EMORY COLLEGE OF ARTS AND SCIENCES

Department of Biology

November 10, 2023

Dear editor,

With this letter, we are resubmitting our research article entitled “Transmission bottleneck size estimation from *de novo* viral genetic variation”. We thank the two reviewers for their thoughtful comments and suggestions as well as the handling editor for their invitation to submit a revision. We have addressed the reviewers’ comments and suggestions and believe that these edits and additions have substantially improved our manuscript. Below, please find a point-by-point response to the reviewer comments.

With best regards,

Katia Koelle

Professor
Department of Biology
Emory University

Emory University
Rollins Research Center
1510 Clifton Rd. NE, Suite 2006
Atlanta, GA 30322

Email katia.koelle@emory.edu

Associate editor's comments to the author:
Both reviewers clearly agree this paper merits publication, but suggest some clarifications, particularly about some of the assumptions made by the model. I am sure the authors can submit a revised version that will address these issues.

We thank the editor for this encouragement and for the choice of thoughtful reviewers, who we believe have provided excellent suggestions for improving our submitted work. We have revised the manuscript to respond to these suggestions and hope that this revised manuscript is now suitable for publication in *Molecular Biology and Evolution*.

Reviewer: 1
Comments to the Author
In this paper, Shi and collaborators present a statistical approach for estimating transmission bottleneck sizes from viral sequence data from donor-recipient pairs. Their approach works by computing the probability for a certain number of genetically distinct clonal lineages to establish and grow within the recipient. They test their approach in simulations, demonstrating an ability to recover true model parameters. They then apply this method to study data from Influenza A and SARS-CoV-2 transmission pairs, confirming prior reports of small transmission bottleneck sizes. The authors observe that their model provides a somewhat better fit to data for both IAV and SARS-CoV-2 when using R0 values for intrahost replication that are lower than those reported in the literature.

This paper addresses an important topic. The authors argue that some of the assumptions they have made (neutrality of mutations, infinite sites) are unlikely to substantially bias their results, and I agree with their assessment. Some aspects of the modeling are clear, but others are not explicit enough to fully describe how the model is applied to data. While the methods the authors develop are interesting, it would be helpful to make a stronger case for this approach versus other alternatives. My specific comments are below.

We thank the reviewer for their careful reading of our manuscript.

1. The mathematical model described in the supplement is clear, but it is not entirely obvious how the model has been applied to data. This was my understanding of the procedure: the authors compare deep sequencing data from a donor and recipient. They search for genetic differences between the donor and recipient sequences, where nearly all sequences in the donor at one site have a particular nucleotide, and nearly all sequences in the recipient at the same site have a different nucleotide. Here, "nearly all" is defined by the variant-calling threshold. This number of differences is then considered to be the number of clonal lineages that established in the recipient, which can be used to estimate the bottleneck size as described.
Is this description correct? If so, it would be helpful for the authors to make this point explicit. Model fitting is clear when the number of clonal lineages is defined, as in simulations, but how the clonal lineages are determined in actual data was unclear.

Yes, the reviewer's description is correct. We have added a short subsection to the Results section entitled 'Application to empirical data' that explicitly describes this process of calling clonal variants. This subsection is placed immediately prior to the first empirical application (influenza A virus) and is highlighted in blue in the revised text.

2. If the procedure above is correct, then I am confused about what would happen in a situation where more than one viral lineage establishes within a recipient. Assuming that each lineage has different mutations, and that no one lineage fixes, it seems that none of these lineages would be counted because their respective variants would be polymorphic in the virus population.

Yes, the reviewer is correct. In a given donor-recipient pair, if the number of viral lineages that establish is two or more (that is, the transmission bottleneck size N_b is 2 or more), there cannot be any clonal mutations that are observed. This situation is depicted in both Figure 1B and Figure 1D. When a single viral lineage establishes (that is, the transmission bottleneck size N_b is 1), there may be clonal mutations that are observed (e.g., Figure 1C, when a wild-type lineage gives rise to a mutant lineage and goes extinct and all other wild-type lineages go extinct) or there might not be (e.g., when a single wild-type viral lineage establishes). We have edited the text in the subsection titled 'The stochastic within-host model' to clarify these outcomes. Please see the text highlighted in blue in this section.

3. Related to the point above, it seems that it could happen that the same variant is observed in both the donor and recipient, but it could be counted as a new clonal lineage if the variant frequency in the donor was just below the variant-calling threshold. Based on Figure S2, it seems that this might happen reasonably often for SARS-CoV-2? Depending on the variant-calling threshold and the mutation rate, it could be much more likely that a low-frequency variant was transmitted rather than it arising de novo in the recipient, but this appears not to be accounted for in the model.

We count a site as harboring a clonal mutation if the site is monomorphic in both the donor and in the recipient, but with different alleles. The reviewer is therefore correct in that there is a possibility that a variant is present in the donor at a frequency that falls below the variant-calling threshold and that it fixes in the recipient following transmission, and that we thus incorrectly call it a clonal mutation that arose *de novo* in the recipient. We now elaborate in detail on this possibility and steps that can be taken to guard against the erroneous calling of clonal variants in a new subsection in the Results section entitled 'Guarding against erroneously called clonal variants.' This subsection follows the influenza A virus (IAV) and SARS-CoV-2 applications. One analysis that we propose is to perform a variant-calling threshold sensitivity analysis. We have therefore moved previous text from the IAV and SARS-CoV-2 sections that performs this sensitivity analysis into this subsection. (We feel that this works better due to also additional analyses that are now included in the influenza A virus and SARS-CoV-2 subsections, in response to reviewer comments.) A second analysis that we propose is to examine the frequencies of the clonal variants identified at a specified variant calling threshold. If any of these variants are present in the donor above a frequency of 0%, they could be removed from the

dataset and the inference approach re-applied to this modified dataset. We perform this analysis on the IAV and SARS-CoV-2 datasets and find that our results are not substantially impacted. Finally, we note that an iSNV that is present at low frequency in a donor and observed as fixed in a recipient is itself evidence of a tight transmission bottleneck, with additions to the Supplemental Material to show this quantitatively.

4. As noted by the authors, the "WT lineage," consisting of multiple transmitted viruses, could be genetically diverse. How can the authors distinguish between diversity of viruses transmitted from the donor and new mutations that arise in the recipient in real data?

Genetic diversity that is present in the initial particles that are transmitted to the recipient can be distinguished based on genetic variation present at those sites in the donor. If a donor viral population harbors an iSNV at a certain site (above the specified variant calling threshold), then if that allele is fixed at that site in the recipient, we assume that the allele derived from the donor.

We believe we addressed the remainder of this question in our response to question (3) above.

5. Here, the authors assume that viral sequences do not recombine and that linkage is complete. However, IAV can undergo reassortment, which could reduce linkage disequilibrium between genetic variation on different gene segments. For other viruses such as HIV, recombination is extensive. Could the authors comment on how recombination or reassortment would influence their results?

This is a great point. We have added text to the Discussion section to address this point, in the paragraph that focuses on limitations and assumptions of the presented inference approach. We first reiterate that our approach assumes that linkage is complete. For IAV, although reassortment is possible, we now cite some of our previous work that the effective rate of reassortment appears to be limited for IAV in human infections (Sobel Leonard et al. 2017a). Further, if an allele that arose *de novo* ends up fixing, it has to do so very rapidly, when viral population sizes are still small. This reduces the likelihood that much genetic diversification has occurred by the time of fixation and that cellular coinfection between different genetic variants has led to recombination. We further mention that if recombination/reassortment does occur frequently during the initial stages of an infection, it would act to bias our bottleneck size estimates to be low.

6. It is interesting that, when fitting the model to data, the authors have chosen to fix R0 for viral replication within the host and fit the mutation rate. Naively, one might expect that the mutation rate is easier to determine a priori from data. One might also expect that the mutation rates vary less between individuals – given that times from infection to symptom onset and viral load vary between individuals, R0 values may be heterogeneous. Is the model robust to heterogeneity in R0 values, and if not, can the authors explain why this is not a concern when applying the method to data?

This is also a really interesting comment. First, we note that our estimates of λ (the mean number of initial particles in the recipient) and mutation rate μ are largely insensitive to within-host R_0. Maximum likelihood estimates of λ do not change across the broad range of R_0 values considered in our sensitivity analyses (the range spans R_0 = 4.4 to 37.7 for IAV (Figure S1) and R_0 = 2.6 to 14.9 for SARS-CoV-2 (Figure S4)). Mutation rate estimates increase slightly with higher R_0 values assumed. Because λ estimates do not change across these broad ranges of R_0, this suggests that our inferences are robust to interindividual heterogeneity in R_0 values.

In discussing the reviewer's comment, however, we decided to include two additional analyses, first as applied to our simulated clonal variant dataset (Figure 3A). We first asked whether we could jointly estimate λ and R_0 if we set the mutation rate to its true value of μ = 0.40 mutations per genome per infection cycle. We now show these results in Figure 3I,J,K. These analyses indicate that the true values of λ and R_0 can be recovered in this case. We then asked whether we could simultaneously estimate λ, the within-host basic reproduction number R_0, and the mutation rate μ. We now show these results in Supplemental Figure S3 and allude to this figure from the main text. Unfortunately, when we try to estimate all three of these parameters, there are issues with identifiability. As such, either the within-host R_0 or the mutation rate μ has to be set to a reasonable literature estimate in order to be able to estimate the remaining two parameters.

Based on these results, we included additional analyses on our IAV and our SARS-CoV-2 datasets. The IAV analysis set the mutation rate for IAV to 1.75 mutations per genome per infection cycle, based on the estimate of this parameter in Pauly et al. (2017) *eLife*, and jointly estimated within-host R_0 and λ. The SARS-CoV-2 analysis set the mutation rate for SARS-CoV-2 to 0.03 mutations per genome per infection cycle, based on the estimate in Amicone et al. (2022) *EMPH*, and jointly estimated within-host R_0 and λ. These results are shown in Supplemental Figure S2 and are alluded to from the main text.

7. Here, several alternative approaches to estimating transmission bottleneck sizes from data are mentioned. The authors note several possible issues that could bias the results of these approaches toward inferring smaller bottleneck sizes, which their method aims to address. In applications to real data, however, they recover very similar results to those previously obtained using other methods. It would be helpful for the authors to discuss this point, and to clarify the advantage of their approach versus other methods, since the potential bias of the other methods apparently does not have a large effect on the results. Is little bias seen in other methods because (evidently) transmission bottleneck sizes are already very small?

Regarding the method discussed in the paper, there are multiple ways that this could be justified versus other approaches. One simple example would be to test the performance of multiple methods on simulated data, where it may be possible to observe the bias in other methods that the authors describe. Another way would be to explicitly point out data sets that can be analyzed using the authors' method, which would not be possible to analyze using alternatives.

We thank the reviewer for this comment. We have added a paragraph of text to the Discussion section to clarify this point. Specifically, this text indicates that our current findings corroborate the previous findings using existing methods based on shared genetic variation, but that this did not necessarily need to be the case. If we found a lack of clonal variants in recipients, this would have pointed towards large transmission bottlenecks, which would have been at odds with results from the existing methods that have the potential to vastly underestimate transmission bottleneck sizes.

Finally, we decided against formally evaluating the performance of multiple methods against a single common simulated dataset. This is because we already describe the reasons for why existing methods would be biased (non-random sampling of the donor's viral population and the occurrence of genetic drift during acute infections) and we feel that the effects of these factors are straightforward enough in terms of their effects to not warrant a full comparative analysis, especially given the length of the current manuscript.

Minor comments

8. In several places, the paper states that changes in variant frequencies within the recipient would tend to lead to underestimates of the transmission bottleneck size. It would be helpful to provide some intuition for this assertion.

Thank you for this comment. We believe that we have addressed this comment in our response to questions (7) above.

9. In Figure 4A and E, it is not specified which bars represent data and which ones represent the model. Ideally, all figures should come with a self-contained legend that indicates clearly what is being plotted.

We have edited the figure legend text for panels 4A and 4E to specify that the dark green bars correspond to the empirical distribution and that the light green bars correspond to the expectation distribution.

Reviewer: 2
Comments to the Author
Shi et al introduce a new method for estimation of the size of the bottleneck for infectious disease transmission, using sites without within-host diversity rather than, as in most previous methods, those with such diversity in the source. This is a polished and well-conceived manuscript. The method is a very worthwhile addition to inference algorithms for bottleneck sizes, which should complement existing methods well. It definitely should be published, but I have one fairly major concern with this iteration.

We thank the reviewer for their careful reading of our manuscript.

I find the assumption of a Poisson-distributed – not even a zero-truncated Poisson-distributed! –

number of initial viral particles to be a strange choice. When lambda is estimated to be 0.01 for both real viruses, over 99% of the probability mass is put on zero wild-type particles, i.e. there being no infection at all. (0.01 is also the lower limit of the values of lambda that were even considered, so the situation might actually be even more extreme. It surely is not good practice to just stick with a lower bound when your estimate is literally that lower bound, if there are no rounding issues here.) I struggle with the idea that this method can make reliable inferences along those lines as no such examples are in the data. The authors wisely do not attempt to define the circumstances of an infectious contact, so what this number actually means is very unclear, but the success probability of infection is not something this method or these datasets were ever well placed to answer. You would have to start with a very, very good argument for why that number is in fact Poisson-distributed, and what we have is instead handwavy (lines 301, 353). Indeed, we have a noted divergence of actual results to those expected from the fitted model in both cases, which the authors attribute to literature estimates of R0 being too high. Not to consider the possibility that the distribution of initial viral particles could be misspecified is a significant omission. I would suggest that since this method will only ever be used on transmission pairs, the distribution of founding particles should be one whose minimum value is 1.

We thank the reviewer for this comment. First, we would like to note that a zero-truncated Poisson distribution would yield the same results for mean N and mean N_b as the Poisson distribution we use. This is because these distributions, conditional on successful infection, are identical. That is, the light green bars shown in Figure 3A would be the same if the original distribution were a zero-truncated Poisson distribution rather than the Poisson distribution we show in dark green bars in Figure 3A. In practice, it therefore does not matter whether the distribution of founding particles is a probability mass function with values that are constrained to be ≥ 1 versus a probability mass function with values that are ≥ 0, since infections with 0 particles are never successful. We prefer using a non-negative probability mass function ($N \geq 0$), though, over a probability mass function that is constrained to be $N \geq 1$. This is because we can think of the initial viral particle distribution as being the distribution of initial viral particles that a contact of a donor is faced with. Only a subset of these contacts get infected. These distributions correspond nicely to the ones shown in Figure 3A.

To further hammer the point home, previous papers for SARS-CoV-2 do usually find the bottleneck to be 1 but larger estimates are not infrequent. The variance of the bottleneck size is a pertinent question, with consequences for, e.g., the sensitivity of using N_b>=2 as a test for a transmission pair. Here the variance of N is forced to be 0.01, due to the equality of mean and variance in a Poisson distribution. The probability of N>=2 under that distribution is about 5E-5, or if we condition on an infection actually taking place (i.e. N>=1) about 5E-3. The probability of N_b>=2 conditioned on N>=1 will be smaller still. (As the possibility of lambda being smaller than 0.01 is not explored, even this might be too large.) Since the assumptions surrounding the distribution of N are so stringent and questionable, I do not find this persuasive at all and would not like to see it start to be cited as an estimate of the probability of wider bottlenecks in

influenza or SARS-CoV-2.

We greatly appreciate this comment and can see the reviewer's argument. In response to it, we have refit the influenza A virus and SARS-CoV-2 datasets under two additional, and different, assumptions. The first one is that the distribution of initial viral population sizes is distributed according to a negative binomial distribution with an overdispersion parameter set at a small value (k = 0.1), corresponding to high levels of overdispersion in the initial particle number distribution (variance in N >> mean). We find with this analysis that the mean of this distribution is similarly very small for both viruses; Figures 5B and 5E). The second one assumes that the distribution of initial viral population sizes is distributed according to a bimodal distribution with a certain proportion p having a single viral particle successfully establish infection and the remaining proportion (1-p) having a very large number of viral particles successfully establishing infection. We assume by 'very large' that these latter infections do not result in any clonal variants, and that all clonal variants therefore need to arise from the $N = 1$ proportion of the infections. We find with this analysis that the proportion p is very close to one for both viruses; Figures 5C and 5F), indicating that for the majority of infections, the bottleneck size would be 1.

We present these two analyses to point out to the reader exactly what the reviewer is suggesting: that there could be wide bottlenecks in some cases and we cannot conclude that all transmission events have small bottleneck sizes. The data only indicate that a large proportion of transmissions would have to start with a bottleneck size of 1 (this is a robust conclusion based on the three different initial distributions we consider: the Poisson distribution, the negative binomial distribution, and the bimodal distribution). We have added these analyses to a new subsection at the end of the Results section.

There are assumptions here both that mutations have had time to fix before sampling, and that the population is ever-growing. Both seem at least somewhat questionable. What are the consequences if some mutations that would have fixed eventually had not at the time of sampling? What happens if you are sampling after peak viral load and the viral population is on the way down? I do not think these are problems that are fatal to the method, as all methods of this type must make some quite serious assumptions, but they are worth some consideration.

We thank the reviewer for this comment and now address these possibilities in the new subsection under Results entitled 'Guarding against the erroneous calling of clonal variants'.

The authors may want to clarify what is meant by a mutant lineage, in particular that further mutations to a mutant lineage do not themselves produce further mutant lineages. The statement that there are infinite numbers of mutant lineages in the second scenario (line 148) is also perhaps a little confusing since you are always time-constrained by sampling. The limit is infinity but in practice you cannot have gotten there.

We thank the reviewer for this suggestion and now explicitly define what a mutant lineage is in the Methods text.

Twice in the text we are informed that the distinction between N and N_b is to be outlined "later" (line 135) or "below" (line 185) but this seems to be missing from the main text; it can be understood from the supplement but should be spelled out in less mathematical language.

We thank the reviewer for this suggestion and have now added text to the main manuscript (in Methods subsection 'The stochastic within-host model' and in Results subsection 'Application to simulated data' to make the distinctions between N, N_b, mean N, and mean N_b all more clear.

The "etc?" in line 187 is presumably a typo.

We have removed the "etc" from this line and other lines, replacing it with '…'