CORONAVIRUS

Comment on "Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2"

Michael A. Martin[1] and Katia Koelle[2,3]*

A reanalysis of SARS-CoV-2 deep sequencing data from donor-recipient pairs indicates that transmission bottlenecks are very narrow (one to three virions).

In their recent research article (*1*), Popa *et al.* combined epidemiological and viral genetic data to characterize the transmission dynamics of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in Austria between February and April 2020. The genetic data they analyzed comprised >500 deep-sequenced virus samples. Beyond using consensus-level SARS-CoV-2 sequences to infer transmission clusters within Austria and to examine the role that Austria played in seeding regional epidemics elsewhere in Europe, the authors used their sequenced samples to characterize mutational dynamics within hosts and along short transmission chains. Although we believe that the findings from their consensus-level genetic analysis are robust, we here revisit their analyses of mutational dynamics at the below-the-consensus level. Specifically, we consider their estimates of the viral transmission bottleneck size, defined as the number of virions that successfully seed infection in a recipient individual following infection from a donor individual. Equivalently, it is the number of viral particles from a person who transmits the infection that contribute genetically to the viral population in the recipient who contracts it. From our reanalysis, we conclude that transmission bottleneck sizes for SARS-CoV-2 are not on the order of 1000 virions as concluded by the authors but instead much smaller.

Our decision to revisit Popa *et al.*'s conclusions on transmission bottleneck sizes stems from certain patterns present in some of their figures. First, inferred bottleneck size estimates using a 3% variant calling threshold were bimodal, with 14 of the 39 transmission pairs having an inferred bottleneck size (N_b) of <10 and the remaining 25 pairs having N_b estimates of 115 to 5000 (their fig. S4G). Further, when a 1% variant calling threshold was used, only a single transmission pair retained an N_b estimate of <10 (their figure 5B). In an attempt to understand these patterns, we first reanalyzed their deep sequencing data and recalled variants using their pipeline. In the analyses presented below, we use these recalled variant frequencies, which appear to be highly similar to those presented in Popa *et al.* based on the plots published as part of their article that show the frequencies of called variants in donor individuals against those in recipient individuals in their identified transmission pairs (10.5281/zenodo.4247401).

As expected, re-estimation of transmission bottleneck sizes at variant calling thresholds of 1 and 3% yielded similar results to those shown in (*1*) (Fig. 1A, fig. S1, A and B, and data file S1). During this analysis, we noticed that bottleneck size estimates dropped, sometimes precipitously, when going from a 1 to a 3% cutoff for each of the 13 transmission pairs that had donors with a maximum intrahost single-nucleotide variant (iSNV) frequency of >6% ($P = 0.004$, paired *t* test; Fig. 1A). Because increasing the variant calling threshold would remove low-frequency iSNVs from the analysis, these consistent decreases in N_b estimates could come about if low-frequency donor iSNVs pointed toward bottleneck sizes being large, whereas high-frequency donor iSNVs instead pointed toward bottleneck sizes being small. Examination of low-frequency iSNVs across donor-recipient pairs indicates a high degree of congruence between their frequencies (Fig. 1B, inset, and fig. S2 in data file S2), which would suggest wide transmission bottlenecks. In contrast, high-frequency donor iSNVs rarely appeared to be transmitted to their corresponding recipient (fig. S2), suggesting narrow transmission bottlenecks.

To come to terms with these conflicting patterns, we considered genetic variation that appeared de novo in recipient hosts. This genetic variation appears in the donor-against-recipient variant frequency plots as iSNVs that are absent from a donor but present in a corresponding recipient. When a de novo variant is observed as fixed in a recipient sample, we should not observe any shared iSNVs between a donor and a recipient that are present in the recipient at subclonal (that is, not fixed) frequencies unless within-host recombination occurred extremely rapidly or the fixed de novo variant arose multiple times in different genetic backgrounds. However, in the transmission pairs analyzed in Popa *et al.*, shared subclonal iSNVs are observed in several transmission pairs where there is also a fixed de novo variant present in the recipient. The transmission pair CoV_162 ➔ CoV_161 provides an example (Fig. 1B). This means that the low-frequency iSNVs shared between CoV_162 and CoV_161 are either spurious or that they arose independently in the recipient (that is, they are homoplasies). Although iSNV homoplasies have been documented in a number of recent SARS-CoV-2 studies (*2*, *3*), we believe that these low-frequency iSNVs in the Popa *et al.* transmission pairs are likely spurious, potentially arising from systematic issues related to the sequencing protocol. This is because these low-frequency iSNVs occur at extremely similar frequencies between the donor sample and the recipient sample (Fig. 1B, inset, and fig. S2), which is unlikely if the iSNVs were homoplasies. In either case, however, the low-frequency shared iSNVs in transmission pair CoV_162 ➔ CoV_161 and in other transmission pairs with fixed de novo variants in the recipient could only constitute transmitted genetic variation under scenarios that are highly implausible from a biological perspective and as such

[1] Graduate Program in Population Biology, Ecology, and Evolution, Emory University, Atlanta, GA 30322, USA. [2] Department of Biology, Emory University, Atlanta, GA 30322, USA. [3] Emory-UGA Center of Excellence for Influenza Research and Surveillance (CEIRS), Atlanta GA 30322, USA.
*Corresponding author. Email: katia.koelle@emory.edu

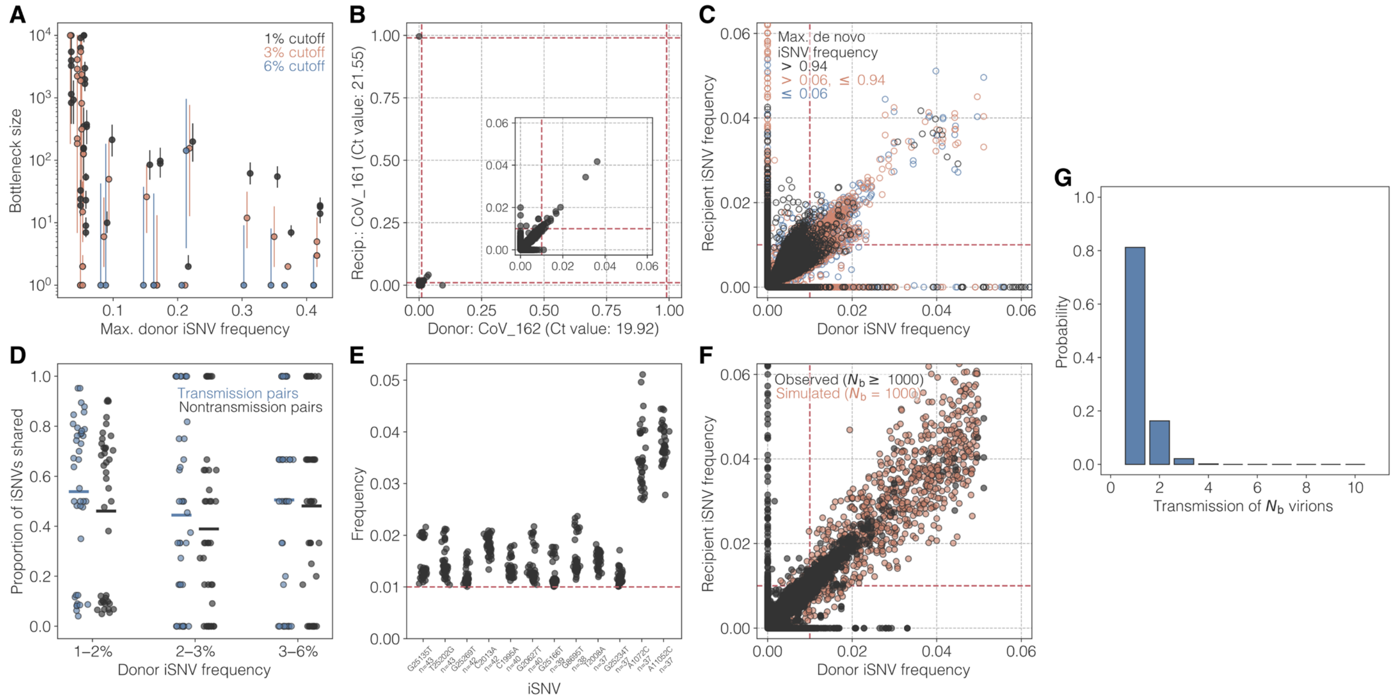

Fig. 1. Transmission bottleneck sizes and patterns of shared viral genetic diversity between the transmission pairs studied in Popa *et al*. (**A**) Bottleneck size estimates for 39 epidemiologically confirmed SARS-CoV-2 transmission pairs using variant calling thresholds of 1% ([0.01, 0.99]), 3% ([0.03, 0.97]), and 6% ([0.06, 0.94]). Estimates are based on all iSNVs that passed the quality-filtering thresholds. Maximum likelihood estimates are indicated by a colored circle and vertical lines show 95% CIs. Estimates are plotted according to the maximum iSNV frequency observed in the donor of the transmission pair. Estimates from the same transmission pair are offset slightly on the *x* axis to aid in visualization. iSNV frequencies are based on variant calling relative to donor-specific reference sequences. (**B**) All iSNVs observed in either the donor or the recipient of the epidemiologically confirmed CoV_162 ➔ CoV_161 transmission pair. Note the de novo variant in the recipient (C26894U) that is fixed within individual 161 and absent from individual 162. Inset highlights low-frequency iSNVs. Red dashed lines show the 1% variant calling threshold. iSNV frequencies are based on variant calling relative to donor-specific reference sequences. Here, all iSNVs that passed quality filtering thresholds are shown, regardless of whether they fell above or below the 1% variant calling threshold. (**C**) Low-frequency iSNVs observed across transmission pairs. Transmission pairs are classified as belonging to one of three groups according to the maximum de novo iSNV frequency observed in the recipient. Red dashed lines show the 1% variant calling threshold, and iSNV frequencies are based on variant calling relative to donor-specific reference sequences. (**D**) Proportion of iSNVs identified in a donor that are shared with a recipient host. Blue dots show proportions for epidemiologically linked pairs, whereas black dots show proportions for random, epidemiologically unlinked donor-recipient pairs. Random pairs were generated such that the random recipient was not a member of the same family as the focal donor or recipient and was not a known recipient of that donor sample. iSNVs were binned on the basis of their frequency in the donor: [0.01, 0.02), [0.02, 0.03), and [0.03, 0.06). Allele frequencies are based on variant calling relative to Wuhan/Hu-1 (9). Differences between the epidemiologically linked and unlinked distributions were assessed using the Kolmogorov-Smirnov test. This test failed to find significant differences between these distributions in the 1 to 2% donor frequency group ($P = 0.389$), 2 to 3% donor frequency group ($P = 0.752$), and 3 to 6% frequency group ($P > 0.999$). (**E**) Top 12 most abundantly shared iSNVs among the 43 samples involved in the 39 transmission pairs. iSNVs are ordered by the number of samples in which they were found (*n*). Each dot represents the allele frequency of that iSNV in a given sample. Red dotted line shows the 1% variant calling threshold. Allele frequencies are based on variant calling relative to Wuhan/Hu-1. (**F**) Patterns of shared viral genetic diversity between transmission pairs under the assumption of a large bottleneck of $N_b = 1000$ (red dots). Black dots show all iSNVs observed in either the donor or recipient for all transmission pairs with an estimated bottleneck size of ≥1000 at a variant calling threshold of 1%. (**G**) Probability of a transmission bottleneck of size N_b based on bottleneck size estimation using a variant calling threshold of 6% and data from all 13 transmission pairs with one or more iSNVs above this 6% threshold. The probability that a transmission involves a bottleneck size of 1, 2, or 3 virions exceeds 99%.

should be excluded from a transmission bottleneck analysis involving these transmission pairs.

A comprehensive analysis of all transmission pairs identified in Popa *et al.* indicates that patterns of low-frequency shared genetic variation are quantitatively highly similar across transmission pairs. To illustrate this, we categorized transmission pairs into three groups: transmission pairs with de novo fixed variants in the recipient (here defined as >94% in frequency), transmission pairs with de novo high-frequency (6–94%) variants in the recipient, and transmission pairs with only low-frequency de novo variants (≤6%). Figure 1C shows that the shared low-frequency iSNVs across these three groups are quantitatively extremely similar: Most shared iSNVs in each of these groups have frequencies falling between 1 and 2% in the donor, although some have frequencies of up to 6%. Whereas we should expect no transmitted subclonal genetic variation for the transmission pairs falling in the first group, we expect any shared iSNVs between transmission pairs belonging to the second group to have markedly different frequencies between the donor and the recipient because of genetic linkage with the high-frequency de novo variant in the recipient. The third group in principle could have very similar iSNV frequencies if bottleneck sizes were sufficiently large. In contrast to these expectations, Fig. 1C shows that all shared iSNVs (regardless of which group is being considered) are highly congruent in their frequencies between donors and recipients, indicating again that these iSNVs are very likely spurious. When we calculate the proportion of the low-frequency donor iSNVs that are observed in a corresponding recipient (at ≥1%) versus observed in an epidemiologically unlinked recipient, we find that the distribution of these proportions are highly similar (Fig. 1D). This finding again suggests that these shared low-frequency iSNVs do not constitute true shared genetic variation within transmission pairs; if these shared iSNVs were transmitted, we would expect the proportion of shared low-frequency variants to be higher for the corresponding recipient compared to an epidemiologically unlinked one.

Given these findings that shed doubt on low-frequency iSNVs constituting transmitted genetic variation, we decided to quantify the extent to which particular iSNVs were present across the samples used in the transmission pair analyses. We found that 5 iSNVs were present in ≥40 of the 43 samples analyzed at frequencies that fell into a very narrow range (1 to 2.2%) (Fig. 1E). Many other iSNVs were also present across numerous samples (Fig. 1E, fig. S3 in data file S3, and fig. S4), with the frequencies of any particular iSNV being highly similar across the samples that it appears in. This similarity in iSNV frequencies again argues against these low-frequency iSNVs being homoplasies and strongly argues for these iSNVs being spurious. To assess the evidence for this, we plotted the genome location of all variants observed in between 1 and 99% of reads in at least 10 samples against the read depth at those positions (fig. S5). Although these variants do not tend to appear in areas of particularly low sequencing coverage, they do cluster within a small number of sequenced amplicons, which are distributed across the genome (fig. S6).

Last, a comparison between observed patterns of iSNV frequencies between donors and recipients versus those expected under large transmission bottleneck sizes as inferred in Popa *et al.* further argues against the transmission of the low-frequency shared iSNVs. Specifically, observed iSNV frequencies from transmission pairs with inferred bottleneck sizes of $N_b \geq 1000$ show that iSNVs are present in both donor and recipient at highly similar frequencies or are observed exclusively in the donor or recipient (Fig. 1F). On this figure, we overlaid simulated iSNV frequencies under the assumption of a bottleneck size of $N_b = 1000$. Juxtaposition of the observed versus theoretically predicted iSNV frequencies highlights an inconsistency: at N_b values of ~1000, we should expect almost all (at least 96.1%) of the iSNVs present in the donor at ≥2% to be transmitted and also observed above the variant calling threshold of 1% in the recipient. However, only 77.5% of donor iSNVs within the 2 to 6% frequency range were observed in the corresponding recipients at ≥1% frequency. This inconsistency indicates that the low-frequency iSNVs themselves show patterns that cannot be parsimoniously explained by large transmission bottleneck sizes. Moreover, bottleneck sizes of around $N_b = 3000$ are needed to quantitatively reproduce patterns of shared iSNV frequencies (fig. S7); at this bottleneck size, nearly 100% of iSNVs present in the donor at ≥2% should be transmitted to the recipient, but this is not the case.

Given our finding that the shared low-frequency iSNVs called in Popa *et al.* are likely spurious, we re-estimated transmission bottleneck sizes using the beta-binomial method (*4*) at a conservative variant calling threshold of 6% (Fig. 1A, fig. S1C, and data file S1). Increasing the variant calling threshold does not bias bottleneck size estimates, but it does increase statistical uncertainty in the estimated values. At this 6% cutoff, only 13 transmission pairs had one or more donor iSNVs remaining, such that bottleneck sizes could only be estimated for these pairs. The maximum likelihood estimate for N_b was 1 for 12 of these 13 transmission pairs [with the largest upper bound of the 95% confidence interval (CI) being $N_b = 181$ virions]; for the remaining transmission pair (CoV_198 ➔ CoV_230), the estimate was $N_b = 143$ virions (95% CI = 4 to 951). This transmission pair was the only one where a donor iSNV (at a frequency of ~22%) was transmitted to a recipient but remained subclonal (at a frequency of ~17%). Because the confidence intervals around these estimates were large, we also estimated an overall transmission bottleneck size using the data from these 13 transmission pairs. We arrived at an estimate of a mean bottleneck size of 1.21, with three or fewer viral particles successfully seeding infection in >99% of successful transmissions (Fig. 1G). Of note, this estimate depends on patterns of genetic variation observed between donors and recipients of transmission pairs. We here relied on the transmission pairs specified in Popa *et al.*; misspecification of these pairs could result in erroneously small bottleneck estimates.

Our finding of a very tight transmission bottleneck from a reanalysis of the viral deep sequencing data from Popa *et al.* is consistent with conclusions from other recent studies that have quantified SARS-CoV-2 transmission bottleneck sizes in humans (*3*, *5*) and other mammals (*6*). These results indicate that SARS-CoV-2 has a narrow transmission bottleneck, similar in size to that of influenza A viruses (*7*). Small bottleneck sizes also mean that infections generally start off with very little, if any, viral genetic diversity, such that acute infections will likely be characterized by low levels of viral diversity except in instances of superinfection, consistent with other recent studies (*2*, *8*). Our reanalysis thus parsimoniously adds to a growing understanding of SARS-CoV-2 evolution between and within infected individuals.

SUPPLEMENTARY MATERIALS

www.science.org/doi/10.1126/scitranslmed.abh1803
Materials and Methods
Figs. S1 to S7
Data files S1 to S3

REFERENCES AND NOTES

1. A. Popa, J.-W. Genger, M. D. Nicholson, T. Penz, D. Schmid, S. W. Aberle, B. Agerer, A. Lercher, L. Endler, H. Colaço, M. Smyth, M. Schuster, M. L. Grau, F. Martínez-Jiménez, O. Pich, W. Borena, E. Pawelka, Z. Keszei, M. Senekowitsch, J. Laine, J. H. Aberle, M. Redlberger-Fritz, M. Karolyi, A. Zoufaly, S. Maritschnik, M. Borkovec, P. Hufnagl, M. Nairz, G. Weiss, M. T. Wolfinger, D. von Laer, G. Superti-Furga, N. Lopez-Bigas, E. Puchhammer-Stöckl, F. Allerberger, F. Michor, C. Bock, A. Bergthaler, Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. *Sci. Transl. Med.* **12**, eabe2555 (2020).
2. A. L. Valesano, K. E. Rumfelt, D. E. Dimcheff, C. N. Blair, W. J. Fitzsimmons, J. G. Petrie, E. T. Martin, A. S. Lauring, Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts. *PLOS Pathog.* **17**, e1009499 (2021).
3. K. A. Lythgoe, M. Hall, L. Ferretti, M. de Cesare, G. MacIntyre-Cockett, A. Trebes, M. Andersson, N. Otecko, E. L. Wise, N. Moore, J. Lynch, S. Kidd, N. Cortes, M. Mori, R. Williams, G. Vernet, A. Justice, A. Green, S. M. Nicholls, M. A. Ansari, L. Abeler-Dörner, C. E. Moore, T. E. A. Peto, D. W. Eyre, R. Shaw, P. Simmonds, D. Buck, J. A. Todd; Oxford Virus Sequencing Analysis Group (OVSG), T. R. Connor, S. Ashraf, A. da Silva Filipe, J. Shepherd, E. C. Thomson; COVID-19 Genomics UK (COG-UK) Consortium, D. Bonsall, C. Fraser, T. Golubchik, SARS-CoV-2 within-host diversity and transmission. *Science* **372**, eabg0821 (2021).
4. A. Sobel Leonard, D. B. Weissman, B. Greenbaum, E. Ghedin, K. Koelle, Transmission bottleneck size estimation from pathogen deep-sequencing data, with an application to human influenza A virus. *J. Virol.* **91**, e00171-17 (2017).
5. K. Braun, G. K. Moreno, C. Wagner, M. A. Accola, W. M. Rehrauer, D. A. Baker, K. Koelle, D. H. O'Connor, T. Bedford, T. C. Friedrich, L. H. Moncla, Acute SARS-CoV-2 infections

harbor limited within-host diversity and transmit via tight transmission bottlenecks. *PLOS Pathog.* **17**, e1009849 (2021).

6. K. M. Braun, G. K. Moreno, P. J. Halfmann, E. B. Hodcroft, D. A. Baker, E. C. Boehm, A. M. Weiler, A. K. Haj, M. Hatta, S. Chiba, T. Maemura, Y. Kawaoka, K. Koelle, D. H. O'Connor, T. C. Friedrich, Transmission of SARS-CoV-2 in domestic cats imposes a narrow bottleneck. *PLoS Pathog.* **17**, e1009373 (2021).
7. J. T. McCrone, R. J. Woods, E. T. Martin, R. E. Malosh, A. S. Monto, A. S. Lauring, Stochastic processes constrain the within and between host evolution of influenza virus. *eLife* **7**, e35962 (2018).
8. G. Tonkin-Hill, I. Martincorena, R. Amato, A. R. J. Lawson, M. Gerstung, I. Johnston, D. K. Jackson, N. R. Park, S. V. Lensing, M. A. Quail, S. Gonçalves, C. Ariani, M. S. Chapman, W. L. Hamilton, L. W. Meredith, G. Hall, A. S. Jahun, Y. Chaudhry, M. Hosmillo, M. L. Pinckert, I. Georgana, A. Yakovleva, L. G. Caller, S. L. Caddy, T. Feltwell, F. A. Khokhar, C. J. Houldcroft, M. D. Curran, S. Parmar; COVID-19 Genomics UK (COG-UK) Consortium, A. Alderton, R. Nelson, E. Harrison, J. Sillitoe, S. D. Bentley, J. C. Barrett, M. E. Torok, I. G. Goodfellow, C. Langford, D. Kwiatkowski; Wellcome Sanger Institute COVID-19 Surveillance Team, Patterns of within-host genetic diversity in SARS-CoV-2. *eLife* **10**, e66857 (2021).
9. F. Wu, S. Zhao, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, Y.-Z. Zhang, A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).

Acknowledgments: We thank A. Bergthaler and his group for providing clarification on the SARS-CoV-2 deep sequencing data submitted as part of their research article. We also thank C. Bergstrom for a clear definition of transmission bottleneck size and three anonymous reviewers for their insightful recommendations to improve this work. **Funding:** The research reported in this technical comment was supported by National Institute of Allergy and Infectious Diseases Centers of Excellence for Influenza Research and Surveillance (CEIRS) grant HHSN272201400004C and by the NIH National Institute of General Medical Sciences grant 1R01 GM124280-03S1. **Competing interests:** K.K. consults for Moderna on SARS-CoV-2 epidemiology and evolution. **Data and materials availability:** All raw sequencing data used in this study are available from the National Center for Biotechnology Information (NCBI) Sequencing Read Archive (SRA) BioProject #PRJEB39849. Associated metadata and analysis code to recreate the analysis are available at https://doi.org/10.5281/zenodo.5224640.

Submitted 20 February 2021
Accepted 26 September 2021
Published 27 October 2021
10.1126/scitranslmed.abh1803

Citation: M. A. Martin, K. Koelle, Comment on "Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2". *Sci. Transl. Med.* **13**, eabh1803 (2021).