Ser51, Lys57 and Lys61 of one coat protein, and Arg49, Ser51, Ser52, Asn55, Lys57 and Lys61 of the other; Fig. 3b, c). At individual base level, three conserved motifs of RNA–coat protein interactions can be identified (Fig. 3d, e, Supplementary Video 3). First, Thr45 and Ser47 of a coat protein form hydrogen bonds with the base of a partially conserved nucleotide (red in Fig. 3a) in the RNA loop (Fig. 3e), preferably an adenine with two hydrogen bonds (for example, A1757; Fig. 3e), and alternatively a cytosine as in 3 of the 15 cases with only one hydrogen bond (for example, C109; Fig. 3h). Second, in 4 of the 15 cases (magenta in Fig. 3a), Thr45 and Ser47 of the second coat protein also form two hydrogen bonds with an unpaired and sticking-out base (a purine in all these four cases) in the stem region (Fig. 3d, f). However, this purine has a flipped conformation as compared to that in the loop region (compare A1751 in Fig. 3d or G2784 in Fig. 3f with A1757 in Fig. 3e). Third, Tyr85 of one coat protein establishes an aromatic-ring stacking interaction with another base in the loop region (for

[1] Department of Molecular and Medical Pharmacology, University of California, Los Angeles (UCLA), Los Angeles, California 90095, USA. [2] The California NanoSystems Institute (CNSI), UCLA, Los Angeles, California 90095, USA. [3] Department of Microbiology, Immunology and Molecular Genetics, UCLA, Los Angeles, California 90095, USA. [4] State Key Laboratory of Molecular Vaccinology and Molecular Diagnostics, School of Life Sciences, Xiamen University, Xiamen, Fujian 361102, China.

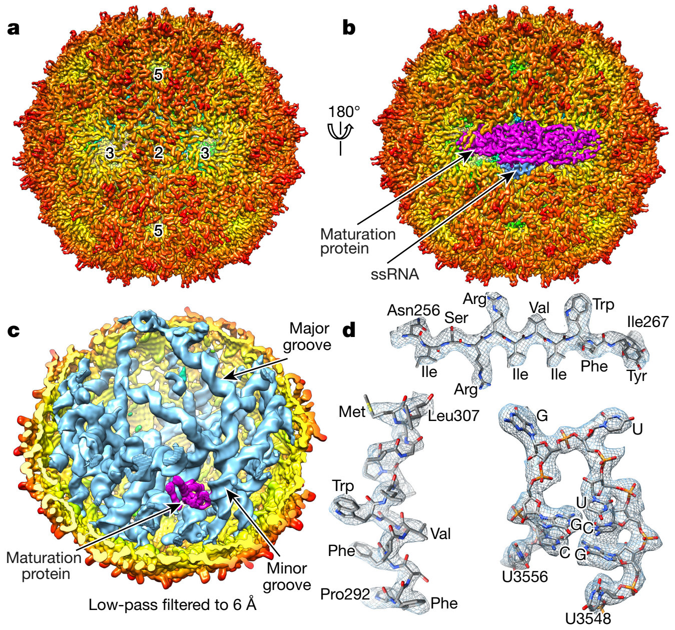


Figure 1 | CryoEM asymmetric reconstruction of MS2 at 3.6 Å resolution. a, **b**, Front (**a**) and back (**b**) views of the cryoEM density map along an icosahedral two-fold symmetry axis with some two-, three- and five-fold axes indicated. The capsid shell is radially coloured with the maturation protein highlighted in magenta and the ssRNA genome inside the capsid in blue. **c**, Cut-open view with half of the capsid shell removed to expose the genome. **d**, Segmented cryoEM densities (mesh) superimposed with their corresponding atomic models (sticks). Top and bottom left, typical β-strand and α-helix densities, respectively, from the maturation protein. Bottom right, part of a maturation protein-bound RNA stem–loop. Purines and pyrimidines are readily distinguishable.

example, U1756; Fig. 3e). This base (green in Fig. 3a) is preferentially a pyrimidine, but is a purine in 4 of the 15 cases. It can be either next to the base that interacts with Thr45 and Ser47 (Fig. 3e) or separated by a spacer (Fig. 3g). The aromatic-ring stacking is usually extended to include another unpaired base (for example, A1754; Fig. 3e) that is also in the loop region, before merging with the base stacking in the stem region (Fig. 3e). In addition to accommodation of diversities in sequence, these stem–loops also utilize the above described non-sequence-specific interactions and conserved motifs to accommodate differences in local environments (Fig. 3i, j).

Although each of the more than 50 stem–loops in the MS2 genome might act as a 'packaging signal' for capsid assembly, as previously proposed[15–17,21], the higher resolution of the above-described 16 stem–loops indicates that they interact more strongly with capsid proteins, and therefore might have more important roles in capsid assembly. Three of these 16 stem–loops are consecutive in sequence (stem–loops 1714–1737, 1746–1764, and 1766–1806, Extended Data Fig. 8), cluster together (stem loops A–C in Extended Data Fig. 5) and bind three neighbouring coat protein dimers—a configuration that would be desirable for nucleating capsid assembly. Indeed, the middle one, which encompasses the start codon of the replicase gene, was proposed to nucleate capsid assembly[12,22]. Interestingly, the 16 stem–loops are non-uniformly distributed, with most of them clustered around the putative nucleation site (Fig. 2a). Because coat protein dimers alone can assemble into both octahedral and wild-type-like icosahedral particles[23], guidance by the RNA stem–loops in the early stage of capsid assembly might be the size-determining factor leading to the assembly of a wild-type, icosahedral capsid.

One maturation protein monomer replaces a coat protein dimer at one of the icosahedral two-fold axes, imparting structural changes to the neighbouring coat proteins (Fig. 4a). The maturation protein structure consists of an α-helix domain (amino acids 140–225, 269–313, and 375–393) with a bundle of six α-helices, and a β-sheet

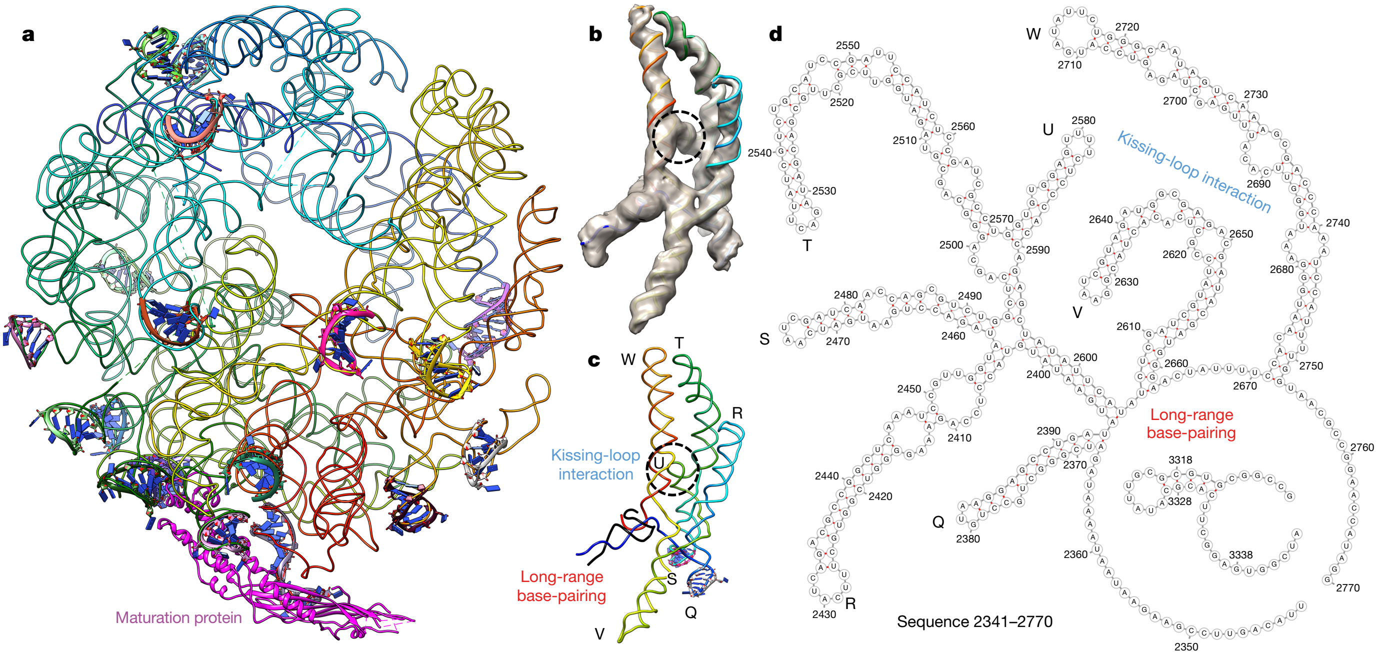


Figure 2 | Modelling the ssRNA genome. a, Backbone structure of the genome (wire) and non-uniform distribution of the high-resolution stem–loops (ribbons). Backbone is rainbow-coloured (blue to red) from 5′ to 3′. **b**–**d**, Example of tracing RNA backbone. Part of the genome density (grey in **b**) is segmented out and superimposed with its backbone model (rainbow-coloured wire, blue to red from 5′ to 3′; **b**, **c**). For each of the two high-resolution stem–loops (ribbons in **c**) contained in this segment, a degenerate sequence was derived on the basis of the resolved bases and used to search against the genome to identify sequence candidates. Each of these short sequence candidates was expanded in both directions to include about 500 bases for secondary structure prediction. The predicted secondary structure was then correlated with the backbone obtained in **b** and only one of these sequence candidates yielded the correct sequence registration of individual stem–loops (indicated by letters Q–W in **c**, **d**). The backbone model reveals kissing-loop and long-range base-pairing interactions as indicated.

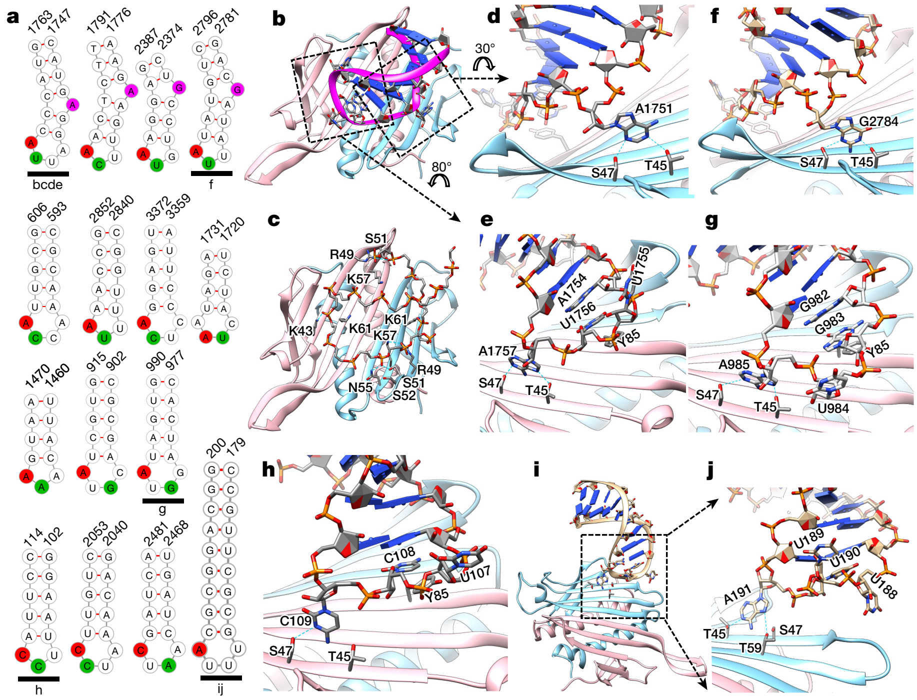


Figure 3 | Conserved interaction motifs between RNA stem–loops and coat protein dimers. a, Secondary structures of RNA stem–loops with nucleotides involved in the three types of conserved interaction motif coloured. Letters beneath some of the stem–loops identify panels in which the atomic model for that stem–loop is shown. **b**–**e**, Atomic model of stem–loop 1747–1763 and its interactions with a coat protein dimer (pink and sky blue ribbons). In **c**, positively charged or polar residues of the coat protein dimer interacting with phosphates of the RNA backbone (sticks) are indicated. Expanded views of the stem (**d**) and loop (**e**) regions show the interaction motifs conserved among the 15 stem–loops. **f**–**j**, Accommodation of diversities in sequence or local environment. Stem–loop 2781–2796 (**f**), viewed in the same orientation as in **d**, shows that a guanine forms the same kind of hydrogen bond with Thr45 and Ser47 as an adenine in **d**. Stem–loops 977–990 (**g**) and 102–114 (**h**), viewed in the same orientation as in **e**, show that a purine (G983 in **g**) instead of a pyrimidine (U1756 in **e**) stacks with Tyr85 and that a pyrimidine (C109 in **h**) forms only one hydrogen bond instead of a purine (A1757 in **e**) forming two hydrogen bonds with Thr45 and Ser47. Stem–loop 179–200 (**i**, **j**) binds to a coat protein dimer from a very different angle owing to steric hindrance of a neighbouring stem–loop (not shown). Nonetheless, the RNA fold and one of the three interaction motifs are conserved, although a hydrogen bond is formed with Thr59 instead of Ser47.

domain (amino acids 1–139, 226–268, and 314–374) with six antiparallel β-strands sandwiched between an N-terminal loop and a helix–loop–helix motif (Fig. 4b). The maturation protein is slightly tilted off the capsid surface with its α-helix domain inserting into the capsid lumen and its β-sheet domain projecting out (Fig. 4a, Supplementary Video 1).

The maturation protein has extensive (more than 20) interactions with one of the 16 high-resolution stem–loops of the genome structure (stem–loop 3540–3563 at the 3′ end of the genome; Fig. 4c, d). These interactions can be categorized into four types (Fig. 4e–h). First, phosphates in the RNA backbone can form charge–charge interactions with basic residues in the maturation protein (Lys50 and Arg43, 62, 358, and 229; Fig. 4e, f), or form hydrogen bonds with polar residues (Tyr60, Trp264, and Gln152; Fig. 4e, f, h) or even with an amide of the protein backbone (Val141; Fig. 4f). Second, there are two cases of RNA base stacking with an aromatic ring of the protein side chain: U3553 with Phe4 and U3557 with Trp54 (Fig. 4f, g). Third, there are multiple hydrogen bonds between RNA bases and protein side chains, primarily serines and threonines. These include interactions between U3549 and Thr47, U3552 and Asn45 (Fig. 4e), G3554 and Ser258/Thr324, and G3555 and Ser58 (Fig. 4f). Last, seven hydrogen bonds are formed between RNA bases and carboxyl or amide groups of the protein backbone in the long loop connecting the first two β-strands of the maturation protein (Fig. 4e, g). In this regard, Pro57 and Pro59 (Fig. 4g) might be important for maintaining the twisted conformation of the protein backbone, preventing the formation of β-strand interactions in this segment, and freeing those carboxyl and amide groups for hydrogen bonding with the RNA bases. It is also noteworthy that the base pairing in the maturation protein-bound state of this stem–loop is not optimal. If the maturation protein binds to the RNA after the stem–loop has been folded, four base pairs have to be melted and two new ones formed (Fig. 4d) to accommodate the RNA–maturation protein interactions described above. How this dynamic process is achieved remains to be determined.

This maturation protein-interacting stem–loop also interacts with coat proteins, albeit much less extensively and in a manner that bears no similarity to that of the other 15 stem–loop–coat protein dimer structures. Two phosphates in the stem region of this stem–loop interact with the same coat protein (yellow in Fig. 4h): one through a hydrogen bond with Asn27 and the other through a charge–charge interaction with Arg49. Base C3548, which sticks out from the loop region, forms a hydrogen bond with the terminal carboxylic acid group of Tyr129 of a second coat protein (beige in Fig. 4h). It may also stack with the Tyr129 aromatic ring of a third coat protein (purple in Fig. 4h).

Apart from its extensive interactions with the 3′ end stem–loop, the maturation protein also binds the backbone of three other stem–loops with positively charged Arg and Lys residues (Fig. 5a, b). Three arginines in the β-sheet domain of the maturation protein bind a very short stem–loop (3524–3532) in the 3′ untranslated region (3′-UTR) of the RNA. The other six arginines and two lysines are clustered at the tip of the α-helix bundle and bind stem–loops 1766–1806 and 1960–1995, with Arg188 inserted into the minor groove of stem–loop 1960–1995 (Fig. 5a).

During infection, the maturation protein is responsible for attaching the MS2 virion to the bacterial F-pilus and delivering the genome into the host[8]. Fitting our MS2 model into the structure of the MS2–F-pilus complex[19] revealed interactions between the F-pilus and the

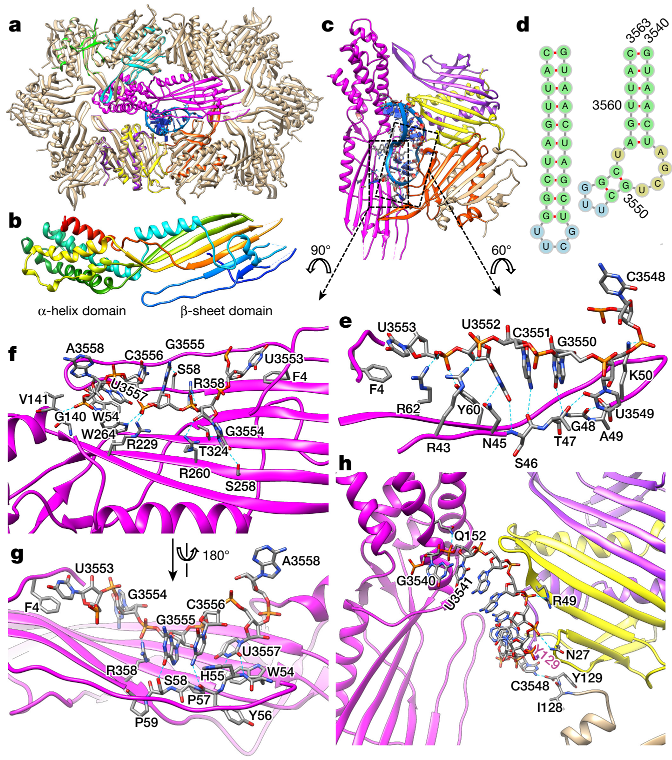


Figure 4 | Maturation protein and its interactions with the 3′-end stem–loop. a, Incorporation of maturation protein into the capsid shell. The maturation protein (magenta) replaces a coat protein dimer at a two-fold symmetry axis and induces structural changes of neighbouring coat proteins. Atomic models of the changed coat proteins (coloured ribbons) are superimposed with those of coat proteins at other two-fold symmetry axes (beige ribbons) that are unaffected by the maturation protein. **b**, Maturation protein model, rainbow-coloured (blue to red from N to C terminus). **c**, Binding of the 3′-end stem–loop to the maturation protein and neighbouring coat proteins as viewed inside the capsid. **d**, Base-pairing in the 3′-end stem–loop as observed in our structure interacting with maturation protein (right), or theoretically as a free stem–loop (left). **e**–**h**, Details of the interactions between the 3′-end stem–loop and the maturation protein or coat proteins. **h**, Expanded view of the RNA stem region with half of the stem hidden for clarity.

helix–loop–helix motif on one side of the β-sheet domain of the maturation protein. All of the β-strands were nearly parallel to the pilus axis (Fig. 5c), resulting in slight tilting of the virion, as observed previously[19,24] (Fig. 5d). Our results shed light on why genome release occurs only when MS2 binds to an F-pilus on a living bacterium[25], but not when it binds to a detached F-pilus[8]. Considering that F-pili in living bacteria are highly dynamic, with extension and retraction accompanied by rotation[26], the driving force for MS2 genome delivery might come from the dynamics of, rather than the binding to, the F-pilus. As the pilus retracts, the virion would get stuck outside the cell owing to its relatively large size, but the maturation protein and its tightly bound ssRNA genome are pulled together out of the capsid shell, leading to the delivery of this ribonucleoprotein complex into the host. Indeed, the maturation protein is delivered into the host cell with the genome during infection[27,28] while the empty MS2 capsid is left outside[29], and infectious ribonucleoprotein complexes can be reconstituted by mixing the maturation protein and ssRNA genome of MS2[30]. One might even imagine that after the proposed primitive 'RNA world', such a minimalist ribonucleoprotein complex could constitute a simple replicator to set the stage for evolution towards more sophisticated complexes. Selective pressure towards better fitness of the replicator naturally could have led to the acquisition of a protein coat to protect the RNA from a hostile environment.

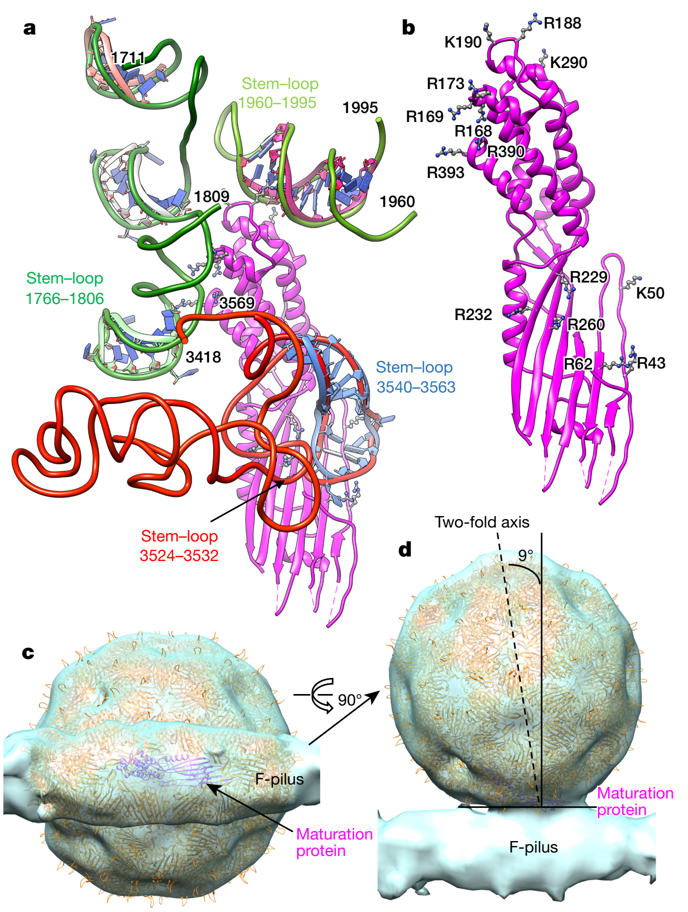


Figure 5 | Binding of maturation protein to genome and bacterial F-pilus. a, Overview of maturation protein (magenta) with surrounding RNA stem–loops (wires coloured as in Fig. 2a) shown in the same orientation as in Fig. 3c. **b**, As in **a** but without the RNA backbone model, showing the distribution of positively charged arginine and lysine residues in the maturation protein that bind RNA stem–loops. **c**, **d**, Fitting of our atomic models (ribbons) of the MS2 virion into a tomographic reconstruction (EMD-2365, semitransparent surface) of MS2 attached to bacterial F-pilus[19]. The maturation protein projects obliquely out from the capsid surface, resulting in a slight tilt (approximately 9°) of the MS2 virion when attached to the F-pilus (**d**).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 August 2016; accepted 26 October 2016.

Published online 19 December 2016.

1. Jiang, W. *et al.* Structure of epsilon15 bacteriophage reveals genome organization and DNA packaging/injection apparatus. *Nature* **439,** 612–616 (2006).
2. Lander, G. C. *et al.* The structure of an infectious P22 virion shows the signal for headful DNA packaging. *Science* **312,** 1791–1795 (2006).
3. Catalano, C. E. *Viral Genome Packaging Machines: Genetics, Structure, and Mechanism* (Springer, 2005).
4. Basnak, G. *et al.* Viral genomic single-stranded RNA directs the pathway toward a *T*=3 capsid. *J. Mol. Biol.* **395,** 924–936 (2010).
5. Sun, S., Rao, V. B. & Rossmann, M. G. Genome packaging in viruses. *Curr. Opin. Struct. Biol.* **20,** 114–120 (2010).
6. Perlmutter, J. D. & Hagan, M. F. Mechanisms of virus assembly. *Annu. Rev. Phys. Chem.* **66,** 217–239 (2015).
7. Klug, A. The tobacco mosaic virus particle: structure and assembly. *Philos. Trans. R. Soc. Lond. B* **354,** 531–535 (1999).
8. Valentine, R. C. & Strand, M. Complexes of F-pili and RNA bacteriophage. *Science* **148,** 511–513 (1965).
9. Fiers, W. *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260,** 500–507 (1976).

10. Schmidt, B. F., Berkhout, B., Overbeek, G. P., van Strien, A. & van Duin, J. Determination of the RNA secondary structure that regulates lysis gene expression in bacteriophage MS2. *J. Mol. Biol.* **195,** 505–516 (1987).
11. Poot, R. A., Tsareva, N. V., Boni, I. V. & van Duin, J. RNA folding kinetics regulates translation of phage MS2 maturation gene. *Proc. Natl Acad. Sci. USA* **94,** 10110–10115 (1997).
12. Valegård, K., Murray, J. B., Stockley, P. G., Stonehouse, N. J. & Liljas, L. Crystal structure of an RNA bacteriophage coat protein-operator complex. *Nature* **371,** 623–626 (1994).
13. Convery, M. A. *et al.* Crystal structure of an RNA aptamer-protein complex at 2.8 Å resolution. *Nat. Struct. Biol.* **5,** 133–139 (1998).
14. Rowsell, S. *et al.* Crystal structures of a series of RNA aptamers complexed to the same protein target. *Nat. Struct. Biol.* **5,** 970–975 (1998).
15. Borodavka, A., Tuma, R. & Stockley, P. G. Evidence that viral RNAs have evolved for efficient, two-stage packaging. *Proc. Natl Acad. Sci. USA* **109,** 15769–15774 (2012).
16. Patel, N. *et al.* Revealing the density of encoded functions in a viral RNA. *Proc. Natl Acad. Sci. USA* **112,** 2227–2232 (2015).
17. Rolfsson, Ó. *et al.* Direct evidence for packaging signal-mediated assembly of bacteriophage MS2. *J. Mol. Biol.* **428,** 431–448 (2016).
18. Valegård, K., Liljas, L., Fridborg, K. & Unge, T. The three-dimensional structure of the bacterial virus MS2. *Nature* **345,** 36–41 (1990).
19. Dent, K. C. *et al.* The asymmetric structure of an icosahedral virus bound to its receptor suggests a mechanism for genome release. *Structure* **21,** 1225–1234 (2013).
20. Golmohammadi, R., Valegård, K., Fridborg, K. & Liljas, L. The refined structure of bacteriophage MS2 at 2.8 A resolution. *J. Mol. Biol.* **234,** 620–639 (1993).
21. Dykeman, E. C., Stockley, P. G. & Twarock, R. 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 (2013).
22. Witherell, G. W., Gott, J. M. & Uhlenbeck, O. C. Specific interaction between RNA phage coat proteins and RNA. *Prog. Nucleic Acid Res. Mol. Biol.* **40,** 185–220 (1991).
23. Plevka, P., Tars, K. & Liljas, L. Crystal packing of a bacteriophage MS2 coat protein mutant corresponds to octahedral particles. *Protein Sci.* **17,** 1731–1739 (2008).
24. Toropova, K., Stockley, P. G. & Ranson, N. A. Visualising a viral RNA genome poised for release from its receptor complex. *J. Mol. Biol.* **408,** 408–419 (2011).
25. Danziger, R. E. & Paranchych, W. Stages in phage R17 infection. 3. Energy requirements for the F-pili mediated eclipse of viral infectivity. *Virology* **40,** 554–564 (1970).
26. Clarke, M., Maddera, L., Harris, R. L. & Silverman, P. M. F-pili dynamics by live-cell imaging. *Proc. Natl Acad. Sci. USA* **105,** 17978–17981 (2008).
27. Krahn, P. M., O'Callaghan, R. J. & Paranchych, W. Stages in phage R17 infection. VI. Injection of A protein and RNA into the host cell. *Virology* **47,** 628–637 (1972).
28. Kozak, M. & Nathans, D. Fate of maturation protein during infection by coliphage MS2. *Nat. New Biol.* **234,** 209–211 (1971).
29. Silverman, P. M. & Valentine, R. C. The RNA injection step of bacteriophage f2 infection. *J. Gen. Virol.* **4,** 111–124 (1969).
30. Shiba, T. & Miyake, T. New type of infectious complex of *E. coli* RNA phage. *Nature* **254,** 157–158 (1975).

Supplementary Information is available in the online version of the paper.

Acknowledgements This project was supported in part by grants from the National Institutes of Health (GM071940, DE025567, DE023591, CA177322 and AI094386) and National Science Foundation (DMR-1548924). We acknowledge the use of instruments at the Electron Imaging Center for Nanomachines (supported by UCLA and by instrumentation grants from the NIH (1S10OD018111, 1U24GM116792) and NSF (DBI-1338135)). X.D. and Z.L. were supported in part by fellowships from the China Scholarship Council. We appreciate critical reading of the manuscript by F. Guo.

Author Contributions X.D., Z.H.Z. and R.S. designed the project; X.D., Z.L. and S.S. prepared the sample and acquired cryoEM data; X.D. solved the structure; X.D. and M.L. built the model; X.D., M.L., Y.D., Z.H.Z. and R.S. interpreted the results; X.D. and Z.H.Z. wrote the paper; and all authors contributed to the editing of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Z.H.Z. (Hong.Zhou@UCLA.edu) for cryoEM and modelling or R.S. (RSun@mednet.ucla.edu) for virology.

Reviewer Information *Nature* thanks W. Dai and J. E. Johnson for their contribution to the peer review of this work.

METHODS

Sample preparation. Coliphage MS2 (ATCC 15597-B1) and its host *E. coli* strain C-3000 (ATCC 15597) were obtained from American Type Culture Collection (ATCC). Bacteria were cultured in ATCC-recommended broth medium at 37 °C and infected with MS2 phage at middle log-phase. After complete lysis of the bacteria within a few hours, the lysate was centrifuged at 10,000*g* for 15 min to remove cell debris, and then centrifuged again at 100,000*g* for 4 h to pellet the phage particles. The pellet was resuspended in buffer containing 50 mM Tris (pH7.5), 150 mM NaCl, 5 mM $CaCl_2$ and 5 mM $MgCl_2$. The suspension was applied to an OptiPrep (Sigma-Aldrich) density gradient (10–50% w/v with 10% steps) and centrifuged at 100,000*g* overnight. The phage band was clearly visible and was collected, diluted with buffer, and centrifuged at 100,000*g* for 4 h. The purified phage particles were resuspended in 100 μl buffer. To prepare the cryoEM sample, an aliquot of 2.5 μl phage sample was applied to a Quantifoil grid, blotted with filter paper and plunge-frozen in liquid ethane with an FEI Vitrobot.

Cryo-electron microscopy and movie preprocessing. CryoEM images were collected with Leginon[31] on an FEI Titan Krios electron microscope equipped with a Gatan imaging filter (GIF) and K2 Summit direct detection camera. A nominal magnification of 130,000× was used, giving a pixel size of 1.06 Å at the sample level. A slit width of 20 eV was set for the energy filter. Movies were recorded with the K2 camera operating in counting mode with an electron dose rate below $8e^-$ per pixel per s. An accumulated dose of $50e^-$ per Å^2 on the sample was fractionated into a movie stack of 29 image frames.

For each of the total 6,080 movies recorded, the frames were aligned for drift correction with the method described previously[32]. The images averaged from all 29 frames were used for initial model searching and structure refinement, while images averaged from the first 14 frames were used for calculation of the final density map. The defocus value was set to −2 μm during the imaging session, and was determined with CTFFIND3 (ref. 33) to be in the range of −0.5 to −3 μm in the images. Particles were picked with Ethan[34] and preprocessed with EMAN[35].

Asymmetric model generation. Two-dimensional classification of particle images was performed with refine2d of EMAN. The resulted 2D class averages were used to calculate an initial model with the *starticos* command of EMAN[35]. The resulting reconstruction consisted of two layers of smooth, featureless spherical shells, with the outer and inner layers corresponding to the averaged densities of the viral capsid and packaged RNA, respectively. This initial reconstruction was used to run 3D classification of the images using Relion[36], with C1 symmetry applied. Within fewer than 10 iterations, one of the emerged 3D classes showed densities of separated RNA chains inside the capsid. Three-dimensional classification was then started over again with the converged RNA-containing structure as the initial model. All of the resulting 10 classes showed similar structures with prominent RNA densities, suggesting that the RNA genome was indeed well organized inside the capsid of MS2 phage and the initial model was solid.

It is noteworthy that two recent attempts at cryoEM reconstruction of a segmented dsRNA virus relied on subtraction of the icosahedrally symmetric capsid contribution from the raw cryoEM images to minimize interference of the symmetric capsid in the orientation search for the asymmetric components[37,38]. Our success in generating the asymmetric model of MS2 without applying such computationally demanding subtraction methods suggests that the asymmetric RNA genome has sufficient power to drive asymmetric orientation search even in the presence of icosahedral capsid signal. This simpler computational strategy opens the door to modelling viral genomes and genome–capsid interactions in spherical viruses.

Structure refinement. The dataset was divided into two random halves and refined separately against the Relion-generated model using Frealign[39], considering that Frealign demands much less computational resources than Relion in processing such a large dataset. The refinement procedure included five rounds of grid search (mode 3) with data points lower than 20 Å resolution, followed by several iterations of local refinement (mode 1) with gradually increasing resolution range. Particle images binned for 4× or 2× were used throughout the refinement procedure to speed up calculations. The unbinned particle images from averaging 14 of the 29 frames in the drift-corrected image stacks were used in the last few iterations of the refinement. A final density map was calculated by merging the two half datasets, containing a total number of 339,718 particles. The average resolution was determined based on the 'gold-standard' FSC (Fourier shell correlation) = 0.143 criterion[40]. Local resolutions were assessed with ResMap[41] (Extended Data Fig. 1b).

Atomic model building for proteins. The atomic model of the maturation protein was built *ab initio* with Coot[42]. To model the coat proteins, the crystallographic structure with PDB ID 2MS2 (ref. 20) was fitted into the density map with Chimera[43], and then manually adjusted with Coot. Only 6 of the 178 copies of the coat protein needed partial modification. All the models were refined with the Phenix real space refinement program[44].

Backbone tracing and atomic model building for RNA. In modelling the 16 high-resolution RNA stem–loops, we took advantage of the available RNA moiety model in the crystallographic structure 2B2G[45]. This model was first fitted into our cryoEM densities, mutated into the genuine sequence, and then manually adjusted in Coot. To trace the backbone of the genome, we low-pass filtered the density map to 6 Å resolution so that the RNA backbone became visible and was manually traced with the baton mode in Coot. For some RNA densities that are weak in the final density map (that is, flexible) but show better quality in some of the 10 classes from 3D classification, the density map of the best class was used to guide the backbone tracing.

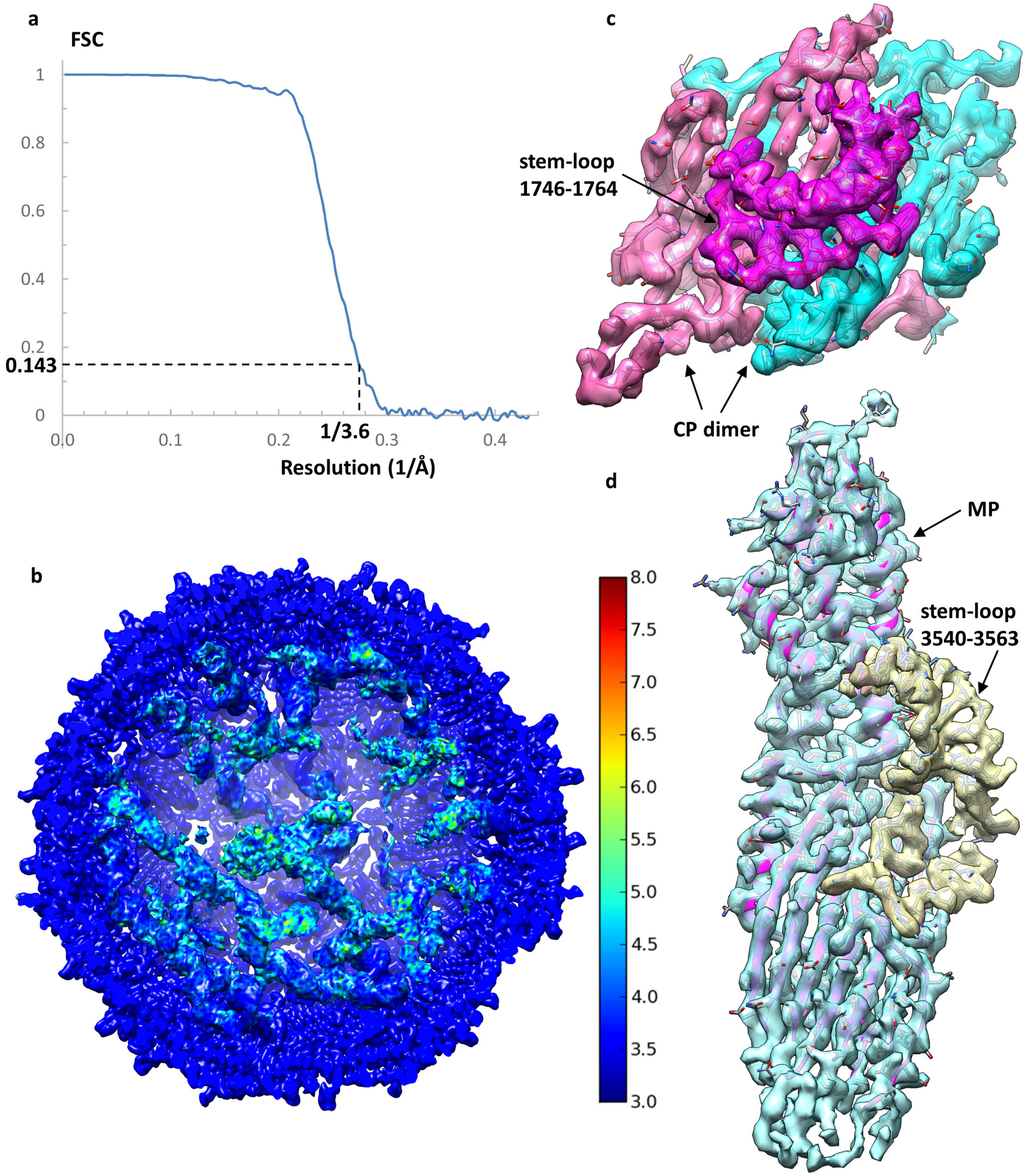
Three constraints were used in conjunction to determine the sequences of high-resolution stem–loops and to trace the genome backbone simultaneously. First, at 3.6 Å resolution, purines and pyrimidines are readily distinguishable based on the fact that the densities of purines are relatively fattier than those of pyrimidines (Fig. 1d). Therefore, a degenerate sequence can be derived for each stem–loop: R = A/G or Y = C/U was assigned to each nucleotide (nt) in high quality regions, while *N* = A/G/C/U was assigned to nucleotides that had poor density. This degenerate sequence was then searched against the MS2 genome with the JDSA program[46]. With a typical length of 9–12 nt for a high quality stem–loop, and considering that bases in the stem region should be paired (G–U was counted as paired in addition to the Watson–Crick pairs), the degenerate search usually produced fewer than three candidates, and a single hit was not uncommon. Constraint from the backbone tracing was then used to further narrow down the candidate sequence and/or to confirm the assignment. For two tandem stem–loops in the path of the tracing, if their assigned sequences are also in tandem and have a distance in the genome agreeing with that in the tracing, then both assignments are likely to be correct. A third constraint we used to validate the sequence assignments and backbone tracing was the secondary structure prediction of the RNA sequence. Although the full-length MS2 genome of 3,569 bases was too long for most RNA secondary structure prediction algorithms, we found that prediction results for sequences in the length of a few hundred bases are consistent among different prediction software and thus more reliable. Therefore, we divided the entire genome into several segments, predicted the secondary structure of each segment with the RNAfold webserver[47], and compared it with the corresponding backbone as identified by the assigned sequences of the high-resolution stem–loops contained in that segment. The predicted secondary structure and traced backbone matched for most of the genome (Fig. 2c, d, Extended Data Figs 3–7), thus confirming the validity of the backbone tracing and sequence assignments of the high-resolution stem–loops. From the correspondence between the traced backbone and the predicted secondary structure, we can also roughly assign sequences for most of the low-resolution stem–loops.

Modelling the high-resolution stem–loops individually was challenging owing to the relatively low resolution of the genome structure. Ambiguity may arise from multiple results of the degenerate sequence search in some cases and also from the generation of the degenerate sequence itself (that is, ambiguity in assigning purine or pyrimidine to some of the nucleotides based on features of the density at the current resolution). Tracing the backbone of the genome based solely on the 6 Å resolution map is impractical because there are numerous junctions and crossovers in the genome (see Fig. 2b for example). Our strategy combines the two levels of structural information with secondary structure prediction of the genome sequence to eliminate ambiguities by trial and error, greatly improving the reliability of the model. This strategy is generally applicable for modelling genome organization in many other viruses.

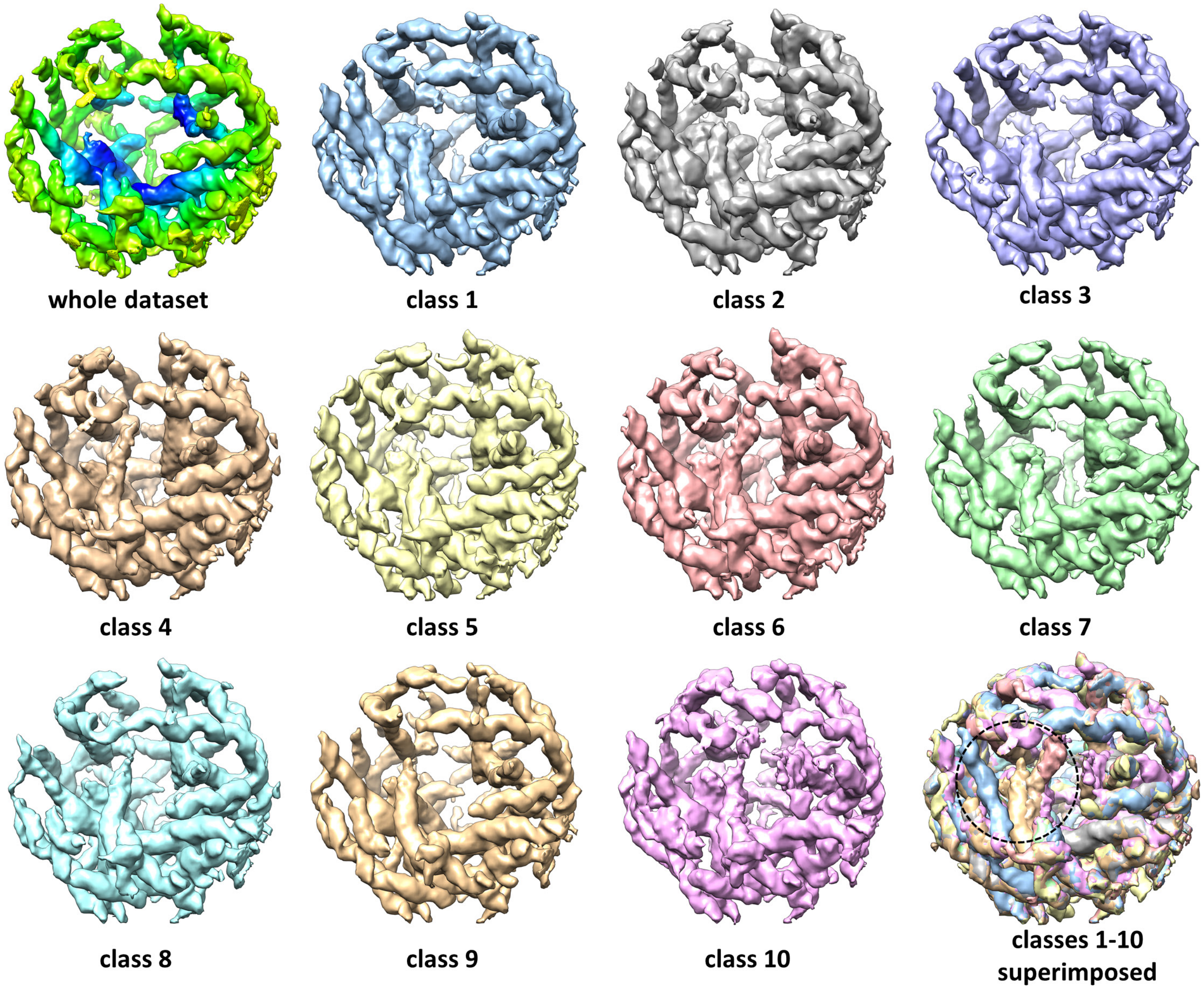
Data availability. The cryoEM density map and the atomic models have been deposited in EMDB and PDB under the accession numbers EMD-8397 and 5TC1, respectively. The traced backbone model and the annotated secondary structure of the MS2 genome are available as Supplementary Information. All other data are available from the corresponding author upon reasonable request.

31. Suloway, C. *et al.* Automated molecular microscopy: the new Leginon system. *J. Struct. Biol.* **151,** 41–60 (2005).
32. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat. Methods* **10,** 584–590 (2013).
33. Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol.* **142,** 334–347 (2003).
34. Kivioja, T., Ravantti, J., Verkhovsky, A., Ukkonen, E. & Bamford, D. Local average intensity-based method for identifying spherical particles in electron micrographs. *J. Struct. Biol.* **131,** 126–134 (2000).
35. Ludtke, S. J., Baldwin, P. R. & Chiu, W. EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.* **128,** 82–97 (1999).

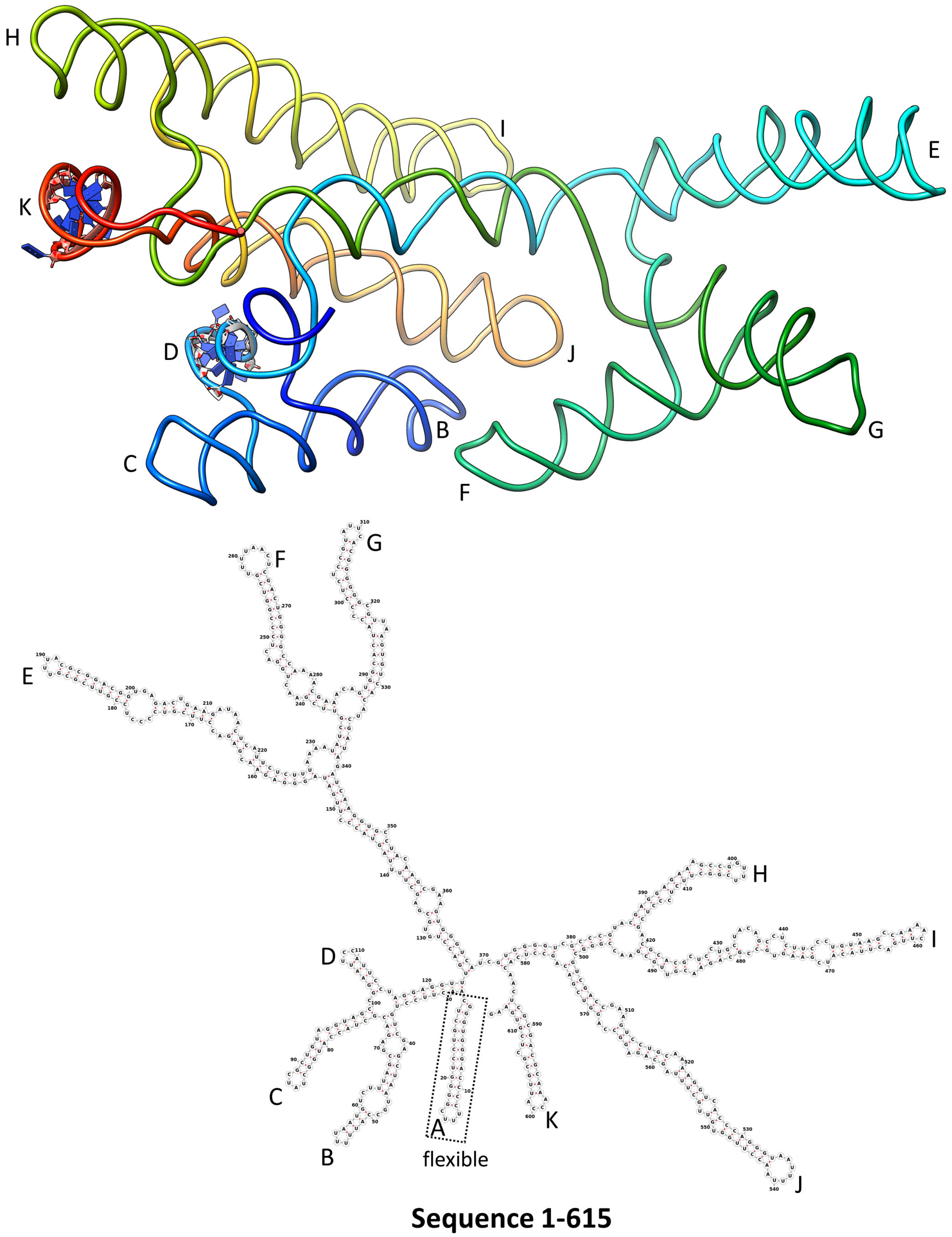
36. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180,** 519–530 (2012).
37. Zhang, X. *et al. In situ* structures of the segmented genome and RNA polymerase complex inside a dsRNA virus. *Nature* **527,** 531–534 (2015).
38. Liu, H. & Cheng, L. Cryo-EM shows the polymerase structures and a nonspooled genome within a dsRNA virus. *Science* **349,** 1347–1350 (2015).
39. Grigorieff, N. FREALIGN: high-resolution refinement of single particle structures. *J. Struct. Biol.* **157,** 117–125 (2007).
40. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333,** 721–745 (2003).
41. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11,** 63–65 (2014).
42. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr*. **66,** 486–501 (2010).
43. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
44. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 213–221 (2010).
45. Horn, W. T. *et al.* Structural basis of RNA binding discrimination between bacteriophages Qbeta and MS2. *Structure* **14,** 487–495 (2006).
46. Lim, C. Y. *et al.* The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev.* **18,** 1606–1617 (2004).
47. Lorenz, R. *et al.* ViennaRNA package 2.0. *Algorithms Mol. Biol.* **6,** 26 (2011).



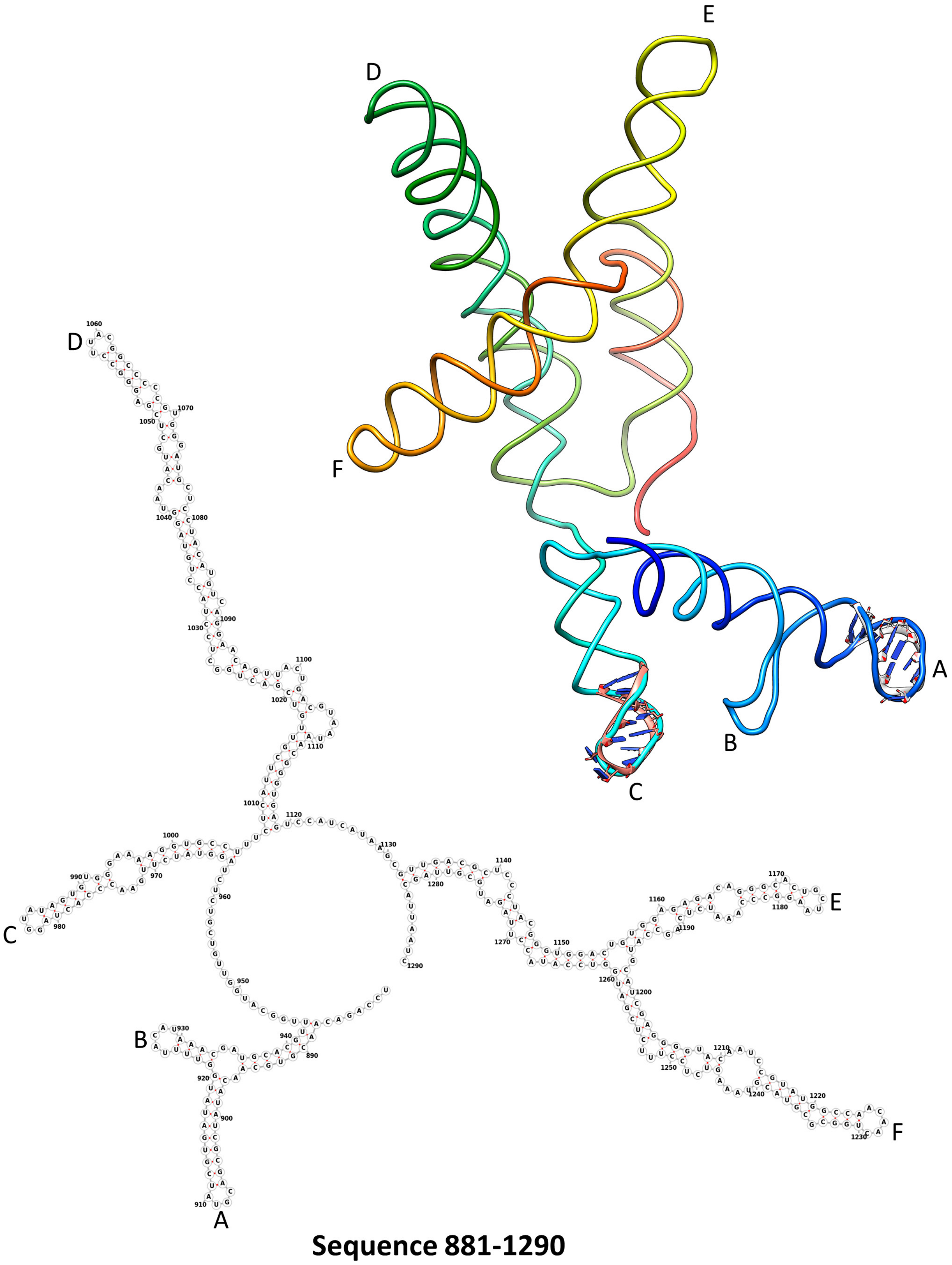
Extended Data Figure 1 | Resolution assessment of the cryoEM reconstruction. a, 'Gold-standard' FSC curve of the cryoEM reconstruction. The average resolution of the final density map is 3.6 Å as determined by the FSC = 0.143 criterion[40]. **b**, Local resolution assessed by ResMap[41]. Density voxels are coloured according to their local resolution as defined in the colour scale on the right. Only half of the capsid is shown to expose the RNA densities inside. **c**, **d**, CryoEM densities of a coat protein dimer (**c**) or the maturation protein (**d**) with their bound RNA stem–loops to show quality of the density map. In both cases, the cryoEM densities are semitransparent to show the fitted atomic models of the protein and RNA.



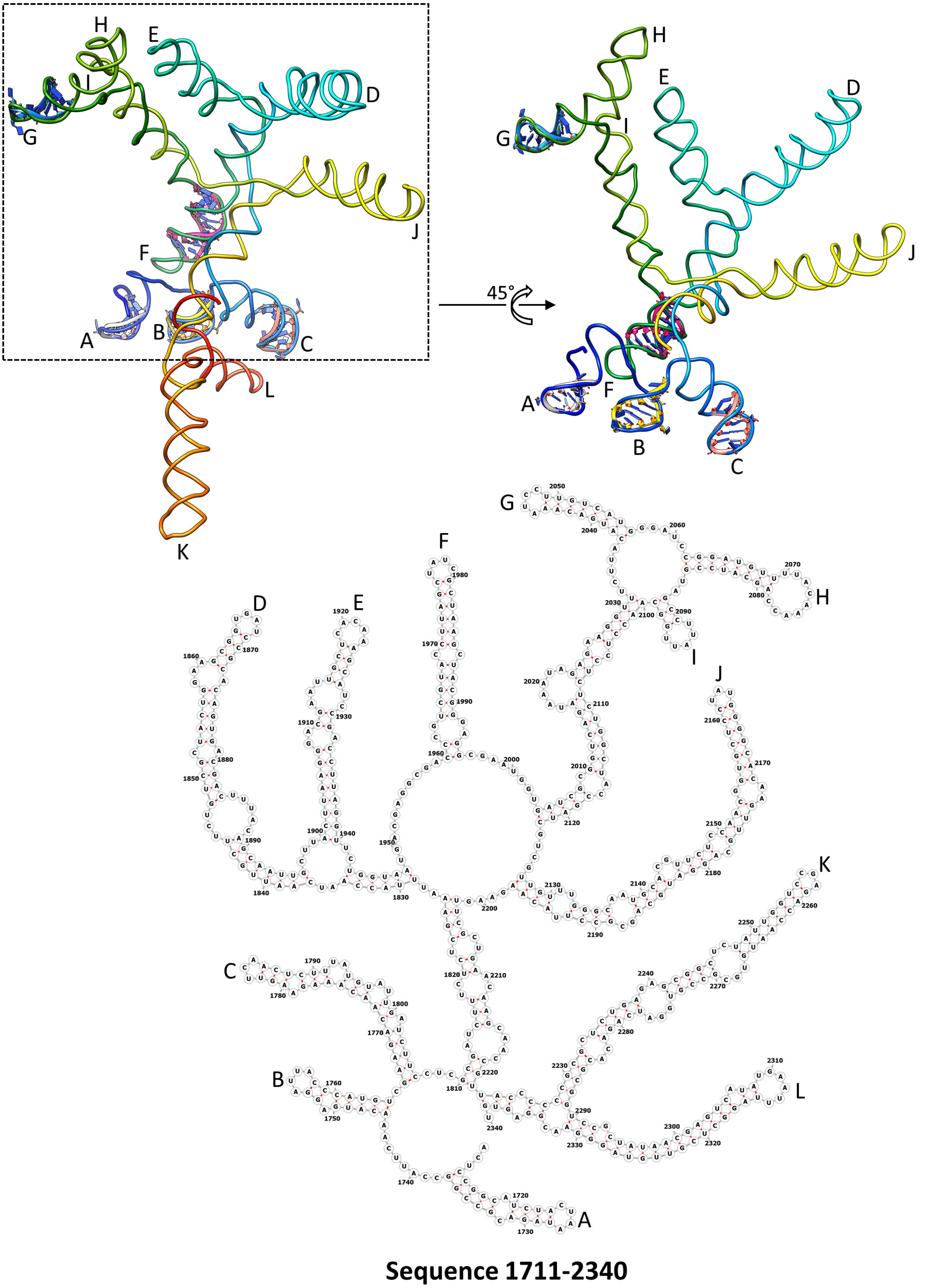
Extended Data Figure 2 | Three-dimensional classification. The entire dataset of the cryoEM images were subjected to 3D classification and refinement starting from a single initial model of the asymmetric reconstruction. Ten classes were arbitrarily set. The resulting density maps were compared with the reconstruction of the whole dataset and with each other. The overall structures of the ten classes are almost identical, except for small regions as exemplified by the region enclosed in the dashed circle in the superimposed map. The RNA fragments of these regions have multiple conformations, and are thus not traced in our model. Overall, we were able to trace an RNA density amounting to 80% of the genome.



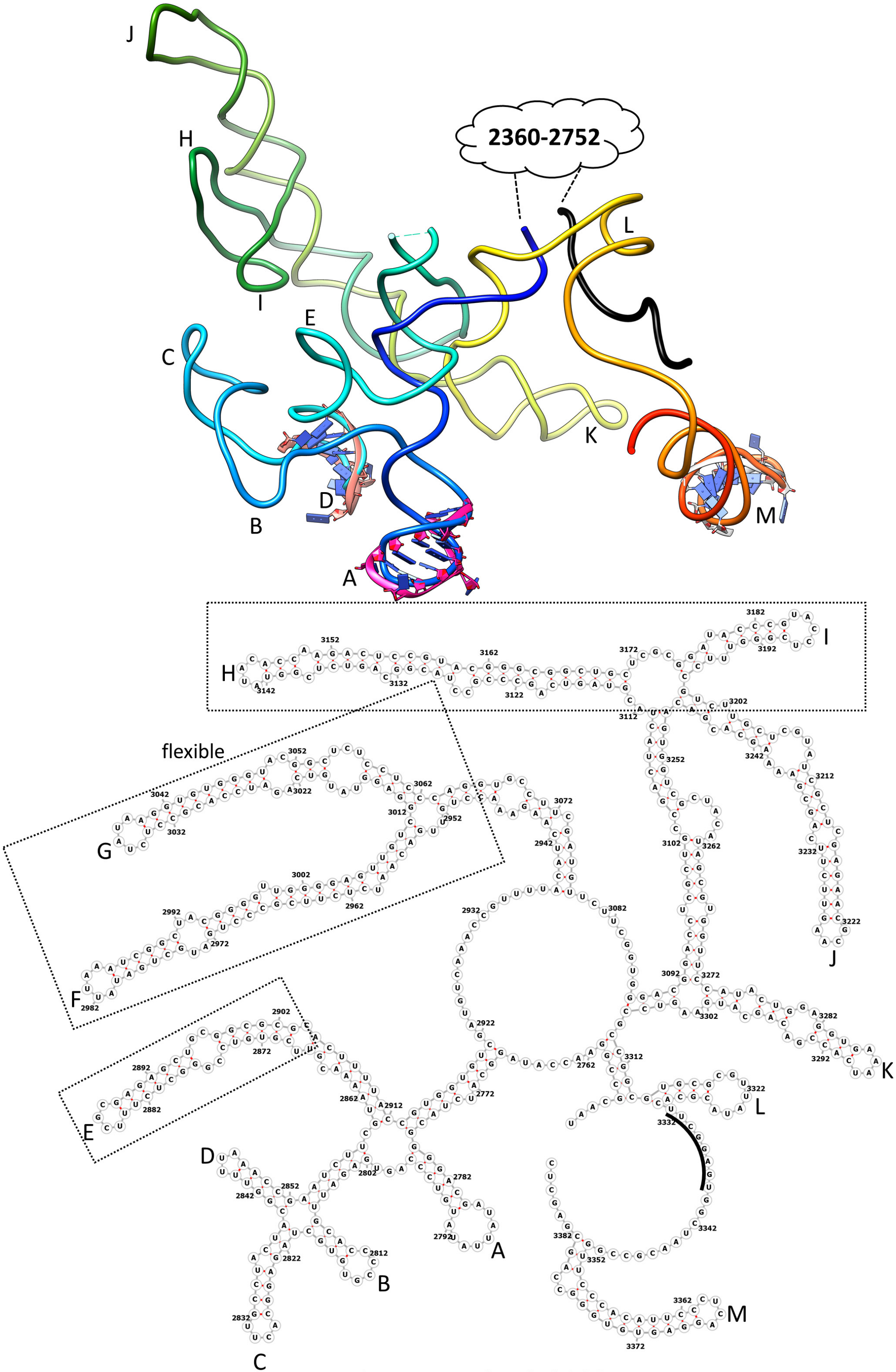
Extended Data Figure 3 | Backbone model of MS2 genome segment 1–615. Part of the traced backbone model of MS2 genome (top panel; rainbow-coloured blue to red from 5′ to 3′) is compared with the predicted secondary structure (bottom panel) of genome sequence 1–615. Matching stem–loops in the two are marked with the same letter. Atomic models of high-resolution stem–loops (ribbons in top panel) contained in the segment are also shown. Some of the base pairings in the predicted secondary structure have been modified to make it more consistent with the observed structure. Dashed box in the bottom panel denotes flexible stem–loop that is not well resolved in the cryoEM density map and thus not traceable for the backbone.



Extended Data Figure 4 | Backbone model of MS2 genome segment 881–1290. Part of the traced backbone model of MS2 genome (top panel; rainbow-coloured blue to red from 5′ to 3′) is compared with the predicted secondary structure (bottom panel) of genome sequence 881–1290. Matching stem–loops in the two are marked with the same letter. Atomic models of high-resolution stem–loops (ribbons in top panel) contained in the segment are also shown. Some of the base pairings in the predicted secondary structure have been modified to make it more consistent with the observed structure.

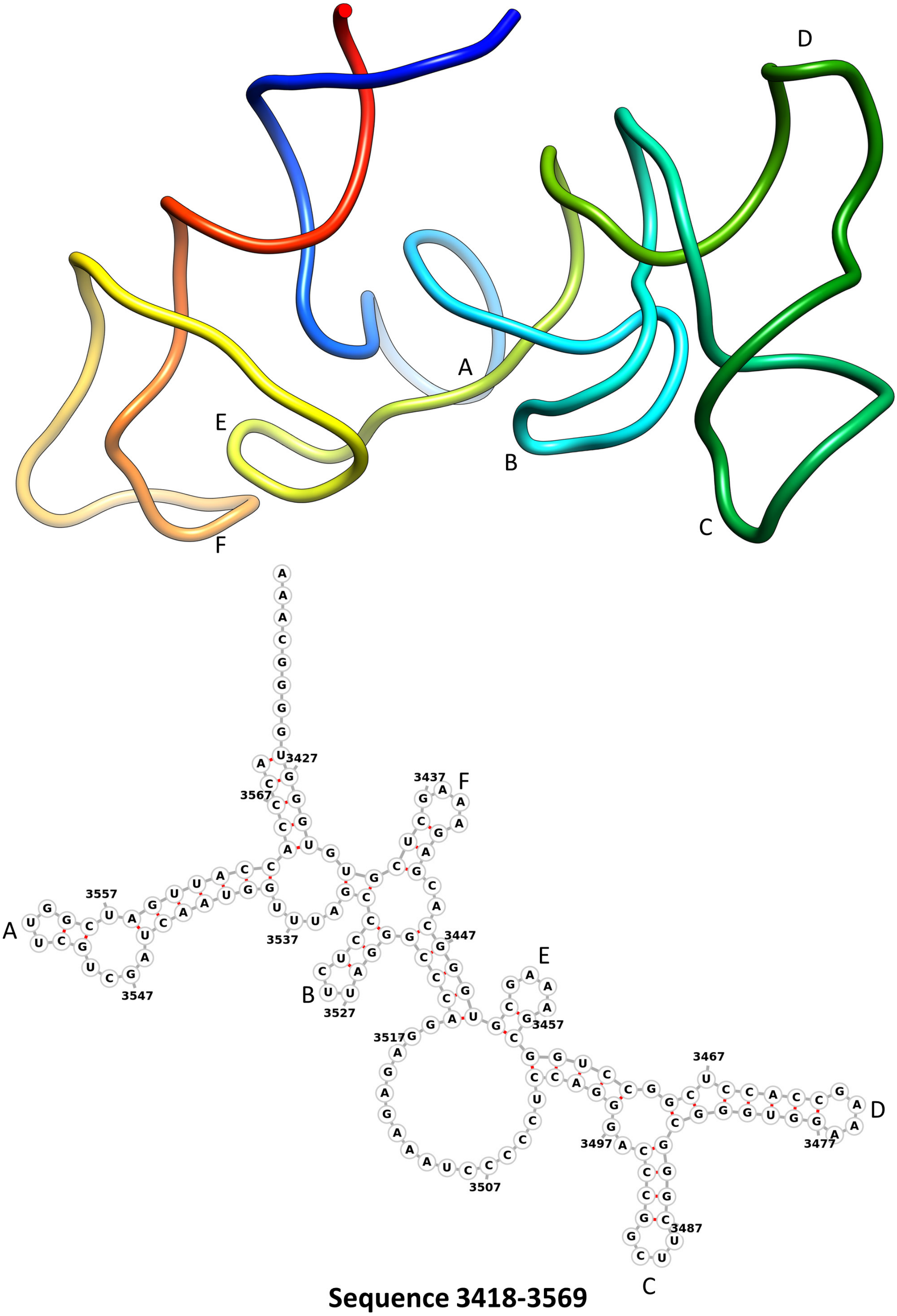


Extended Data Figure 5 | Backbone model of MS2 genome segment 1711–2340. Part of the traced backbone model of MS2 genome (top panels; rainbow-coloured blue to red from 5′ to 3′) is compared with the predicted secondary structure (bottom panel) of genome sequence 1711–2340. Matching stem–loops in the two are marked with the same letter. Atomic models of high-resolution stem–loops (ribbons in top panels) contained in the segment are also shown. Some of the base pairings in the predicted secondary structure have been modified to make it more consistent with the observed structure.

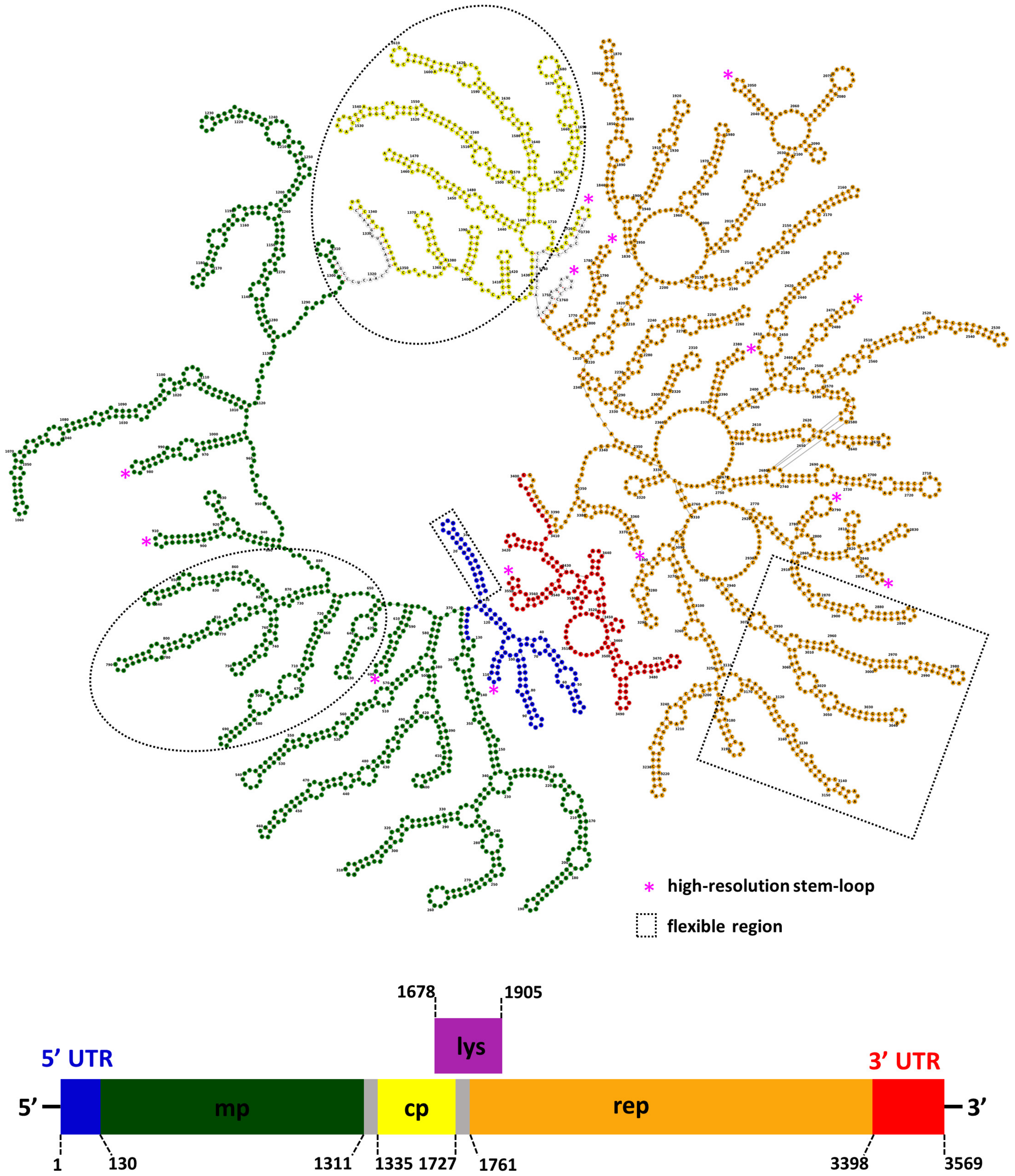


Sequence 2753-3388

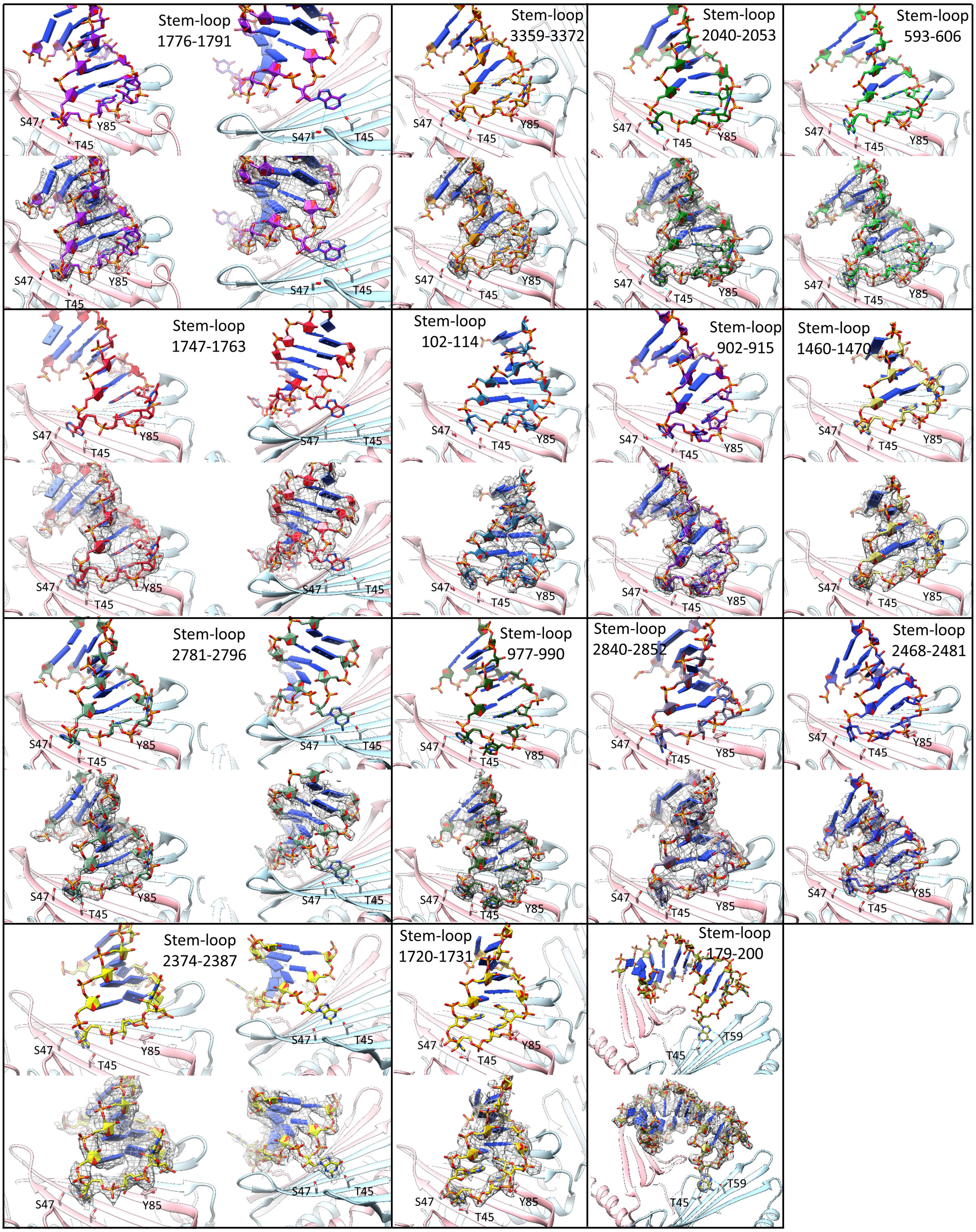
Extended Data Figure 6 | Backbone model of MS2 genome segment 2753–3388. Part of the traced backbone model of MS2 genome (top panel; rainbow-coloured blue to red from 5′ to 3′) is compared with the predicted secondary structure (bottom panel) of genome sequence 2753–3388. Matching stem–loops in the two are marked with the same letter. Atomic models of high-resolution stem–loops (ribbons in top panel) contained in the segment are also shown. Some of the base pairings in the predicted secondary structure have been modified to make it more consistent with the observed structure. Dashed boxes in the bottom panel denote flexible stem–loops that are not well resolved in the cryoEM density map and thus not traceable for the backbone. Black wire in the top panel denotes RNA segment 2341–2359 that has long-range base-pairing interactions (also illustrated in Fig. 2d) with this segment, and the pairing bases are marked with black arc in the bottom panel.



Extended Data Figure 7 | Backbone model of MS2 genome segment 3418–3569. Part of the traced backbone model of MS2 genome (top panel; rainbow-coloured blue to red from 5′ to 3′) is compared with the predicted secondary structure (bottom panel) of genome sequence 3418–3569. Matching stem–loops in the two are marked with the same letter. Some of the base pairings in the predicted secondary structure have been modified to make it more consistent with the observed structure.



Extended Data Figure 8 | Secondary structure of the MS2 genome. Secondary structures of all genome segments in Fig. 2d and Extended Data Figs 3–7 are assembled to show the secondary structure of the entire MS2 genome. The genome sequences are coloured according to the genes encoded as depicted in the schematic diagram at the bottom, except for the lysis gene which overlaps with the coat protein gene and the replicase gene. The star signs denote the positions of the 16 high-resolution stem–loops. Segments enclosed with dotted boxes or ellipses are flexible.



Extended Data Figure 9 | CryoEM densities (mesh) and atomic models (stick) of the 15 high-resolution RNA stem–loops that interact with coat protein dimers (ribbon).