INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Farheen N, Thattai M. 2019 Frustration and fidelity in influenza genome assembly. *J. R. Soc. Interface* **16**: 20190411. http://dx.doi.org/10.1098/rsif.2019.0411

Received: 15 June 2019
Accepted: 15 October 2019

Subject Category:
Life Sciences–Physics interface

Subject Areas:
biophysics, computational biology

Keywords:
segmented virus, influenza, self-assembly, network evolution

Author for correspondence:
Mukund Thattai
e-mail: thattai@ncbs.res.in

†Present address: Brandeis University, Waltham, MA 02454, USA.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.4706870.

THE ROYAL SOCIETY PUBLISHING

Frustration and fidelity in influenza genome assembly

Nida Farheen[1,†] and Mukund Thattai[2]

[1]Indian Institute of Science Education and Research, Pune 411008, India
[2]Simons Centre for the Study of Living Machines, National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bangalore 560065, India

MT, 0000-0002-2558-6517

The genome of the influenza virus consists of eight distinct single-stranded RNA segments, each encoding proteins essential for the viral life cycle. When the virus infects a host cell, these segments must be replicated and packaged into new budding virions. The viral genome is assembled with remarkably high fidelity: experiments reveal that most virions contain precisely one copy of each of the eight RNA segments. Cell-biological studies suggest that genome assembly is mediated by specific reversible and irreversible interactions between the RNA segments and their associated proteins. However, the precise inter-segment interaction network remains unresolved. Here, we computationally predict that tree-like irreversible interaction networks guarantee high-fidelity genome assembly, while cyclic interaction networks lead to futile or frustrated off-pathway products. We test our prediction against multiple experimental datasets. We find that tree-like networks capture the nearest-neighbour statistics of RNA segments in packaged virions, as observed by electron tomography. Just eight tree-like networks (of a possible 262 144) optimally capture both the nearest-neighbour data and independently measured RNA–RNA binding and co-localization propensities. These eight do not include the previously proposed hub-and-spoke and linear networks. Rather, each predicted network combines hub-like and linear features, consistent with evolutionary models of interaction gain and loss.

1. Introduction

The influenza virus is notable in having a segmented genome, spread across eight RNA strands [1]. This segmented organization directly impacts influenza biology and evolution. The negative-sense genomic RNA is transcribed in an infected cell's nucleus to form positive-sense RNA, which undergoes both translation (to synthesize viral proteins) and replication (to form new genomic RNA). Segmentation allows genomic re-assortment, contributing to the emergence of novel influenza strains [2,3]. However, it also complicates the assembly and packaging of the complete viral genome into new virions [4]. Genomic RNA strands associate with specific viral proteins (nucleoprotein (NP) and the polymerase complex PB2, PB1 and PA) to form rod-like viral ribonucleoprotein segments (vRNPs). Over 10 000 vRNPs are synthesized within 4 hours post infection; these are packaged into nascent viral capsids at the plasma membrane, generating over 1000 virions per hour [5]. Since each vRNP segment encodes essential proteins, all eight segments must be assembled and packaged to generate an infectious virion [1,6]. Electron microscopy (EM) and fluorescence *in situ* hybridization (FISH) studies have shown that over 80% of new virions contain the complete genome, with each vRNP present in precisely one copy [7–9].

How does the influenza virus assemble its genome with such high fidelity? Genome assembly takes place as the vRNPs are trafficked to the plasma membrane [1,4]. The selective packaging model [6,7,10,11] posits that vRNPs bind to one another non-randomly via specific RNA–RNA and RNA–protein interactions; it is the resulting vRNP clusters that are packaged into virions. Consistent with

this idea, mutations in the conserved RNA terminal regions cause defects in genome packaging [12–16]. Genomic RNA strands are seen to physically bind *in vitro* and *in vivo* [17–19] via RNA base-pairing interactions [20–22] and the interactions mediated by vRNP-associated proteins [23–25]. In addition to the equilibrium RNA–RNA binding measured *in vitro*, live-cell imaging reveals that some inter-segment interactions are irreversible on the time scale of an infection [26]. Distinct vRNP segments are seen to co-localize inside the cytoplasm of infected cells [26,27]. EM tomography of virions shows that the eight vRNPs are arranged as parallel rods, with electron-dense regions that suggest tight lateral interactions [19].

These studies strongly support the existence of specific interactions between vRNP segments. However, they do not reveal the core interaction network that primarily drives genome assembly. Many possible interaction networks have been suggested, including a hub-and-spoke network (with a central 'master segment') and a linear network (looping to form a 'daisy chain') [10]. To our knowledge, none of these hypotheses have been rigorously tested against the measured interaction data. Here, we approach this problem by first exploring the influence of the inter-segment interaction network on the fidelity of genome assembly. We focus on the irreversible interactions, which create key decision points between correct and incorrect assembly pathways. Reversible and non-specific interactions [28] can play a role in stabilizing vRNP clusters already formed via irreversible interactions; we do not consider them here. By combining theoretical considerations with experimental datasets of virion structure, RNA–RNA binding and vRNP co-localization, we identify a handful of specific inter-segment interactions as the primary drivers of high-fidelity viral genome assembly.

2. Results

2.1. Routes to high-fidelity genome assembly

We first explore the dynamics of the selective packaging model, in which genome assembly is driven by specific inter-segment interactions. The efficiency of a self-assembly reaction is typically measured by its yield: the fraction of total input material that is correctly assembled. A better measure for our purposes is fidelity: the fraction of output clusters that contain precisely one copy of each of the eight genomic RNA segments. Fidelity corresponds to the experimentally measured fraction of new budding virions that are infectious, assuming that clusters are uniformly packaged into viral capsids.

Irreversible interactions correspond to energetically favourable contacts between specific binding sites on the vRNP segments; these interactions could be orientationally rigid or flexible. We assume binding sites are organized such that two vRNP segments of the same type cannot bind to one another, and a given type of vRNP segment can bind to at most one copy of a given other type of vRNP segment. To assemble eight vRNP segments, we require a minimum of seven interactions. Networks with precisely seven interactions are acyclic (tree like), while those with more than seven interactions must include cycles (closed paths). There are $8^6 = 262\,144$ tree-like networks (oeis.org/A000272) and over 250 million cyclic networks (oeis.org/A001187) that could potentially connect eight vRNP segments. Given an interaction network, we can model genome assembly as a stochastic chemical reaction (Methods). We start with a pool containing all vRNP types in equal amounts. We then allow the growth of clusters through pairwise aggregation, mediated by specific interactions between vRNP segments belonging to each cluster. We assume all allowed aggregation reactions occur by mass action with identical rate constants. Once no further aggregation events are possible, we calculate the final fidelity of the assembly reaction.

Cyclic interaction networks constitute the vast majority of possible networks. If the interactions in such networks are orientationally flexible, cycles will drive the futile synthesis of long polymers (figure 1*a*; *X*-*Y*-*Z*-*X*-*Y* …). Such futile reactions can be prevented by making the interactions orientationally rigid: the desired cluster with one copy of each segment is then stable because its binding sites are all either occupied or occluded. However, this introduces a new problem: once all the monomeric vRNP segments are depleted, the assembly reaction gets stuck at frustrated oligomeric states (figure 1*b*; even though *Y* and *Z* can aggregate, *X*-*Y* and *X*-*Z* cannot since both copies of *X* compete to occupy the same position). This type of frustration has been observed in a broad class of self-assembly processes [29]. One way to prevent this is to use a fixed order of assembly, by tuning the aggregation rates (rapidly make *X*-*Y*, and then slowly make *X*-*Y*-*Z*). However, vRNPs appear to aggregate in many possible orders (though some might be preferred [26,27]). In this situation, cyclic interaction networks will always show low fidelity, owing to vRNPs being trapped in futile or frustrated off-pathway clusters. By contrast, tree-like networks of irreversible interactions will always show 100% fidelity (figure 1*c*,*d*; all aggregates reach state *Y*-*X*-*Z*), even when there is no fixed order, independent of the rates of aggregation, and whether interactions are flexible or rigid. This is surprising since tree-like interaction networks locally resemble cyclic interaction networks. A simple proof (Methods) shows that these results are completely general for tree-like and cyclic networks, regardless of the specific network topology. This strongly suggests that the core interaction network, which drives genome assembly, should be tree like.

2.2. Inferring interaction networks from experimental data

EM tomography shows that the eight vRNP segments (henceforth numbered 1–8; figure 2*a*) are arranged in a characteristic '7 + 1' pattern within virions, with seven vRNPs on the periphery surrounding a central vRNP (figure 2*b*). Using vRNP length as a proxy for segment identity, all but the three longest segments (1, 2, 3) can be distinguished from one another [8]. The relative positions of the segments are found to vary from virion to virion, suggesting that interactions are orientationally flexible (figure 2*b*; see Methods for segment positions in the 30 EM-tomography-observed virions [8]). However, certain vRNP pairs are more likely than others to appear as nearest neighbours. Segment interactions can be further resolved by the SPLASH (sequencing of psoralen crosslinked, ligated and selected hybrids) technique [30], which uses crosslinking and RNA sequencing to infer base-paired nucleotides in RNA complexes. SPLASH can be used to score the propensity of interaction between vRNPs in purified virions [20]. Both the EM tomography [8] and SPLASH [20] data are obtained for influenza strain A/WSN/33 (H1N1) in MDCK cells, allowing them to be directly compared.

Given a proposed interaction network, we assign it two types of scores based on experimental data (Methods). To

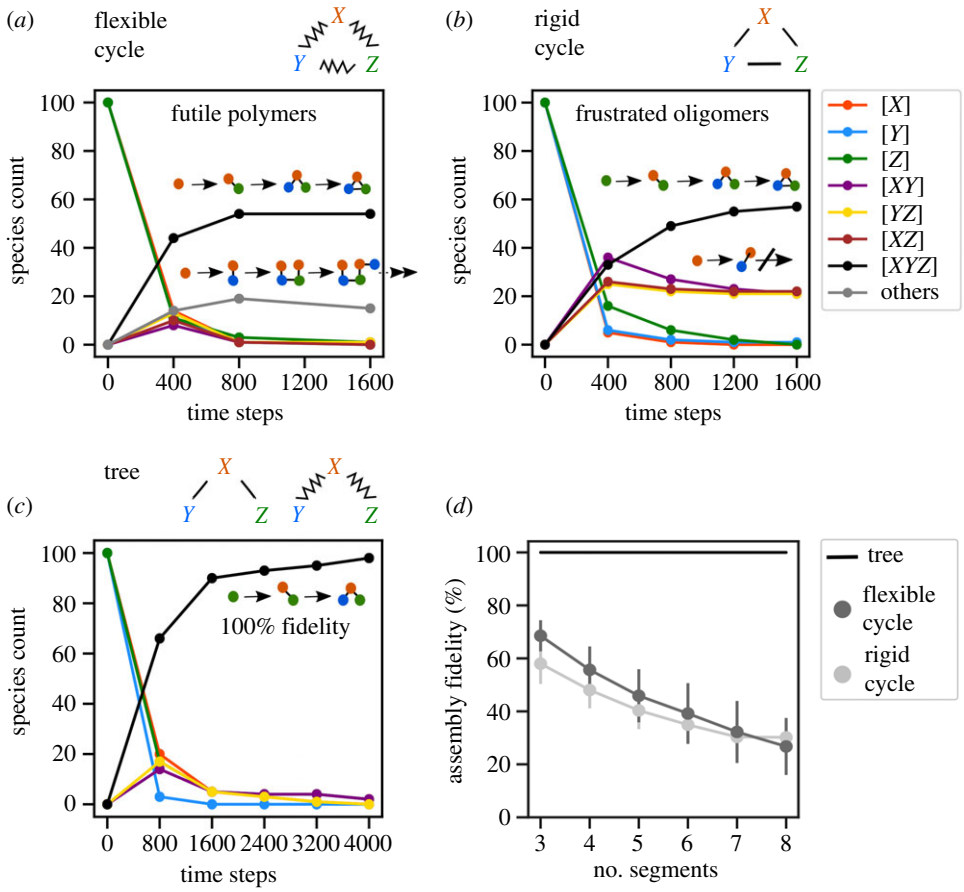

Figure 1. Stochastic simulation of genome assembly. We consider a toy model in which three segments (*X*, *Y*, *Z*) must assemble to form a desired cluster (*XYZ*). (*a–c*) Simulating time-dependent aggregation dynamics. The underlying interaction networks are indicated above each plot (zigzag edges show orientationally flexible interactions, straight edges show orientationally rigid interactions). The simulation starts with 100 copies of each segment and continues until no further reactions are possible. Counts of each possible reaction species over time from a single simulation are shown as curves of different colours (legend). Schematics show a few sample aggregation reactions, focusing on clusters that are present at long times. The fidelity is defined as the fraction of final clusters that are of the desired type (*XYZ*, black curves). (*a*) Flexible cyclic network. Flexible interactions allow the futile synthesis of long polymers. Segments are trapped within these futile clusters, reducing fidelity. (*b*) Rigid cyclic network. Once all monomeric segments are depleted, the remaining oligomers cannot bind to one another since identical segments compete to occupy the same spatial position. This is known as frustration. Segments are trapped within these frustrated oligomers, reducing fidelity. (*c*) Tree-like network. Once sufficient time has passed, all segments aggregate to produce the desired cluster *XYZ*, and no other cluster types are present. This reaction has 100% fidelity. (*d*) We consider systems with varying numbers of segments, whose interaction network is either a tree or a single long cycle. We compute the mean (±s.d.) fidelity over 500 stochastic simulations. Both flexible and rigid cycles show decreasing fidelity with an increasing segment number. Trees always show 100% fidelity. These results generalize to all tree-like and cyclic networks, regardless of the size and topology (proof in Methods). (Online version in colour.)

obtain the 'RNA contacts' score, we simply sum SPLASH scores for all the bonds present in the interaction network (figure 2*c*,*d*: left; there are two scores corresponding to two SPLASH replicates). A higher 'RNA contacts' score indicates better agreement with the SPLASH data. The 'stretched bonds' score is more involved, since there are six possible assignments of segments 1, 2 and 3 for each virion observed by EM tomography (figure 2*c*,*d*: right). For a given virion, we select the assignment that permits the most bonds between nearest neighbours; to obtain the 'stretched bonds' score, we then sum the number of stretched (non-nearest neighbour) bonds across the 30 observed virions. A lower 'stretched bonds' score indicates better agreement with the virion nearest-neighbour data; an interaction network that captures all the observed nearest-neighbour occurrences would have a score of zero. Note that a cyclic network, compared with any tree-like sub-network, will have a better (greater than or equal to) 'RNA contacts' score and a worse (greater than or equal to) 'stretched bonds' score.

We first calculated 'stretched bonds' scores for every possible cyclic and tree-like interaction network. The best overall network was a tree (figure 3*a*) with a 'stretched bonds' score of 34 (13 virions had two stretched bonds, eight virions had one and nine virions had none). The 131 best networks were tree like, while all cyclic networks had 'stretched bonds' scores of 44 or worse. To estimate the statistical significance of this result, we generated 1000 shuffled datasets, in each of which the peripheral segments of all 30 EM-tomography-observed virions were randomly permuted. We then determined the best tree-like network for each shuffled dataset (figure 3*a*). The most common network thus found was the hub-and-spoke network, with a 'stretched bonds' score of 48, seen 737 out of 1000 times (by definition the hub-and-spoke score does not vary when virions are peripherally shuffled). Across 1000 shuffled datasets, no tree had a 'stretched bonds' score of less than 42. We can conclude that far more than 1000 random draws are required before we find a shuffled virion dataset with a tree matching the 'stretched bonds' score of 34 found for real virions. This proves both that segment organization in real virions is highly non-random (p-value less than 0.001 for the null hypothesis that peripheral segments are randomly ordered)

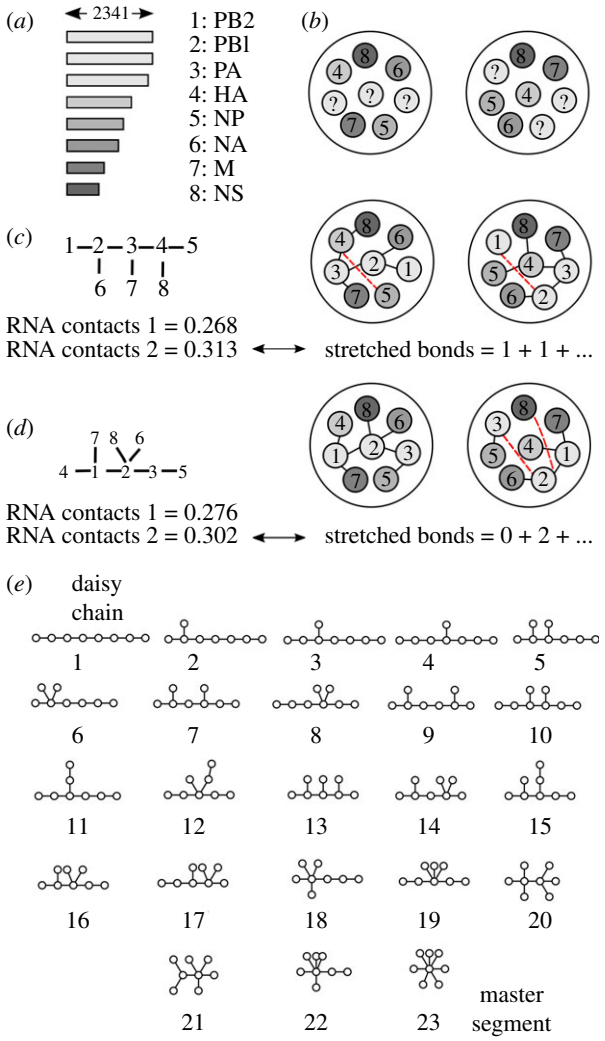

Figure 2. Genome packaging and inter-segment interaction networks. (*a*) The influenza genome is made up of eight distinct RNA segments. Bar lengths show the number of nucleotides on each segment. Segments are conventionally numbered in order of decreasing length and named according to the viral proteins they encode. (*b*) Each RNA segment is associated with viral proteins (NP, PB2, PB1 and PA) to form a rod-like viral ribonucleoprotein (vRNP) segment. Segments are packaged within virions as parallel rods in a '7 + 1' arrangement (as seen along the rod axes). EM tomography cannot distinguish between segments 1, 2 and 3; these are labelled '?'. The precise segment arrangement varies from virion to virion; two actual examples are shown here, out of 30 measured virions (Methods). (*c*,*d*) Given an interaction network (left), we assign it two types of scores (Methods). The 'RNA contacts' score is the RNA–RNA interaction propensity measured by SPLASH, summed over each inter-segment bond in the network (two 'RNA contacts' scores correspond to two SPLASH replicates). Higher 'RNA contacts' scores indicate better agreement with the SPLASH data. The 'stretched bonds' score is the total number of stretched (non-nearest neighbour) bonds, summed across 30 virions observed by EM tomography. For each virion, we use the assignment of segments 1, 2 and 3 that minimizes the number of stretched bonds (red lines). Lower 'stretched bonds' scores indicate better agreement with the nearest-neighbour data. (*e*) The 23 possible unlabelled tree-like network topologies for eight segments: topology 1 is the linear network which forms a 'daisy chain'; topology 23 is the hub-and-spoke network with a central 'master segment'. (Online version in colour.)

and that a tree-like interaction network captures the relevant nearest-neighbour statistics.

For the remainder of our analysis, we focus on the 262 144 possible tree-like networks. These fall into 23 topological classes (figure 2*e*): topology 1 is the linear network (daisy chain) and topology 23 is the hub-and-spoke network (master segment) [10]. We can represent each tree-like network as a point on a scatter plot (figure 3*b*), with the horizontal axis showing its 'stretched bonds' score and the vertical axis showing its 'RNA contacts' score (points corresponding to the two SPLASH replicates are labelled X and Y). Trees that are higher and to the left dominate (i.e. are strictly better than) trees that are lower and to the right. Trees not dominated by any other tree, which jointly optimize the two scores, constitute the 'Pareto front' (the upper-left envelope of the scatter plot). The leftmost tree on the Pareto front (X_1/Y_1, also shown in figure 3*a*) has the best possible 'stretched bonds' score of 34 but a poor 'RNA contacts' score. If we try to improve the 'RNA contacts' score, the 'stretched bonds' score gradually worsens until the shoulder value of 42 (Y_6) past which it rapidly worsens. This (combined with the fact that 997 out of 1000 shuffled datasets have 'stretched bonds' scores above 42; figure 3*a*) suggests that we should only consider Pareto trees with 'stretched bonds' scores of 42 or less. There are only eight such tree-like networks (figure 3*c*), a massive reduction from the 262 144 initial possibilities. These eight Pareto trees have a median diameter of four interactions (compared with two for hub-and-spoke and seven for linear) and a median max-degree of five interactions (compared with seven for hub-and-spoke and two for linear). Segments 5, 6 and 8 are always tips; one among 1, 2 or 3 is always a hub. Across these Pareto trees, all but one interaction connects the set {1, 2, 3} with the set {4, 5, 6, 7, 8}.

2.3. Comparison with segment co-localization during infection

We reasoned that segment pairs predicted to directly interact are more likely to co-localize during infection. The co-localization index of all 28 segment pairs has been measured at 8 hours post infection using FISH probes [26]. We compared this co-localization index against the number of times each of the 28 segment pairs is observed in the predicted Pareto trees (figure 3*f*). We find that these two quantities are significantly correlated (Methods): their Kendall rank correlation coefficient is 0.40 (p-value = 0.0034). This provides strong independent support for our prediction that the core interaction network is tree like. (The same co-localization data have also been used to probe the time ordering of aggregation [27], which we do not address here.)

2.4. Evolution of interaction networks

Our theoretical model of high-fidelity assembly suggests that the interaction network should be tree like, but does not select a preferred tree topology. Genetic re-assortment studies show that inter-segment interactions evolve as the viral strains diverge [2]. This process can make certain interaction network topologies more abundant than others in segmented viral genomes.

Consider first a simple model (figure 4*a*) in which a tree-like interaction network is grown by connecting new segments to randomly chosen existing segments. The probability of obtaining a certain tree topology can be calculated by enumerating all possible growth orders starting from a single segment; this is equivalent to counting all consistent ways to label the segments of a given tree from the oldest to the

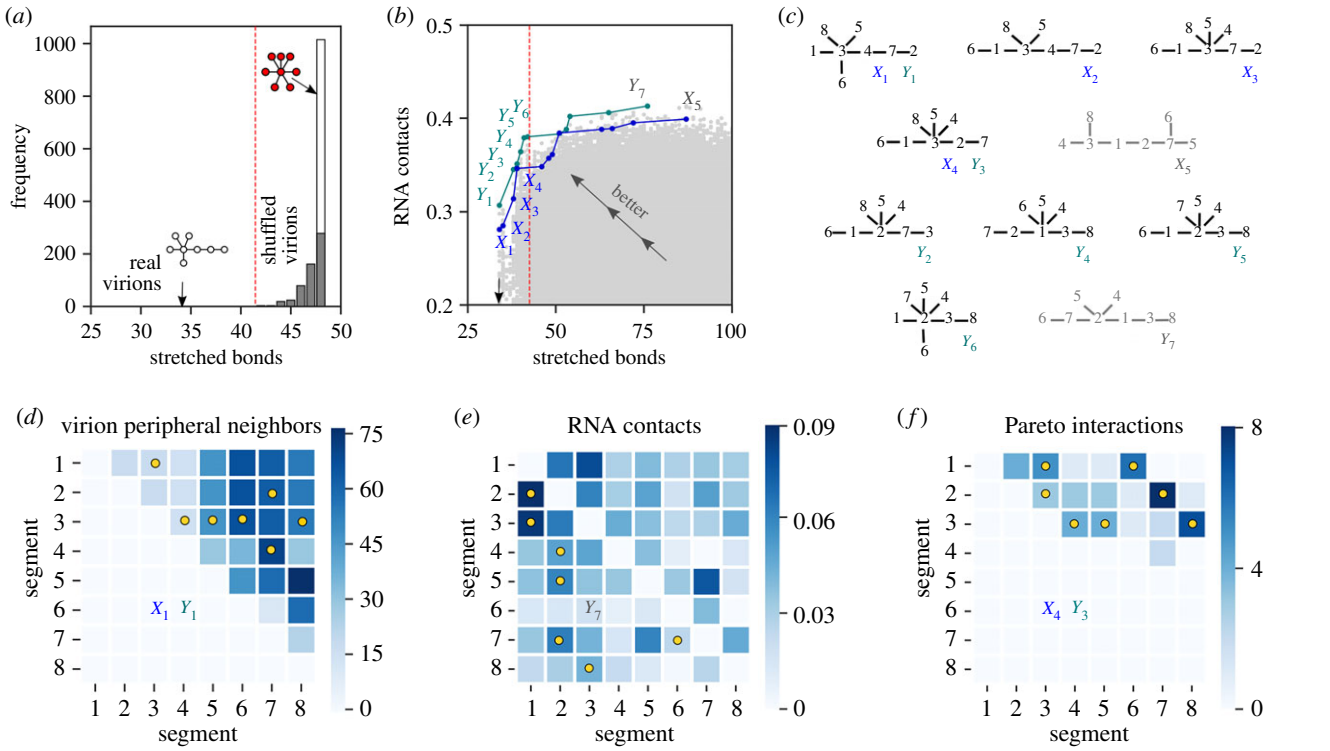

Figure 3. Inferring interaction networks from experimental data. (*a*) We calculated 'stretched bonds' scores (based on segment nearest-neighbour occurrences within 30 virions; figure 2*c*,*d*; right) for all possible interaction networks on eight segments (over 250 million networks). The best network was a tree (topology 18; white network) with a score of 34. We repeated the same analysis for tree-like networks alone (262 144 networks) using synthetic datasets in which peripheral segments of each virion were randomly shuffled. The histogram shows the distribution of best scores obtained for 1000 synthetic datasets, each containing 30 shuffled virions; the lowest score on shuffled data was 42 (seen three out of 1000 times; the vertical red line shows this cut-off value). The white bar shows instances in which the best network was hub-and-spoke (topology-23; red network) with a score of 48 (seen 737 out of 1000 times). (*b*) Scatter plot of 'stretched bonds' scores and 'RNA contacts' scores for all 262 144 possible tree-like networks on eight segments. Each grey point corresponds to a single tree; each tree is represented by two points corresponding to two SPLASH replicates (figure 2*c*,*d*: left). Trees that are higher and to the left show better agreement with experimental data. The 'stretched bonds' score and 'RNA contacts' score are jointly optimized by trees on the Pareto front, which form the upper-left envelope of the scatter plot. We show two Pareto fronts (blue, *X* labels; green, *Y* labels) corresponding to two SPLASH replicates. The black arrow shows the best 'stretched bonds' score of 34; the vertical red line shows the cut-off 'stretched bonds' score of 42. (*c*) The eight Pareto trees with 'stretched bonds' scores of 42 or less, corresponding to labels in figure 3*b*. We also show the two right-most trees on the Pareto front, which have the best 'RNA contacts' scores (grey). (*d*–*f*) Inter-segment associations seen in different data sources. For context, yellow dots show inter-segment bonds in specific Pareto trees, indicated by their labels from figure 3*c*. (*d*) Frequencies with which segment pairs are observed as peripheral nearest neighbours across 30 virions; each virion is represented by all six possible assignments of segments 1, 2 and 3, so the maximum possible score is 180. Since the data are symmetric, only the upper-triangular portion is shown. (*e*) Inter-segment RNA–RNA contact propensities measured by SPLASH. The upper-triangular and lower-triangular portions represent SPLASH scores for two different replicates. (*f*) The number of times each inter-segment interaction is observed among all eight Pareto trees with 'stretched bonds' scores of 42 or less. Since the data are symmetric, only the upper-triangular portion is shown. The Pareto interaction map is much more sparse than the virion peripheral neighbour map and the RNA contact map. (Online version in colour.)

newest (Methods). Under this 'gain-only' model, the probability is 4.0×10^{-4} for a hub-and-spoke network, 0.013 for a linear network and 0.10 for the most probable topology-8 network (figure 4*c*). A more realistic scenario is one in which interactions can be gained or lost (figure 4*b*). If we start with a viral population in which a given tree topology is fixed, a gain-plus-loss event can generate a new tree topology that has a chance of sweeping to fixation (we assume that all trees have equal fitness, all cyclic networks have low but non-zero fitness and all disconnected networks have zero fitness). This can be modelled as a Markov chain whose equilibrium distribution gives the probability that the population has a given tree topology (Methods). Under this 'gain/loss' model, the probability is 2.2×10^{-5} for a hub-and-spoke network, 0.088 for a linear network and 0.16 for the most probable topology-3 network (figure 4*c*)

While these abstract models cannot capture the evolutionary dynamics of real viral populations, they do make certain robust predictions. We can be confident that the highly symmetric hub-and-spoke network is extremely unlikely to arise via the random gain and loss of interactions, unless it is specifically selected. More generally, we ought to expect interaction networks that combine both hub-like and linear features, rather either purely hub-and-spoke or purely linear networks. The five topologies represented among the eight Pareto trees are entirely consistent with this expectation (figure 4*c*).

3. Discussion

The mechanism by which the influenza virus packages its genome is a natural instance of a self-assembly process. There is growing interest in exploring the general principles of self-assembly across contexts: in complex multi-component biological systems such as ribosomes [31] and viruses [11]; and in synthetic systems such as colloidal aggregates and DNA tiles [32]. Kinetic assembly processes, in which aggregation reactions are driven out of equilibrium, allow greater control

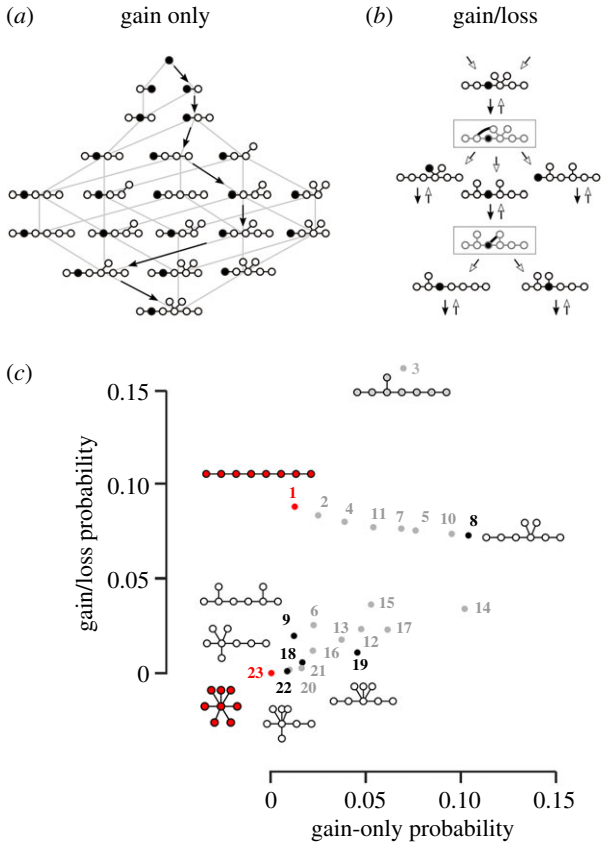

Figure 4. Evolution of tree-like interaction networks. (*a*) Under the 'gain-only model' new segments are attached to randomly selected existing segments (black arrowhead = segment gain). The probability of obtaining given tree topology is calculated by summing all possible growth paths from all possible starting segments. In this example, there are 42 possible growth paths from the initial segment to the final tree. Similar calculations were done for all distinct initial segments and final trees. (*b*) Under the 'gain/loss' model, trees evolve through repeated gain-plus-loss events (black arrowhead = interaction gain, white arrowhead = interaction loss) and rapidly transition via cyclic states (grey boxes) to new tree topologies. This is a limiting case of evolution under successive population-genetic sweeps. It can be represented as a Markov chain that encodes transition rates between the 23 possible tree topologies. The probability of obtaining a given tree topology is found from the equilibrium distribution of the Markov chain. (*c*) Probabilities of tree topologies under the 'gain-only' and 'gain/loss' evolutionary scenarios. Numbers correspond to the 23 tree topologies shown in figure 2*e*. We show the five topologies represented among the eight Pareto trees (white network and figure 3*c*). We also highlight the linear (daisy chain) and hub-and-spoke (master segment) topologies (red networks) as well as high-probability topology 3 (grey network). The hub-and-spoke network is exceedingly unlikely under either evolutionary model. (Online version in colour.)

and higher yield of desired final products [31,32]. However, out-of-equilibrium irreversible reactions can also lead the system down futile or frustrated paths [29]. Here, we identify a simple design principle—a tree-like network of irreversible interactions—that provably achieves perfect fidelity.

Our approach is distinct from the challenge of finding the time ordering of aggregation [27], which by definition is a tree whose nodes are clustered states and whose directed edges are reactions. The objects of our study are networks that could be cyclic or tree like, whose nodes are segments and whose undirected edges are physical interactions. Our stochastic growth model (figure 1) does not select or require a single growth order; all growth orders consistent with the interaction network are permitted. The prediction that the network is tree like then follows from the requirement of fidelity. This is due to a surprising property of tree-like interaction networks: they inexorably funnel growing aggregates towards the desired final product. Moreover, this is achieved at the highest possible rate, since every aggregation reaction is productive (unlike equilibrium binding/unbinding reactions).

The purely theoretical preference for a tree-like interaction network allows us to extract useful information from the measured interaction propensities of vRNP segments. Since both our primary experimental datasets (EM tomography of nearest neighbours [8] and SPLASH RNA–RNA interaction measurements [20]) correspond to packaged virions, they cannot distinguish the irreversible interactions that drive genome assembly from weaker ones that stabilize the final assembled genome, or even incidental nearest-neighbour associations. This is why the virion peripheral neighbour map (figure 3*d*) and RNA contacts map (figure 3*e*) are both fairly dense. By requiring the interaction network to be tree like and focusing only on Pareto trees, we remove such false positives to predict a sparse set of core interactions (figure 3*f*).

Further studies are needed to select the correct tree from among the predicted Pareto trees. Mutational studies could directly probe interacting regions, such as by swapping packaging signals between different RNA segments [15]. For example, the virus grows poorly when packaging signals of segments 1, 3, 5 and 7 are replaced with that of segment 6, but grows well when the same swap is done for segments 2, 4 and 8 [15]. This is most consistent with Pareto tree X_3, in which segments 1, 3 and 7 are internal, while the remaining segments are tips. In the predicted Pareto trees, almost all direct interactions involve segments 1, 2 or 3 (which encode vRNP-associated polymerase proteins PB2, PB1 and PA), whereas there are almost no direct interactions between segments 4, 6 and 7 (which encode capsid proteins HA, NA and M). This could enhance the re-assortment of the immunogenic capsid protein varieties between different influenza strains [3]. We will need a principled approach to incorporate information from these and other disparate data sources. Nevertheless, multiple lines of evidence (measurements of segment co-localization [26,27]; non-random nearest-neighbour propensities in packaged virions [8]; and viral growth rates upon swapping packaging signals [15]) support our hypothesis that the core interaction network underlying influenza genome assembly is tree like. Our results not only provide insight into the dynamics of infection, but also have implications for understanding how new influenza strains emerge via genomic re-assortment and evolution.

4. Methods

4.1. Stochastic simulation of genome assembly

We model genome assembly using a Monte Carlo simulation with discrete time steps. An interaction network on M segment types is specified, with either orientationally flexible or orientationally rigid bonds. We assume two segments of the same type cannot bind to one another, and a given segment type can bind to at most one copy of a given other segment type. We initialize the simulation with N copies of each segment type. As the simulation proceeds, the segments irreversibly aggregate into clusters. Within each cluster, we form satisfied bonds between every pair of segments that can interact. At each time step, we select two of the clusters at random. We aggregate them into a single large cluster if a pair of segments, one from each cluster,

has an unsatisfied bond. For the rigid bonds case, we must also check that the two clusters do not both include the same segment type, since these would compete to occupy the same spatial position. We continue the simulation until no further aggregation events are possible. The final fidelity is then calculated as the fraction of clusters that are of the desired type, containing precisely one copy of each segment type.

4.2. Proof that tree-like networks have 100% fidelity

The proof is by contradiction. We are given a flexible or rigid tree-like interaction network on M segment types, and given N copies of each segment type. The maximal cluster has exactly one copy of each of the M segment types. Suppose the final fidelity at the end of the aggregation process is less than 100%. There must be at least one cluster C with fewer than M segment types. Since the interaction network is connected, there must be at least one pair of segment types X and Y that interact, such that X is in C and Y is not in C. So, the copy of X in C has an unsatisfied bond with Y. Since no further aggregation events are possible, every copy of Y either has a satisfied bond with a copy of X or (for the rigid bonds case) has an unsatisfied bond with X but belongs to a cluster C' that cannot aggregate with C. This implies that C and C' both include some segment type Z. Any such Z is (directly or indirectly) connected to X in C and to Y in C'. Since we already know that X and Y can directly interact, this would mean the interaction network contains a cycle, which we know is not the case. Therefore, no such cluster C' exists. The only remaining possibility is that all N copies of Y have a satisfied bond with X. Since a given segment type can bind to at most one copy of a given other segment type, this implies that all N copies of X have a satisfied bond with Y, contradicting our assertion that there is at least one copy of X that has an unsatisfied bond with Y. This completes the proof.

4.3. Proof that cyclic networks have less than 100% fidelity

We are given a flexible or rigid cyclic interaction network on M segment types, and given N copies of each segment type. To show that the average final fidelity is less than 100%, it is sufficient to show that there is at least one possible trajectory of the stochastic aggregation process with final fidelity less than 100%. First, we treat the flexible case. Consider any length-L cycle in the interaction network, containing a pair of segment types X and Y that interact. We can grow two linear $(L-1)$-sized clusters C and C' by single-segment-addition reactions so that C contains one copy each of all L segment types except Y, and C' contains one copy each of all L segment types except X. C and C' can then aggregate via the X-Y interaction to form a futile cluster containing two copies each of $(L-1)$ segment types. This ensures the final fidelity is less than 100%. Next we treat the rigid case. Since a given segment type can bind to at most one copy of a given other segment type, the minimal length of a cycle is 3. Consider any length-3 cycle in the interaction network, containing segments X, Y and Z (the proof for a length-L cycle for any $L \geq 3$ is identical). If N is odd, first form a single cluster XYZ by single-segment-addition reactions, leaving an even number of copies of each segment type X, Y and Z. If N is even, proceed to the next step. Let half of each segment type aggregate with its 'left neighbour' and the remaining half with its 'right neighbour' to form dimers XY, YZ, XZ. No further aggregation reactions are possible, since for any pair of dimers both will include the same segment type. The dimers are said to be frustrated. This ensures that the final fidelity of at least one trajectory is less than 100%, so the average final fidelity is less than 100%.

4.4. Scoring interaction networks

'RNA contacts' score (figure 2*c*,*d*: left): from SPLASH measurements, we obtain the total number of interactions observed between each segment pair (data from electronic supplementary material, table 2; datasets 3,4 in [20]). We normalize this across all segment pairs to obtain an RNA–RNA contact propensity. The 'RNA contacts' score of a given interaction network is the sum of contact propensities of the interacting segment pairs.

'Stretched bonds' score (figure 2*c*,*d*: right): from EM tomography, we obtain the arrangement of segments within packaged virions [8]. Segments 1, 2 and 3 cannot be distinguished by this method; for a given interaction network, we find the assignment of 1, 2 and 3 within each virion that has the most interactions between nearest neighbours. The 'stretched bonds' score of the network is the sum of the number of stretched (non-nearest neighbour) bonds across virions. The observed segment arrangements for 30 virions are shown below (data from figure 3; fig. S4 in [8]). Each string represents a single virion, starting with the central segment and moving clockwise over peripheral segments; '?' represents segments 1, 2 or 3.

4:87??65?, 4:86?57??, 4:857?6??, 4:85?6??, 4:8?75?6?,
4:8?75??6, 4:8??5?6, 4:8?6?75?, 4:8?6?7?5, 4:8?5?7?6,
4:8???65?, 4:8???5?6, ?:8754?6?, ?:8746?5?, ?:86547??,
?:8647??5, ?:86?57?4, ?:86?47?5, ?:8564??7, ?:856?7?4,
?:856?47?, ?:8547?6?, ?:85?6?74, ?:847?6?5, ?:84657??,
?:8?674?5, ?:8?6?745, ?:8?546?7, ?:8?5?746, ?:8?4756?

4.5. Comparison of predicted segment interactions with measured segment co-localization

The presence of specific segments in cellular RNA foci at 8 hours post infection has been measured using FISH probes [26]. The Pearson correlation coefficient of dual-probe intensities serves as a co-localization index for the corresponding segment pair. Here, we use the mean Pearson correlation coefficient for each segment pair, averaged across measurements of multiple cells (data from fig. 3 in [26]). We compare this against the number of occurrences of these segment pairs in the eight Pareto trees. Across the 28 segment pairs, we find a Kendall rank correlation coefficient of $\tau = 0.40$ between these two quantities. To estimate statistical significance, we carry out the same analysis using randomly permuted data. The τ values from the random data are greater than or equal to the observed value of 0.40 in 340 out of 100 000 random replicates (p-value = 0.0034).

4.6. Modelling the evolution of tree-like interaction networks

Gain-only model (figure 4*a*): we are given segments labelled 1, … , N (this is an arbitrary label unrelated to vRNP segment identity). At each step of the process, we take a tree with L segments and add segment $L+1$ to a randomly chosen segment in the tree. We start with segment 1 and stop when we reach a tree with N segments. We record the final tree topology, ignoring segment labels. The statistical weight of a given tree topology under this process can be calculated as follows. List all distinct segment types, up to isomorphism (e.g. the hub-and-spoke topology has only two distinct segment types). Pick a segment type, root the tree at this segment and label it 1. Calculate the number of distinct ways, up to isomorphism, to label the remaining segments 2, … , N such that label values always increase along every branch. Summing this number over all distinct root segments gives the statistical weight of a given tree topology. To get the gain-only probability, we normalize this by the combined statistical weight of all possible tree topologies.

Gain/loss model (figure 4*b*): we model transitions between N-segment tree topologies as a discrete Markov chain. We are

given a starting tree with N arbitrarily labelled segments and $(N-1)$ inter-segment interactions. There are $(N-1)(N-2)/2$ possible new inter-segment interactions. Adding a single new interaction gives a network with a single cycle. Removing any interaction in the cycle gives back a tree (e.g. removing the newly added interaction gives back the original tree). Summing over all possible gain-plus-loss events and ignoring segment labels, we find the transition probability from the initial tree topology to any other tree topology. The gain/loss probability over tree topologies is the equilibrium distribution of this Markov chain.

Data accessibility. We have provided a supporting data file as electronic supplementary material.

Authors' contributions. M.T. conceived the project. N.F. carried out the analysis. N.F. and M.T. wrote the paper.

Competing interests. We declare we have no competing interests.

Funding. We received no funding for this study.

Acknowledgements. We acknowledge M. S. Madhusudhan for his critical feedback. We thank Mark Miller and other participants of the 'Entropy, Information and Order in Soft Matter' Program (ICTS/eiosm2018/08) at the International Centre for Theoretical Sciences for their useful discussions.

References

1. Bouvier NM, Palese P. 2008 The biology of influenza viruses. *Vaccine* **26**, D49–D53. (doi:10.1016/j.vaccine.2008.07.039)
2. McDonald SM, Nelson MI, Turner PE, Patton JT. 2016 Reassortment in segmented RNA viruses: mechanisms and outcomes. *Nat. Rev. Microbiol.* **14**, 448–460. (doi:10.1038/nrmicro.2016.46)
3. Lowen AC. 2018 It's in the mix: reassortment of segmented viral genomes. *PLoS Pathog.* **14**, e1007200. (doi:10.1371/journal.ppat.1007200)
4. Lakdawala SS, Fodor E, Subbarao K. 2016 Moving on out: transport and packaging of influenza viral RNA into virions. *Annu. Rev. Virol.* **3**, 411–427. (doi:10.1146/annurev-virology-110615-042345)
5. Frensing T, Kupke SY, Bachmann M, Fritzsche S, Gallo-Ramirez LE, Reichl U. 2016 Influenza virus intracellular replication dynamics, release kinetics, and particle morphology during propagation in MDCK cells. *Appl. Microbiol. Biotechnol.* **100**, 7181–7192. (doi:10.1007/s00253-016-7542-4)
6. Fujii Y, Goto H, Watanabe T, Yoshida T, Kawaoka Y. 2003 Selective incorporation of influenza virus RNA segments into virions. *Proc. Natl Acad. Sci. USA* **100**, 2002–2007. (doi:10.1073/pnas.0437772100)
7. Nakatsu S, Sagara H, Sakai-Tagawa Y, Sugaya N, Noda T, Kawaoka Y. 2016 Complete and incomplete genome packaging of influenza A and B viruses. *mBio* **7**, e01248-16. (doi:10.1128/mBio.01248-16)
8. Noda T, Sugita Y, Aoyama K, Hirase A, Kawakami E, Miyazawa A, Sagara H, Kawaoka Y. 2012 Three-dimensional analysis of ribonucleoprotein complexes in influenza A virus. *Nat. Commun.* **3**, 639. (doi:10.1038/ncomms1647)
9. Chou Yy, Vafabakhsh R, Doganay S, Gao Q, Ha T, Palese P. 2012 One influenza virus particle packages eight unique viral RNAs as shown by FISH analysis. *Proc. Natl Acad. Sci. USA* **109**, 9101–9106. (doi:10.1073/pnas.1206069109)
10. Hutchinson EC, von Kirchbach JC, Gog JR, Digard P. 2010 Genome packaging in influenza A virus. *J. Gen. Virol.* **91**, 313–328. (doi:10.1099/vir.0.017608-0)
11. Twarock R, Bingham RJ, Dykeman EC, Stockley PG. 2018 A modelling paradigm for RNA virus assembly. *Curr. Opin. Virol.* **31**, 74–81. (doi:10.1016/j.coviro.2018.07.003)
12. Kobayashi Y, Dadonaite B, van Doremalen N, Suzuki Y, Barclay WS, Pybus OG. 2016 Computational and molecular analysis of conserved influenza A virus RNA secondary structures involved in infectious virion production. *RNA Biol.* **13**, 883–894. (doi:10.1080/15476286.2016.1208331)
13. Hutchinson EC, Curran MD, Read EK, Gog JR, Digard P. 2008 Mutational analysis of cis-acting RNA signals in segment 7 of influenza A virus. *J. Virol.* **82**, 11 869–11 879. (doi:10.1128/JVI.01634-08)
14. Hutchinson EC, Wise HM, Kudryavtseva K, Curran MD, Digard P. 2009 Characterization of influenza A viruses with mutations in segment 5 packaging signals. *Vaccine* **27**, 6270–6275. (doi:10.1016/j.vaccine.2009.05.053)
15. Gao Q, Chou YY, Doganay S, Vafabakhsh R, Ha T, Palese P. 2012 The influenza A virus PB2, PA, NP, and M segments play a pivotal role during genome packaging. *J. Virol.* **86**, 7043–7051. (doi:10.1128/JVI.00662-12)
16. Gerber M, Isel C, Moules V, Marquet R. 2014 Selective packaging of the influenza A genome and consequences for genetic reassortment. *Trends Microbiol.* **22**, 446–455. (doi:10.1016/j.tim.2014.04.001)
17. Gavazzi C, Isel C, Fournier E, Moules V, Cavalier A, Thomas D, Lina B, Marquet R. 2012 An *in vitro* network of intermolecular interactions between viral RNA segments of an avian H5N2 influenza A virus: comparison with a human H3N2 virus. *Nucleic Acids Res.* **41**, 1241–1254. (doi:10.1093/nar/gks1181)
18. Gavazzi C, Yver M, Isel C, Smyth RP, Rosa-Calatrava M, Lina B, Moules V, Marquet R. 2013 A functional sequence-specific interaction between influenza A virus genomic RNA segments. *Proc. Natl Acad. Sci. USA* **110**, 16 604–16 609. (doi:10.1073/pnas.1314419110)
19. Fournier E *et al.* 2011 A supramolecular assembly formed by influenza A virus genomic RNA segments. *Nucleic Acids Res.* **40**, 2197–2209. (doi:10.1093/nar/gkr985)
20. Dadonaite B, Gilbertson B, Knight ML, Trifkovic S, Rockman S, Laederach A, Brown LE, Fodor E, Bauer DLV. 2019 The structure of the influenza A virus genome. *Nat. Microbiol.* **4**, 1781–1789. (doi:10.1038/s41564-019-0513-7)
21. Lenartowicz E *et al.* 2016 Self-folding of naked segment 8 genomic RNA of influenza A virus. *PLoS ONE* **11**, e0148281. (doi:10.1371/journal.pone.0148281)
22. Ruszkowska A, Lenartowicz E, Moss WN, Kierzek R, Kierzek V. 2016 Secondary structure model of the naked segment 7 influenza A virus genomic RNA. *Biochem. J.* **473**, 4327–4348. (doi:10.1042/BCJ20160651)
23. Moreira ÉA *et al.* 2016 A conserved influenza A virus nucleoprotein code controls specific viral genome packaging. *Nat. Commun.* **7**, 12861. (doi:10.1038/ncomms12861)
24. Li Z *et al.* 2009 Mutational analysis of conserved amino acids in the influenza A virus nucleoprotein. *J. Virol.* **83**, 4153–4162. (doi:10.1128/JVI.02642-08)
25. Bolte H, Rosu ME, Hagelauer E, García-Sastre A, Schwemmle M. 2019 Packaging of the influenza A virus genome is governed by a plastic network of RNA/protein interactions. *J. Virol.* **93**, e01861-18. (doi:10.1128/JVI.01861-18)
26. Lakdawala SS *et al.* 2014 Influenza A virus assembly intermediates fuse in the cytoplasm. *PLoS Pathog.* **10**, e1003971. (doi:10.1371/journal.ppat.1003971)
27. Majarian TD, Murphy RF, Lakdawala SS. 2019 Learning the sequence of influenza A genome assembly during viral replication using point process models and fluorescence in situ hybridization. *PLoS Comp. Biol.* **15**, e1006199. (doi:10.1371/journal.pcbi.1006199)
28. Venev SV, Zeldovich KB. 2013 Segment self-repulsion is the major driving force of influenza genome packaging. *Phys. Rev. Lett.* **110**, 098104. (doi:10.1103/PhysRevLett.110.098104)
29. Madge J, Bourne D, Miller MA. 2018 Controlling fragment competition on pathways to addressable self-assembly. *J. Phys. Chem. B* **122**, 9815–9825. (doi:10.1021/acs.jpcb.8b08096)
30. Aw JGA *et al.* 2016 In vivo mapping of eukaryotic RNA interactomes reveals principles of higher-order organization and regulation. *Mol. Cell* **62**, 603–617. (doi:10.1016/j.molcel.2016.04.028)
31. Bunner AE, Beck AH, Williamson JR. 2010 Kinetic cooperativity in *Escherichia coli* 30S ribosomal subunit reconstitution reveals additional complexity in the assembly landscape. *Proc. Natl Acad. Sci. USA* **107**, 5417–5422. (doi:10.1073/pnas.0912007107)
32. Murugan A, Zou J, Brenner MP. 2015 Undesired usage and the robust self-assembly of heterogeneous structures. *Nat. Commun.* **6**, 6203. (doi:10.1038/ncomms7203)