Article:
Tetter, S, Terasaka, N, Steinauer, A et al. (10 more authors) (2021) Evolution of a virus-like architecture and packaging mechanism in a repurposed bacterial protein. Science, 372 (6547). pp. 1220-1224. ISSN 0036-8075

https://doi.org/10.1126/science.abg2822

Title: Evolution of a virus-like architecture and packaging mechanism in a repurposed bacterial protein

Authors: Stephan Tetter[1]†‡, Naohiro Terasaka[1]§‡, Angela Steinauer[1]‡, Richard J. Bingham[2], Sam Clark[2], Andrew P. Scott[3], Nikesh Patel[3], Marc Leibundgut[4], Emma Wroblewski[3], Nenad Ban[4], Peter G. Stockley[3], Reidun Twarock[2], Donald Hilvert[1]*

Affiliations:

[1]Laboratory of Organic Chemistry, ETH Zurich, 8093 Zurich, Switzerland

[2]Departments of Mathematics and Biology, University of York, York, YO10 5DD, UK

[3]Astbury Centre for Structural Molecular Biology, University of Leeds, Leeds, LS2 9JT, UK

[4]Institute of Molecular Biology and Biophysics, ETH Zurich, 8093 Zurich, Switzerland

†Present address: MRC Laboratory of Molecular Biology, Cambridge, CB2 0QH, UK

§Present address: Department of Chemistry, Graduate School of Science, The University of Tokyo, Tokyo, Japan

‡These authors contributed equally to this work

*Correspondence to: donald.hilvert@org.chem.ethz.ch

Abstract

Viruses are ubiquitous pathogens of global impact. Prompted by the hypothesis that their earliest progenitors recruited host proteins for virion formation, we have used stringent laboratory evolution to convert a bacterial enzyme lacking affinity for nucleic acids into an artificial nucleocapsid that efficiently packages and protects multiple copies of its own encoding mRNA. Revealing remarkable convergence on the molecular hallmarks of natural viruses, the accompanying changes reorganized the protein building blocks into an interlaced 240-subunit icosahedral capsid impermeable to nucleases, while emergence of a robust RNA stem-loop packaging cassette ensured high encapsidation yields and specificity. In addition to evincing a plausible evolutionary pathway for primordial viruses, these findings highlight practical strategies for developing non-viral carriers for diverse vaccine and delivery applications.

One Sentence Summary

A bacterial protein evolved to efficiently package and protect its own genome begins to resemble a natural virus.

Main Text

Understanding the origins and evolutionary trajectories of viruses is a fundamental scientific challenge (*1*). Even the simplest virions, optimized for genome propagation over billions of years of evolution, require co-assembly of many copies of a single protein with an RNA or DNA molecule to afford a closed-shell container of defined size, shape, and symmetry. Strategies for excluding competing host nucleic acids and protecting the viral genome from nucleases are also needed. While recreating such properties in non-viral containers is challenging (*2–6*), capsids generated by bottom-up design are promising as customizable tools for delivery and display (*7–9*).

Previous efforts to produce artificial nucleocapsids that encapsulate their own genetic information have utilized natural and computationally designed protein cages possessing engineered cationic interiors (*5*, *6*). However, even after directed evolution only ~10% of the resulting particles contained the full-length target RNA, underscoring the difficulties associated with packaging and protecting nucleic acids in a cell. In addition to competition from abundant host nucleic acids, genome degradation by cellular RNases is problematic owing to slow assembly, cage dynamics and/or porosity. Here we show that complementary adaptations of cargo and container can be harnessed to address these challenges and recapitulate the structural and packaging properties of natural viruses.

Our starting point was a previously evolved nucleocapsid, derived from *Aquifex aeolicus* lumazine synthase (AaLS), a bacterial enzyme that naturally forms 60-subunit nanocontainers but has no inherent affinity for nucleic acids (*10*). AaLS was redesigned by circular permutation and appending the arginine-rich peptide λN+, which tightly binds an RNA stem-loop called BoxB (*11*, *12*) (Fig. S1A). The resulting nucleocapsid variant, NC-1 (previously called λcpAaLS), was subsequently evolved via intermediate NC-2 (λcpAaLS-β16) to NC-3 (λcpAaLS-α9) by selecting for variants that capture capsid-encoding mRNA transcripts flanked by BoxB tags. Nevertheless, only one in eight of the NC-3 capsids packaged the full-length RNA genome (*6*).

To improve NC-3's packaging properties, we mutagenized its gene by error-prone PCR and subjected the library to three cycles of expression, purification, and nuclease challenge, followed by re-amplification of the surviving mRNA. Selection stringency was steadily increased in each cycle by decreasing nuclease size (60 kDa benzonase → 14 kDa RNase A → 11 kDa RNase T1) and extending nuclease exposure from 1 to 4 hours. This strategy ensured 1) efficient assembly of RNA-containing capsids, 2) protection of the cargo from nucleases, and 3) enrichment of variants that package the full-length mRNA (Fig. 1A). The best variant, NC-4, had nine new mutations, three of which were silent (Figs. S2,S3).

After optimizing protein production and purification, we compared NC-4 to its precursors. Particle heterogeneity decreased notably over the course of evolution from NC-1 so that NC-4 assembles into homogeneous capsids (Fig. S1B,C), with protein yields after purification that increased by an order of magnitude in the last evolutionary step (~35 mg NC-4/L medium versus

~3 mg NC-3/L). Additionally, nuclease resistance steadily improved. NC-1 RNA is almost completely degraded upon treatment with either benzonase or RNase A, whereas NC-2 protects small amounts of full-length mRNA from benzonase but not RNase A (Fig. S1D). In contrast, both NC-3 and NC-4 protect most of their encapsidated RNA from both nucleases (Fig. S1D,E). Importantly, NC-4 also packages its own full-length mRNA with improved specificity. While earlier generations encapsidate a broad size range of RNA species (400–2000 nt), NC-4 binds one major species corresponding to the 863 nt-long capsid mRNA (Fig. 1B, left). Long-read direct cDNA sequencing confirmed the decrease in encapsidated host RNA (Fig. 1C), which was largely ribosomal (Fig. S4). The simultaneous increase in genome packaging efficiency over the four generations is clearly evident in gels stained with the fluorogenic dye DFHBI-1T, which binds the Broccoli aptamers (*13*) introduced with the BoxB tags (*6*) (Fig. 1B, right).

The fraction of full-length genome relative to total encapsidated RNA was quantified by real-time PCR to be (2±2)% for NC-1, (6±5)% for NC-2, (24±12)% for NC-3 and (64±11)% for NC-4 (Fig. 1D). When NC-4 was further purified by ion-exchange chromatography to remove incomplete or poorly-assembled capsids, (87±19)% of the RNA corresponded to the full-length genome. Given the total number of encapsidated nucleotides (~2500), NC-4 packages on average 2.5 full-length mRNAs per capsid, a dramatic improvement compared to its precursors and other artificial nucleocapsids (*5*, *6*). This packaging capacity suggests that the evolved capsid could readily accommodate substantially longer RNAs, such as more complex genomes or large RNA molecules of medical interest.

Improved genome packaging and protection were accompanied by major structural transformations. The cavity of the starting 16 nm diameter AaLS scaffold is too small to package an 863 nt-long RNA (*2*, *6*). However, addition of the λN+ peptide to circularly permuted AaLS afforded expanded capsids with diameters in the 20–30 nm range, which were subsequently evolved toward uniform ~30 nm diameter particles (Fig. S1C). To elucidate the nature of these changes, we turned to cryo-EM.

Characterization of the initial NC-1 design revealed a range of assemblies of varying size and shape (Fig. S5A,B). Although particle heterogeneity and aggregation complicated single-particle reconstruction, two expanded structures with tetrahedral symmetry were successfully obtained (Fig. 2A). Like the wild-type protein, both are composed entirely of canonical lumazine synthase pentamers (Fig. 3A), but they possess large, keyhole-shaped pores (~4 nm wide) through which nucleases could diffuse. One capsid is a 180-mer (Fig. S5C–F, Table S1) that closely resembles a previously characterized AaLS variant possessing a negatively charged lumen (*14*, *15*). The other NC-1 structure is an unprecedented 120-mer (Fig. S5G–I, Table S1). It features wild-type-like pentamer-pentamer interactions as well as inter-pentamer contacts characteristic of its 180-mer sibling (Fig. 2A). At the monomer level, the major deviation from the AaLS fold is seen in a short helix (residues 67–74) and adjacent loop (residues 75–81) (Fig. 2B,C). In AaLS, this region is involved in lumenal interactions between the pentameric building blocks at the threefold-symmetry axes. In NC-1 chains that are not involved in wild-type-like pentamer-pentamer

contacts, this loop assumes altered conformations and is resolved to lower local resolution (Figs. 2B,C, S5E,H).

The second-generation variant NC-2, obtained after benzonase challenge, is also polymorphic and aggregation-prone. Several distinct morphologies were identified by 2D-classification (Fig. S6), one of which was reconstructed as a tetrahedrally symmetric 180-mer (4.5 Å, Fig. 2A, Table S1) that superimposes on the analogous NC-1 structure. Four mutations (I58V, G61D, V62I, and I191F) shorten two strands of the core beta-sheet and, indirectly, further increase disorder in neighboring residues 66–81 (Fig. 2B,C). These changes disfavor wild type-like pentamer-pentamer interactions, explaining the absence of smaller capsids with more tightly packed capsomers (*16*). Structural heterogeneity and particle aggregation precluded reconstruction of additional structures that may contribute to the benzonase-resistance phenotype.

The ability of NC-3 and NC-4 to protect their cargo from RNases significantly smaller than the pores in the parental structures suggests a novel solution to nuclease resistance. In fact, three-dimensional reconstructions of NC-3 (7.0 Å) and NC-4 (3.0 Å) (Fig. S7, Table S1) yielded superimposable structures that are markedly different from any previously characterized AaLS derivative (Fig. 2A). Both capsids form icosahedrally symmetric 240-mers that feature smaller pores (~2.5 nm) than their progenitors. The pentagonal vertices align with AaLS pentamers, and are surrounded by 30 hexagonal patches (Figs. 3B, S8). This architecture is typical of T=4 virus capsids, in which a single protein chain assumes four similar, quasi-equivalent conformations, repeated with icosahedral symmetry to afford a closed container with increased volume (*17*).

The most striking feature of our evolved cages is a 3D-domain swap (*18*), which links neighboring monomers and reorganizes the structure into trimeric building blocks (Figs. 3B, S8). As reported for some viral capsids (*19–22*), such interlacing may enhance particle stability. This rearrangement was made possible by a hinge around residues 62–66, which permits dissociation of the N-terminal helix and strand of each subunit from the core, allowing it to dock onto a neighboring subunit in the trimeric capsomer. An elongated alpha-helix extends C-terminally from this hinge, formed by fusing the short helix (residues 67–74) to the following helix by ordering of the intervening loop (residues 75–81) (Figs. 2C, S9A). Slight variations in the hinge angles allow the subunits to occupy four quasi-equivalent positions within the expanded icosahedral lattice (*23*) (Figs. 3C, S8D-G) and repurpose the inter-pentamer interfaces of the wild-type scaffold for penton-hexon contacts (Fig. S9). Such flexible hinges might similarly be exploited for the rational design of large (T>1) capsid assemblies from a single protein chain, an as yet unmet challenge due to the difficulty of designing proteins capable of adopting several distinct conformations.

The smaller pores in the NC-3 and NC-4 shells provide a compelling explanation for nuclease resistance. The structurally unresolved λN+ peptides, which line the lumenal edge of these openings, likely further restrict access to the cage interior. Nevertheless, the superimposable structures do not account for the differences in packaging efficiency between NC-3 and NC-4.

Although a lysine to arginine mutation that appeared in the λN+ peptide of NC-4 is known to increase affinity to the BoxB tags ~3-fold (*11*), the effects of reverting this mutation are modest (Fig. S10), indicating that other factors must be at play.

An NC-4 variant lacking the RNA-binding peptide still assembles into capsids, but the yields decrease ~2-fold and the resulting particles are heterogeneous in both size and shape (Fig. S11), suggesting a potential role for RNA in capsid formation. Some viruses that package single-stranded RNA genomes utilize multiple stem-loop packaging signals to orchestrate capsid assembly and ensure cargo specificity within the crowded confines of the cell (*24*). Could the evolution of additional RNA packaging signals in the NC-4 genome explain its superiority to NC-3? Besides the originally introduced BoxB tags (*6*), BB1 and BB2, both genomes have 37 BoxB-like URxRxRR (R=purine) and URxR sequences (*25*) (Table S2). In order to determine whether any of these serve as packaging signals, we used synchrotron X-ray footprinting (XRF). Synchrotron radiation generates hydroxyl radicals, which cleave the RNA backbone. Because base-pairing and intermolecular interactions, such as with protein, decrease local cleavage propensity, XRF provides a means to map intermolecular interactions and RNA secondary structure (*26*).

Footprints for packaged NC-3 and NC-4 RNA show that only BB1, BB2, and 11 out of 37 BoxB-like motifs exhibit low XRF reactivity (Table S2). Furthermore, XRF-informed prediction of RNA secondary structure ensembles (*27*, *28*) indicates that only seven of these motifs (BB1, BB2, and potential packaging signals PS1–5) are presented as stem-loops with significant frequency (Figs. S12A, S13A). Assuming that interactions with the λN+ peptides stabilize the stem-loops, comparison of their display frequency in encapsulated versus free RNA pinpoints which of these motifs might serve as packaging signals.

In NC-3, the secondary structure predictions (Figs. 4A,C,E, S12) indicate that the original high-affinity BoxB tags are more frequently displayed as stem-loops in free transcripts than in capsids (96% vs. 63% for BB1 and 75% vs. 52% for BB2). Although the five lower affinity PS1–5 motifs are displayed more frequently upon encapsulation, their broad distribution, coupled with modest display of the high-affinity tags, contrasts with natural viruses, which appear to utilize narrow clusters of packaging signals surrounding an efficiently displayed, high-affinity stem-loop to initiate capsid assembly (*24*). The lack of robust assembly instructions may explain why 72% of the RNA packaged in NC-3 is ribosomal. Ribosomal RNA is compact, abundant and also possesses multiple BoxB-like signals (Fig. S4C,D), which may allow it to function as an alternative nucleation hub for capsid assembly.

In NC-4, four of the seven potential packaging signals are significantly populated as stem loops in packaged genomes (PS1, BB1, PS2, PS4) and all are clustered at the 5'-end of the transcript. Notably, BB1 is displayed in 99% of all packaged RNA folds (Figs. 4B,D, S13). The low reactivities observed for the four URxR sub-motifs within the capsid (Fig. 4F) imply that they are in contact with protein. Robust display of a high-affinity packaging signal within a cassette of lower affinity motifs (PS1, PS2, and PS4) is reminiscent of nucleation complexes

found in Satellite Tobacco Necrosis Virus (*29*), MS2 phage (*30*), and Hepatitis B Virus (*31*). This finding suggests that NC-4 similarly evolved a key hallmark of RNA packaging signal-mediated assembly. Genome-encoded packaging instructions likely foster selective RNA encapsulation as well as rapid, efficient capsid assembly (*32*), providing a compelling explanation for the improved properties of the evolved cage (Fig. 4G). Encapsulation of alternative or longer, more complex genomes may similarly benefit from optimization of RNA sequence and structure.

Successful conversion of a bacterial enzyme into a nucleocapsid that packages and protects its own encoding mRNA with high efficiency and selectivity shows how primordial self-replicators could have recruited host proteins for virion formation (*1*). While we started from a capsid-forming enzyme, similar pathways could be envisioned for smaller oligomeric proteins, where cargo protection would provide the evolutionary driving force towards shell formation. The convergence on structural properties characteristic of natural RNA viruses through co-evolution of capsid and cargo is striking. Introduction of destabilizing mutations into the starting protein was key to the dramatic remodeling of the protein shell, providing the molecular heterogeneity needed to depart from the initial, energetically stable, architectural solution and converge on a regular, 240-subunit, closed-shell icosahedral assembly. At the same time, evolution of multiple RNA packaging motifs that can cooperatively bind the coat proteins likely guided specificity and efficient assembly. While such constructs are themselves attractive as customizable and potentially safe alternatives to natural viruses for gene delivery and vaccine applications, the lessons learned from their evolution may also inform ongoing efforts to tailor the properties of natural viruses for more effective gene therapy (*33*).

References and Notes

1. M. Krupovic, V. V Dolja, E. V Koonin, Origin of viruses: primordial replicators recruiting capsids from hosts. *Nat. Rev. Microbiol.* **17**, 449–458 (2019).
2. Y. Azuma, T. G. W. Edwardson, N. Terasaka, D. Hilvert, Modular Protein Cages for Size-Selective RNA Packaging in Vivo. *J. Am. Chem. Soc.* **140**, 566–569 (2018).
3. T. G. W. Edwardson, T. Mori, D. Hilvert, Rational engineering of a designed protein cage for siRNA delivery. *J. Am. Chem. Soc.* **140**, 10439–10442 (2018).
4. S. Lilavivat, D. Sardar, S. Jana, G. C. Thomas, K. J. Woycechowsky, In vivo encapsulation of nucleic acids using an engineered nonviral protein capsid. *J. Am. Chem. Soc.* **134**, 13152–13155 (2012).
5. G. L. Butterfield, M. J. Lajoie, H. H. Gustafson, D. L. Sellers, U. Nattermann, D. Ellis, J. B. Bale, S. Ke, G. H. Lenz, A. Yehdego, R. Ravichandran, S. H. Pun, N. P. King, D. Baker, Evolution of a designed protein assembly encapsulating its own RNA genome. *Nature* **552**, 415–420 (2017).
6. N. Terasaka, Y. Azuma, D. Hilvert, Laboratory evolution of virus-like nucleocapsids from nonviral protein cages. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 5432–5437 (2018).

7. T. G. W. Edwardson, D. Hilvert, Virus-Inspired Function in Engineered Protein Cages. *J. Am. Chem. Soc.* **141**, 9432–9443 (2019).

8. K. A. Cannon, J. M. Ochoa, T. O. Yeates, High-symmetry protein assemblies: patterns and emerging applications. *Curr. Opin. Struct. Biol.* **55**, 77–84 (2019).

9. A. C. Walls, B. Fiala, A. Schäfer, S. Wrenn, M. N. Pham, M. Murphy, L. V. Tse, L. Shehata, M. A. O'Connor, C. Chen, M. J. Navarro, M. C. Miranda, D. Pettie, R. Ravichandran, J. C. Kraft, C. Ogohara, A. Palser, S. Chalk, E. C. Lee, K. Guerriero, E. Kepl, C. M. Chow, C. Sydeman, E. A. Hodge, B. Brown, J. T. Fuller, K. H. Dinnon, L. E. Gralinski, S. R. Leist, K. L. Gully, T. B. Lewis, M. Guttman, H. Y. Chu, K. K. Lee, D. H. Fuller, R. S. Baric, P. Kellam, L. Carter, M. Pepper, T. P. Sheahan, D. Veesler, N. P. King, Elicitation of Potent Neutralizing Antibody Responses by Designed Protein Nanoparticle Vaccines for SARS-CoV-2. *Cell* **183**, 1367–1382 (2020).

10. R. Ladenstein, M. Fischer, A. Bacher, The lumazine synthase/riboflavin synthase complex: shapes and functions of a highly variable enzyme system. *FEBS J.* **280**, 2537–2563 (2013).

11. R. J. Austin, T. Xia, J. Ren, T. T. Takahashi, R. W. Roberts, Designed arginine-rich RNA-binding peptides with picomolar affinity. *J. Am. Chem. Soc.* **124**, 10966–10967 (2002).

12. G. Di Tomasso, P. Lampron, P. Dagenais, J. G. Omichinski, P. Legault, The ARiBo tag: A reliable tool for affinity purification of RNAs under native conditions. *Nucleic Acids Res.* **39**, e18 (2011).

13. G. S. Filonov, C. W. Kam, W. Song, S. R. Jaffrey, In-gel imaging of RNA processing using broccoli reveals optimal aptamer expression strategies. *Chem. Biol.* **22**, 649–660 (2015).

14. E. Sasaki, D. Böhringer, M. Van De Waterbeemd, M. Leibundgut, R. Zschoche, A. J. R. R. Heck, N. Ban, D. Hilvert, Structure and assembly of scalable porous protein cages. *Nat. Commun.* **8**, 14663 (2017).

15. F. P. Seebeck, K. J. Woycechowsky, W. Zhuang, J. P. Rabe, D. Hilvert, A simple tagging system for protein encapsulation. *J. Am. Chem. Soc.* **128**, 4516–4517 (2006).

16. M. S. Fornasari, D. A. Laplagne, N. Frankel, A. A. Cauerhff, F. A. Goldbaum, J. Echave, Sequence Determinants of Quaternary Structure in Lumazine Synthase. *Mol. Biol. Evol.* **21**, 97–107 (2004).

17. D. L. D. Caspar, A. Klug, Physical principles in the construction of regular viruses. *Cold Spring Harb. Symp. Quant. Biol.* **27**, 1–24 (1962).

18. M. J. Bennett, M. P. Schlunegger, D. Eisenberg, 3D domain swapping: A mechanism for oligomer assembly. *Protein Sci.* **4**, 2455–2468 (1995).

19. C. Qu, L. Liljas, N. Opalka, C. Brugidou, M. Yeager, R. N. Beachy, C. M. Fauquet, J. E. Johnson, T. Lin, 3D domain swapping modulates the stability of members of an icosahedral virus group. *Structure* **8**, 1095–1103 (2000).

20. Z. Sun, K. El Omari, X. Sun, S. L. Ilca, A. Kotecha, D. I. Stuart, M. M. Poranen, J. T. Huiskonen, Double-stranded RNA virus outer shell assembly by bona fide domain-swapping. *Nat. Commun.* **8**, 14814 (2017).

21. R. Sánchez-Eugenia, A. Durana, I. López-Marijuan, G. A. Marti, D. M. A. Guérin, X-ray structure of Triatoma virus empty capsid: Insights into the mechanism of uncoating and RNA release in dicistroviruses. *J. Gen. Virol.* **97**, 2769–2779 (2016).

22. G. Squires, J. Pous, J. Agirre, G. S. Rozas-Dennis, M. D. Costabel, G. A. Marti, J. Navaza, S. Bressanelli, D. M. A. Guérin, F. A. Rey, Structure of the Triatoma virus capsid. *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1026–1037 (2013).

23. M. Bonjack-Shterengartz, D. Avnir, The enigma of the near-symmetry of proteins: Domain swapping. *PLoS One* **12**, e0180030 (2017).

24. R. Twarock, P. G. Stockley, RNA-mediated virus assembly: Mechanisms and consequences for viral evolution and therapy. *Annu. Rev. Biophys.* **48**, 495–514 (2019).

25. P. Legault, J. Li, J. Mogridge, L. E. Kay, J. Greenblatt, NMR structure of the bacteriophage λ N peptide/boxB RNA complex: Recognition of a GNRA fold by an arginine-rich motif. *Cell* **93**, 289–299 (1998).

26. Y. Hao, J. Bohon, R. Hulscher, M. C. Rappé, S. Gupta, T. Adilakshmi, S. A. Woodson, Time-resolved hydroxyl radical footprinting of RNA with X-rays. *Curr. Protoc. Nucleic Acid Chem.* **73**, e52 (2018).

27. Y. Ding, C. Y. Chan, C. E. Lawrence, Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res.* **32**, W135–W141 (2004).

28. K. E. Deigan, T. W. Li, D. H. Mathews, K. M. Weeks, Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 97–102 (2009).

29. R. J. Ford, A. M. Barker, S. E. Bakker, R. H. Coutts, N. A. Ranson, S. E. V Phillips, A. R. Pearson, P. G. Stockley, Sequence-specific, RNA-protein interactions overcome electrostatic barriers preventing assembly of satellite tobacco necrosis virus coat protein. *J. Mol. Biol.* **425**, 1050–1064 (2013).

30. Ó. Rolfsson, S. Middleton, I. W. Manfield, S. J. White, B. Fan, R. Vaughan, N. A. Ranson, E. Dykeman, R. Twarock, J. Ford, C. C. Kao, P. G. Stockley, Direct evidence for packaging signal-mediated assembly of bacteriophage MS2. *J. Mol. Biol.* **428**, 431–448 (2016).

31. N. Patel, S. J. White, R. F. Thompson, R. Bingham, E. U. Weiß, D. P. Maskell, A. Zlotnick, E. C. Dykeman, R. Tuma, R. Twarock, N. A. Ranson, P. G. Stockley, HBV RNA pre-genome encodes specific motifs that mediate interactions with the viral core protein that promote nucleocapsid assembly. *Nat. Microbiol.* **2**, 17098 (2017).

32. E. C. Dykeman, P. G. Stockley, R. Twarock, Solving a Levinthal's paradox for virus assembly identifies a unique antiviral strategy. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 5361–5366 (2014).

33. C. Li, R. J. Samulski, Engineering adeno-associated virus vectors for gene therapy. *Nat. Rev. Genet.* **21**, 255–272 (2020).

34. M. Jain, H. E. Olsen, B. Paten, M. Akeson, The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239 (2016).

35. J. D. Perlmutter, M. F. Hagan, Mechanisms of virus assembly. *Annu. Rev. Phys. Chem.*

66, 217–239 (2015).

36. J. Z. Porterfield, A. Zlotnick, A simple and general method for determining the protein and nucleic acid content of viruses by UV absorbance. *Virology* **407**, 281–288 (2010).

37. E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, A. Bairoch, in *The Proteomics Protocols Handbook*, J. M. Walker, Ed. (Humana Press, Totowa, NJ, 2005), vol. 112, pp. 571–607.

38. J. F. Greisch, S. Tamara, R. A. Scheltema, H. W. R. Maxwell, R. D. Fagerlund, P. C. Fineran, S. Tetter, D. Hilvert, A. J. R. Heck, Expanding the mass range for UVPD-based native top-down mass spectrometry. *Chem. Sci.* **10**, 7163–7171 (2019).

39. T. Beck, S. Tetter, M. Künzle, D. Hilvert, Construction of Matryoshka-type structures from supercharged protein nanocages. *Angew. Chem. Int. Ed.* **54**, 937–940 (2015).

40. J. Zivanov, T. Nakane, B. O. Forsberg, D. Kimanius, W. J. H. Hagen, E. Lindahl, S. H. W. Scheres, New tools for automated high-resolution cryo-EM structure determination in RELION-3. *Elife* **7**, e42166 (2018).

41. S. Q. Zheng, E. Palovcak, J. P. Armache, K. A. Verba, Y. Cheng, D. A. Agard, MotionCor2: Anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).

42. K. Zhang, Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).

43. P. Emsley, K. Cowtan, Coot: Model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).

44. D. Liebschner, P. V. Afonine, M. L. Baker, G. Bunkoczi, V. B. Chen, T. I. Croll, B. Hintze, L. W. Hung, S. Jain, A. J. McCoy, N. W. Moriarty, R. D. Oeffner, B. K. Poon, M. G. Prisant, R. J. Read, J. S. Richardson, D. C. Richardson, M. D. Sammito, O. V. Sobolev, D. H. Stockwell, T. C. Terwilliger, A. G. Urzhumtsev, L. L. Videau, C. J. Williams, P. D. Adams, Macromolecular structure determination using X-rays, neutrons and electrons: Recent developments in Phenix. *Acta Crystallogr. D Struct. Biol.* **75**, 861–877 (2019).

45. F. Karabiber, J. L. McGinnis, O. V Favorov, K. M. Weeks, QuShape: Rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *RNA* **19**, 63–73 (2013).

46. R. Twarock, A. Luque, Structural puzzles in virology solved with an overarching icosahedral design principle. *Nat. Commun.* **10**, 4414 (2019).

Acknowledgments

We acknowledge technical support from the Functional Genomics Center Zurich, Miroslav Peterek and Peter Tittmann (Scientific Center for Optical and Electron Microscopy, ETH Zurich), Daniel Böhringer (Cryo-EM Knowledge Hub, ETH Zurich) for help with cryoEM Data collection and elaboration, and Oliver Allemann for help with optimization of nucleocapsid expression and purification. We thank DNA Sequencing & Services (MRC I PPU, School of Life Sciences, University of Dundee, Scotland, www.dnaseq.co.uk) for DNA sequencing.
Funding: DH thanks the Swiss National Science Foundation and the ETH Zurich for financial support. RT acknowledges funding via an EPSRC Established Career Fellowship (EP/R023204/1) and a Royal Society Wolfson Fellowship (RSWF/R1/180009). RT & PGS acknowledge support from a Joint Wellcome Trust Investigator Award (110145 & 110146), and PGS also thanks the NSLS-II at the Brookhaven Synchrotron for the award of slots for data collection at Beamline 17-BM, together with Erik Farquhar for his expert assistance during these visits. NT acknowledges support from the Human Frontier Science Program (LT000426/2015-L). AS is the recipient of a Marie Skłodowska-Curie Individual Fellowship (LEVERAGE mRNA). **Author contributions:** NT performed laboratory evolution, AS, ST, and NT biochemical characterization of nucleocapsids, and ST cryo-EM analysis with help from ML. APS, NP, and EW performed, and RJB, SC, PGS, and RT analyzed X-ray footprinting experiments. ST, AS, and DH wrote the manuscript with input from all other authors.
Competing interests: Authors declare no competing interests. **Data and materials availability:** The principal data supporting the findings of this study are provided in the figures and Supplementary Information. Additional data that support the findings of this study are available from the corresponding author upon request. Cryo-EM maps and atomic models have been deposited in the Electron Microscopy Data Bank (EMDB) and wwPDB, respectively, with the following accession codes: EMDB-11631 and PDB 7A4F (NC-1 120-mer), EMDB-11632 and PDB 7A4G (NC-1 180-mer), EMDB-11633 and PDB 7A4H (NC-2), EMDB-11634 and PDB 7A4I (NC-3), and EMDB-11635 and PDB 7A4J (NC-4).

Supplementary Materials

Materials and Methods

Figures S1-S13

Tables S1-S3

References (*36-46*)

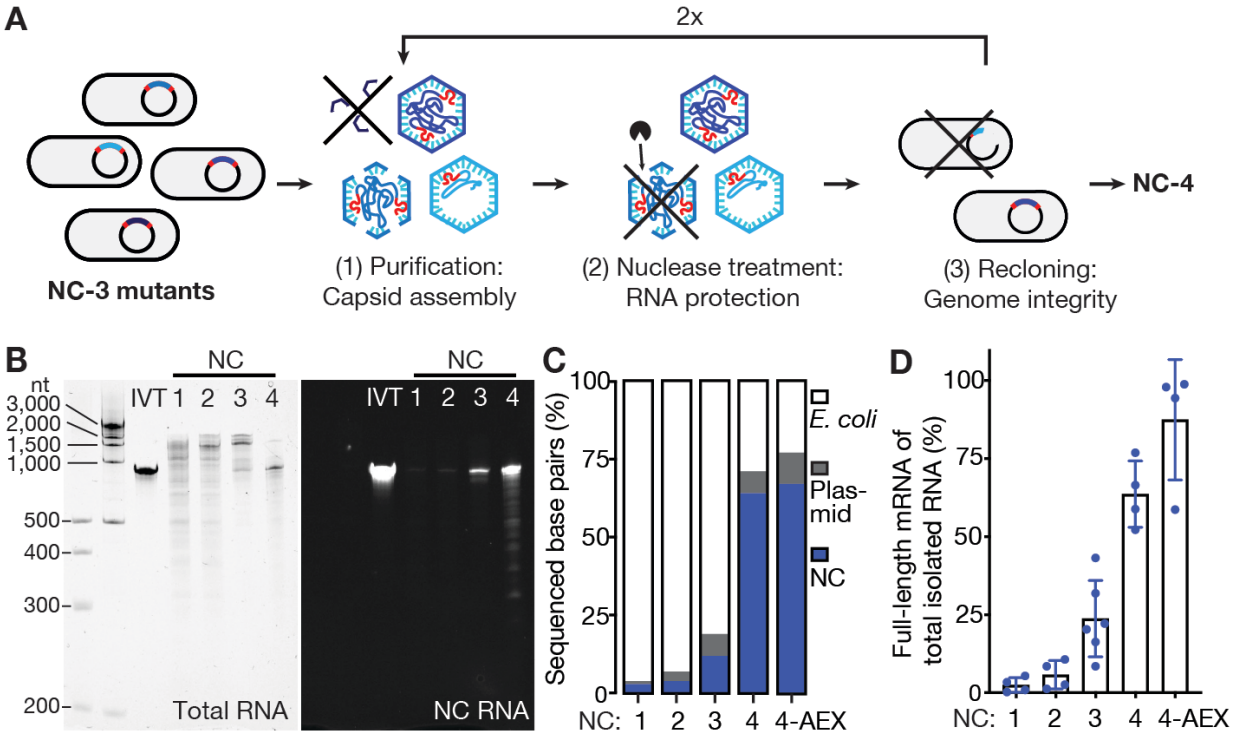


Fig. 1. NC-4 packages its genome with high selectivity. (**A**) Laboratory evolution: a library of NC-3 mutants generated by error-prone PCR was expressed in Escherichia coli and purified by affinity and size exclusion chromatography. This step recovers assembled capsids. The purified capsid library was then treated with nucleases to enrich for capsids that protect their RNA cargo. Finally, the RNA was extracted from capsids, reverse-transcribed, and re-cloned into the original expression vector. This step selects for capsids that contain full-length genomes. (**B**) Denaturing PAGE (5%) of NC-1 to NC-4 stained for total RNA with GelRed (left) and the fluorogenic dye DFHBI-1T (right), which selectively binds the broccoli aptamer present in the 5'- and 3'-untranslated regions of the mRNA genome (NC RNA). IVT, in vitro-transcribed reference mRNA. (**C**) RNA identities and their relative abundance were determined by Oxford Nanopore Sequencing (*34*) for all four capsids, including anion-exchanged NC-4 (4-AEX), and assigned to three main categories: bacterial RNA (E. coli), nucleocapsid mRNA (NC), and RNA originating from other plasmid-associated genes (plasmid). The encapsulated E. coli genes are primarily rRNA (Fig. S4). (**D**) The fraction of total extracted RNA corresponding to the full-length mRNA genome was determined by real-time quantitative PCR (mean of at least two biological replicates, each measured in two separate laboratories, error bars represent the standard deviation of the mean).

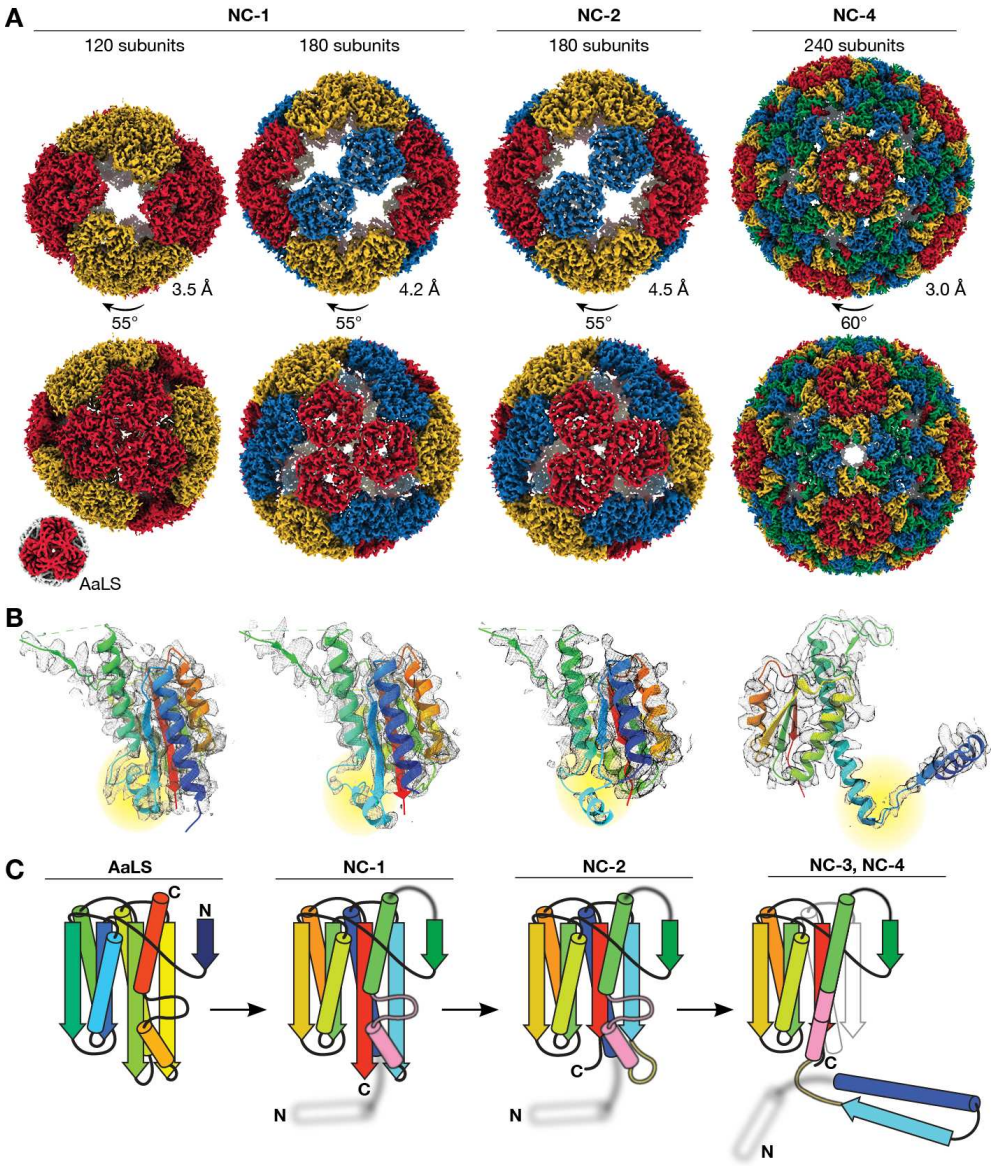


Fig. 2. Structural evolution towards virus-like nucleocapsids. (**A**) Maps are shown for the tetrahedrally symmetric NC-1 and NC-2 structures with symmetry-related pentamers in the same color, and the icosahedral *T*=4 NC-4 capsid with the four quasi-equivalent chains highlighted in different colors. The lower resolution NC-3 capsid (7.0 Å, Fig. S7) resembles NC-4. Wild-type AaLS (*10*) is shown for comparison (not to scale). Resolutions were estimated by Fourier shell correlation (0.143 threshold). (**B**) Fits of single chains (rainbow; N-terminus to C-terminus from blue to red) in the electron density of the capsids above show the evolution of the monomer. Residues 66 to 81 are highlighted (yellow). Clear density is seen for this segment in NC-1 protomers involved in AaLS-like inter-pentamer contacts. In other chains, as in the 180-mer NC-1 structure, this region is less well resolved. In NC-2 the nearby beta-sheet is also perturbed, further enhancing the flexibility of this region. In NC-3 and NC-4, this segment rearranges into an extended helix that supports the domain swap. (**C**) Rainbow-colored models depict the changes in the protein fold. The helix (67–74) and loop (75–81) that undergo a major rearrangement are colored in pink, and the hinge loop (62–66) in yellow; the structurally unresolved RNA-binding peptide is depicted as a blurry white helix.

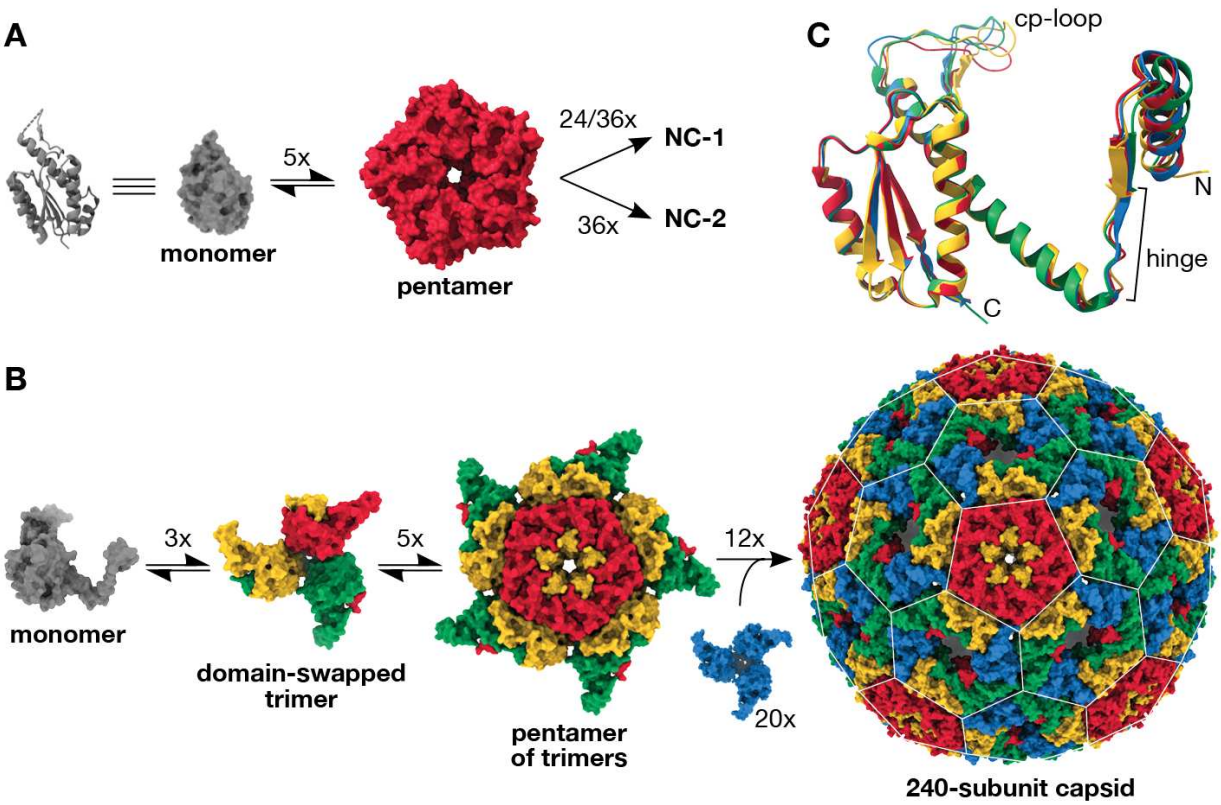
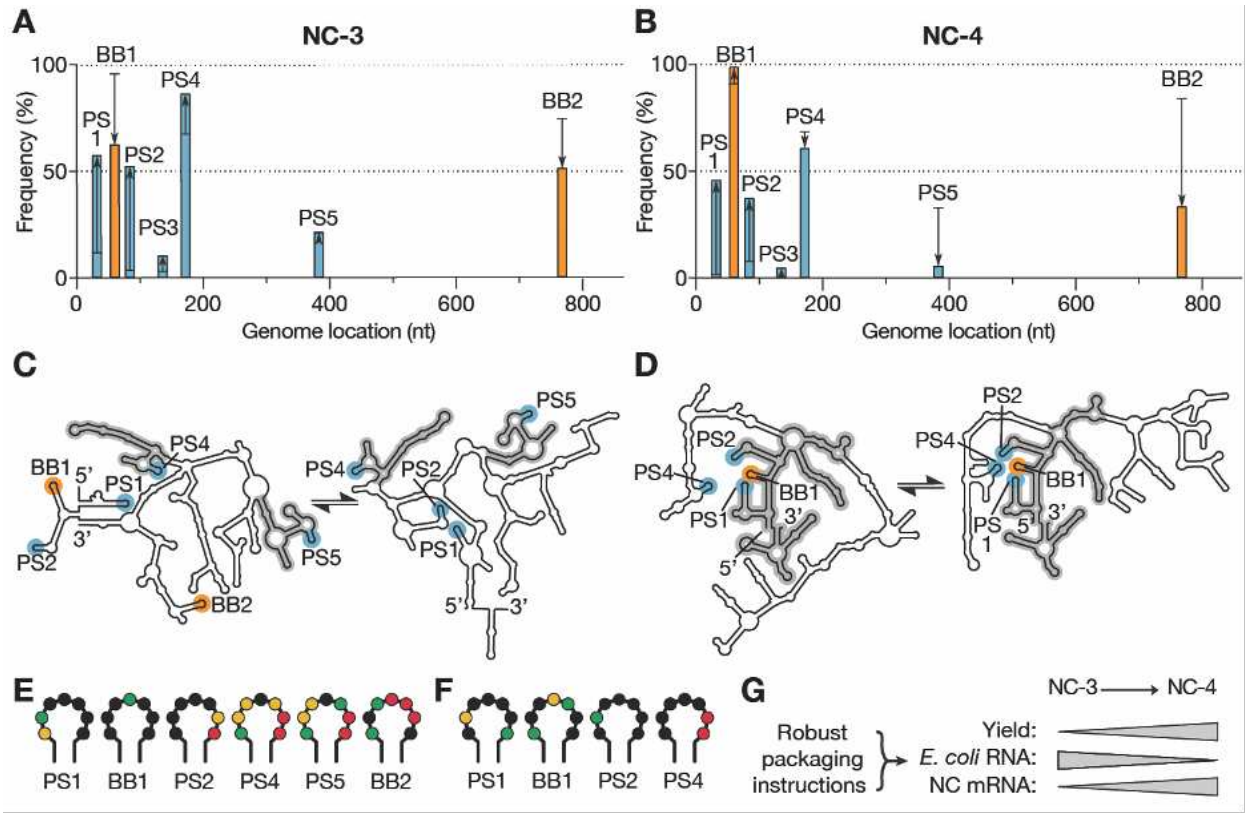


Fig. 3. Virus-like architecture by protomer reorganization. (**A**) The assembly of 120- and 180-subunit NC-1 and NC-2 cages from monomers (cartoon and surface shown in grey) presumably proceeds via AaLS-like pentamers. (**B**) Based on the assembly mechanisms of other viral capsids (*35*), the T=4 capsids likely arise from domain-swapped trimeric building blocks that further combine into pentamers. Combining the latter with additional domain-swapped trimers (blue) would afford the complete 240-subunit capsid. The pentagonal and hexagonal faces of the icosahedrally symmetric capsid are highlighted by a white lattice. (**C**) Assembly of the T=4 icosahedral NC-3 and NC-4 structures requires the subunits to adopt different, quasi-equivalent conformations. An overlay of the four quasi-equivalent chains of NC-4, colored as in panel B, shows that the hinge region provides flexibility for subtle adjustments in the relative orientation of the flanking segments. Additional differences are visible in the poorly resolved surface loop introduced by circular permutation (cp-loop), which interacts with the neighboring subunit in both pentamers and hexamers via a single short beta strand.



425

426 **Fig. 4. Virus-like genome packaging mediated by packaging signals.** (**A**,**B**) XRF reactivities were used
427 to calculate how frequently the seven packaging signal candidates occur in stem-loops in the NC-3 (A)
428 and NC-4 (B) mRNA genomes; 1000 sample folds were generated for each of 1116 combinations of
429 reactivity offsetting and scaling factors (see Figs. S12,S13). Cumulative display frequencies of the motifs
430 as stem-loop are plotted against genome position for the packaged transcripts (bars), with the high-affinity
431 BoxB tags highlighted in orange, BoxB-like PS1–5 motifs in blue, and the respective in vitro-transcribed
432 RNA indicated by black lines; arrows show the increase or decrease observed upon packaging. (**C**,**D**)
433 Two consensus folds predicted for packaged NC-3 and NC-4 mRNA (see also Figs. S12,S13). Secondary
434 structure features shared between the respective folds are highlighted in grey. These structures indicate
435 more extensive fold conservation in NC-4, as well as more robust display of a packaging cassette
436 comprising PS1, BB1, PS2, and PS4, than in NC-3. (**E**,**F**) Reactivities of the URxRxRR motifs displayed
437 in the packaged NC-3 (E) and NC-4 (F) RNA folds depicted in panels (C) and (D), respectively.
438 Reactivity follows the order: red (high)→ yellow→ green→ black (low). The four packaging signal
439 candidates in NC-4 show low reactivities, consistent with protection by capsid protein. (**G**) The evolution
440 of a packaging cassette that steers efficient capsid assembly around the target RNA provides a compelling
441 explanation for the improved properties of NC-4 compared to NC-3.
442

Supplementary Materials for

Evolution of a virus-like architecture and packaging mechanism in a repurposed bacterial protein

Stephan Tetter†, Naohiro Terasaka†, Angela Steinauer†, Richard J. Bingham, Sam Clark, Andrew P. Scott, Nikesh Patel, Marc Leibundgut, Emma Wroblewski, Nenad Ban, Peter G. Stockley, Reidun Twarock, Donald Hilvert*.

†These authors contributed equally.
*Correspondence to: donald.hilvert@org.chem.ethz.ch

This PDF file includes:

Materials and Methods

The nomenclature for the previously published nucleocapsids (*6*) was simplified to make the evolutionary relationship between the different variants clearer. NC-1 corresponds to λcpAaLS, the original nucleocapsid generated by circular permutation of AaLS and addition of the λN+ peptide to the new lumenal N-terminus; NC-2 is the best variant obtained after one round of optimization, λcpAaLS-β16; and NC-3 is the best performing variant from the second evolutionary round, λcpAaLS-α9 (*6*). The final variant, NC-4, was evolved in the current study.

Library construction by error-prone PCR

Error-prone PCR was carried out using the JBS Error-Prone Kit (#PP-102, Jena Bioscience) according to the manufacturer's protocol. Two primers (primer 1: 5'- GCG GAT AAC AAT TCC CCT CTA GAG; primer 2: 5'- GGG TTA TGC TAG TTA TTG CTC AGC G) were used with pMG-dB-λcpAaLS-α9 (*6*) as a template. PCR products were purified using the Zymoclean Gel DNA Recovery Kit (#D4001, Zymo Research). Both the products and the acceptor vector (pMG-dB) were doubly digested at their NdeI and XhoI restriction sites. The DNA fragments were purified using the DNA Clean & Concentrator-5 (#D4013, Zymo Research), ligated with T4 DNA ligase (#M0202, NEB), and purified again using the same kit. The capsid library (~1 µg ligation product) was transformed into electrocompetent *Escherichia coli* XL1-Blue cells by electroporation. The cells were incubated in 50 mL Luria-Bertani (LB) medium for 1 hour at 37 °C. The library size (~3 x 10^6 mutants) was determined by plating serial dilutions of the cell suspension onto LB-agar plates containing ampicillin (50 µg/mL). To the remaining cells, LB medium and ampicillin (50 µg/mL) were added to the original volume of 50 mL. Cells were cultured overnight at 37 °C and 230 rpm. The next day, plasmid DNA was extracted using the ZR Plasmid Miniprep Classic Kit (#D4016, Zymo Research).

Directed evolution of NC-4 from NC-3

Evolution of NC-4 was based on a plasmid library generated by error-prone PCR as described above. The library was subjected to three iterative cycles of selection. Each cycle involved transformation of *E. coli* cells, expression of the nucleocapsid variants, isolation, and purification by affinity and size, nuclease treatment, RNA extraction, reverse transcription, and re-ligation of the surviving variants into the vector backbone. Nuclease selection stringency was increased in each cycle.

Initially, electrocompetent *E. coli* BL21(DE3)-gold cells were transformed with the plasmid library and incubated in 4 mL LB medium for 1 hour at 37 °C. After adding ampicillin (50 µg/mL), cells were cultured at 37 °C and 230 rpm for an additional 6 hours. This 4-mL culture was then transferred to 400 mL LB medium containing ampicillin (50 µg/mL) and cultured as before until the OD_{600} reached 0.4–0.6, at which point protein production was induced by adding isopropyl ß-D-1-thiogalactopyranoside (IPTG) to a final concentration of 0.2 mM. Cells were cultured at 20 °C and 230 rpm for 16 hours, then harvested by centrifugation at 5,000 *g* and 4 °C for 10 min. The pellet was stored at -20 °C until purification. For purification, cells were resuspended in 15 mL lysis buffer (50 mM sodium phosphate buffer, pH 7.4) containing 1 M NaCl and 20 mM imidazole supplemented with lysozyme (1 mg/mL; #A3711, AppliChem) and DNase I (10 µg/mL; #A3778, AppliChem). The mixture was incubated at room temperature for 1 hour. After lysis by sonication and clearance by centrifugation at 9,500 *g* and 25 °C for 25 min, the supernatant was loaded onto 2 mL of Ni(II)-NTA agarose resin (QIAGEN) in a gravity flow column. Beads were washed with lysis

buffer containing 1 M NaCl and 20 mM imidazole, and protein was eluted with lysis buffer containing 200 mM NaCl and 500 mM imidazole. The buffer was exchanged to 50 mM sodium phosphate buffer (pH 7.4), 5 mM EDTA (storage buffer) containing 200 mM NaCl, using an Amicon Ultra-15 centrifugal filter unit (30 kDa MWCO, Merck Millipore). Capsids were further purified by size-exclusion chromatography (SEC) on a Superose 6 increase 10/300 GL (GE Healthcare) in storage buffer with 200 mM NaCl. Proteins were purified at room temperature.

For the selection of capsids that protect their RNA genome from nucleases, a solution of capsids containing approximately 2 µg of total RNA was treated at 37 °C for 1 hour with benzonase (2.5 U/µL; #101654, Merck Millipore) in 250 µL of storage buffer supplemented with 5 mM $MgCl_2$. RNA was then extracted with TRIzol reagent (#15596026, Invitrogen) and dissolved in water. The resulting RNA sample was incubated with RQ1 DNase (#M6101, Promega) in the manufacturer's reaction buffer at 37 °C for 1 hour and subsequently purified by phenol-chloroform extraction and ethanol precipitation. From this RNA, complementary DNA (cDNA) was prepared by reverse transcription with primer 3 (primer 3: 5'- GCG GAT AAC AAT TCC CCT CTA GAG) using SuperScript III reverse transcriptase (#18080044, Invitrogen) according to the manufacturer's protocol. The resulting cDNA was amplified in 30 PCR cycles with Phusion High-Fidelity DNA polymerase (#M0530, NEB) using primers 3 and 4 (primer 4: 5'- GGG TTA TGC TAG TTA TTG CTC AGC G). Purified DNA was digested with NdeI and XhoI restriction enzymes and ligated into the pMG-dB acceptor vector.

The resulting plasmid library was then employed in the next cycle, carried out as above, but in the presence of RNase A (10 µg/mL; #R4875, Sigma-Aldrich) instead of benzonase. The surviving variants were then subjected to a third cycle, with two modifications of the selection protocol. First, gel filtration was carried out on a HiPrep 16/60 Sephacryl-S400 (GE Healthcare) column, which has poorer resolution than the previously used column, but its higher exclusion volume allows efficient removal of larger aggregates. Second, nuclease treatment was extended to 4 hours and performed with a mixture of RNase A (10 µg/mL) and 2 vol% of an RNase cocktail enzyme mix from Thermo (#AM2288). From variants surviving the third selection cycle, 12 clones were picked, sequenced, and produced in *E. coli*. After affinity purification and SEC, variant NC-4 was chosen for detailed analysis based on its yield, RNA packaging, minimal aggregation, and structural homogeneity.

Production and purification of NC-1, NC-2, NC-3, NC-4, NC-4*, and Δλ-NC-4

All NCs were produced in *E. coli* BL21(DE3)-gold cells. Two-liter Erlenmeyer flasks containing 800 mL LB medium were inoculated with 8 mL overnight cultures and incubated at 37 °C and 200 rpm until the OD_{600} reached 0.5–0.7. Protein production was induced by adding IPTG to a final concentration of 0.5 mM. Cells were cultured at 25 °C for 18 hours and then harvested by centrifugation at 6,000 *g* and 15 °C for 20 min. The cell pellet from one 800-mL culture was resuspended in 20 mL LB medium, transferred and split into two 50-mL Falcon tubes. The medium used for transfer was removed by centrifugation at 4,000 *g* and 15 °C for 10 min, decanted, and aliquots of the cell pellet were frozen in liquid nitrogen and stored at −20 °C until purification. For purification, a cell pellet corresponding to 400 mL of culture volume was resuspended in 20 mL lysis buffer (50 mM sodium phosphate buffer at pH 7.4) containing 20 mM imidazole, and either 200 mM (NC-3), 500 mM (NC-1 and NC-2), or 1 M (NC-4, NC-4*, and Δλ-NC-4) NaCl. The lysis buffer was supplemented with lysozyme (1 mg/mL). The mixture was incubated at room temperature for 20 min on an orbital shaker. After lysis by sonication (5 cycles of 1 min on, 1 min off, with amplitude = 80 and cycle = 60, UP200S sonicator,

Hielscher Ultrasonics GmbH) and clearance by centrifugation at 8,500 *g* and 15 °C for 25 min, the supernatant was loaded onto 3 mL of Ni(II)-NTA agarose resin in a gravity flow column. After incubation for 10 min and washing with lysis buffer containing 20 mM imidazole, NCs were eluted with elution buffer (50 mM sodium phosphate buffer at pH 7.4, 500 mM imidazole) containing 200 (NC-3) or 500 mM (NC-1, NC-2, NC-4, NC-4*, and Δλ-NC-4) NaCl. The eluted fractions were concentrated and buffer-exchanged into storage buffer containing 200 mM (NC-3 and NC-4) or 500 mM (NC-1 and NC-2) NaCl using Amicon Ultra-15 centrifugal filter units (100 kDa MWCO, Merck Millipore). Protein capsids were further purified by SEC at room temperature using a Superose 6 increase 10/300 GL column equilibrated in storage buffer containing 200 (NC-3, NC-4, NC-4*, and Δλ-NC-4) or 500 mM (NC-1, NC-2) NaCl. Purified fractions were pooled, concentrated, aliquoted, and either analyzed immediately, or frozen in liquid nitrogen and stored at –80 °C. Where stated, NC-4 was further purified by anion exchange chromatography at room temperature using a MonoQ 10/100 column (Pharmacia Biotech). The mobile phase consisted of storage buffer containing 200–1000 mM NaCl.

For NC-3 and NC-4 variants, protein and RNA concentrations were measured by UV absorbance and deconvoluted using a previously reported protocol (*36*). For NC-1 and NC-2, this calculation could not be applied, likely because scattering from aggregating particles skewed absorbance values. Extinction coefficients for proteins were calculated using the ExPASy ProtParam tool (*37*). Wild-type AaLS was produced and purified as previously reported (*38*).

Negative-stain transmission electron microscopy (TEM)

Negative-stain TEM was performed as reported previously (*39*). Briefly, TEM grids (#01814-F, Ted Pella, Inc.) were negatively glow discharged at 15 mA for 45 s with a Pelco easiGlow Glow Discharge Cleaning System. After FPLC purification, grids were incubated with the capsid solution (10 μM monomer in storage buffer containing 200 mM NaCl) for 1 min, washed twice with doubly distilled water (ddH_2O), and once with TEM staining solution (2% wt/vol aqueous uranyl acetate, pH 4), after which the grids were incubated with staining solution for 10 s, dried, and imaged using a TFS Morgagni 268 microscope.

In vitro transcription of reference mRNAs

Reference messenger RNAs (mRNAs) for real-time quantitative PCR (RT-qPCR) were prepared by runoff *in vitro* transcription. DNA templates were prepared by PCR with primer 5 (primer 5: 5'- GCG AAA TTA ATA CGA CTC ACT ATA G) and primer 6 (primer 6: 5'- CAA AAA ACC CCT CAA GAC CC) from plasmids pMG-dB-NC-1 to pMG-dB-NC-4 using the LongAmp Taq assay (#M0287, NEB). PCR-amplified templates were gel-purified using the DNA Clean & Concentrator-5 kit. *In vitro* transcription reactions were performed using T7 RNA polymerase (#EP0111, Thermo Scientific) according to the manufacturer's protocol. Template DNA was digested by RQ1 DNase and RNA was precipitated with isopropanol. RNA samples were purified twice by denaturing polyacrylamide electrophoresis (PAGE). Briefly, preparative urea PAGE gels (20 cm x 16 cm x 0.1 mm) were prepared in Tris/borate/EDTA (TBE) buffer supplemented with 8 M urea and 5% polyacrylamide. Polymerization was initiated using TEMED (8 μL per 10 mL gel solution) and APS (10% in water, 90 μL per 10 mL gel solution). RNA bands were visualized by UV shadowing and excised with a scalpel. The gel pieces were crushed with a pipet tip and the RNA was extracted in water containing 0.3 M NaCl overnight at room temperature. The next day, the RNA was purified by ethanol precipitation and dissolved in water. RNA quality and purity were assessed by measuring A260/A280 and A260/A230 ratios

(for pure RNA, both ratios are ≥2.0) and by analytical PAGE gels. RNA concentrations were measured using the Qubit RNA HS assay (#Q32852, Invitrogen).

Extraction of nucleocapsid RNA and RT-qPCR

RNA was extracted from 100-µL or 200-µL aliquots of purified NCs containing a total amount of 5–10 µg RNA using the RNeasy Mini kit (#74104, QIAGEN) following the manufacturer's instructions. RNA standards were prepared by *in vitro* transcription as described above. After ensuring that RNA samples were free of contaminants by absorbance, concentrations from extracted RNAs and *in vitro*-transcribed standards were measured with the Qubit RNA HS Assay. cDNA of the capsid's genome was prepared by reverse transcription with primer 7 (5'- CCA AGG GGT TAT GCT AGT TAT TGC TCA GC) and SuperScript III reverse transcriptase (#18080044, Invitrogen) according to the manufacturer's protocol. After the reverse transcription reaction, RNase H (#18021014, Invitrogen) was added to digest RNA transcripts. Immediately following the reverse transcription reaction, dilutions of the cDNA were mixed with KOD SYBR qPCR Master Mix (#QKD-201, TOYOBO), primers 8 (5'- TGT GAG CGG ATA ACA ATT CCC CTC) and 9 (5'- GGG TTA TGC TAG TTA TTG CTC AGC G), and ROX reference dye according to the manufacturer's protocol. cDNA was amplified in 40 PCR cycles on a StepOnePlus thermocycler (Applied Biosystems) employing the thermocycler-specific PCR conditions provided in the qPCR mix manual. Absolute amounts of full-length genome were determined using standard curves prepared with cDNA originating from highly pure *in vitro*-transribed reference RNAs. The full RT-qPCR experiments to quantify the fraction of full-length mRNA in the total isolated RNA were repeated in two separate laboratories (by Angela Steinauer at ETH Zurich and by Naohiro Terasaka at the University of Tokyo) to ensure reproducibility.

Long read sequencing

Nanopore sequencing was performed as described previously (*6*). Oxford Nanopore Technology relies on polyadenylated RNAs. Therefore, the extracted NC RNAs were polyadenylated using *E. coli* poly(A) polymerase (#M0276, NEB) and purified using the RNA Clean & Concentrator-5 kit (#R1015, Zymo Research). cDNA libraries were prepared with the Direct cDNA Sequencing Kit (#SQK-DCS109, Oxford Nanopore Technologies) and Native Barcoding Kit 1D (#EXP-NBD104, Oxford Nanopore Technologies) following the manufacturer's protocols. Sequencing was carried out in a flow cell (#FLO-MIN106) using the 72-h 1D protocol. Base calling and de-multiplexing were performed using Oxford Nanopore Technology's Guppy Basecalling Software (version 3.2.10+aabd4ec). Adapter sequences of demultiplexed reads were removed using Porechop (version v0.2.4, https://github.com/rrwick/Porechop). Reads were mapped to the plasmid reference genome and to the *E. coli* genome (RefSeq: NC_000913.3) using Minimap2 (version 2.17 (r941), https://github.com/lh3/minimap2). Index reference files containing the pMG plasmid genomes and the *E. coli* genome were prepared using samtools (version 1.10, https://github.com/samtools/). Index reference files and mapped reads were imported into CLC Genomics Workbench (version 12.0, QIAGEN Bioinformatics). Alignments were sorted and the read sequences and lengths corresponding to the most abundant gene classes were extracted using samtools. For each gene, we calculated the sum of all gene-specific base pairs and compared it to the sum of all recorded base pairs.

Nuclease challenge assay

Aliquots of nucleocapsids containing a total amount of 5–10 μg RNA were treated in 50 mM sodium phosphate buffer at pH 7.4, 200 mM NaCl, 5 mM EDTA, 5 mM $MgCl_2$ either lacking nuclease or supplemented with benzonase (2.5 U/μL; 101654, Merck Millipore) or RNase A (10 μg/mL; #R4875, Sigma-Aldrich). The concentration of RNA and protein was held constant at 80 ng RNA/μL, which corresponds to about ~5 μg protein/μL. Aliquots were challenged with the respective nucleases at 37 °C for the indicated time periods after which samples were frozen in liquid nitrogen and stored at –80 °C. For analysis, RNA was extracted from the capsid as previously described (*5*). A nuclease-treated NC solution (100 μL) was mixed with TRIzol (500 μL), vortexed for 3–5 s, and left on ice for 10 min. Then, chloroform (100 μL) was added, the samples were vortexed again, and the two phases were separated by centrifugation at 15,000 *g* for 20 min. The upper, aqueous layer (~300 μL) was carefully transferred to a clean Eppendorf tube and mixed with an equal volume of 20% ethanol in nuclease-free water. This extraction step was essential to remove nuclease contamination. Subsequently, the solution was transferred to an RNeasy Mini spin column, and RNA purified according to the manufacturer's protocol.

RNA stability was visualized on denaturing PAGE gels. Analytical urea PAGE gels (8.3 cm x 7.3 cm x 0.1 mm) were prepared in TBE buffer supplemented with 8 M urea and 8% polyacrylamide. Polymerization was initiated using TEMED (8 μL per 10 mL gel solution) and ammonium persulfate (APS) (10% in water, 90 μL per 10 mL gel solution). Gels were loaded with equal volumes of extracted RNA. The NC genome was selectively visualized using the fluorogenic dye DFHBI-1T (#446461, United States Biological), which fluoresces upon binding to the Broccoli aptamer that is part of the BoxBr tags (*13*). Total RNA was visualized using GelRed (#41002, Biotium).

Cryo-electron microscopy: data collection and image processing

Freshly purified NCs were concentrated in storage buffer. NC-1 and NC-2 eluted as two major peaks from the SEC column, both of which were pooled for analysis. These variants were concentrated to a 280-nm absorbance of 20–30, as the protein concentration could not be estimated accurately due to the high absorbance ratio of 260/280 nm mentioned above. NC-3 and NC-4 were concentrated to 4–5 mg/mL. Copper-supported holey carbon grids (R2/2 Cu 400, Quantifoil) were negatively glow discharged at 15 mA for 15 s with a Pelco easiGlow Glow Discharge Cleaning System. Then, 3.5 μL of sample were applied and blotted with a vitrobot (FEI) for 12 to 14 s at 25 blot strength, 100% humidity, and 22 °C. Grids were plunged into liquid ethane and stored in liquid nitrogen.

Initial screening for all capsids and data collection for NC-3 were performed with a TFS Tecnai F20 equipped with a Falcon II direct electron detector (FEI). Movies of 7 frames were collected at a total dose of 40 electrons per $Å^2$ and a magnification of 62,000x (1.8 Å pixel size). Defocus ranged from –1.8 to –3.3 μm. NC-1, NC-2, and NC-4 data collection was performed on a Titan Krios equipped with a Falcon III direct electron detector (FEI). Movies of 40 frames were collected at a dose of 60 electrons per $Å^2$ and a magnification of 130,000x (1.1 Å pixel size). NC-4 was collected in electron counting, NC-1 and NC-2 in integration mode. Defocus ranged from –0.8 to –2.7 μm.

All single-particle reconstructions were performed in Relion 3.0 (*40*). Motion correction was performed with MotionCor2 (*41*) implemented in Relion, contrast transfer function (CTF) estimation with GCTF (*42*). Good micrographs were selected based on metadata values and

manual inspection. For NC-1, NC-2, and NC-3, early classifications were performed with CTF ignored up to the first peak to avoid grouping into a few, featureless classes.

Reconstruction of NC-1 and NC-2 structures was complicated by heterogeneity and aggregation. 2D classification was performed with multiple different mask sizes in order to obtain classes with distinct features for differently sized species. These classes were then used for the generation of initial 3D models of the tetrahedrally symmetric 120-mer (NC-1) and 180-mers (NC-1 and NC-2). For the final reconstruction though, as shown in Figs. S5 and S6, 2D classification was performed with a single large mask and size differences mainly separated subsequently in 3D classification, as this procedure led to higher particle numbers and consequently better resolution. We tried to reconstruct additional 3D structures from the heterogenous particles, but no other reasonable models could be obtained. Single- or multi-reference 3D classification based on known AaLS-derived structures, such as the 240-subunit capsid, or hollow spheres were not successful. The inability to extract further structures from the samples likely reflects substantial aggregation, low particle numbers of individual capsid architectures, shape irregularities, and lower symmetry.

In contrast to NC-1 and NC-2 particles, NC-3 and NC-4, were better behaved and more homogeneous, making data elaboration according to standard procedures (*40*) fairly straightforward. Good 2D classes were used to generate initial models with imposed icosahedral symmetry. The best classes from 3D classification, masked around the capsid shells, were further refined. Further processing steps are described in Fig. S7.

Model building and refinement were performed in Coot 0.8.9.2 (*43*), Phenix 1.18 (*44*), and Pymol 2.0. Electron density maps from 3D refinement, postprocessing in Relion, and autosharpening in Phenix were used during model building. NC-1, NC-2, and NC-4 models were based on a crystal structure of the wild-type lumazine synthase (PDB-ID: 1hqk).

Atomic models were initially built into the asymmetric units and refined. After symmetry expansion, the full capsids were refined with non-crystallographic symmetry constraints to reflect the symmetry imposed during reconstruction. Experimental data versus model geometry were weighted in Phenix to optimize both electron density fit and geometry. While the core fold of the protomers in the tetrahedrally symmetric capsids were well resolved, the maps for the segment encompassing residues 66–81 displayed lower local resolution in subunits where this area is exposed towards the capsid openings and not in contact with neighboring protomers. These segments were built by repositioning the known structural elements of the wild-type protein as rigid groups and remodeling the flanking linkers according to visible density and chemical constraints, although multiple alternative conformations may exist. The pseudo-atomic NC-3 model is based on the structure of NC-4 with reversion of the mutations and an additional cycle of refinement to satisfy geometric and steric constraints. More information on data collection and model building is found in Table S1.

XRF data collection and analysis

XRF experiments were performed as described (*26*) and analyzed using QuShape (*45*) modified to incorporate sample replicate comparisons. *In vitro*-transcribed and NC-packaged RNAs were exposed in triplicate to X-ray pulses of 25 or 50 ms at the National Synchrotron Light Source II, beamline 17-BM XFP at Brookhaven National Laboratory (Upton, NY). Packaged RNA was subsequently extracted from the protein shell by standard techniques (*26*).

Nucleotide modification propensity (reactivity) is directly related to residue mobility, and thus reflects base pairing and inter-molecular contacts. Reactivity was quantitated post-exposure by

capillary electrophoresis sequencing using three dye-labeled primers (primer 10: 5'- CCA AGG GGT TAT GCT AGT TAT TGC TCA GC; primer 11: ATG CTA CGA TAC CGA AAC GAA GGC; primer 12: 5'- CTC GAT AGC CTG TTC CAA GGT G) that cover ~90% of the genome, including the coding region of the structural protein. Pairwise Pearson correlation coefficients (PCC) for normalized replicates gave best correlations overall at 50 ms exposure (Table S3), and the respective data were therefore further analyzed. XRF footprints showing normalized reactivities for each nucleotide were generated for both the free and packaged state using established protocols (*27*, *28*).

To find potential packaging signals (PSs), we looked for sequences similar to the UGxAxAA motif (x, any nucleotide) and the UGxA submotif, which are known to bind the λN+ peptide (*25*). Eighteen occurrences of URxRxRR (R, any purine) and 21 of URxRxxx were identified in the mRNA transcripts of both NC-3 and NC-4 (Table S2). Contact with the RNA-binding peptide would result in lowered reactivity. Of the 39 identified sites, only 13 (highlighted in grey in Table S2) displayed such low reactivity levels (green and black colored nts in Table S2), including the two copies of the BoxB sequence (BB1 & BB2).

XRF reactivity levels were used as constraints to weight RNA secondary structure predictions. We used a modification of the RNA folding algorithm S-fold that includes such data via a scaling factor (*m*) and an offset (*b*) to generate a statistical sample of secondary structures from the Boltzmann ensemble of RNA secondary structures (*27*). Typically, the (*m*,*b*) combination that best represents a known secondary structure element within a probed RNA is identified, and that combination is used to predict the overall secondary structure (*28*). Because the reported stem-loop of the BoxBr tag contains C-G base pairs that stabilize the stem (*12*), it occurs with high probability in the ensembles for many (*m*,*b*) combinations, and it is therefore insufficiently discriminatory to identify a unique (*m*,*b*) combination. We therefore computed 1000 statistical (Boltzmann-weighted) sample folds for all 1116 (*m*,*b*) combinations for *m* values between 0 and 7 and *b* values between 0 and -6, in increments of 0.2. Computing multiple folds per (*m,b*) combination takes into consideration that large RNAs occur as ensembles of secondary structures with comparable folding free energies. For any of the sites to act as packaging signals, they must be presented with sufficient frequency in the ensemble. In order to identify trends, the maximum, minimum and average frequency of stem-loops overlapping with the identified motifs were computed over all sample folds and all (*m*,*b*) combinations tested, both for the *in vitro* transcript and packaged RNA. Of the 13 identified sites, only seven (BB1, BB2, and PS1–5; Table S2) appear as part of a loop in either NC-3 or NC-4 with significant frequency (>50% of the folds in >50 of the (*m*,*b*) combinations).

For the calculation of representative folds of the packaged NC-3 and NC-4 mRNA, (*m*,*b*) values were chosen taking into account contributions from all PSs which were preferentially displayed in the packaged over free transcripts via their cumulative normalized frequencies of occurrence (Figs. S12,S13). Two local probability maxima were identified for both packaged mRNAs, and structures computed for both. Maximum ladder distances for these folds (Figs. S12,S13) show that evolution did not select for genome compactness, likely because the mRNA is considerably smaller than the packaging capacity of the evolved T=4 capsids.

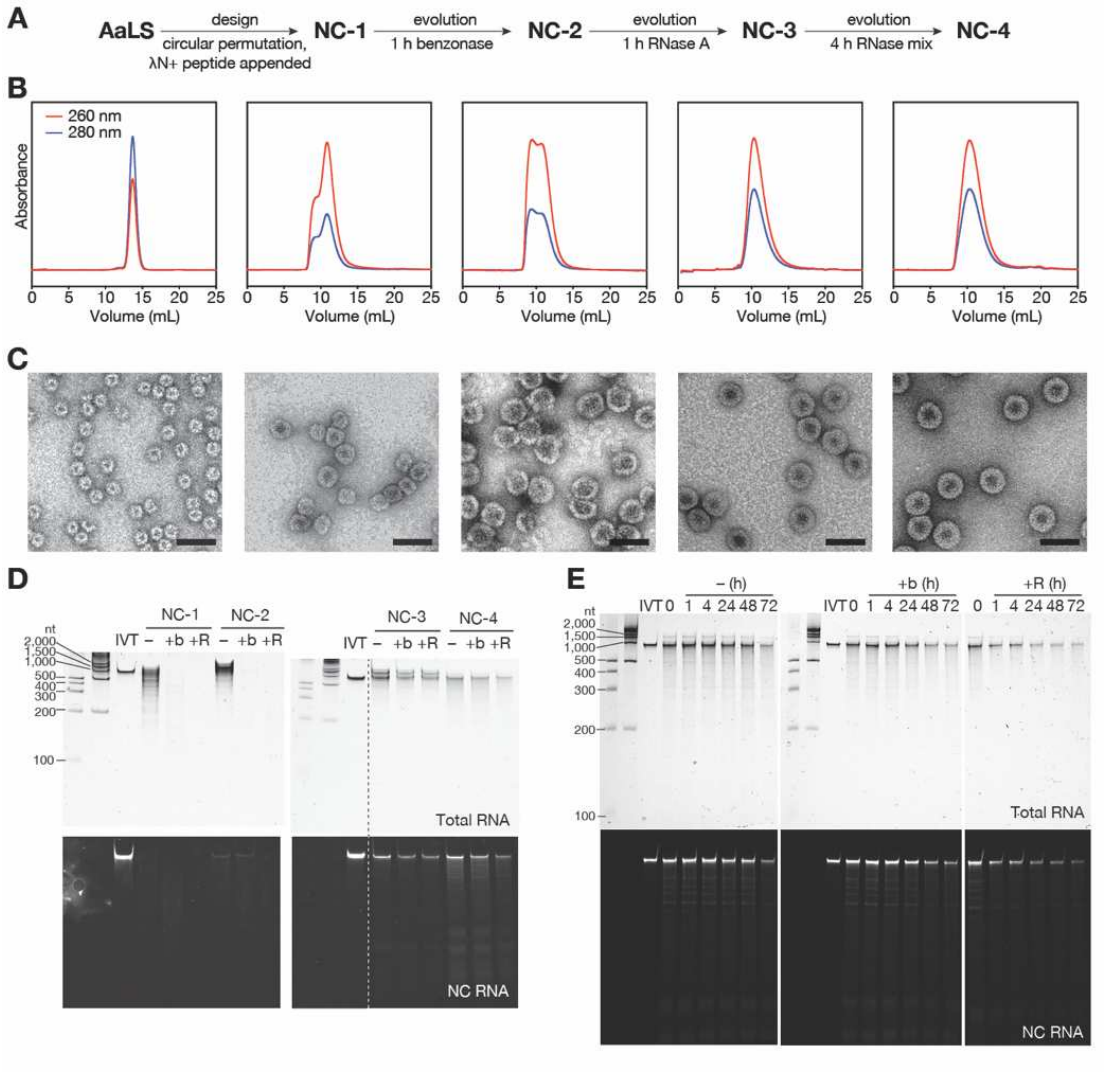


Fig. S1. Nucleocapsid design and evolution.

(**A**) NC-1 was generated by circular permutation of AaLS and addition of the λN+ peptide to the new N-terminus (*6*). Directed evolution over three generations with increasingly stringent nuclease challenge in each step yielded NC-4. As explained in the Materials and Methods section, the previously described nucleocapsids were renamed to clarify their evolutionary relationships. (**B**) Size-exclusion chromatograms of purified, re-injected nucleocapsids (column: Superose 6 increase 10/300 GL). (**C**) Transmission electron micrographs of purified capsids. Scale bar: 50 nm. (**D**) Capsid stability towards nucleases: Purified NCs were incubated without nuclease (–), or treated with benzonase (+b), or RNase A (+R) for 1 hour at 37 °C. RNA was extracted and equal volumes were loaded onto a denaturing PAGE (8%) gel. Total RNA was stained with GelRed, nucleocapsid mRNA (NC RNA) was visualized with DFHBI-1T, a small molecule that fluoresces upon binding to the broccoli aptamer present in the 5'- and 3'-untranslated regions of the capsid mRNA. IVT = in vitro-transcribed reference mRNA. The dashed line indicates two non-concurrent portions of the same gel image. (**E**) Purified NC-4 was treated with nucleases as in (D) for the indicated number of hours and analyzed on a denaturing PAGE (5%) gel.

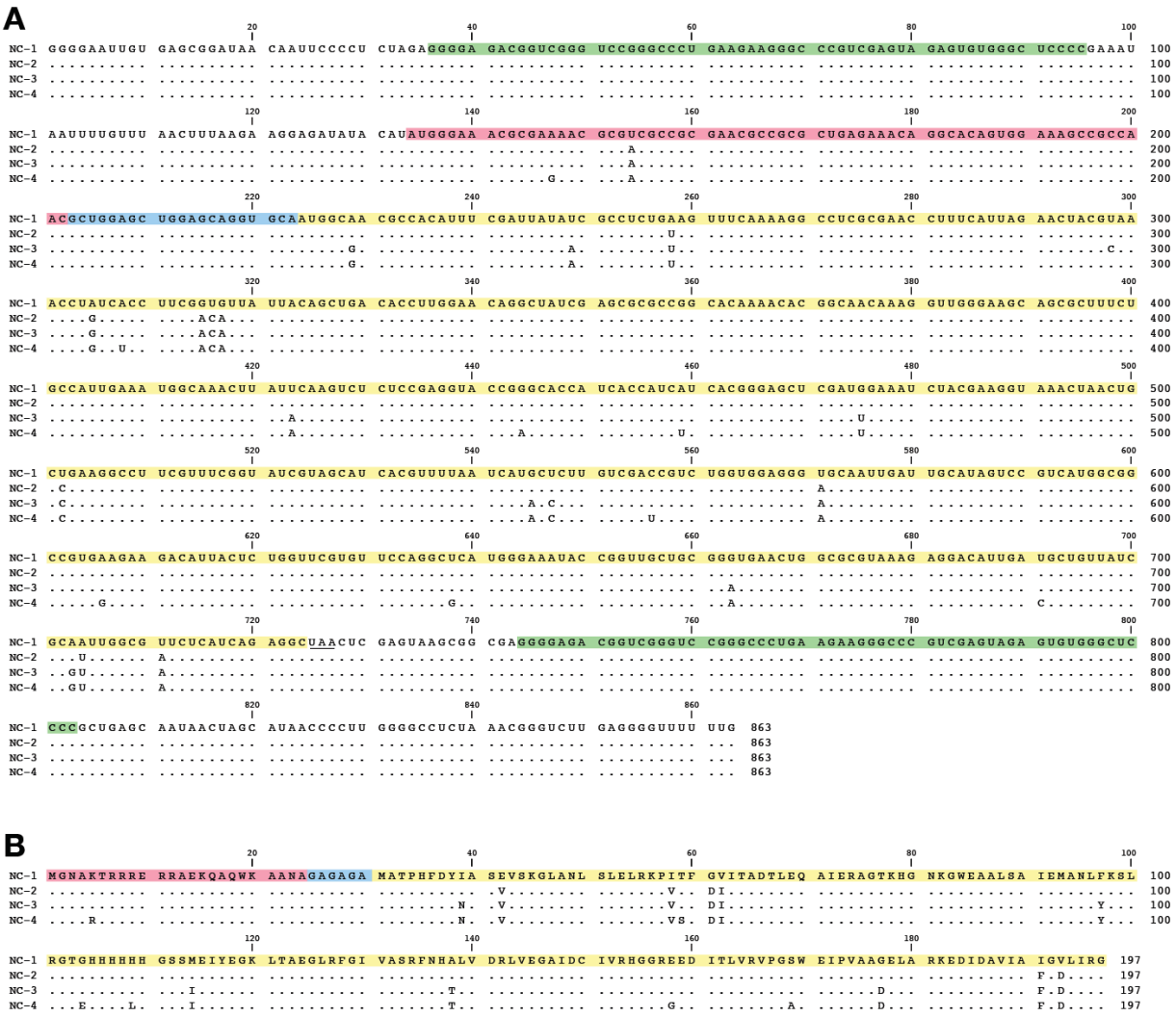


Fig. S2. Sequence alignment of NC-1 to NC-4.

(**A**) mRNA and (**B**) protein sequences of NC-1 to NC-4 (green = BoxBr tags, magenta = λN+ peptide, blue = (GlyAla)-linker, yellow = cpAaLS).

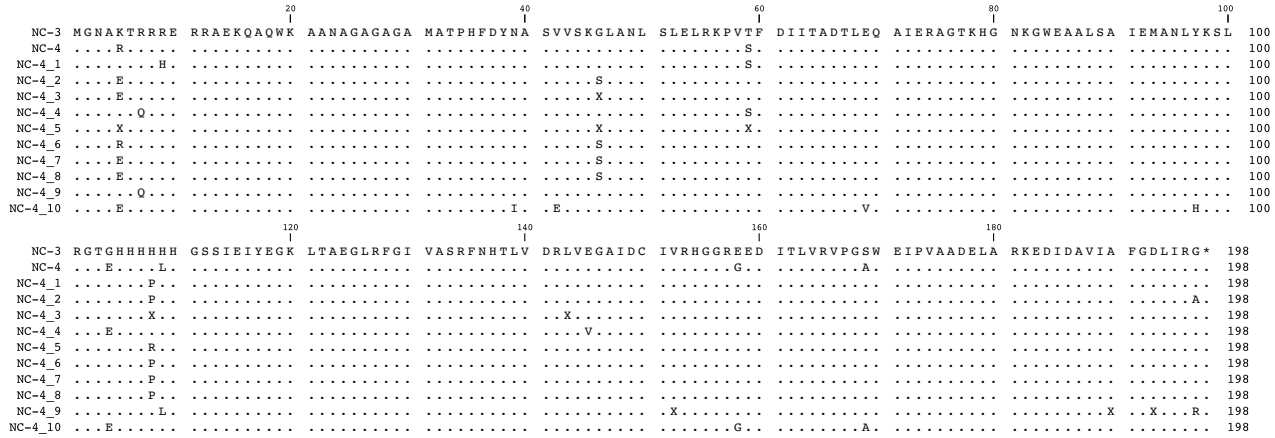
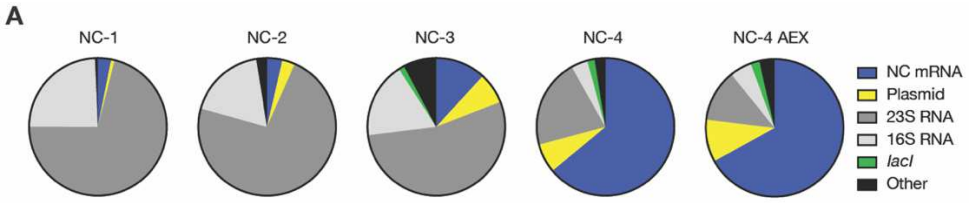


Fig. S3. Fourth generation nucleocapsids.

Protein sequences of the nucleocapsids that were isolated in the final round of selection are shown. X indicates sequencing ambiguity; for NC-4_11, sequencing failed. As noted in the text, NC-4 yielded homogeneous, well-behaved particles and was selected for detailed characterization. Like all the variants, it contains mutations in the N-terminal λN+ peptide and in the external loop introduced by circular permutation, which may influence RNA binding affinity and capsid assembly/stability, respectively.



B

NC-1		NC-2		NC-3		NC-4		NC-4 AEX	
Gene	% bp	Gene	% bp	Gene	% bp	Gene	% bp	Gene	% bp
NC-1	3.100	NC-2	3.600	NC-3	11.800	NC-4	64.000	NC-4	67.000
Plasmid	0.830	Plasmid	3.008	Plasmid	7.320	Plasmid	6.864	Plasmid	9.945
23S rRNA	71.128	23S rRNA	72.704	23S rRNA	53.952	23S rRNA	21.015	23S rRNA	12.293
16S rRNA	24.250	16S rRNA	18.166	16S rRNA	17.829	16S rRNA	3.851	16S rRNA	5.103
*lacI**	0.090	*lacI**	0.269	*lacI**	1.335	*lacI**	1.757	*lacI**	2.078
Other:		Other:		Other:		Other:		Other:	
rnpB	0.003	*rnpB*	0.004	*gltA*	0.199	*gltA*	0.278	*gltA*	0.306
ssrA	0.002	*ssrA*	0.004	*rnpB*	0.093	*acnB*	0.149	*acnB*	0.227
Other genes†	0.597	*gltA*	0.004	*gpmA*	0.062	*gpmA*	0.077	*idc*	0.096
		acnB	0.004	*ssrA*	0.039	*serS*	0.048	*rnpB*	0.095
		idc	0.002	*acnB*	0.035	*rnpB*	0.039	*gpmA*	0.084
		serS	0.002	*dadA*	0.017	*idc*	0.039	*aspC*	0.054
		acpP	0.001	*acpP*	0.012	*dadA*	0.030	*serS*	0.045
		dadA	0.001	*aspC*	0.011	*ychF*	0.025	*dadA*	0.044
		gpmA	0.001	*idc*	0.008	*aspC*	0.024	*acpP*	0.030
		Other genes†	2.230	*ychF*	0.005	*acpP*	0.010	*ychF*	0.021
				serS	0.003	*ssrA*	0.003	*ssrA*	0.010
				Other genes†	7.281	Other genes†	1.792	Other genes†	2.569
Total	100.000	Total	100.000	Total	100.000	Total	100.000	Total	100.000
#Reads	104,325	#Reads	81,784	#Reads	48,669	#Reads	92,166	#Reads	68,253

*The *lacI* gene is present in both the plasmid and the bacterial genome.
†Sum of all remaining bps aligned with the *E. coli* genome present at <0.001%.

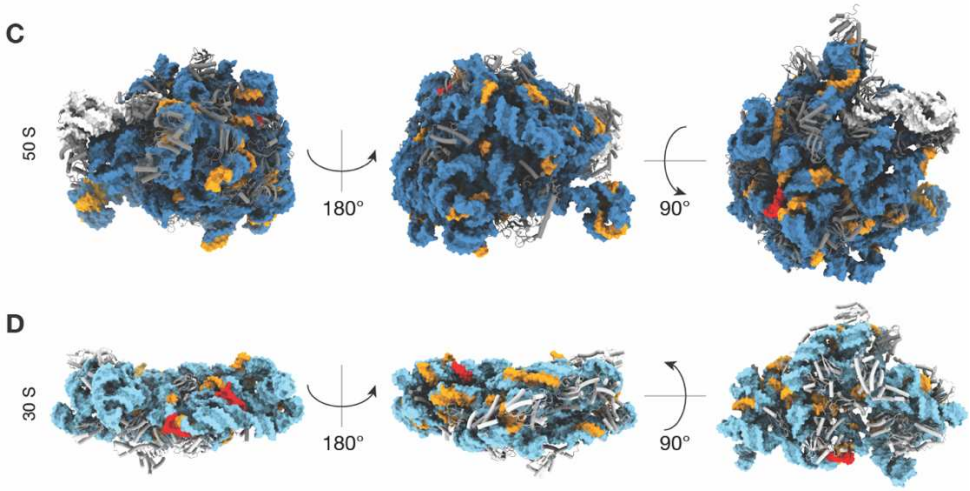


Fig. S4. RNA cargo identified by long-read sequencing.

(**A**) Pie chart representations of main gene classes identified by nanopore sequencing (*34*). (**B**) Relative fraction of identified gene classes in percent calculated by adding gene-specific base pairs (bps) and comparing them to the sum of all recorded bps. (**C** and **D**) UGxAxAA (red) and URxRxRR (orange) motifs are mapped onto the 50S (C) and 30S (D) subunits of the *E. coli* ribosome (PDB: 5h5u). The 23S rRNA is colored in dark blue, the 16S rRNA in light blue, accessory proteins are shown as cartoons. The ubiquity and compactness of ribosomes, together with interactions between some of the exposed BoxB-like RNA motifs with the λN+ peptide, may explain competitive encapsidation of the ribosome by the nucleocapsids.

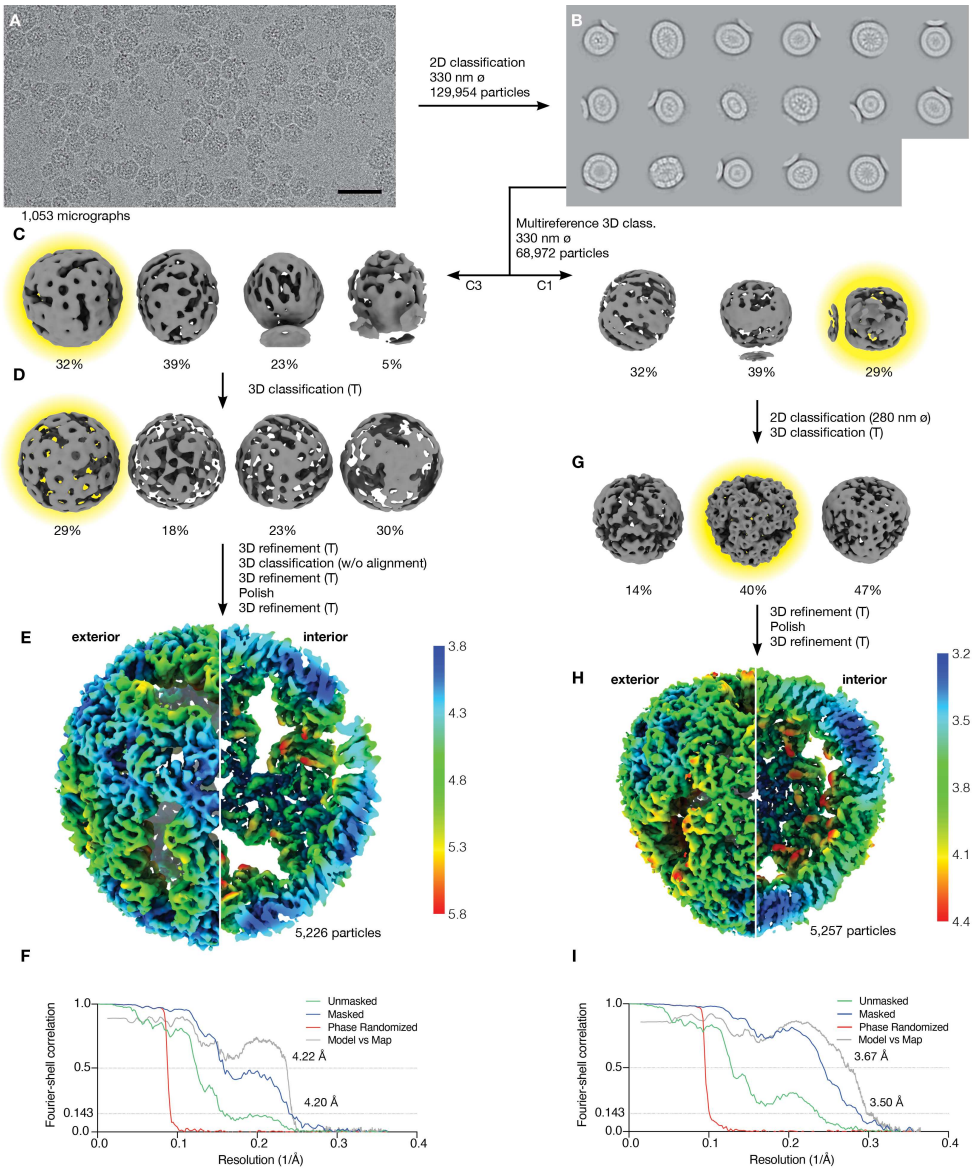


Fig. S5. Single particle reconstruction of NC-1 structures.

(**A**) 1,053 movies were analyzed for reconstruction of NC-1 (scale bar: 50 nm). (**B**) From these, 129,954 particles were picked and classified in 2D. (**C**) Size differences of the heterogeneous particles were initially classified in 3D with multiple references and lower symmetry (C1 and C3). References included the NC-4 structure (Fig. S7). (**D** and **G**), Particles were further classified with tetrahedral symmetry, using initial models generated beforehand by 2D classification with tight masks and then imposition of tetrahedral symmetry on classes showing distinct features. As indicated, further 2/3D classifications of classes highlighted in yellow, polishing, and refinement with imposed tetrahedral symmetry led to the final structures. (**E** and **H**) Refined maps, colored by local resolution, of the 180-mer (5,226 particles, 4.20 Å) (E) and the 120-mer (5,257 particles, 3.50 Å) (H). (**F** and **I**), Gold-standard Fourier-shell correlation curves for the 180-mer (F) and 120-mer (I).

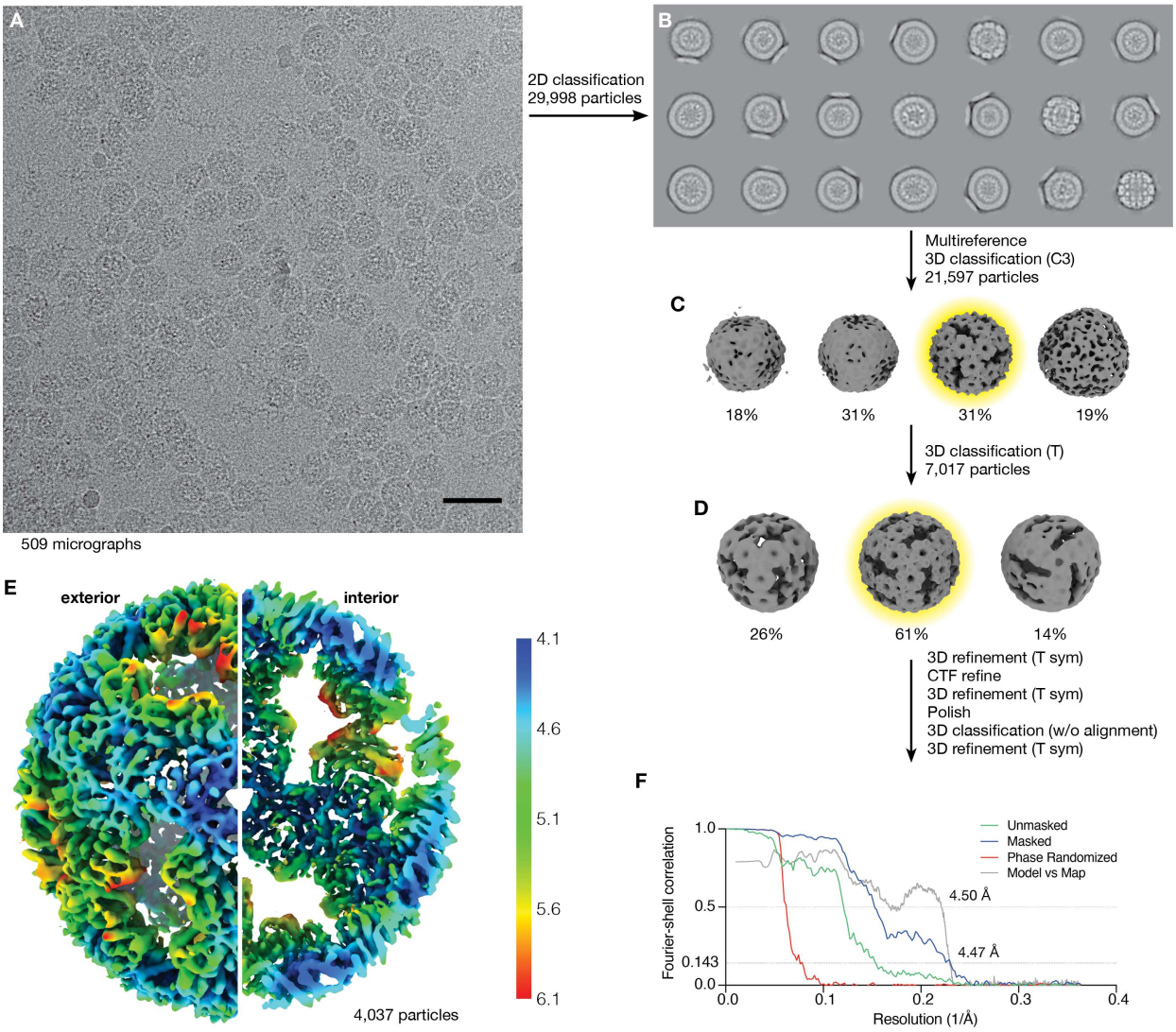
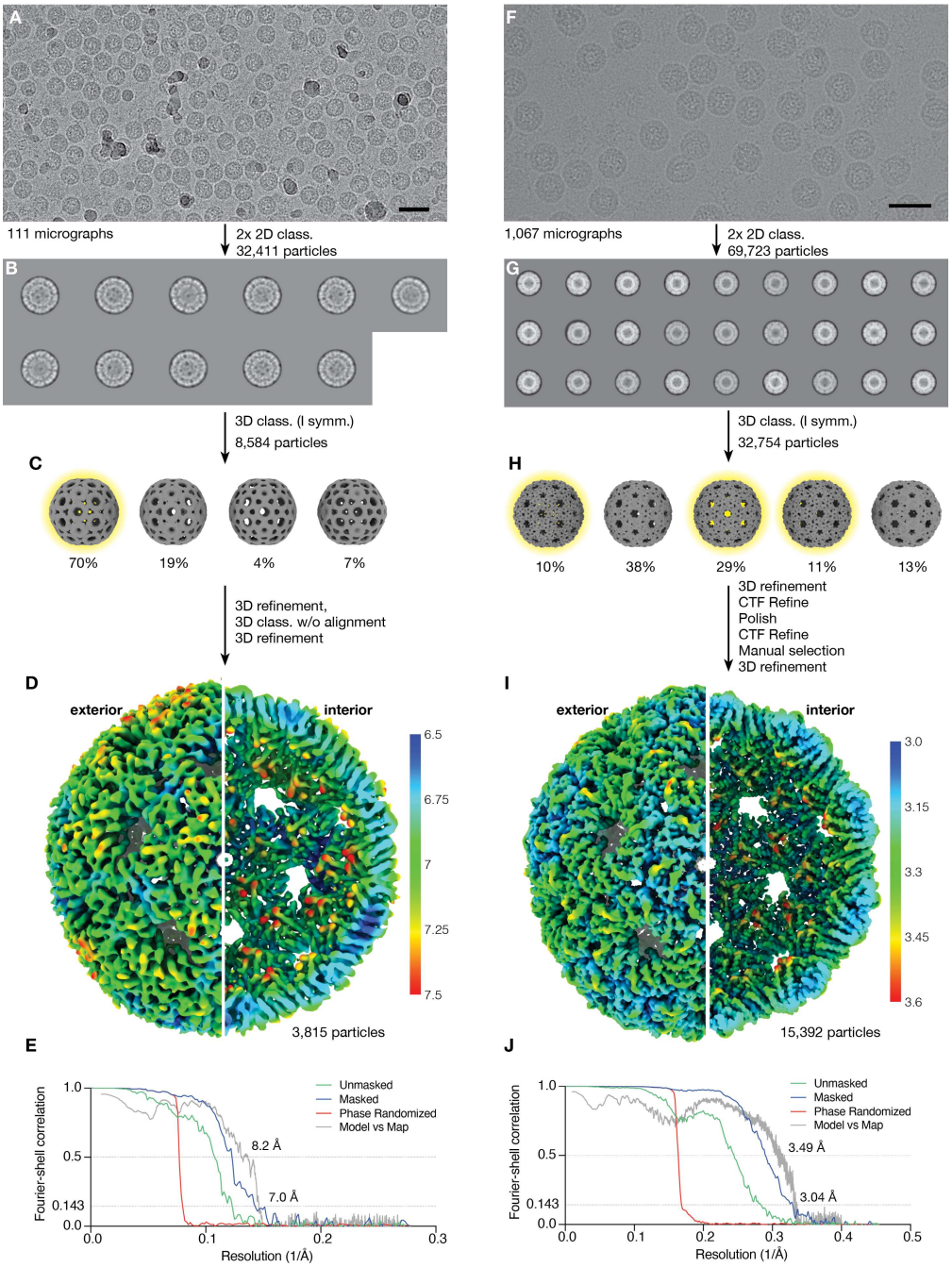


Fig. S6. Single particle reconstruction of NC-2.

(**A**) 509 movies were used to analyze NC-2 (scale bar: 50 nm). (**B**) 29,998 particles were classified in 2D. (**C**) Non-junk particles were further classified in 3D with multiple references (including the structure of NC-4, Fig. S7) with C3-symmetry. (**D**) Further 3D classification (with tetrahedral symmetry) of classes highlighted in yellow, contrast transfer function (CTF) refinement, polishing, and refinement with tetrahedral symmetry of individual classes led to the final structures. (**E**) Refined map, colored by local resolution (4,037 particles, 4.47 Å). (**F**) Gold-standard Fourier-Shell correlation curves.



841

842 **Fig. S7. Single particle reconstruction of NC-3 (A–E) and NC-4 (F–J).**

843 (**A** and **F**) Sample movies from NC-3 (A) and NC-4 (F) micrographs (scale bar: 50 nm). (**B** and **G**)
844 Particles were classified in 2D, and symmetric classes with clear features processed further. (**C** and **H**)
845 Initial model generation and 3D classification were performed with imposition of icosahedral symmetry.
846 For NC-4, multiple highly similar 3D classes (highlighted in yellow) were pooled. Particles were further
847 processed as indicated. (**D**) Postprocessed map of NC-3 (3,815 particles, 7.0 Å). (**I**) Refined map of NC-4
848 (15,392 particles, 3.04 Å). (**E** and **J**) Gold-standard Fourier-Shell correlation curves.
849

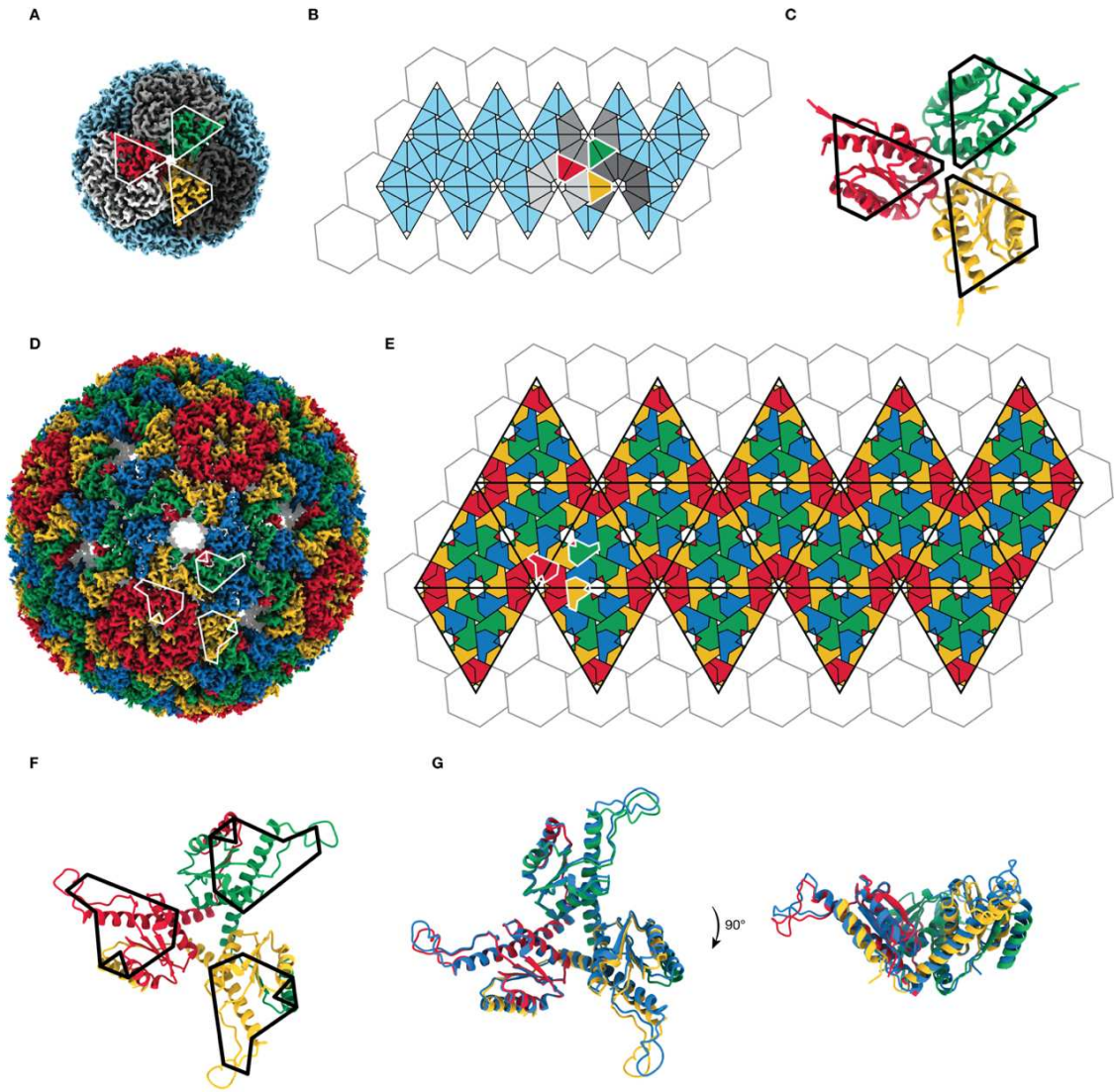


Figure S8: Architectural adaptation from T=1 to T=4.

(**A**) AaLS (EMDB:3538) is shown with pentamers highlighted in shades of grey, and three symmetric monomers highlighted in red, green and yellow. (**B**) Representation of the capsid in panel A on a gyrated hexagonal lattice (*46*). (**C**) Three C3-symmetric AaLS monomers outlined by black (C) or white (A and B) solid lines barely interact in the wildtype fold. (**D**) NC-4 structure. A trimer composed of quasi-symmetric monomers is highlighted in red, green, and yellow in NC-4. (**E**) A lattice representation of NC-4 shows that its intricate architecture is still based on the same gyrated hexagonal lattice as AaLS. Differences arise due to the domain swap and the addition of the external loop introduced by circular permutation. (**F**) Three quasi-symmetric monomers, outlined by black (F) or white (D and E) lines, interact closely after the domain swap. (**G**) An overlay of the two different sets of trimers (Fig. 3B) composing NC-4 further highlights the quasi-equivalence between chains that compose the 240-mer. The "blue" trimer has a slightly more acute hinge angle than the red/green/yellow trimers, which allows it to adapt to its location between three hexagonal patches.

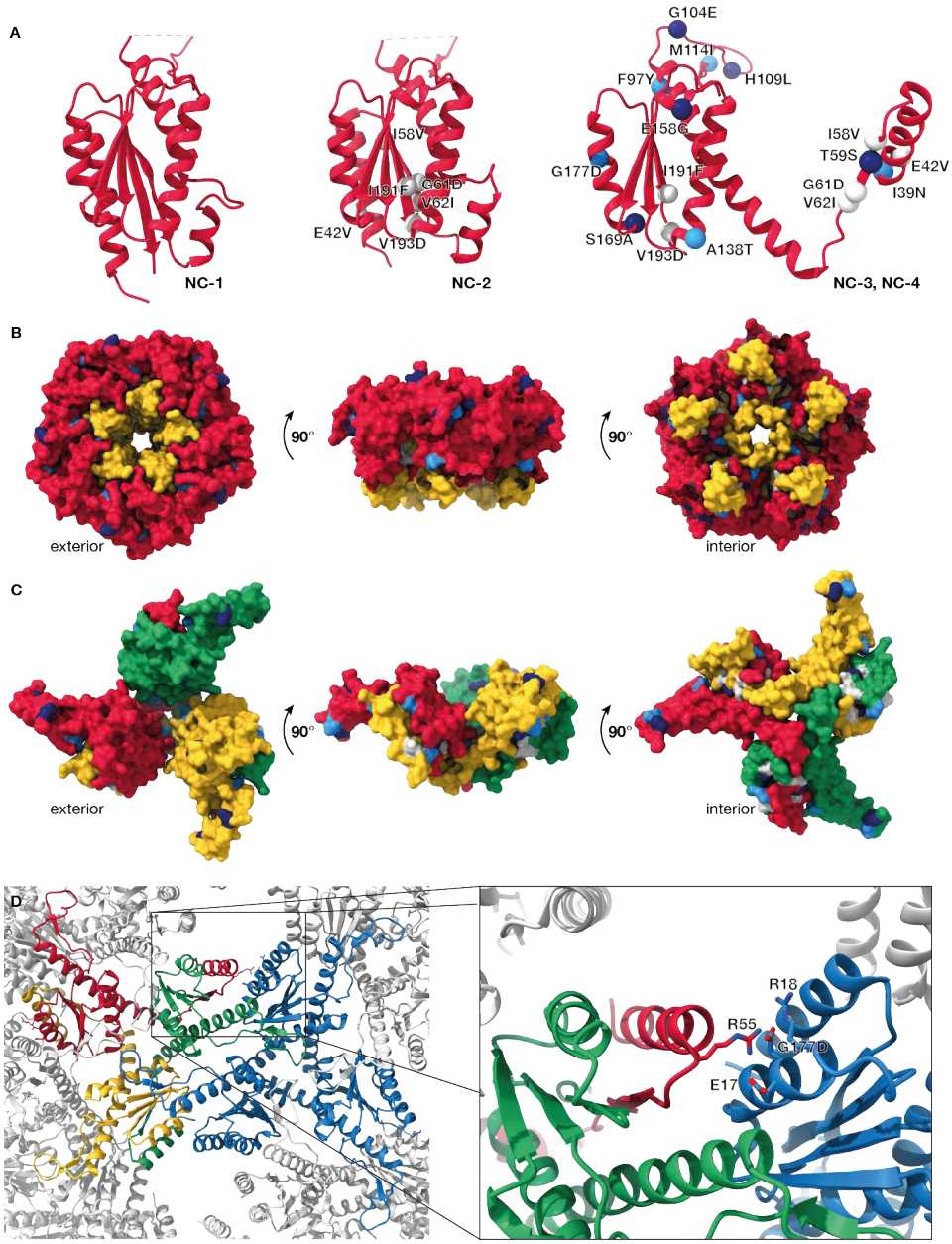
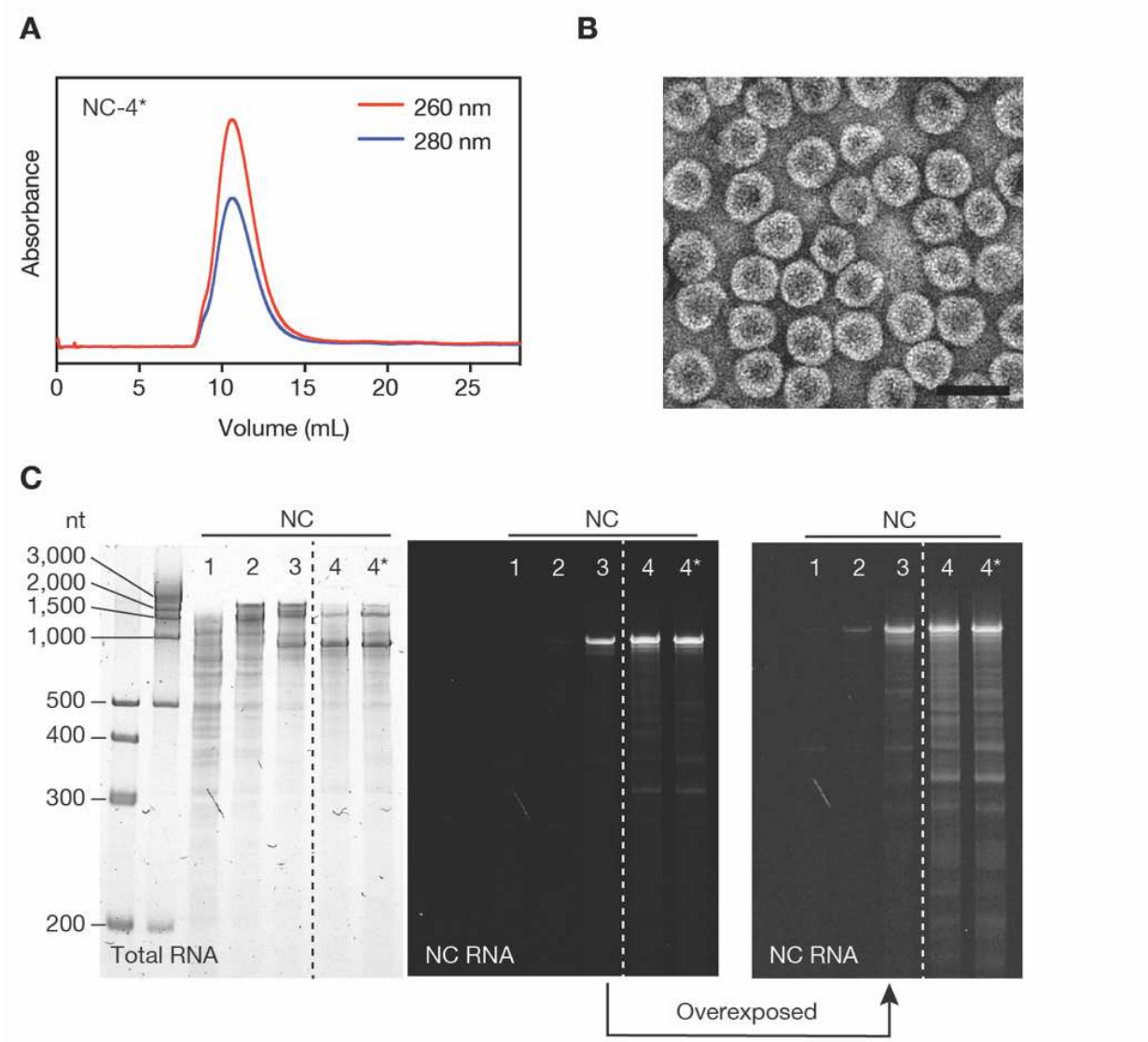


Figure S9: Mutations from NC-1 to NC-4

(**A**) Subunits of NC-1, NC-2, and NC-4 are shown with mutations indicated as spheres and color coded according to their order of appearance (NC-2, white; NC-3, light blue; NC-4, dark blue). (**B**) Surface representation of a pentamer excised from the NC-4 structure shown in Fig. 2, which comprises the N-terminal fragments of the yellow subunits (residues 1–74) and the C-terminal fragments of the red subunits (76–197). Because the penton-hexon interfaces contain relatively few mutations, the original AaLS interfaces are largely preserved upon expansion from a T=1 to a T=4 structure. (**C**) Surface representation of an NC-4 trimer. The mutations introduced into NC-2 (white) are located at the interfaces between the domain-swapped subunits, whereas subsequent mutations (light and dark blue) are located mainly on the interior and exterior capsid surfaces and the inter-trimer interfaces. (**D**) The G177D mutation, which first appeared in NC-3, may stabilize trimer-trimer interactions in the expanded structures. The aspartate side chain inserts into a network of charged residues and likely forms a salt bridge with Arg55 on a neighboring trimer.



878

879 **Fig. S10. Reversion of the K5R mutation in NC-4.**

880 (**A**) Size-exclusion chromatograms of purified, re-injected nucleocapsid (column: Superose 6 increase
881 10/300 GL). (**B**) Transmission electron micrograph of purified R5K NC-4 (NC-4*). Scale bar: 50 nm. (**C**)
882 RNA was extracted from each nucleocapsid generation and equal amounts were loaded onto a denaturing
883 PAGE (5%) gel. Total RNA was stained with GelRed, NC RNA was visualized with DFHBI-1T. The
884 DFHBI-1T-stained gel on the right was overexposed to better visualize faint bands. The dashed line
885 indicates two non-concurrent portions of the same gel image.

886

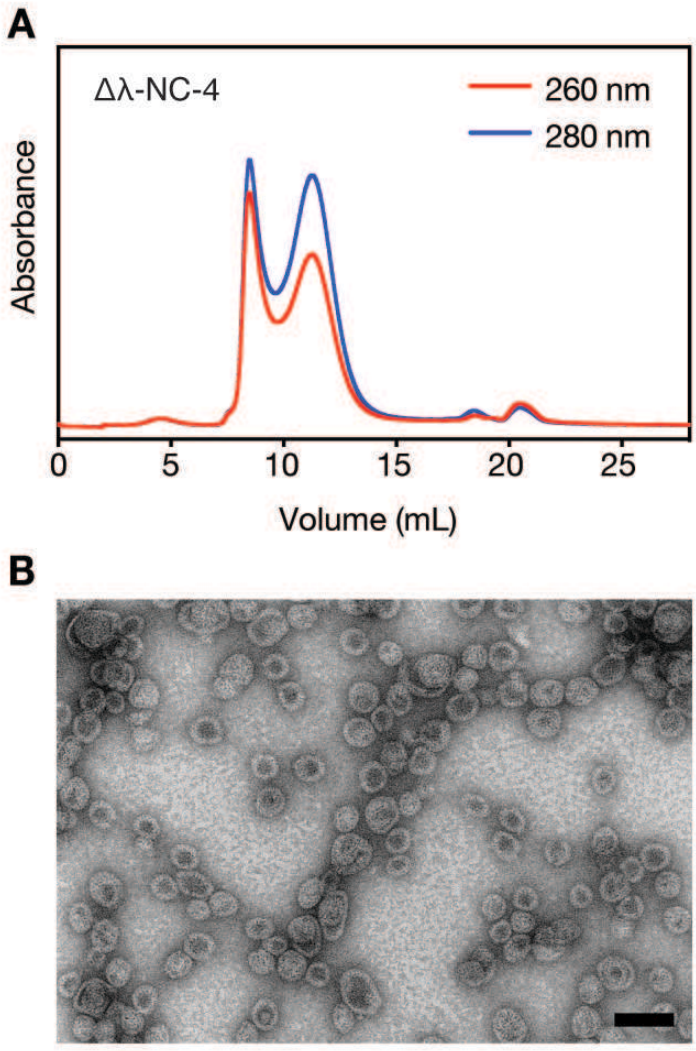
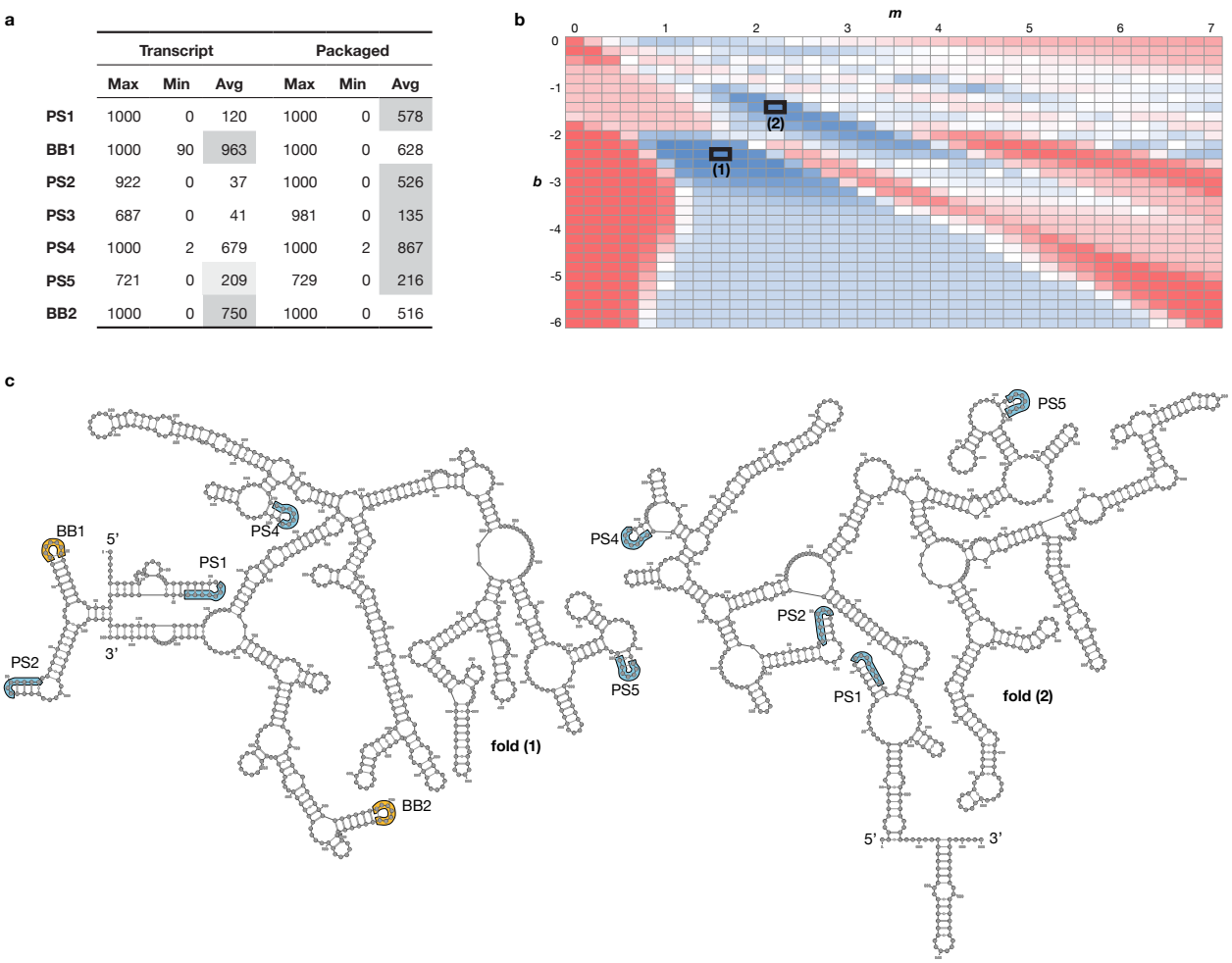


Fig. S11. Removal of the RNA-binding motif in NC-4.

(**A**) Size-exclusion chromatograms of purified, re-injected NC-4 capsid lacking the λN+ peptide and the (GlyAla)-linker (residues 1–30; Δλ-NC-4) (column: Superose 6 increase 10/300 GL). (**B**) Transmission electron micrograph of purified Δλ-NC-4. Scale bar: 50 nm.



	Transcript			Packaged		
	Max	Min	Avg	Max	Min	Avg
PS1	1000	0	120	1000	0	578
BB1	1000	90	963	1000	0	628
PS2	922	0	37	1000	0	526
PS3	687	0	41	981	0	135
PS4	1000	2	679	1000	2	867
PS5	721	0	209	729	0	216
BB2	1000	0	750	1000	0	516

Fig. S12. Secondary structure prediction based on XRF data for NC-3.

(**A**) The minimum, maximum and average number of times a given SL occurs in an ensemble of 1000 sample folds, generated using a modified version of the S-fold algorithm that includes the XRF data via a scaling factor *m* and an offset *b*, was computed for each of 1116 (*m*,*b*) combinations (see panel B). The seven potential packaging signals listed occur in at least 50% of the 1000 sample folds for at least 50 of the 1116 (*m*,*b*) combinations. Stem-loops for which the average number increases for packaged mRNA compared with the free transcript, or vice versa, are highlighted; if the lower value is within 85%, it is highlighted in lighter shade. (**B**) Stem-loops with entries highlighted in packaged RNA in A are collectively optimized via a cost function given by the sum of the normalized (by their maximal number of occurrence) frequencies for these SLs in an ensemble of 1000 sample folds for each (*m*,*b*) combination. This identifies (*m*,*b*) values for which their occurrence is locally maximally aligned with the trend in the tables in A. The (*m*,*b*) combinations for which the cost function is maximal are: *m*=1.6, *b*=-2.4 (77.6%, labelled 1) and *m*=2.2, *b*=-1.4 (75.7%, labelled 2). (**C**) The predicted folds corresponding to these (*m*,*b*) combinations, represented as cartoons in Figure 4C, are shown with their full sequence. The maximum ladder distances are 98 for fold (1), and 102 for fold (2).

a

	Transcript			Packaged		
	Max	Min	Avg	Max	Min	Avg
PS1	429	0	16	1000	0	460
BB1	1000	11	914	1000	794	990
PS2	900	0	80	1000	0	376
PS3	585	0	31	605	0	58
PS4	1000	1	691	1000	0	609
PS5	731	0	334	697	0	82
BB2	1000	0	844	1000	0	336

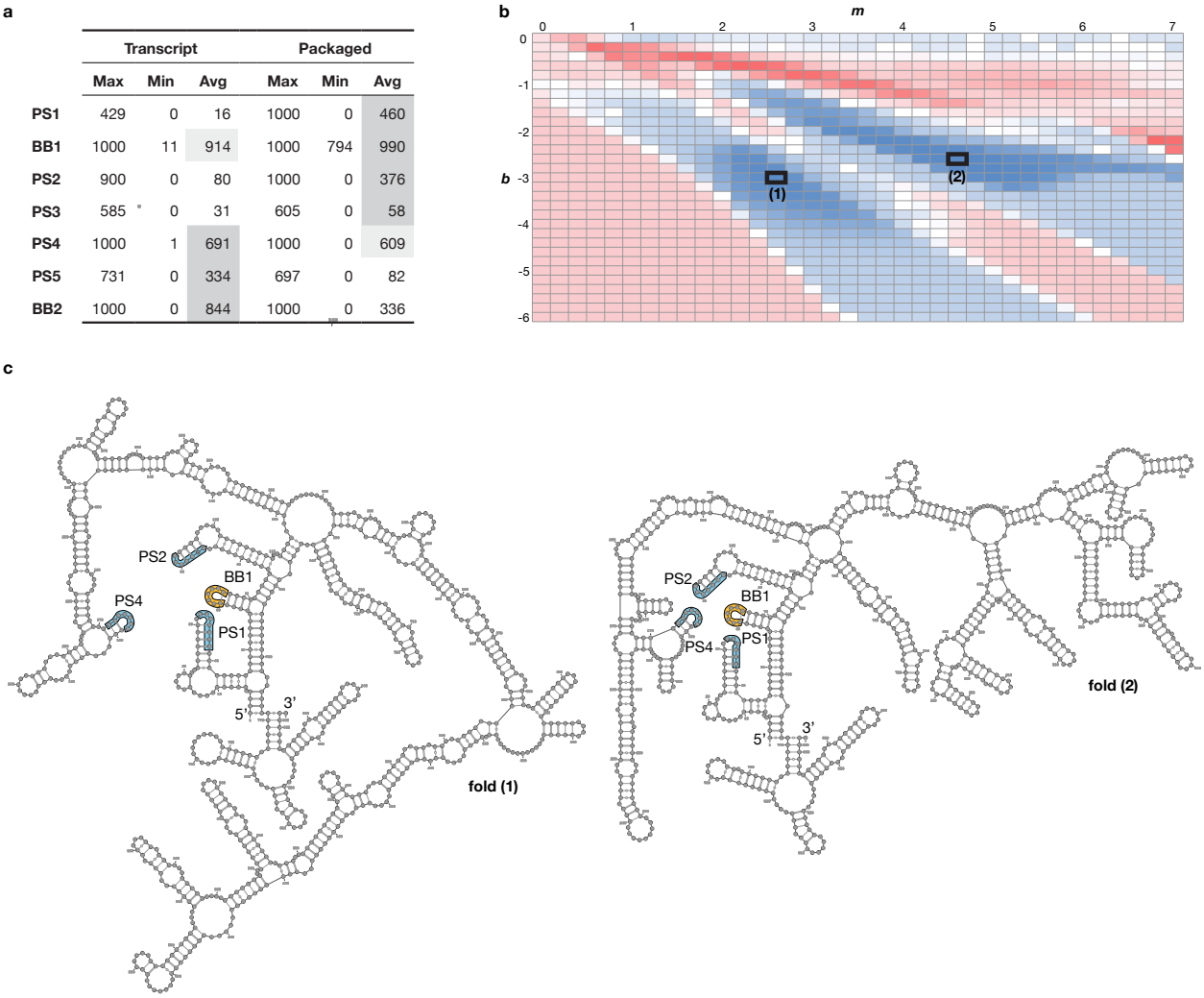


Fig. S13. Secondary structure prediction based on XRF data for NC-4.

(**A**) The minimum, maximum and average number of times a given SL occurs in an ensemble of 1000 sample folds, generated using a modified version of the S-fold algorithm that includes the XRF data via a scaling factor *m* and an offset *b*, was computed for each of 1116 (*m*,*b*) combinations (see panel B). The seven potential packaging signals listed occur in at least 50% of the 1000 sample folds for at least 50 of the 1116 (*m*,*b*) combinations. Stem-loops for which the average number increases for packaged mRNA compared with free transcript, or vice versa, are highlighted; if the lower value is within 85%, it is highlighted in lighter shade. (**B**) Stem-loops with entries highlighted in packaged RNA in A are collectively optimized via a cost function given by the sum of the normalized (by their maximal number of occurrence) frequencies for these SLs in an ensemble of 1000 sample folds for each (*m*,*b*) combination. This identifies (*m*,*b*) values for which their occurrence is locally maximally aligned with the trend in the tables in (A). The (*m*,*b*) combinations for which the cost function is maximal are: *m*=4.6, *b*=-2.6 (79.9%, labelled 1) and *m*=2.6, *b*=-3.0 (79.0%, labelled 2). (**C**) The predicted folds corresponding to these (*m*,*b*) combinations, represented as cartoons in Figure 4D, are shown with their full sequence. The maximum ladder distances are 125 for fold (1), and 105 for fold (2).

	NC-1 120-mer	NC-1 180-mer	NC-2	NC-3	NC-4
EMDB map entry	11631	11632	11633	11634	11635
PDB coordinate entry	7A4F	7A4G	7A4H	7A4I	7A4J
Data Collection and reconstruction					
Microscope model	FEI Titan Krios	FEI Titan Krios	FEI Titan Krios	FEI Tecnai F20	FEI Titan Krios
Detector model	Falcon III	Falcon III	Falcon III	Falcon II	Falcon III
# of Micrographs collected	1481	1481	848	134	1080
Magnification	130 000x	130 000x	130 000x	62 000x	130 000x
Voltage (kV)	300	300	300	200	300
Electron dose (e-/A2)	60	60	60	40	60
Pixel Size (Å)	1.1	1.1	1.1	1.8	1.1
Defocus range (μm)	-0.8 to -2.6	-0.8 to -2.6	-0.8 to -2.6	-1.8 to -3.3	-0.8 to -2.6
Symmetry imposed	T	T	T	I1	I1
# of Micrographs used	1 053	1 053	509	111	1 067
Initial particle images	129 954	129 954	29 998	32 411	69 723
Final particle images	5 257	5 226	4 037	3815	15 392
Resolution (Å) (at FSC = 0.143	3.50	4.20	4.47	7.04	3.04
Map sharpening B-factor (Å2)	-31	-52	-141	-550	-66
Model building					
Starting model	1hqk	1hqk	1hqk	7a4j	1hqk
Composition					
Chains	120	180	180	240	240
Atoms	141432	205680	204396	290520	289560
Protein residues	18540	27120	26904	37860	37860
Water	0	0	0	0	0
Ligands	0	0	0	0	0
Bonds (RMSD)					
Length (Å) (# > 4σ)	0.002 (0)	0.002 (0)	0.002 (0)	0.002 (0)	0.002 (0)
Angles (°) (# > 4σ)	0.430 (12)	0.383 (34)	0.396 (0)	0.466 (189)	0.402 (0)
MolProbity score	1.83	1.46	1.67	1.86	1.58
Clash score	6.30	6.41	6.06	5.65	3.77
Ramachandran plot (%)					
Outliers	0	0	0	0	0
Allowed	2.36	2.27	1.81	2.57	1.82
Favored	97.64	97.73	98.19	97.43	98.18
Ramachandran plot Z-score					
whole	1.60 (0.06)	2.89 (0.06)	3.99 (0.05)	1.36 (0.04)	0.40 (0.04)
helix	2.31 (0.05)	2.64 (0.05)	3.81 (0.04)	2.05 (0.04)	1.02 (0.04)
sheet	0.45 (0.08)	1.44 (0.07)	2.32 (0.07)	1.59 (0.07)	1.39 (0.07)
loop	0.02 (0.10)	0.95 (0.09)	0.73 (0.08)	1.04 (0.05)	1.28 (0.05)
Rotamer outliers (%)	3.40	1.14	2.76	3.80	3.50
Cβ outliers (%)	0	0	0	0	0
Peptide plane (%)					
Cis proline/general	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Twisted proline/general	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
CaBLAM outliers (%)	0.07	0.93	1.1	1.14	0.98
ADP (B-factors)					
Iso/Aniso (#)	141432/0	205680/0	204396/0	290520/0	289560/0
min/max/mean	45.97/167.52/91.60	14.53/153.78/58.76	34.68/253.39/95.22	96.48/410.28/208.34	41.87/138.99/82.97
Occupancy					
Mean	1	1	1	1	1
occ = 1 (%)	100	100	100	100	100
Box					
Lengths (Å)	248.88, 250.25, 250.25	301.12, 301.12, 303.88	303.88, 302.50, 301.12	333.00, 333.00, 333.00	324.50, 324.50, 324.5
Angles (°)	90.00, 90.00, 90.00	90.00, 90.00, 90.00	90.00, 90.00, 90.00	90.00, 90.00, 90.00	90.00, 90.00, 90.00
Model vs. Data					
CC (mask)	0.84	0.78	0.73	0.8	0.81
CC (box)	0.78	0.72	0.69	0.82	0.69
CC (peaks)	0.72	0.65	0.58	0.69	0.64
CC (volume)	0.83	0.75	0.72	0.8	0.8
Resolution range (Å)	3.2-4.4	3.8-5.8	4.1-6.1	6.4-8.5	3.0-3.6

930

931 **Table S1. Cryo-EM data**

URxRxRR			
Position	**NC-3**	**NC-4**	
32	UAGAGGG	UAGAGGG	**PS1**
60	UGAAGAA	UGAAGAA	**BB1**
116	UAAGAAG	UAAGAAG	
133	UAUGGGA	UAUGGGA	
135	UGGGAAA	UGGGAAA	**PS3**
172	UGAGAAA	UGAGAAA	**PS4**
188	UGGAAAG	UGGAAAG	
294	UACGCAA	UACGUAA	
383	UGGGAAG	UGGGAAG	**PS5**
420	UAUACAA	UAUACAA	
482	UACGAAG	UACGAAG	
564	UGGAGGG	UGGAGGG	
581	UGCAUAG	UGCAUAG	
604	UGAAGAA	UGGAGAA	
641	UGGGAAA	UGGGAAA	
676	UAAAGAG	UAAAGAG	
734	UAAGCGG	UAAGCGG	
768	UGAAGAA	UGAAGAA	**BB2**

URxRxxx			
Position	**NC-3**	**NC-4**	
79	UAGAGUG	UAGAGUG	
84	UGUGGGC	UGUGGGC	**PS2**
86	UGGGCUC	UGGGCUC	
127	UAUACAU	UAUACAU	
129	UACAUAU	UACAUAU	
205	UGGAGCU	UGGAGCU	
211	UGGAGCA	UGGAGCA	
220	UGCAAUG	UGCAAUG	
245	UAUAACG	UAUAACG	
256	UGUAGUU	UGUAGUU	
288	UAGAACU	UAGAACU	
298	CAAACCU	UAAACCU	
322	UACAGCU	UACAGCU	
336	UGGAACA	UGGAACA	
406	UGAAAUG	UGAAAUG	
422	UACAAGU	UACAAGU	
474	UUGAAAUC	UUGAAAUC	
490	UAAACUA	UAAACUA	
631	UCCAGGC	UCCAGGC	
658	UGCGGAU	UGCGGAU	
664	UGAACUG	UGAACUG	

Table S2. BoxB-like sequence motifs within the NC-3 and NC-4 genomes.

Genome positions of nucleotide strings fulfilling the search motif together with the color-coded reactivities (black, green, orange and red from low to high) in NC-3 and NC-4. The 13 sequences that show sufficiently low reactivity to potentially act as packaging signals are highlighted in grey. Of these, seven motifs occur in a stem loop in over half of the sample folds for at least 50 *m*,*b* combinations tested. Two are the BoxBr tags (BB1, BB2), introduced by design, and the other five potential packaging signals are designated PS1-PS5 in the order they appear in the sequence.

	NC-3						NC-4					
	Transcript			*In situ*			Transcript			*In situ*		
Primer	A	B		A	B		A	B		A	B	
12	0.964		B	0.718		B	0.927		B	0.899		
	0.940	0.930	C	0.774	0.869	C	0.963	0.950	C	0.893	0.976	
	A	B		A	B		A	B		A	B	
11	0.932		B	0.919		B	0.931		B	0.975		
	0.919	0.981	C	0.983	0.917	C	0.715	0.725	C	0.949	0.954	
	A	B		A	B		A	B		A	B	
10	0.866		B	0.865		B	0.954		B	0.952		
	0.893	0.951	C	0.927	0.918	C	0.841	0.768	C	0.957	0.954	

Table S3. Pairwise Pearson correlation coefficients (PCCs) for normalized replicates at 50 ms exposure.

PCCs for triplicate primer extensions, analyzed by primer region and RNA. A, B, and C represent the individual replicates. Lower values imply greater variability in the respective mRNA segment.