

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41

# Transmission Bottleneck Size Estimation from Pathogen Deep-Sequencing Data, with an Application to Human Influenza A Virus

Bottleneck size inference from pathogen deep-sequencing data

Ashley Sobel Leonard<sup>1</sup>, Daniel Weissman<sup>2</sup>, Benjamin Greenbaum<sup>3</sup>, Elodie Ghedin<sup>4</sup>, Katia Koelle<sup>1,\*</sup>

<sup>1</sup> Department of Biology, Duke University, Durham, NC 27701

<sup>2</sup> Department of Physics, Emory University, Atlanta, GA 30322

<sup>3</sup> Tisch Cancer Institute, Departments of Medicine, Oncological Sciences, and Pathology, Icahn School of Medicine at Mount Sinai, New York, NY, 10029.

<sup>4</sup> Center for Genomics & Systems Biology, Department of Biology, and College of Global Public Health, New York University, 12 Waverly Place, New York, NY 10003

\* Corresponding author: [katia.koelle@duke.edu](mailto:katia.koelle@duke.edu)

42

43 **Abstract**

44

45         The bottleneck governing infectious disease transmission describes the size of the  
46 pathogen population transferred from a donor to a recipient host. Accurate quantification  
47 of the bottleneck size is of particular importance for rapidly evolving pathogens such as  
48 influenza virus, as narrow bottlenecks would limit the extent of transferred viral genetic  
49 diversity and, thus, have the potential to slow the rate of viral adaptation. Previous studies  
50 have estimated the transmission bottleneck size governing viral transmission through  
51 statistical analyses of variants identified in pathogen sequencing data. The methods used  
52 by these studies, however, did not account for variant calling thresholds and stochastic  
53 dynamics of the viral population within recipient hosts. Because these factors can skew  
54 bottleneck size estimates, we here introduce a new method for inferring transmission  
55 bottleneck sizes that explicitly takes these factors into account. We compare our method,  
56 based on beta-binomial sampling, with existing methods in the literature for their ability to  
57 recover the transmission bottleneck size of a simulated dataset. This comparison  
58 demonstrates that the beta-binomial sampling method is best able to accurately infer the  
59 simulated bottleneck size. We then apply our method to a recently published dataset of  
60 influenza A H1N1p and H3N2 infections, for which viral deep sequencing data from  
61 inferred donor-recipient transmission pairs are available. Our results indicate that  
62 transmission bottleneck sizes across transmission pairs are variable, yet that there is no  
63 significant difference in the overall bottleneck sizes inferred for H1N1p and H3N2. The  
64 mean bottleneck size for influenza virus in this study, considering all transmission pairs,  
65 was  $N_b = 196$  (95% confidence interval 66-392) virions. While this estimate is consistent

66 with previous bottleneck size estimates for this dataset, it is considerably higher than the  
67 bottleneck sizes estimated for influenza from other datasets.

68

## 69 **Author Summary**

70 The transmission bottleneck size describes the size of the pathogen population  
71 transferred from the donor to recipient host at the onset of infection and is a key factor in  
72 determining the rate at which a pathogen can adapt within a host population. Recent  
73 advances in sequencing technology have enabled the bottleneck size to be estimated from  
74 pathogen sequence data, though there is not yet a consensus on the statistical method to  
75 use. In this study, we introduce a new approach for inferring the transmission bottleneck  
76 size from sequencing data that accounts for the criteria used to identify sequence variants  
77 and stochasticity in pathogen replication dynamics. We show that the failure to account for  
78 these factors may lead to underestimation of the transmission bottleneck size. We apply  
79 this method to a previous dataset of human influenza A infections, showing that  
80 transmission is governed by a loose transmission bottleneck and that the bottleneck size is  
81 highly variable across transmission events. This work advances our understanding of the  
82 bottleneck size governing influenza infection and introduces a method for estimating the  
83 bottleneck size that can be applied to other rapidly evolving RNA viruses, such as norovirus  
84 and RSV.

85

## 86 **Introduction**

87 Infectious disease transmission relies on the transfer of a pathogenic organism from  
88 one host to another. This transfer is characterized by a transmission bottleneck, defined as  
89 the size of the founding pathogen population in the recipient host. Accurate quantification

90 of transmission bottleneck sizes for pathogenic organisms is critical for several reasons.  
91 First, bottleneck sizes impact levels of genetic diversity in recipient hosts, and thereby  
92 impact the rate at which pathogens can adapt to host populations, with smaller bottleneck  
93 sizes slowing rates of adaptation [1,2]. Second, when cooperative interactions occur within  
94 a pathogen population (e.g., [3,4]), or when viral complementation and cellular coinfection  
95 are critical for producing viral progeny (e.g., [5]), bottleneck sizes will necessarily impact  
96 initial pathogen replication rates, with larger bottleneck sizes enabling the occurrence of  
97 these interactions and thus facilitating within-host replication. Finally, transmission  
98 bottleneck sizes impact the ability to accurately reconstruct who-infected-whom during an  
99 ongoing epidemic [6], such that estimation of the transmission bottleneck size can point to  
100 cases which may be problematic, and for which a certain class of phylodynamic inference  
101 methods (such as [7]) might be particularly useful.

102         The transmission bottleneck size has been estimated for a number of pathogenic  
103 organisms, including pathogens of plants [8–13] and animals [14–22]. While these  
104 estimates have relied on the distribution of pathogen types in the infection recipients, as  
105 determined by molecular and phenotypic markers or Sanger sequencing of the pathogen  
106 population in donor and recipient hosts, deep sequencing data have recently started to be  
107 used to gauge transmission bottleneck sizes [23–29]. Some of these studies have  
108 characterized the general magnitude of transmission bottlenecks size, with results  
109 indicating that narrow, selective bottlenecks tend to govern the transmission dynamics of  
110 viral pathogens that are ill-adapted to their recipient hosts [24–26]. Studies that have  
111 instead gauged transmission bottleneck sizes of well-adapted viral pathogens using deep  
112 sequencing data have, in contrast, generally found that they tend to be loose, with many

113 virions initiating infection [23,28,29]. While many of these studies focus on assessing how  
114 “loose” or “narrow” a transmission bottleneck is, other studies have attempted to more  
115 quantitatively estimate transmission bottleneck sizes. One approach relied on the use of  
116 barcoded influenza virus during experimental transmission studies in small mammals, with  
117 results indicating that the route of transmission greatly impacts the size of the bottleneck  
118 [27].

119 In natural infections, it is not feasible to rely on barcoded or otherwise marked  
120 pathogens. In these cases, statistical approaches have therefore instead been used to  
121 quantify bottleneck sizes [28,30]. Two studies have used the Kullback-Leibler divergence  
122 index (developed in [30]) to estimate the viral effective population size initiating infection  
123 from deep sequencing data [28,30]. One of these studies quantified the transmission  
124 effective population size for ebola in human-to-human infections [30]. The other quantified  
125 this transmission effective population size for human influenza A viruses [28]. A second  
126 statistical approach used by [28] makes use of a single-generation population genetic  
127 Wright-Fisher model to estimate the effective viral population size initiating infection.  
128 While this approach similarly showed that the effective population size following influenza  
129 virus transmission in natural human-to-human infection is large, this model yielded  
130 quantitatively different results from the Kullback-Leibler approach. Further, in both of  
131 these studies, it is not clear how the effective population size relates to the transmission  
132 bottleneck size. It is worth noting, however, that the effective population size is generally  
133 considered to be an underestimate of the true population size as it represents the minimum  
134 population size necessary to establish observed levels of genetic diversity.

135 Both of these approaches [28,30] analyze only variants that are identified as present  
136 in both the donor and the recipient. But the absence of a donor variant in a recipient host is  
137 also informative, and ignoring such missing variants can significantly bias transmission  
138 bottleneck size estimates. Another limitation of both approaches is that they do not consider  
139 the effect stochastic dynamics early in infection may have on variant frequencies in the  
140 recipient. To address these concerns, we here introduce a new method for estimating the  
141 transmission bottleneck size of pathogens. This method accounts for stochastic dynamics  
142 occurring during viral replication in the recipient and further accounts for variant calling  
143 thresholds that are used in calling a variant present or absent in a sample. We refer to this  
144 method as the beta-binomial sampling method, based upon the method's derived likelihood  
145 expression. Using a simulated dataset, we compare the beta-binomial sampling method to  
146 two methods of bottleneck size inference that are present (in some form) in the current  
147 literature: the presence/absence method and the binomial sampling method. This  
148 comparison demonstrates that the beta-binomial sampling method is able to recover the  
149 true bottleneck size of the simulated dataset, whereas the 2 other methods infer biased  
150 estimates by failing to account for variant calling thresholds or stochastic dynamics in the  
151 recipient host. Finally, we apply the beta-binomial sampling method to an existing next  
152 generation sequencing dataset of influenza A virus infections to estimate the transmission  
153 bottleneck size in natural human-to-human flu transmission.

## 154 **Models**

155 Figure 1 provides a schematic of the data that are used for inferring transmission  
156 bottleneck sizes in the approaches we consider in this study. Deep sequencing data consist  
157 of short reads at various sites in the genome, obtained from both the infected donor and the

158 recipient at, generally, a single time point for each individual. The short read data are used  
159 to identify viral variants in the donor and recipient hosts. Comparison of these variants'  
160 frequencies across donor-recipient transmission pairs allows us to infer the transmission  
161 bottleneck size ( $N_b$ ), the number of virions comprising the founding viral population at the  
162 onset of infection in the recipient host. We specifically define  $N_b$  as the number virions that  
163 successfully establish lineages that persist to sampling; there may be additional virions that  
164 transiently replicate in the recipient host but quickly die out.

165         Given the extent of sequencing error in deep-sequencing data, there can be a high  
166 degree of noise in the short read data and, thereby, in the extent of polymorphism present  
167 at nucleotide sites. To limit spurious identification of variants arising from sequencing  
168 noise, it is common practice to use criteria, such as a variant calling threshold, to validate  
169 identified variants [31]. The variant calling threshold is the minimum frequency at which  
170 a variant can almost certainly be distinguished from background sequencing error. This  
171 threshold frequency may be chosen according to generally accepted error rates for a  
172 specific sequencing platform, error rates informed by a control run, or error rates based on  
173 the concordance of variant calls from replicate sequence runs. For the commonly used  
174 Illumina sequencing platforms, variant thresholds tend to fall in the range of 0.5-3% [24–  
175 26,28,32–35]. Conservative variant calling cutoffs are often used, as they ensure that  
176 sequencing artifacts are excluded. However, conservative frequency cutoffs may have  
177 effects on transmission bottleneck size analyses due to variants that are not called in the  
178 recipient host, despite being present. Such 'false negatives' in the recipient have the  
179 potential to skew inferred transmission bottleneck size towards inappropriately low values.

180 We present methods for inferring the transmission bottleneck size from deep  
181 sequencing data, paying special attention to the effects of ‘false negative’ variant calls. We  
182 first introduce the beta-binomial sampling method that we have developed for bottleneck  
183 size inference, which further incorporates the effects of stochastic pathogen dynamics in  
184 recipient hosts. For comparison, we then summarize two existing methods of bottleneck  
185 size inference in the literature: the presence/absence method and the binomial sampling  
186 method. Of note, all three of these methods assume that the genetic diversity of the  
187 pathogen is entirely neutral, such that selection does not impact variant frequency  
188 dynamics. These methods further assume independence between variant sites. We address  
189 the limitations of these assumptions in the Discussion.

190 **Bottleneck size inference allowing for stochastic pathogen dynamics in the recipient**  
191 **host.**

192 The beta-binomial sampling method for inferring the bottleneck size allows variant  
193 allele frequencies in the recipient host to change between the time of founding and the time  
194 of sampling (see Figure 1) as the result of stochastic pathogen replication dynamics early  
195 in infection. We consider two implementations of the beta-binomial sampling method: an  
196 approximate version that assumes infinite read depth and an exact version that incorporates  
197 sampling noise arising from a finite number of reads. The derivation of the beta-binomial  
198 sampling method can be found in the Methods.

199 In the approximate version, the likelihood of a transmission bottleneck size  $N_b$ ,  
200 given variant frequency data at site  $i$ , is given by:

$$L(N_b)_i = \sum_{k=0}^{N_b} p\_beta(v_{R,i}|k, N_b - k) p\_bin(k|N_b, v_{D,i}), \quad (1)$$



201 where  $v_{R,i}$  is the variant frequency at site  $i$  in the recipient and  $p\_beta(v_{R,i}|k, N_b - k)$  is  
202 given by the beta probability density function parameterized with shape parameters  $k$  and  
203  $N_b - k$ , and evaluated at  $v_{R,i}$ . The term  $p\_bin(k|N_b, v_{D,i})$  denotes the binomial distribution  
204 evaluated at  $k$  and parameterized with  $N_b$  number of trials and a success probability of  $v_{D,i}$ ,  
205 where  $v_{D,i}$  is the variant frequency at site  $i$  in the donor. If the donor variant at site  $i$  is not  
206 detected in the recipient, this may be because it is truly absent from the recipient or because  
207 it falls below the variant calling threshold. To allow for both of these possibilities, the  
208 likelihood that the transmission bottleneck size is  $N_b$ , given that the variant at site  $i$  was not  
209 detected, is given by:

$$L(N_b)_i = \sum_{k=0}^{N_b} [p\_beta\_cdf(v_{R,i} < T | k, N_b - k) p\_bin(k|N_b, v_{D,i})], \quad (2)$$

210 where  $T$  is the variant calling threshold and  $p\_beta\_cdf(v_{R,i} < T | k, N_b - k)$  is given by the  
211 beta cumulative distribution function evaluated at the variant calling threshold.

212 In the exact version of the beta-binomial sampling method, we incorporate  
213 sampling error by modifying equations (1) and (2) to consider the number of variant reads  
214 and the number of total reads at variant site  $i$ ,  $R_{var,i}$  and  $R_{tot,i}$ , respectively. The likelihood  
215 expression for the bottleneck size at site  $i$  becomes:

$$L(N_b)_i = \sum_{k=0}^{N_b} p\_betabin(R_{var,i}|R_{tot,i}, k, N_b) p\_bin(k|N_b, v_{D,i}), \quad (3)$$

216 where  $p\_betabin(R_{var,i}|R_{tot,i}, k, N_b)$  is given by the beta-binomial probability density  
217 function evaluated at  $R_{var,i}$  and parameterized with  $R_{tot,i}$  number of trials and parameters  
218  $k$  and  $N_b$ . If the donor-identified variant at site  $i$  is not detected in the recipient, we again  
219 construct the likelihood that allows for this variant to either be absent from the recipient or  
220 below the variant calling threshold:

$$L(N_b)_i = \sum_{k=0}^{N_b} p_{betabin\_cdf}(R_{var,i} < TR_{tot,i} | k, N_b) p_{bin}(k | N_b, v_{D,i}), \quad (4)$$

221 where, in this case,  $p_{betabin\_cdf}(R_{var,i} < TR_{tot,i} | k, N_b)$  is given by the beta-binomial  
222 cumulative distribution function evaluated at the number of reads that would qualify as  
223 falling at the variant calling threshold.

224 We expect that the maximum likelihood estimate (MLE) of  $N_b$  inferred with the  
225 approximate method will converge to the MLE of  $N_b$  inferred with the exact method when  
226 read coverage is high. The benefit of using the approximate version, when appropriate, is  
227 that the incorporation of sampling error is computationally intensive.

228 Once transmission bottleneck sizes have been estimated using either the  
229 approximate or exact beta-binomial sampling method, the probability that a variant  
230 is truly present/absent in the recipient and the probability that a variant is simply  
231 called present/absent in the recipient (under the assumption of infinite coverage) can  
232 be determined for any given donor variant frequency.

### 233 **Existing methods for inferring transmission bottleneck sizes**

234 **Presence/absence method of bottleneck size inference.** The simplest approach to  
235 estimating transmission bottleneck sizes from pathogen deep-sequencing data is to  
236 calculate variant frequencies in donor hosts and then use information on the  
237 presence/absence of these variants in recipient hosts to quantify bottleneck size. Studies  
238 that have adopted this approach include [9,36]. Given a variant  $i$  present at frequency  $v_{D,i}$   
239 in the donor, and a founding population size of  $N_b$ , the probability that the variant was not  
240 transferred to the recipient is simply given by  $(1 - v_{D,i})^{N_b}$  [9,36]. Correspondingly, the  
241 probability that at least one virion in the founding population carried the variant allele is

242 given by  $1 - (1 - v_{D,i})^{N_b}$ . From these expressions, the likelihood of the founding  
243 population size  $N_b$  in a donor-recipient pair is simply calculated by multiplying the  
244 probabilities of the observed outcomes across the variant sites:

$$L(N_b) = \prod_{j=1}^{V_{absent}} (1 - v_{D,j})^{N_b} \prod_{k=1}^{V_{present}} 1 - (1 - v_{D,k})^{N_b}, \quad (5)$$

245 where  $j$  indexes the viral variants that are absent in the recipient,  $k$  indexes the viral variants  
246 that are present in the recipient,  $V_{absent}$  is the total number of variants which are called  
247 absent in the recipient, and  $V_{present}$  is the total number of variants that are called present in  
248 the recipient. The total number of variants identified in the donor is given by  $V_{absent} +$   
249  $V_{present}$ .

250 The presence/absence method considers only the detection of donor-identified  
251 variants in the recipient host and, therefore, is especially prone to the effects of false  
252 negative variants. Moreover, accounting for the variant calling threshold to ameliorate  
253 these effects is not possible with this method. Due to the inability of this method to account  
254 for false negatives, we expect that the transmission bottleneck estimates inferred with the  
255 presence/absence method will be considerably lower than the bottleneck size estimates  
256 inferred by the beta-binomial sampling method.

257 **Binomial sampling method of bottleneck size inference.** The second approach, or class  
258 of approaches, from the literature for inferring transmission bottleneck sizes is based on a  
259 binomial sampling process. Studies that have adopted this general kind of approach include  
260 [28,30]. We describe a version of this approach that parallels the beta-binomial sampling  
261 method we described above. The binomial sampling approach makes use of donor-  
262 identified variant frequencies in the donor and both the number of variant reads and the

263 number of total reads in the recipient, at each donor-identified variant site. The likelihood  
264 expression for the bottleneck size, given these data at site  $i$ , is given by:

$$L(N_b)_i = \sum_{k=0}^{N_b} p\_bin\left(R_{var,i} \middle| R_{tot,i}, \frac{k}{N_b}\right) p\_bin(k|N_b, v_{D,i}), \quad (6)$$

265 where  $p\_bin\left(R_{var,i} \middle| R_{tot,i}, \frac{k}{N_b}\right)$  is given by the binomial probability density function  
266 evaluated at  $R_{var,i}$ . The term  $p\_bin(k|N_b, v_{D,i})$  is again given by the binomial distribution.  
267 For variants called as absent in the recipient host, the likelihood of the transmission  
268 bottleneck size is given:

$$L(N_b)_i = \sum_{k=0}^{N_b} p\_bin\_cdf\left(R_{var,i} < TR_{tot,i} \middle| \frac{k}{N_b}\right) p\_bin(k|N_b, v_{D,i}), \quad (7)$$

269 where  $p\_bin\_cdf$  is the binomial cumulative distribution function. Derivation of the  
270 binomial sampling method can be found in the Methods section.

271 The sole difference between the beta-binomial sampling method and the binomial  
272 sampling method is that the binomial sampling method does not account for stochastic  
273 dynamics of the pathogen early on in the recipient. These stochastic dynamics enable the  
274 frequencies of variants in a recipient at the time of sampling to differ from those at the time  
275 of founding (Figure 1). Because the binomial sampling method does not incorporate this  
276 source of frequency variation, we expect there to be smaller frequency deviations between  
277 variants in donor-recipient pairs under the assumption of a single-generation binomial  
278 sampling model as compared to a model that allows for these stochastic dynamics, for a  
279 given bottleneck size. To explain a given pattern of donor-recipient frequency pairs,  $N_b$   
280 estimates are thus expected to be significantly lower for the binomial sampling method  
281 than for the beta-binomial sampling method. Application of the binomial sampling method

282 will therefore yield a conservative (lower-bound) estimate of  $N_b$ , as previously remarked  
283 upon [30].

## 284 **Results**

### 285 **Results on simulated data.**

286 To examine the ability of the three methods described above to accurately infer  
287 transmission bottleneck sizes, we used a simulated dataset of one donor-recipient pair  
288 (Methods). The dataset was generated under the assumption of stochastic pathogen  
289 dynamics in the recipient host between the time of infection and the time of sampling.  
290 While this assumption matches the assumption for the beta-binomial sampling method, we  
291 feel that it is also biologically the most realistic assumption. In this dataset, 109 out of the  
292 500 donor-identified simulated variants were called absent in the recipient host (Figure  
293 2A). The majority of these variants were present in the recipient host, but below our variant  
294 calling threshold of 3% and, therefore, were ‘false negatives’. The beta-binomial sampling  
295 method, as expected, recovers the true bottleneck size of 50 virions (Figure 2B). In contrast,  
296 both the presence/absence method (Figure 2C) and the binomial sampling method (Figure  
297 2D) significantly underestimate the simulated bottleneck size. The underlying reasons for  
298 these methods’ inability to recover the true bottleneck size differ. For the presence/absence  
299 method, this underestimation can be attributed to ‘false negative’ variant calls. For the  
300 binomial sampling method, we were able to statistically account for the variant calling  
301 threshold effects; the underestimation of this method, therefore, is solely attributed to this  
302 method not accounting for stochastic pathogen dynamics in the recipient. The binomial  
303 sampling method instead assumes deterministic viral growth from the time of founding to  
304 the time of sampling (see Methods). Because more sampling stochasticity is present at

305 smaller bottleneck sizes, the binomial sampling method underestimates the simulated  
306 bottleneck size in its attempt to reproduce observed variation in variant frequencies by  
307 inappropriately constricting  $N_b$ .

308         Given that the binomial sampling model and the beta-binomial model fit to the same  
309 data, the relative performance of these models can be assessed using model selection  
310 approaches. The maximum likelihood obtained using the beta-binomial sampling method  
311 was significantly higher than the maximum likelihood obtained using the binomial  
312 sampling method (Figure 2B,D; legend), indicating that the beta-binomial sampling model  
313 is statistically preferred over the binomial sampling model. We can further take into  
314 consideration the smoothness of the likelihood curves in our choice of model, with multi-  
315 modal/rugged likelihood curves being undesirable outcomes. In Figure 2E, we plot the  
316 likelihood curves for one variant under the likelihood expression of the beta-binomial  
317 sampling method and under the expression of the binomial sampling method. The rugged  
318 likelihood surface of the binomial sampling model arises because of this method's stringent  
319 assumption that variant frequencies remain fixed between the time of infection of the  
320 recipient and the time of sampling. In contrast, the beta-binomial sampling method allows  
321 for stochastic changes in variant frequencies during viral growth, relaxing the assumption  
322 that the viral population at the time of sampling needs to perfectly reflect the founding viral  
323 population. As a result, likelihood curves of the beta-binomial sampling model do not show  
324 large differences in likelihood values for small differences in  $N_b$ , further indicating that the  
325 beta-binomial sampling model is preferable.

326         Given an estimate of the transmission bottleneck size, the probability that a variant  
327 is transferred to a recipient host can be calculated using the expression  $1 - (1 - v_{D,i})^{N_b}$ ,

328 where  $v_{D,i}$  is the frequency of variant  $i$  present in the donor host and  $N_b$  is the bottleneck  
329 size estimate. In Figure 3A, we plot this probability of variant transfer over a range of donor  
330 variant frequencies for the simulated dataset. In this figure, we further plot ‘observed’  
331 probabilities of variant transfer, given a variant calling threshold of 3% on the simulated  
332 dataset. Finally, we plot in this figure the ‘observed’ probabilities of variant transfer as  
333 predicted under the beta-binomial sampling method, evaluated at the transmission  
334 bottleneck size estimated. We see, first, that the true probabilities of variant transfer greatly  
335 exceed those that are observed in the dataset given the variant calling threshold of 3%.  
336 However, the method’s calculated predictions of observed variant transfer probabilities fall  
337 within the 95% confidence intervals for the probabilities of variant transfer observed in the  
338 dataset.

339 As described in the Models section, the exact beta-binomial sampling method we  
340 developed accounts for sampling noise arising from finite read coverage. If we ignore  
341 sampling noise, we can estimate bottleneck sizes more rapidly using the approximate  
342 method, described by equations (1) and (2). In Figure 3B we show bottleneck size estimates  
343 over a range of different coverage levels for both the exact and approximate beta-binomial  
344 sampling methods. At high coverage levels (>200 reads), both implementations of the beta-  
345 binomial sampling method yield similar bottleneck size estimates and are able to recover  
346 the simulated bottleneck size of 50 virions. For lower levels of coverage, however, this  
347 approximation starts to fail and will lead to a considerable underestimation of  $N_b$ , indicating  
348 that the approximate beta-binomial sampling method is inappropriate for low coverage  
349 levels. We also note that even at high coverage, a slight overestimation of the bottleneck  
350 size is apparent. The overestimation can be attributed to the rare false positive identification

351 of variants in the recipient (instances of a variant that is absent in the recipient being called  
352 present) and, more generally, a slight inflation of variant frequencies with sequencing error.  
353 Overestimation no longer occurs when these methods are applied to datasets that are  
354 simulated in the absence of sequence error (results not shown).

### 355 **Transmission Bottleneck Size Estimation for Human Influenza A Virus**

356 We first applied the beta-binomial sampling method for inferring transmission  
357 bottleneck sizes to the influenza A/H1N1p transmission pairs identified in a previously  
358 studied influenza NGS dataset described in detail in [28]. We point the reader to this  
359 previous publication for details on the dataset, including coverage levels, how transmission  
360 pairs were inferred, etc. Poon and Song et al. [28] estimated the mean effective population  
361 size for all H1N1p transmission pairs to be  $N_e = 192$  virions (mean s.d. range 114-276).  
362 The approach considered the combined set of variants that were present at frequencies  $\geq 1\%$   
363 and that were shared by 8 identified household donor-recipient pairs (a total of 26 variants).  
364 In contrast to their analysis, we estimated transmission bottleneck sizes for each of 9  
365 transmission pairs separately, using a minimum variant frequency cutoff of 3% to call  
366 variants. We used a 3% cutoff based on concordance results from replicate sequencing  
367 runs, described in [28]. The less conservative 1% cutoff used by Poon and Song et al. [28]  
368 to estimate effective population size was chosen to allow for more sites to be included in  
369 their analysis. Our analysis, using a total of 289 variants, estimated MLE bottleneck sizes  
370 ranging from 49 to 276 virions across the H1N1p transmission pairs (Figure 4A). The  
371 bottleneck sizes inferred by the approximate beta-binomial sampling method did not differ  
372 significantly from those inferred by the exact method for any of the transmission pairs.  
373 This was expected, given high coverage levels across variant sites.



374 To summarize our results for the bottleneck size estimates for the H1N1p  
375 transmission pairs, we estimated parameters of a negative binomial distribution using all  
376 of the variant frequencies across the transmission pairs (see Methods). This negative  
377 binomial distribution was chosen because our results shown in Figure 4A indicated that the  
378 variance in transmission bottleneck sizes is likely to exceed the mean. We further fit a  
379 Poisson distribution to these same data, and the negative binomial distribution was  
380 statistically preferred over the Poisson distribution using AIC, indicating that, while a  
381 single infection may be initiated by a Poisson-distributed number of virions, different  
382 infections are likely to be initiated by founding population sizes that vary in their mean.  
383 The MLE of the negative binomial distribution's parameters was  $r = 5$  and  $p = 0.966$ ,  
384 resulting in a mean H1N1p transmission bottleneck size of  $N_b = 142$ , and a 95% range of  
385 54-262 virions (Figure 4A). While our overall bottleneck size estimates were consistent  
386 with the estimates from Poon and Song et al. using a much more limited number of variants,  
387 our analysis further shows that the transmission bottleneck sizes varied considerably  
388 between transmission pairs.

389 We next used the beta-binomial sampling method to infer the transmission  
390 bottleneck sizes for each of the H3N2 transmission pairs of the influenza NGS dataset.  
391 Poon and Song et al. estimated the mean effective population size for H3N2 to be  $N_e = 248$   
392 (mean s.d. range 45-457), again using a combined set of variants that were present at  
393 frequencies  $\geq 1\%$  and that were shared by 6 identified household donor-recipient pairs (a  
394 total of 81 variants). Our analysis, considering each of the 7 identified H3N2 transmission  
395 pairs separately, inferred MLE bottleneck sizes ranging from 107 to 370 virions across the  
396 transmission pairs using a total of 621 variants (Figure 4B). Again, as expected, the  $N_b$

397 sizes inferred by the approximate beta-binomial sampling method did not differ  
398 significantly from those inferred using the exact beta-binomial sampling method. We again  
399 fit a negative binomial distribution to all of the variants across the transmission pairs and  
400 estimated MLE parameters of  $r = 9$  and  $p = 0.966$ , resulting in a mean H3N2 transmission  
401 bottleneck size of  $N_b = 256$ , and a 95% range of 131-413 virions (Figure 4B). We again  
402 observed that the overall bottleneck size estimate for H3N2 was consistent with Poon et  
403 al.'s estimate, though the bottleneck size estimates varied considerably between  
404 transmission pairs.

405 **Overall influenza A transmission bottleneck sizes.** We next sought to determine whether  
406 influenza A/H1N1p and influenza A/H3N2 virus subtypes statistically differed from one  
407 another in bottleneck sizes. We found that the H1N1p and H3N2 distributions of  
408 transmission bottleneck size MLEs did not differ significantly from one another ( $p = 0.15$   
409 using the Kolmogorov-Smirnov test). Given this finding, we fit a negative binomial  
410 distribution to all of the variants across both subtype datasets, arriving at a MLE of  $r = 4$   
411 and  $p = 0.980$  for the parameters of the negative binomial distribution. These parameters  
412 correspond to a mean bottleneck size of  $N_b = 196$  and a 95% range of 66-392 virions  
413 (Figures 4A, 4B). We show the probability density function for this negative binomial  
414 distribution in Figure 5A. We further plot the expected probability of variant transfer for  
415 this bottleneck size estimate (Figure 5B), similar to what we show for the simulated dataset  
416 in Figure 3A. Finally, we plot the probability of observed variant transfer under this  $N_b$   
417 estimate, under the assumptions of the beta-binomial sampling model. The agreement  
418 between the probability of observed variant transfer and the empirical data indicate that

419 variant calling thresholds again make it appear that variant transfer from donor to recipient  
420 is much less likely than it is, given bottleneck size estimates based on variant frequencies.

## 421 **Discussion**

422 Here, we have introduced a new method for estimating the transmission bottleneck  
423 size of pathogens from next generation sequencing data from donor-recipient pairs. We  
424 have further analyzed how well this beta-binomial sampling method performs in  
425 comparison to two existing methods in the literature: the presence/absence method and the  
426 binomial sampling method. Using a simulated dataset, we have demonstrated that both the  
427 presence/absence method and the binomial sampling method (for different reasons)  
428 systematically underestimate the transmission bottleneck size and that the latter can lead  
429 to undesirable rugged likelihood curves. In contrast, the beta-binomial sampling method,  
430 as expected, is able to recover the simulated bottleneck size (Figure 2B) and is able to  
431 accurately predict the probability that a donor variant would be identified in a recipient  
432 host under a given variant calling threshold (Figure 3A). Application of the beta-binomial  
433 sampling method to a previously published H1N1p and H3N2 NGS dataset showed a high  
434 degree of heterogeneity between bottleneck size estimates across transmission pairs  
435 (Figure 4). A negative binomial distribution was fit to all of the variants, yielding an overall  
436 mean  $N_b$  of 196 virions and a 95% range of 66 – 382 virions (Figures 4A, 4B, 5A).

437 The bottleneck sizes that we estimated for the H1N1p and H3N2 transmission pairs  
438 are close to the previous estimates of the effective population size  $N_e$  arrived at by Poon  
439 and Song et al. for this dataset [28], although we were able to further estimate transmission  
440 bottleneck sizes by transmission pair and our method was able to make use of a much larger  
441 number of identified variants. Our bottleneck size estimates are consistent with the more

442 qualitative observations of loose transmission bottlenecks for influenza A virus  
443 transmission in horses [20,22,23], pigs [20,21], and dogs [19]. Our  $N_b$  estimates, however,  
444 are considerably larger than the previous bottleneck sizes estimated for this virus by Varble  
445 et al. [27], Frise et al. [29], and McCaw et al. [17]. Varble et al.'s experimental study  
446 showed that the route of transmission affected the bottleneck size, with contact  
447 transmission giving rise to larger bottlenecks. They found that, of the 71-100 distinct viral  
448 tags, only 7-24 of those tagged viruses were detected in the recipients following infection  
449 via direct contact [27]. The number of distinct viral tags, however, might reflect the lower  
450 limit of the bottleneck size because it is possible that more than one virion passing through  
451 the bottleneck would have the same tag. Frise et al. reported a mean bottleneck size of 28.2  
452 infectious genomes for contact transmission of an efficiently transmitted H1N1 strain in  
453 ferrets, though they were unable to identify an upper limit to the bottleneck size confidence  
454 interval [29]. Both of these previous estimates are much larger than the earlier estimate of  
455 3.8 virions estimated by McCaw et al. for contact transmission of H1N1 in ferrets [17].  
456 While other studies exist that have estimated the transmission bottleneck size in the context  
457 of viral adaptation to a new host species [24–26], comparisons with these studies  
458 are inappropriate because these bottlenecks are subject to strong selective forces, which  
459 considerably narrow the transmission bottleneck size [37].

460         The  $N_b$  estimates for influenza virus transmission in the dataset described here, both  
461 by our study and Poon and Song et al.'s original analysis, are considerably higher than  
462 previous quantitative estimates of IAV's bottleneck size for contact transmission  
463 [17,27,29]. Notably, these previous estimates of  $N_b$  were arrived at using data from  
464 experimental ferret infections. With a recent analysis showing that secondary attack rates

465 in ferret studies are considerably higher than human secondary attack rates, controlling for  
466 infecting subtype [38], one possibility for these discrepancy is that ferrets and other small  
467 mammals may require fewer influenza A virions to successfully initiate infection.

468 McCaw et al.'s bottleneck size estimate, in particular, was significantly lower than  
469 our  $N_b$  estimates for contact transmission of influenza virus [17]. One possible explanation  
470 for the low  $N_b$  estimate is that the 'competitive mixture' method they used to calculate  
471 bottleneck size considers only two viral populations, analogous to the estimates derived  
472 from a single variant in the methods we considered. The competitive mixture method is,  
473 thus, highly susceptible to fluctuations between donor and recipient variant frequencies  
474 arising from stochastic viral dynamics in the recipient. Thus, for the same reason that the  
475 binomial sampling method we described here underestimates bottleneck sizes, we would  
476 expect this competitive mixture method to considerably underestimate bottleneck sizes.  
477 Yet, this method is free of one of the necessary assumptions made for each the three  
478 methods that we considered, namely that the variants considered are independent. The  
479 independence assumption is clearly violated in this data set given extensive genetic linkage  
480 within influenza gene segments [39]. We can, however, somewhat control for the effects  
481 of linkage by selecting only one variant per gene segment. This data-thinning approach still  
482 assumes independence across gene segments that, while not ideal, may be supported by  
483 recent experimental evidence showing high levels of reassortment *in vitro* [40]. If intrahost  
484 reassortment occurs at similar rates *in vivo*, then sampling only one variant per gene  
485 segment should remove much of the bias due to linkage.

486 The methods we considered make other assumptions that may have also impacted  
487 transmission bottleneck size estimates. These assumptions include that: (1) donor-

488 identified variants did not originate *de novo* in any recipient hosts, (2) variants were  
489 biallelic, and (3) variants were selectively neutral. Significant levels of *de novo* evolution  
490 of variants in recipient hosts would artificially increase estimated bottleneck sizes.  
491 Therefore, these methods may not be appropriate for pathogens causing chronic infections,  
492 such as HIV, where sampling of the recipient host can occur years after infection initiation.  
493 However, we do not expect substantial *de novo* evolution of variants to occur over the  
494 course of an acute influenza infection based on recent findings [37] and the observation  
495 that the vast majority of recipient-identified variants were also present in the donor.  
496 Therefore, we do not expect this assumption to have significantly influenced our bottleneck  
497 size estimates for influenza virus.

498         We also do not expect that the second assumption—that loci are biallelic—to have  
499 biased our bottleneck size estimates. This is because no sites used in our bottleneck size  
500 calculations contained more than one variant allele above our variant calling threshold of  
501 3%. This assumption, however, could be removed in future uses of the beta-binomial  
502 sampling method by appropriately modifying the likelihood expressions to account for  
503 more than one variant per site.

504         The third assumption of selective neutrality is the one that could greatly affect the  
505 accuracy of our bottleneck size estimates if not met. Selection, either for or against a  
506 variant, would lead to larger differences in variant frequency between a donor and a  
507 recipient host than would be expected for neutral variants. Larger differences in variant  
508 frequencies would bias the estimated transmission bottleneck sizes towards smaller values.  
509 Thus, our bottleneck size estimates, which assume neutrality, are necessarily conservative  
510 estimates.

511 In this study, we have developed a new statistical approach that can be used to  
512 accurately infer transmission bottleneck sizes for acute viral infections, such as influenza,  
513 RSV, and norovirus, using NGS data from identified donor-recipient pairs. This beta-  
514 binomial sampling method accounts for the possibility of ‘false negatives’ variants that are  
515 not called as present due to necessary variant calling thresholds. The method further  
516 accounts for changes in variant frequencies between the time of recipient infection and the  
517 time of pathogen sampling from the recipient that arise due to stochastic replication  
518 dynamics early in infection. Given the importance of the transmission bottleneck size in  
519 regulating the rate of pathogen evolution at the level of the host population, estimation of  
520 the transmission bottleneck size is a necessary component in the analysis of pathogens  
521 important to public health. Though methods such as viral tagging to estimate the bottleneck  
522 size for experimental infections exist, these techniques are not applicable for natural  
523 infections. Hence this work provides a strong foundation for future estimation of bottleneck  
524 sizes from viral sequence data that, importantly, can be applied to clinical samples.

## 525 **Materials and Methods**

526 **Development of the beta-binomial sampling method.** Here, we derive the beta-binomial  
527 sampling method for inferring transmission bottleneck sizes from pathogen NGS data. The  
528 final likelihood expressions for this method are provided in equations (3) and (4). As  
529 described above, the method allows variant frequencies in the recipient host to change  
530 between infection and sampling (Figure 1) due to stochastic pathogen dynamics occurring  
531 during the process of replication. More concretely, early in the infection when there are  
532 only a small number of replicating virions, stochasticity in viral growth is expected to have  
533 a large effect. For a stochastic birth-death process with a constant birth rate  $\lambda$  and a constant

534 death rate  $\mu$ , the probability mass function for the viral population size originating from a  
535 single virion that successfully establishes is given by [41]:

$$P(N_k(t) = k) = (1 - \eta_t)\eta_t^{k-1}, k \geq 1, \quad (\text{M1})$$

536 where  $t$  is the time of sampling and  $\eta_t = \frac{\lambda(e^{(\lambda-\mu)t}-1)}{\lambda e^{(\lambda-\mu)t}-\mu}$ . For the bursty replication that  
537 characterizes many viruses, (M1) is still approximately true at long times with an adjusted  
538 value of  $\eta_t$ .

539 The population sizes stemming from each of the  $N_b$  founding virions, contingent on  
540 their successful establishment, are thus geometrically-distributed random variables. As  
541 these population sizes are likely to be very large at the time of sampling, we can  
542 approximate them as being exponentially-distributed random variables. Under this  
543 approximation, the distribution of the fractions of the population that descend from each of  
544 the founding virions is distributed as Dirichlet(1,1,...1), with  $N_b$  1's, one for each ancestor.  
545 A subset  $k$  of these founder virions carry the variant allele; the remaining subset of these  
546 founder virions ( $N_b-k$ ) carry the reference allele. Collapsing the Dirichlet distribution yields  
547 that the fraction of the population carrying the variant allele is distributed as Beta( $k, N_b-k$ ).  
548 Remarkably, this fraction does not depend on the within-host viral birth rate  $\lambda$ , the death  
549 rate  $\mu$ , the time of sampling  $t$ , or the burstiness of replication. To get the overall likelihood  
550 of population bottleneck size  $N_b$ , we simply have to consider all possible scenarios of how  
551 many virions out of the total  $N_b$  virions transferred carried the variant allele. Under the  
552 assumption that the founding pathogen population is randomly sampled from the pathogen  
553 population of the donor host, the probability that the founding population of  $N_b$  virions  
554 carries  $k$  variant alleles is given by the binomial distribution  $p_{bin}(k|N_b, v_{D,i}) \equiv$



555  $\Pr(X = k|N_b, v_{D,i}) = \binom{N_b}{k} (v_{D,i})^k (1 - v_{D,i})^{N_b - k}$ , where the number of trials is given by  
556  $N_b$  and the success probability is given by  $v_{D,i}$ , the frequency of variant  $i$  in the donor. Thus,  
557 the overall likelihood of population bottleneck size  $N_b$  for variant  $i$  is given by:

$$L(N_b)_i = \sum_{k=0}^{N_b} p\_beta(v_{R,i}|k, N_b - k) p\_bin(k|N_b, v_{D,i}), \quad (\text{M2})$$

558 where  $v_{R,i}$  is the frequency of variant  $i$  in the recipient and the term  $p\_beta(v_{R,i}|k, N_b - k)$   
559 is given by the beta probability density function, evaluated at  $v_{R,i}$ . This expression is  
560 provided in the main text as equation (1).

561 Accommodating sampling noise arising from a finite number of reads is simple,  
562 leading to minor modifications to the above equation (M2):

$$L(N_b)_i = \sum_{k=0}^{N_b} p\_betabin(R_{var,i}|R_{tot,i}, k, N_b) p\_bin(k|N_b, v_{D,i}), \quad (\text{M3})$$

563 where  $R_{var,i}$  is the number of reads of the variant allele in the recipient sample at site  $i$ , and  
564  $R_{tot}$  is the total number of reads at that site. The term  $p\_betabin(R_{var,i}|R_{tot,i}, k, N_b)$  is  
565 given by the beta-binomial distribution evaluated at  $R_{var,i}$  and parameterized with  $R_{tot,i}$  as  
566 number of trials and parameters  $k$  and  $N_b$ . Expression (M3), reproduced in the main text as  
567 equation (3), thus incorporates both noise from the sampling process itself and from the  
568 process of stochastic pathogen growth. The overall likelihood of bottleneck size  $N_b$  for a  
569 transmission pair is simply the product of the site-specific likelihoods.

570 As previously mentioned, we expect that variant calling thresholds will impact the  
571 likelihood calculations used in the bottleneck size estimation. These thresholds will force  
572 some variant alleles in the recipient viral population to be called absent when they are  
573 actually present at frequencies below the value of the chosen threshold. Since true absence

574 of a variant allele is more likely at smaller bottleneck sizes, conservative variant calling  
575 thresholds will bias  $N_b$  estimates to lower values. Simply excluding variants that are called  
576 absent from the analysis, however, will also bias bottleneck size estimates, this time  
577 towards higher values. To get around this, we do not recommend simply lowering the  
578 variant calling threshold because NGS sequencing error can give rise also to false positives,  
579 thereby inappropriately inflating bottleneck size estimates. Instead, we recommend  
580 accommodating below-threshold variants in the following way. For a donor-identified  
581 variant  $i$  that is called absent in the recipient (whether truly absent or just called absent),  
582 the likelihood of the transmission bottleneck size is given by the following expression:

$$L(N_b)_i = \sum_{k=0}^{N_b} p\_beta\_cdf(v_{R,i} < T|k, N_b - k) p\_bin(k|N_b, v_{D,i}), \quad (\text{M4})$$

583 where  $T$  is the variant calling threshold (e.g., of 3%) and  $p\_beta\_cdf(v_{R,i} < T|k, N_b - k)$   
584 is given by the beta cumulative distribution function evaluated at the variant calling  
585 threshold. This expression is reproduced in the main text as equation (2). We can again  
586 incorporate the effects of sampling noise by considering the number of reads at the variant  
587 site with the expression:

$$L(N_b)_j = \sum_{k=0}^{N_b} [p\_betabin\_cdf(R_{var,i} < TR_{tot,i}|k, N_b) p\_bin(k|N_b, v_{D,i})], \quad (\text{M5})$$

588 where, in this case,  $p\_betabin\_cdf(R_{var,j} < TR_{tot,j}|k, N_b)$  is given by the beta-binomial  
589 cumulative distribution function evaluated at the number of reads that would qualify as  
590 falling at the variant calling threshold. This expression reproduces equation (4) of the main  
591 text.

592 Once the transmission bottleneck sizes have been estimated using the beta-binomial  
593 sampling method, the probability of true variant presence/absence in the recipient host can

594 be determined for any given donor variant frequency. Similarly, the probability that a  
595 variant is called present/absent can be determined for any given donor frequency  $v_{D,i}$ , given  
596 a sufficiently high read count in the recipient host. Given a high read count, the probability  
597 that a variant is called present in the recipient is given by:  $\sum_{k=0}^{N_b} [1 - p\_beta\_cdf(v_{R,i} <$   
598  $T | k, N_b - k)] p\_bin(k | N_b, v_{D,i})$ .

### 599 **The binomial sampling method.**

600 In contrast to the beta-binomial sampling method, the binomial sampling method  
601 implicitly assumes that the infecting virus population is subject to deterministic dynamics  
602 between the time of infection and the time at which the recipient virus is sampled and, thus,  
603 that the sampled pathogen population in the recipient perfectly reflects the founding  
604 pathogen population under the common assumption of selective neutrality. The founding  
605 pathogen population is, as in the beta-binomial sampling method, assumed to be randomly  
606 sampled from the pathogen population of the donor host. The site-specific likelihood of the  
607 transmission bottleneck size  $N_b$  is therefore given by:

$$L(N_b)_i = \sum_{k=0}^{N_b} [p\_bin(R_{var,i} | R_{tot,i}, \frac{k}{N_b}) p\_bin(k | N_b, v_{D,i})], \quad (M6)$$

608 where  $p\_bin(R_{var,i} | R_{tot,i}, f_k) = \binom{R_{tot,i}}{R_{var,i}} \left(\frac{k}{N_b}\right)^{R_{var,i}} \left(1 - \frac{k}{N_b}\right)^{R_{tot,i} - R_{var,i}}$ . This expression  
609 reproduces equation (6) in the main text. The overall likelihood of transmission bottleneck  
610 size  $N_b$  is calculated by multiplying across all site-specific likelihoods.

611 The above expression incorporates sampling noise, which is important when only  
612 a small number of reads are available. With an increasing number of reads, sampling noise  
613 necessarily goes down, making  $p\_bin\left(R_{var,i} | R_{tot,i}, \frac{k}{N_b}\right) \approx 0$  in cases where  $\frac{R_{var,i}}{R_{tot,i}} \neq \frac{k}{N_b}$ .  
614 This will result in dramatic differences in likelihood values between small values of  $N_b$ ,

615 and more generally, multi-modal likelihood curves that are very sensitive to specific variant  
616 frequencies in the recipient host.

617 One basic issue with this approach is therefore the assumption of where differences  
618 in variant frequencies across donor-recipient pairs stem from. Under this model, any  
619 observed differences are due to the presence of a transmission bottleneck because it  
620 assumes that the sampled pathogen population in the recipient perfectly reflects the  
621 founding pathogen population. This assumption is met under a scenario of deterministic,  
622 and neutral, viral population dynamics between the time of the transmission event and the  
623 time of pathogen sampling from the recipient host. For example, if we assume deterministic  
624 exponential growth from the time of the transmission event to the time of sampling, the  
625 dynamics of the viral population that carries the variant allele is given by  $N_v(t) =$   
626  $N_v(0)e^{rt}$ , and, similarly, the dynamics of the viral population that carries the reference  
627 allele is given by  $N_r(t) = N_r(0)e^{rt}$ . At the time of the transmission event ( $t = 0$ ), the  
628 fraction of the viral population that carries the variant allele is given by  $k/N_b$ . At time  $t$ ,  
629 the fraction of the viral population that carries the variant allele is given by  $N_v(t)/(N_v(t) +$   
630  $N_r(t))$ , which simplifies to  $k/N_b$ .

631 The bottleneck size estimates inferred with the binomial sampling method are again  
632 subject to the effects of ‘false negative’ variant calls. We can modify the binomial sampling  
633 method to incorporate the variant call threshold in a way similar to how the threshold  
634 frequency was incorporated into the beta-binomial sampling method. For a donor-  
635 identified variant  $i$  that is called absent in the recipient (whether truly absent or just called  
636 absent), the likelihood of the transmission bottleneck size is:

$$L(N_b)_i = \sum_{k=0}^{N_b} [p\_bin\_cdf \left( R_{var,i} < TR_{tot,i} \left\lfloor \frac{k}{N_b} \right\rfloor \right) p\_bin(k|N_b, v_{D,i})]. \quad (M7)$$

637 This expression reproduces equation (7) in the main text. The probability that the number  
638 of variant reads falls below the level required for the variant to be called present is given  
639 by the binomial cumulative distribution function:

$$640 \quad p\_bin\_cdf \left( R_{var,i} < \lfloor TR_{tot,i} \rfloor \left\lfloor \frac{k}{N_b} \right\rfloor \right) = \sum_{j=0}^{\lfloor TR_{tot,i} \rfloor} \binom{R_{tot,i}}{j} \left( \frac{k}{N_b} \right)^j \left( 1 - \frac{k}{N_b} \right)^{R_{tot,i}-j},$$

641 where  $\lfloor TR_{tot,i} \rfloor$  is the largest integer smaller than  $TR_{tot,i}$ .

642 As with the beta-binomial sampling method, once transmission bottleneck sizes  
643 have been estimated using the binomial sampling method, the probability of true variant  
644 presence/absence in the recipient host can be determined for any given donor variant  
645 frequency. Similarly, the probability that a variant is called present/absent can be  
646 determined for any given donor frequency  $v_{D,i}$ , provided information on the total read count  
647 in the recipient. Specifically, in the case of a high number of reads, the probability that a  
648 variant is called present (whether it is absent or present in the recipient host) is given by  
649  $\sum_{k=0}^{N_b} B(k, N_b, T) p\_bin(k|N_b, v_{D,i})$ , where  $B(k, N_b, T)$  is a Boolean function that  
650 evaluates to 1 if  $\frac{k}{N_b} > T$  and 0 otherwise.

### 651 **Simulated deep-sequencing data**

652 To illustrate the use of the methods used to estimate  $N_b$ , we generated a mock deep-  
653 sequencing dataset via simulation. For this dataset, we assumed a single donor-recipient  
654 pair, with 500 independent donor-identified variants. Independently for both the donor and  
655 the recipient, we drew the total number of reads at each of the 500 sites from a normal  
656 distribution with a mean of 500 reads and a standard deviation of 100 reads. Draws from

657 the normal distribution were rounded to the nearest integer and those that fell at 0 or below  
658 were discarded. For the donor, we then first determined “true” variant frequencies at each  
659 of these sites by drawing from an exponential distribution with mean frequency of 0.08.  
660 Variants with observed frequencies below the variant calling threshold of 0.03 or above  
661 0.50 were discarded. To determine the number of variant reads at a given site in the donor,  
662 we drew from a binomial distribution with the number of trials being the total read count  
663 at that site in the donor and the probability of success being given by that site’s “true”  
664 variant frequency in the donor. We then incorporated sequencing error by again using  
665 draws from binomial distributions. Specifically, we determined the number of “true”  
666 reference reads in the donor that were misclassified as variant reads and the number of  
667 “true” variant reads in the donor that were correctly classified as variant reads, based on an  
668 assumed sequencing error rate of 1%. The total number of observed variant reads at a given  
669 site in a donor was then calculated as the sum of the misclassified reference reads and the  
670 correctly classified variant reads. Observed variant frequencies in the donor were then  
671 calculated by dividing the number of observed variant reads by the total number of  
672 observed reads at each site. In this manner, we simulated 500 variants, with observed  
673 frequencies in the range of 3-50%. The lower bound value of 3% was our assumed variant  
674 calling threshold; the upper bound value of 50% coincided with a variant allele always  
675 being the minority allele.

676 For the recipient, we simulated the total number of variant reads at each site by first  
677 simply determining at each site the number of virions in the founding population that  
678 carried the variant allele, under the assumption of a transmission bottleneck size of  $N_b =$   
679 50. This was done by, at each site, drawing from a binomial distribution with the number

680 of trials being  $N_b$  and the probability of success being the “true” variant frequency at that  
681 site in the donor. For the simulated data set, we first determined the “true” fraction of the  
682 viral population carrying the variant allele at the time of sampling by drawing from a beta  
683 distribution with the shape parameter being the number of variant alleles in the founder  
684 population and the scale parameter being the difference between the founding population  
685 size of  $N_b$  and the number of variant alleles in the founder population. The “true” number  
686 of variant reads was then determined by drawing from a binomial distribution with the  
687 number of trials being the total number of reads at that site and the probability of success  
688 being the fraction of the population at the time of sampling that carried the variant allele.  
689 We then obtained the total number of variant reads at a given site in a recipient by  
690 introducing sequencing error to the “true” number of variant reads and the “true” number  
691 of reference reads.

## 692 **Application to Influenza A deep sequencing data**

693 We applied the three methods for bottleneck size inference described in in the *Models*  
694 section to the influenza A deep-sequencing data examined in [28]. In this study, Poon and  
695 colleagues identified donor-recipient transmission pairs based on household information  
696 and the genetic similarities between the viral populations in infected hosts. We base our  
697 analyses on these already-identified transmission pairs. In some cases, there were several  
698 members of the household who became infected. In this subset of cases, rather than  
699 considering all feasible pairwise combinations of who-infected-whom, we assumed that  
700 the index case transmitted to multiple household members. With this assumption, the 9  
701 identified transmission pairs for influenza A subtype H1N1p were 681\_V1(0) →  
702 681\_V3(2), 684\_V1(0) → 684\_V2(3), 712\_V1(0) → 712\_V1(4), 742\_V1(0) →

703 742\_V3(3), 751\_V1(0) → 751\_V3(1), 751\_V1(0) → 751\_V2(3), 751\_V1(0) →  
704 751\_V2(4), 779\_V1(0) → 779\_V2(1), 779\_V1(0) → 779\_V1(2), where  $X\_VY(Z)$  refers to  
705 household number  $X$ , visit number  $Y$ , subject  $Z$ , and the arrow demarcates transmission  
706 from donor to recipient. The 7 identified transmission pairs for influenza A subtype H3N2  
707 were 689\_V1(0) → 689\_V2(2), 720\_V1(0) → 720\_V2(1), 734\_V1(0) → 734\_V3(2),  
708 739\_V1(0) → 739\_V2(2), 739\_V1(0) → 739\_V2(3), 747\_V1(0) → 747\_V2(2), and  
709 763\_V1(0) → 763\_V2(3). The deep-sequencing data are publically available from [28] and  
710 from <https://www.synapse.org/#!Synapse:syn8033988>. We called variants and determined  
711 variant frequencies from these data using VarScan [42,43], using a variant calling threshold  
712 of 3%, mean quality score of 20, and a p-value of 0.05. We provide variants and their  
713 frequencies as used in this study as a supplementary data file.

714 **Calculation of overall transmission bottleneck sizes across transmission pairs.** To  
715 calculate transmission bottleneck sizes over multiple transmission pairs, we did not simply  
716 take the sum of log-likelihoods across transmission pairs. Taking simply the sum would  
717 inappropriately give greater weight to transmission pairs with a larger number of donor-  
718 identified variants. To weight each of the transmission pairs equally, we scaled the log-  
719 likelihood of each transmission pair based on the number of variants identified in that  
720 transmission pair, such that the overall log-likelihood was given by  $\sum_{p=1}^N \frac{n_{max}}{n_p} \log L_p(N_b)$ ,  
721 where  $N$  is the number of transmission pairs,  $n_p$  is the number of donor-identified variants  
722 in transmission pair  $p$ ,  $n_{max} = \max(n_p)$ , and  $\log L_p(N_b)$  are the log-likelihoods across  $N_b$   
723 values in transmission pair  $p$ .

724



## 725 **Acknowledgements**

726 This work was funded by MIDAS CIDID U54-GM111274, supporting KK and ASL.

727 ASL was further supported by the Duke Medical Scientists Training Program grant T32

728 GM007171. EG was supported by U01 AI111598.

729

## 730 **Figure captions**

731 **Figure 1. Schematic showing virus transmission from donor to recipient host.** The

732 number of virions that initiate infection in the recipient host is defined as the transmission

733 bottleneck size or founding population size  $N_b$ . The viral sampling process is shown, with

734 deep sequencing of the viral population resulting in reads that carry polymorphisms at

735 certain nucleotide sites. The nucleotide read-outs at any site can be used to estimate variant

736 frequencies. Dashed horizontal lines in the variant frequency plots denote the variant

737 calling cutoff or threshold. The goal is to estimate  $N_b$  given data on variant frequencies in

738 the donor, and in the recipient, the total number of reads and the number of variant reads

739 at each of the variant sites identified in the donor.

740 **Figure 2.** Estimated transmission bottleneck sizes for a simulated NGS dataset. (A)

741 Scatterplot showing the frequencies of donor-identified variants against corresponding

742 frequencies of these variants in the recipient. Points in black are variants that are called

743 present in the recipient host. Points in grey are variants that are called absent in the recipient

744 host. Black line shows where  $v_{\text{donor}} = v_{\text{recipient}}$ . Gray lines show the variant calling threshold

745 of 3%. (B) The beta-binomial sampling method's log-likelihood curve over a range of  $N_b$

746 values. Maximum likelihood estimate (MLE) = 55 virions (95% CI = 47-64 virions).

747 Likelihood at MLE = -1972.7. (C) The presence/absence method's log-likelihood curve  
748 over a range of  $N_b$  values. MLE = 19 virions (95% CI = 16-22 virions). (D) The binomial  
749 sampling method's log-likelihood curve over a range of  $N_b$  values. MLE = 32 virions (95%  
750 CI = 28-36 virions). Likelihood at MLE = -1981.8. In (B)-(D), vertical black lines show  
751 the true transmission bottleneck size of  $N_b = 50$ . Vertical colored lines show the MLE, and  
752 shaded areas show the 95% confidence interval, determined using the likelihood ratio test.  
753 (E) Likelihood surfaces for a single variant present in the recipient at a frequency of 16.9%  
754 under the beta-binomial sampling model and the binomial sampling model.

755 **Figure 3.** Additional results from application of the beta-binomial sampling method to the  
756 simulated dataset. (A) The probability of a donor-identified variant being either transferred  
757 or observed as transferred ("called") in a recipient host, as a function of donor variant  
758 frequencies. Observed probabilities of donor-identified variants being called in a recipient  
759 host are shown in black, calculated directly from the simulated dataset using 3% frequency  
760 bins. 95% confidence intervals assume the probability of variant transfer follows a  
761 binomial distribution with the number of trials being the number of donor-identified  
762 variants present in a frequency bin and the success probability given by the calculated  
763 probability of transferred variants observed in the frequency bin. Probabilities of donor-  
764 identified variants being truly present in a recipient host are shown in purple, given  
765 bottleneck size estimates from the beta-binomial sampling method. Probabilities of donor-  
766 identified variants being called present in a recipient host are shown in gray, given  
767 bottleneck size estimates from the beta-binomial sampling method. (B)  $N_b$  estimates for  
768 simulated datasets that differ in coverage levels. At each coverage level, 5 datasets were  
769 generated, under the same parameters and assumptions as the dataset shown in Figure 2A.

770 The exact beta-binomial sampling method and the approximate version of this method were  
771 both used to estimate  $N_b$  for each dataset.  $N_b$  maximum likelihood estimates and 95%  
772 confidence intervals are shown, in purple for the exact beta-binomial sampling method and  
773 in pink for the approximate method.

774 **Figure 4.** Transmission bottleneck sizes estimated for influenza A virus transmission pairs  
775 H1N1p (A) and H3N2 (B).  $N_b$  estimates are shown for the exact beta-binomial sampling  
776 method (purple) and the approximate version of this method (pink). Bars show mean and  
777 95% CI, calculated using the likelihood ratio test. Overall transmission bottleneck sizes  
778 estimated across H1N1p transmission pairs (*'H1N1p'*, teal), across H3N2 transmission  
779 pairs (*'H3N2'*, teal), and across both subtypes (*'N<sub>b</sub>'*, orange), under the assumption of a  
780 negative binomial distribution, are also shown. Previous Poon et al. estimates are further  
781 shown (*'N<sub>e</sub>'*) for H1N1p and H3N2 (black). Bars for the Poon et al. estimates show mean  
782 estimated effective population sizes and mean s.d. ranges.

783

784 **Figure 5.** Overall influenza A bottleneck size estimates and probabilities of variant transfer  
785 under these estimates. (A) The negative binomial probability density function describing  
786 overall transmission bottleneck sizes across H1N1p and H3N2 viral subtypes,  
787 parameterized with the MLE values of  $r = 4$  and  $p = 0.980$ . Vertical black lines show the  
788 95% range of this distribution. The MLE bottleneck size estimates for the H3N2 (orange)  
789 and H1N1 (green) transmission pairs are shown above the pdf. (B) The probability of a  
790 donor-identified variant either being transferred or identified (“called”) in the recipient  
791 host, as a function of donor variant frequency. Probabilities of a donor variant being present  
792 in a recipient host are shown in purple, given bottleneck size estimates provided by the

793 negative binomial distribution shown in (A). Probabilities of donor identified variants  
794 being called present in a recipient host, given these same bottleneck size estimates and the  
795 assumptions of the beta-binomial sampling models, are shown in gray. The empirical  
796 probabilities of donor-identified variants being called in a recipient, as calculated from the  
797 combined H1N1p and H3N2 datasets over 3% frequency bins, are shown in black.

798

799

800

801

802

803

804

805

806

807

808

809

810

811

812

813

## 814 **Bibliography**

- 815 1. Gutiérrez S, Michalakis Y, Blanc S. Virus population bottlenecks during within-  
816 host progression and host-to-host transmission. *Curr Opin Virol.* 2012;2: 546–555.  
817 doi:10.1016/j.coviro.2012.08.001
- 818 2. Geoghegan JL, Senior AM, Holmes EC. Pathogen population bottlenecks and  
819 adaptive landscapes: overcoming the barriers to disease emergence. *Proc R Soc B*  
820 *Biol Sci.* 2016;283: 20160727. doi:10.1098/rspb.2016.0727
- 821 3. Skums P, Bunimovich L, Khudyakov Y. Antigenic cooperation among intrahost  
822 HCV variants organized into a complex network of cross-immunoreactivity. *Proc*  
823 *Natl Acad Sci U S A.* 2015;112: 6653–8. doi:10.1073/pnas.1422942112
- 824 4. Xue KS, Hooper KA, Ollodart AR, Dingens AS, Bloom JD. Cooperation between  
825 distinct viral variants promotes growth of h3n2 influenza in cell culture. *Elife.*  
826 2016;5: 1–15. doi:10.7554/eLife.13974
- 827 5. Brooke CB, Ince WL, Wrammert J, Ahmed R, Wilson PC, Bennink JR, et al. Most  
828 influenza A virions fail to express at least one essential viral protein. *J Virol.*  
829 2013;87: 3155–3162. doi:10.1128/JVI.02284-12
- 830 6. Worby CJ, Lipsitch M, Hanage WP. Within-Host Bacterial Diversity Hinders  
831 Accurate Reconstruction of Transmission Networks from Genomic Distance Data.  
832 *PLoS Comput Biol.* 2014;10. doi:10.1371/journal.pcbi.1003549
- 833 7. De Maio N, Wu C-H, Wilson DJ. SCOTTI: Efficient Reconstruction of  
834 Transmission within Outbreaks with the Structured Coalescent. 2016; Available:  
835 <http://arxiv.org/abs/1603.01994>
- 836 8. Hall JS, French R, Hein GL, Morris TJ, Stenger DC. Three distinct mechanisms

- 837 facilitate genetic isolation of sympatric wheat streak mosaic virus lineages.  
838 Virology. 2001;282: 230–6. doi:10.1006/viro.2001.0841
- 839 9. Sacristán S, Malpica JM, Fraile A, García-Arenal F. Estimation of Population  
840 Bottlenecks during Systemic Movement of Tobacco Mosaic Virus in Tobacco  
841 Plants. J Virol. 2003;77: 9906–9911. doi:10.1128/JVI.77.18.9906-9911.2003
- 842 10. Moury B, Fabre F, Senoussi R. Estimation of the number of virus particles  
843 transmitted by an insect vector. Proc Natl Acad Sci U S A. 2007;104: 17891–6.  
844 doi:10.1073/pnas.0702739104
- 845 11. Betancourt M, Fereres A, Fraile A, García-Arenal F. Estimation of the effective  
846 number of founders that initiate an infection after aphid transmission of a  
847 multipartite plant virus. J Virol. 2008;82: 12416–12421. doi:10.1128/JVI.01542-08
- 848 12. Zwart MP, Daròs JA, Elena SF. One is enough: In vivo effective population size is  
849 dose-dependent for a plant RNA virus. PLoS Pathog. 2011;7.  
850 doi:10.1371/journal.ppat.1002122
- 851 13. Fabre F, Moury B, Johansen EI, Simon V, Jacquemond M, Senoussi R. Narrow  
852 Bottlenecks Affect Pea Seedborne Mosaic Virus Populations during Vertical Seed  
853 Transmission but not during Leaf Colonization. PLoS Pathog. 2014;10: e1003833.  
854 doi:10.1371/journal.ppat.1003833
- 855 14. Smith DR, Adams AP, Kenney JL, Wang E, Weaver SC. Venezuelan equine  
856 encephalitis virus in the mosquito vector *Aedes taeniorhynchus*: Infection initiated  
857 by a small number of susceptible epithelial cells and a population bottleneck.  
858 Virology. 2008;372: 176–186. doi:10.1016/j.virol.2007.10.011
- 859 15. Zwart MP, Hemerik L, Cory JS, de Visser JAGM, Bianchi FJJ a, Van Oers MM, et

- 860 al. An experimental test of the independent action hypothesis in virus-insect  
861 pathosystems. *Proc Biol Sci.* 2009;276: 2233–2242. doi:10.1098/rspb.2009.0064
- 862 16. van der Werf W, Hemerik L, Vlak JM, Zwart MP. Heterogeneous host  
863 susceptibility enhances prevalence of Mixed-Genotype Micro-Parasite infections.  
864 *PLoS Comput Biol.* 2011;7. doi:10.1371/journal.pcbi.1002097
- 865 17. McCaw JM, Arinaminpathy N, Hurt AC, McVernon J, McLean AR. A  
866 Mathematical Framework for Estimating Pathogen Transmission Fitness and  
867 Inoculum Size Using Data from a Competitive Mixtures Animal Model. *PLoS*  
868 *Comput Biol.* 2011;7. doi:10.1371/journal.pcbi.1002026
- 869 18. Forrester NL, Guerbois M, Seymour RL, Spratt H, Weaver SC. Vector-Borne  
870 Transmission Imposes a Severe Bottleneck on an RNA Virus Population. *PLoS*  
871 *Pathog.* 2012;8: e1002897. doi:10.1371/journal.ppat.1002897
- 872 19. Hoelzer K, Murcia PR, Baillie GJ, Wood JLN, Metzger SM, Osterrieder N, et al.  
873 Intra-host Evolutionary Dynamics of Canine Influenza Virus in Naïve and Partially  
874 Immune Dogs. *J Virol.* 2010;84: 5329–5335. doi:10.1128/JVI.02469-09
- 875 20. Stack JC, Murcia PR, Grenfell BT, Wood JLN, Holmes EC. Inferring the inter-  
876 host transmission of influenza A virus using patterns of intra-host genetic  
877 variation. *Proc R Soc B Biol Sci.* 2012;280: 20122173–20122173.  
878 doi:10.1098/rspb.2012.2173
- 879 21. Murcia PR, Hughes J, Battista P, Lloyd L, Baillie GJ, Ramirez-Gonzalez RH, et al.  
880 Evolution of an Eurasian Avian-like Influenza Virus in Naïve and Vaccinated Pigs.  
881 *PLoS Pathog.* 2012;8: e1002730. Available: [http://www.ncbi.nlm-nih-](http://www.ncbi.nlm.nih-gov.proxy.lib.duke.edu/pmc/articles/PMC3364949/?tool=pmcentrez&report=abstr)  
882 [gov.proxy.lib.duke.edu/pmc/articles/PMC3364949/?tool=pmcentrez&report=abstr](http://www.ncbi.nlm-nih-gov.proxy.lib.duke.edu/pmc/articles/PMC3364949/?tool=pmcentrez&report=abstr)

- 883 act
- 884 22. Hughes J, Allen RC, Baguelin M, Hampson K, Baillie GJ, Elton D, et al.  
885 Transmission of Equine Influenza Virus during an Outbreak Is Characterized by  
886 Frequent Mixed Infections and Loose Transmission Bottlenecks. *PLoS Pathog.*  
887 2012;8: e1003081. doi:10.1371/journal.ppat.1003081
- 888 23. Murcia PR, Baillie GJ, Daly J, Elton D, Jervis C, Mumford JA, et al. Intra- and  
889 Interhost Evolutionary Dynamics of Equine Influenza Virus. *J Virol.* 2010;84:  
890 6943–6954. doi:10.1128/JVI.00112-10
- 891 24. Wilker PR, Dinis JM, Starrett G, Imai M, Hatta M, Nelson W, et al. Selection on  
892 hemagglutinin imposes a bottleneck during mammalian transmission of reassortant  
893 H5N1 influenza viruses. *Nat Commun.* 2013;4.  
894 doi:10.1038/ncomms3636.Selection
- 895 25. Zaraket H, Baranovich T, Kaplan BS, Carter R, Song M-S, Paulson JC, et al.  
896 Mammalian adaptation of influenza A(H7N9) virus is limited by a narrow genetic  
897 bottleneck. *Nat Commun.* Nature Publishing Group; 2015;6: 6553.  
898 doi:10.1038/ncomms7553
- 899 26. Moncla LH, Zhong G, Nelson CW, Dinis JM, Mutschler J, Hughes AL, et al.  
900 Selective Bottlenecks Shape Evolutionary Pathways Taken during Mammalian  
901 Adaptation of a 1918-like Avian Influenza Virus. *Cell Host Microbe.* Elsevier Inc.;  
902 2016;19: 169–180. doi:10.1016/j.chom.2016.01.011
- 903 27. Varble A, Albrecht RA a, Backes S, Crumiller M, Bouvier NMM, Sachs D, et al.  
904 Influenza A virus transmission bottlenecks are defined by infection route and  
905 recipient host. *Cell Host Microbe.* 2014;16: 691–700.



- 906           doi:10.1016/j.chom.2014.09.020
- 907   28.   Poon LLM, Song T, Rosenfeld R, Lin X, Rogers MB, Zhou B, et al. Quantifying  
908           influenza virus diversity and transmission in humans. *Nat Genet. Nature*  
909           Publishing Group; 2016;1: 1–6. doi:10.1038/ng.3479
- 910   29.   Frise R, Bradley K, van Doremalen N, Galiano M, Elderfield RA, Stilwell P, et al.  
911           Contact transmission of influenza virus between ferrets imposes a looser  
912           bottleneck than respiratory droplet transmission allowing propagation of antiviral  
913           resistance. *Sci Rep. Nature Publishing Group*; 2016;6: 29793.  
914           doi:10.1038/srep29793
- 915   30.   Emmett KJ, Lee A, Khiabani H, Rabadan R. High-resolution Genomic  
916           Surveillance of 2014 Ebolavirus Using Shared Subclonal Variants. *PLoS Curr.*  
917           2015;7: 1–17.  
918           doi:10.1371/currents.outbreaks.c7fd7946ba606c982668a96bcba43c90
- 919   31.   Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, et al. Best  
920           practices for evaluating single nucleotide variant calling methods for microbial  
921           genomics. *Front Genet.* 2015;6: 235. doi:10.3389/fgene.2015.00235
- 922   32.   Ghedin E, Holmes EC, DePasse J V., Pinilla LT, Fitch A, Hamelin M-E, et al.  
923           Presence of Oseltamivir-Resistant Pandemic A/H1N1 Minor Variants Before Drug  
924           Therapy With Subsequent Selection and Transmission. *J Infect Dis.* 2012;206:  
925           1504–1511. doi:10.1093/infdis/jis571
- 926   33.   Van den Hoecke S, Verhelst J, Vuylsteke M, Saelens X. Analysis of the genetic  
927           diversity of influenza A viruses using next-generation DNA sequencing. *BMC*  
928           Genomics. 2015;16. doi:10.1186/s12864-015-1284-z

- 929 34. Lakdawala SS, Jayaraman A, Halpin RA, Lamirande EW, Shih AR, Stockwell TB,  
930 et al. The soft palate is an important site of adaptation for transmissible influenza  
931 viruses. *Nature*. Nature Publishing Group, a division of Macmillan Publishers  
932 Limited. All Rights Reserved.; 2015;526: 122–5. doi:10.1038/nature15379
- 933 35. Dinis JM, Florek NW, Fatola OO, Moncla LH, Mutschler JP, Charlier OK, et al.  
934 Deep sequencing reveals potential antigenic variants at low frequency in influenza  
935 A-infected humans. *J Virol*. 2016;2013: JVI.03248-15. doi:10.1128/JVI.03248-15
- 936 36. Sacristán S, Díaz M, Fraile A, García-Arenal F. Contact transmission of Tobacco  
937 mosaic virus: a quantitative analysis of parameters relevant for virus evolution. *J*  
938 *Virol*. 2011;85: 4974–4981. doi:10.1128/JVI.00057-11
- 939 37. Sobel Leonard A, McClain MT, Smith GJD, Wentworth DE, Halpin RA, Lin X, et  
940 al. Deep Sequencing of Influenza A Virus from a Human Challenge Study Reveals  
941 a Large Founder Population Size and Rapid Intrahost Viral Evolution. *J Virol*.  
942 2016;
- 943 38. Buhnerkempe MG, Gostic K, Park M, Ahsan P, Belser JA, Lloyd-Smith JO.  
944 Mapping influenza transmission in the ferret model to transmission in humans.  
945 *Elife*. 2015;4. doi:10.7554/eLife.07969
- 946 39. Boni MF, Zhou Y, Taubenberger JK, Holmes EC. Homologous Recombination Is  
947 Very Rare or Absent in Human Influenza A Virus. *J Virol*. 2008;82: 4807–4811.  
948 doi:10.1128/JVI.02683-07
- 949 40. Ince WL, Gueye-Mbaye A, Bennink JR, Yewdell JW. Reassortment Complements  
950 Spontaneous Mutation in Influenza A Virus NP and M1 Genes To Accelerate  
951 Adaptation to a New Host. *J Virol*. 2013;87: 4330–4338. doi:10.1128/JVI.02749-

- 952 12
- 953 41. Kendall DG. On the Generalized “Birth-and-Death” Process. *Ann Math Stat.*
- 954 1948;19: 1–15.
- 955 42. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al.
- 956 VarScan: variant detection in massively parallel sequencing of individual and
- 957 pooled samples. *Bioinformatics.* 2009;25: 2283–2285.
- 958 doi:10.1093/bioinformatics/btp373
- 959 43. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan
- 960 2: Somatic mutation and copy number alteration discovery in cancer by exome
- 961 sequencing. *Genome Res.* 2012;22: 568–576. doi:10.1101/gr.129684.111
- 962









