AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology®

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

Ashley Sobel Leonard,[a] Daniel B. Weissman,[b] Benjamin Greenbaum,[c] Elodie Ghedin,[d] Katia Koelle[a]

Department of Biology, Duke University, Durham, North Carolina, USA[a]; Department of Physics, Emory University, Atlanta, Georgia, USA[b]; Tisch Cancer Institute, Departments of Medicine, Oncological Sciences, and Pathology, Icahn School of Medicine at Mount Sinai, New York, New York, USA[c]; Center for Genomics and Systems Biology, Department of Biology, and College of Global Public Health, New York University, New York, New York, USA[d]

ABSTRACT The bottleneck governing infectious disease transmission describes the size of the pathogen population transferred from the donor to the recipient host. Accurate quantification of the bottleneck size is particularly important for rapidly evolving pathogens such as influenza virus, as narrow bottlenecks reduce the amount of transferred viral genetic diversity and, thus, may decrease the rate of viral adaptation. Previous studies have estimated bottleneck sizes governing viral transmission by using statistical analyses of variants identified in pathogen sequencing data. These analyses, however, did not account for variant calling thresholds and stochastic viral replication dynamics within recipient hosts. Because these factors can skew bottleneck size estimates, we introduce a new method for inferring bottleneck sizes that accounts for these factors. Through the use of a simulated data set, we first show that our method, based on beta-binomial sampling, accurately recovers transmission bottleneck sizes, whereas other methods fail to do so. We then apply our method to a data set of influenza A virus (IAV) infections for which viral deep-sequencing data from transmission pairs are available. We find that the IAV transmission bottleneck size estimates in this study are highly variable across transmission pairs, while the mean bottleneck size of 196 virions is consistent with a previous estimate for this data set. Furthermore, regression analysis shows a positive association between estimated bottleneck size and donor infection severity, as measured by temperature. These results support findings from experimental transmission studies showing that bottleneck sizes across transmission events can be variable and influenced in part by epidemiological factors.

IMPORTANCE The transmission bottleneck size describes the size of the pathogen population transferred from the donor to the recipient host and may affect the rate of pathogen adaptation within host populations. Recent advances in sequencing technology have enabled bottleneck size estimation from pathogen genetic data, although there is not yet a consistency in the statistical methods used. Here, we introduce a new approach to infer the bottleneck size that accounts for variant identification protocols and noise during pathogen replication. We show that failing to account for these factors leads to an underestimation of bottleneck sizes. We apply this method to an existing data set of human influenza virus infections, showing that transmission is governed by a loose, but highly variable, transmission bottleneck whose size is positively associated with the severity of infection of the donor. Beyond advancing our understanding of influenza virus transmission, we hope that

Received 30 January 2017 **Accepted** 21 April 2017

Accepted manuscript posted online 3 May 2017

Citation Sobel Leonard A, Weissman DB, Greenbaum B, Ghedin E, Koelle K. 2017. Transmission bottleneck size estimation from pathogen deep-sequencing data, with an application to human influenza A virus. J Virol 91:e00171-17. https://doi.org/10.1128/JVI.00171-17.

Editor Douglas S. Lyles, Wake Forest University

Address correspondence to Katia Koelle, katia.koelle@duke.edu.

this work will provide a standardized statistical approach for bottleneck size estimation for viral pathogens.

KEYWORDS bottleneck, influenza A virus, next-generation sequencing

Infectious disease transmission relies on the transfer of a pathogenic organism from one host to another. This transfer is characterized by a transmission bottleneck, defined as the size of the founding pathogen population in the recipient host. Accurate quantification of transmission bottleneck sizes for pathogenic organisms is critical for several reasons. First, bottleneck sizes impact levels of genetic diversity in recipient hosts and thereby impact the rate at which pathogens can adapt to host populations, with smaller bottleneck sizes decreasing rates of adaptation (1, 2). Second, when cooperative interactions occur within a pathogen population (e.g., see references 3 and 4) or when viral complementation and cellular coinfection are critical for producing viral progeny (e.g., see reference 5), bottleneck sizes will necessarily impact initial pathogen replication rates, with larger bottleneck sizes enabling the occurrence of these interactions and thus facilitating within-host replication. Finally, transmission bottleneck sizes impact the ability to accurately reconstruct who infected whom during an ongoing epidemic (6), such that estimation of the transmission bottleneck size can point to cases which may be problematic and for which a certain class of phylodynamic inference methods (see reference 7) might be particularly useful.

The transmission bottleneck size has been estimated for a number of pathogenic organisms, including pathogens of plants (8–13) and animals (14–22). While those estimates relied on the distribution of pathogen types in infection recipients, as determined by molecular and phenotypic markers or Sanger sequencing of the pathogen population in donor and recipient hosts, deep-sequencing data have recently started to be used to gauge transmission bottleneck sizes (23–29). Some of those studies characterized the general magnitude of transmission bottleneck sizes, with results indicating that narrow, selective bottlenecks tend to govern the transmission dynamics of viral pathogens that are ill adapted to their recipient hosts (24–26). Studies that instead gauged transmission bottleneck sizes of well-adapted viral pathogens using deep-sequencing data have, in contrast, generally found that they tend to be loose, with many virions initiating infection (23, 28, 29). While many of those studies focus on assessing how “loose” or “narrow” a transmission bottleneck is, other studies have attempted to quantitatively estimate transmission bottleneck sizes. One approach relied on the use of barcoded influenza virus during experimental transmission studies in small mammals, with results indicating that the route of transmission greatly impacts the size of the bottleneck (27).

In natural infections, it is not feasible to rely on barcoded or otherwise marked pathogens. In these cases, statistical approaches have therefore instead been used to quantify bottleneck sizes (28, 30). Two studies have used the Kullback-Leibler divergence index (developed in reference 30) to estimate the viral effective population size initiating infection from deep-sequencing data (28, 30). One of those studies quantified the transmission effective population size for Ebola virus in human-to-human infections (30). The other study quantified this transmission effective population size for human influenza A viruses (IAVs) (28). A second statistical approach used previously (28) makes use of a single-generation population genetic Wright-Fisher model to estimate the effective viral population size initiating infection. While this approach similarly showed that the effective population size following influenza virus transmission in natural human-to-human infection is large, this model yielded quantitatively different results from those of the Kullback-Leibler approach. Furthermore, in both of those studies, it is not clear how the effective population size relates to the transmission bottleneck size. It is worth noting, however, that the effective population size is generally considered to be an underestimate of the true population size, as it represents the minimum population size necessary to establish observed levels of genetic diversity.

Both of these approaches (28, 30) analyze only variants that are identified as being

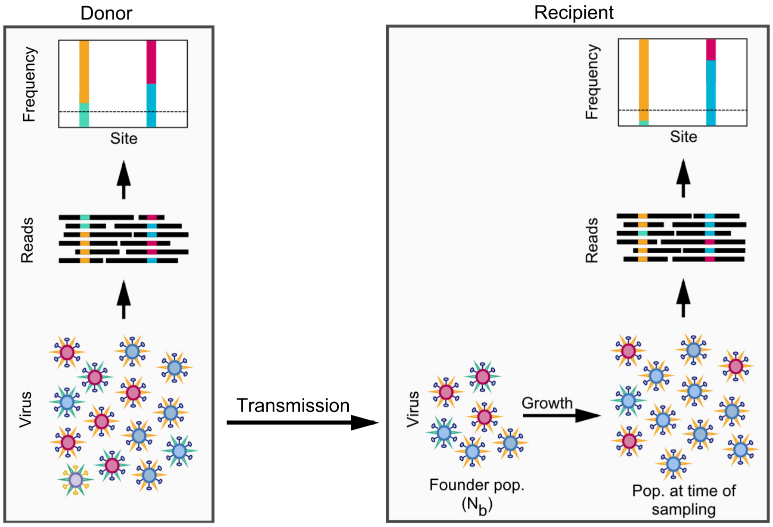

FIG 1 Schematic showing virus transmission from donor to recipient host. The number of virions that initiate infection in the recipient host is defined as the transmission bottleneck size or founding population size, N_b. The viral sampling process is shown, with deep sequencing of the viral population resulting in reads that carry polymorphisms at certain nucleotide sites. The nucleotide readouts at any site can be used to estimate variant frequencies. Dashed horizontal lines in the variant frequency plots denote the variant calling cutoff or threshold. The goal is to estimate N_b given data on variant frequencies in the donor and in the recipient, the total number of reads, and the number of variant reads at each of the variant sites identified in the donor.

present in both the donor and the recipient. However, the absence of a donor variant in a recipient host is also informative, and ignoring such missing variants can significantly bias transmission bottleneck size estimates. Another limitation of both approaches is that they do not consider the effect that stochastic dynamics early in infection may have on variant frequencies in the recipient. To address these concerns, here, we introduce a new method for estimating the transmission bottleneck size of pathogens. This method accounts for stochastic dynamics occurring during viral replication in the recipient and further accounts for variant calling thresholds that are used in calling a variant present or absent in a sample. In addition, this method has the ability to estimate a bottleneck size for individual transmission pairs. We refer to this method as the beta-binomial sampling method, based upon this method's derived likelihood expression. Using a simulated data set, we compare the beta-binomial sampling method to two methods of bottleneck size inference that are present (in some form) in the literature: the presence/absence method and the binomial sampling method. This comparison demonstrates that the beta-binomial sampling method is able to recover the true bottleneck size of the simulated data set, whereas the 2 other methods infer biased estimates by failing to account for variant calling thresholds or stochastic dynamics in the recipient host. Finally, we apply the beta-binomial sampling method to an existing next-generation sequencing (NGS) data set of influenza A virus infections to estimate the transmission bottleneck size in natural human-to-human flu transmission.

Models. Figure 1 provides a schematic of the data that are used for inferring transmission bottleneck sizes in the approaches that we consider in this study. Deep-sequencing data consist of short reads at various sites in the genome, obtained from both the infected donor and the recipient at, generally, a single time point for each individual. The short-read data are used to identify viral variants in the donor and recipient hosts. Comparison of these variants' frequencies across donor-recipient transmission pairs allows us to infer the transmission bottleneck size (N_b), the number of virions comprising the founding viral population at the onset of infection in the recipient host. We specifically define N_b as the number of virions that successfully establish lineages that persist to the sampling time point. There may, however, be additional virions that transiently replicate in the recipient host but quickly die out and are therefore not included in N_b.

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

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

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

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

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

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

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

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

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

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

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

$$L(N_b)_i = \sum_{k=0}^{N_b} p\_betabin\_cdf(R_{var,i} < TR_{tot,i}|R_{tot,i},\ k,\ N_b - k)\ p\_bin(k|N_b,\ \nu_{D,i}) \tag{4}$$

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

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

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

Existing methods for inferring transmission bottleneck sizes. (i) Presence/absence method of bottleneck size inference. The simplest approach to estimating transmission bottleneck sizes from pathogen deep-sequencing data is to calculate variant frequencies in donor hosts and then use information on the presence/absence of these variants in recipient hosts to quantify the bottleneck size. Studies that have adopted this approach have been reported previously (9, 36). Given a variant, *i*, present at frequency $\nu_{D,i}$ in the donor and a founding population size of N_b, the probability that the variant was not transferred to the recipient is simply given by $(1 - \nu_{D,i})^{N_b}$ (9, 36). Correspondingly, the probability that at least one virion in the founding population carried the variant allele is given by $1 - (1 - \nu_{D,i})^{N_b}$. From these expressions, the likelihood of the founding population size of N_b in a donor-recipient pair is simply calculated by multiplying the probabilities of the observed outcomes across the variant sites:

$$L(N_b) = \prod_{j=1}^{V_{absent}} (1 - \nu_{D,j})^{N_b} \prod_{k=1}^{V_{present}} [1 - (1 - \nu_{D,k})^{N_b}] \tag{5}$$

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

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

(ii) Binomial sampling method of bottleneck size inference. The second approach, or class of approaches, from the literature for inferring transmission bottleneck sizes is based on a binomial sampling process. Studies that have adopted this general kind of approach have been reported previously (28, 30). We describe a version of this approach that parallels the beta-binomial sampling method that we describe above. The binomial sampling approach makes use of donor-identified variant frequencies in the donor and both the number of variant reads and the number of total reads in the recipient, at each donor-identified variant site. The likelihood expression for the bottleneck size, given these data at site *i*, is given by

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

where $p\_bin(R_{var,i}|R_{tot,i}, k/N_b)$ is given by the binomial probability density function evaluated at $R_{var,i}$. The term $p\_bin(k|N_b, \nu_{D,i})$ is again given by the binomial distribution. For variants called as absent in the recipient host, the likelihood of the transmission bottleneck size is given as

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

where *p_bin_cdf* is the binomial cumulative distribution function. The derivation of the binomial sampling method can be found in Materials and Methods.

The sole difference between the beta-binomial sampling method and the binomial sampling method is that the binomial sampling method does not account for stochastic dynamics of the pathogen early on in the recipient. These stochastic dynamics enable the frequencies of variants in a recipient at the time of sampling to differ from those at the time of founding (Fig. 1). Because the binomial sampling method does not incorporate this source of frequency variation, we expect there to be smaller frequency deviations between variants in donor-recipient pairs under the assumption of a single-generation binomial sampling model than in a model that allows for these stochastic dynamics, for a given bottleneck size. To explain a given pattern of donor-recipient frequency pairs, N_b estimates are thus expected to be significantly lower for the binomial sampling method than for the beta-binomial sampling method. Application of the binomial sampling method will therefore yield a conservative (lower-bound) estimate of N_b, as previously remarked upon (30).

RESULTS

Results on simulated data. To examine the abilities of the three methods described above to accurately infer transmission bottleneck sizes, we used a simulated data set of one donor-recipient pair (see Materials and Methods). The data set was generated under the assumption of stochastic pathogen dynamics in the recipient host between the time of infection and the time of sampling. While this assumption matches the assumption for the beta-binomial sampling method, we feel that it is also biologically the most realistic assumption. In this data set, 109 out of the 500 donor-identified simulated variants were called absent in the recipient host (Fig. 2A). The majority of these variants were present in the recipient host but below our variant calling threshold of 3% and therefore were false negatives. The beta-binomial sampling method, as expected, recovers the true bottleneck size of 50 virions (Fig. 2B). In contrast, both the presence/absence method (Fig. 2C) and the binomial sampling method (Fig. 2D) significantly underestimate the simulated bottleneck size. The underlying reasons for these methods' inability to recover the true bottleneck size differ. For the presence/absence method, this underestimation can be attributed to false-negative variant calls. For the binomial sampling method, we were able to statistically account for the variant calling threshold effects; the underestimation of this method, therefore, is attributed solely to this method not accounting for stochastic pathogen dynamics in the recipient. The binomial sampling method instead assumes deterministic viral growth from the time of founding to the time of sampling (see Materials and Methods). Because more

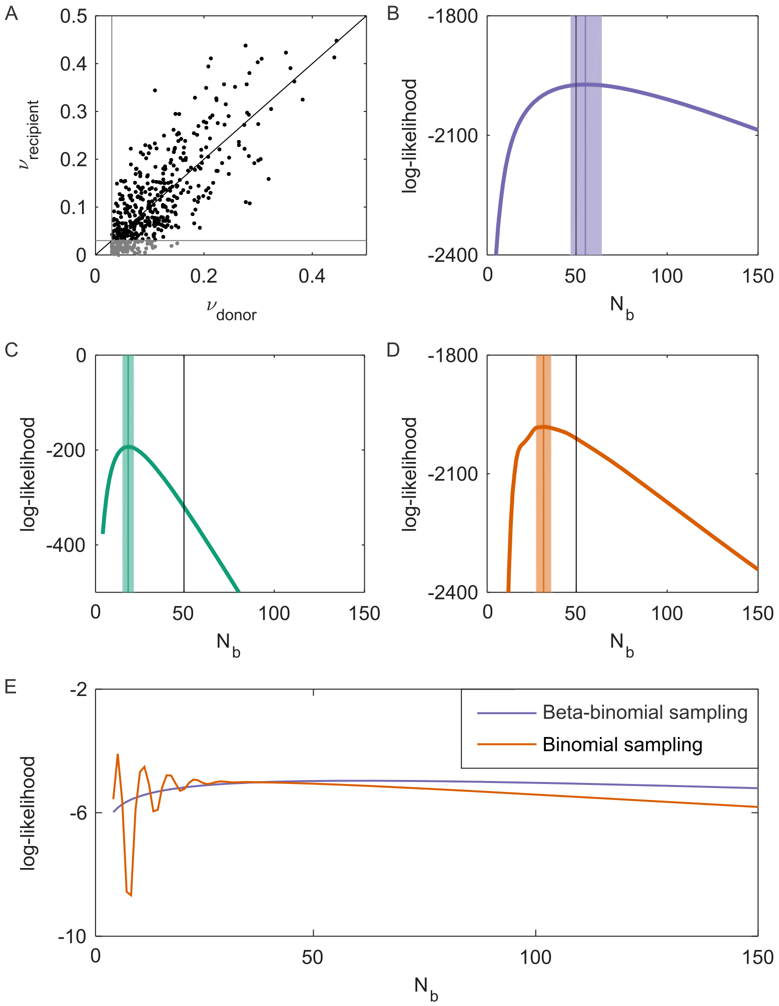

FIG 2 Estimated transmission bottleneck sizes for a simulated NGS data set. (A) Scatterplot showing the frequencies of donor-identified variants against the corresponding frequencies of these variants in the recipient. Points in black are variants that are called present in the recipient host. Points in gray are variants that are called absent in the recipient host. The black line shows where ν_{donor} equals $\nu_{\text{recipient}}$. Gray lines show the variant calling threshold of 3%. (B) Log-likelihood curve for the beta-binomial sampling method over a range of N_b values. The MLE is 55 virions (95% CI = 47 to 64 virions). The likelihood at MLE equals −1,972.7. (C) Log-likelihood curve for the presence/absence method over a range of N_b values. The MLE equals 19 virions (95% CI = 16 to 22 virions). (D) Log-likelihood curve for the binomial sampling method over a range of N_b values. The MLE equals 32 virions (95% CI = 28 to 36 virions). The likelihood at MLE equals −1,981.8. In panels B to D, vertical black lines show the true transmission bottleneck size, N_b, of 50. Vertical colored lines show the MLEs, and shaded areas show the 95% confidence intervals, determined by using the likelihood ratio test. (E) Likelihood surfaces for a single variant present in the recipient at a frequency of 16.9% under the beta-binomial sampling model and the binomial sampling model.

sampling stochasticity is present at smaller bottleneck sizes, the binomial sampling method underestimates the simulated bottleneck size in its attempt to reproduce the observed variation in variant frequencies by inappropriately constricting N_b.

Given that the binomial sampling model and the beta-binomial model were fit to the same data, the relative performances of these models can be assessed by using model selection approaches. The maximum likelihood obtained by using the beta-binomial sampling method was significantly higher than the maximum likelihood obtained by using the binomial sampling method (Fig. 2B and D), indicating that the beta-binomial sampling model is statistically preferred over the binomial sampling model. We can further take into consideration the smoothness of the likelihood curves in our choice of model, with multimodal/rugged likelihood curves being undesirable

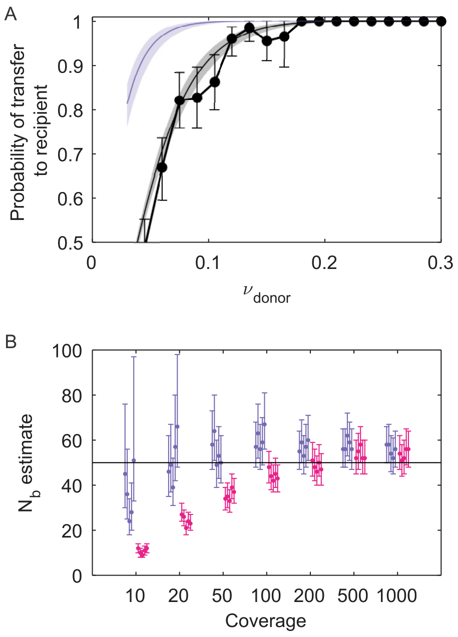

FIG 3 Additional results from application of the beta-binomial sampling method to the simulated data set. (A) Probability of a donor-identified variant being either transferred or observed as transferred ("called") in a recipient host, as a function of donor variant frequencies. The observed probabilities of donor-identified variants being called in a recipient host are shown in black, calculated directly from the simulated data set using 3% frequency bins. The 95% confidence intervals assume that the probability of variant transfer follows a binomial distribution with the number of trials being the number of donor-identified variants present in a frequency bin and the success probability given by the calculated probability of transferred variants observed in the frequency bin. Probabilities of donor-identified variants being truly present in a recipient host are shown in purple, given bottleneck size estimates from the beta-binomial sampling method. Probabilities of donor-identified variants being called present in a recipient host are shown in gray, given bottleneck size estimates from the beta-binomial sampling method. (B) N_b estimates for simulated data sets that differ in coverage levels. At each coverage level, 5 data sets were generated under the same parameters and assumptions as those for the data set shown in Fig. 2A. Both the exact beta-binomial sampling method and the approximate version of this method were used to estimate N_b for each data set. N_b maximum likelihood estimates and 95% confidence intervals are shown in purple for the exact beta-binomial sampling method and in pink for the approximate method.

outcomes. In Fig. 2E, we plot the likelihood curves for one variant under the likelihood expression of the beta-binomial sampling method and under the expression of the binomial sampling method. The rugged likelihood surface of the binomial sampling model arises because of this method's stringent assumption that variant frequencies remain fixed between the time of infection of the recipient and the time of sampling. In contrast, the beta-binomial sampling method allows for stochastic changes in variant frequencies during viral growth, relaxing the assumption that the viral population at the time of sampling needs to perfectly reflect the founding viral population. As a result, likelihood curves of the beta-binomial sampling model do not show large differences in likelihood values for small differences in N_b, further indicating that the beta-binomial sampling model is preferable.

Given an estimate of the transmission bottleneck size, the probability that a variant is transferred to a recipient host can be calculated by using the expression $1 - (1 - \nu_{D,i})^{N_b}$, where $\nu_{D,i}$ is the frequency of variant *i* present in the donor host and N_b is the bottleneck size estimate. In Fig. 3A, we plot this probability of variant transfer over a range of donor variant frequencies for the simulated data set. In this figure, we further plot "observed" probabilities of variant transfer, using a variant calling threshold of 3% for the simulated data set. Finally, in Fig. 3A, we plot the observed probabilities of variant transfer as predicted under the beta-binomial sampling method, evaluated at the transmission bottleneck size estimated. We see first that the true probabilities of variant transfer greatly exceed those that are observed in the data set given the variant calling threshold of 3%. However, this method's calculated predictions of observed variant transfer probabilities fall within the 95% confidence intervals (CIs) for the probabilities of variant transfer observed in the data set.

As described in the introduction, the exact beta-binomial sampling method that we developed accounts for sampling noise arising from finite read coverage. If we ignore sampling noise, we can estimate bottleneck sizes more rapidly using the approximate method, described by equations 1 and 2. In Fig. 3B, we show bottleneck size estimates over a range of different coverage levels for both the exact and approximate beta-binomial sampling methods. At high coverage levels ($>$200 reads), both implementations of the beta-binomial sampling method yield similar bottleneck size estimates and are able to recover the simulated bottleneck size of 50 virions. For lower levels of coverage, however, this approximation starts to fail and will lead to a considerable underestimation of N_b, indicating that the approximate beta-binomial sampling method is inappropriate for low coverage levels. We also note that even at high coverage levels, a slight overestimation of the bottleneck size is apparent for both the beta-binomial and the approximate beta-binomial sampling methods. This overestimation can be attributed to the rare false-positive identification of variants in the recipient (instances of a variant that is absent in the recipient being called present) and, more generally, a slight inflation of variant frequencies with sequencing error. Overestimation no longer occurs when these methods are applied to data sets that are simulated in the absence of sequence error (results not shown).

Transmission bottleneck size estimation for human influenza A virus. We first applied the beta-binomial sampling method for inferring transmission bottleneck sizes to the influenza A/H1N1p virus transmission pairs identified in an influenza virus NGS data set described in detail previously (28). We point the reader to this previous report for details on the data set, including coverage levels and how transmission pairs were inferred, etc. Poon et al. (28) estimated the mean effective population size for all H1N1p transmission pairs, N_e, to be equal to 192 virions (mean standard deviation range, 114 to 276 virions). This approach considered the combined set of variants that were present at frequencies of $\geq$1% and that were shared by 8 identified household donor-recipient pairs (a total of 26 variants). In contrast to that analysis, we estimated transmission bottleneck sizes for each of the 9 transmission pairs separately, using a minimum variant frequency cutoff of 3% to call variants. We used a 3% cutoff based on concordance results from replicate sequencing runs, as described previously (28). The less conservative 1% cutoff used by Poon et al. (28) to estimate the effective population size was chosen to allow for more sites to be included in their analysis. Our analysis, using a total of 289 variants, estimated MLE bottleneck sizes ranging from 49 to 276 virions across the H1N1p transmission pairs (Fig. 4A). The bottleneck sizes inferred by the approximate beta-binomial sampling method did not differ significantly from those inferred by the exact method for any of the transmission pairs. This was expected, given high coverage levels across variant sites.

To summarize our results for the bottleneck size estimates for the H1N1p transmission pairs, we estimated parameters of a negative binomial distribution using all of the variant frequencies across the transmission pairs (see Materials and Methods). This negative binomial distribution was chosen because our results shown in Fig. 4A indicated that the variance in transmission bottleneck sizes is likely to exceed the mean. We further fit a Poisson distribution to the same data, and the negative binomial distribution was statistically preferred over the Poisson distribution using the Akaike information criterion (AIC) indicating that while a single infection may be initiated by a Poisson-distributed number of virions, different infections are likely to be initiated by founding population sizes that vary in their means. The MLE values for the negative binomial distribution's parameters were an *r* value of 5 and a *p* value of 0.966, resulting in a mean H1N1p transmission bottleneck size, N_b, of 142 and a 95% range of 54 to 262 virions (Fig. 4A). While our overall bottleneck size estimates were consistent with the estimates of Poon et al. using a much more limited number of variants, our analysis

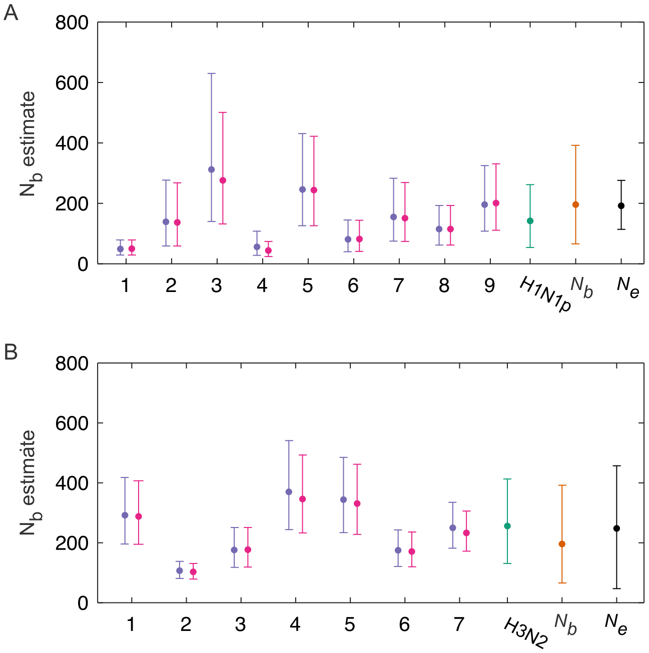

FIG 4 Transmission bottleneck sizes estimated for influenza A virus H1N1p (A) and H3N2 (B) transmission pairs. N_b estimates are shown for the exact beta-binomial sampling method (purple) and the approximate version of this method (pink). Bars show means and 95% CIs, calculated by using the likelihood ratio test. Overall transmission bottleneck sizes estimated across H1N1p transmission pairs ("H1N1p") (teal), across H3N2 transmission pairs ("H3N2") (teal), and across both subtypes ("N_b") (orange), under the assumption of a negative binomial distribution, are also shown. Previous estimates by Poon et al. (28) are also shown ("N_e") for H1N1p and H3N2 (black). Bars for the estimates by Poon et al. show mean estimated effective population sizes and mean standard deviation ranges.

further shows that the transmission bottleneck sizes varied considerably between transmission pairs.

We next used the beta-binomial sampling method to infer the transmission bottleneck sizes for each of the H3N2 transmission pairs of the influenza virus NGS data set. Poon et al. estimated the mean effective population size, N_e, for H3N2 to be 248 virions (mean standard deviation range of 45 to 457 virions), again using a combined set of variants that were present at frequencies of ≥1% and that were shared by 6 identified household donor-recipient pairs (a total of 81 variants). Our analysis, considering each of the 7 identified H3N2 transmission pairs separately, inferred MLE bottleneck sizes ranging from 107 to 370 virions across the transmission pairs, using a total of 621 variants (Fig. 4B). Again, as expected, the N_b sizes inferred by the approximate beta-binomial sampling method did not differ significantly from those inferred by using the exact beta-binomial sampling method. We again fit a negative binomial distribution to all of the variants across the transmission pairs and estimated MLE parameters of an r value of 9 and a p value of 0.966, resulting in a mean H3N2 transmission bottleneck size, N_b, of 256 virions and a 95% range of 131 to 413 virions (Fig. 4B). We again observed that the overall bottleneck size estimate for H3N2 was consistent with the estimate by Poon et al., although the bottleneck size estimates varied considerably between transmission pairs.

Overall influenza A virus transmission bottleneck sizes. We next sought to determine whether influenza A/H1N1p and influenza A/H3N2 virus subtypes statistically differed from one another in bottleneck sizes. We found that the H1N1p and H3N2 distributions of transmission bottleneck size MLEs did not differ significantly from one another ($P = 0.15$ using the Kolmogorov-Smirnov test). Given this finding, we fit a negative binomial distribution to all of the variants across the data sets of both subtypes, arriving at MLE parameters of an r value of 4 and a p value of 0.980 for the parameters of the negative binomial distribution. These parameters correspond to a

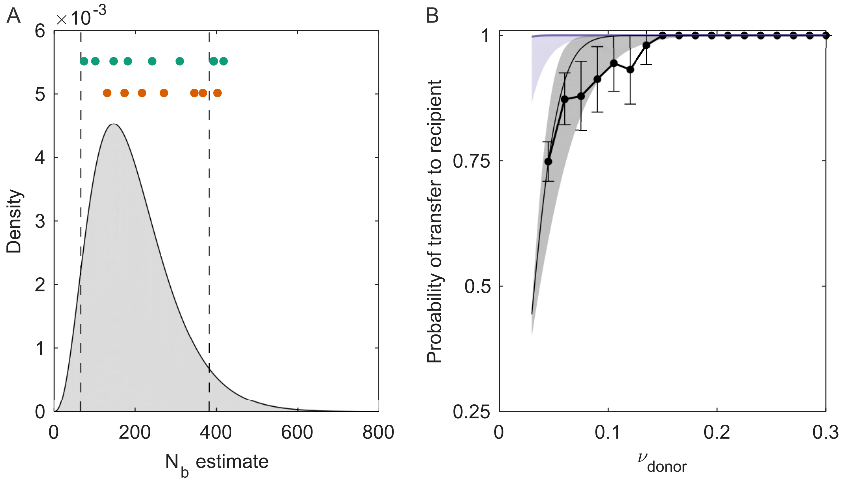

FIG 5 Overall influenza A virus bottleneck size estimates and probabilities of variant transfer under these estimates. (A) The negative binomial probability density function (pdf) describing overall transmission bottleneck sizes across H1N1p and H3N2 viral subtypes, parameterized with MLE values of an *r* value of 4 and a *p* value of 0.980. Vertical black lines show the 95% range of this distribution. The MLE bottleneck size estimates for the H3N2 (orange) and H1N1 (green) transmission pairs are shown above the pdf. (B) Probability of a donor-identified variant being either transferred or identified (called) in the recipient host as a function of donor variant frequency. Probabilities of a donor variant being present in a recipient host are shown in purple, given bottleneck size estimates provided by the negative binomial distribution shown in panel A. Probabilities of donor-identified variants being called present in a recipient host, given these same bottleneck size estimates and the assumptions of the beta-binomial sampling models, are shown in gray. The empirical probabilities of donor-identified variants being called in a recipient, as calculated from the combined H1N1p and H3N2 data sets over 3% frequency bins, are shown in black.

mean bottleneck size, N_b, of 196 virions and a 95% range of 66 to 382 virions (Fig. 4A and B). We show the probability density function for this negative binomial distribution in Fig. 5A. We further plot the expected probability of variant transfer for this bottleneck size estimate (Fig. 5B), similar to what we show for the simulated data set in Fig. 3A. Finally, we plot the probability of observed variant transfer under this N_b estimate, under the assumptions of the beta-binomial sampling model. The agreement between the probability of observed variant transfer and the empirical data indicates that variant calling thresholds again make it appear that variant transfer from donor to recipient is much less likely than it is, given bottleneck size estimates based on variant frequencies.

Relationship between donor temperature and estimated bottleneck size. Given the extent of variation in bottleneck size estimates across transmission pairs, we next considered whether certain characteristics of the donor may account for some of the observed variation. Available metadata for donor individuals included demographic data (age and gender), 2009 vaccination status, oseltamivir treatment, temperature measurements, and symptom scores (available at http://web.hku.hk/~bcowling/influenza/HK_H1N1_study.htm). Symptom scores were calculated as the number of symptoms present at the time of measurement, with considered symptoms being headache, sore throat, cough, myalgia, runny nose, and phlegm. As possible explanatory variables, we limited our analysis to temperature measurements and symptom scores. This is because, with the exception of antiviral treatment, a clear hypothesis relating any of these metadata variables to inferred transmission bottleneck sizes was lacking. We did not consider antiviral treatment as a possible explanatory variable because the time at which the antiviral was administered relative to the time of transmission was unknown.

We determined the relationship between inferred bottleneck sizes and both donor temperature and symptoms using multiple-linear-regression analysis. Specifically, we used the maximum donor temperature and the maximum symptom score as predictors of the inferred bottleneck size. The results of this regression indicated that donor symptoms were not a significant predictor of inferred transmission bottleneck sizes

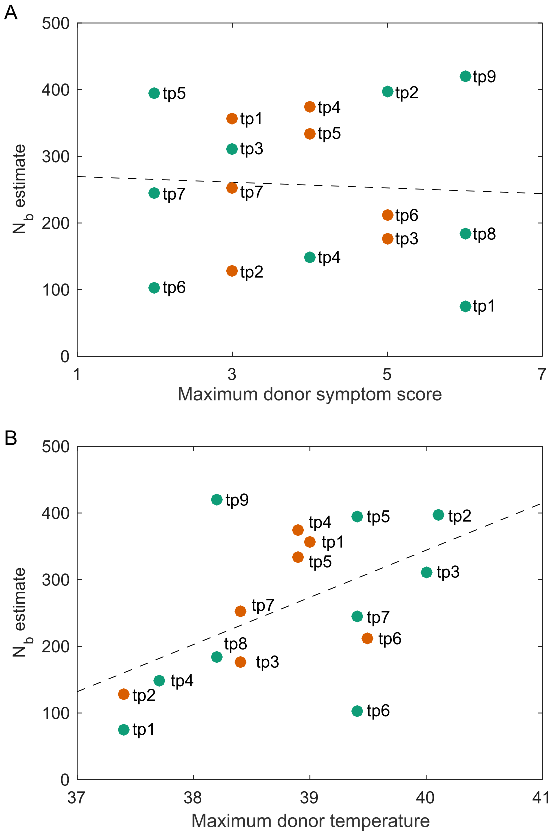

FIG 6 Relationships between transmission bottleneck size estimates and characteristics of the donor host. (A) Relationship between inferred transmission bottleneck sizes and the donor's maximum symptom score. (B) Relationship between inferred transmission bottleneck sizes and the donor's maximum temperature. In panels A and B, points are labeled by transmission pair, with green points denoting H1N1p transmission pairs and orange points denoting H3N2 transmission pairs. Dashed lines show marginal linear regressions calculated from the multiple linear regression using both maximum donor temperature and maximum donor symptom score as predictor variables. Maximum donor temperature was found to be a significantly positive predictor of N_b ($r = 0.79$; $P = 0.035$), while maximum donor symptoms were not predictive of N_b ($r = 0.13$; $P = 0.52$).

($P = 0.52$) (Fig. 6A). However, donor temperature was found to be a significant predictor of inferred bottleneck sizes ($P = 0.035$) (Fig. 6B) at a significance level of an α value of 0.05, with higher donor temperatures being positively associated with larger transmission bottleneck sizes. Using a highly conservative significance level of an α value of 0.025, determined by applying the Bonferroni correction for multiple comparisons, this P value falls slightly above the level of significance. We note, however, that the Bonferroni correction applied to results of multiple-linear-regression analyses has been shown to be overly conservative (37).

DISCUSSION

Here, we have introduced a new method for estimating the transmission bottleneck size of pathogens from next-generation sequencing data from donor-recipient pairs. We have further analyzed how well this beta-binomial sampling method performs in comparison to two existing methods in the literature: the presence/absence method and the binomial sampling method. Using a simulated data set, we have demonstrated that both the presence/absence method and the binomial sampling method (for different reasons) systematically underestimate the transmission bottleneck size and that the latter can lead to undesirable rugged likelihood curves. In contrast, the beta-binomial sampling method, as expected, is able to recover the simulated bottleneck size (Fig. 2B) and is able to accurately predict the probability that a donor variant would be identified in a recipient host under a given variant calling threshold (Fig. 3A).

Application of the beta-binomial sampling method to a previously reported H1N1p and H3N2 NGS data set showed a high degree of heterogeneity between bottleneck size estimates across transmission pairs (Fig. 4). A negative binomial distribution was fit to all of the variants, yielding an overall mean N_b of 196 virions and a 95% range of 66 to 382 virions (Fig. 4A and B and 5A).

The bottleneck sizes that we estimated for the H1N1p and H3N2 transmission pairs are close to previous estimates of the effective population size, N_e, arrived at by Poon et al. for this data set (28), although we were able to further estimate transmission bottleneck sizes by transmission pair, and our method was able to make use of a much larger number of identified variants. Our bottleneck size estimates are consistent with the more qualitative observations of loose transmission bottlenecks for influenza A virus transmission in horses (20, 22, 23), pigs (20, 21), and dogs (19). Our N_b estimates, however, are considerably larger than the previous bottleneck sizes estimated for this virus by Varble et al. (27), Frise et al. (29), and McCaw et al. (17). The experimental study by Varble et al. showed that the route of transmission affected the bottleneck size, with contact transmission giving rise to larger bottlenecks. Those researchers found that, of the 71 to 100 distinct viral tags, only 7 to 24 of these tagged viruses were detected in the recipients following infection via direct contact (27). The number of distinct viral tags, however, might reflect the lower limit of the bottleneck size because it is possible that more than one virion passing through the bottleneck would have the same tag. Frise et al. reported a mean bottleneck size of 28.2 infectious genomes for contact transmission of an efficiently transmitted H1N1 strain in ferrets, although they were unable to identify an upper limit to the bottleneck size confidence interval (29). Both of those estimates are much larger than the earlier estimate of 3.8 virions by McCaw et al. for contact transmission of H1N1 in ferrets (17). While there are other studies that have estimated the transmission bottleneck size in the context of viral adaptation to a new host species (24–26), comparisons with those studies are inappropriate because these bottlenecks are subject to strong selective forces, which considerably narrow the transmission bottleneck size (38).

The N_b estimates for influenza virus transmission in the data set described here, in both our study and the original analysis by Poon et al., are considerably higher than previous quantitative estimates of the bottleneck size for contact transmission of IAV (17, 27, 29). Notably, those previous estimates of N_b were arrived at by using data from experimental ferret infections. With a recent analysis showing that secondary attack rates in ferret studies are considerably higher than human secondary attack rates, controlling for infecting subtype (39), one possibility for these discrepancies is that ferrets and other small mammals may require fewer influenza A virions to successfully initiate infection.

In particular, the bottleneck size estimate by McCaw et al. was significantly lower than our N_b estimates for contact transmission of influenza virus (17). One possible explanation for the low N_b estimate is that the "competitive-mixture" method that those authors used to calculate bottleneck size considers only two viral populations, analogous to the estimates derived from a single variant in the methods that we considered. The competitive-mixture method is thus highly susceptible to fluctuations between donor and recipient variant frequencies arising from stochastic viral dynamics in the recipient. Thus, for the same reason that the binomial sampling method that we describe here underestimates bottleneck sizes, we would expect this competitive-mixture method to considerably underestimate bottleneck sizes. However, this method is free of one of the necessary assumptions made for each the three methods that we considered, namely, that the variants considered are independent. The independence assumption is clearly violated in this data set given the extensive genetic linkage within influenza virus gene segments (40). We can, however, somewhat control for the effects of linkage by selecting only one variant per gene segment. This data-thinning approach still assumes independence across gene segments that, while not ideal, may be supported by recent experimental evidence showing high levels of reassortment *in* *vitro* (41). If intrahost reassortment occurs at similar rates *in vivo*, then sampling of only one variant per gene segment should remove much of the bias due to linkage.

The methods that we considered make other assumptions that may also have impacted transmission bottleneck size estimates. These assumptions include that (i) donor-identified variants did not originate *de novo* in any recipient hosts, (ii) variants were biallelic, and (iii) variants were selectively neutral. Significant levels of *de novo* evolution of variants in recipient hosts would artificially increase estimated bottleneck sizes. Therefore, these methods may not be appropriate for pathogens causing chronic infections, such as HIV, where sampling of the recipient host can occur years after the initiation of infection. However, we do not expect substantial *de novo* evolution of variants to occur over the course of an acute influenza virus infection based on recent findings (38) and the observation that the vast majority of recipient-identified variants were also present in the donor. Therefore, we do not expect this assumption to have significantly influenced our bottleneck size estimates for influenza virus.

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

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

In addition to confirming the large transmission bottleneck size for IAV in this data set, we have shown that estimated bottleneck sizes vary considerably across transmission pairs. This observation is in agreement with data from previous influenza virus transmission studies in ferrets. Those studies showed that the bottleneck stringency for IAV is greatly influenced by the route of transmission, with contact transmission being much looser than respiratory/airborne transmission (27, 29). Our method's ability to infer bottleneck sizes of individual transmission pairs means that such analyses could potentially distinguish the route of transmission. Moreover, our analysis identified an association between the severity of infection of the donor, as measured by temperature, and the size of the transmission bottleneck, where more severe infections were associated with larger bottlenecks. This finding is intriguing, given that a previous study showed a positive relationship between host temperature and viral load during early infection (42), and another study showed a positive relationship between host temperature and nasal shedding (43). Our finding suggests that donor viral load and/or nasal shedding levels may impact transmission bottleneck sizes. Our finding that donor symptom scores do not explain any variation in bottleneck size estimates across transmission pairs is perhaps not surprising, given that some of the symptoms included in the score (e.g., headache) are unlikely to contribute to donor infectiousness.

In this study, we have developed a new statistical approach that can be used to accurately infer transmission bottleneck sizes for acute viral infections, such as influenza virus, respiratory syncytial virus (RSV), and norovirus, using NGS data from identified donor-recipient pairs. This beta-binomial sampling method accounts for the possibility of false-negative variants that are not called as present due to necessary variant calling thresholds. This method further accounts for changes in variant frequencies between the time of infection of the recipient and the time of pathogen sampling from the recipient that arise due to stochastic replication dynamics early in infection. Given the importance of the transmission bottleneck size in regulating the rate of pathogen evolution at the level of the host population, estimation of the transmission bottleneck size is a necessary component in the analysis of pathogens important to public health.

Although methods such as viral tagging to estimate the bottleneck size for experimental infections exist, these techniques are not applicable to natural infections. Hence, this work provides a strong foundation for future estimations of bottleneck sizes from viral sequence data that, importantly, can be applied to clinical samples.

MATERIALS AND METHODS

Development of the beta-binomial sampling method. Here, we derive the beta-binomial sampling method for inferring transmission bottleneck sizes from pathogen NGS data. The final likelihood expressions for this method are provided in equations 3 and 4. As described above, this method allows variant frequencies in the recipient host to change between infection and sampling (Fig. 1) due to stochastic pathogen dynamics occurring during the process of replication. More concretely, early in infection, when there are only a small number of replicating virions, stochasticity in viral growth is expected to have a large effect. For a stochastic birth-death process with a constant birth rate, λ, and a constant death rate, μ, the probability mass function for the viral population size originating from a single virion that successfully establishes infection (44) is given by

$$P\,(N_k\,(t) \;=\; k) \;=\; (1 \;-\; \eta_t)\,\eta_t^{\,k\,-\,1},\; k \;\geq\; 1 \tag{8}$$

where t is the time of sampling and $\eta_t \;=\; \dfrac{\lambda\,(e^{(\lambda\,-\,\mu)^{t}\,-1})}{\lambda e^{(\lambda-\mu)^{t}} - \,\mu}$. For the bursty replication that characterizes many viruses, equation 8 is still approximately true at long times with an adjusted value of η_t.

The population sizes stemming from each of the N_b founding virions, contingent on their successful establishment, are thus geometrically distributed random variables. As these population sizes are likely to be very large at the time of sampling, we can approximate them as being exponentially distributed random variables. Under this approximation, the distribution of the fractions of the population that descend from each of the founding virions is Dirichlet(1,1,. . .1), with N_b 1's, one for each ancestor. A subset, k, of these founder virions carries the variant allele; the remaining subset of these founder virions ($N_b - k$) carries the reference allele. Collapsing the Dirichlet distribution yields that the fraction of the population carrying the variant allele is distributed as Beta(k, $N_b - k$). Remarkably, this fraction does not depend on the within-host viral birth rate, λ; the death rate, μ; the time of sampling, t; or the burstiness of replication. To obtain the overall likelihood of population bottleneck size, N_b, we simply have to consider all possible scenarios of how many virions out of the total N_b virions transferred carried the variant allele. Under the assumption that the founding pathogen population is randomly sampled from the pathogen population of the donor host, the probability that the founding population of N_b virions carries k variant alleles is given by the binomial distribution $p\_bin\,(k|N_b,\,\nu_{D,i}) \equiv \mathrm{Pr}\,(X \;=\; k|N_b,\,\nu_{D,i}) \;=\; \binom{N_b}{k}\,(\nu_{D,i})^{k}\,(1 \;-\; \nu_{D,i})^{N_b\,-\,k}$, where the number of trials is given by N_b and the success probability is given by $\nu_{D,i}$, the frequency of variant i in the donor. Thus, the overall likelihood of population bottleneck size, N_b, for variant i is given by equation 1, where $\nu_{R,i}$ is the frequency of variant i in the recipient and the term $p\_beta(\nu_{R,i}|k,\, N_b \,-\, k)$ is given by the beta probability density function, evaluated at $\nu_{R,i}$.

Accommodating sampling noise arising from a finite number of reads is simple, leading to minor modifications to the above-described equation (equation 1), resulting in equation 3, where $R_{var,i}$ is the number of reads of the variant allele in the recipient sample at site i and R_{tot} is the total number of reads at that site. The term $p\_betabin(R_{var,i}|R_{tot,i},\, k,\, N_b{-}k)$ is given by the beta-binomial distribution evaluated at $R_{var,i}$ and parameterized with $R_{tot,i}$ as the number of trials and parameters k and N_b. Equation 3 thus incorporates noise both from the sampling process itself and from the process of stochastic pathogen growth. The overall likelihood of bottleneck size N_b for a transmission pair is simply the product of the site-specific likelihoods.

As mentioned above, we expect that variant calling thresholds will impact the likelihood calculations used in the bottleneck size estimation. These thresholds will force some variant alleles in the recipient viral population to be called absent when they are actually present at frequencies below the value of the chosen threshold. Since a true absence of a variant allele is more likely at smaller bottleneck sizes, conservative variant calling thresholds will bias N_b estimates toward lower values. Simply excluding variants that are called absent from the analysis, however, will also bias bottleneck size estimates, this time toward higher values. To get around this, we do not recommend simply lowering the variant calling threshold because NGS sequencing errors can also give rise to false positives, thereby inappropriately inflating bottleneck size estimates. Instead, we recommend accommodating below-threshold variants in the following way. For a donor-identified variant, i, that is called absent in the recipient (whether truly absent or just called absent), the likelihood of the transmission bottleneck size is given by equation 2, where T is the variant calling threshold (e.g., of 3%) and $p\_beta\_cdf(\nu_{R,i} < T|k,\, N_b - k)$ is given by the beta cumulative distribution function evaluated at the variant calling threshold. We can again incorporate the effects of sampling noise by considering the number of reads at the variant site with equation 4, where, in this case, $p\_betabin\_cdf(R_{var,i} \,<\, TR_{tot,i}|R_{tot,i},\; k,\; N_b{-}k)$ is given by the beta-binomial cumulative distribution function evaluated at the number of reads that would qualify as falling at the variant calling threshold.

Once the transmission bottleneck sizes have been estimated by using the beta-binomial sampling method, the probability of the true presence/absence of a variant in the recipient host can be determined for any given donor variant frequency. Similarly, the probability that a variant is called present/absent can be determined for any given donor frequency, $\nu_{D,i}$, given a sufficiently high read count in the recipient host. Given a high read count, the probability that a variant is called present in the recipient is given by $\Sigma_{k=0}^{N_b}[1 - p\_beta\_cdf\,(v_{R,i} < T|k, N_b - k)]\,p\_bin\,(k|N_b, v_{D,i})$.

The binomial sampling method. In contrast to the beta-binomial sampling method, the binomial sampling method implicitly assumes that the infecting virus population is subject to deterministic dynamics between the time of infection and the time at which the recipient virus is sampled and, thus, that the sampled pathogen population in the recipient perfectly reflects the founding pathogen population under the common assumption of selective neutrality. The founding pathogen population is, as in the beta-binomial sampling method, assumed to be randomly sampled from the pathogen population of the donor host. The site-specific likelihood of the transmission bottleneck size, N_b, is therefore given by equation 6, where $p\_bin\,(R_{var,i}|R_{tot,i}, f_k) = \binom{R_{tot,i}}{R_{var,i}}\left(\frac{k}{N_b}\right)^{R_{var,i}}\left(1 - \frac{k}{N_b}\right)^{R_{tot,i} - R_{var,i}}$. The overall likelihood of the transmission bottleneck size, N_b, is calculated by multiplying across all site-specific likelihoods.

The above-described expression incorporates sampling noise, which is important when only a small number of reads are available. With an increasing number of reads, the sampling noise necessarily goes down, making $p\_bin(R_{var,i}|R_{tot,i}, k/N_b) \approx 0$ in cases where $R_{var,i}/R_{tot,i} \neq k/N_b$. This will result in dramatic differences in likelihood values between small values of N_b and, more generally, multimodal likelihood curves that are very sensitive to specific variant frequencies in the recipient host.

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

The bottleneck size estimates inferred with the binomial sampling method are again subject to the effects of false-negative variant calls. We can modify the binomial sampling method to incorporate the variant call threshold in a way similar to how the threshold frequency was incorporated into the beta-binomial sampling method. For a donor-identified variant, *i*, that is called absent in the recipient (whether truly absent or just called absent), the likelihood of the transmission bottleneck size is explained by equation 7. The probability that the number of variant reads falls below the level required for the variant to be called present is given by the binomial cumulative distribution function $p\_bin\_cdf\left(R_{var,i} < \lfloor TR_{tot,i}\rfloor \middle| R_{tot,i}, \frac{k}{N_b}\right) = \Sigma_{j=0}^{\lfloor TR_{tot,i}\rfloor}\binom{R_{tot,i}}{j}\left(\frac{k}{N_b}\right)^{j}\left(1 - \frac{k}{N_b}\right)^{R_{tot,i}-j}$, where $\lfloor TR_{tot,i}\rfloor$ is the largest integer smaller than $TR_{tot,i}$.

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

Simulated deep-sequencing data. To illustrate the use of the methods used to estimate N_b, we generated a mock deep-sequencing data set via simulation. For this data set, we assumed a single donor-recipient pair, with 500 independent donor-identified variants. Independently for both the donor and the recipient, we drew the total number of reads at each of the 500 sites from a normal distribution with a mean of 500 reads and a standard deviation of 100 reads. Draws from the normal distribution were rounded to the nearest integer, and those that fell at 0 or below were discarded. For the donor, we then first determined "true" variant frequencies at each of these sites by drawing from an exponential distribution with a mean frequency of 0.08. Variants with observed frequencies below the variant calling threshold of 0.03 or above 0.50 were discarded. To determine the number of variant reads at a given site in the donor, we drew from a binomial distribution with the number of trials being the total read count at that site in the donor and the probability of success being given by that site's true variant frequency in the donor. We then incorporated sequencing error by again using draws from binomial distributions. Specifically, we determined the number of true reference reads in the donor that were misclassified as variant reads and the number of true variant reads in the donor that were correctly classified as variant reads, based on an assumed sequencing error rate of 1%. The total number of observed variant reads at a given site in a donor was then calculated as the sum of the misclassified reference reads and the correctly classified variant reads. Observed variant frequencies in the donor were then calculated by dividing the number of observed variant reads by the total number of observed reads at each site. In this manner, we simulated 500 variants, with observed frequencies in the range of 3 to 50%. The lower bound value of 3% was our assumed variant calling threshold; the upper bound value of 50% coincided with a variant allele always being the minority allele.

For the recipient, we simulated the total number of variant reads at each site by first simply determining, at each site, the number of virions in the founding population that carried the variant allele, under the assumption of a transmission bottleneck size, N_b, of 50. This was done, at each site, by drawing from a binomial distribution with the number of trials being N_b and the probability of success being the true variant frequency at that site in the donor. For the simulated data set, we first determined the true fraction of the viral population carrying the variant allele at the time of sampling by drawing from a beta distribution with the shape parameter being the number of variant alleles in the founder population and the scale parameter being the difference between the founding population size of N_b and the number of variant alleles in the founder population. The true number of variant reads was then determined by drawing from a binomial distribution with the number of trials being the total number of reads at that site and the probability of success being the fraction of the population at the time of sampling that carried the variant allele. We then obtained the total number of variant reads at a given site in a recipient by introducing sequencing error to the true number of variant reads and the true number of reference reads.

Application to influenza A virus deep-sequencing data. We applied the three methods for bottleneck size inference described in the introduction to influenza A virus deep-sequencing data examined previously (28). In that study, Poon and colleagues identified donor-recipient transmission pairs based on household information and the genetic similarities between the viral populations in infected hosts. We base our analyses on these previously identified transmission pairs. In some cases, there were several members of a household who became infected. In this subset of cases, rather than considering all feasible pairwise combinations of who infected whom, we assumed that the index case transmitted the infection to the remaining household members. With this assumption, the 9 identified transmission pairs for influenza A virus subtype H1N1p were 681_V1(0) → 681_V3(2), 684_V1(0) → 684_V2(3), 712_V1(0) → 712_V1(4), 742_V1(0) → 742_V3(3), 751_V1(0) → 751_V3(1), 751_V1(0) → 751_V2(3), 751_V1(0) → 751_V2(4), 779_V1(0) → 779_V2(1), and 779_V1(0) → 779_V1(2), where *X*_V*Y*(*Z*) refers to household *X*, visit *Y*, and subject *Z* and the arrow demarcates transmission from the donor to the recipient. The 7 identified transmission pairs for influenza A virus subtype H3N2 were 689_V1(0) → 689_V2(2), 720_V1(0) → 720_V2(1), 734_V1(0) → 734_V3(2), 739_V1(0) → 739_V2(2), 739_V1(0) → 739_V2(3), 747_V1(0) → 747_V2(2), and 763_V1(0) → 763_V2(3). The deep-sequencing data are publically available (28) (see https://www.synapse.org/#!Synapse:syn8033988). We called variants and determined variant frequencies from these data using VarScan (45, 46), using a variant calling threshold of 3%, a mean quality score of 20, and a *P* value of 0.05. We provide variants and their frequencies used in this study in Data Set S1 in the supplemental material.

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

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/JVI.00171-17.

SUPPLEMENTAL FILE 1, XLSX file, 0.1 MB.

ACKNOWLEDGMENTS

This work was funded by MIDAS CIDID U54-GM111274, supporting K.K. and A.S.L. A.S.L. was further supported by Duke Medical Scientists Training Program grant T32 GM007171. E.G. was supported by grant U01 AI111598.

REFERENCES

1. Gutiérrez S, Michalakis Y, Blanc S. 2012. Virus population bottlenecks during within-host progression and host-to-host transmission. Curr Opin Virol 2:546–555. https://doi.org/10.1016/j.coviro.2012.08.001.
2. Geoghegan JL, Senior AM, Holmes EC. 2016. Pathogen population bottlenecks and adaptive landscapes: overcoming the barriers to disease emergence. Proc Biol Sci 283:20160727. https://doi.org/10.1098/rspb.2016.0727.
3. Skums P, Bunimovich L, Khudyakov Y. 2015. Antigenic cooperation among intrahost HCV variants organized into a complex network of cross-immunoreactivity. Proc Natl Acad Sci U S A 112:6653–6658. https://doi.org/10.1073/pnas.1422942112.
4. Xue KS, Hooper KA, Ollodart AR, Dingens AS, Bloom JD. 2016. Cooperation between distinct viral variants promotes growth of H3N2 influenza in cell culture. eLife 5:e13974. https://doi.org/10.7554/eLife.13974.
5. Brooke CB, Ince WL, Wrammert J, Ahmed R, Wilson PC, Bennink JR, Yewdell JW. 2013. Most influenza A virions fail to express at least one essential viral protein. J Virol 87:3155–3162. https://doi.org/10.1128/JVI.02284-12.
6. Worby CJ, Lipsitch M, Hanage WP. 2014. Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. PLoS Comput Biol 10:e1003549. https://doi.org/10.1371/journal.pcbi.1003549.
7. De Maio N, Wu C-H, Wilson DJ. 2016. SCOTTI: efficient reconstruction of transmission within outbreaks with the structured coalescent. PLoS Comput Biol 12:e1005130. https://doi.org/10.1371/journal.pcbi.1005130.
8. Hall JS, French R, Hein GL, Morris TJ, Stenger DC. 2001. Three distinct mechanisms facilitate genetic isolation of sympatric wheat streak mosaic virus lineages. Virology 282:230–236. https://doi.org/10.1006/viro.2001.0841.
9. Sacristán S, Malpica JM, Fraile A, García-Arenal F. 2003. Estimation of

population bottlenecks during systemic movement of tobacco mosaic virus in tobacco plants. J Virol 77:9906–9911. https://doi.org/10.1128/JVI.77.18.9906-9911.2003.

10. Moury B, Fabre F, Senoussi R. 2007. Estimation of the number of virus particles transmitted by an insect vector. Proc Natl Acad Sci U S A 104:17891–17896. https://doi.org/10.1073/pnas.0702739104.
11. Betancourt M, Fereres A, Fraile A, García-Arenal F. 2008. Estimation of the effective number of founders that initiate an infection after aphid transmission of a multipartite plant virus. J Virol 82:12416–12421. https://doi.org/10.1128/JVI.01542-08.
12. Zwart MP, Daròs JA, Elena SF. 2011. One is enough: in vivo effective population size is dose-dependent for a plant RNA virus. PLoS Pathog 7:e1002122. https://doi.org/10.1371/journal.ppat.1002122.
13. Fabre F, Moury B, Johansen EI, Simon V, Jacquemond M, Senoussi R. 2014. Narrow bottlenecks affect pea seedborne mosaic virus populations during vertical seed transmission but not during leaf colonization. PLoS Pathog 10:e1003833. https://doi.org/10.1371/journal.ppat.1003833.
14. Smith DR, Adams AP, Kenney JL, Wang E, Weaver SC. 2008. Venezuelan equine encephalitis virus in the mosquito vector Aedes taeniorhynchus: infection initiated by a small number of susceptible epithelial cells and a population bottleneck. Virology 372:176–186. https://doi.org/10.1016/j.virol.2007.10.011.
15. Zwart MP, Hemerik L, Cory JS, de Visser JAGM, Bianchi FJJA, Van Oers MM, Vlak JM, Hoekstra RF, Van der Werf W. 2009. An experimental test of the independent action hypothesis in virus-insect pathosystems. Proc Biol Sci 276:2233–2242. https://doi.org/10.1098/rspb.2009.0064.
16. van der Werf W, Hemerik L, Vlak JM, Zwart MP. 2011. Heterogeneous host susceptibility enhances prevalence of mixed-genotype microparasite infections. PLoS Comput Biol 7:e1002097. https://doi.org/10.1371/journal.pcbi.1002097.
17. McCaw JM, Arinaminpathy N, Hurt AC, McVernon J, McLean AR. 2011. A mathematical framework for estimating pathogen transmission fitness and inoculum size using data from a competitive mixtures animal model. PLoS Comput Biol 7:e1002026. https://doi.org/10.1371/journal.pcbi.1002026.
18. Forrester NL, Guerbois M, Seymour RL, Spratt H, Weaver SC. 2012. Vector-borne transmission imposes a severe bottleneck on an RNA virus population. PLoS Pathog 8:e1002897. https://doi.org/10.1371/journal.ppat.1002897.
19. Hoelzer K, Murcia PR, Baillie GJ, Wood JLN, Metzger SM, Osterrieder N, Dubovi EJ, Holmes EC, Parrish CR. 2010. Intrahost evolutionary dynamics of canine influenza virus in naive and partially immune dogs. J Virol 84:5329–5335. https://doi.org/10.1128/JVI.02469-09.
20. Stack JC, Murcia PR, Grenfell BT, Wood JLN, Holmes EC. 2013. Inferring the inter-host transmission of influenza A virus using patterns of intra-host genetic variation. Proc Biol Sci 280:20122173. https://doi.org/10.1098/rspb.2012.2173.
21. Murcia PR, Hughes J, Battista P, Lloyd L, Baillie GJ, Ramirez-Gonzalez RH, Ormond D, Oliver K, Elton D, Mumford JA, Caccamo M, Kellam P, Grenfell BT, Holmes EC, Wood JLN. 2012. Evolution of an Eurasian avian-like influenza virus in naive and vaccinated pigs. PLoS Pathog 8:e1002730. https://doi.org/10.1371/journal.ppat.1002730.
22. Hughes J, Allen RC, Baguelin M, Hampson K, Baillie GJ, Elton D, Newton JR, Kellam P, Wood JLN, Holmes EC, Murcia PR. 2012. Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. PLoS Pathog 8:e1003081. https://doi.org/10.1371/journal.ppat.1003081.
23. Murcia PR, Baillie GJ, Daly J, Elton D, Jervis C, Mumford JA, Newton R, Parrish CR, Hoelzer K, Dougan G, Parkhill J, Lennard N, Ormond D, Moule S, Whitwham A, McCauley JW, McKinley TJ, Holmes EC, Grenfell BT, Wood JLN, Parrish R, Hoelzer K, Dougan G, Parkhill J, Lennard N, Ormond D, Moule S, McCauley JW, McKinley TJ, Holmes EC, Grenfell BT, Wood JLN, Parrish CR, Whitwham A. 2010. Intra- and interhost evolutionary dynamics of equine influenza virus. J Virol 84:6943–6954. https://doi.org/10.1128/JVI.00112-10.
24. Wilker PR, Dinis JM, Starrett G, Imai M, Hatta M, Nelson CW, O'Connor DH, Hughes AL, Neumann G, Kawaoka Y, Friedrich TC. 2013. Selection on hemagglutinin imposes a bottleneck during mammalian transmission of reassortant H5N1 influenza viruses. Nat Commun 4:2636. https://doi.org/10.1038/ncomms3636.
25. Zaraket H, Baranovich T, Kaplan BS, Carter R, Song M-S, Paulson JC, Rehg JE, Bahl J, Crumpton JC, Seiler J, Edmonson M, Wu G, Karlsson E, Fabrizio T, Zhu H, Guan Y, Husain M, Schultz-Cherry S, Krauss S, McBride R, Webster RG, Govorkova EA, Zhang J, Russell CJ, Webby RJ. 2015. Mammalian adaptation of influenza A(H7N9) virus is limited by a narrow genetic bottleneck. Nat Commun 6:6553. https://doi.org/10.1038/ncomms7553.
26. Moncla LH, Zhong G, Nelson CW, Dinis JM, Mutschler J, Hughes AL, Watanabe T, Kawaoka Y, Friedrich TC. 2016. Selective bottlenecks shape evolutionary pathways taken during mammalian adaptation of a 1918-like avian influenza virus. Cell Host Microbe 19:169–180. https://doi.org/10.1016/j.chom.2016.01.011.
27. Varble A, Albrecht RAA, Backes S, Crumiller M, Bouvier NMM, Sachs D, García-Sastre A, TenOever BRR. 2014. Influenza A virus transmission bottlenecks are defined by infection route and recipient host. Cell Host Microbe 16:691–700. https://doi.org/10.1016/j.chom.2014.09.020.
28. Poon LLM, Song T, Rosenfeld R, Lin X, Rogers MB, Zhou B, Sebra R, Halpin RA, Guan Y, Twaddle A, DePasse JV, Stockwell TB, Wentworth DE, Holmes EC, Greenbaum B, Peiris JSM, Cowling BJ, Ghedin E. 2016. Quantifying influenza virus diversity and transmission in humans. Nat Genet 48:195–200. https://doi.org/10.1038/ng.3479.
29. Frise R, Bradley K, van Doremalen N, Galiano M, Elderfield RA, Stilwell P, Ashcroft JW, Fernandez-Alonso M, Miah S, Lackenby A, Roberts KL, Donnelly CA, Barclay WS. 2016. Contact transmission of influenza virus between ferrets imposes a looser bottleneck than respiratory droplet transmission allowing propagation of antiviral resistance. Sci Rep 6:29793. https://doi.org/10.1038/srep29793.
30. Emmett KJ, Lee A, Khiabanian H, Rabadan R. 9 February 2015. High-resolution genomic surveillance of 2014 ebolavirus using shared subclonal variants. PLoS Curr https://doi.org/10.1371/currents.outbreaks.c7fd7946ba606c982668a96bcba43c90.
31. Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, Keim P, Morrow JB, Salit ML, Zook JM. 2015. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. Front Genet 6:235. https://doi.org/10.3389/fgene.2015.00235.
32. Ghedin E, Holmes EC, DePasse JV, Pinilla LT, Fitch A, Hamelin M-E, Papenburg J, Boivin G. 2012. Presence of oseltamivir-resistant pandemic A/H1N1 minor variants before drug therapy with subsequent selection and transmission. J Infect Dis 206:1504–1511. https://doi.org/10.1093/infdis/jis571.
33. Van den Hoecke S, Verhelst J, Vuylsteke M, Saelens X. 2015. Analysis of the genetic diversity of influenza A viruses using next-generation DNA sequencing. BMC Genomics 16:79. https://doi.org/10.1186/s12864-015-1284-z.
34. Lakdawala SS, Jayaraman A, Halpin RA, Lamirande EW, Shih AR, Stockwell TB, Lin X, Simenauer A, Hanson CT, Vogel L, Paskel M, Minai M, Moore I, Orandle M, Das SR, Wentworth DE, Sasisekharan R, Subbarao K. 2015. The soft palate is an important site of adaptation for transmissible influenza viruses. Nature 526:122–125. https://doi.org/10.1038/nature15379.
35. Dinis JM, Florek NW, Fatola OO, Moncla LH, Mutschler JP, Charlier OK, Meece JK, Belongia EA, Friedrich TC. 2016. Deep sequencing reveals potential antigenic variants at low frequencies in influenza A virus-infected humans. J Virol 90:3355–3365. https://doi.org/10.1128/JVI.03248-15.
36. Sacristán S, Díaz M, Fraile A, García-Arenal F. 2011. Contact transmission of Tobacco mosaic virus: a quantitative analysis of parameters relevant for virus evolution. J Virol 85:4974–4981. https://doi.org/10.1128/JVI.00057-11.
37. Mundfrom DJ, Perrett JJ, Schaffer J, Piccone A, Roozeboom M. 2006. Bonferroni adjustments in tests for regression coefficients. Mult Linear Regres Viewp 32:1–6.
38. Sobel Leonard A, McClain MT, Smith GJD, Wentworth DE, Halpin RA, Lin X, Ransier A, Stockwell TB, Das SR, Gilbert AS, Lambkin-Williams R, Ginsburg GS, Woods CW, Koelle K. 2016. Deep sequencing of influenza A virus from a human challenge study reveals a selective bottleneck and only limited intrahost genetic diversity. J Virol 90:11247–11258. https://doi.org/10.1128/JVI.01657-16.
39. Buhnerkempe MG, Gostic K, Park M, Ahsan P, Belser JA, Lloyd-Smith JO. 2015. Mapping influenza transmission in the ferret model to transmission in humans. eLife 4:e07969. https://doi.org/10.7554/eLife.07969.
40. Boni MF, Zhou Y, Taubenberger JK, Holmes EC. 2008. Homologous recombination is very rare or absent in human influenza A virus. J Virol 82:4807–4811. https://doi.org/10.1128/JVI.02683-07.
41. Ince WL, Gueye-Mbaye A, Bennink JR, Yewdell JW. 2013. Reassortment complements spontaneous mutation in influenza A virus NP and M1

genes to accelerate adaptation to a new host. J Virol 87:4330–4338. https://doi.org/10.1128/JVI.02749-12.

42. Lau LLH, Cowling BJ, Fang VJ, Chan K-H, Lau EHY, Lipsitch M, Cheng CKY, Houck PM, Uyeki TM, Peiris JSM, Leung GM. 2010. Viral shedding and clinical illness in naturally acquired influenza virus infections. J Infect Dis 201:1509–1516. https://doi.org/10.1086/652241.

43. Song D, Moon H, Jung K, Yeom M, Kim H, Han S, An D, Oh J, Kim J, Park B, Kang B. 2011. Association between nasal shedding and fever that influenza A (H3N2) induces in dogs. Virol J 8:1. https://doi.org/10.1186/1743-422X-8-1.

44. Kendall DG. 1948. On the generalized "birth-and-death" process. Ann Math Stat 19:1–15. https://doi.org/10.1214/aoms/1177730285.

45. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25:2283–2285. https://doi.org/10.1093/bioinformatics/btp373.

46. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22:568–576. https://doi.org/10.1101/gr.129684.111.