CORONAVIRUS

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

Michael D. Nicholson[1], Lukas Endler[2], Alexandra Popa[2], Jakob-Wendelin Genger[2], Christoph Bock[2,3], Franziska Michor[4,5,6,7,8,9]*, Andreas Bergthaler[2]*

Further analysis of SARS-CoV-2 genome sequencing data identifies several highly recurrent genetic variants with low allele frequencies, which, if filtered out, provide estimates consistent with tighter transmission bottlenecks.

We thank Martin and Koelle for highlighting the challenges of inferring severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) transmission bottlenecks from viral genomics data in their Technical Comment (*1*). Eventually, controlled infections with defined numbers of SARS-CoV-2 virions are expected to provide direct experimental clarification of infectivity (*2*). However, such "human challenge trials" are ethically and practically difficult, and although experiments have been launched (*3*), no results have yet been published. Biomathematical approaches can be used to infer transmission bottleneck sizes through the tracking of low-frequency variants based on deep viral genome sequencing data and confirmed infector-infectee pairs. We established such a dataset comprising more than 400 SARS-CoV-2 genomes with epidemiological information from the first wave of the COVID-19 pandemic in Austria, when infection rates in the general population were low and it was often possible to reconstruct high confidence infector-infectee pairs through contract tracing.

Applying the mathematical model of viral transmission presented in (*4*) to our data, we estimated SARS-CoV-2 bottleneck sizes to be on average 10^3 viral particles [25 and 75% quartiles: 3.5 and 1763 at an allele frequency (AF) cutoff of 3%] across 39 epidemiologically resolved transmission pairs (*5*). Martin and Koelle reanalyzed our data and found similar results, confirming our previous observations. Yet, they raised concerns regarding the threshold of detection for including low-frequency variants for bottleneck size estimation and suggested that increasing the frequency cutoff to 6% was required to avoid technical artifacts. This higher threshold reduced the number of pairs for which the analysis was feasible to 13 of 39 possible pairs and resulted in bottleneck estimates of 1 virion for 12 pairs and 143 virions for the remaining pair.

In response to the commentary of Martin and Koelle, we reanalyzed our data in additional ways that complement those presented in our manuscript. Our current results provide strong support for the validity of our identified low-frequency variants, with one potential caveat: We identified several highly shared yet low-frequency variants that could, in principle, be due to technical or biological artifacts. Despite the shared variants comprising less than 0.004% of variants detected at >3% AF (18 of 5132), we found that when we excluded these shared variants, the resulting bottleneck estimates were altered and generally lower than those derived from the unfiltered set of low-frequency variants. After exclusion of these variants, we were still able to provide bottleneck estimates for 29 of 39 transmission pairs. Below, we outline our methodology and revised analysis in detail and also discuss controls that should be considered when performing bottleneck estimates.

As described in our paper (*5*), we designed and validated our analysis pipelines to deal with the technical challenges of detecting and analyzing low-frequency variants. This approach included processing of biological and technical replicates, performing titration experiments to account for varying amounts of viral RNA present in a given sample, and using stringent parameters for variant calling pipelines, with realignment in the vicinity of indels to prevent calling errors (*5*). We assessed the detection limit of our sequencing and alternative AF calling pipeline with both a titration series of patient samples and synthetic SARS-CoV-2 RNA genomes. We also confirmed the reliability and reproducibility of the results obtained with our analysis pipeline using independent sequencing runs of two patient samples harboring different viral loads and included one sample as a quality control in all subsequent sequencing runs to account for sequencing batch effects. Furthermore, a comparison of the mutation spectra of low-frequency variants (2 to 50%) identified similar mutational signatures as compared to the pool of high-confidence fixed variants across the genome. However, they differed from the pattern of mutational signatures observed for variants at frequencies below 1% that were subsequently treated as background noise (*5*).

Together, these steps led to the conclusion that a cutoff of 1 or 2% for the AF resulted in high-confidence variants. Conscious of the effect of differing cutoffs on the bottleneck size estimates, we demonstrated the robustness of our bottleneck size estimates for AF cutoffs of 1, 2, and 3%. Across 43 samples in 39 transmission pairs, 86 variants were detected in at least one sample with a frequency of at least 3% (*5*). Increasing this cutoff to 6%, as suggested by Martin and Koelle, cuts the number of variants considered for bottleneck size estimates by a third (57 variants remaining).

A recent study (*6*) observed 18 sites that were present at AF in the range of 3 to 50% in more than 20 of 1313 (or ~1.5%) of sequenced SARS-CoV-2 genome samples. These sites were termed "highly shared," and the authors chose to mask these variants for

[1] Cancer Research UK Edinburgh Centre, Institute of Genetics and Cancer, University of Edinburgh, Crewe Road, Edinburgh EH4 2XU, UK. [2] CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, 1090 Vienna, Austria. [3] Institute of Artificial Intelligence, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, 1090 Vienna, Austria. [4] Department of Data Science, Dana-Farber Cancer Institute, Boston, MA 02215, USA. [5] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02215, USA. [6] Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA. [7] The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [8] The Ludwig Center at Harvard, Boston, MA 02115, USA. [9] Center for Cancer Evolution, Dana-Farber Cancer Institute, Boston, MA 02115, USA.

*Corresponding author. Email: michor@jimmy.harvard.edu (F.M.); abergthaler@cemm.oeaw.ac.at (A.B.)

their bottleneck analysis. To examine the validity of the same "prevalence threshold" of 1.5% for our dataset, we carried out a simulation analysis incorporating key features of our data (Materials and Methods). Under a null model where all 424 samples are unlinked, we assessed the number of sites that are expected to be mutated in several samples by chance. We found that 5% of simulations contained a site mutated in 6 of 424 samples (1.4%), whereas less than 0.3% of simulations yielded a site mutated in greater than 1.5% of samples (Fig. 1A). Therefore, we initially adopted the same prevalence threshold as (*6*) of 1.5% (7 of 424 samples).

We detected 18 sites with allele frequencies in the range 3 to 50% in more than 6 of 424 samples (Fig. 1B). Six of these variants were present in transmission pairs with a 3% low-frequency threshold and therefore affected the bottleneck analysis. With these six variants excluded, we were able to estimate the bottleneck size for 29 of 39 transmission pairs with a low-frequency variant cutoff of 3% (Fig. 1C). These estimates resulted in a bottleneck size of 1 virion for 27 of 29 pairs and 8 virions and 58 virions for the other 2 pairs, respectively. With a prevalence threshold of 10%, which resulted in the exclusion of two variants, the bottleneck estimates for the 29 pairs remained the same, and we were able to analyze 1 additional pair, which resulted in a bottleneck estimate of 1 virion. Thus, the exclusion of only two variants (positions 1072 and 11052), which have not been flagged as problematic according to a continuously updated database (*7*), led to a substantial reduction in bottleneck size estimates. These findings highlight the sensitivity of the bottleneck estimation method (*4*) to highly prevalent variants at low allele frequencies. Although these results are similar to those of Martin and Koelle, our method enabled analysis of a larger set of transmission pairs through evidence-guided exclusion of sites and thereby preservation of valuable sequence information.

Such approaches may be effective at removing a certain type of potential artefact. However, they come with the risk of introducing systematic biases by eliminating specifically those variants that are consistent across infector-infectee pairs; for example, a variant present in several longitudinal samples and also shared across an infector-infectee pair could be excluded. Furthermore, the large confidence intervals of these estimates, especially those reporting a single virion as the transmission bottleneck, highlight the limit in precision of the bottleneck size estimation method in the case of low intrasample genetic diversity. Notably, our study focused on many samples as part of epidemiologically resolved clusters and analyzed SARS-CoV-2 genomes early in the pandemic, when masks or physical distancing measures were not yet implemented. Many of the investigated transmission pairs occurred as part of large gatherings, during activities in closed spaces for extended time periods or with frequent personal interactions. In this context, a productive infection by multiple viral particles from infector to infectee seems plausible and is also supported by another recent study (*8*).

Future bottleneck size analysis should incorporate not only a frequency cutoff for variants, but also a threshold for how prevalent a variant is across samples. On the basis of our data, we believe a frequency cutoff of 3% is accurate for datasets of similar quality. How to select an appropriate prevalence threshold is less clear;

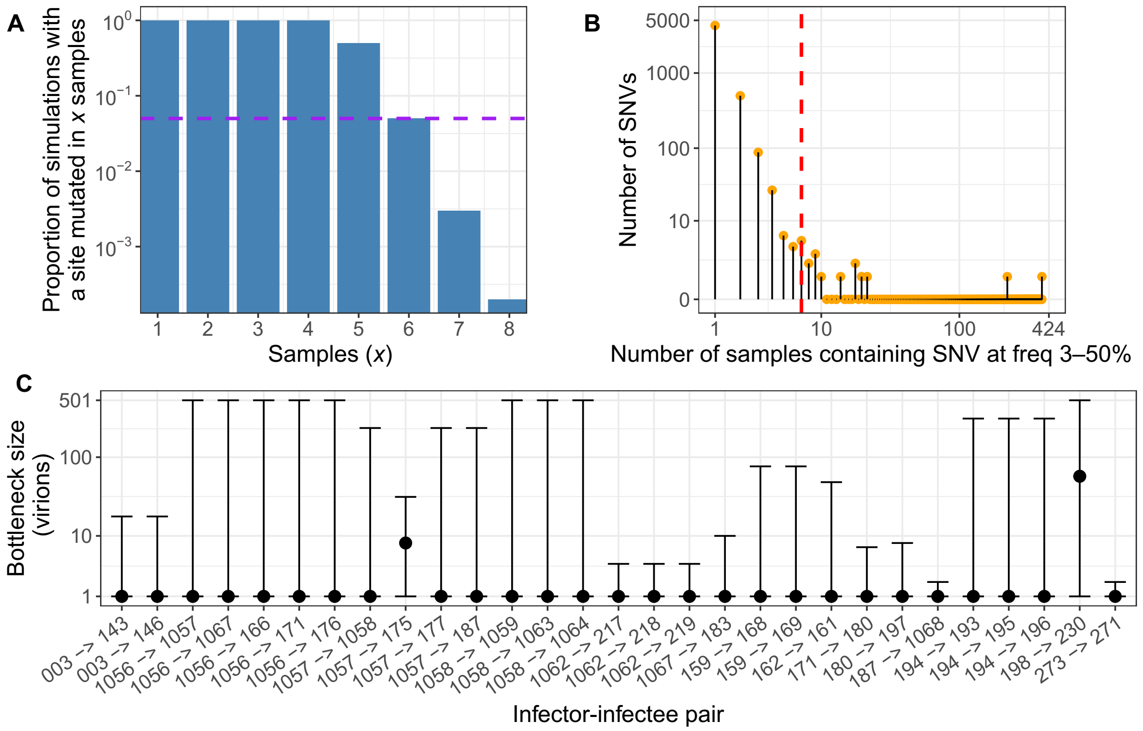

Fig. 1. Detection of highly prevalent variants and their impact on bottleneck estimates. (**A**) Simulation analysis under a null model of variant sites shows that it is unlikely to observe a mutated site in more than ~1.5% (7 of 424) samples. Horizontal dashed line denotes 0.05. (**B**) The number of single-nucleotide variants (SNVs) in a given number of samples at allele frequencies between 0.03 and 0.5. The red vertical line corresponds to 7 of 424 samples (~1.5%); variants to the right of this line are termed "highly shared". (**C**) Bottleneck size estimates after masking six highly shared variants. Dots are maximum likelihood estimates, and whiskers are 95% confidence intervals capped at 501.

however, we hope that the null model presented in this report provides a starting point for further investigations, such as more complex simulations including realistic infection networks and de novo mutations. In our own data, variants may be observed in several samples at low frequency because the samples were acquired longitudinally from the same patients or due to transmission; thus, a variant observed in several samples at low frequency does not necessarily mean it is spurious. Additional data are necessary to fully understand the provenance and consequences of these variants. Furthermore, a stringent threshold may make it difficult or impossible to detect large bottlenecks even when they are a biological reality.

In our view, currently the most suitable solution is manual curation of the low-frequency variants to exclude technical artefacts while retaining as many bona fide (high-confidence) low-frequency variants as possible in the dataset. Confidence in the results of such a curation approach can be increased by comparing mutational signatures and performing careful technical control experiments. Our expanding knowledge about the genetic diversity of SARS-CoV-2 and the evolutionary dynamics of variants will continue to shape future analyses and highlight potential limitations (*9*). We hope that our dataset and the discussion provided here will motivate further research into the challenges of and solutions to inferring SARS-CoV-2 transmission bottlenecks.

MATERIALS AND METHODS

Null model for highly prevalent sites

For each simulation, we generated a synthetic batch of 424 samples. Each synthetic sample received a number of mutations sampled, with replacement, from the distribution of variant counts in the frequency range [0.03, 0.5] in the patient data. We determined the proportion of unique variants in the frequency range [0.1, 1] from the patient data that had a mutated base A, C, G, and U. The mutations for a synthetic sample were then placed on the SARS-CoV-2 genome according to these base proportions, with each site mutated at most once. For a batch of synthetic samples, we then determined the number of sites that were mutated in x samples. This procedure was performed for 10,000 simulated batches.

REFERENCES AND NOTES

1. M. A. Martin, K. Koelle, Reanalysis of deep-sequencing data from Austria points towards a small SARS-COV-2 transmission bottleneck on the order of one to three virions. *bioRxiv*, 2021.02.22.432096 (2021).
2. T. Kirby, COVID-19 human challenge studies in the UK. *Lancet Respir. Med.* **8**, e96 (2020).
3. University of Oxford, Human challenge trial launches to study immune response to COVID-19 (2021); www.ox.ac.uk/news/2021-04-19-human-challenge-trial-launches-study-immune-response-covid-19.
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. 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).
6. 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, 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).
7. N. G. N. De Maio, C. Walker, R. Borges, L. Weilguny, G. Slodkowicz, Issues with SARS-CoV-2 sequencing data (2020); https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473.
8. M. Prentiss, A. Chu, K. K. Berggren, Superspreading events without superspreaders: Using high attack rate events to estimate N_o for airborne transmission of COVID-19. *medRxiv*, 2020.10.21.20216895 (2020).
9. 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).

Acknowledgments: We thank M. A. Martin and K. Koelle for their valuable contribution and the reviewers for their constructive comments. We are also grateful to the entire team involved in the previous study (*5*). **Funding:** M.D.N. is a cross-disciplinary post-doctoral fellow supported by funding from CRUK Brain Tumour Centre of Excellence Award (C157/A27589). J.-W.G. is supported by a DOC fellowship of the Austrian academy of Sciences. A.B. was supported by an ERC Starting Grant (European Union's Horizon 2020 research and innovation program, grant agreement numbers 677006). The project was funded, in part, by the Vienna Science and Technology Fund (WWTF) as part of the WWTF COVID-19 Rapid Response Funding 2020 (to A.B.) and supported by a collaboration with the Austrian Agency for Health and Food Safety (AGES). **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Data related to this study has been made available within the previous study by Popa *et al.* (*5*).

Submitted 5 May 2021
Accepted 11 August 2021
Published 27 October 2021
10.1126/scitranslmed.abj3222

Citation: M. D. Nicholson, L. Endler, A. Popa, J.-W. Genger, C. Bock, F. Michor, A. Bergthaler, Response to comment on "Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2". *Sci. Transl. Med.* **13**, eabj3222 (2021).

Science Translational Medicine

Response to comment on “Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2”

Michael D. NicholsonLukas EndlerAlexandra PopaJakob-Wendelin GengerChristoph BockFranziska MichorAndreas Bergthaler

Sci. Transl. Med., 13 (617), eabj3222. • DOI: 10.1126/scitranslmed.abj3222

Science Translational Medicine (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Translational Medicine* is a registered trademark of AAAS.