Learning the sequence of influenza A genome assembly during viral replication using point process models and fluorescence in situ hybridization

Timothy D. Majarian,[1,2*] Robert F. Murphy,[1-4#] and Seema S. Lakdawala[5#]

[1]Computational Biology Department, [2]Department of Biological Sciences, [3]Department of Biomedical Engineering, [4]Machine Learning Department, Carnegie Mellon University
[5]Department of Microbiology and Molecular Genetics, University of Pittsburgh School of Medicine

*Present address: Broad Institute of MIT and Harvard

#Corresponding authors: murphy@andrew.cmu.edu, Lakdawala@pitt.edu

Abstract

Within influenza virus infected cells, viral genomic RNA are selectively packed into progeny virions, which predominantly contain a single copy of 8 viral RNA segments. Intersegmental RNA-RNA interactions are thought to mediate selective packaging of each viral ribonucleoprotein complex (vRNP). Clear evidence of a specific interaction network culminating in the full genomic set has yet to be identified. Using multi-color fluorescence *in situ* hybridization to visualize four vRNP segments within a single cell, we developed image-based models of vRNP-vRNP spatial dependence. These models were used to construct likely sequences of vRNP associations resulting in the full genomic set. Our

results support the notion that selective packaging occurs during cytoplasmic transport and identifies the formation of multiple distinct vRNP sub-complexes that likely form as intermediate steps toward full genomic inclusion into a progeny virion. The methods employed demonstrate a statistically driven, model based approach applicable to other interaction and assembly problems.

Author Summary

Influenza virus consists of eight viral ribonucleoproteins (vRNPs) that are assembled by infected cells to produce new virions. The process by which all eight vRNPs are assembled is not yet understood. We therefore used images from a previous study in which up to four vRNPs had been visualized in the same cell to construct models that measure how well the subcellular distribution of one vRNP can be predicted from one or more other vRNPs. We used the likelihood of these models as an estimate of the extent of association between vRNPs and thereby constructed likely sequences of vRNP assembly that would produce full virions. Our work identifies the formation of multiple distinct vRNP sub-complexes that likely form as intermediate steps toward production of a virion. The results may be of use in designing strategies to interfere with virus assembly. We also anticipate that the approach may be useful for studying other assembly processes, especially for complexes with modest affinities and more components than can be visualized simultaneously.

Introduction

47 Influenza A virus, part of the *orthomyxoviridae* family, has a segmented genome of eight
48 distinct viral RNA segments coding at least 11 major proteins and several auxiliary
49 peptides. Most notable of the 11 viral proteins are hemagglutinin (HA) and neurominidase
50 (NA), the canonical *H* and *N* in influenza strain designation. A segmented genome offers
51 potential evolutionary advantages during viral replication, in the form of reassortment. The
52 exchange of genomic material between two distinct viral strains in a co-infected cell often
53 causes a massive shift within the genome, greatly increasing viral genetic diversity. In only
54 the last century, influenza pandemics of 1957 (H2N2), 1968 (H3N2), and 2009 (H1N1)
55 have emerged through reassortment of at least two viruses (1). Segmented genomes also
56 come with an inherent mechanistic challenge: ensuring that progeny receive a complete
57 and unique set of genomic segments. Random packaging is possible but comes at the cost of
58 producing an overwhelming majority of progeny that are not viable. Evidence suggests that
59 there exists an active mechanism of selective packaging ensuring progeny viability through
60 efficient and accurate genomic packaging (2).
61
62 Within a virion, the viral genome is organized into individual viral ribonucleoprotein
63 (vRNP) complexes, composed of the viral RNA, virally encoded nucleoprotein (NP), and a
64 heterotrimeric polymerase complex made up of PB1, PB2 and PA. The vRNP structure is
65 known classically as a helical panhandle, where the RNA is wrapped around NP with RNA
66 bases exposed and the 3' and 5' ends associated with the polymerase complex (2). We have
67 recently demonstrated that the structure of vRNP is more complex than the classically
68 depicted "beads-on-a-string" schematic. The NP and viral RNA associate in a non-uniform
69 manner, where regions of the RNA are unbound by NP and capable of forming complex

70 structures (3). Macro-organization of vRNPs within the viral capsid show tight packaging,
71 organized in a '7+1' orientation, a single vRNP center shaft surrounded by 7 others (4-6).
72
73 Given the need to package one copy of all 8 vRNP segments, there is strong evidence for the
74 existence of a selective packaging mechanism prior to budding (2, 4-7). Intersegmental
75 RNA-RNA interactions have been proposed to mediate selective packaging of all eight
76 influenza vRNP segments. Specifically, the 5' and 3' non-coding regions of each segment are
77 highly conserved, essential for efficient packaging, and may be critical for the proposed
78 RNA-RNA interactions (2, 8-11). Previous studies using *in vitro* transcribed viral RNA have
79 successfully demonstrated RNA-RNA binding. These studies also show that a set of eight
80 reproducible interactions can suffice to form a full genome complex (12-14). However,
81 these studies were performed on viral RNA in the absence of NP which would alter the
82 regions of viral RNA accessible for the proposed RNA-RNA interactions. Recent *in vivo*
83 experiments have also pointed towards RNA-RNA interaction as the driving mechanism for
84 selective packaging (15). While there is consensus of its existence, a clear network of
85 segment interactions leading to a full genome complex has yet to be elucidated.
86
87 We and others have previously used fluorescence *in situ* hybridization (FISH) studies to
88 visualize the intracellular localization of multiple vRNP segments during a productive viral
89 infection (16, 17). These studies have shown that distinct vRNP segments colocalize within
90 the cytoplasm. Based on our previous studies, we proposed a model whereby vRNP
91 segments are exported from the nucleus, the site of vRNP synthesis, as subcomplexes that
92 undergo further assembly en route to the plasma membrane through dynamic fusion or

93 colocalization events (17). We also performed a series of binary colocalization
94 comparisons, which would postulate a simple linear vRNP interaction network. However,
95 no such network was identified, suggesting the presence of a more complex vRNP
96 interaction network, including higher order subcomplexes as intermediate steps (17).
97
98 The lack of a method to image all eight vRNP segments within a single infected cell has
99 limited our ability to examine the precise spatial relationship between segments. In this
100 study, we developed a model based approach, rooted in point process theory, to quantify
101 vRNP spatial dependence as a metric for RNA-RNA interaction using multi-color FISH
102 images. A spatial point process model is a statistical model that captures the probability
103 that an event will happen at any position within a given geometry. The probability can
104 depend on the spatial distributions of other events or objects, often referred to as
105 covariates or factors. In this context, an event is the observance of fluorescent signal
106 (corresponding to one or more vRNPs) at a given location. We have previously used point
107 process modeling to construct generative models reflecting the spatial dependencies
108 between various punctate cell organelles and other cellular structures, such as the cell and
109 nuclear membranes, microtubules and the endoplasmic reticulum (18, 19). Related
110 analytical methods have been used to analyze clustering of molecules reflecting spatial
111 dependencies between copies of the same structure (20, 21) Here we use point process
112 models to present a statistically rigorous analysis of intracellular influenza A genome
113 assembly dynamics from a spatial perspective during a productive viral infection.
114

115 For an imaged cell, the locations of each vRNP segment can be represented as "realizations"
116 of an underlying probability density specific to that vRNP, calculated over the cytoplasm,
117 and dependent on the spatial proximity to various other structures within the cell. That is,
118 at any given position within the cell, there exists a probability of observing a single
119 segment, of a particular identity, that depends on the positions of cellular organelles and
120 other vRNP segments. By modeling the locational densities of each vRNP, we can therefore
121 learn a dependency network from which likely vRNP interactions are implied.
122
123 FISH imaging yields many distinct observations of individual cells, each with multiple vRNP
124 patterns observed. These point patterns, when viewed as realizations of point processes,
125 allow for statistical learning of spatial dependencies over many replicates for each vRNP
126 segment. By defining a set of covariates as the minimum distances between a given
127 segment and other segments observed within the same cell, we produce a set of models
128 that describe the spatial relationship between distinct segments. The extent to which two
129 (or more) vRNPs are dependent upon each other (e.g., likely to be found at nearby
130 positions) is reflected in the likelihood that images of those vRNPs would have been
131 produced by a learned model: high likelihoods signify that the positions of a given vRNP
132 can be predicted well from the others in that model. Model likelihoods can then be seen as
133 a metric of vRNP interaction, with higher likelihoods indicating more probable vRNP
134 association. These likelihoods can then be used to construct the most probable sequence of
135 interactions to form the full genome with methods borrowed from evolutionary tree
136 construction (22).
137

138 **Results:**

139

140 **Imaging influenza infected cells**

141 To investigate the spatial dependence between vRNP segments, we utilized images from a
142 previous study of infected cells stained for different combinations of vRNP segments using
143 four-color FISH (17). The data set included 14 different probe combinations that covered
144 all pairwise vRNP associations and 32 out of 56 possible triple vRNP combinations
145 (Supplemental Table 1 and 2). Prior to initiating pattern analysis, each image was
146 processsed to remove background, isolate individual cells, and find punctate vRNPs. Figure
147 1a and b presents representative raw and segmented images with detected points.

148

149 **Assessing randomness of the spatial distributions of vRNP**

150 We began by determining the extent to which the individual vRNPs could be considered
151 randomly distributed throughout the cell, since our search for spatial dependencies
152 between different vRNPs would be illogical if their positions were all random. We defined
153 the null hypothesis as complete spatial randomness, as would be exhibited by a
154 homogenous Poisson process, and estimated a p-value for this hypothesis using Monte
155 Carlo methods. Briefly, for a given vRNP in a particular cell, we started by segmenting the
156 cytoplasm and counting the number of points observed within. We then simulated
157 (generated) many random point configurations, of the same number of points, within the
158 segmented cytoplasm. We calculated a test statistic for each random configuration (see
159 Methods) and used these to assign a p-value to the test statistic for the observed pattern for
160 that cell; these were averaged for each vRNP. As shown in Table 1, every vRNP segment

161 showed significant deviation, with low variance, from a spatially random distribution. This
162 indicates the presence of a spatial trend and some dependency on cellular structures for
163 their location.
164
165 **Modeling cell and nuclear membrane dependence**
166 Given that the localizations of vRNPs are not random, we sought to determine if some
167 relationship to the cell and nuclear membranes could explain the spatial trend of vRNP
168 segments. Three covariates were defined for each position in the cell (minimum distance to
169 the cell membrane, minimum distance to the nuclear membrane, and the minimum ratio of
170 cell to nuclear distance for each point) and models were created for each vRNP using each
171 of the covariates separately. We estimated the accuracy of each model by cross validated
172 likelihood calculation (see Methods). It is important to note that, because the initial
173 pseudolikelihoods were used only to fit parameters, actual likelihoods can be compared
174 between models of different complexity.
175
176 Dependency on the cell membrane produced the lowest likelihood (worst) models relative
177 to all others (Table 2). Congregation of vRNP to the apical cell membrane is a potential
178 skewing factor. The many observed points within close proximity to the membrane
179 potentially caused low fitted probability of observing other points closer to the nucleus.
180 Cell shape differences combined with lower resolution in the z-dimension may have also
181 influenced these models.
182

183 For most vRNPs, models incorporating nuclear distance showed higher likelihoods than
184 those using cell membrane distance (Table 2). Some vRNP segments, such as PB2, PB1, PA
185 and NS, showed greater dependency on nuclear distance than the others. The ratio of cell
186 membrane to nuclear distance produced the best (highest likelihood) models in all cases
187 (Table 2). This ratio was a more accurate descriptor of location across varying cell shapes,
188 providing the most cogent model for dependence on host cell structures. Ratio models for
189 PB2, PB1, PA, and NS yielded higher likelihoods than other vRNP segments, showing a
190 greater dependence on both the nuclear and cell membrane location.
191
192 **Inter-segment dependence in pairs, triplets, and quadruplets**
193 Every pair of vRNP segments was observed in at least three cells, allowing for modeling of
194 vRNP-vRNP dependence among all segment pairs. For each pair of vRNP segments, two
195 models were trained, switching the dependent (primary) and independent (secondary)
196 patterns. Since the minimum inter-pattern distance measure is not commutative, the two
197 models need not be the same. More simply, a given vRNP location may depend on its
198 proximity to another vRNP, but the converse is not necessarily true.
199
200 In most cases, pairwise inter-pattern distance models (Table 3) had higher likelihoods than
201 those produced with the cell and nuclear membrane factors (Table 2). Some pairs showed
202 higher likelihoods than others, specifically PB1-NP, NP-PB1, PA-NP, HA-PB1, and PB1-M.
203 These pairings also showed relatively low variance in cross-validated likelihood estimates
204 (Supplemental Table 3). Many of the highest likelihood vRNP pairs are observed between
205 segments encoding the viral polymerase components, which is not surprising given the

206 dependence between these proteins both structurally and functionally (23-25). Figure 2
207 highlights the pairwise models by indiciating related vRNP pairs that are above -0.2 log
208 likelihood. Overall, vRNPs showed the greatest dependence on PA and PB1, with the most
209 connections between these two segments and the others, and the least dependence on NA,
210 which had the fewest connections.
211
212 Extending the pairwise models, the ratio of cell to nuclear distance was added as another
213 covariate. Observed point location in these models depended both on inter-pattern
214 distance and relative distance to the cell and nuclear membranes. Improvements in model
215 likelihood were seen in some models (Table 4), yet overall there was an increase in the
216 variance of our estimates (Supplemental Table 4). Overfitting to the cell and nuclear
217 distance may be the cause of this increased variance.
218
219 Thirty-two of 56 possible triplet vRNP segment complexes (referred to as triplets) were
220 present within the dataset analyzed and could therefore be modeled within the same cell.
221 This process allows for modeling of complex vRNP relations containing more than two
222 vRNP segments. Three models were trained for each observed triplet, one for each protein
223 depending on the other two. The highest likelihood model was chosen. Model likelihoods
224 tended to be higher than that of either pair model for a given vRNP (Table 5). Consistently,
225 we observed that the most likely triplet models for each vRNP included the five most likely
226 vRNP pairings (PB1-NP, NP-PB1, PA-NP, HA-PB1, and PB1-M) observed in pair likelihood
227 models (Table 3). Model likelihoods for all triplets observed can be found in Supplemental
228 Table 5.

229

230 A small subset of all possible four vRNP segment complexes (referred to as quadruplets)
231 were contained within the dataset and were included in our modeling of the complex
232 interactions. Again, considerable overlap is seen between the most likely quadruplet
233 models and both triplet and pair models (Table 6). The consistency in vRNP composition
234 between observed likely quadruplet, triples, and pairs further validates the spatial
235 dependence of vRNP segments and the presence of vRNP subcomplexes as influenza A
236 assembly intermediates. Model likelihoods for all quadruplets observed can be found in
237 Supplemental Table 6.

238

239 One of our major goals was to attempt to extrapolate higher order relationships from the
240 pair, triple and quadruplet models. To project how accurately this might be done, we
241 compared our observed triple and quadruple models with predicted models calculated
242 from only the pair models. Figure 3 illustrates that most of our observed triples and
243 quadruplets could be predicted accurately from the pair models, with a r^2 value of 0.266.
244 The exceptions are the triple and quadruple clusters containing the PB2-NA pair with
245 either NS, PB1 or both. These combinations were predicted to occur more often than was
246 observed in the experimental data sets (Figure 3). The pairwise likelihood of NA with PB2
247 is >-0.5, which is consistent with the observed likelihood and may be the driving force
248 behind the score compared to the other pairwise interactions. Without the triples and
249 quadruples containing PB2-NA, the r^2 value increases to 0.442. Overall, these data support
250 the notion that pairwise point process models can be used to predict higher order
251 complexes with reasonable accuracy.

252

253

254 **Interaction network construction**

255 Given these results, we sought to create a model of the set of interactions (complexes) that
256 occur at all steps in the formation of a complete viral genome. To do this, we represented
257 all possible interactions between vRNPs as a weighted, directed, acyclic graph with nodes
258 labeled by each unique, unordered set of vRNPs. A path from all single nodes to the root
259 represents a set of vRNP interactions resulting in a full genome complex (see Methods). For
260 illustration, if all eight vRNPs directly formed a complete virion without forming any
261 intermediate complexes (a highly unlikely scenario), this would be represented by weights
262 of 1 for the connection between each vRNP and the complete virion and weights of zero for
263 all other complexes. We then proceeded to estimate the weight for each possible vRNP
264 complex containing at least 2 and up to eight segments based upon either the observed or
265 predicted likelihoods (see Methods). Dynamic programming was then used to find the most
266 probable path within the graph (Figure 4a). This tree produces two distinct clusters HA, M,
267 NA, NS (hereafter referred to as C1) and PB1, PB2, NP (C2) that merge with PA as a final
268 step towards forming the full genome.

269

270 To assess the contribution of the observed quadruplets and triplets on this particular
271 network, we performed the network construction both without observed quadruplets
272 (Figure 4b) and without observed quadruplets and triplets (Figure 4c). Interestingly,
273 exclusion of all observed quadruplets did not drastically alter the network, with both C1

274 and C2 still being present. However, networks built only with observed pairs resulted in a
275 quite different network lacking larger subcomplexes (Figure 4c).
276
277 Given that the dynamic programming approach yields only a single most likely tree, we
278 sought to assess the robustness of that inferred tree to potential inaccuracy in the
279 estimated dependencies. We therefore assessed the stability of the tree by adding various
280 amounts of random noise to the estimated likelihoods (see Methods, Figure 4d).
281 Surprisingly, the addition of noise of up to a third of the mean log-likelihood had little effect
282 on the network found.
283
284 Since the vRNP sets constituting C1 and C2 had been directly measured (reactions L and B,
285 respectively), we also considered the possibility that the presence of measured (rather
286 than inferred) likelihoods for them drove network construction and biased our results only
287 towards those sets of vRNAs that we observed. If this is not the case, removal of all
288 observations of these subcomplexes would ideally not significantly change the final
289 network; C1 and C2 should be recaptured from other interaction models. Removal of data
290 derived from images of the sets of vRNPs in C1 and C2 (Figure 4e) yielded a network that
291 largely overlapped with Figure 4a. Here we see HA, M, NS, *PB1* in place of C1 and NP, *PA*,
292 PB2 in place of C2. This same tree was generated with the removal of only C1.
293
294 Removal of only C2 again yielded a network with substantial overlap to Figure 4a. C1 was
295 replaced by HA, M, NA, NS with NP, PA, PB1, PB2 replacing C2 (Figure 4f). Together, these
296 constructed networks demonstrate both the stability of the primary interaction network

297 and the importance of triplet and quadruplet observations in predicting higher order vRNA
298 complexes.
299
300 **Discussion:**
301 Influenza A vRNP segments selectively assemble within an infected cell to produce progeny
302 virions containing one copy of all eight segments. The mechanism driving selective
303 assembly is still largely unknown, but RNA-RNA interaction between the segments has
304 been proposed (12-14). In this study, we have modeled the *in vivo* spatial dependencies of
305 influenza A vRNP segments using multi-color fluorescent *in situ* images to generate
306 possible vRNP-vRNP interaction networks and propose a new perspective on genome
307 packaging during viral replication. By utilizing a rigorous statistical framework, we have
308 extended the efforts of previous groups and present a novel method for construction of
309 vRNP interaction networks based on their precise spatial information within actively
310 infected cells. Using spatial proximity as a proxy for physical interaction, point process
311 modeling in this study suggests clear spatial dependenencies between certain vRNP
312 segments, and confirms our previous observations of subcomplex formation during
313 cytoplasmic transport (18).
314
315 Previous studies using *in vitro* transcribed RNA suggest multiple interactions between
316 vRNP segments are expected (2, 4, 12, 14, 26). Similarly, the modeled likelihoods show
317 potentially multiple interactions for each vRNP segment. Electron tomography studies have
318 revealed a conserved '7+1' supramolecular structure within the viral interior, where a

319 center shaft is surrounded by seven remaining segments (2, 4, 26), demonstrating an
320 ordered process in genome assembly.
321
322 Our results suggest two candidates for the center shaft or 'master segment': PA and PB1.
323 From the 6 constructed networks, PA is the most variable, switching between
324 subcomplexes. This would be expected if a segment was evenly dependent on most others,
325 a possible signature of a core segment. PB1 also behaves similarly. Both PB1 and PA show
326 the highest average pairwise model likelihood over all other segments.
327
328 A few vRNP pairs were seen often in the most likely pairwise, triplet, and quadruplet
329 models and in the final constructed network. PB1-PB2, HA-M, and HA-PA show
330 commutative relationships, appearing in the top most likely triplet and quadruplet models
331 for each vRNP within the pair. Some non-commutative dependencies were also seen, most
332 notably in NP depending on PB2 in both triplet and quadruplet models without PB2-NP
333 dependence.
334
335 By using dynamic programming, we were able to generate the most likely interaction
336 network based on all models. Subtle variations in the vRNP interaction network were
337 observed by the addition of noise and exclusion of higher order datasets, suggesting that
338 there may exist multiple interaction pathways all acting at once in a single cell. Some of the
339 variation in our results may also be due to the presence of some lower specificity vRNA-
340 vRNA interactions also occurring during genome assembly. This would explain some of the
341 highly likely pairwise interactions observed that were not included in the final interaction

342 network, such as PB1 and HA. A system with multiple possible interaction networks could
343 be potentially advantageous in both ensuring packaging of a full genome in viral particles
344 and in future reassortment events. For example, reassortment of one vRNA segment during
345 a coinfection may decrease the binding efficacy necessary for a single interaction network
346 but a second network may serve to rescue viral viability while also increasing genetic
347 diversity.
348
349 With this, we also point out the consistent presence of the PB2, PB1 and NP cluster. Since
350 these three proteins function together, with PA, to form the viral polymerase, which
351 promote vRNA replication and transcription, this cluster is potentially important. Viruses
352 capable of packaging complementary polymerase segments into progeny virions will have
353 a fitness adavantage compared to viruses with polymerase mismatches. This phenotype
354 has been observed in experimental reassortment experiments between 2009 H1N1
355 pandemic and seasonal H3N2 viruses(23, 25).
356
357 More generally, the computational pipeline presented may function as a useful tool in
358 elucidating spatial dependency over a wide range of biological phenomena observable
359 through microscopy. As we have demonstrated, point process models can capture key
360 aspects of point distributions while retaining a basis in probability. Network construction
361 synthesizes the results derived through modeling into a cogent, most-likely set of spatial
362 interactions or dependencies.
363

364 Future work in spatial modeling of influenza A packaging should incorporate a temporal
365 dimension through live cell imaging. Point process models have been adapted for
366 spatiotemporal modeling (27). Incorporation of intracellular markers important for
367 influenza A vRNP transport may increase the accuracy of models. Influenza A vRNP
368 transport from the nucleus, the site of vRNP synthesis, to the plasma membrane is a
369 complex process utilizing a variety of host proteins (28). Rab11A-containing vesicles are
370 thought to be the primary mode of transport although there is evidence that a Rab11A-
371 independent mechanism exists (29-31). A recent study has implicated the ER to mediating
372 transport of vRNP segments through anchored Rab11A proteins(32). Novel methods in
373 fluorescent multiplexing (33) present an exciting opportunity for observing many cellular
374 structures which will provide a holistic image the spatial dependence of vRNP segments
375 upon subcellular structures and to each other.
376
377 **Methods:**
378 **Multi-color Fluorescent In Situ Hybridization Images**
379 This study used previously published multi-color FISH images from (17) that were
380 generated at the National Institutes of Health. Briefly, MDCK cells were infected with
381 recombinant WSN/1933 H1N1 for 8 hours and then fixed and stained with FISH probes,
382 obtained from Biosearch technologies, against four distinct vRNP segments. DAPI was
383 included as well to label DNA. Multi-color FISH samples were imaged on an Leica SP5
384 white light laser to ensure spectral separation of the five colors (Dapi, Alexa 488, Quasar
385 570, Cal Fluor Red 590, Quasar 670). The specificity of the probes and spectral separation
386 between fluorophores was previously confirmed (17).

387

388 **Image preprocessing, cell segmentation and point detection**

389 All image preprocessing was performed in MATLAB, v. 2015a. Prior to point detection and
390 segmentation, image noise was removed by convolution with a Gaussian mask. Each image
391 channel was denoised individually. As described in (17), each image was captured with
392 0.17 um z-step size spanning the entire cell volume, defined for each cell using both the
393 nuclear and FISH staining to define the apical cell membrane. Each image was captured
394 with a pixel size of ~50x50x168 nm.

395

396 Individual vRNP segments were identified by finding connected areas of signal within the
397 denoised image. The center of each object was taken as the maximum intensity pixel
398 within, yielding a set of point coordinates.

399

400 Since there was no fluorescent tag for cell membrane components, we estimated the cell
401 boundary using the vRNP images. We assumed that fluorescent signal would be denser
402 within the cell cytoplasm than in the surrounding area with cell-cell junctions subtly
403 defined by 'valleys', curves of low signal density whose normal vectors point towards
404 increasing density. The gradient of point density over the image was then used to segment
405 individual cells from both the background and each other by the mean-shift algorithm (34).
406 Hand segmentation was also used to ensure accurate cell membrane segmentation. In most
407 cases, the convex hull of all points within an identified cell region then defined the cell
408 membrane. The nucleus was segmented through simple thresholding and smoothing.

409

410

411 **Point processes and patterns**

412

413 Within each cell, the vRNP segments form a point pattern, **x** = {x_1, x_2, …, x_n}, where *n* is the
414 number of observed points and each x_i is the 3-dimensional coordinate vector for point *i*.
415 The point pattern is defined over a bounded region, *W*, the segmented cell cytoplasm. **x** is
416 then viewed as a realization of some random point process **X**, the generating distribution
417 for all patterns of the particular vRNP identity. The process **X** then represents the
418 culmination of biological factors that determines vRNP segment location, eg. nuclear
419 export, directed transport over the microtubule network, inhibition by other organelles,
420 etc.

421

422 To model **X**, first define its locational density, *f*, a function over the cytoplasm, where *f(*u*)* is
423 the probability of observing a point at position **u**. The simplest model for this density is the
424 Homogenous Poisson, characterized by complete randomness over space, or uniform
425 probability:

426

427 $$f(\boldsymbol{x}|n) = \frac{1}{Z}\lambda^n \qquad (1)$$

428

429 where λ determines point density, and Z a normalizing constant. In the point process
430 literature, λ is referred to as *intensity*, but we refer to it as point density to avoid confusion
431 with fluorescence intensity.

432

433 **Characterizing spatial randomness**

434

435 Hypothesis testing for spatial randomness, comparing an observed pattern to an
436 expectation under a Poisson assumption, was used as an initial motivation for further
437 modeling. Ripley's K-function describes the number of neighboring points within a given
438 distance (r) of each observed point in a pattern, a measure of clustering or inhibition of
439 points within space. Under a Poisson assumption,

440

441 $$K_{poi}(r) = \pi r^2 \quad (2)$$

442

443 The expected K-function value can then be used to measure the difference between a given
444 pattern and that of a Homogenous Poisson with an equal number of points. To assess this
445 for observed point patterns, we used the test statistic

446

447 $$T = \int_0^{r_{max}} \left(\widehat{K}(r) - K_{poi}(r)\right)^2 dr \quad (3)$$

448

449 where $\widehat{K}(r)$ is the estimated K-function for a given pattern and r a radius defining the point
450 neighborhood. To provide a background distribution for this statistic for a given cell
451 geometry, we generated 100 samples from a Homogenous Poisson process defined over its
452 cell cytoplasm and calculated the test statistic. We then obtained a p-value for the
453 hypothesis that an observed distribution was drawn from a homogenous Poisson process:

454

455 $$p = \frac{1+\sum_{i=1}^{m} I(t_{obs} \geq t_i)}{m+1} \quad (4)$$

456

457 where m is the number of samples drawn. The p-values were averaged over all cells for a

458 given vRNP.

459

460 **<u>Modeling spatial dependence on cellular structures</u>**

461

462 An Inhomogenous Poisson Process, in its simplest form, is characterized by a spatially

463 dependent locational density. Per point factors (eg: distance to cell membrane) determine

464 the probability of a point occurring at a given location. As above, let **X** be some point

465 process with realization **x**, and define $s(\mathbf{x})$, an n x k matrix where each $s_{i,j}$ is some factor j

466 that can be calculated for each point x_i. The Inhomogenous Poisson point density function

467 is:

468

$$\lambda(x_i; \theta) = \left(exp(\theta^T s(x_i)\right)$$

469

470 where θ is the k-dimensional parameter vector. The likelihood of the model given point

471 data is then defined by:

472

$$l(\boldsymbol{x}; \theta) = \left(\prod_{i=1}^{n} \lambda(x_i)\right) \exp\left(-\int_W \lambda(u) du\right)$$

473

To quantify vRNP dependence on the cell and nuclear membranes, we define the first factor (f_1) as the ratio of nuclear to cell distance calculated per point:

$$f_1(\boldsymbol{x}) = {d_n(\boldsymbol{x})}/{d_c(\boldsymbol{x})}$$

<u>Modeling spatial dependence between vRNPs</u>

The second factor, minimum inter-pattern point distance, was used to explore whether there is vRNP-vRNP interaction between vRNPs of different types. For each observed vRNP segment of type *l*, and some other vRNP point pattern of type *j*, the minimum inter-pattern point distance is:

$$f_2\left(x_i^l | l \neq t\right) = \min_j d\left(x_i^l, x_j^t\right)$$

For each vRNP type, we use a 'one depends on all' schema. That is, each model with the same set of vRNPs is not equivalent but exclusively represents a single vRNP pattern (the primary pattern) depending on others (the secondary patterns). Models were constructed for every unique pair, triplet, and quadruplet of vRNP types.

For each model, all instances of its unique components were gathered, all images with tags for every vRNP in the model. Parameters of the model were then fit for each instance using

494 the maximum pseudolikelihood (35). Hold one out cross validation was used to assess
495 quality of fit over all instances yielding cross-validated pseudo-likelihood estimated
496 parameters and variances. These fitted parameters were then used to calculate a true
497 model likelihood as in Li et al (19).
498
499 **Interaction network assembly**
500
501 We formulate the task of finding the most probable set of vRNA interactions yielding a full
502 genome as a graph problem. First, let the set of nodes in the graph be every possible
503 combination of vRNAs ranging in size from a single vRNA to all 8 segments. Directed edges
504 in the graph connect subset to super set nodes. We weight each edge of the graph by the
505 likelihood of the most likely model that results in the superset and contains the subset as a
506 primary or secondary segment (Figure 1). Edges that represent models that have not been
507 observed remain unweighted.
508
509 From this graph, a set of interactions yielding the full genome is a tree with leaves as the
510 vRNA singletons and root as the full 8-mer of vRNA segments. In these trees, we require
511 each node to have at least two incoming edges, barring the leaf nodes. The set of incoming
512 edges to a given node must emanate from nodes whose labels, when combined, exactly
513 equal the given node. A set of incoming edges essentially 'combines' nodes, directly
514 corresponding to the physical event of vRNP segments binding. The most likely set of
515 interactions is also the most highly weighted tree. The problem of finding the most
516 probable interaction network then reduces to finding the highest weighted tree in the

517 graph. This problem can easily be solved through dynamic programming within a paradigm
518 often used in evolutionary tree construction (22).
519
520 The likelihoods of models involving multiple vRNP segments can be taken as a measure of
521 spatial dependence of primary segments on secondary segments. vRNP segments that truly
522 interact are expected to be highly dependent either unidirectionally or bidirectionally. The
523 cross validated model likelihoods can be used to weight groups of edges in the graph. Since
524 the graph also has edges that were not observed in actual images (eg: any complex of more
525 than 4 vRNAs), composite or inferred likelihoods were generated through recursion in a
526 dynamic programming scheme. For each unobserved complex, we took the most likely set
527 of observed models that could generate the complex.
528
529 For a given unobserved complex $N = \{N_1, N_2, N_3, \ldots\}$,
530
531 $$L(N) = \max(L(N_1|N_2,N_3,\ldots), L(N_2|N_1,N_3,\ldots), \ldots)$$
532
533 Then for some likelihood in the above, $L(N_1|N_2)$ where $N_1 = (n^1_1, \ldots n^1_k)$ and $N_2 = (n^2_1, \ldots$
534 $n^2_l)$, $|N_1| = k$, $|N_2| = l$,
535
536
$$L\big((N_1|N_2)\big) = \sum_{i=1}^{k} \underset{N_j^2 \in \, all\ unique\ sets\ of\ N_2}{\text{average}} (L(n_i^1|N_j^2))$$
537

538 **Interaction network stability analysis**

539

540 To assess the stability of the generated interaction network, two procedures of
541 perturbation were performed: introduction of noise and removing certain high likelihood
542 models. We simulated noise as a normal random distribution centered at 0 with varying
543 standard deviations, termed the noise levels. Under increasing noise levels, each observed
544 model likelihood was amended with a noise value drawn from this distribution prior to
545 interaction network construction. For each noise level, we simulated 100 trials and tallied
546 the proportion of output trees that contained each possible edge. Trees that align with the
547 non-noisy assembly signal stability while highly variable trees signal instability.

548

549 **Reproducible Research Archive**

550

551 All images, derived data and source code will be available upon publication at
552 http://murphylab.cbd.cmu.edu/software.

553

554 **Funding Acknowledgment:** This study was supported in part by National Institutes of
555 Health grants AI108600-01, GM090033, and GM103712 and by National Science
556 Foundation grant MCB1616492 .

557

558 **Competing Interests:** The authors declare that there are no competing interests.

559

560 **REFERENCES**

561

562 1. Morens DM, Taubenberger JK. Pandemic influenza: certain uncertainties. Rev Med
563 Virol. 2011;21(5):262-84.
564 2. Hutchinson EC, von Kirchbach JC, Gog JR, Digard P. Genome packaging in influenza A
565 virus. J Gen Virol. 2010;91(Pt 2):313-28.
566 3. Lee N, Le Sage V, Nanni AV, Snyder DJ, Cooper VS, Lakdawala SS. Genome-wide
567 analysis of influenza viral RNA and nucleoprotein association. Nucleic Acids Res.
568 2017;45(15):8968-77.
569 4. Noda T, Sagara H, Yen A, Takada A, Kida H, Cheng RH, et al. Architecture of
570 ribonucleoprotein complexes in influenza A virus particles. Nature. 2006;439(7075):490-2.
571 5. Chou YY, Vafabakhsh R, Doganay S, Gao Q, Ha T, Palese P. One influenza virus
572 particle packages eight unique viral RNAs as shown by FISH analysis. Proc Natl Acad Sci U S
573 A. 2012;109(23):9101-6.
574 6. Noda T, Kawaoka Y. Packaging of influenza virus genome: robustness of selection.
575 Proc Natl Acad Sci U S A. 2012;109(23):8797-8.
576 7. Isel C, Munier S, Naffakh N. Experimental Approaches to Study Genome Packaging of
577 Influenza A Viruses. Viruses. 2016;8(8).
578 8. Fujii Y, Goto H, Watanabe T, Yoshida T, Kawaoka Y. Selective incorporation of
579 influenza virus RNA segments into virions. Proc Natl Acad Sci U S A. 2003;100(4):2002-7.
580 9. Marsh GA, Rabadan R, Levine AJ, Palese P. Highly conserved regions of influenza a
581 virus polymerase gene segments are critical for efficient viral RNA packaging. J Virol.
582 2008;82(5):2295-304.

583 10. Essere B, Yver M, Gavazzi C, Terrier O, Isel C, Fournier E, et al. Critical role of
584 segment-specific packaging signals in genetic reassortment of influenza A viruses. Proc Natl
585 Acad Sci U S A. 2013;110(40):E3840-8.
586 11. Gerber M, Isel C, Moules V, Marquet R. Selective packaging of the influenza A
587 genome and consequences for genetic reassortment. Trends Microbiol. 2014;22(8):446-55.
588 12. Fournier E, Moules V, Essere B, Paillart JC, Sirbat JD, Cavalier A, et al. Interaction
589 network linking the human H3N2 influenza A virus genomic RNA segments. Vaccine.
590 2012;30(51):7359-67.
591 13. Fournier E, Moules V, Essere B, Paillart JC, Sirbat JD, Isel C, et al. A supramolecular
592 assembly formed by influenza A virus genomic RNA segments. Nucleic Acids Res.
593 2012;40(5):2197-209.
594 14. Gavazzi C, Isel C, Fournier E, Moules V, Cavalier A, Thomas D, et al. An in vitro
595 network of intermolecular interactions between viral RNA segments of an avian H5N2
596 influenza A virus: comparison with a human H3N2 virus. Nucleic Acids Res.
597 2013;41(2):1241-54.
598 15. Gilbertson B, Zheng T, Gerber M, Printz-Schweigert A, Ong C, Marquet R, et al.
599 Influenza NA and PB1 Gene Segments Interact during the Formation of Viral Progeny:
600 Localization of the Binding Region within the PB1 Gene. Viruses. 2016;8(8).
601 16. Chou YY, Heaton NS, Gao Q, Palese P, Singer R, Lionnet T. Colocalization of different
602 influenza viral RNA segments in the cytoplasm before viral budding as shown by single-
603 molecule sensitivity FISH analysis. PLoS Pathog. 2013;9(5):e1003358.

604 17. Lakdawala SS, Wu Y, Wawrzusin P, Kabat J, Broadbent AJ, Lamirande EW, et al.
605 Influenza a virus assembly intermediates fuse in the cytoplasm. PLoS Pathog.
606 2014;10(3):e1003971.
607 18. Johnson GR, Li J, Shariff A, Rohde GK, Murphy RF. Automated Learning of Subcellular
608 Variation among Punctate Protein Patterns and a Generative Model of Their Relation to
609 Microtubules. PLoS Comput Biol. 2015;11(12):e1004614.
610 19. Li Y, Majarian TD, Naik AW, Johnson GR, Murphy RF. Point process models for
611 localization and interdependence of punctate cellular structures. Cytometry A.
612 2016;89(7):633-43.
613 20. Griffie J, Shlomovich L, Williamson DJ, Shannon M, Aaron J, Khuon S, et al. 3D
614 Bayesian cluster analysis of super-resolution data reveals LAT recruitment to the T cell
615 synapse. Sci Rep. 2017;7(1):4077.
616 21. Yunta ML, Lagache T, Santi-Rocca J, Bastin P, Olivo-Marin JC, editors. A statistical
617 analysis of spatial clustering along cell filaments using Ripley's K function. 2014 IEEE 11th
618 International Symposium on Biomedical Imaging (ISBI); 2014 April 29 2014-May 2 2014.
619 22. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of
620 progressive multiple sequence alignment through sequence weighting, position-specific
621 gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22(22):4673-80.
622 23. Li C, Hatta M, Watanabe S, Neumann G, Kawaoka Y. Compatibility among polymerase
623 subunit proteins is a restricting factor in reassortment between equine H7N7 and human
624 H3N2 influenza viruses. J Virol. 2008;82(23):11880-8.
625 24. Te Velthuis AJ, Fodor E. Influenza virus RNA polymerase: insights into the
626 mechanisms of viral RNA synthesis. Nat Rev Microbiol. 2016;14(8):479-93.

627 25. Phipps KL, Marshall N, Tao H, Danzy S, Onuoha N, Steel J, et al. Seasonal H3N2 and
628 2009 Pandemic H1N1 Influenza A Viruses Reassort Efficiently but Produce Attenuated
629 Progeny. J Virol. 2017;91(17).
630 26. Noda T, Sugita Y, Aoyama K, Hirase A, Kawakami E, Miyazawa A, et al. Three-
631 dimensional analysis of ribonucleoprotein complexes in influenza A virus. Nat Commun.
632 2012;3:639.
633 27. Illian JB, Soerbye, S, Rue, H & Hendrichsen, D. Using INLA to fit a complex point
634 process model with temporally varying effects – a case study Journal of Environmental
635 Statistics. 2012;3(7).
636 28. Lakdawala SS, Fodor E, Subbarao K. Moving On Out: Transport and Packaging of
637 Influenza Viral RNA into Virions. Annu Rev Virol. 2016;3(1):411-27.
638 29. Bruce EA, Digard P, Stuart AD. The Rab11 pathway is required for influenza A virus
639 budding and filament formation. J Virol. 2010;84(12):5848-59.
640 30. Eisfeld AJ, Kawakami E, Watanabe T, Neumann G, Kawaoka Y. RAB11A is essential
641 for transport of the influenza virus genome to the plasma membrane. J Virol.
642 2011;85(13):6117-26.
643 31. Nturibi E, Bhagwat AR, Coburn S, Myerburg MM, Lakdawala SS. Intracellular
644 Colocalization of Influenza Viral RNA and Rab11A is Dependent upon Microtubule
645 Filaments. J Virol. 2017;91(19).
646 32. de Castro Martin IF, Fournier G, Sachse M, Pizarro-Cerda J, Risco C, Naffakh N.
647 Influenza virus genome reaches the plasma membrane via a modified endoplasmic
648 reticulum and Rab11-dependent vesicles. Nat Commun. 2017;8(1):1396.

649 33. Lin JR, Fallahi-Sichani M, Sorger PK. Highly multiplexed imaging of single cells using
650 a high-throughput cyclic immunofluorescence method. Nat Commun. 2015;6:8390.
651 34. Comaniciu D, Meer P. Mean shift: A robust approach toward feature space analysis.
652 IEEE Transactions on pattern analysis and machine intelligence. 2002;24(5):603-19.
653 35. Baddeley A, Turner R. Practical Maximum Pseudolikelihood for Spatial Point
654 Patterns. Australian & New Zealand Journal of Statistics. 2000;42(3):283-322.
655
656

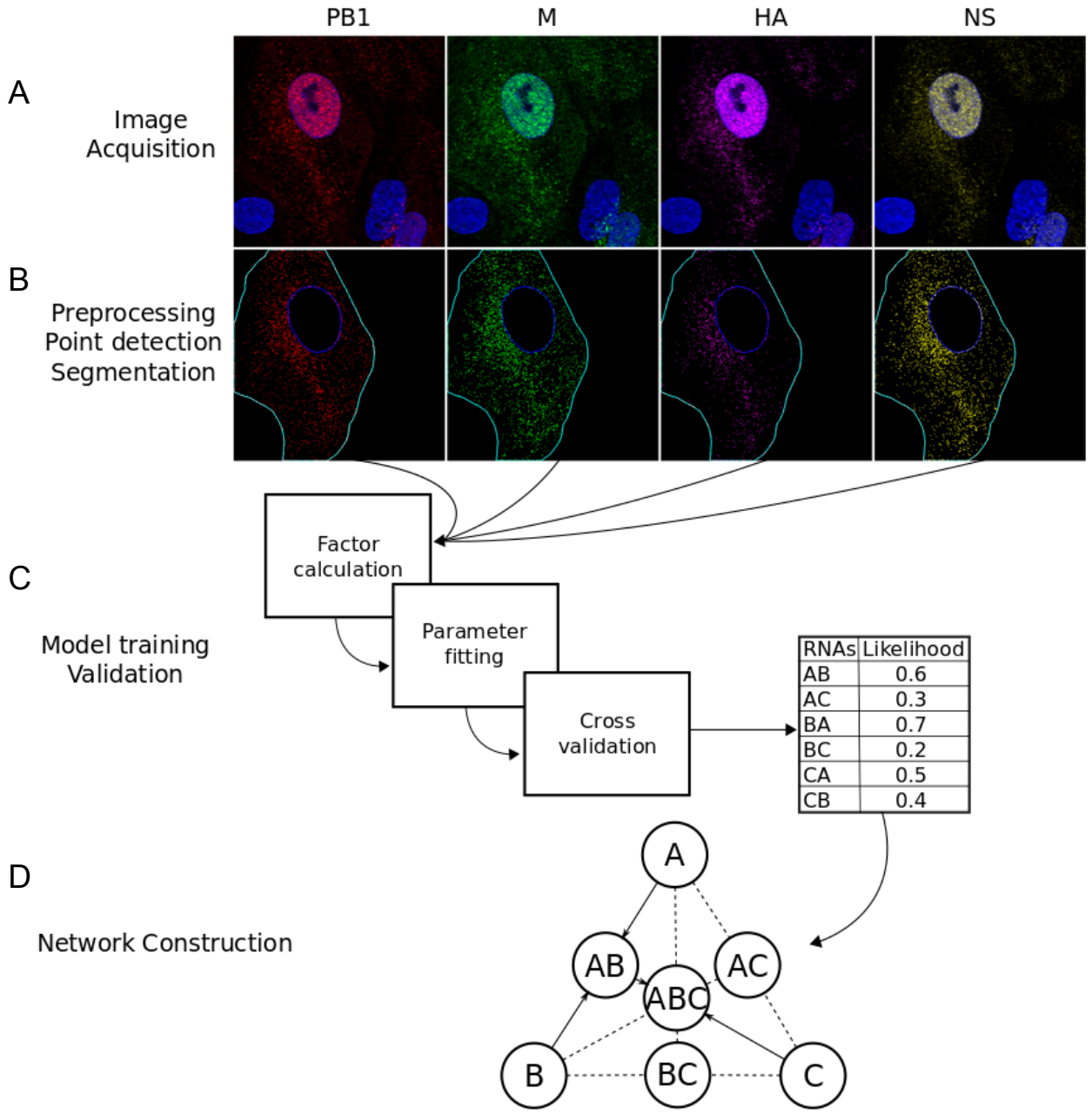

Figure 1: Illustration of the analysis pipeline. (A) Unprocessed fluorescence channels for cells FISH tagged for four vRNA segments and stained with DAPI. (B) Preprocessed image channels after filtering, nuclear and cell membrane segmentation, and point identification. The 3D images in (A) and (B) are displayed as maximum value projections. (C) Factors are calculated for each channel in each image, model parameters are fit, and a likelihood generated. (D) Dynamic programming is used to generate a most likely set of interactions yielding a full genome. Edges are possible interactions, solid lines indicate edges that are part of the most likely set of interactions. In this toy example, the most likely set of interactions is A and B binding to form AB, and then C binding to AB.

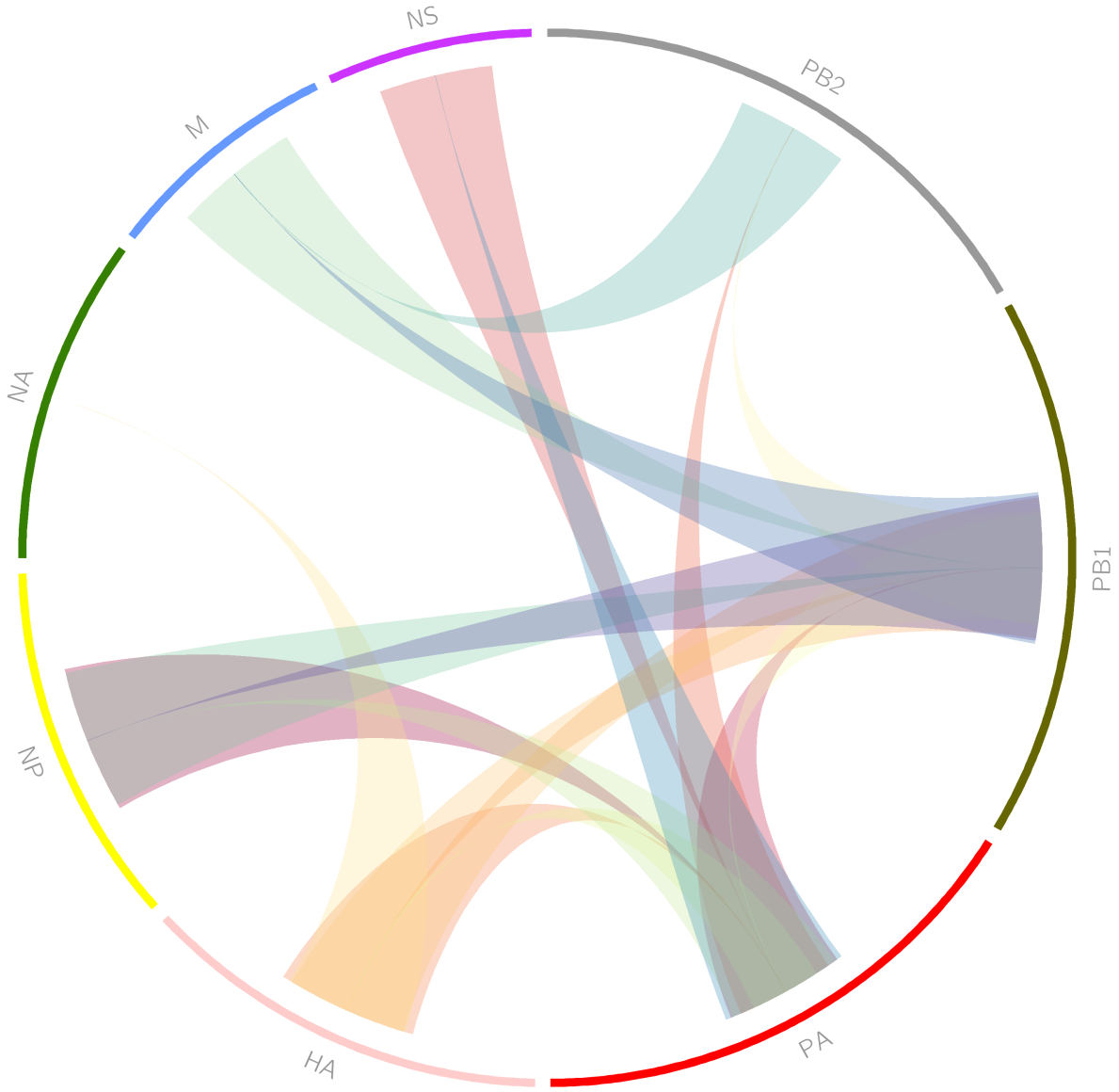

Figure 2: Pairwise interactions learning from MDCK cell images. Each vRNA segment is arranged around the perimeter. Bands connect vRNA segments that interact with high likelihood. The width of the base of each band is proportional to the likelihood of the interaction model, larger bases correspond to more likely pairings. Only interactions whose log likelihoods are above -0.2 are shown.

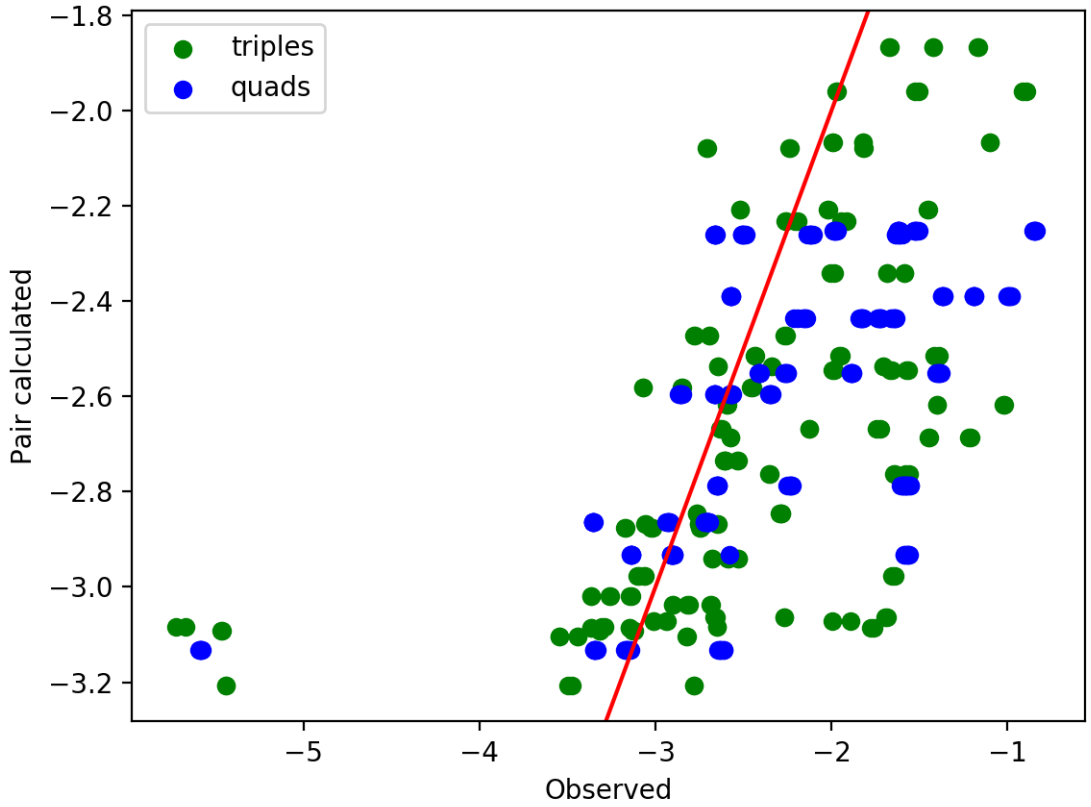

Figure 3: Comparison of composite and observed model likelihoods for triplet and quadruplet models. To ensure that the recursion in our dynamic programming scheme accurately predicts model likelihoods, we compared triplet and quadruplet observed likelihoods to those generated from only pairwise models. Ideally, all points would fall on the red line, matching observed and predicted values perfectly. The cluster in the lower left corner represent interactions that are predicted to occur with high likelihood but in reality, do not. All points within this cluster contain the same four vRNA segments (PB1, PB2, NA, NS)

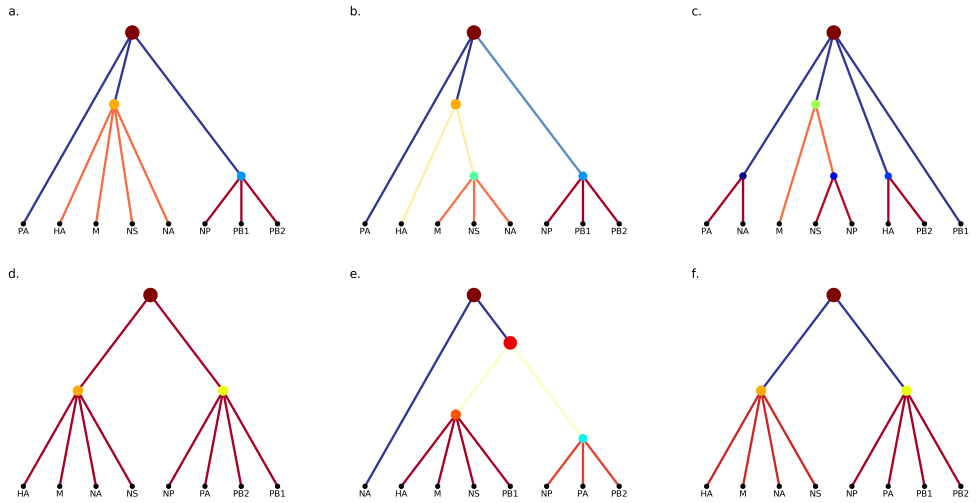

Figure 4: Constructed interaction networks. Using dynamic programming, we constructed the single most likely set of interactions yeilding a full genome. First, we used all model likelihoods (a) to build a master tree. Hotter edge colors correspond to more likely interactions. Since the images contain only a subset of the possible triplets and quadruplet, we assessed how much this affected the master tree by removing all quadruplets (b) or both quadruplets and triplets (c). To test the stability of the master tree, trees were constructed with random noise added to each observed model likelihood before performing the dynamic programing. Multiple noise levels were tested up to 1/3 of the mean model likelihood. In all cases, two possible trees were generated: (a) or (d). Given the two observed clusters in the master tree, we next asked whether the composite likelihood function could recapture their interactions without actually observing them. Removing all models involving HA, M, NS, NA (C1) and NP, PB1, PB2 (C2) yielded (e). Removing only C2 yielded (f).

Table 1: Monte Carlo p-values and variances for the hypothesis that the spatial distribution of each vRNA segment is distributed by a homogeneous Poisson process.

	Mean p-value	Variance
HA	0.00671	0.00015
PB1	0.00524	0.00034
PB2	0.0034	0.00011
NA	0.00249	0.00003
M	0.00005	0
PA	0.00467	0.0002
NP	0.01185	0.00036
NS	0.00306	0.00007

Table 2: Cross validated negative log-likelihoods for models dependent on distance to different reference locations.

	Cell Model	Nuclear Model	Cell+Nuclear Model
PB1	2.398	0.408	2.073
PB2	2.362	0.410	0.746
PA	2.414	0.470	1.252
HA	2.412	2.447	5.210
NP	2.506	2.501	5.236
NA	2.493	2.350	1.748
M	2.554	2.585	3.490
NS	2.397	0.666	1.931

Table 3: Cross validated negative log-likelihoods for pairwise interaction models.

	PB1	PB2	PA	HA	NP	NA	M	NS
PB1		0.241	0.203	0.205	0.125	0.403	0.171	0.291
PB2	0.387		0.268	0.284	0.315	0.561	0.208	0.398
PA	0.221	0.220		0.229	0.155	0.322	0.281	0.232
HA	0.128	0.333	0.184		0.272	0.233	0.258	0.288
NP	0.180	0.318	0.194	0.290		0.252	0.275	0.341
NA	0.325	0.320	0.309	0.300	0.274		0.285	0.300
M	0.225	0.329	0.284	0.289	0.297	0.293		0.291
NS	0.277	0.343	0.222	0.288	0.325	0.344	0.289	

Table 4: Cross validated negative log-likelihoods for pairwise interaction models that also include cell-to-nuclear distance ratio.

	PB1	PB2	PA	HA	NP	NA	M	NS
PB1		0.59	0.40	3.98	0.14	2.25	1.67	1.72
PB2	0.45		0.28	0.41	0.36	0.58	0.22	0.51
PA	0.37	0.31		1.77	0.56	1.08	0.91	1.17
HA	2.73	0.47	1.08		4.11	3.84	3.50	0.82
NP	0.19	0.56	0.49	2.54		5.36	6.98	0.72
NA	1.42	2.50	0.73	3.36	11.05		3.50	1.01
M	0.71	0.80	0.85	3.09	7.02	3.34		0.92
NS	2.15	0.41	0.99	0.88	0.53	0.72	1.12	

Table 5: Cross validated log-likelihoods for the top two triple models with each vRNA segment as the primary pattern. Pairwise model likelihoods for the members of each model are shown in the right columns.

RNA	CV Likelihood	PB1	PB2	PA	HA	NP	NA	M	NS
PB1									
PB1 M									
PB2	0.168		0.241					0.171	
PB1 M PA	0.176			0.203				0.171	
PB2									
PB2 M									
PB1	0.164	0.387						0.208	
PB2 PA M	0.175			0.268				0.208	
PA									
PA M PB2	0.273		0.220					0.281	
PA PB1 M	0.336	0.221						0.281	
HA									
HA NS PA	0.313			0.184					0.288
HA NP PB2	0.326		0.333			0.272			
NP									
NP PB1									
PB2	0.180	0.180	0.318						
NP PA PB1	0.302	0.180		0.194					
NA									
NA HA NP	0.255				0.300	0.274			
NA M HA	0.333				0.300			0.285	
M									
M PB2 PA	0.236		0.329	0.284					
M PB1 PA	0.257	0.225		0.284					
NS									
NS M PB2	0.300		0.343					0.289	
NS HA PA	0.356			0.222	0.288				

Table 6: Cross validated negative log-likelihoods for the top two quadruple models with each vRNA segment as the primary pattern. Pairwise model negative log-likelihoods for the members of each model are shown in the right columns.

RNA	CV Likelihood	PB1	PB2	PA	HA	NP	NA	M	NS
PB1									
PB1 M PA PB2	0.154		0.241	0.203				0.171	
PB1 PB2 NA NS	0.313		0.241				0.403		0.291
PB2									
PB2 PA PB1 M	0.159	0.387		0.268				0.208	
PB2 NS HA M	0.256				0.284			0.208	0.398
PA									
PA PB2 PB1 NP	0.202	0.221	0.220			0.155			
PA HA NS M	0.241				0.228			0.281	0.232
HA									
HA PA NS M	0.171			0.184				0.258	0.288
HA M NA NP	0.188					0.272	0.233	0.258	
NP									
NP PA PB2 PB1	0.234	0.180	0.318	0.194					
NP HA NA M	1.289				0.290		0.252	0.275	
NA									
NA HA NP M	0.242				0.300	0.274		0.285	
NA M PA NS	0.365			0.309				0.285	0.300
M									
M PB1 PB2 PA	0.245	0.225	0.329	0.284					
M NS HA PA	0.246			0.284	0.289				0.291
NS									
NS M PA HA	0.229			0.222	0.288			0.289	
NS PB2 M HA	0.340		0.343		0.288			0.289	

764 Supp. Table 1: FISH imaging probe sets. The leftmost column is the experiment label with
765 the right 4 columns showing which vRNA segments were tagged. Each experiment also
766 used DAPI staining.

Reaction	Probe 1	Probe 2	Probe 3	Probe 4
A	M	HA	PB2	NS
B	PB1	PB2	PA	NP
C	PB1	M	HA	NS
D	M	PA	HA	NS
E	M	HA	NA	NP
F	PB1	PB2	NA	NS
G		PA	NA	
H			NP	NS
I	PB1	M	PB2	PA
J		M	NA	NP
K		PB2	HA	NP
L	M	HA	NA	NS
M	M	NA	HA	PA
N	M	PB2	NP	
Rxn 1	M	PA	NA	NS

767
768
769

Supp. Table 2: Observed triplets within the reaction set. 32 of 56 possible triplets were observed.

Observed triplets	Reaction
PB2, PB1, PA	B, I
PB2, PB1, NP	B
PB2, PB1, NA	F
PB2, PB1, M	I
Pb2, PB1, NS	F
PB2, PA, NP	B
PB2, PA, M	I
PB2, HA, NP	K
PB2, HA, M	A
PB2, HA, NS	A
PB2, NP, NS	N
PB2, NA, NS	F
PB2, M, NS	A
PB1, PA, NP	B
PB1, PA, M	I
PB1, HA, M	C
Pb1, HA, NS	C
PB1, NA, NS	F
PB1, M, NS	C
PA, HA, NA	M
PA, HA, M	D, M
PA, HA, NS	D
PA, NA, M	M, Rxn 1
PA, NA, NS	Rxn 1
PA, M, NS	Rxn 1
HA, NP, NA	E
HA, NP, M	E
HA, NA, M	E, L, M
HA, NA, NS	L
HA, M, NS	L
NP, NA, M	E, J
NA, M, NS	L, Rxn 1

Supp. Table 3: Cross validated log-variances for pairwise interaction models with cell-to-nuclear distance ratio.

(large)

Supp. Table 4: Variance in the cross validated negative log-likelihoods for pairwise interaction models with cell-to-nuclear distance ratio shown in Table 4.

	PB1	PB2	PA	HA	NP	NA	M	NS
PB1		0.0172	0.0025	0.0012	0.0010	0.0002	0.0012	0.0067
PB2	0.0552		0.0072	0.0310	0.0086	0.0028	0.0025	0.0919
PA	0.0055	0.0091		0.0054	0.0035	0.0061	0.0055	0.0103
HA	0.0018	0.0175	0.0011		0.0052	0.0064	0.0119	0.0146
NP	0.0006	0.0489	0.0015	0.0029		0.0060	0.0079	0.0003
NA	0.0002	0.0001	0.0023	0.0561	0.0143		0.0066	0.0025
M	0.0022	0.0146	0.0064	0.0146	0.0104	0.0037		0.0068
NS	0.0068	0.0033	0.0083	0.0038	0.0004	0.0011	0.0042	

Supp. Table 5: Cross validated log-likelihoods for all observed triplets.

(large)

Supp. Table 6: Cross validated log-likelihoods for all observed quadruplets.

(large)