LETTER

doi:10.1038/nature25157

Evolution of a designed protein assembly encapsulating its own RNA genome

Gabriel L. Butterfield[1,2,3]*, Marc J. Lajoie[1,2]*, Heather H. Gustafson[4,5], Drew L. Sellers[4,5,6], Una Nattermann[1,2,7], Daniel Ellis[1,2,3], Jacob B. Bale[1,2,3], Sharon Ke[4], Garreck H. Lenz[8], Angelica Yehdego[9], Rashmi Ravichandran[1,2], Suzie H. Pun[4,5], Neil P. King[1,2] & David Baker[1,2,10]

The challenges of evolution in a complex biochemical environment, coupling genotype to phenotype and protecting the genetic material, are solved elegantly in biological systems by the encapsulation of nucleic acids. In the simplest examples, viruses use capsids to surround their genomes. Although these naturally occurring systems have been modified to change their tropism[1] and to display proteins or peptides[2–4], billions of years of evolution have favoured efficiency at the expense of modularity, making viral capsids difficult to engineer. Synthetic systems composed of non-viral proteins could provide a 'blank slate' to evolve desired properties for drug delivery and other biomedical applications, while avoiding the safety risks and engineering challenges associated with viruses. Here we create synthetic nucleocapsids, which are computationally designed icosahedral protein assemblies[5,6] with positively charged inner surfaces that can package their own full-length mRNA genomes. We explore the ability of these nucleocapsids to evolve virus-like properties by generating diversified populations using *Escherichia coli* as an expression host. Several generations of evolution resulted in markedly improved genome packaging (more than 133-fold), stability in blood (from less than 3.7% to 71% of packaged RNA protected after 6 hours of treatment), and *in vivo* circulation time (from less than 5 minutes to approximately 4.5 hours). The resulting synthetic nucleocapsids package one full-length RNA genome for every 11 icosahedral assemblies, similar to the best recombinant adeno-associated virus vectors[7,8]. Our results show that there are simple evolutionary paths through which protein assemblies can acquire virus-like genome packaging and protection. Considerable effort has been directed at 'top-down' modification of viruses to be safe and effective for drug delivery and vaccine applications[1,9,10]; the ability to design synthetic nanomaterials computationally and to optimize them through evolution now enables a complementary 'bottom-up' approach with considerable advantages in programmability and control.

What minimal features are required for a synthetic system to encapsulate its own genome and to evolve biological functionality similar to viruses? In the nearly 40 years since the first high-resolution structure of an icosahedral virus[11], the structures and functions of a wide array of viral capsids have been characterized. This has inspired efforts to reengineer naturally occurring protein containers[12] and to design new polypeptides[13] to package biological molecules. For example, the naturally occurring, non-viral protein container lumazine synthase was evolved in *E. coli* to sequester a toxic protein[14]. However, there have been no reports of non-viral containers that can encapsulate their own genomes and evolve in complex biochemical environments outside of cells.

We recently reported the design, with atomic-level accuracy, of two-component, 120-subunit icosahedral protein assemblies with internal volumes large enough to package biological macromolecules[5]. These highly stable and engineerable assemblies[5,6] in principle could be redesigned to package their own genomes: bicistronic mRNAs encoding the two protein subunits. We investigated this possibility by modifying two assemblies with accessible protein termini and no large pores, I53-47 and I53-50[5], either by introducing positively charged residues on their interior surfaces (I53-47-v1 and I53-50-v1; Fig. 1a and Extended Data Table 1a) or by genetically fusing the Tat RNA-binding peptide from bovine immunodeficiency virus[15] to the interior-facing C terminus of one subunit (I53-50-Btat and I53-47-Btat). After expression and intracellular assembly in *E. coli* (Fig. 1b), intact protein assemblies were purified from cell lysates using immobilized metal affinity chromatography (IMAC) and size exclusion chromatography (SEC). The assemblies eluted as a single peak at the same retention volume as the original design[5] (Extended Data Fig. 1), and intact particles were observed by negative-stain transmission electron microscopy (Fig. 1c, Extended Data Fig. 1a). After purification, the assemblies were incubated with RNase A for 10 min at 20 °C to degrade any RNA not protected inside the synthetic capsid-like proteins. Nucleic acid and protein co-migrated on native agarose gels (Fig. 1d, e, Extended Fig. 1b, c), suggesting that the protected nucleic acid was encapsulated in the protein assembly. Nucleic acid extraction followed by reverse transcription quantitative PCR (RT–qPCR) and Sanger sequencing confirmed that full-length RNA genomes were packaged and protected from RNase by I53-50-v1 and I53-50-Btat but not by the original I53-50 design (Fig. 1f); all versions of I53-47 could package their genomes (Extended Data Fig. 1d). In all cases, RT–PCR products were only obtained upon addition of reverse transcriptase, indicating that the protected nucleic acids were RNA and not DNA. We refer to these designed RNA–protein complexes as synthetic nucleocapsids.

To investigate whether synthetic nucleocapsids can evolve, we generated combinatorial libraries of variants and selected for improved genome packaging and fitness against nuclease challenge. Nine positions on the interior surfaces of I53-50-v1 and I53-50-Btat were mutated to positive, negative, or uncharged polar amino acids (Supplementary Table 1) to produce variants with a wide range of interior charge distributions. We performed three rounds of selection comprising expression, purification, RNase challenge, RNA recovery, and re-cloning (Fig. 2a). The RNA recovered from the selected population after each round was reverse-transcribed and sequenced on an Illumina MiSeq. The net interior charge of the evolved population converged to narrow distributions around 388 ± 87 (mean $\pm$ s.d.

[1]Institute for Protein Design, University of Washington, Seattle, Washington 98195, USA. [2]Department of Biochemistry, University of Washington, Seattle, Washington 98195, USA. [3]Graduate Program in Molecular and Cellular Biology, University of Washington, Seattle, Washington 98195, USA. [4]Department of Bioengineering, University of Washington, Seattle, Washington 98195, USA. [5]Molecular Engineering and Sciences Institute, University of Washington, Seattle, Washington 98195, USA. [6]Institute for Stem Cell and Regenerative Medicine, University of Washington, Seattle, Washington 98109, USA. [7]Graduate Program in Biological Physics, Structure & Design, University of Washington, Seattle, Washington 98195, USA. [8]College of Arts & Sciences, University of Washington, Seattle, Washington 98195, USA. [9]School of Public Health, University of Washington, Seattle, Washington 98195, USA. [10]Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA.
*These authors contributed equally to this work.

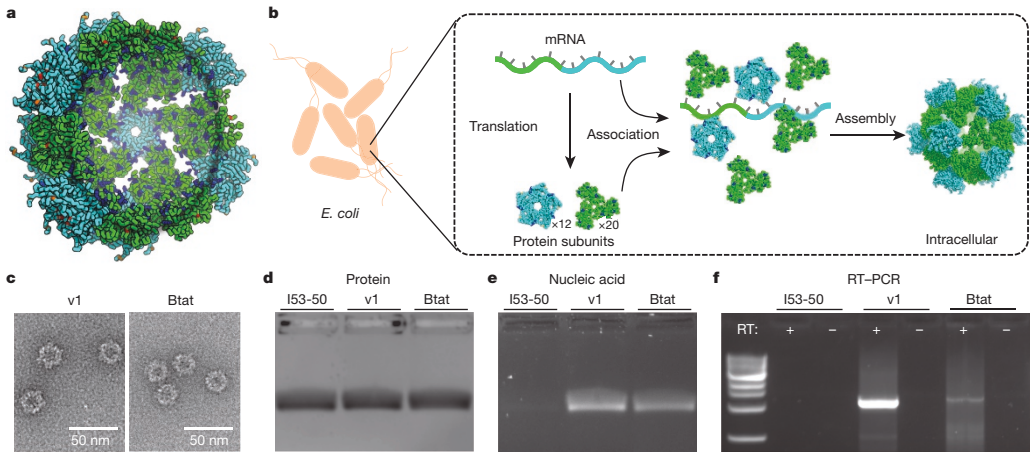


Figure 1 | Biochemical characterization of synthetic nucleocapsids. **a**, Design model of I53-50-v1. Trimeric subunits are coloured green and pentameric subunits are coloured cyan. Mutations with respect to the original I53-50 protein assembly[7] are coloured blue (increases in positive charge and/or decreases in negative charge (for example, E→N, N→K, E→K)), orange (no change in charge (for example, E→D, N→T, K→R)), or red (decreases in positive charge and/or increases in negative charge (for example, N→E, K→N, K→E)). **b**, Synthetic nucleocapsids encapsulate their own mRNA genomes while assembling into icosahedral capsids inside *E. coli* cells. **c**, Negative-stain electron micrographs of I53-50-v1 (positively charged interior) and I53-50-Btat (RNA binding Tat peptide from bovine immunodeficiency virus). Micrographs shown are representative of the entire sample tested on between one and three different grids, each at a different concentration. **d**, **e**, Synthetic nucleocapsids were purified, treated with RNase A, and electrophoresed on non-denaturing 1% agarose gels then stained with Coomassie (protein, **d**) and SYBR gold (nucleic acid, **e**). Nucleic acids co-migrated with capsid proteins for I53-50-v1 and I53-50-Btat, but not for the original I53-50. **f**, Full-length synthetic nucleocapsid genomes were recovered from each sample by RT–PCR. The left-most lane is NEB 1-kb DNA ladder. Plus and minus symbols indicate PCR performed on templates prepared with and without reverse transcriptase (RT), respectively, confirming that I53-50-v1 and I53-50-Btat package their own full-length RNA genomes. This procedure is part of our standard quality control for synthetic nucleocapsids and has been performed reproducibly more than 10 times.

of the population) in the absence of Btat and 662 ± 91 (480 of which are from 60 copies of Btat) in the presence of Btat (Fig. 2b). A total of 1,170 different variants exhibited higher enrichment than I53-50-v1 (Fig. 2c); there are evidently many solutions to the genome-packaging problem. The presence or absence of the positively charged Btat peptide influenced the identities of beneficial mutations: all except two of the lysine residues were beneficial in the absence of Btat (Fig. 2d), whereas most lysine residues were disfavoured in the presence of Btat (Fig. 2e). We combined the substitutions from one of the most highly enriched variants from the library lacking Btat (Fig. 2c; trimeric subunit: K178N, K183N, E189K; pentameric subunit: K123N, H125K) with the most enriched substitution from a separate library of mutants in the trimer-pentamer interface (pentameric subunit: E24F; Supplementary Fig. 2 and Supplementary Table 1a) to produce I53-50-v2, which exhibited improved genome packaging efficiency as assessed by RT–qPCR (Extended Data Fig. 2a). The net interior charge did not change between I53-50-v1 and I53-50-v2—the improved genome packaging and protection results from reconfiguration of the position of the charges (Fig. 2f). I53-50-v2 outperformed the best variants from the I53-50-Btat library (Extended Data Fig. 2a), so we focused on I53-50-v2 for subsequent evolution experiments.

The ability to evolve the nucleocapsids enabled comprehensive mapping of how each residue affects the fitness of the 2.5-megadalton complex comprising 22,920 amino acids and 1,370 RNA bases. We produced a deep mutational scanning library[16,17] of I53-50-v2 with every residue in each protein subunit substituted with each of the 20 amino acids, and performed two consecutive rounds of selection with two biological replicates. Selection in the first round was performed at room temperature with 10 μg ml^{-1} RNase A for 10 min to deplete non-assembling variants from the population, and selection in the second round was at 37 °C for 1 h with either 10 μg ml^{-1} RNase A or heparinized mouse whole blood. Each biological replicate of the naive, round 1 and round 2 populations was sequenced on an Illumina MiSeq, and enrichment values were calculated from the fraction of the population corresponding to each variant before and after selection (Fig. 3a, b; 7,156 out of the possible 7,240 single mutants were observed with at least 10 counts in the pre-selection population). The enrichments of individual mutations were correlated between the RNase A and mouse whole blood selections (Fig. 3c), suggesting that similar mechanisms underlie the increased genome protection in both cases.

Evaluating the enrichment values in the context of the I53-50 design model (Fig. 3d–g) provides insight into the features important for genome encapsulation and protection. I53-50 is composed of 20 trimers and 12 pentamers; the hydrophobic protein cores, intra-oligomer interfaces, and designed inter-oligomer interface were conserved (Fig. 3d, Supplementary Fig. 3), probably because proteins bearing mutations that disrupt the assembly fail to protect their genomes and are removed from the population. Strong selective pressure also operated on the electrostatics of the surface lining the pore between trimeric subunits of I53-50-v2; all highly depleted residues were lysines or arginines, whereas the nearby glutamate (residue E4) was highly conserved (Fig. 3f, g). Lysine removal around the pore also occurred in the earlier transition from I53-50-v1 to I53-50-v2 (K179N in the trimer and K124N in the pentamer; Fig. 2d, Supplementary Fig. 4). Positively charged residues near the pores may compromise genome protection either by promoting protrusion of the encapsulated RNA from the interior of the icosahedral assembly, thereby rendering it susceptible to RNases, or by destabilizing the assembly through electrostatic repulsion between trimeric subunits. To test whether several of the most enriched mutations could be combined to produce a synthetic nucleocapsid with superior fitness, a combinatorial library was constructed containing charged and uncharged polar residues at positions where positively charged residues were deleterious in the deep mutational scanning data (trimeric subunit: K2, K8, K9, K11 and K61). After selection in 10 μg ml^{-1} RNase A at 37 °C for 1 h, the six most enriched variants were tested individually to evaluate their improvements over I53-50-v2 (Extended Data Fig. 2b, c). The one best protected under these conditions was designated I53-50-v3 (trimeric subunit: K2T, K9R, K11T and K61D). The failure of an assembly-defective variant to protect its genome (I53-50-v3-KO; trimeric subunit: V29R, pentameric subunit: A38R; Supplementary Fig. 5) confirmed that encapsulation was required for RNA protection.

We next investigated whether synthetic nucleocapsids can evolve with part of their lifecycle inside an animal. As long circulation times

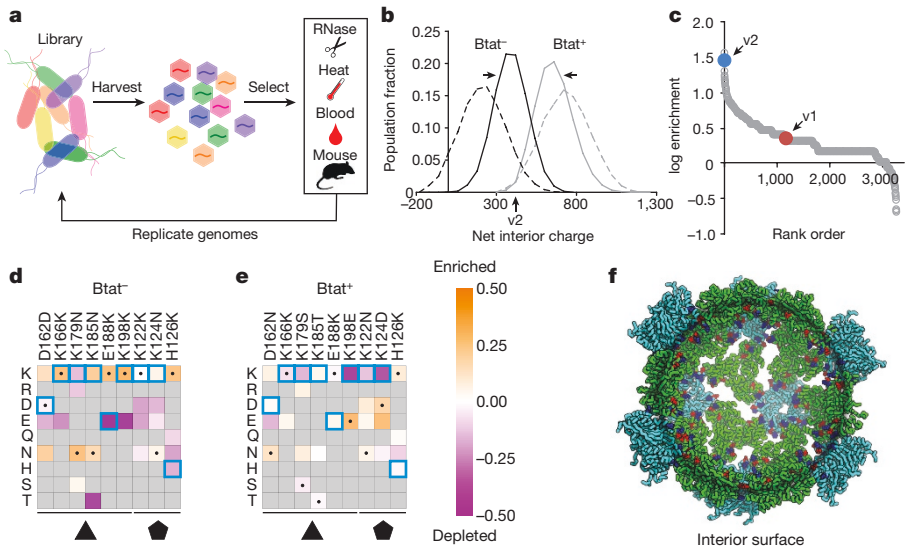


Figure 2 | Evolution of RNA packaging. a, A library of plasmids encoding synthetic nucleocapsid variants is transformed into *E. coli*. Each cell in the population produces a unique synthetic nucleocapsid variant. Nucleocapsids are purified *en masse* from cell lysates and challenged (for example, RNase, heat, blood and mouse circulation). The capsid-protected mRNA is then recovered and amplified using RT–qPCR, re-cloned into a plasmid library, and transformed into *E. coli* for another generation. **b**–**f**, Combinatorial libraries targeting nine residues on the interior surface of I53-50 (Supplementary Table 1) were used to investigate how the interior surface charge affects RNA packaging in the presence or absence of a positively charged RNA binding peptide (Btat). Three rounds of evolution were performed with two independent biological replicates. **b**, The evolved populations converged towards narrow distributions of interior net charge: Btat$^-$ library from 215 ± 114 (mean $\pm$ s.d.) to 388 ± 87, Btat$^+$ library from 733 ± 119 to 662 ± 91. The net interior charge of each variant was calculated from its sequence by summing the positive and negative residues on the interior surface. Black lines are without Btat and grey lines are with Btat; dashed lines are naive populations and solid lines are round three selected populations. These results represent the combined population distribution of two independent evolutionary trajectories. **c**, Rank order list of variants observed in both biological replicates; 1,170 unique variants outperformed I53-50-v1. I53-50-v2 was created based on the second most highly enriched variant from the Btat- library. **d**, **e**, Heat map of log enrichments for each mutation explored in the combinatorial surface charge optimization library. All except two of the lysine residues were beneficial in the absence of the positively charged Btat, whereas most lysine residues were disfavoured in the presence of Btat. Purple or orange indicate mutations that were depleted or enriched in the selected population, respectively. Blue squares and black dots indicate the I53-50-v1 starting sequence and the I53-50-v2 selected sequence, respectively. **f**, Design model of I53-50-v2. Although the net interior surface charge did not change from I53-50-v1 to I53-50-v2, the spatial configuration of charged residues impacted genome packaging efficiency (see Fig. 4a). Colouring is as described in Fig. 1a.

are desirable for *in vivo* applications such as drug delivery, we decided to focus on this property. We hypothesized that the histidine tag might mediate undesired interactions *in vivo*, so we created cleavable versions that were used for all subsequent experiments (see Methods). We produced two populations of synthetic nucleocapsids, one displaying hydrophilic 60-residue polypeptides of varying compositions intended to mimic viral glycosylation or PEGylation[18] (Supplementary Table 2) and another with 14 exterior surface positions combinatorially mutated to polar charged and uncharged amino acids (D, E, N, Q, K and R; Supplementary Table 1). We administered each population of nucleocapsids to mice by retro-orbital injection ($n = 5$ mice for hydrophilic peptides and 6 for surface positions), and evaluated the survival of each member of the population *in vivo* by blood draws from the tail vein at successive time points. From both libraries, several distinct sequences markedly improved circulation times. An optimal amino acid composition emerged in the hydrophilic peptide library (Extended Data Fig. 3a–c). Arbitrary polypeptides with similar amino acid composition (for example, 4.5 repeats of PETSPASTEPEGS or 4 repeats of PESTGAPGETSPEGS) increased the circulation time, whereas other polypeptides composed of different amino acids (for example, 12 repeats of ESESG) did not (Extended Data Fig. 3d, e). From the exterior surface library (Extended Data Fig. 4a, b), we isolated several variants exhibiting markedly enhanced circulation time compared to I53-50-v3 (Extended Data Fig. 4c, d) and found that most contained the substitution E67K in the pentameric subunit. We generated I53-50-v4 by incorporating E67K along with a set of other consensus mutations that were enriched in the selected population of synthetic nucleocapsids and may also contribute to increased expression and stability (Extended Data Table 1a; as the hydrophilic polypeptides reduced nucleocapsid yield, they were not included).

We next sought to determine what fraction of the I53-50-v4 synthetic nucleocapsids are filled, and with which RNAs. Negative-stain electron microscopy analysis of 15,119 particles showed that most I53-50-v4 nucleocapsids are more electron-dense—probably due to encapsulated nucleic acid—than the unfilled I53-50-v0 assemblies (Extended Data Fig. 5). Quantification of bulk RNA and protein indicated that there is approximately one nucleocapsid genome equivalent (1,433 nucleotides) of total RNA encapsulated per 6.6 I53-50-v1 capsids and 4.8 I53-50-v4 capsids (Extended Data Table 1b). RNA sequencing (RNA-seq) of I53-50-v4 capsids showed that approximately 74% of this RNA was derived from the nucleocapsid genome, suggesting one genome equivalent of nucleocapsid mRNA per 6.5 capsids (Fig. 4e, f). Independent quantification by RT–qPCR indicated one full-length genome per 11 capsids. The difference between these two estimates probably reflects the inclusion of genome fragments by RNA-seq (Extended Data Fig. 6). Although capsid genomes are modestly enriched and ribosomal RNA is depleted in nucleocapsids relative to cells (Fig. 4e, f), I53-50-v4 does not exhibit increased specificity for its genome relative to I53-50-v1 (Extended Data Fig. 7a). Instead, packaging correlates strongly with expression level (Extended Data Fig. 7b), accounting for the encapsulation of a modest amount of host cell RNA. The ability to package arbitrary RNA sequences combined with the ability to assemble *in vitro* from purified subunits[5] could make synthetic nucleocapsids the basis for a highly flexible platform for RNA delivery.

Like modern viruses, our evolved synthetic nucleocapsids exhibit genome packaging, nuclease protection, and sustained circulation *in vivo*. Each evolutionary step (Extended Data Table 1a and Extended Data Fig. 8) improved the particular property under selection without compromising gains from previous steps (Fig. 4). Negative-stain electron micrographs of I53-50-v1, I53-50-v2, I53-50-v3 and I53-50-v4

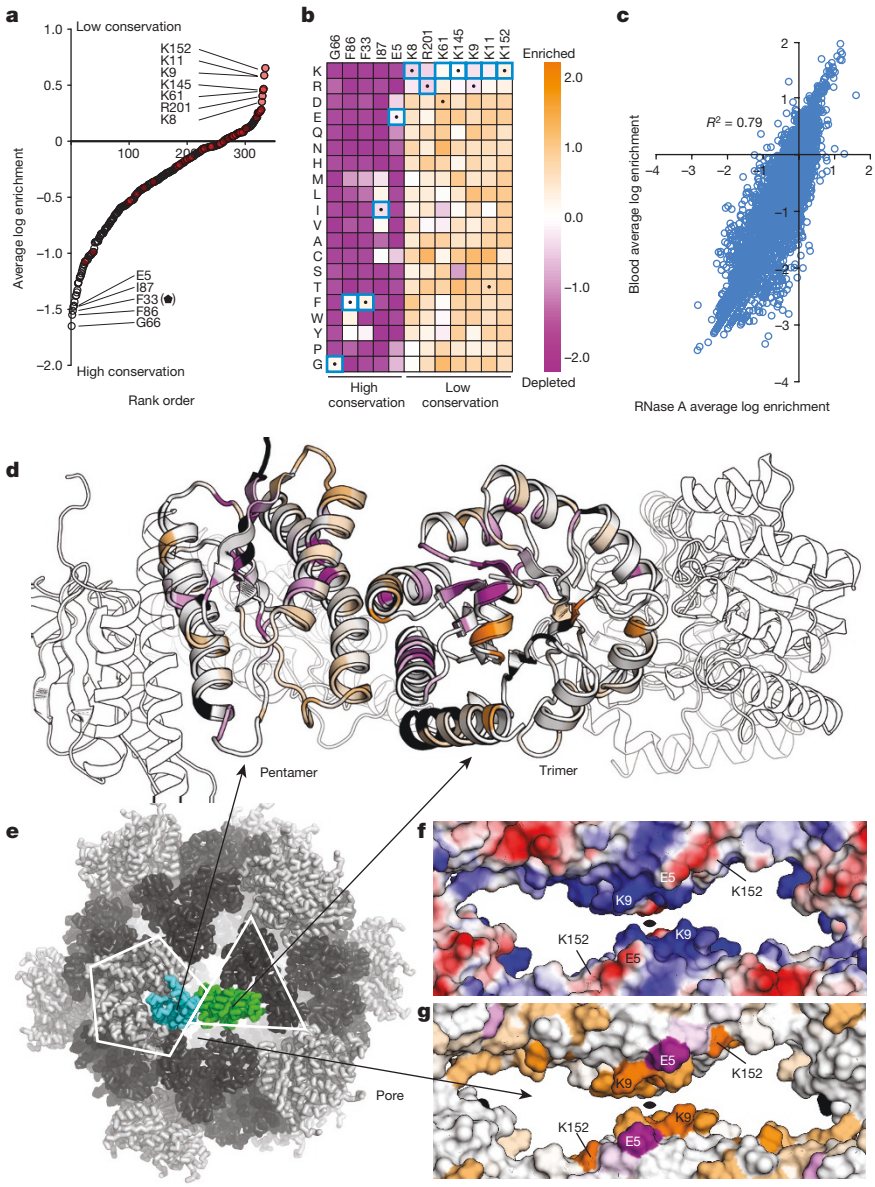


Figure 3 | Synthetic nucleocapsid fitness landscape. Deep mutational scanning of I53-50-v2 enables comprehensive mapping of how each residue affects nucleocapsid fitness. **a**, Average log enrichments for all 20 amino acids at each position in the 2.5-megadalton capsid revealed that many native lysine and arginine residues (red circles) favour being mutated to other amino acids. **b**, Heat map of log enrichments for all amino acids at the positions exhibiting the highest and lowest conservation. Purple or orange indicate mutations that were depleted or enriched in the selected population, respectively. Blue squares and black dots indicate the I53-50-v2 starting sequence and the I53-50-v3 selected sequence, respectively. **c**, Average log enrichment was highly correlated between the RNase A (10 μg ml^{-1} RNase A, 37 °C, 1 h) and heparinized mouse blood (37 °C, 1 h) selections, indicating that the beneficial mutations shared a common mechanism for improving nucleocapsid stability (RNase A versus blood: Pearson correlation $r = 0.79$; one biological replicate of each selection condition for the trimeric subunit and two independent biological replicates for the pentameric subunit). **d**, The core and interface residues of the capsid pentameric and trimeric subunits are more highly conserved than the surface residues. The colour spectrum in **d** and **g** represents the average log enrichment of all 20 amino acids at the indicated position and is rescaled relative to that in **c** for clarity (purple is conserved and orange is highly mutated; see Methods). **e**, I53-50 design model with pentameric subunit (cyan), trimeric subunit (green) and pore indicated. **f**, **g**, Surface electrostatics (−, red; +, blue) (**f**) and sequence conservation (**g**) show that lysine and arginine residues are highly depleted in the capsid pore during evolution (in particular, trimeric subunit residues K8, K9, K11, K61, K145 and K152). By contrast, the negatively charged E4 is highly conserved. These data suggest that positively charged residues in the capsid pore are deleterious for RNA packaging and protection.

showed that the functional improvements introduced by evolution did not compromise the designed icosahedral architecture (Supplementary Fig. 6), and dynamic light scattering indicated uniform populations of nucleocapsids around the expected size (radius = 13.5 nm; Extended Data Fig. 9). The I53-50-v1 design provided a starting point for evolution, inefficiently packaging its own full-length genome. Evolving the interior surface produced I53-50-v2, which packages roughly 1 RNA genome for every 14 capsids, rivalling the best recombinant adeno-associated viruses (rAAVs)[8,9] (Fig. 4d). Subsequently, evolving the capsid pore for improved stability resulted in I53-50-v3, which protects 44% of its RNA when challenged by RNase A (10 μg ml^{-1}, 37 °C, 6 h) and 82% of its RNA when challenged by mouse whole blood (37 °C, 6 h), whereas I53-50-v2 only protects 1.0% and 1.2%, respectively (Fig. 4a, b). Evolving the exterior surface of the capsid in circulation in live mice produced I53-50-v4, with a more than 54-fold increase in the circulation half-life—from less than 5 min to roughly 4.5 h—relative to I53-50-v3 (Fig. 4c). To further characterize the difference in behaviour between these two nucleocapsids, we determined the relative

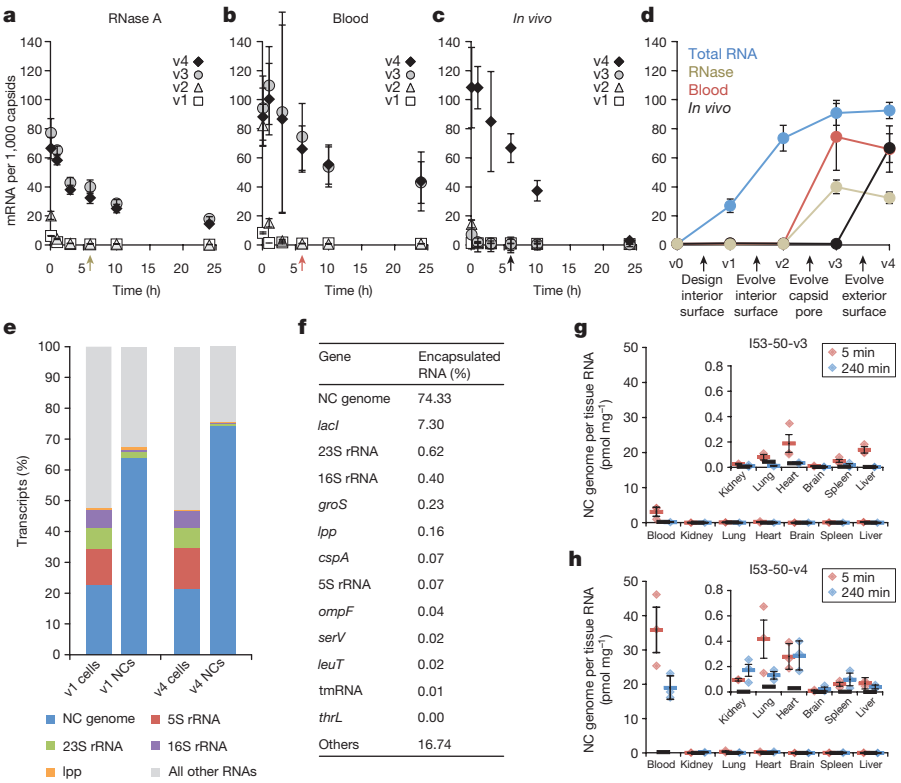


Figure 4 | Increased fitness of evolved synthetic nucleocapsids. Evolution markedly increases the property under selection without compromising previously evolved properties. **a**–**c**, Time courses of full-length RNA genomes per 1,000 capsids isolated after challenge with 10 μg ml^{-1} RNase A at 37 °C ($n = 3$ independent reactions) (**a**), heparinized mouse whole blood at 37 °C ($n = 3$ independent reactions) (**b**), and *in vivo* circulation of live mice ($n = 5$ biologically independent animals) (**c**). Error bars denote s.e.m. **d**, Summary of improved nucleocapsid properties, including total packaged RNA (10 μg ml^{-1} RNase A for 10 min at 20 °C to degrade non-encapsulated RNA, $n = 3$ independent reactions). The coloured arrows in **a**–**c** indicate the 6-h time point represented in the summary plot. Five synthetic nucleocapsids were tested: I53-50-v0 (original assembly that did not package its full length mRNA), I53-50-v1 (design with positive interior surface for packaging RNA), I53-50-v2 (evolution-optimized interior surface), I53-50-v3 (evolution-optimized residues lining the capsid pore), and I53-50-v4 (evolution-optimized exterior surface for increased circulation in living mice). Evolution resulted in efficient genome encapsulation for I53-50-v2 and its derivatives (approximately 1 RNA genome per 14 icosahedral capsids for I53-50-v2), protection from blood for I53-50-v3 and I53-50-v4 (82% and 71% protection, respectively), and increased circulation half-life for I53-50-v4 (~4.5 h serum half-life). Full-length RNA genomes were quantified by RT–qPCR, capsid proteins were quantified by Qubit, and genomes per capsid were calculated based on these values by dividing the number of genomes by the number of capsids. **e**, Nucleocapsid genomes are enriched and ribosomal RNA is depleted in nucleocapsids. **f**, Top 13 RNA transcripts encapsulated in I53-50-v4. Nucleocapsid genomes account for more than 74% of the packaged transcripts. **g**, **h**, The relative biodistribution of intact I53-50-v3 (**g**) and I53-50-v4 (**h**) nucleocapsids was evaluated by RT–qPCR of their full-length genomes recovered from mouse organs collected 5 min or 4 h after retro-orbital injection ($n = 3$ biologically independent animals at each time point for each nucleocapsid, I53-50-v3 and I53-50-v4). Red (5 min) and blue (240 min) bars represent the mean of three biologically independent animals, error bars denote the s.e.m., and thick black bars represent the detection limit of the assay. No obvious tissue tropism was observed for either nucleocapsid. At 4 hours after the injection, I53-50-v3 had largely disappeared, whereas I53-50-v4 remained predominantly in the blood with lower levels in the other tissues.

biodistribution of intact nucleocapsids by RT–qPCR of full-length genomes at both 5 min and 4 h. As expected, no obvious tissue tropism was observed for either nucleocapsid. There is no substantial intact I53-50-v3 remaining in any organs by 4 h after injection, consistent with the rapid elimination of I53-50-v3 compared to I53-50-v4 (Fig. 4g, h).

This work demonstrates that by acquiring positive charge on its interior, an otherwise inert self-assembling protein nanomaterial can package its own RNA genome and evolve under selective pressure. Starting from this blank slate, evolution uncovered several simple mechanisms to improve complex properties such as genome packaging, nuclease resistance and *in vivo* circulation time. This suggests paths by which viruses could have arisen from protein assemblies that adopted simple mechanisms to package their own genetic information. Modern viruses are much more complex, having evolved under selective pressure to minimize genome size and to optimize multiple capsid functions required for a complete viral life cycle. However, this makes it difficult to change one property (for example, alter tropism or remove epitopes for pre-existing antibodies[19,20]) without compromising other functions. By contrast, the simplicity of our synthetic nucleocapsids should allow them to be further engineered more freely. Combining the evolvability of viruses with the accuracy and control of computational protein design, synthetic nucleocapsids can be custom-designed and then evolved to optimize function in complex biochemical environments.

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 19 July; accepted 21 November 2017.
Published online 13 December 2017.

1. Deverman, B. E. *et al.* Cre-dependent selection yields AAV variants for widespread gene transfer to the adult brain. *Nat. Biotechnol.* **34,** 204–209 (2016).
2. Chackerian, B., Caldeira, Jdo, C., Peabody, J. & Peabody, D. S. Peptide epitope identification by affinity selection on bacteriophage MS2 virus-like particles. *J. Mol. Biol.* **409,** 225–237 (2011).
3. Smith, G. P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228,** 1315–1317 (1985).
4. Söderlind, E., Simonsson, A. C. & Borrebaeck, C. A. Phage display technology in antibody engineering: design of phagemid vectors and *in vitro* maturation systems. *Immunol. Rev.* **130,** 109–124 (1992).

5. Bale, J. B. *et al.* Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science* **353,** 389–394 (2016).
6. Hsia, Y. *et al.* Design of a hyperstable 60-subunit protein icosahedron. *Nature* **535,** 136–139 (2016).
7. Drouin, L. M. *et al.* Cryo-electron microscopy reconstruction and stability studies of the wild type and the R432A variant of adeno-associated virus type 2 reveal that capsid structural stability is a major factor in genome packaging. *J. Virol.* **90,** 8542–8551 (2016).
8. Sommer, J. M. *et al.* Quantification of adeno-associated virus particles and empty capsids by optical density measurement. *Mol. Ther.* **7,** 122–128 (2003).
9. Pascual, E. *et al.* Structural basis for the development of avian virus capsids that display influenza virus proteins and induce protective immunity. *J. Virol.* **89,** 2563–2574 (2015).
10. Waehler, R., Russell, S. J. & Curiel, D. T. Engineering targeted viral vectors for gene therapy. *Nat. Rev. Genet.* **8,** 573–587 (2007).
11. Harrison, S. C., Olson, A. J., Schutt, C. E., Winkler, F. K. & Bricogne, G. Tomato bushy stunt virus at 2.9 Å resolution. *Nature* **276,** 368–373 (1978).
12. Lilavivat, S., Sardar, D., Jana, S., Thomas, G. C. & Woycechowsky, K. J. *In vivo* encapsulation of nucleic acids using an engineered nonviral protein capsid. *J. Am. Chem. Soc.* **134,** 13152–13155 (2012).
13. Hernandez-Garcia, A. *et al.* Design and self-assembly of simple coat proteins for artificial viruses. *Nat. Nanotechnol.* **9,** 698–702 (2014).
14. Wörsdörfer, B., Woycechowsky, K. J. & Hilvert, D. Directed evolution of a protein container. *Science* **331,** 589–592 (2011).
15. Puglisi, J. D., Chen, L., Blanchard, S. & Frankel, A. D. Solution structure of a bovine immunodeficiency virus Tat-TAR peptide-RNA complex. *Science* **270,** 1200–1203 (1995).
16. Starita, L. M. & Fields, S. Deep mutational scanning: a highly parallel method to measure the effects of mutation on protein function. *Cold Spring Harb. Protoc.* **2015,** 711–714 (2015).
17. Whitehead, T. A. *et al.* Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.* **30,** 543–548 (2012).
18. Knop, K., Hoogenboom, R., Fischer, D. & Schubert, U. S. Poly(ethylene glycol) in drug delivery: pros and cons as well as potential alternatives. *Angew. Chem. Int. Edn Engl.* **49,** 6288–6308 (2010).
19. Hui, D.J. *et al.* AAV capsid $CD8^+$ T-cell epitopes are highly conserved across AAV serotypes. *Mol. Ther. Methods Clin. Dev.* **2,** 15029 (2015).
20. Mingozzi, F. *et al.* $CD8^+$ T-cell responses to adeno-associated virus capsid in humans. *Nat. Med.* **13,** 419–422 (2007).

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

Acknowledgements We thank R. Chari for RNA-seq advice; S. Bustin for RT–qPCR advice; E. Gray and N. Arroyo for heparinized mouse blood; D. Veesler, J. Kollman and M. Johnson for EM advice; Y. Hsia for DLS advice; C. Walkey, Y. Hsia, G. Rocklin, J. Nelson, A. Chatterjee, S. Kosuri, G. Church, J. Bloom and A. Hessel for suggestions. This work was supported by the Howard Hughes Medical Institute (D.B.), the Bill and Melinda Gates Foundation (D.B. and N.P.K., grant number OPP1118840), the Defense Advanced Research Projects Agency (D.B. and N.P.K., grant number W911NF-15-1-0645), and the NIH (S.H.P., grant number NIH1R01CA177272; D.L.S., grant number NIH1R21NS099654-01A1). G.L.B. was supported by a National Science Foundation Graduate Fellowship. M.J.L. was supported by a Washington Research Foundation Innovation Postdoctoral Fellowship and a Cancer Research Institute Irvington Fellowship from the Cancer Research Institute. H.H.G. was supported by an NIH training grant (NIH5T32HL0071312). U.N. was supported in part by a PHS National Research Service Award (T32GM007270) from NIGMS.

Author Contributions G.L.B. and M.J.L. designed the research and the experimental approach with guidance from N.P.K. and D.B.; G.L.B. and M.J.L. performed the evolution, nucleocapsid characterization, Illumina sequencing, and data analysis; H.H.G. and D.L.S. designed and performed the *in vivo* mouse experiments, and samples were processed by G.L.B. and M.J.L.; U.N. designed, performed, and analysed electron microscopy experiments; D.E. and J.B.B. designed the starting protein assemblies that were subsequently used for RNA packaging; S.K., G.H.L., A.Y. and R.R. assisted with cloning and protein purification; S.H.P., N.P.K. and D.B. supervised the research; G.L.B. and M.J.L. wrote the manuscript and produced the figures with guidance from H.H.G., D.L.S., U.N., S.H.P., N.P.K. and D.B.; G.L.B., M.J.L., H.H.G., D.L.S., U.N., J.B.B., S.H.P., N.P.K. and D.B. revised the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to D.B. (dabaker@uw.edu).

Reviewer Information *Nature* thanks D. Schaffer and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size.

Solutions and buffers. Lysogeny broth (LB): autoclave 10 g tryptone, 5 g yeast extract, 5 g NaCl, 1 l dH_2O. LB agar plates: autoclave LB with 15 g l^{-1} bacto agar. Terrific broth (TB): autoclave 12 g tryptone, 24 g yeast extract, 4 ml glycerol, 950 ml dH_2O separately from KPO_4 salts (23.14 g KH_2PO_4, 125.31 g K_2HPO_4, 1 l dH_2O); mix 950 ml broth with 50 ml KPO_4 salts at room temperature. Antibiotics: kanamycin (50 μg ml^{-1} final). Inducers: β-D-1-thiogalactopyranoside (IPTG, 500 μM final). Tris-buffered saline with imidazole (TBSI): 250 mM NaCl, 20 mM imidazole, 25 mM Tris-HCl, pH 8.0. Lysis buffer: TBSI supplemented with 1 mg ml^{-1} lysozyme (Sigma, L6876, from chicken egg), 1 mg ml^{-1} DNase I (Sigma, DN25, from bovine pancreas), and 1 mM phenylmethanesulfonyl fluoride (PMSF). Elution buffer: 250 mM NaCl, 500 mM imidazole, 25 mM Tris-HCl, pH 8.0. Phosphate-buffered saline (PBS): 150 mM NaCl, 20 mM $NaPO_4$, pH 7.4. 20× lithium borate buffer (use at 1×): 1 l dH_2O, 8.3 g lithium hydroxide monohydrate, 36 g boric acid. Tris-glycine buffer: 25 mM Tris-HCl, 192 mM glycine, 0.1% SDS, pH 8.3.

DNA cloning by PCR mutagenesis and isothermal assembly. Synthetic genes encoding I53-50 and I53-47[5] were amplified using Kapa High Fidelity Polymerase according to manufacturer's protocols with primers incorporating the desired mutations or the Btat peptide. The resulting amplicons were isothermally assembled[21] with PCR-amplified or restriction-digested (NdeI and XhoI) pET29b fragments and transformed into chemically competent *E. coli* XL1-Blue cells. Monoclonal colonies were verified by Sanger sequencing. Plasmid DNA was purified using a Qiagen miniprep kit and transformed into chemically competent *E. coli* BL21(DE3)* cells for protein expression.

Kunkel mutagenesis. Kunkel mutagenesis was performed as previously described[22]. In brief, *E. coli* CJ236 was transformed with the desired pET vector and then infected with bacteriophage M13K07. Single-stranded DNA (ssDNA) was purified from PEG/NaCl-precipitated bacteriophage using a Qiaprep M13 kit. Oligonucleotides were phosphorylated for 1 h with T4 polynucleotide kinase (NEB, M0201) and annealed to purified ssDNA plasmids. For routine cloning, annealing was performed using a temperature ramp from 95 °C to 25 °C over 30 min. For library generation, annealing mixtures were denatured at 95 °C for 2 min, followed by annealing for 5 min at either 55 °C (220 bp agilent oligonucleotides) or 50 °C (all other oligonucleotides). Oligonucleotides were extended using T7 DNA polymerase (NEB) for 1 h at 20 °C and transformed into *E. coli* as described for either routine cloning or library generation.

Transformation of DNA libraries. Plasmid DNA libraries generated as described above by isothermal assembly or kunkel mutagenesis were purified by SPRI purification[23] and electrotransformed into *E. coli* DH10B (Invitrogen 18290-015) to produce libraries with at least 10× coverage. Transformed libraries were grown as lawns on LB agar plates containing 50 μg ml^{-1} kanamycin. Additionally, a tenfold dilution series of the transformed library was spotted onto an additional plate to assess library size. After 12–18 h of growth, the resulting lawn of cells was scraped from the plate into 1 ml of LB and pelleted at 16,000 *g* for 30 s. Plasmid DNA was purified directly from this cell pellet using a Qiagen miniprep kit and electrotransformed into *E. coli* BL21(DE3)* with a minimum of 10× coverage of the library. The resulting bacterial lawns were then lifted from plates in 1mL TB and inoculated directly into expression cultures.

Deep mutational scanning library design, amplification, and purification. For the deep mutational scanning library, the DNA sequence encoding the two components of I53-50-v2 was divided into 7 windows of up to 159 bp. For each window, a pool of oligonucleotides was synthesized to mutate every residue of I53-50-v2 in the specified window (Agilent SurePrint Oligonucleotide Library Synthesis, OLS). Each oligonucleotide encoded a single amino acid change using the most common codon in *E. coli* for that amino acid. To disambiguate bona fide mutations from sequencing and reverse transcription errors, mutagenic oligonucleotides included silent mutations on either side of the mutagenized position. Each of the 7 oligonucleotide pools was amplified from the OLS pool using primers that annealed to constant regions flanking the mutagenic sequences. Reaction progress was monitored by SYBR green fluorescence on a Bio-Rad CFX96 to prevent overamplification. The resulting amplicons were then PAGE-purified and subjected to an additional round of amplification and SPRI purification. A final PCR reaction was set up with only the reverse primer to perform linear amplification of the desired primer sequence (50 cycles of temperature cycling were performed to generate a DNA sample highly enriched for the reverse strand). This sample was then purified using a Qiagen QIAquick PCR Purification Kit. The resulting pool of single stranded oligonucleotides was then used in a kunkel reaction as described above for library generation.

Hydrophilic polypeptide library design, amplification, and purification. The hydrophilic polypeptide library was generated by alternating sets of hydrophilic amino acids (DE, ST, QN, GE, EK, ES, EQ, EP or PAS) with a guest residue (A, S, T, E, D, Q, N, K, R, P, G, L or I) introduced between every 1, 2 or 5 occurrences to generate a final peptide of 59 amino acids in length. An additional 21 peptides were generated by splitting known hydrophilic peptides[24,25] into 59 amino acid chunks or concatenating one of their primary repeating units. All polypeptide sequences were reverse translated to DNA using codon frequencies found in *E. coli* K12[26], and flanking sequences were added for amplification. These oligonucleotide sequences were synthesized using Agilent OLS technology. After amplification, flanking regions were removed using the AgeI and HindIII restriction enzymes, and the insert was cloned onto the C terminus of the I53-50-v3 pentamer subunit by ligation (T4 ligase, NEB M0202, final concentration: 40 U $μl^{-1}$, 1× T4 ligase buffer with 1 mM ATP). The resulting DNA was SPRI purified and transformed as described above.

Protein expression and purification. *E. coli* BL21(DE3)* expression cultures were grown to an optical density (600 nm) of 0.6 in 500 ml TB supplemented with 50 μg ml^{-1} kanamycin at 37 °C with shaking at 225 r.p.m. Expression was induced by the addition of IPTG (500 μM final). Expression proceeded for 4 h at 37 °C with shaking at 225 r.p.m. Cultures were harvested by centrifugation at 5,000*g* for 10 min and stored at −80 °C.

Cell pellets were resuspended in TBSI and lysed by sonication or homogenization using a Fastprep96 with lysing matrix B. Lysate was clarified by centrifugation at 24,000*g* for 30 min and passed through 2 ml of nickel-nitrilotriacetic acid agarose (Ni-NTA) (Qiagen, 30250), washed 3 times with 10 ml TBSI, and eluted in 3 ml of elution buffer, of which only the second and third milliltre were kept. EDTA was immediately added to 5 mM final concentration to prevent Ni-mediated aggregation.

For *in vitro* evolution (Figs 1–3, Extended Data Figs 1, 2, 3d, 4c and Supplementary Figs 2–8) and all experiments involving hydrophilic polypeptides, synthetic nucleocapsids were prepared with a C-terminal histidine tag on the pentameric subunit. For these constructs, purification proceeded immediately from Ni-NTA elution to size exclusion chromatography (SEC) using a Superose 6 Increase column (GE Healthcare, 29-0915-96) in TBSI. Nucleocapsids with hydrophilic polypeptides intended for mouse experiments were subjected to endotoxin removal as described below prior to SEC.

For all *in vivo* evolution experiments (Fig. 4, Extended Data Fig. 4a, d and Supplementary Figs 8, 9), synthetic nucleocapsids were prepared with a N-terminal, thrombin cleavable histidine tag on the pentameric subunit to allow scarless removal. This was done to allow removal of the affinity tag for *in vivo* use and to prevent the divalent cation-dependent aggregation observed in the C-terminal histidine-tagged constructs. After elution from Ni-NTA, these samples were dialysed into PBS and treated with thrombin at a final concentration of 0.00264 U $μl^{-1}$ for 90 min at 20 °C to remove the histidine tag. Thrombin was inactivated by addition of PMSF (1 mM final concentration), and nucleocapsids were purified by SEC using a Superose 6 Increase column in PBS.

Endotoxin was removed from all samples intended for animal studies. Endotoxin removal was performed after thrombin cleavage by addition of Triton X-114 (1% final concentration vol/vol) followed by incubation at 4 °C for 5 min, incubation at 37 °C for 5 min, and centrifugation at 24,000 *g* at 37 °C for 2 min. The supernatant was then removed, incubated 4 °C for 5 min, incubated at 37 °C for 5 min, and centrifuged at 24,000*g* at 37 °C for 2 min to ensure optimal endotoxin removal before continuing with SEC purification in PBS.

Gel electrophoresis. Native agarose gels: agarose gels were prepared using 1% ultrapure agarose (Invitrogen) in lithium borate buffer. For synthetic nucleocapsid samples, 20 μl purified synthetic nucleocapsids were treated with 10 μg ml^{-1} RNase A (20 °C for 10 min), mixed with 4 μl 6× loading dye (NEB B7025S, no SDS), and electrophoresed at 100 V for 45 min. Gels were then stained with SYBR gold (Thermo Fischer Scientific, S11494) for RNA followed by Gelcode (Thermo Fischer Scientific, 24590) for protein.

DNA gels: 1% agarose gels were prepared containing SYBR Safe (Invitrogen) according to the manufacturer's protocols.

Protein SDS–PAGE: SDS–PAGE was performed using 4–20% polyacrylamide gels (Bio-Rad) in Tris-glycine buffer.

RNA purification and reverse transcription. RNA was purified using TRIzol (Thermo Fischer Scientific, 15596018) and a Qiagen RNeasy kit (Qiagen, 74106) according to the manufacturers' instructions. In brief, 100 μl synthetic nucleocapsid samples were mixed vigorously with 500 μl TRIzol. 100 μl chloroform was added and mixed vigorously, and then the solution was centrifuged for 10 min at 24,000*g*. Then, 150 μl of the aqueous phase was mixed with 150 μl of 100% ethanol, transferred to a RNeasy spin column for purification according to the manufacturer's instructions, and eluted in 50 μl nuclease-free dH_2O. For samples intended for absolute quantification (including RNA standards), yeast tRNA was added to 100 ng $μl^{-1}$ final concentration to ensure consistent sample complexity.

Reverse transcription was carried out using Thermoscript reverse transcriptase for 1 h at 53 °C according to the manufacturer's instructions, with the

only modifications being that a gene-specific primer (skpp_reverse) was used. Thus, a 10 μl reaction contained: 1 μl dNTPs (10 mM each), 1 μl DTT (100 μM), 1 μl Thermoscript reverse transcriptase, 2 μl cDNA synthesis buffer, 1 μl RNaseOut, 1 μl skpp_reverse (10 μM), 2 μl purified RNA template, and 1 μl nuclease-free dH_2O. Controls lacking reverse transcriptase were set up identically except with the substitution of nuclease-free dH_2O in place of Thermoscript reverse transcriptase.

Quantitative PCR. Quantitative PCR was performed in a 10 μl reaction using a Kapa High Fidelity PCR kit (Kapa Biosystems, KK2502) according to the manufacturer's instructions with the addition of SYBR green at 1× concentration and 0.5 μM forward and reverse primers (skpp_fwd and skpp_Offset_Rev) for quantification of nucleocapsid RNA. Thermocycling and C_q calculations were performed on a Bio-Rad CFX96 with the following protocol: 5 min at 95 °C, then 40 cycles of: 98 °C for 20 s, 64 °C for 15 s, 72 °C for 90 s.

Allele-specific qPCR was performed using Kapa 2G Fast polymerase readymix along with 1× SYBR green, 3 μl of 100× diluted cDNA template, and 0.5 μM each of the forward and reverse allele-specific primers for each construct. Thermocycling and C_q calculations were performed on a Bio-Rad CFX96 with the following protocol: 5 min at 95 °C, then 40 cycles of: 95 °C for 15 s, 58 °C for 15 s and 72 °C for 90 s.

Absolute quantification of full-length RNA per protein capsid was calculated from C_q values using a linear fit ($-\log([\mathrm{RNA}]) = m \times (C_q) + b$, where m is the slope and b is the y-intercept of a standard curve that consists of *in vitro* transcribed nucleocapsid RNA). *In vitro* transcription was performed using a NEB HiScribe T7 high yield RNA synthesis kit (NEB, E2040S) according to the manufacturer's protocols. Excess DNA was degraded using RNase-free DNase I (NEB, M0303), and RNA was purified using Agencourt RNAClean XP (Beckman Coulter, A63987) according to manufacturer protocols. The concentration of this standard was measured using a Qubit RNA HS Assay Kit (Life Technologies, Q32852), and a tenfold dilution series was prepared in nuclease-free dH_2O supplemented with 100 ng μl^{-1} yeast tRNA. The dilution series samples were then processed in parallel with the synthetic nucleocapsid samples using the RNA purification and reverse transcription protocol above, and run on the same qPCR plate as the samples to be quantified.

In the pooled samples used to compare the fitness of I53-50-v1, I35-50-v2, I53-50-v3 and I53-50-v4, the total amount of full-length nucleocapsid genome was quantified by qPCR performed with skpp_fwd and skpp_rev using the Kapa High Fideltiy PCR kit as described above. Subsequently, the relative fraction of RNA corresponding to each version was determined by allele specific qPCR as described above using allele-specific primers (Supplementary Table 6) unique to each version. Absolute quantification was with respect to a standard curve for each version prepared as described above. The fractional RNA content from each version was then multiplied by total amount of full-length genomes.

***In vitro* synthetic nucleocapsid selection conditions.** The total amount of RNA packaged in nucleocapsids was evaluated by treating 100 μl synthetic nucleocapsids with 10 μg ml^{-1} RNase A at 20 °C for 10 min ('total RNA') so as to degrade non-encapsulated RNA. Reaction buffer was PBS for N-terminal histidine-tagged constructs or TBSI for C-terminal histidine-tagged constructs. More stringent RNase protection assays were performed with 10 μg ml^{-1} RNase A at 37 °C for the specified duration ('RNase'). Protection from blood was assessed by diluting synthetic nucleocapsids 1:10 in heparinized mouse whole blood (collected from the vena cava of mice sacrificed using a lethal dose of avertin and stabilized in 6 U ml^{-1} heparin) and incubating at 37 °C for the specified duration ('blood'). Samples were then centrifuged at 24,000*g* for 2 min before adding the supernatant to TRIzol. RNA was purified as described in the RNA Purification and RT–qPCR sections. All reactions were quenched by adding the sample directly to 500 μl TRIzol.

General information about mouse work. 6–8-week-old BALB/c mice were selected randomly and retro-orbitally injected with 150 μl of synthetic nucleocapsids. All mice were female to minimize any unknown variability in tissue distribution bias attributed to animal sex. No blinding was performed. The Institutional Animal Care and Use Committee (IACUC) at the University of Washington authorized all animal work in accordance with ethical animal use and regulations.

***In vivo* synthetic nucleocapsid selection conditions.** Synthetic nucleocapsid libraries containing either hydrophilic polypeptides (104 μg ml^{-1}, 3.7×10^{12} injected particles per mouse) or exterior surface mutations (570 μg ml^{-1}, 2.0×10^{13} injected particles per mouse) were created and selected for circulation time in live mice. Five mice per library underwent retro-orbital injections and tail lancet blood draws at 5, 10, 15 and 30 min, with a final euthanization and blood draw at 60 min. Following Illumina MiSeq sequencing of the selected nucleocapsid libraries, the circulation times of several selected variants (I53-50-v1, I53-50-v2, I53-50-v3, 10 hydrophilic polypeptide variants, and 4 surface mutation variants were pooled to 570 μg ml^{-1} total protein) were compared in 5 mice with tail lancet blood draws at 5, 15, 30, 60 and 120 min, submental collection[10] at 4 h, and final euthanization and blood draw at 6 h. I53-50-v4 was created based on the consensus sequence of the most common residues in the library after *in vivo* selection.

Synthetic nucleocapsid characterization for Fig. 4a–d. I53-50-v1, I53-50-v2, I53-50-v3 and I53-50-v4 were expressed in *E. coli* BL21(DE3)*, harvested by cell lysis, purified by IMAC, dialysed into PBS, cleaved by thrombin, subjected to endotoxin removal, and purified by SEC as described above. The protein concentrations for each sample were determined using a Qubit Protein Assay kit (Thermo Fisher Scientific, Q33211), and samples were mixed to give a final concentration of 170 μg ml^{-1} nucleocapsid protein for each version (680 μg ml^{-1} total). This pool was split into four different samples that were each subjected to the total RNA, RNase, blood and *in vivo* selection conditions described above ($n = 3$ independent replicates for each *in vitro* selection condition). For *in vivo* selections, 150 μl of the pool was injected retro-orbitally (2.4×10^{13} particles per mouse), and tail lancet draws were performed at 5 min, 1 h, 3 h and 6 h, submental collection[10] at 10 h, and final euthanization and blood draw at 24 h. While I53-50-v0 was not included in this nucleocapsid pool, we independently showed that it does not package detectable RNA.

Synthetic nucleocapsid biodistribution. I53-50-v3 and I53-50-v4 were injected into 6 mice each. Animals were then euthanized after either 5 min or 4 h (3 animals per nucleocapsid version at each time point). After blood was collected, animals were exsanguinated by transcardial perfusion with saline. Half of each bisected organ and 20 μl of whole blood were collected into tubes containing 500 μl TRIzol and homogenized. RNA was purified, total tissue RNA was measured by either absorbance at $A_{260\,nm}$ (organs) or Qubit RNA HS Assay Kit (blood, owing to its lower total RNA), and full-length nucleocapsid genomes were quantitated by RT–qPCR as described above.

Negative-stain electron microscopy specimen preparation, data collection and data processing. 6 μl of purified protein (I53-50-v0, I53-50-v1, I53-50-v2, I53-50-v3, I53-50-v4, I53-50-Btat, I53-47-v0, I53-47-v1 and I53-47-Btat) at 0.04–0.30 mg ml^{-1} were applied to glow discharged, carbon-coated 300-mesh copper grids (Ted Pella), washed with Milli-Q water and stained with 0.75% uranyl formate as described previously[27]. Screening and sample optimization was performed on a 100 kV Morgagni M268 transmission electron microscope (FEI) equipped with an Orius charge-coupled device (CCD) camera (Gatan). Data were collected with Leginon automatic data-collection software[28] on a 120 kV Tecnai G2 Spirit transmission electron microscope (FEI) using a defocus of 1 μm with a total exposure of 30 e^- Å^{-2}. All final images were recorded using an Ultrascan 4000 4 k × 4 k CCD camera (Gatan) at 52,000 × magnification at the specimen level. For data collection used in two-dimensional class averaging, the dose of the electron beam was 80 e^- Å^{-2}, and micrographs were collected with a defocus range between 1.0 and 2.0 μm. Coordinates for unique particles (7,979 for I53-50-v0 and 7,130 for I53-50-v4) were obtained for averaging using EMAN2[29].

Illumina sequencing sample preparation for evolution experiments. Evolution experiments were analysed by performing targeted RNA-seq on full-length nucleocapsid genomes surviving the specified selection condition (RT–qPCR using skpp_reverse as the reverse transcription primer and using skpp_fwd and skpp_Offset_Rev as the qPCR primers). The starting populations and selected populations were evaluated by sequencing nucleocapsid genomes extracted from producer cells or nucleocapsids, respectively. After SPRI purification, two sequential qPCR reactions were performed using Kapa HiFi polymerase to add sequencing adapters and barcodes, respectively. qPCR reactions were monitored by SYBR green fluorescence and terminated before completion so as to prevent over-amplification. The resulting amplicons were purified using SPRI purification or a Qiagen QIAquick Gel Extraction Kit. The resulting amplicons were then denatured and loaded into a Miseq 600 cycle v3 (Illumina) kit and sequenced on an Illumina MiSeq according to the manufacturer's instructions.

Illumina sequencing sample preparation for comprehensive RNA-seq. The composition of encapsulated RNA was evaluated by performing comprehensive RNA-seq on total RNA from producer cells (representing expression levels) and nucleocapsids (representing encapsulated RNA). RNA was extracted using TRIzol and purified using a Direct-zol RNA MiniPrep Plus kit (Zymo Research, R2072) with on-column DNase digestion. The purified RNA was quantitated using a Qubit RNA HS Assay Kit, and 100 ng of RNA was used to prepare each RNA-seq library with a NEBNext Ultra RNA Library Prep Kit for Illumina (NEB, E7530S). Each library was PCR amplified using Kapa HiFi polymerase to add sequencing barcodes before being pooled for sequencing. The resulting libraries were then denatured and loaded into an Illumina NextSeq 500/550 High Output Kit v2 (75 cycles) and sequenced on an Illumina NextSeq according to the manufacturer's instructions.

Sequencing analysis for evolution experiments. Raw sequencing reads were converted to fastq format and parsed into separate files for each sequencing barcode using the Generate Fastq workflow on the Illumina MiSeq. Forward and reverse reads were combined using the read_fuser script from the enrich package[30].

For all libraries, enrichment values were calculated as the change in fraction of the library corresponding to each linked sequence (rank order of variants)

or unlinked substitutions (heat maps) that were observed at least 10 times in the naive library. The base 10 logarithm of each value was then taken in order to give enrichment values that more symmetrically span enrichment and depletion.

For the charge optimization library, the total interior charge of each variant was calculated by summing the number of lysine and arginine residues, and subtracting the number of aspartate and glutamate residues determined to be on the interior surface of the capsid by visual inspection of the design model. For the deep mutational scanning library, substitutions were only counted if they contained the expected silent mutation barcodes as described in the oligonucleotide design section. This greatly reduces the effect of both RT–PCR errors and sequencing errors because instead of a minimum of one error allowing a miscalled amino acid mutation, a minimum of three errors are required for a mutation to be miscalled.

Heat maps were generated using a custom MatPlotLib[31] script by mapping the calculated log enrichment values onto a LinearSegmentedColormap (purple, white, orange; rgb = (0.75,0,0.75), (1,1,1), (1.0,0.5,0)) using the 'pcolormesh' function. The minimum and maximum values of the colour-mesh were set as shown in each figure to fully utilize the dynamic range of the colour-map. A pymol session coloured by the average log enrichment of all 20 amino acids at each position was created by substituting average log enrichment values for B-factors in the pdb file and running the command: 'spectrum b, purple_white_white_orange, minimum = −1.5, maximum = 0.6'. Note that this is rescaled relative to the colouring of individual residues because the averages span a smaller range than the individual values, and thus a different colour range is needed to clearly differentiate values.

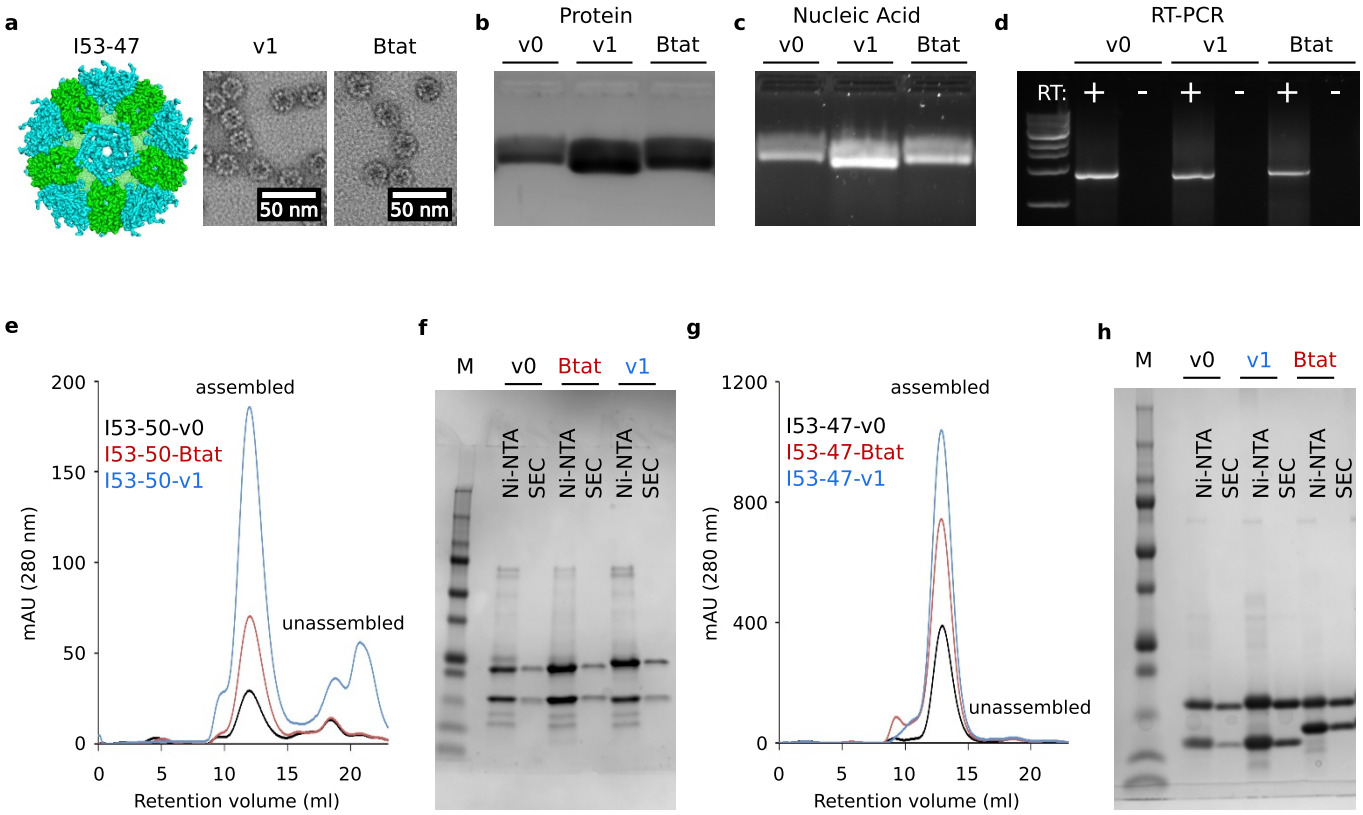
Sequencing analysis for comprehensive RNA-seq. RNA-seq data were converted from bcl format to fastQ format using the Illumina bcl2fastq script. HISAT2[32] converted fastQ to sam, and SAMtools[33] converted sam files to sorted bam files. StringTie[34] was used to calculate gene expression as transcripts per kilobase million (TPM).

Dynamic light scattering. Dynamic light scattering was performed on a DynaPro NanoStar (Wyatt) DLS setup. I53-50-v0, I53-50-v1, and I53-50-v4 were evaluated with 0.2 mg ml^{-1} of nucleocapsid protein in PBS at 25 °C. Data analysis was performed using DYNAMICS v7 (Wyatt) with regularization fits.

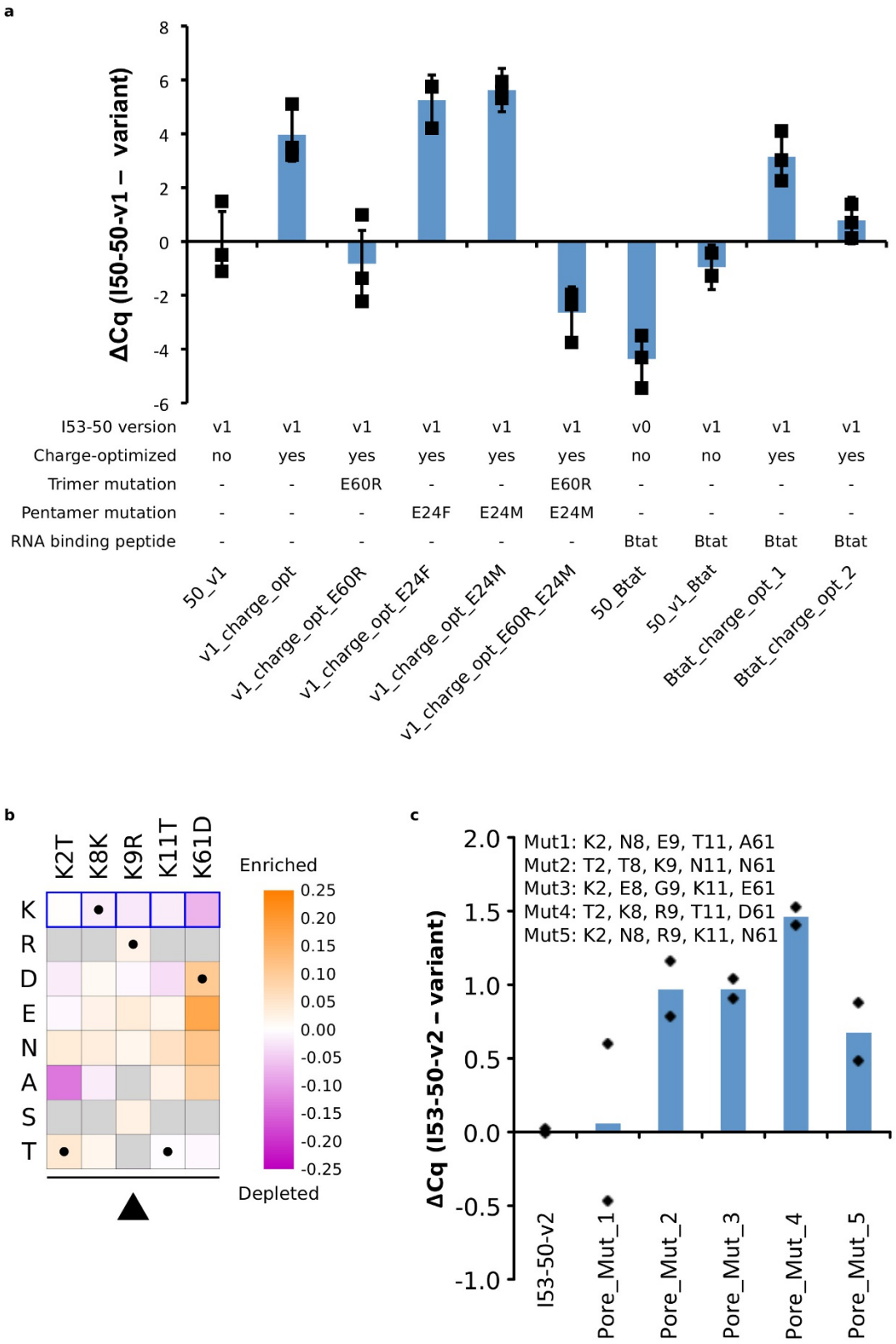
Code availability. Custom scripts for Illumina sequencing analysis are available on github (https://github.com/mlajoie/Synthetic_Nucleocapsid).

Data availability. Nucleocapsid genome files (GenBank format) and design models (pdb format) are available on github (https://github.com/mlajoie/Synthetic_Nucleocapsid). All raw sequencing data from Figs 2–4, Extended Data Figs 2–4, and Supplementary Figs 3, 7–9 are available at the NCBI Sequence Read Archive under BioProject accession PRJNA417493. Source data for Figs 1d–f, 4a–d, Extended Data Figs 1b–d, 3d, e, 4c, d and Supplementary Fig. 8 are provided with the paper. All other raw data not included in the manuscript are available from the corresponding author upon request.

21. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6,** 343–345 (2009).
22. Kunkel, T. A. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proc. Natl Acad. Sci. USA* **82,** 488–492 (1985).
23. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22,** 939–946 (2012).
24. Alvarez, P., Buscaglia, C. A. & Campetella, O. Improving protein pharmacokinetics by genetic fusion to simple amino acid sequences. *J. Biol. Chem.* **279,** 3375–3381 (2004).
25. Schellenberger, V. *et al.* A recombinant polypeptide extends the in vivo half-life of peptides and proteins in a tunable manner. *Nat. Biotechnol.* **27,** 1186–1190 (2009).
26. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **41,** D36–D42 (2013).
27. Nannenga, B. L., Iadanza, M. G., Vollmar, B. S. & Gonen, T. Overview of electron crystallography of membrane proteins: crystallization and screening strategies using negative stain electron microscopy. *Curr. Protoc. Protein Sci.* Chapter 17, Unit17.15 (2013).
28. Suloway, C. *et al.* Automated molecular microscopy: the new Leginon system. *J. Struct. Biol.* **151,** 41–60 (2005).
29. Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157,** 38–46 (2007).
30. Fowler, D. M., Araya, C. L., Gerard, W. & Fields, S. Enrich: software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics* **27,** 3430–3431 (2011).
31. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9,** 90–95 (2007).
32. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12,** 357–360 (2015).
33. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
34. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protocols* **11,** 1650–1667 (2016).

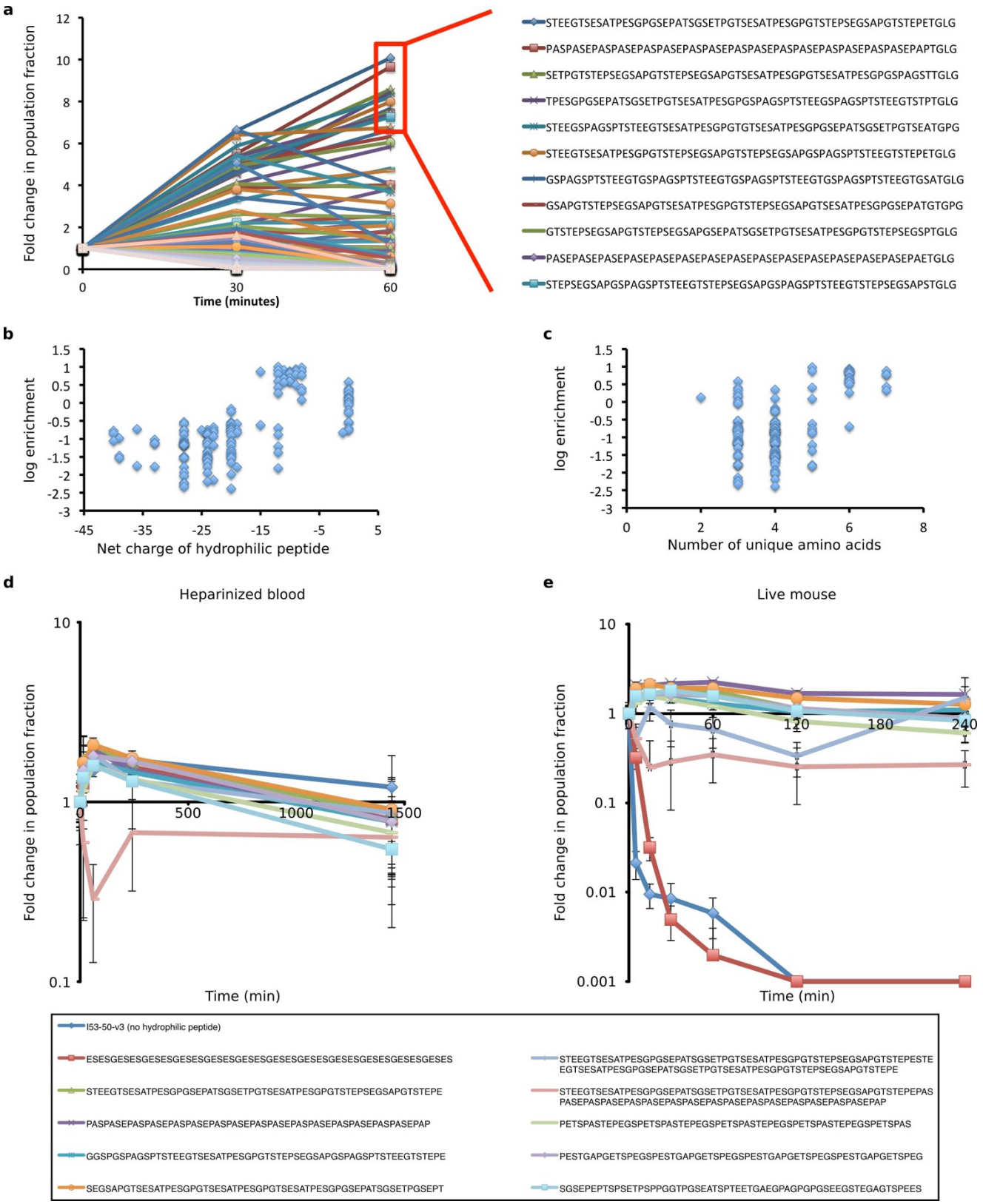


Extended Data Figure 1 | I53-47 nucleocapsids and SEC. a, Design model of I53-47 and negative-stain electron micrographs of I53-47-v1 (designed positively charged interior) and I53-47-Btat (BIV Tat RNA-binding peptide translationally fused to the C terminus of the capsid trimeric subunit). Micrographs shown are representative of the entire sample tested on between one and three different grids, each at different concentrations. **b**, Synthetic nucleocapsids were Ni-NTA-purified, RNase-treated, and electrophoresed on non-denaturing 1% agarose gels. The gels were stained with Coomassie (protein; **b**) and SYBR gold (nucleic acid, **c**). Nucleic acids co-migrated with capsid proteins for all three versions of I53-47, suggesting that all versions package nucleic acid. **d**, Full-length synthetic nucleocapsid genomes were recovered from each sample by RT–qPCR. Plus and minus symbols indicate PCR performed on templates prepared with and without reverse transcriptase, respectively, confirming that all versions package their own full-length RNA genomes. This procedure is part of our standard quality control for synthetic nucleocapsids and has been performed reproducibly more than 10 times, including once on the I53-47 nucleocapsid shown here. **e**–**h**, SEC of nucleocapsids. RNA-packaging capsids show identical SEC retention volume as the original published capsid[5]. Three versions of I53-50 and I53-47 were analysed: v0 is the original published design, v1 has the designed positively charged interior, and Btat has the BIV Tat RNA-binding peptide translationally fused to the C terminus of the capsid trimer subunit. **e**, SEC traces of I53-50 capsids were performed on a GE superose 6 increase column. **f**, SDS–PAGE of samples before and after SEC purification shows both subunits in the expected 1:1 stoichiometry. **g**, **h**, SEC traces and SDS–PAGE for I53-47 capsids. This procedure is part of our standard quality control for synthetic nucleocapsids and has been performed reproducibly more than 10 times.



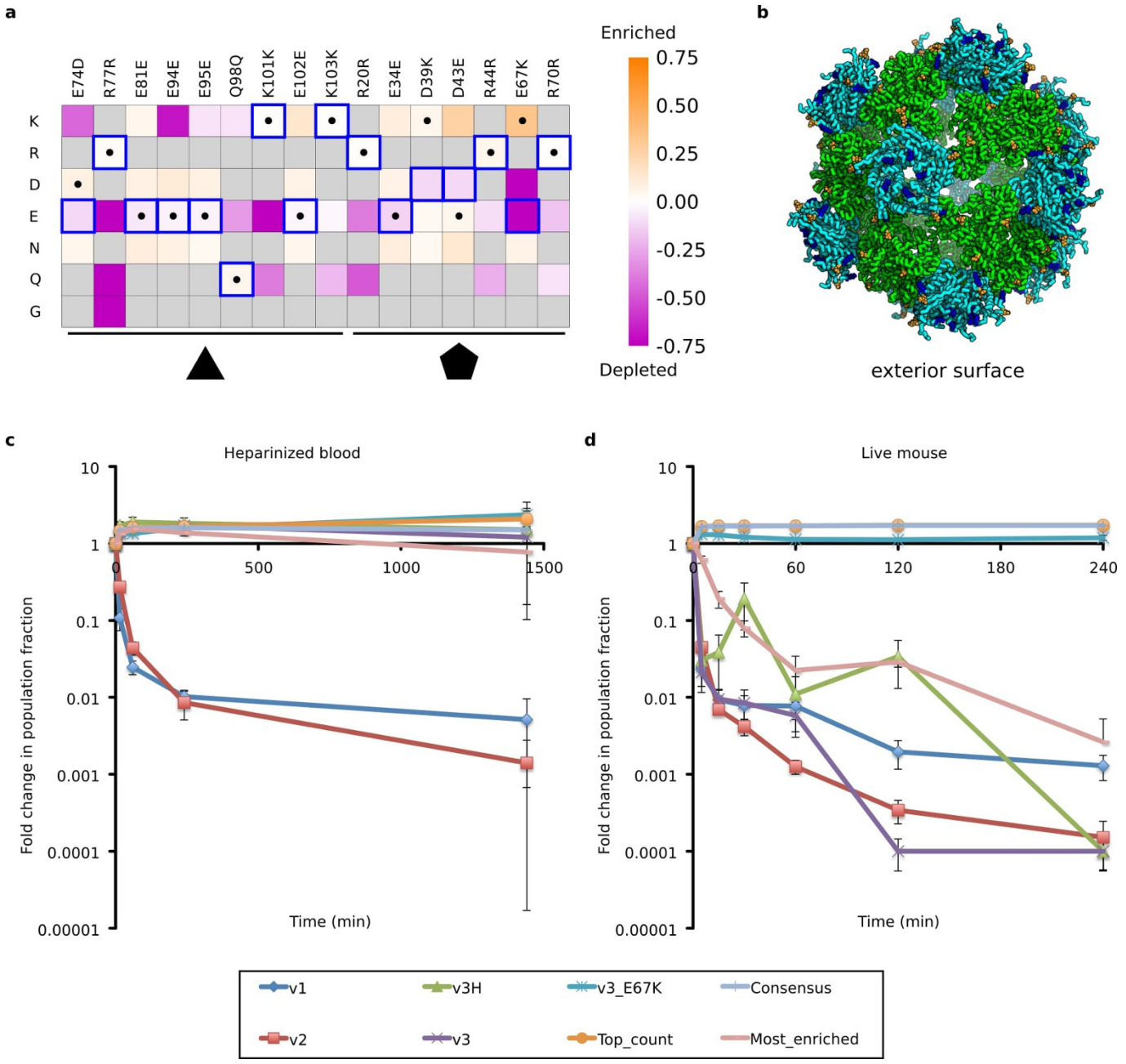
Extended Data Figure 2 | See next page for caption.

Extended Data Figure 2 | Top synthetic nucleocapsid candidates for I53-50-v2 and I53-50-v3. a, Top candidate testing to choose I53-50-v2 with improved genome packaging. New variants were created rationally based on the best sequences from the evolved interior charge optimization (Fig. 2) and interface (Supplementary Fig. 2) libraries. The amount of packaged full-length mRNA was compared for each of these nucleocapsids. Each nucleocapsid was expressed, purified by IMAC, and treated with 10 μg ml^{-1} RNase A at 20 °C for 10 min in triplicate. RT–qPCR was used to determine the relative amount of full-length mRNA packaged in each variant. C_q values are reported relative to those of I53-50-v1 ($C_{qI53\text{-}50\text{-}v1} - C_{qvariant}$). The charge-optimized variant with E24F was chosen as I53-50-v2 based on these data. In the absence of a discernable difference in packaging between E24M and E24F, E24F was selected owing to the apparent preference for hydrophobic residues at that position (Supplementary Fig. 2). Data points represent the values of three independent biological replicates, and error bars represent s.e.m. **b**, **c**, Top candidate testing to choose I53-50-v3 with improved nuclease resistance. **b**. Heat map of log enrichments for each mutation explored in the combinatorial library to remove positively charged residues near the nucleocapsid pore. A single round of selection (10 μg ml^{-1} RNase A, 37 °C, 1 h) was performed. Purple and orange indicate mutations that were depleted or enriched in the selected population, respectively. Blue squares and black dots indicate the I53-50-v2 starting sequence and I53-50-v3 selected sequence, respectively. **c**, Enriched variants selected from the combinatorial library were expressed, purified by IMAC and SEC, and treated with 10 μg ml^{-1} RNase A at 37 °C for 1 h in duplicate. RT–qPCR was used to determine the relative amount of full-length mRNA packaged in each variant. C_q values are reported relative to those of I53-50-v2 ($C_{qI53\text{-}50\text{-}v2} - C_{qvariant}$). Data points represent the values of two independent biological replicates, and bars represent the mean of these values. The variant labelled Pore_Mut_4 was chosen as I53-50-v3 based on this data.

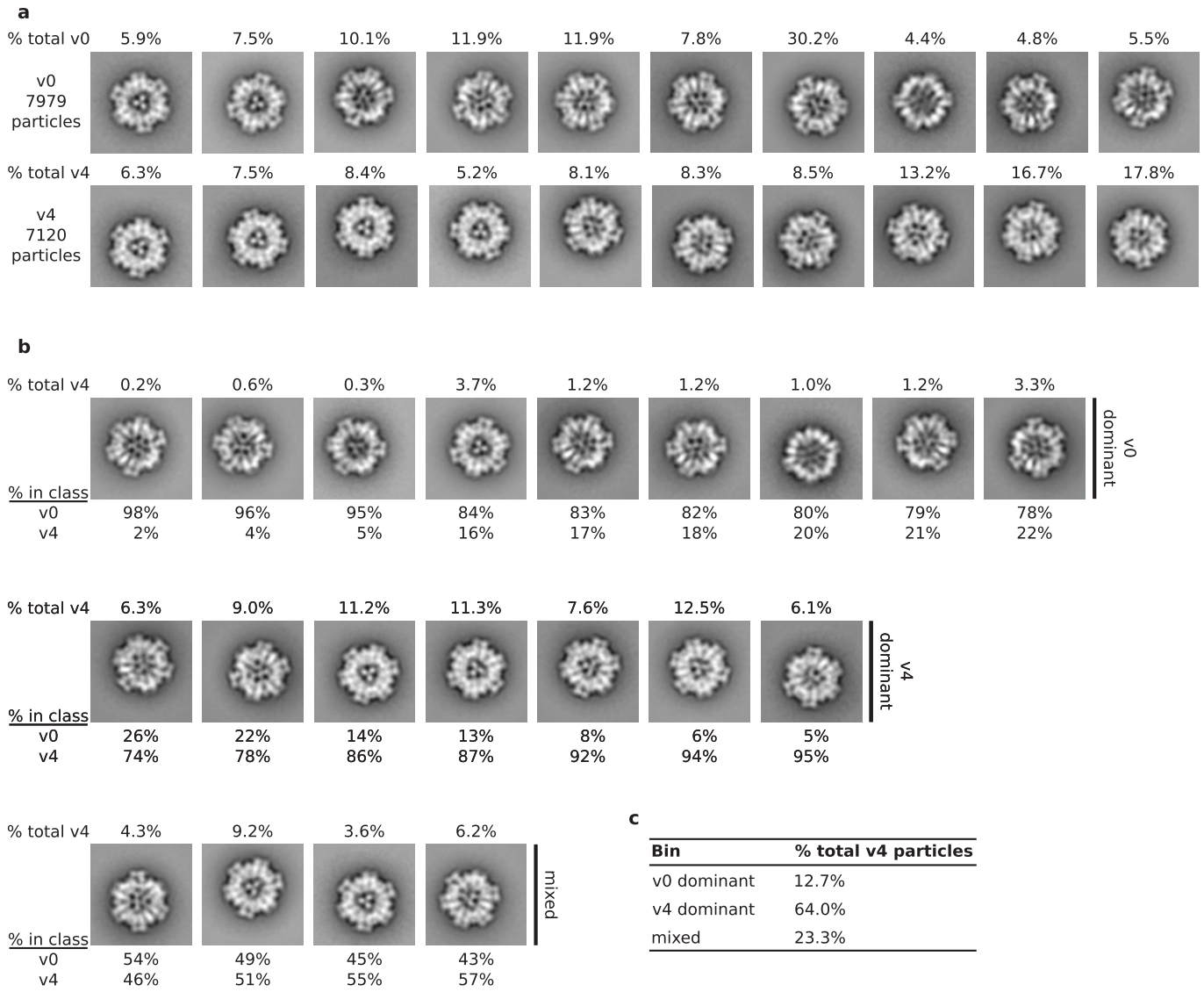


Extended Data Figure 3 | See next page for caption.

Extended Data Figure 3 | Evolution and performance of nucleocapsids modified with hydrophilic polypeptides *in vitro* or *in vivo*. **a**, The change in population fraction corresponding to each variant was calculated from Illumina MiSeq counts for the input pool ($t = 0$), RNA recovered from circulation after 30 min ($n = 3$ biologically independent mice), and RNA recovered from circulation after 60 min ($n = 2$ biologically independent mice). **b**, Scatter plot of $\log_{10}$ enrichment of each hydrophilic polypeptide versus its net charge as calculated from the total number of charged residues in its sequence. **c**, Scatter plot of $\log_{10}$ enrichment of each polypeptide versus the number of unique amino acids in its sequence. **d**, Each of 11 variants were individually expressed and purified by IMAC before being pooled (equal protein concentration) and purified *en masse* by SEC. The resulting nucleocapsid pool was then incubated in heparinized whole blood at 37 °C ($n = 3$ independent reactions per time point). RNA was recovered at the indicated time points, and the fraction of each variant was determined by Illumina MiSeq counts taken at each time point. **e**. The same nucleocapsid pool used in **d** was injected retro-orbitally into mice ($n = 5$ biologically independent mice). RNA content was then assessed as in **d** using RNA isolated from tail vein draws at the indicated time points. All variants exhibit high stability in blood; however, the unmodified I53-50-v3 nucleocapsid (no polypeptide, blue) and a negative control polypeptide (ESESG, red) are cleared rapidly from circulation *in vivo*. Error bars represent s.e.m. The lower error bar for the pink data point at 15 min is not shown because its s.e.m. is nearly equivalent to its value.



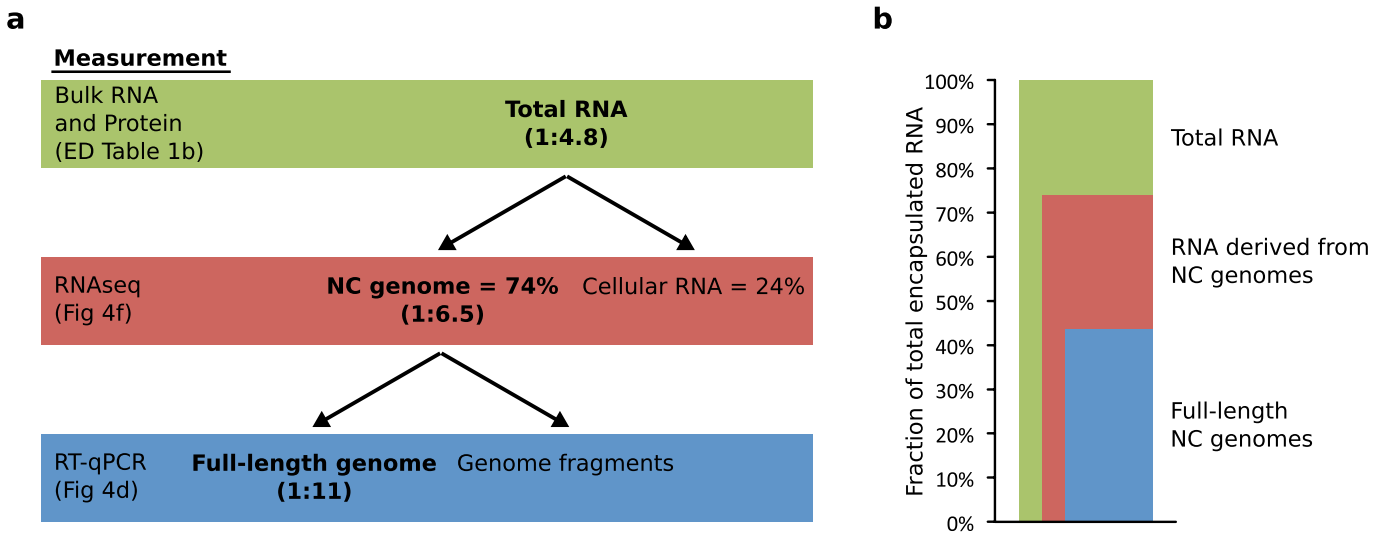
Extended Data Figure 4 | Evolution and performance of nucleocapsids with exterior surface mutations *in vitro* or *in vivo*. **a**, Heat map of log enrichments between the injected pool and RNA recovered from the tail vein 60 min later. Purple and orange indicate mutations that were depleted or enriched in the selected population, respectively. Blue squares and black dots indicate the I53-50-v3 starting sequence and I53-50-v4 selected sequence, respectively. Residues not in the designed combinatorial library are coloured grey. Note the strong enrichment of the E67K mutation and corresponding depletion of the native E67 allele. **b**, Design model of I53-50-v4. Colouring is as described in Fig. 1a. **c**, Four variants were tested: a consensus sequence based on the most common residue at each position after selection in mouse circulation (consensus, I53-50-v4), the full-length sequence with the greatest fold increase in population fraction (Most_enriched), the sequence with the most total counts (Top_count), and I53-50-v3 with only the E67K mutation (v3_E67K). Previous versions (I53-50-v1, I53-50-v2 and I53-50-v3) were also included as benchmarks. Each variant was individually expressed and purified by IMAC before being pooled (equal protein concentration) and purified *en masse* by SEC. The resulting nucleocapsid pool was then incubated in whole blood ($n = 3$ independent reactions per time point). RNA was recovered at the indicated time points, and the fraction of each variant was determined by Illumina MiSeq counts taken at each time point. **d**, The same nucleocapsid pool used in **c** was injected retro-orbitally into mice ($n = 5$ biologically independent mice). I53-50-v3 was evaluated with (v3) and without (v3H) the H6Q and H9Q mutations, and both variants were found to have similar behaviour. Error bars represent s.e.m.



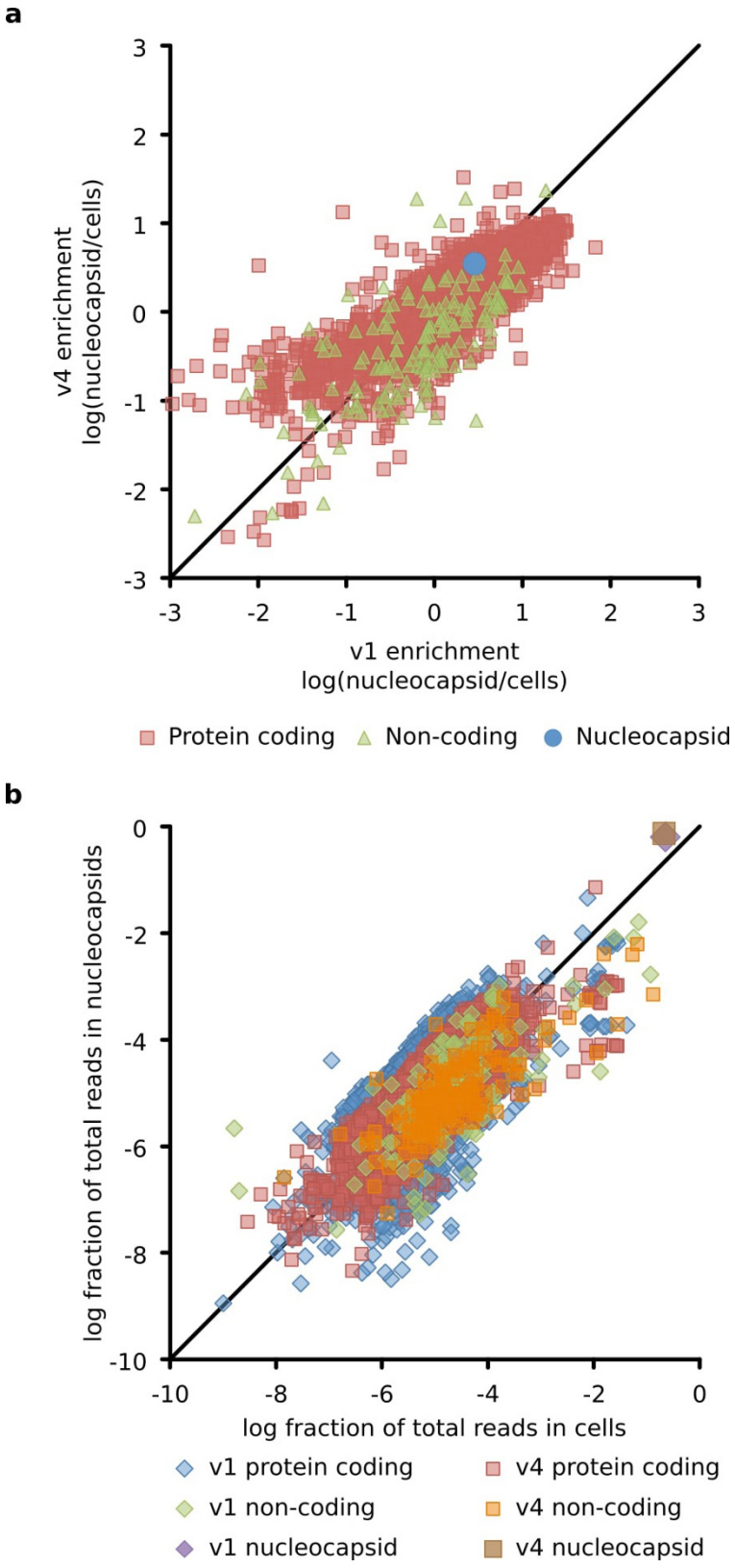
c

Bin	% total v4 particles
v0 dominant	12.7%
v4 dominant	64.0%
mixed	23.3%

Extended Data Figure 5 | Negative-stain transmission electron microscopy class averages. a, Two-dimensional class averages of I53-50-v0 (7,979 particles) and I53-50-v4 (7,120 particles) data sets showing the percentage of the total particles present in each class. I53-50-v4 nucleocapsids are on average denser than unfilled I53-50-v0 assemblies, especially near the inner surface of the capsid. **b**, All I53-50-v0 and I53-50-v4 particles from **a** were combined into a single set (15,119 particles), and 20 class averages were made from the combined data. Class averages were grouped into three bins (v0 dominant has $\leq$ 25% I53-50-v4, v4 dominant has $\geq$ 74% I53-50-v4, and mixed has the rest) and arranged from left to right with increasing fraction of I53-50-v4 particles (shown below each class). The v0 dominant classes appear more similar to the I53-50-v0 class averages in **a**, while the v4 dominant classes appear more similar to the I53-50-v4 class averages. The percentage of the complete I53-50-v4 dataset found in each class is shown above each class average. **c**, Table presenting the bins into which I53-50-v4 particles were assigned. We found that 64% of I53-50-v4 particles were present in the v4 dominant classes, which also seem to be more filled than the v0-dominant classes. Although TEM cannot determine the nature of the contents, encapsulated RNA is plausible.

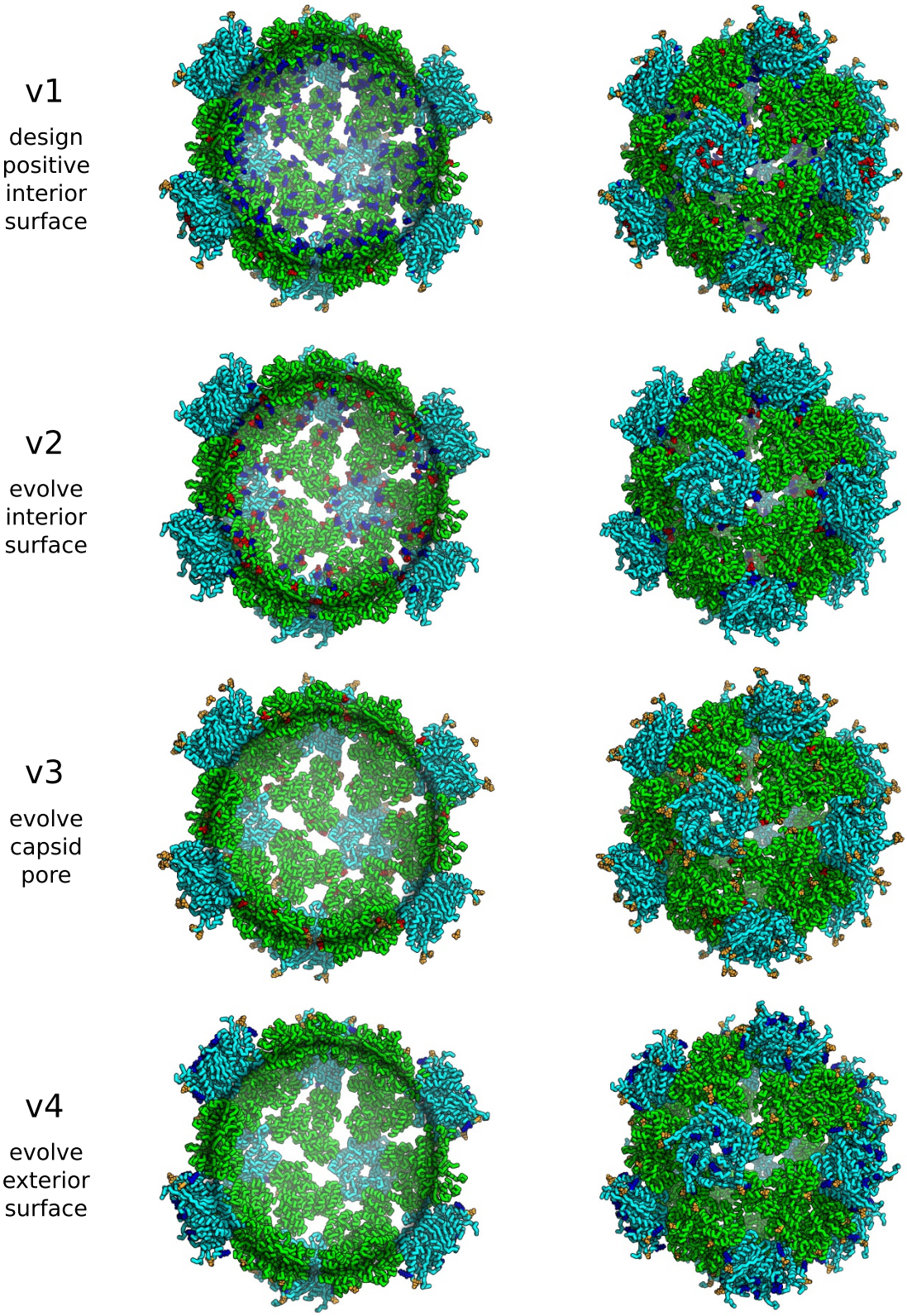


Extended Data Figure 6 | Summary of encapsulated RNA composition analysis. a, Flow chart explaining the relationship between bulk RNA measurements and RT–qPCR quantification. Bulk RNA measurements also account for cellular RNA and nucleocapsid genome fragments, whereas RT–qPCR only quantifies full-length genomes. Ratios of nucleocapsid genomes to capsids are based on these measurements and are reported in parentheses. **b**, Stacked bar plot describing the fractions of total encapsulated RNA that are full-length nucleocapsid genomes or fragments thereof.

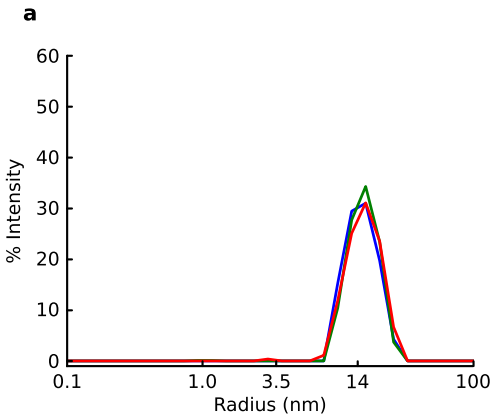


Extended Data Figure 7 | See next page for caption.

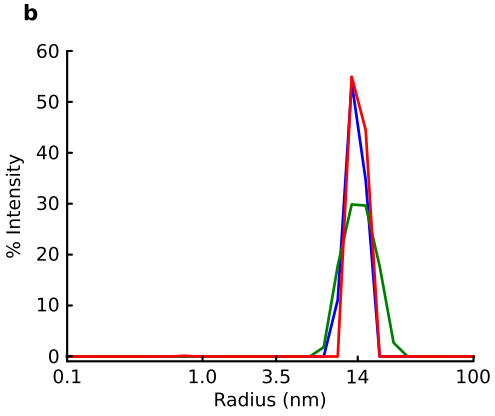
Extended Data Figure 7 | Packaging correlates strongly with expression level in producer cells. a, log enrichment (fraction packaged in nucleocapsid divided by fraction produced in cells) for I53-50-v4 versus I53-50-v1. Each point represents a unique RNA (red squares are protein coding mRNAs, green triangles are non-coding RNAs such as ribosomal RNA, and the blue circle is the nucleocapsid genomic RNA). No increase in specificity was observed over the course of evolution from the rationally designed I53-50-v1 to the *in vivo* circulating I53-50-v4. This is not surprising because no attempt was made to evolve increased specificity. The diagonal line is $y = x$. **b**, log fraction of total reads in nucleocapsids versus log fraction of total reads in cells shows that packaging correlates strongly with expression level (Pearson values for I53-50-v1 and I53-50-v4 are 0.83 and 0.86, respectively). Each point represents a unique RNA. The diagonal line is $y = x$. RNAs above the line are enriched in nucleocapsids, and RNAs below the line are depleted in nucleocapsids. Although the nucleocapsid genome is slightly enriched, its high packaging yield seems to arise because T7 RNA polymerase floods the cell with genomes, thereby increasing the chance that the capsid randomly packages the genome. Conversely, ribosomal RNA may be restricted from nucleocapsids because intact ribosomes are too large to be encapsulated. All data points represent the average of two independent biological replicates.



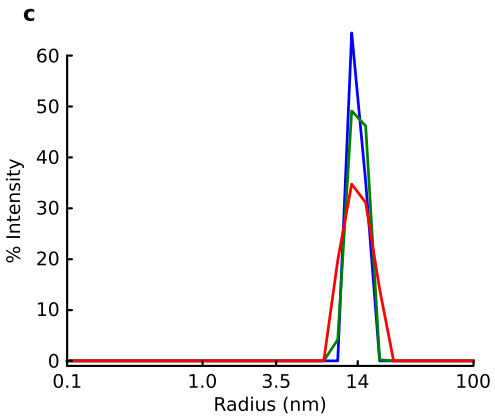
Extended Data Figure 8 | Design models of synthetic nucleocapsid versions 1 to 4. Trimer subunits are coloured green and pentamer subunits are coloured cyan. Mutations with respect to the previous version are coloured blue (increases in positive charge and/or decreases in negative charge (for example, E→N, N→K and E→K)), orange (no change in charge (for example, E→D, N→T and K→R)), or red (decreases in positive charge and/or increases in negative charge (for example, N→E, K→N, K→E)).



Replicate	Radius (nm)	Polydispersity %
v0-A	14.49	11.3
v0-B	14.49	9.44
v0-C	14.50	8.85



Replicate	Radius (nm)	Polydispersity %
v1-A	13.36	7.36
v1-B	13.65	3.57
v1-C	13.60	7.07



Replicate	Radius (nm)	Polydispersity %
v4-A	13.47	2.35
v4-B	13.47	5.73
v4-C	13.61	3.41

Extended Data Figure 9 | Dynamic light scattering of nucleocapsids. Dynamic light scattering (DLS) was performed on synthetic nucleocapsids and fitted with regularization analysis, confirming uniform populations of nucleocapsids around the expected size. **a**, I53-50-v0 has a C-terminal histidine tag. **b**, I53-50-v1 has an N-terminal histidine tag that was cleaved before DLS. **c**, I53-50-v4 has an N-terminal histidine tag that was cleaved before DLS. The experiment was independently repeated three times (data for independent replicates are shown in the figure).

Extended Data Table 1 | Amino acid substitutions and quantification of nucleocapsid genomes

a

Version	Changes in trimer with respect to previous version	Changes in pentamer with respect to previous version
I53-50-v1	T126D, E166K, S179K, T185K, A195K, E198K	Y9H, A38R, S105D, D122K, D124K
I53-50-v2	K179N, K185N, E188K	E24F, K124N, H126K
I53-50-v3	K9R, K11T, K61D	H6Q, H9Q
I53-50-v4	E74D	D39K, D43E, E67K

b

Sample	Protein (μg ml^{-1})	Total encapsulated RNA (ng μl^{-1})	Capsids (M)	Total RNA (M)	Capsids/ Genome equiv.	% RNA is NC genome	Capsids/ genome
I53-50-v0 (rep 1)	184	bd	7.4E-08	bd	bd	bd	bd
I53-50-v0 (rep 2)	188	bd	7.6E-08	bd	bd	bd	bd
I53-50-v1 (rep 1)	436	14.0	1.7E-07	3.0E-08	5.7	64%	8.9
I53-50-v1 (rep 2)	504	12.3	2.0E-07	2.6E-08	7.5	64%	11.7
I53-50-v4 (rep 1)	217	8.0	8.5E-08	1.7E-08	5.0	74%	6.7
I53-50-v4 (rep 2)	217	8.7	8.5E-08	1.9E-08	4.6	74%	6.2

a, All amino acid substitutions made for each version relative to the previous version. **b**, Genomes per nucleocapsid by bulk RNA and protein measurements. bd, below detection. Capsid molecular masses: v0 = 2,479.440 kDa, v1 = 2,544.300 kDa, v4 = 2,539.320 kDa. Total RNA was calculated by assigning nucleocapsid genome molecular mass to total encapsulated RNA: v0 = 443.618 kDa, v1 = 464.212 kDa, v4 = 463.971 kDa. Genome equivalents relates to total encapsulated RNA (including cellular RNA). The percentage of RNA mapping to the nucleocapsid genome was determined by RNA-seq analysis.

nature research

Corresponding author(s): David Baker

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

Three replicates were performed for in vitro selections. Five biological replicates were performed for all in vivo experiments to control for mouse-to-mouse variability. These replicates resulted in SEMs that were small enough that clear differences in nucleocapsid performance were observed.

2. Data exclusions

Describe any data exclusions.

One animal was excluded during hydrophilic peptide library evolution due to a failed retro-orbital injection.

3. Replication

Describe whether the experimental findings were reliably reproduced.

All experimental findings were reliably reproduced.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Mice were randomly assigned to each group.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

No blinding was performed.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☒	☐	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	Electron microscopy class averaging was performed using EMAN2. DLS data was analyzed using DYNAMICS v7 (Wyatt). Paired-end reads were combined using the read_fuser script from the enrich package. Variant counts for evolution experiments were analyzed using a custom python script that is available on github. RNAseq analysis was performed using the following published software: bcl2fastq script from Illumina, HISAT2, SAMtools, and StringTie.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	We intend to deposit our nucleocapsid variants at addgene.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used in this study.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used in the study.
b. Describe the method of cell line authentication used.	N/A
c. Report whether the cell lines were tested for mycoplasma contamination.	N/A
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	N/A

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	6 – 8 week old Balbc mice were selected randomly and retro-orbitally injected with 150 µL of synthetic nucleocapsids. All mice were female to minimize any unknown variability in tissue distribution bias attributed to animal sex. The Institutional Animal Care and Use Committee (IACUC) at the University of Washington authorized all animal work in accordance with ethical animal use and regulations.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	This study did not involve human research participants.