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@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 spatial point process 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.

45 **Introduction**

46 Influenza A virus, part of the *orthomyxoviridae* family, has a segmented genome of eight
47 distinct viral RNA segments coding at least 11 major proteins and several auxiliary peptides.
48 Most notable of the 11 viral proteins are hemagglutinin (HA) and neurominidase (NA), the
49 canonical *H* and *N* in influenza strain designation. A segmented genome offers potential
50 evolutionary advantages during viral replication, in the form of reassortment. The exchange of
51 genomic material between two distinct viral strains in a co-infected cell often causes a shift
52 within the genome, greatly increasing viral genetic diversity. In only the last century, influenza
53 pandemics of 1957 (H2N2), 1968 (H3N2), and 2009 (H1N1) have emerged through
54 reassortment of at least two viruses [1]. Segmented genomes also come with an inherent
55 mechanistic challenge: ensuring that progeny receive a complete set of genomic segments.
56 Random packaging is possible but comes at the cost of producing an overwhelming majority of
57 progeny that are not viable [2]. Evidence suggests that there exists an active mechanism of
58 selective packaging ensuring progeny viability through efficient and accurate genomic
59 packaging [3, 4].

60

61 Within a virion, the viral genome is organized into individual viral ribonucleoprotein (vRNP)
62 complexes, composed of the viral RNA, virally encoded nucleoprotein (NP), and a
63 heterotrimeric polymerase complex made up of PB1, PB2 and PA. The vRNP structure is known
64 classically as a helical panhandle, where the RNA is wrapped around NP with RNA bases
65 exposed and the 3' and 5' ends associated with the polymerase complex [3]. We have recently
66 demonstrated that the structure of vRNP is more complex than the classically depicted "beads-

67 on-a-string" schematic. The NP and viral RNA associate in a non-uniform manner, where regions
68 of the RNA are unbound by NP and capable of forming complex structures [5]. Macro-
69 organization of vRNPs within the viral capsid show tight packaging, organized in a '7+1'
70 orientation, a single vRNP center shaft surrounded by seven others (4-6).
71
72 Given the need to package one copy of all eight vRNP segments, there is strong evidence for
73 the existence of a selective packaging mechanism prior to budding [3, 6-9]. Intersegmental
74 RNA-RNA interactions have been proposed to mediate selective packaging of all eight influenza
75 vRNP segments. The 5' and 3' regions of each segment have been implicated to be essential for
76 efficient packaging and may be critical for the proposed RNA-RNA interactions [3, 10-13].
77 Previous studies using *in vitro* transcribed viral RNA have successfully demonstrated RNA-RNA
78 binding. These studies also show that a set of eight reproducible interactions can suffice to
79 form a full genome complex [14-16]. However, these studies were performed on viral RNA in
80 the absence of NP which would alter the regions of viral RNA accessible for the proposed RNA-
81 RNA interactions. Recent *in vivo* experiments have also pointed towards RNA-RNA interaction
82 as the driving mechanism for selective packaging [17]. While there is consensus of its existence,
83 a clear network of segment interactions leading to a full genome complex has yet to be
84 elucidated.
85
86 We and others have previously used fluorescence *in situ* hybridization (FISH) studies to visualize
87 the intracellular localization of multiple vRNP segments during a productive viral infection [18,
88 19]. These studies, performed eight hours post infection to capture an initial infection cycle and

89 reduce complication of cytoplasmic vRNP from a subsequent infection, have shown that distinct
90 vRNP segments colocalize within the cytoplasm. Based on our previous studies, we proposed a
91 model whereby vRNP segments are exported from the nucleus, the site of vRNP synthesis, as
92 subcomplexes that undergo further assembly en route to the plasma membrane through
93 dynamic fusion or colocalization events [19]. We also performed a series of binary
94 colocalization comparisons, which would postulate a simple linear vRNP interaction network.
95 However, no such network was identified, suggesting the presence of a more complex vRNP
96 interaction network, including higher order subcomplexes as intermediate steps [19].
97
98 The lack of a method to image all eight vRNP segments within a single infected cell has limited
99 our ability to examine the precise spatial relationship between segments. In this study, we
100 developed a model based approach, rooted in point process theory, to quantify vRNP spatial
101 dependence as a metric for RNA-RNA interaction using multi-color FISH images. A spatial point
102 process model is a statistical model that captures the probability that an event will happen at
103 any position within a given geometry. The probability can depend on the spatial distributions
104 of other events or objects, often referred to as covariates or factors. In this context, an event is
105 the observance of fluorescent signal (corresponding to one or more vRNPs) at a given location.
106 We have previously used point process modeling to construct generative models reflecting the
107 spatial dependencies between various punctate cell organelles and other cellular structures,
108 such as the cell and nuclear membranes, microtubules and the endoplasmic reticulum [20, 21].
109 Related analytical methods have been used to analyze clustering of molecules reflecting spatial
110 dependencies between copies of the same structure [22, 23] Here we use point process models

111 to present a statistically rigorous analysis of intracellular influenza A genome assembly
112 dynamics from a spatial perspective during a productive viral infection.
113
114 For an imaged cell, the locations of each vRNP segment can be represented as “realizations” of
115 an underlying probability density specific to that vRNP, calculated over the cytoplasm, and
116 dependent on the spatial proximity to various other structures within the cell. That is, at any
117 given position within the cell, there exists a probability of observing a single segment, of a
118 particular identity, that depends on the positions of cellular organelles and other vRNP
119 segments. By modeling the locational densities of each vRNP, we can therefore learn a
120 dependency network from which likely vRNP interactions are implied.
121
122 FISH imaging yields many distinct observations of individual cells, each with multiple vRNP
123 patterns observed. These point patterns, when viewed as realizations of point processes, allow
124 for statistical learning of spatial dependencies over many replicates for each vRNP segment. By
125 defining a set of covariates as the minimum distances between a given segment and other
126 segments observed within the same cell, we produce a set of models that describe the spatial
127 relationship between distinct segments. The extent to which two (or more) vRNPs are
128 dependent upon each other (e.g., likely to be found at nearby positions) is reflected in the
129 likelihood that images of those vRNPs would have been produced by a learned model: high
130 likelihoods signify that the positions of a given vRNP can be predicted well from the others in
131 that model. Model likelihoods can then be seen as a metric of vRNP interaction, with higher
132 likelihoods indicating more probable vRNP association. These likelihoods can then be used to

133 construct the most probable sequence of interactions to form the full genome with methods
134 borrowed from evolutionary tree construction [24].
135
136 **Results:**
137
138 **Imaging influenza virus infected cells**
139 To investigate the spatial dependence between vRNP segments, we utilized images from a
140 previous study of infected cells stained for different combinations of vRNP segments using four-
141 color FISH at eight hours post infection [19]. The data set included 14 different probe
142 combinations that covered all pairwise vRNP associations and 32 out of 56 possible triple vRNP
143 combinations (Supplemental Table 1 and 2). Prior to initiating pattern analysis, each image was
144 processed to remove background, isolate individual cells, and find punctate vRNP segments.
145 Figure 1a and b presents representative raw and segmented images with detected points.
146
147 **Assessing randomness of the spatial distributions of vRNP**
148 We began by determining the extent to which the individual vRNPs could be considered
149 randomly distributed throughout the cell, since our search for spatial dependencies between
150 different vRNPs would be illogical if their positions were all random. We defined the null
151 hypothesis as complete spatial randomness, as would be exhibited by a homogenous Poisson
152 process, and estimated a p-value for this hypothesis using Monte Carlo methods. Briefly, for a
153 given vRNP in a particular cell, we started by segmenting the cytoplasm and counting the
154 number of points observed within. We then simulated (generated) many random point

155 configurations, of the same number of points, within the segmented cytoplasm. We calculated
156 a test statistic for each random configuration (see Methods) and used these to assign a p-value
157 to the test statistic for the observed pattern for that cell; these were averaged for each vRNP.
158 As shown in Table 1, every vRNP segment showed significant deviation (at the $p < 0.05$ level),
159 with low variance, from a spatially random distribution. This indicates the presence of a spatial
160 trend and some dependency on cellular structures for their location.
161
162 **Modeling cell and nuclear membrane dependence**
163 Given that the localizations of vRNPs are not random, we sought to determine if some
164 relationship to the cell and nuclear membranes could explain the spatial trend of vRNP
165 segments. Three covariates were defined for each position in the cell (minimum distance to the
166 cell membrane, minimum distance to the nuclear membrane, and the minimum ratio of cell to
167 nuclear distance for each point) and models were created for each vRNP using each of the
168 covariates separately. We estimated the accuracy of each model by cross validated likelihood
169 calculation (see Methods), where values closer to 0 are more likely. It is important to note that,
170 because actual likelihoods were calculated, they can be compared between models of different
171 complexity.
172
173 Dependency on the cell membrane produced the lowest likelihood (worst) models relative to all
174 others (Table 2) for all vRNP segments. Congregation of vRNP to the apical cell membrane is a
175 potential skewing factor, since there will be a high concentration of all segments at this cellular
176 location. The many observed points within close proximity to the membrane potentially caused

177 low fitted probability of observing other points closer to the nucleus. Cell shape differences
178 combined with lower resolution in the z-dimension may have also influenced these models.
179
180 For most vRNPs, models incorporating nuclear distance showed higher likelihoods than those
181 using cell membrane distance (Table 2). Some vRNP segments, such as PB2, PB1, PA and NS,
182 showed greater dependency on nuclear distance than the other segments. The ratio of cell
183 membrane to nuclear distance generally produced models better than cell membrane alone,
184 but they were worse than nuclear alone in all cases except NA (Table 2). Thus nuclear distance
185 provides the most cogent model for dependence of vRNP localization on a cellular structure.
186
187 **Inter-segment dependence in pairs, triplets, and quadruplets**
188 Every pair of vRNP segments was observed in at least three cells, allowing for modeling of
189 vRNP-vRNP dependence among all segment pairs. For each pair of vRNP segments, two models
190 were trained, switching the dependent (primary) and independent (secondary) patterns. Since
191 the minimum inter-pattern distance measure is not commutative, the two models need not be
192 the same. More simply, a given vRNP location may depend on its proximity to another vRNP,
193 but the converse is not necessarily true. This can be observed when one vRNP segment is
194 found in two distinct locations, and a second vRNP segment is found in only one of those
195 locations. The second vRNP segment would be observed to have a high likelihood for being
196 predicted from the first, but the first would not have a high likelihood of being predicted from
197 the second.
198

199 In all cases, pairwise inter-pattern distance models (Table 3) for a given vRNP had higher
200 likelihoods than those produced for that vRNP with only the cell and/or nuclear membrane
201 factors (Table 2). Some pairs showed higher likelihoods than others, such as PB1-NP, PB1-M,
202 PA-NP, HA-PB1, and NP-PB1. These pairings also showed relatively low variance in cross-
203 validated likelihood estimates (Supplemental Table 3). Many of the highest likelihood vRNP
204 pairs are observed between segments encoding the viral polymerase components (PB2, PB1, PA
205 and NP). Given the dependence between these proteins both structurally and functionally [25-
206 27], it is possible that these segments evolved a dependence upon each other to ensure their
207 joint packaging in a virion (a possibility that would require extensive additional work to
208 examine). Figure 2 highlights the pairwise models in Table 4 by illustrating vRNP pairs that are
209 below 0.23 log likelihood. Overall, vRNPs showed the greatest dependence on PA and PB1, with
210 the most connections between these two segments and the others. The least dependence was
211 observed for NA and NS vRNP segments.
212
213 Extending the pairwise models, nuclear distance was added as another covariate. Minor
214 improvements in model likelihood were seen in some models (Table 4) with others showing
215 decreased likelihood. Overall there was an increase in the variance of our estimates
216 (Supplemental Table 4). Overfitting to nuclear distance may be the cause of this increased
217 variance.
218
219 Thirty-two of 56 possible triplet vRNP segment complexes (referred to as triplets) were present
220 within the dataset analyzed and could therefore be modeled within the same cell. This process

221 allows for modeling of complex vRNP relations containing more than two vRNP segments.
222 Three models were trained for each observed triplet, one for each vRNP depending on the
223 other two. The highest likelihood model was chosen. Model likelihoods tended to be higher
224 than that of either pair model for a given vRNP (Table 5). We observed that the most likely
225 triplet models for each vRNP included four of the five most likely vRNP pairings observed in pair
226 likelihood models (Table 3), (namely PB1-NP, PA-NP, HA-PB1, and NP-PB1). However, not all of
227 the most likely pairwise associations (e.g., PB1-M) were observed in the most likely triplet
228 models, highlighting that models based solely on pairwise comparisons might not accurately
229 represent larger order vRNP complexes. Model likelihoods for all triplets observed can be found
230 in Supplemental Table 5.
231
232 A small subset of all possible four vRNP segment complexes (referred to as quadruplets) were
233 contained within the dataset and models for predicting one of the four from the other three
234 were also constructed. Again, considerable overlap is seen between the most likely quadruplet
235 models and both triplet and pair models (Table 6), such as PB2-PB1-PA and PB1-PA. The
236 consistency in vRNP composition between observed likely quadruplet, triples, and pairs further
237 validates the spatial dependence of vRNP segments and the presence of vRNP subcomplexes as
238 influenza A assembly intermediates. Model likelihoods for all quadruplets observed can be
239 found in Supplemental Table 6.
240
241 One of our major goals was to attempt to extrapolate higher order relationships from the pair,
242 triple and quadruplet models. To project how accurately this might be done, we compared our

243 observed triple and quadruple models with predicted models calculated from any of the four
244 pair models (vRNP pairs (Table 3), vRNP pairs + nucleus (Table 4), vRNP pairs +cellular
245 membrane, and vRNP pairs +nuclear/cell membrane ratio) (Figure 3a). We also compared
246 observed quadruplet likelihoods to those predicted using both the pair and triple models,
247 similar to the pair models any of the 4 models including the various cellular features were
248 included (Figure 3b). Most of our observed triples and quadruplets could be predicted with
249 modest accuracy from the pair models but the overall r^2 value was only 0.302. This was due to
250 inaccuracy in predicting the triple and quadruple clusters containing the PB2-NA pair with
251 either NS, PB1 or both. These combinations were predicted to occur much more often than
252 was observed in the experimental data sets. The pairwise likelihood of NA with PB2 is >-0.6,
253 which is consistent with the observed likelihood and may be the driving force behind the score
254 compared to the other pairwise interactions. Analysis of the predicted triples and quadruples
255 after removing the PB2-NA containing clusters resulted in a r^2 value increases to 0.642 (we
256 cannot justify this removal on a statistical basis but rather on the observation that the
257 combinations with the poorest predictions all involved the same pair). Predicting quadruples
258 from pairs and triples resulted in a r^2 value of 0.450, and with removal of the same quadruple it
259 rose to 0.714. Overall, these data support the notion that pairwise point process models can be
260 used to predict higher order complexes with moderate accuracy, and that prediction accuracy
261 increases if higher order models are used.
262
263
264 **Interaction network construction**

265 Given these results, we sought to create a model of the set of vRNP interactions (complexes)
266 that occur at all steps in the formation of a complete viral genome. To do this, we represented
267 all possible interactions between vRNP segments as a weighted, directed, acyclic graph with
268 nodes labeled by each unique, unordered set of vRNP segments. A path from all single nodes to
269 the root represents a set of vRNP interactions resulting in a full genome complex (see
270 Methods). For illustration, if all eight vRNP segments directly formed a complete set without
271 forming any intermediate complexes (a highly unlikely scenario), this would be represented by
272 weights of one for the connection between each vRNP and the complete vRNP supramolecular
273 complex containing all eight segments and weights of zero for all other complexes. Such a
274 scenario would result in a tree with the root connected to each of the leaves
275
276 In our analysis, we estimated the weight for each possible intermediate vRNP complex
277 assembly subcomplex to contain at least two and up to eight segments based upon either the
278 observed or predicted likelihoods (see Methods). Dynamic programming, an efficient method
279 for recursively finding the highest scoring path to a given node, was then used to find the most
280 probable path within the graph using the highest likelihood trained model for a given nodes
281 from those with or without cell and/or nuclear features (Figure 4a). The resultant tree produces
282 two distinct clusters HA, M, NA, NS (hereafter referred to as C1) and PB1, PB2, NP (C2) that
283 merge with PA as a final step towards forming the full genome. This network suggests three
284 steps to form a complete set of all eight vRNP segments: 1) formation of C1, 2) formation of C2
285 and 3) addition of PA to C1 and C2 to form the final product.
286

287 To assess whether a particular vRNP localization dataset was driving this network, we
288 performed a series of network analyses with data excluded. First, the contribution of observed
289 quadruplets and triplets was assessed by excluding observed quadruplets (Figure 4b) and
290 without observed quadruplets and triplets (Figure 4c). Interestingly, exclusion of all observed
291 quadruplets did not drastically alter the network, with both C1 and C2 still being present,
292 although formation of C1 required two steps. However, networks built only with observed pairs
293 (excluding triplets and quadruplets) resulted in a quite different network lacking larger
294 subcomplexes (Figure 4c), reflecting the fact that pairs were only able to make approximate
295 estimates of higher order interactions
296
297 Given that the dynamic programming approach yields only a single most likely tree, we sought
298 to assess the robustness of the inferred tree to potential inaccuracy in the estimated
299 dependencies. We therefore assessed the stability of the tree by adding various amounts of
300 random noise to the estimated likelihoods (see Methods). Surprisingly, the addition of noise of
301 up to a quarter of the mean log-likelihood had no effect on the network found, yielding the
302 same network (Figure 4a). As noise levels increased to half of the mean log-likelihood, one
303 other tree was generated (Figure 4f). This demonstrates the stability limit of the construction
304 but notably this revised tree still contains clusters that resemble the original. If the original tree
305 was based on biased likelihoods from imaging artifacts, the addition of noise would have
306 resulted in an altered tree network at lower noise levels and with higher frequency, but since
307 this did not occur, we are confident in the accuracy of our original likelihood estimations.
308

309 Finally, to exclude the possibility that datasets directly measuring C1 and C2 (reactions L and B,
310 respectively) were biasing the network towards inclusion of those sets of vRNPs, we performed
311 a network analysis of a dataset where the measured likelihoods of these two clusters was
312 excluded. Removal of data derived from images of the sets of vRNPs in C1 and C2 (Figure 4e)
313 yielded networks that resemble Figure 4a, indicating that even in the absence of data from
314 these clusters, similar clusters will form during assembly. Removal of each cluster alone was
315 recapitulated by lower order interactions. Removal of C1 [HA, M, NA, NS] data resulted in the
316 same tree presented in figure 4b, where HA, M, NA, NS form in two steps from a triplet and a
317 single vRNP segment. Exclusion of C2 [NP, PB1, PB2} result in a tree similar to figure 4d, where
318 C2 cluster forms from a pair and a single segment (Figure 4d). Removal of data was also
319 combined with the addition of noise to further test stability. Predominantly, the same networks
320 were generated with and without noise. As noise levels increased in the exclusion
321 constructions, a tree with HA, M, NS, *PB1* in place of C1 and NP, *PA*, PB2 in place of C2, where
322 PB1 and PA were not observed in the C1 or C2 original clusters respectively (Figure 4f).
323 Together, these constructed networks suggest a primary interaction scheme that can be further
324 experimentally tested and the importance of triplet and quadruplet observations in predicting
325 higher order vRNP complexes. Other, very low frequency trees that were generated in the
326 presence of high levels of noise are located in the reproducible research archive.
327

328 **Discussion:**

329 Influenza A vRNP segments selectively assemble within an infected cell to produce progeny
330 virions containing one copy of all eight segments. The mechanism driving selective assembly is

331 still largely unknown, but RNA-RNA interaction between the segments has been proposed [14-
332 16]. In this study, we have modeled the *in vivo* spatial dependencies of influenza A vRNP
333 segments using multi-color fluorescent *in situ* images to generate possible vRNP-vRNP
334 interaction networks and propose a new perspective on genome packaging during viral
335 replication. By utilizing a rigorous statistical framework, we have extended the efforts of
336 previous groups and present a novel method for construction of vRNP interaction networks
337 based on their precise spatial information within actively infected cells. Using spatial proximity
338 as a proxy for physical interaction, point process modeling in this study suggests clear spatial
339 dependencies between certain vRNP segments, and confirms our previous observations of
340 subcomplex formation during cytoplasmic transport (18).
341
342 Previous studies using *in vitro* transcribed RNA suggest multiple interactions between vRNP
343 segments are expected [3, 6, 14, 16, 28]. Similarly, the modeled likelihoods show potentially
344 multiple interactions for each vRNP segment. Electron tomography studies have revealed a
345 conserved '7+1' supramolecular structure within the viral interior, where a center shaft is
346 surrounded by seven remaining segments [3, 6, 28], demonstrating an ordered process in
347 genome assembly.
348
349 Our results suggest two candidates for the center shaft or 'master segment': PA and PB1. From
350 the six constructed networks, PA is the most variable in that it is least often paired early in
351 assembly but, with the addition of noise and exclusion of data, may interact with both primary
352 observed clusters. This would be expected if a segment was evenly dependent on most others,

353 a possible signature of a core segment. PB1 also behaves similarly but to a lesser degree. Both
354 PB1 and PA show the highest average pairwise model likelihood over all other segments.
355
356 A few vRNP pairs were seen often in the most likely pairwise, triplet, and quadruplet models
357 and in the final constructed network. PB1-PB2 and HA-M, show commutative relationships,
358 appearing in one of the top most likely triplet and quadruplet models for each vRNP within the
359 pair. Some non-commutative dependencies were also seen, most notably in NP depending on
360 PB2 in both triplet and quadruplet models with little PB2-NP dependence.
361
362 By using dynamic programming, we were able to generate the most likely interaction network
363 based on all models. While this network was surprising robust to the addition of noise to the
364 model likelihoods, subtle variations in the vRNP interaction network were observed with the
365 exclusion of higher order datasets, suggesting that there may exist multiple interaction
366 pathways all acting at once in a single cell. Some of the variation in our results may also be due
367 to the presence of some lower specificity vRNA-vRNA interactions also occurring during
368 genome assembly. This would explain some of the highly likely pairwise interactions observed
369 that were not included in the final interaction network, such as PB1 and HA. In addition, vRNP
370 networks could also change over the course of a viral infection when packaging fidelity may be
371 compromised [29]. Our current data set only considers one time point, eight hours post
372 infection, which represents the time of initial virion release from infected cells, and would miss
373 alternate networks present at later time points. Therefore, the analysis presented here may
374 only capture a snapshot of a dynamic assembly process that could change throughout an

375 infection. A system with multiple possible interaction networks could be potentially
376 advantageous in both ensuring packaging of a full genome in viral particles and in future
377 reassortment events. For example, reassortment of one vRNA segment during a coinfection
378 may decrease the binding efficacy necessary for a single interaction network but a second
379 network may serve to rescue viral viability while also increasing genetic diversity.
380
381 With this, we also point out the consistent presence of the PB2, PB1 and NP cluster. Since these
382 three proteins function together, with PA, to form the viral polymerase, which promote vRNA
383 replication and transcription, this cluster is potentially important. Viruses capable of packaging
384 complementary polymerase segments into progeny virions will have a fitness advantage
385 compared to viruses with polymerase mismatches. This phenotype has been observed in
386 experimental reassortment experiments between 2009 H1N1 pandemic and seasonal H3N2
387 viruses[25, 27].
388
389 More generally, the computational pipeline presented may function as a useful tool in
390 elucidating spatial dependency over a wide range of biological phenomena observable through
391 microscopy. As we have demonstrated, point process models can capture key aspects of point
392 distributions while retaining a basis in probability. Network construction synthesizes the results
393 derived through modeling into a cogent, most-likely set of spatial interactions or dependencies.
394
395 Future work in spatial modeling of influenza A packaging should incorporate a temporal
396 dimension through live cell imaging. Point process models have been adapted for

397 spatiotemporal modeling [30]. Incorporation of intracellular markers important for influenza A
398 vRNP transport may increase the accuracy of models. Influenza A vRNP transport from the
399 nucleus, the site of vRNP synthesis, to the plasma membrane is a complex process utilizing a
400 variety of host proteins [31]. Rab11A-containing vesicles are thought to be the primary mode of
401 transport although there is evidence that a Rab11A-independent mechanism exists [32-34]. A
402 recent study has implicated the ER to mediating transport of vRNP segments through anchored
403 Rab11A proteins [35]. Novel methods in fluorescent multiplexing [36] present an exciting
404 opportunity for observing many cellular structures which will provide a holistic image the
405 spatial dependence of vRNP segments upon subcellular structures and to each other.
406
407 **Methods:**
408 **Multi-color fluorescent in situ hybridization images**
409 This study used previously published multi-color FISH images from [19] that were generated at
410 the National Institutes of Health. Briefly, MDCK cells were infected with recombinant
411 WSN/1933 H1N1 for 8 hours and then fixed and stained with FISH probes, obtained from
412 Biosearch technologies, against four distinct vRNP segments. DAPI was included as well to label
413 DNA. Multi-color FISH samples were imaged on an Leica SP5 white light laser to ensure spectral
414 separation of the five colors (Dapi, Alexa 488, Quasar 570, Cal Fluor Red 590, Quasar 670). The
415 specificity of the probes and spectral separation between fluorophores was previously
416 confirmed [19].
417
418 **Image preprocessing, cell segmentation and point detection**

419 All image preprocessing was performed in MATLAB, v. 2015a. Prior to point detection and
420 segmentation, image noise was removed by convolution with a Gaussian mask. Each image
421 channel was denoised individually. As described in [19], each image was captured with 0.17 um
422 z-step size spanning the entire cell volume, defined for each cell using both the nuclear and
423 FISH staining to define the apical cell membrane. Each image was captured with a pixel size of
424 ~50x50x168 nm.
425
426 Individual vRNP segments were identified by finding connected areas of signal within the
427 denoised image. The center of each object was taken as the maximum intensity pixel within,
428 yielding a set of point coordinates. Note that due to the thickness of the z sections, the
429 apparent distance between points may be an underestimate of the true distance, but this effect
430 is expected to average out when considering many points.
431
432 Since there was no fluorescent tag for cell membrane components, we estimated the cell
433 boundary using the vRNP images. We assumed that fluorescent signal would be denser within
434 the cell cytoplasm than in the surrounding area with cell-cell junctions subtly defined by
435 'valleys', curves of low signal density whose normal vectors point towards increasing density.
436 The gradient of point density over the image was then used to segment individual cells from
437 both the background and each other by the mean-shift algorithm [37]. Hand segmentation was
438 also used to ensure accurate cell membrane segmentation. In most cases, the convex hull of all
439 points within an identified cell region then defined the cell membrane. The nucleus was
440 segmented through simple thresholding and smoothing.

441

442

443 **Point processes and patterns**

444

445 Within each cell, the vRNP segments form a point pattern, **x** = {x_1, x_2, …, x_n}, where *n* is the
446 number of observed points and x_i is the 3-dimensional coordinate vector for point *i*. The point
447 pattern is defined over a bounded region, *W*, the segmented cell cytoplasm. **x** is then viewed as
448 a "realization" (an output) of some random point process **X**, the generating distribution for all
449 patterns of the particular vRNP identity. The process **X** then represents the culmination of
450 biological factors that determines vRNP segment location, eg. nuclear export, directed
451 transport over the microtubule network, inhibition by other organelles, etc.

452

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

457

458 $$f(\boldsymbol{x}|n) = \frac{1}{Z}\lambda^n \qquad (1)$$

459

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

463

464 **Characterizing spatial randomness**

465

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

471

472 $$K_{poi}(r) = \pi r^2 \quad (2)$$

473

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

477

478 $$T = \int_0^{r_{max}} \left(\widehat{K}(r) - K_{poi}(r)\right)^2 dr \quad (3)$$

479

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

485

486 $$p = \frac{1+\sum_{i=1}^{m} I(t_{obs} \geq t_i)}{m+1} \quad (4)$$

487

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

489 vRNP.

490

491 **Modeling spatial dependence on cellular structures**

492

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

494 dependent locational density. Per point "factors" (e.g., distance to cell membrane) are used to

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

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

497 can be calculated for each point x_i. The Inhomogenous Poisson point density function is:

498

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

499

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

501 then defined by:

502

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

503

504

505 To quantify vRNP dependence on cellular structures, we defined the first factor (f_1) as the
506 distance to the nearest point on the nuclear membrane, the distance to the nearest point on
507 the cell membrane, or the ratio of cell to nuclear distance.

508

509

510 **Modeling spatial dependence between vRNPs**

511

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

516

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

517

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

522

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

525 maximum pseudolikelihood [38]. The basic idea is that a given set of parameters can be used to
526 calculate a value that is proportional to the likelihood (which is referred to as a
527 pseudolikelihood) of observing a point at a given position, and the parameters can then be
528 adjusted to give the highest total pseudolikelihood for available observations. Hold one out
529 cross validation was used to assess the quality of fit over all instances yielding cross-validated
530 pseudolikelihood estimated parameters and variances. These fitted parameters were then used
531 to convert the pseudolikelihood to a true model likelihood by estimating the proportionality
532 constant as previously described (19).
533
534 There is some variation in the number of images in which different vRNPs were visualized
535 (Supplemental Table 1). This ranged from 4 for PB1 to 9 for M (the others were present 6 or 7
536 times). As pointed out by a reviewer, this difference raises the possibility that a vRNP that was
537 imaged less frequently might be found to be underrepresented in the extent to which other
538 segments were found to depend upon it. As seen in Table 3, this did not turn out to be the
539 case, as PB1 and M showed similar average likelihoods and more vRNPs showed strong
540 dependence on PB1 than on the others. Similarly, PB1 was not underrepresented in high
541 scoring triples (Table 5).
542
543 **Interaction network assembly**
544
545 We formulate the task of finding the most probable set of vRNA interactions yielding a full
546 genome as a graph problem. First, let the set of nodes in the graph be every possible

547 combination of vRNAs ranging in size from a single vRNA to all 8 segments. Directed edges in
548 the graph connect subset to superset nodes. We weight each edge of the graph by the
549 likelihood of the most likely model (chosen from models with only vRNP dependencies, and
550 those with additional dependencies on either nuclear distance, cell membrane distance, or cell
551 to nuclear distance ratio) that results in the superset and contains the subset as a primary or
552 secondary segment (Figure 1). Edges that represent models that have not been observed
553 remain unweighted.
554
555 From this graph, a set of interactions yielding the full genome is a tree with leaves as the vRNA
556 singletons and root as the full 8-mer of vRNA segments. In these trees, we require each node to
557 have at least two incoming edges, barring the leaf nodes. The set of incoming edges to a given
558 node must emanate from nodes whose labels, when combined, exactly equal the given node. A
559 set of incoming edges essentially 'combines' nodes, directly corresponding to the physical event
560 of vRNP segments binding. The most likely set of interactions is also the most highly weighted
561 tree. The problem of finding the most probable interaction network then reduces to finding the
562 highest weighted tree in the graph. This problem can easily be solved through dynamic
563 programming within a paradigm often used in evolutionary tree construction [24].
564
565 The likelihoods of models involving multiple vRNP segments can be taken as a measure of
566 spatial dependence of primary segments on secondary segments. vRNP segments that truly
567 interact are expected to be highly dependent either unidirectionally or bidirectionally. The cross
568 validated model likelihoods can be used to weight groups of edges in the graph. Since the graph

569 also has edges that were not observed in actual images (eg: any complex of more than 4
570 vRNAs), composite or inferred likelihoods were generated through recursion in a dynamic
571 programming scheme. For each unobserved complex, we took the most likely set of observed
572 models that could generate the complex.
573
574 For a given unobserved complex $N = \{N_1, N_2, N_3, \ldots\}$,
575
576 $$L(N) = \max(L(N_1|N_2,N_3,\ldots), L(N_2|N_1,N_3,\ldots), \ldots)$$
577
578 Then for some likelihood in the above, $L(N_1|N_2)$ where $N_1 = (n^1_1, \ldots n^1_k)$ and $N_2 = (n^2_1, \ldots n^2_l)$,
579 $|N_1| = k$, $|N_2| = l$,
580
581
$$L\big((N_1|N_2)\big) = \sum_{i=1}^{k} \underset{N_j^2 \in \, all\ unique\ sets\ of\ N_2}{\text{average}} (L(n_i^1|N_j^2))$$
582
583 **Interaction network stability analysis**
584
585 To assess the stability of the generated interaction network, two procedures of perturbation
586 were performed: introduction of noise and removing certain high likelihood models. We
587 simulated noise as a normal random distribution centered at 0 with varying standard
588 deviations, termed the noise levels. Under increasing noise levels, each observed model

589 likelihood was amended with a noise value drawn from this distribution prior to interaction
590 network construction. For each noise level, we simulated 100 trials and tallied the proportion of
591 output trees that contained each possible edge. Trees that align with the non-noisy assembly
592 signal stability while highly variable trees signal instability.
593
594 **Reproducible research archive**
595
596 All images, derived data and source code will be available upon publication at
597 http://murphylab.cbd.cmu.edu/software.
598
599 **REFERENCES**
600
601 **Bibliography**
602

603 1. Morens DM, Taubenberger JK. Pandemic influenza: certain uncertainties. Rev Med
604 Virol. 2011;21(5):262-84. Epub 2011/06/28. doi: 10.1002/rmv.689. PubMed PMID:
605 21706672.
606 2. Sidorenko Y, Reichl U. Structured model of influenza virus replication in MDCK cells.
607 Biotechnol Bioeng. 2004;88(1):1-14. doi: 10.1002/bit.20096. PubMed PMID: 15384040.
608 3. Hutchinson EC, von Kirchbach JC, Gog JR, Digard P. Genome packaging in influenza A
609 virus. J Gen Virol. 2010;91(Pt 2):313-28. Epub 2009/12/04. doi: 10.1099/vir.0.017608-0.
610 PubMed PMID: 19955561.
611 4. Nakatsu S, Sagara H, Sakai-Tagawa Y, Sugaya N, Noda T, Kawaoka Y. Complete and
612 Incomplete Genome Packaging of Influenza A and B Viruses. MBio. 2016;7(5). doi:
613 10.1128/mBio.01248-16. PubMed PMID: 27601575; PubMed Central PMCID:
614 PMCPMC5013298.
615 5. Lee N, Le Sage V, Nanni AV, Snyder DJ, Cooper VS, Lakdawala SS. Genome-wide
616 analysis of influenza viral RNA and nucleoprotein association. Nucleic Acids Res.
617 2017;45(15):8968-77. doi: 10.1093/nar/gkx584. PubMed PMID: 28911100.
618 6. Noda T, Sagara H, Yen A, Takada A, Kida H, Cheng RH, et al. Architecture of
619 ribonucleoprotein complexes in influenza A virus particles. Nature. 2006;439(7075):490-2.
620 Epub 2006/01/27. doi: 10.1038/nature04378. PubMed PMID: 16437116.

621 7. Chou YY, Vafabakhsh R, Doganay S, Gao Q, Ha T, Palese P. One influenza virus
622 particle packages eight unique viral RNAs as shown by FISH analysis. Proc Natl Acad Sci U S
623 A. 2012;109(23):9101-6. Epub 2012/05/02. doi: 10.1073/pnas.1206069109. PubMed
624 PMID: 22547828; PubMed Central PMCID: PMC3384215.
625 8. Noda T, Kawaoka Y. Packaging of influenza virus genome: robustness of selection.
626 Proc Natl Acad Sci U S A. 2012;109(23):8797-8. Epub 2012/05/23. doi:
627 10.1073/pnas.1206736109. PubMed PMID: 22615368; PubMed Central PMCID:
628 PMC3384147.
629 9. Isel C, Munier S, Naffakh N. Experimental Approaches to Study Genome Packaging of
630 Influenza A Viruses. Viruses. 2016;8(8). doi: 10.3390/v8080218. PubMed PMID:
631 27517951; PubMed Central PMCID: PMCPMC4997580.
632 10. Fujii Y, Goto H, Watanabe T, Yoshida T, Kawaoka Y. Selective incorporation of
633 influenza virus RNA segments into virions. Proc Natl Acad Sci U S A. 2003;100(4):2002-7.
634 doi: 10.1073/pnas.0437772100. PubMed PMID: 12574509; PubMed Central PMCID:
635 PMCPMC149948.
636 11. Marsh GA, Rabadan R, Levine AJ, Palese P. Highly conserved regions of influenza a
637 virus polymerase gene segments are critical for efficient viral RNA packaging. J Virol.
638 2008;82(5):2295-304. doi: 10.1128/JVI.02267-07. PubMed PMID: 18094182; PubMed
639 Central PMCID: PMCPMC2258914.
640 12. Essere B, Yver M, Gavazzi C, Terrier O, Isel C, Fournier E, et al. Critical role of
641 segment-specific packaging signals in genetic reassortment of influenza A viruses. Proc Natl
642 Acad Sci U S A. 2013;110(40):E3840-8. doi: 10.1073/pnas.1308649110. PubMed PMID:
643 24043788; PubMed Central PMCID: PMCPMC3791739.
644 13. Gerber M, Isel C, Moules V, Marquet R. Selective packaging of the influenza A
645 genome and consequences for genetic reassortment. Trends Microbiol. 2014;22(8):446-55.
646 doi: 10.1016/j.tim.2014.04.001. PubMed PMID: 24798745.
647 14. Fournier E, Moules V, Essere B, Paillart JC, Sirbat JD, Cavalier A, et al. Interaction
648 network linking the human H3N2 influenza A virus genomic RNA segments. Vaccine.
649 2012;30(51):7359-67. Epub 2012/10/16. doi: 10.1016/j.vaccine.2012.09.079. PubMed
650 PMID: 23063835.
651 15. Fournier E, Moules V, Essere B, Paillart JC, Sirbat JD, Isel C, et al. A supramolecular
652 assembly formed by influenza A virus genomic RNA segments. Nucleic Acids Res.
653 2012;40(5):2197-209. Epub 2011/11/15. doi: 10.1093/nar/gkr985. PubMed PMID:
654 22075989; PubMed Central PMCID: PMC3300030.
655 16. Gavazzi C, Isel C, Fournier E, Moules V, Cavalier A, Thomas D, et al. An in vitro
656 network of intermolecular interactions between viral RNA segments of an avian H5N2
657 influenza A virus: comparison with a human H3N2 virus. Nucleic Acids Res.
658 2013;41(2):1241-54. Epub 2012/12/12. doi: 10.1093/nar/gks1181. PubMed PMID:
659 23221636; PubMed Central PMCID: PMC3553942.
660 17. Gilbertson B, Zheng T, Gerber M, Printz-Schweigert A, Ong C, Marquet R, et al.
661 Influenza NA and PB1 Gene Segments Interact during the Formation of Viral Progeny:
662 Localization of the Binding Region within the PB1 Gene. Viruses. 2016;8(8). doi:
663 10.3390/v8080238. PubMed PMID: 27556479; PubMed Central PMCID: PMCPMC4997600.
664 18. Chou YY, Heaton NS, Gao Q, Palese P, Singer R, Lionnet T. Colocalization of different
665 influenza viral RNA segments in the cytoplasm before viral budding as shown by single-
666 molecule sensitivity FISH analysis. PLoS Pathog. 2013;9(5):e1003358. Epub 2013/05/15.

667 doi: 10.1371/journal.ppat.1003358. PubMed PMID: 23671419; PubMed Central PMCID:
668 PMC3649991.
669 19. Lakdawala SS, Wu Y, Wawrzusin P, Kabat J, Broadbent AJ, Lamirande EW, et al.
670 Influenza a virus assembly intermediates fuse in the cytoplasm. PLoS Pathog.
671 2014;10(3):e1003971. doi: 10.1371/journal.ppat.1003971. PubMed PMID: 24603687;
672 PubMed Central PMCID: PMCPMC3946384.
673 20. Johnson GR, Li J, Shariff A, Rohde GK, Murphy RF. Automated Learning of Subcellular
674 Variation among Punctate Protein Patterns and a Generative Model of Their Relation to
675 Microtubules. PLoS Comput Biol. 2015;11(12):e1004614. doi:
676 10.1371/journal.pcbi.1004614. PubMed PMID: 26624011; PubMed Central PMCID:
677 PMCPMC4704559.
678 21. Li Y, Majarian TD, Naik AW, Johnson GR, Murphy RF. Point process models for
679 localization and interdependence of punctate cellular structures. Cytometry A.
680 2016;89(7):633-43. doi: 10.1002/cyto.a.22873. PubMed PMID: 27327612.
681 22. Griffie J, Shlomovich L, Williamson DJ, Shannon M, Aaron J, Khuon S, et al. 3D
682 Bayesian cluster analysis of super-resolution data reveals LAT recruitment to the T cell
683 synapse. Sci Rep. 2017;7(1):4077. doi: 10.1038/s41598-017-04450-w. PubMed PMID:
684 28642595; PubMed Central PMCID: PMCPMC5481387.
685 23. Yunta ML, Lagache T, Santi-Rocca J, Bastin P, Olivo-Marin JC, editors. A statistical
686 analysis of spatial clustering along cell filaments using Ripley's K function. 2014 IEEE 11th
687 International Symposium on Biomedical Imaging (ISBI); 2014 April 29 2014-May 2 2014.
688 24. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of
689 progressive multiple sequence alignment through sequence weighting, position-specific
690 gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22(22):4673-80.
691 25. Li C, Hatta M, Watanabe S, Neumann G, Kawaoka Y. Compatibility among polymerase
692 subunit proteins is a restricting factor in reassortment between equine H7N7 and human
693 H3N2 influenza viruses. J Virol. 2008;82(23):11880-8. doi: 10.1128/JVI.01445-08. PubMed
694 PMID: 18815312; PubMed Central PMCID: PMCPMC2583690.
695 26. Te Velthuis AJ, Fodor E. Influenza virus RNA polymerase: insights into the
696 mechanisms of viral RNA synthesis. Nat Rev Microbiol. 2016;14(8):479-93. doi:
697 10.1038/nrmicro.2016.87. PubMed PMID: 27396566; PubMed Central PMCID:
698 PMCPMC4966622.
699 27. Phipps KL, Marshall N, Tao H, Danzy S, Onuoha N, Steel J, et al. Seasonal H3N2 and
700 2009 Pandemic H1N1 Influenza A Viruses Reassort Efficiently but Produce Attenuated
701 Progeny. J Virol. 2017;91(17). doi: 10.1128/JVI.00830-17. PubMed PMID: 28637755;
702 PubMed Central PMCID: PMCPMC5553182.
703 28. Noda T, Sugita Y, Aoyama K, Hirase A, Kawakami E, Miyazawa A, et al. Three-
704 dimensional analysis of ribonucleoprotein complexes in influenza A virus. Nat Commun.
705 2012;3:639. Epub 2012/01/26. doi: 10.1038/ncomms1647. PubMed PMID: 22273677;
706 PubMed Central PMCID: PMC3272569.
707 29. Brooke CB. Population Diversity and Collective Interactions during Influenza Virus
708 Infection. J Virol. 2017;91(22). Epub 2017/09/01. doi: 10.1128/JVI.01164-17. PubMed
709 PMID: 28855247; PubMed Central PMCID: PMCPMC5660503.
710 30. Illian JB, Soerbye, S, Rue, H & Hendrichsen, D. Using INLA to fit a complex point
711 process model with temporally varying effects – a case study Journal of Environmental
712 Statistics. 2012;3(7).

713 31. Lakdawala SS, Fodor E, Subbarao K. Moving On Out: Transport and Packaging of
714 Influenza Viral RNA into Virions. Annu Rev Virol. 2016;3(1):411-27. doi: 10.1146/annurev-
715 virology-110615-042345. PubMed PMID: 27741407.
716 32. Bruce EA, Digard P, Stuart AD. The Rab11 pathway is required for influenza A virus
717 budding and filament formation. J Virol. 2010;84(12):5848-59. Epub 2010/04/02. doi:
718 10.1128/JVI.00307-10. PubMed PMID: 20357086; PubMed Central PMCID: PMC2876627.
719 33. Eisfeld AJ, Kawakami E, Watanabe T, Neumann G, Kawaoka Y. RAB11A is essential
720 for transport of the influenza virus genome to the plasma membrane. J Virol.
721 2011;85(13):6117-26. Epub 2011/04/29. doi: 10.1128/JVI.00378-11. PubMed PMID:
722 21525351; PubMed Central PMCID: PMC3126513.
723 34. Nturibi E, Bhagwat AR, Coburn S, Myerburg MM, Lakdawala SS. Intracellular
724 Colocalization of Influenza Viral RNA and Rab11A is Dependent upon Microtubule
725 Filaments. J Virol. 2017;91(19). doi: 10.1128/JVI.01179-17. PubMed PMID: 28724771.
726 35. de Castro Martin IF, Fournier G, Sachse M, Pizarro-Cerda J, Risco C, Naffakh N.
727 Influenza virus genome reaches the plasma membrane via a modified endoplasmic
728 reticulum and Rab11-dependent vesicles. Nat Commun. 2017;8(1):1396. doi:
729 10.1038/s41467-017-01557-6. PubMed PMID: 29123131; PubMed Central PMCID:
730 PMCPMC5680169.
731 36. Lin JR, Fallahi-Sichani M, Sorger PK. Highly multiplexed imaging of single cells using
732 a high-throughput cyclic immunofluorescence method. Nat Commun. 2015;6:8390. doi:
733 10.1038/ncomms9390. PubMed PMID: 26399630; PubMed Central PMCID:
734 PMCPMC4587398.
735 37. Comaniciu D, Meer P. Mean shift: A robust approach toward feature space analysis.
736 IEEE Transactions on pattern analysis and machine intelligence. 2002;24(5):603-19.
737 38. Baddeley A, Turner R. Practical maximum pseudolikelihood for spatial point
738 patterns. Australian & New Zealand Journal of Statistics. 2000;42(3):283-322.
739

740

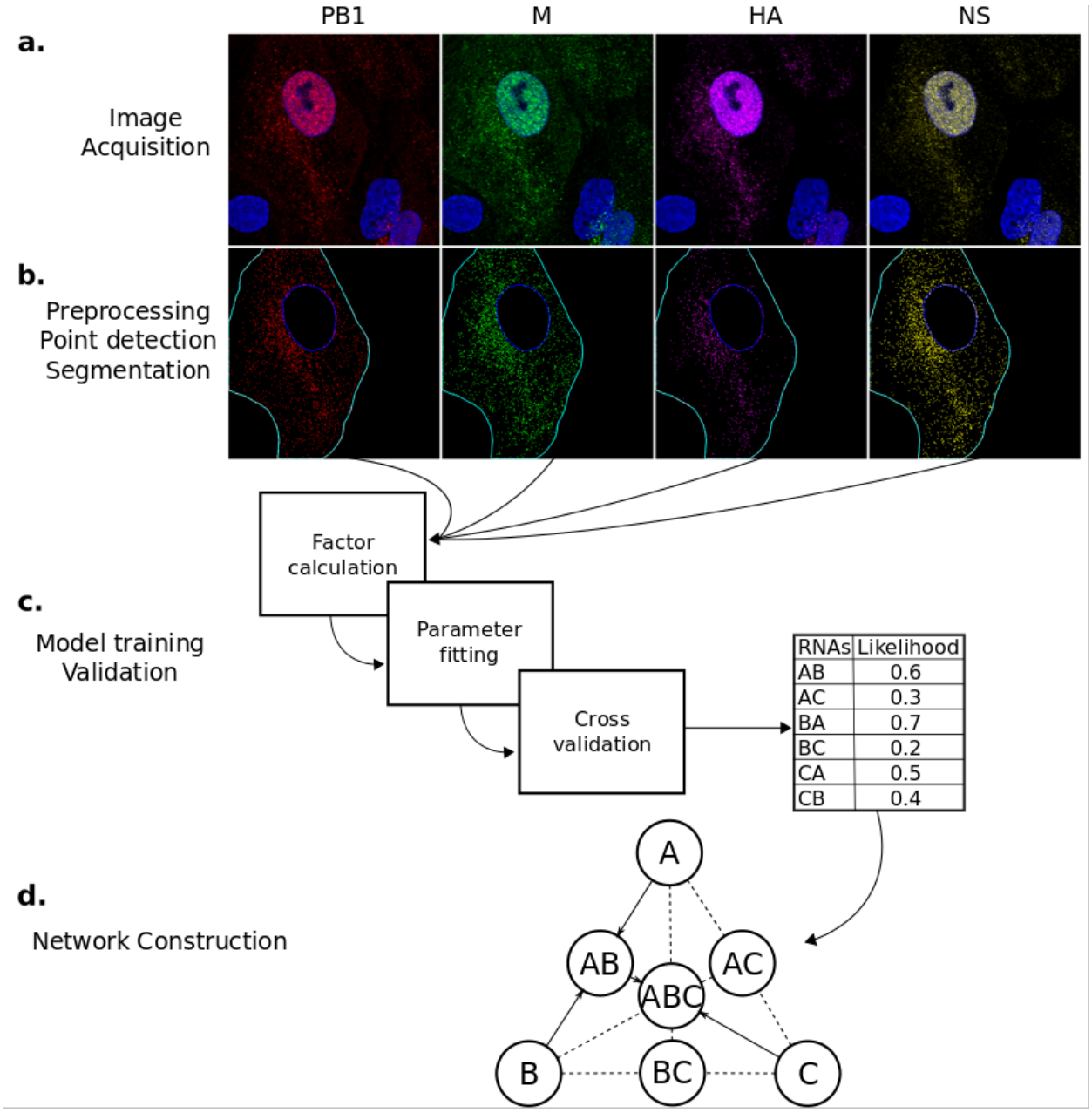


Figure 1: Illustration of the analysis pipeline. (a) Unprocessed fluorescence channels for cells FISH tagged for four vRNA segments and stained with DAPI. (b) Preprocessed image channels after filtering, nuclear and cell membrane segmentation, and point identification. The 3D images in (a) and (b) are 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.

752

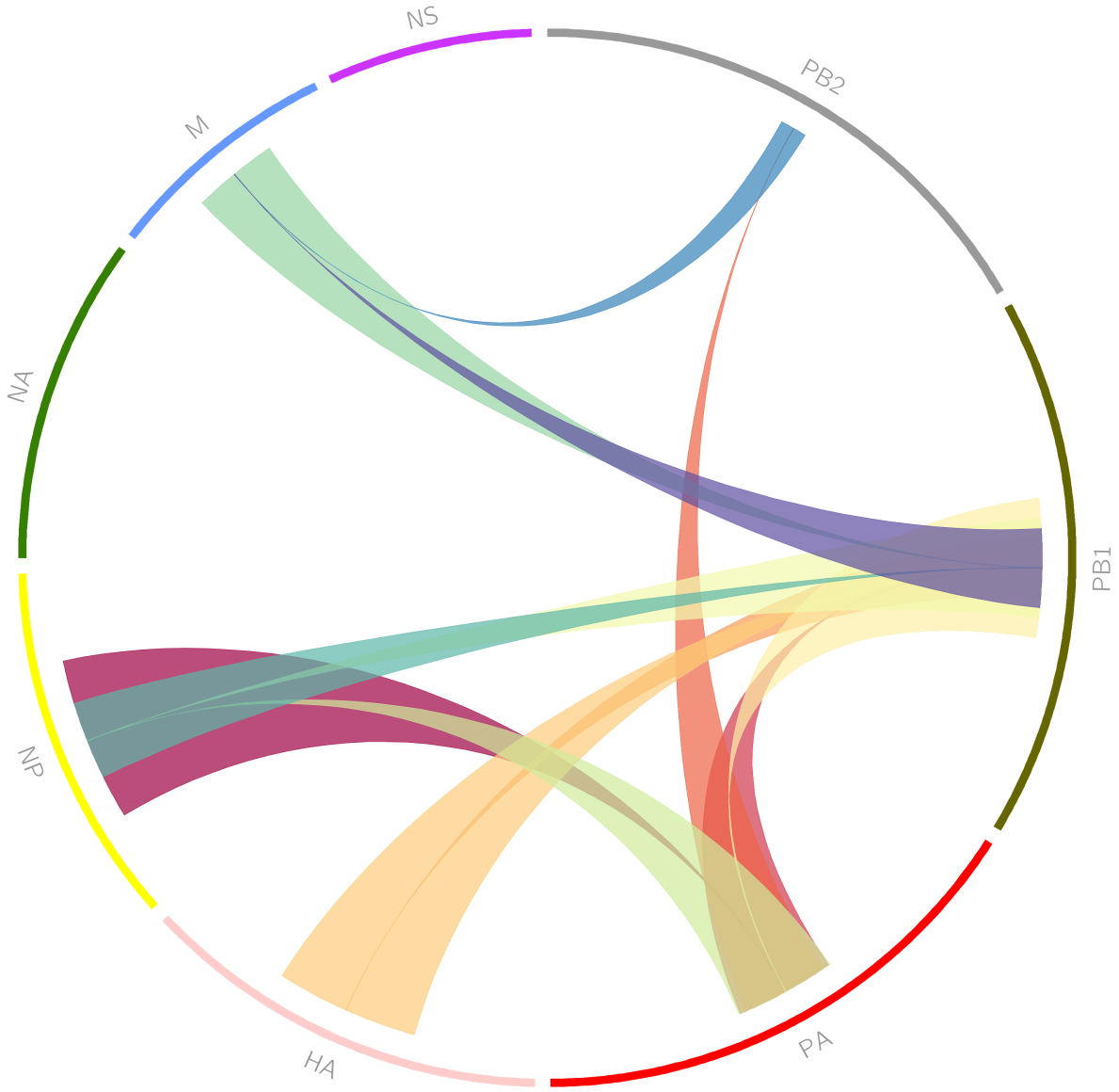
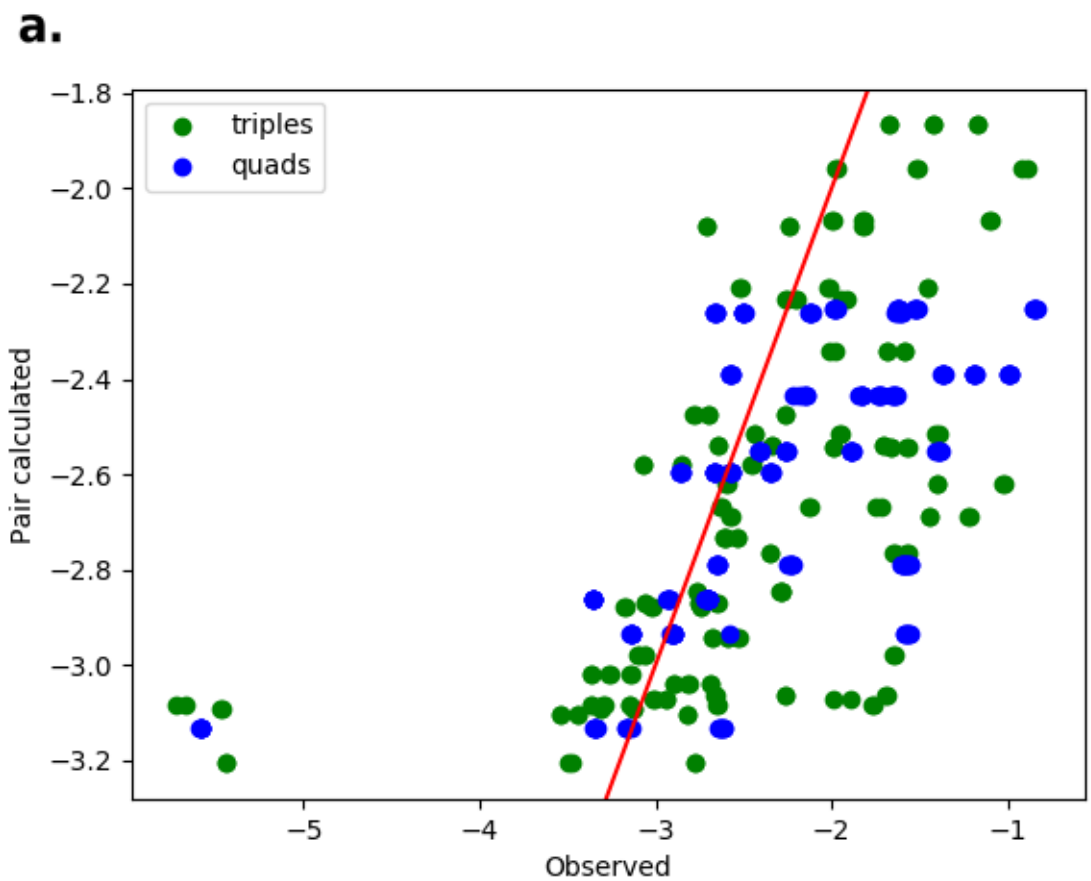
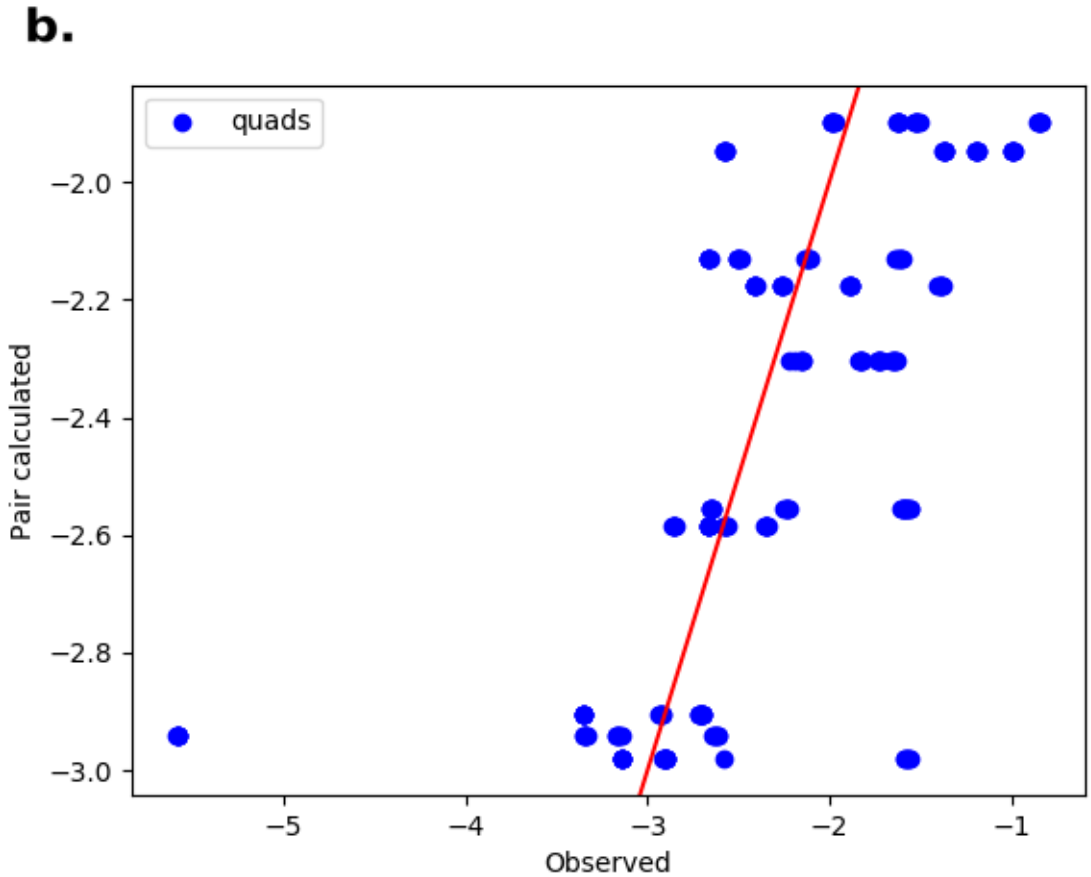


Figure 2: Pairwise interactions learned from MDCK cell images. Each vRNA segment is arranged around the perimeter. Bands connect vRNA segments that interact with high likelihood. The position of the base of each band indicates the independent (secondary) variable and the larger the width of the base the more likely the interaction. Only interactions whose log likelihoods are below 0.23 from Table 4 are shown.


a.
triples
quads
Pair calculated
Observed
−1.8
−2.0
−2.2
−2.4
−2.6
−2.8
−3.0
−3.2
−5
−4
−3
−2
−1



761

762 Figure 3: Comparison of composite and observed model likelihoods for triplet and quadruplet
763 models. To ensure that the recursion in our dynamic programming scheme accurately predicts
764 model likelihoods, we compared triplet and quadruplet observed likelihoods to those
765 generated from only pairwise models (that include all four possible models: vRNP pairs alone,
766 vRNP pairs + nucleus, vRNP pairs + cell membrane, vRNP pairs + nuclear/cell membrane ratio)
767 (a). We also compared observed quadruplet likelihoods to those generated from pairwise and
768 triplet models (using all 4 possible models) (b). Ideally, all points would fall on the red line,
769 matching observed and predicted values perfectly. The cluster in the lower left corner
770 represent interactions that are predicted to occur with high likelihood but in reality, do not. All
771 points within this cluster contain the same four vRNA segments (PB1, PB2, NA, NS). Exclusion of
772 this cluster from our analysis increases the association between predicted and observed triplets
773 and quadruplets.
774

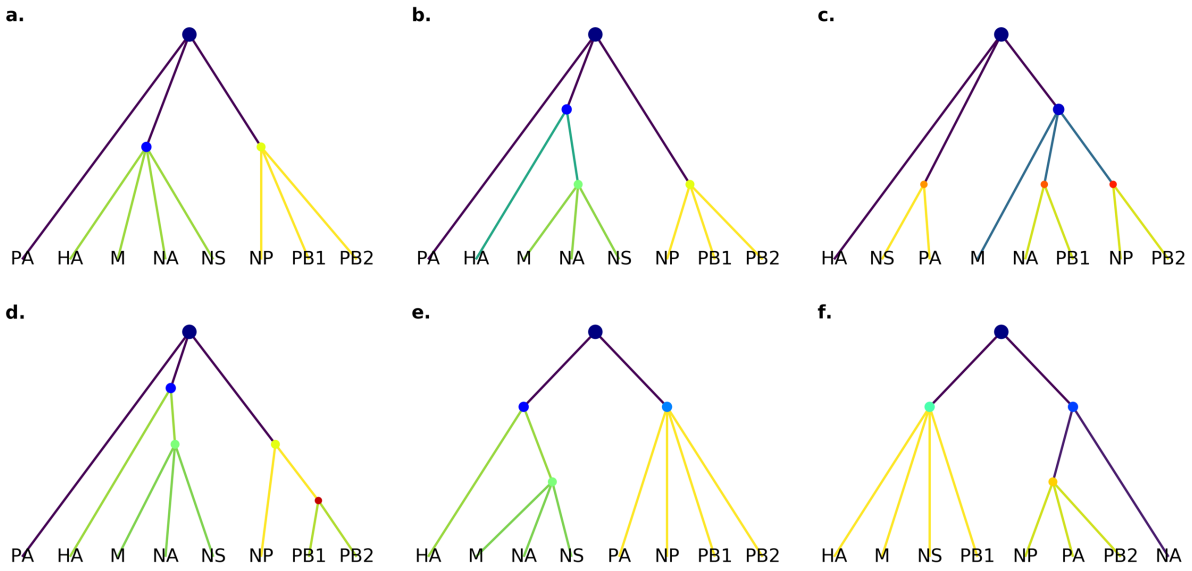


Figure 4: Constructed interaction networks. Using dynamic programming, we constructed the single most likely set of interactions yielding a full genome. First, we used all model likelihoods with and without nuclear and cell features (a) to build a master tree. Hotter edge colors correspond to more likely 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 1/2 of the mean model likelihood with the noise level being the standard deviation of a Gaussian centered at 0. For each level of noise, network construction was performed 100 times. Noise up to 1/4 mean likelihood always resulted in (a). Increasing noise levels yielded (a) on 40% to 60% of constructions. 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 (d). Adding noise up to 1/4 mean likelihood and removing both clusters also generated (d). Removal with higher levels of noise generated (e) for 27% of constructions. In all stability analyses with noise greater than 1/4 mean likelihood, (f) was generated with frequency between 5% and 15%.

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
PB2	0.0034	0.00011
PB1	0.00524	0.00034
PA	0.00467	0.0002
HA	0.00671	0.00015
NP	0.01185	0.00036
NA	0.00249	0.00003
M	0.00005	0
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
PB2	2.5547	0.4153	0.9158
PB1	2.3908	0.408	1.2574
PA	2.4228	0.4643	1.2352
HA	2.443	2.3928	4.396
NP	2.4885	2.4584	4.5447
NA	2.5019	2.4372	2.1754
M	2.4871	2.1294	3.0799
NS	2.3906	0.6666	1.4624

Table 3: Cross validated negative log-likelihoods for pairwise interaction models. The row label indicates the independent (secondary) variable, the vRNP upon which a model predicting the column label (the dependent or primary variable) was constructed.

	PB2	PB1	PA	HA	NP	NA	M	NS
PB2		0.3705	0.2694	0.282	0.3146	0.5605	0.2146	0.3926
PB1	0.2398		0.2018	0.2044	0.1186	0.4015	0.1701	0.2905
PA	0.2276	0.2217		0.2284	0.1502	0.3222	0.2816	0.2315
HA	0.3334	0.1266	0.1838		0.2787	0.232	0.2623	0.2845
NP	0.3438	0.1805	0.1942	0.2895		0.2529	0.275	0.3411
NA	0.3191	0.324	0.3093	0.2947	0.2742		0.2747	0.2982
M	0.3287	0.2236	0.2832	0.2803	0.2975	0.2897		0.2848
NS	0.3427	0.2888	0.2292	0.2838	0.325	0.3436	0.2841	

Table 4: Cross validated negative log-likelihoods for pairwise interaction models that also include nuclear distance. The row label indicates the independent (secondary) variable, the vRNP upon which a model predicting the column label (the dependent or primary variable) was constructed.

	PB2	PB1	PA	HA	NP	NA	M	NS
PB2		0.3817	0.2694	0.2833	0.3145	0.5643	0.2166	0.3977
PB1	0.2427		0.2023	0.2191	0.1215	0.4188	0.1714	0.3097
PA	0.2278	0.2211		0.661	0.1504	0.3238	0.4967	0.5378
HA	0.336	0.13	0.9138		1.944	1.7048	1.2324	0.5447
NP	0.3084	0.1746	0.1949	0.2095		1.6062	1.6495	0.3523
NA	0.3166	0.3317	0.3114	1.7801	1.4877		1.0989	0.3
M	0.33	0.2246	0.5374	1.2838	1.7105	1.1884		0.4153
NS	0.3454	0.2816	0.6259	0.4561	1.7998	0.3466	0.4368	

Table 5: Cross validated negative log-likelihoods for the top two triple 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	PB2	PB1	PA	HA	NP	NA	M	NS
PB2									
PB2 M PB1	0.114		0.382					0.217	
PB2 HA M	0.162				0.283			0.217	
PB1									
PB1 PA									
PB2	0.062	0.243		0.202					
PB1 NP									
PB2	0.071	0.243				0.122			
PA									
PA PB2									
PB1	0.113	0.228	0.222						
PA PB2 NP	0.121	0.228				0.15			
HA									
HA M PB1	0.059		0.13					1.232	
HA PB1 NS	0.111		0.13						0.545
NP									
NP PB1									
PB2	0.141	0.344	0.181						
NP PB2 PA	0.145	0.344		0.195					
NA									
NA HA NP	0.225				1.78	1.488			
NA HA NS	0.246				1.78				0.3
M									
M PB1 NS	0.201		0.225						0.415
M HA PB1	0.201		0.225		1.284				
NS									
NS M PB1	0.169		0.289					0.437	
NS HA PB1	0.184		0.289		0.456				

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	PB2	PB1	PA	HA	NP	NA	M	NS
PB2									
PB2 NS PB1 NA	0.571		0.382				0.564		0.398
PB2 NP PA PB1	0.258		0.382	0.269		0.315			
PB1									
PB1 NS NA PB2	0.347	0.243					0.419		0.31
PB1 PA M PB2	0.16	0.243		0.202				0.171	
PA									
PA NS M HA	0.655			0.438	0.661			0.497	0.538
PA NA M NS	0.272			0.438			0.324	0.497	0.538
HA									
HA PA M NS	0.869			0.914				1.232	0.545
HA NS PB2 M	0.317	0.336						1.232	0.545
NP									
NP HA M NA	0.233				2.095		1.606	1.65	
NP PB2 PA PB1	0.139	0.344	0.181	0.195					
NA									
NA PA NS M	0.272			0.311				1.099	0.3
NA PB1 PB2 NS	0.268	0.319	0.332						0.3
M									
M NS HA PA	0.7			0.537	1.284				0.415
M PA NS NA	0.336			0.537			1.188		0.415
NS									
NS HA PA M	0.705			0.626	0.456			0.437	
NS PB2 M HA	0.340	0.343			0.288			0.289	

833 Supp. Table 1: FISH imaging probe sets. The leftmost column is the experiment label with the
834 right 4 columns showing which vRNA segments were tagged. Each experiment also used DAPI
835 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

836
837
838

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 shown in Table 3.

(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. The row label indicates the independent variable, the vRNP upon which a model predicting the column label (dependent variable) was constructed.

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