medRxiv preprint doi: https://doi.org/10.1101/2024.05.26.24307940; this version posted May 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Estimating R_e and overdispersion in secondary cases from the size of identical sequence clusters of SARS-CoV-2

Emma B. Hodcroft[1,2,3◕¤], Martin S. Wohlfender[1,2,4◕], Richard A. Neher[3,5], Julien Riou[1,2,6], Christian L. Althaus[1,2]

1 Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland
2 Multidisciplinary Center for Infectious Diseases, University of Bern, Bern, Switzerland
3 Swiss Institute of Bioinformatics, Lausanne, Switzerland
4 Graduate School for Cellular and Biomedical Sciences, University of Bern, Bern, Switzerland
5 Biozentrum, University of Basel, Basel, Switzerland
6 Department of Epidemiology and Health Systems, Unisanté, Center for Primary Care and Public Health & University of Lausanne, Lausanne, Switzerland

◕These authors contributed equally to this work.
¤Current Address: Swiss Tropical and Public Health Institute, Allschwil, Switzerland
* emma.hodcroft@unibas.ch

Abstract

The wealth of genomic data that was generated during the COVID-19 pandemic provides an exceptional opportunity to obtain information on the transmission of SARS-CoV-2. Specifically, there is great interest to better understand how the effective reproduction number R_e and the overdispersion of secondary cases, which can be quantified by the negative binomial dispersion parameter k, changed over time and across regions and viral variants. The aim of our study was to develop a Bayesian framework to infer R_e and k from viral sequence data. First, we developed a mathematical model for the distribution of the size of identical sequence clusters, in which we integrated viral transmission, the mutation rate of the virus, and incomplete case-detection. Second, we implemented this model within a Bayesian inference framework, allowing the estimation of R_e and k from genomic data only. We validated this model in a simulation study. Third, we identified clusters of identical sequences in all SARS-CoV-2 sequences in 2021 from Switzerland, Denmark, and Germany that were available on GISAID. We obtained monthly estimates of the posterior distribution of R_e and k, with the resulting R_e estimates slightly lower than resulting obtained by other methods, and k comparable with previous results. We found comparatively higher estimates of k in Denmark which suggests less opportunities for superspreading and more controlled transmission compared to the other countries in 2021. Our model included an estimation of the case detection and sampling probability, but the estimates obtained had large uncertainty, reflecting the difficulty of estimating these parameters simultaneously. Our study presents a novel method to infer information on the transmission of infectious diseases and its heterogeneity using genomic data. With increasing availability of sequences of pathogens in the future, we expect that our method has the potential to provide new insights into the transmission and the overdispersion in secondary cases of other pathogens.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.26.24307940; this version posted May 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Author summary

Pathogen transmission is a stochastic process that can be characterized by two parameters: the effective reproduction number R_e relates to the average number of secondary cases per infectious case in the current conditions of transmission and immunity, and the overdispersion parameter k captures the variability in the number of secondary cases. While R_e can be estimated well from case data, k is more difficult to quantify since detailed information about who infected whom is required. Here, we took advantage of the enormous number of sequences available of SARS-CoV-2 to identify clusters of identical sequences, providing indirect information about the size of transmission clusters at different times in the pandemic, and thus about epidemic parameters. We then extended a previously defined method to estimate R_e, k, and the probability of detection from this sequence data. We validated our approach on simulated and real data from three countries, with our resulting estimates compatible with previous estimates. In a future with increased pathogen sequence availability, we believe this method will pave the way for the estimation of epidemic parameters in the absence of detailed contact tracing data.

Introduction

The COVID-19 pandemic prompted an unprecedented global effort in generating and sharing SARS-CoV-2 sequences. As of March 2024, four years after the first sequences were released, over 16 million full-genome sequences have been shared, almost fifty times the estimated number of full-genome flu sequences generated over decades [1]. At the same time, scientists, governments, and public health authorities were keen to make use of all available methods to better understand the ongoing pandemic and implement appropriate responses. For example, information on the transmission dynamics of SARS-CoV-2, that can be characterized by the basic reproduction number R_0 and the effective reproduction number R_e, was critical to understanding whether current pandemic restrictions were controlling transmission [2]and whether new variants might cause pressure on healthcare systems. As well as estimating R_e from the available case-count data [3, 4], the availability of significant numbers of sequences created the opportunity to estimate this, and other parameters, through sequence- and phylogenetic-based methods as well [5–7].

Historically many of these techniques have been developed and applied to study pathogen dynamics in the past or in outbreak conditions when information such as reliable case counts may not be available [8–11]. However, another benefit of utilizing sequences is the ability to estimate parameters that otherwise require a level of detailed data acquisition that is difficult to obtain. A prime example of this is measuring the heterogeneity in transmission, which can be quantified by the dispersion parameter k. Estimating k normally requires knowledge not only of the number of confirmed cases per day, but the number of secondary cases created by individual cases. In a pandemic or epidemic, k quantifies how much spread may be driven by superspreading events, and thus can inform intervention efforts [12–17]. As part of contact-tracing efforts by many countries, estimates of k using clusters of linked cases have been possible, but due to the detailed level of data most require, have generally been limited to use on a few thousand cases and contacts [18–31], or simulated data [32]. In countries where SARS-CoV-2 cases were well-contained and thus had a high probability of being fully-traced, these datasets may be able to capture transmission dynamics well, but particularly as the pandemic has progressed and case numbers have risen, using relatively small subsets may both not fully capture the larger dynamics and may not be completely traced [33]. Using the large number of readily-available SARS-CoV-2 sequences to estimate values

medRxiv preprint doi: https://doi.org/10.1101/2024.05.26.24307940; this version posted May 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

such as k could overcome these limitations.

Blumberg and Lloyd-Smith's (2013) [34] method provides a way to estimate R_0 (or R_e) and k without requiring contact tracing data, and instead using a transmission chain size distribution. Their framework presumes transmission chains are finite with R_e below 1, or in other words, the transmission chain must "stutter out". This method has been successfully applied to pathogens that spill over from animal reservoirs and don't lead to sustained human-to-human transmission, such as mpox prior to 2022 and MERS-CoV. In MERS-CoV outbreaks during 2013-2014, multiple spillover events from camels and well-traced transmission chains have resulted in data sets well-suited to estimating R_0 and k [35–37]. However, even in instances such as these, where transmission networks are relatively well-recorded, cases may be missed, mis-attributed, or incorrectly considered a part of a transmission chain rather than a separate introduction due to insufficient background case detection. Thus, Blumberg and Lloyd-Smith [38] adapted their method to take into account imperfect case-detection.

In the context of the ongoing COVID-19 pandemic, while individual transmission chains might stop, the state of being in a pandemic implies that overall these chains persist. In the context of the ongoing COVID-19 pandemic, cases can not be readily decomposed into individual limited transmission chains. Instead, transmissions chains aggregate to macroscopic case numbers and the method of stuttering chains by Blumberg and Lloyd-Smith [34] can not be applied. To overcome this limitation, we propose to use clusters of identical sequences as a proxy for transmission chains: even in an ongoing transmission chain, the mutation rate of the virus means that every unique genotype will only exist until the virus mutates and thus is limited in size. To this end, we introduce a new metric, the genomic reproduction number R_g, that relates the reproduction number of cases with identical sequences to the effective reproduction number of cases, R_e. During our work, Tran-Kiem and Bedford (2024) [39] published a related study using the same approach.

In this study, we extended the method by Blumberg and Lloyd-Smith [34] and developed a Bayesian framework to infer R_e and k from viral sequence data. First, we developed a mathematical model of the size distribution of sequence clusters, in which we integrated viral transmission, the mutation rate of the virus, and incomplete case-detection to capture all aspects of the data-generating mechanism. Second, we implemented this model within a Bayesian inference framework, that we validated using simulated data. Third, we identified clusters of identical sequences in all SARS-CoV-2 sequences in 2021 from Switzerland, Denmark, and Germany that were available on GISAID and applied our method to obtain monthly estimates of R_e and k in these countries.

Materials and methods

Data

We downloaded all available SARS-CoV-2 sequences (14.8 million) from GISAID [1] on 7 March 2023. We ran them through the Nextstrain's *ncov-ingest* pipeline [40], which provides a list of all nucleotide mutations relative to the Wuhan-Hu-1/2019 reference (Genbank: MN908947) via Nextclade analysis [41]. Sequences identified as being problematic due to divergence issues or variant sequences reported prior to the variant's origin (date issues) were excluded. We then selected sequences from Switzerland, Denmark, and Germany, three countries with different testing and sequencing strategies resulting in different case-detection and sequencing coverage. The total number of sequences at the start of our analysis was 162,049 for Switzerland, 632,400 for Denmark, and 901,748 for Germany. All sequences are available from GISAID after registration as

medRxiv preprint doi: https://doi.org/10.1101/2024.05.26.24307940; this version posted May 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

EPI_SET_240326pm for Switzerland, EPI_SET_240326mz for Denmark, and EPI_SET_240326uh for Germany (see also S1-S3 Tables).

To identify genetically identical sequences efficiently on a large number of sequences, the list of nucleotide mutations was used as a hash key, with sequences with identical hash keys being classified as a cluster. To minimize the chance that sequences with low numbers of mutations, which may have arisen independently, falsely cluster together, only sequences with more than four mutations were used. Due to the mutation rate of SARS-CoV-2, this limitation only impacted sequences at the very start of the pandemic and before our study period, as most lineages had more than four nucleotide mutations by March/April 2020.

Clustering by lists of nucleotide mutations is efficient but imperfect, as sequencing errors such as lack of coverage and use of ambiguous bases can impact whether sequence mutation lists hash identically. We assumed that such errors are randomly distributed through time and thus have minimal effect on the overall cluster size distribution of such a large number of sequences. It's infeasible to account for all possible sequencing errors, but in an effort to minimize their impact, for all recognised Nextstrain variants, variant-defining mutations as obtained from CoVariants.org [42] were removed and replaced with the variant name. This prevents variant sequences being erroneously separated if they are missing variant-defining mutations. The robustness of the resulting clusters for each country was checked by picking 15 clusters at random, plus the top 10 largest clusters, and identifying them on phylogenetic trees and verifying that the shared mutations appeared only once, forming a monophyletic cluster. The code used for this processing is available at github.com/emmahodcroft/sc2_rk_public.

We grouped the identical sequence clusters into monthly time windows from January to December 2021. We assigned clusters to a given month if at least one sequence was sampled during that month. We focused on this period as the arrival of the SARS-CoV-2 variant of concern (VoC) Alpha in late 2020 led to a scale-up in sequencing efforts in all three countries. Sequencing uptake remained relatively high though 2021 until the arrival of Omicron in late 2021, when a dramatic increase in cases led to a reduction in case ascertainment and sequencing coverage. The number of sequences that were assigned to identical sequence clusters from 1 January 2021 to 31 December 2021 was 96,622 for Switzerland, 267,472 for Denmark, and 355,193 for Germany.

We calculated the sequencing coverage for each month as the ratio of the number of sampled sequences in GISAID over the total number of laboratory-confirmed cases as reported by public health authorities in the respective country. For each country and month we determined among the sequences that have been assigned to identical sequence clusters the number of sequences that have been sampled in the specific month. We directly retrieved daily numbers of newly confirmed cases from public sources of Switzerland [43], Denmark [44] and Germany [45] and summarized them to monthly values. For comparison with our estimates of R_e, we also downloaded estimates that were based on laboratory-confirmed cases for all three countries (github.com/covid-19-Re) [7]. To show as additional information in results figures, we retrieved proportions of viral variants among sequences for all three countries on a bi-weekly basis from CoVariants.org [42].

Mathematical model

We derived a model of the size distribution of identical sequence clusters (Fig 1). To this end, we applied branching process theory to viral transmission, with nodes corresponding to cases and connections between two nodes corresponding to transmission events. We assumed that for each node the number of offspring, i.e. the number of persons to whom an infected individual transmitted the virus, is independent and identically distributed. In addition to the viral transmission process, we

medRxiv preprint doi: https://doi.org/10.1101/2024.05.26.24307940; this version posted May 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

consecutively took into account both the mutation of the virus and the incomplete case-detection. At each node of the branching process, a mutation of the virus could occur, with a constant probability for all nodes. If this happened, then the mutation was passed on to all secondary cases. The mutation process allowed us to divide the overall transmission tree into multiple smaller transmission trees by the viral genome sequence present at the nodes. From there we obtained a model of the distribution of the size of identical sequence clusters according to the parameters governing viral transmission and mutation. A detailed description of all different steps can be found in section 3 of Supplementary Material S1 Text.

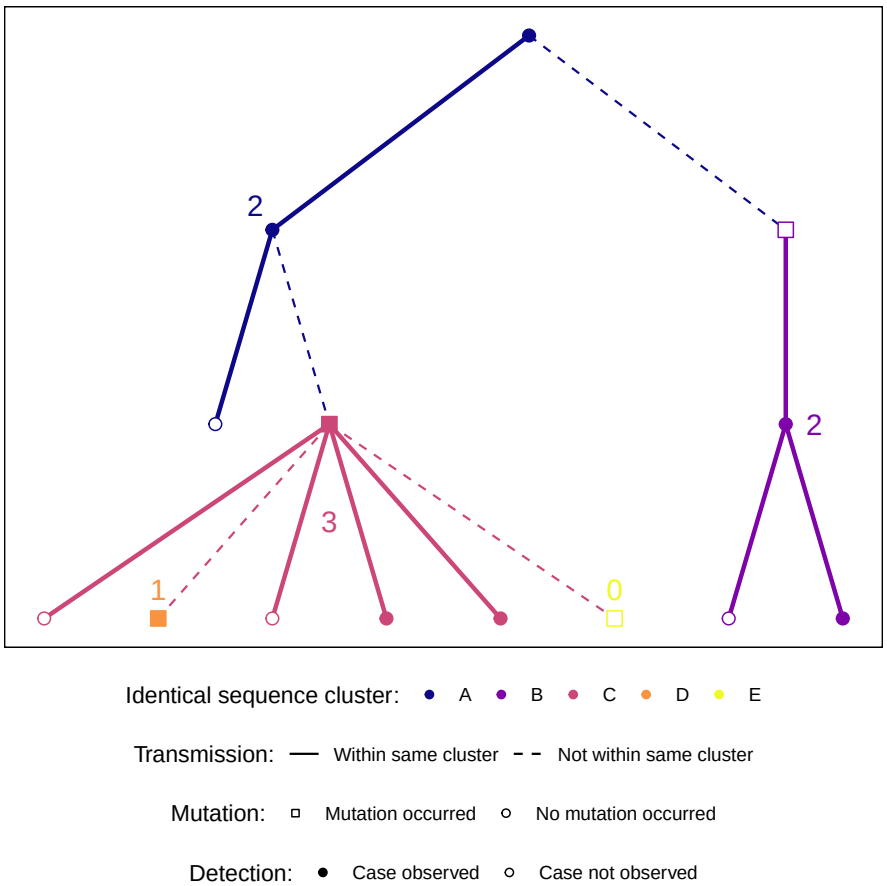

Fig 1. Transmission and identical sequence clusters The branching process model creates transmission clusters. Viral mutation splits the transmission clusters into identical sequence clusters. The numbers indicate the size of the observed identical sequence clusters, i.e. the number of cases in the clusters that were tested and sequenced.

This approach allowed us to adapt a method previously developed by Blumberg and Lloyd-Smith [34] to estimate R_e and k from the distribution of cluster size to our specific case involving the size of identical sequence clusters. This method assumes a negative binomial distribution for the number of secondary cases with mean R_e (interpreted as the average number of secondary cases per infectious case in the current conditions of transmission and immunity) and dispersion parameter k (lower values implying more dispersion, and thus more superspreading events). In each secondary case, the probability that the virus will mutate was defined as:

medRxiv preprint doi: https://doi.org/10.1101/2024.05.26.24307940; this version posted May 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

$$\mu = 1 - e^{-MD/365.25} \tag{1}$$

where M is the expected number of nucleotide mutations observed in a SARS-CoV-2 genome per year and D the mean generation time. For a within-variant rate of 14 mutations per year [46] and a mean generation time of 5.2 days [47], the probability of mutation for each case is 18.1%. Since a new mutation splits the transmission clusters, we introduced a new metric, the *genomic* reproduction number:

$$R_g = (1 - \mu)R_e \tag{2}$$

that is related to the effective reproduction number of cases R_e and corresponds to the average number of secondary cases within a given identical sequence cluster. Since the number of secondary cases follows a negative binomial distribution with mean R_e and dispersion parameter k, the number of secondary cases that belong to the same identical sequence cluster as their source case also follows a negative binomial distribution with mean R_g and dispersion parameter k.

As a last step, we incorporated incomplete case-detection assuming an independent observation process [38]. We computed the probability that a case was detected as the product of the probability that a case was confirmed by a test and the probability that the viral genome of a confirmed case was sequenced.

Bayesian inference

We developed a Bayesian inference framework to estimate the model parameters from the size distribution of identical sequence clusters and implemented in `Stan`, a probabilistic modeling platform using Hamiltonian Monte Carlo methods [48]. We used weakly-informative prior distributions for R_e, k, and the testing probability and an informative prior distribution for the yearly mutation rate M. The mean generation time and the sequencing probability were taken as fixed values. We estimated R_e, k, and the testing probability for monthly time windows from January to December 2021. The posterior samples were summarized by their mean and 95% credible interval (CrI). More details can be found in sections 4 and 6 of Supplementary Material S1 Text. The `R` and `Stan` code files are provided within the `R` package `estRodis`, available on the following GitHub repository: github.com/mwohlfender/estRodis.

Simulations

We conducted a simulation study to validate the model. We generated data of identical sequence clusters for different parameter combinations of R_e, k, and the testing and sequencing probability. To this end, we performed stochastic simulations of transmission using branching processes and then applied mutation and incomplete case-detection. Cases with an identical sequence were grouped in a cluster. We subsequently applied our inference framework to the simulated data and compared the mean of the posterior distribution of the estimated parameters to the true values using the root mean squared error and coverage of the true value by the 95% CrI of the posterior distribution (Supplementary Material S1 Text, section 5).

Results

Based on the new metric of the *genomic* reproduction number R_g, we developed a Bayesian inference model to estimate R_e, k, and the testing probability from data on

medRxiv preprint doi: https://doi.org/10.1101/2024.05.26.24307940; this version posted May 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

the size distribution of identical sequence clusters. We validated this model using simulated data and investigated how combinations of R_e, k, the testing probability, and the sequencing probability could be accurately recovered. To this end, we ran 10 rounds of simulation of 3,000 identical sequence clusters for 240 parameter combinations (Fig 2). Generally, we found that the error between the true and estimated values of R_e and k is minimized for higher testing and sequencing probabilities. Our method provided reliable estimates of R_e and k as long as R_e was below 1.2 (i.e., R_g is below 1 given our assumption for the mutation rate, Fig 2 A and B). When R_e was set to 1.3 (so that R_g is above 1), the inference model could not reliably recover the true value anymore. Overall, the validation showed that our model can be used to estimate R_e and k in specific settings.

We identified clusters of identical SARS-CoV-2 sequences using real-world sequencing data from Switzerland, Denmark, and Germany. From a total of 96,622 sequences from Switzerland, 267,472 from Denmark, and 355,193 from Germany, we identified 58,587, 84,537 and 218,497 clusters of identical sequences, respectively. Small clusters dominated the distribution of cluster size, with 79.7%, 70.8%, and 79.4% of the clusters being of size one in Switzerland, Denmark, and Germany, respectively. The mean cluster size was 1.65 in Switzerland, 3.16 in Denmark, and 1.63 in Germany, which suggests a higher probability of testing and/or sequencing in Denmark compared to the other countries. The sequencing coverage for laboratory-confirmed SARS-CoV-2 cases indeed differed substantially between the three countries over time, with an average sequencing coverage of 10.4% for Switzerland, 38.2% for Denmark, and 6.4% for Germany over 2021. The number of clusters and the mean of the cluster size also varied by month (Fig 3).

We applied the model to the size distribution of identical sequence clusters to obtain monthly estimates of R_e, k, and the testing probability for SARS-CoV-2 in Switzerland, Denmark, and Germany (Fig 4). The estimated R_e values per month for each country across 2021 generally remained below 1, though the values did fluctuate through the year (Fig 4A). The average R_e value for Switzerland, Denmark, and Germany was 0.89 (95% credible interval, CI: 0.76-1.00), 1.04 (95% CI: 0.88-1.16), and 0.92 (95% CI: 0.83-1.00), respectively. The estimates of R_e come with narrow credible intervals. This might be due to the structure of the posterior probability, being highly concentrated around the most likely estimate. Due to the relatively large time window of one month, the underlying trends in R_e as estimated based on laboratory-confirmed cases were not captured consistently using our new method based on identical sequence clusters, notably at time points with transitions between SARS-CoV-2 variants.

Our estimates of the dispersion parameter k varied between the different countries, with average estimates over the 12 month period of 0.17 (95% CI: 0.02-0.30) for Switzerland, 0.38 (95% CI: 0.13-0.60) for Denmark, and 0.15 (95% CI: 0.02-0.27) for Germany. Estimates of k were highest in Denmark, fluctuating around 0.3 to 0.5 from January to October before dropping at the end of the year. Depending on the country and month, the estimates of k have narrower or wider credible intervals. In Switzerland and Germany, the estimates of k were generally lower with values around 0.1 to 0.3 and also dropped at the end of the year. The transition from Alpha to Delta was not associated with a substantial change in the estimates of k.

The estimated testing probability was relatively high and came with more uncertainty. This indicates that the testing probability cannot be precisely estimated from the size distribution of identical sequence clusters using our Bayesian inference model. Thus, we conducted a sensitivity analysis where the testing probability was fixed at 57.6% from January to August 2021 and at 35% from September to December 2021, based on estimates of the case ascertainment using swab positivity in England [49]. This did impact R_e and k estimates to a small degree, generally increasing R_e estimates

medRxiv preprint doi: https://doi.org/10.1101/2024.05.26.24307940; this version posted May 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

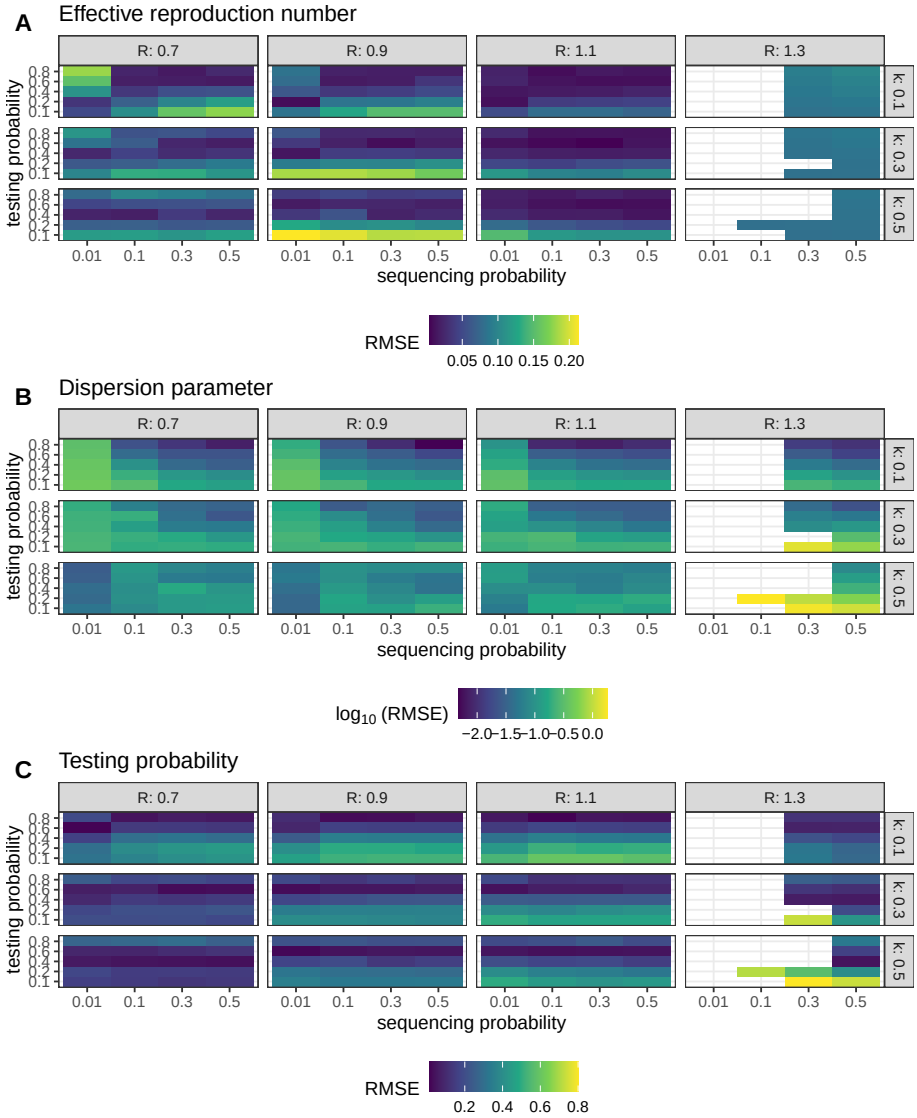

Fig 2. Simulation study of the Bayesian model to estimate R_e and k from the size distribution of identical sequence clusters. A: Error in estimating the effective reproduction number R_e. B: Error in estimating the dispersion parameter k. Error is measured by the root mean squared error (RMSE) between the true value and the estimated mean of the posterior distribution. For each parameter combination, we ran the model 10 times on 3,000 simulated clusters each.

slightly and decreasing k estimates slightly (Supplementary Material S1 Text, section 7).

Discussion

We developed a novel Bayesian model to estimate the effective reproduction number, R_e, and crucially the level of superspreading of an epidemic from viral sequence data, k. To this end, we introduced a new metric, the *genomic* reproduction number, R_g, the mean number of secondary cases created by an infected individual that share the same viral genome. R_g is therefore the analogue of R_e in the context of clusters of identical

medRxiv preprint doi: https://doi.org/10.1101/2024.05.26.24307940; this version posted May 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

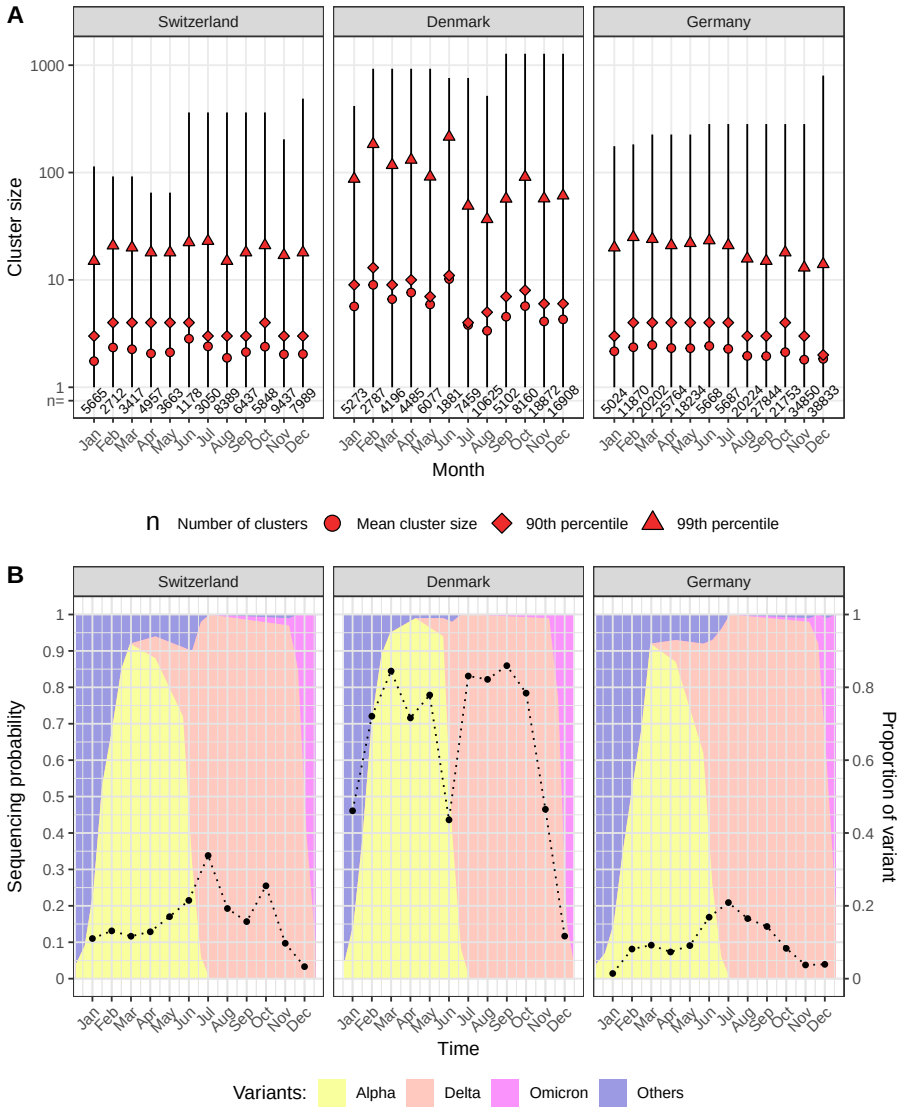

Fig 3. Identical sequence cluster size distribution and sequencing coverage for SARS-CoV-2 in Switzerland, Denmark, and Germany in 2021.
A: Range, mean, 90th percentile and 99th percentile of the identical sequence cluster size distribution and number of clusters by month based on data from GISAID.
B: Sequencing coverage (dots) and proportion of SARS-CoV-2 variants (background colour) by month.

sequences. We applied our model to the size distribution of identical sequence clusters of SARS-CoV-2 in Switzerland, Denmark, and Germany in 2021. We obtained monthly estimates of R_e around or below the epidemic threshold of 1. The estimates of the dispersion parameter k varied substantially by country and month and were typically between 0.1 and 0.5. Together, our study illustrates how the increasing amount of viral sequence data can be used to inform epidemiological parameters of SARS-CoV-2 and potentially other pathogens.

The main strength of our study is the presentation of a new method to estimate epidemiological parameters, specifically the overdispersion in the number of secondary

medRxiv preprint doi: https://doi.org/10.1101/2024.05.26.24307940; this version posted May 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

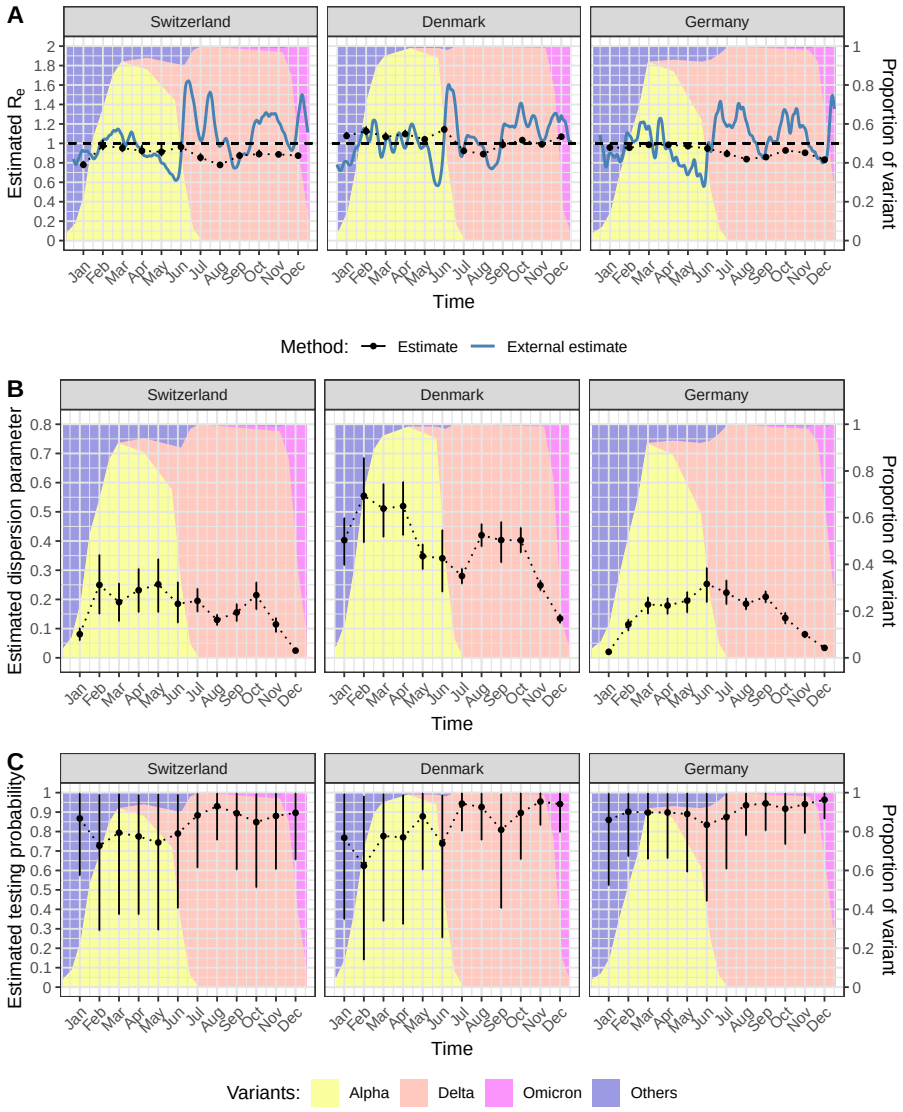

Fig 4. Parameter estimates based on the size distribution of identical SARS-CoV-2 sequence clusters in Switzerland, Denmark and Germany in 2021. A: Effective reproduction number R_e. B: Dispersion parameter k. C: Testing probability. The estimates are based on monthly time windows of identical sequence clusters. For each month are shown the mean posterior distribution and the 95% credible interval (in black). R_e values are compared to external estimates based on laboratory-confirmed cases (in blue, from github.com/covid-19-Re) [7])

cases, from viral sequence data alone. While R_e can often be reliably estimated from counts of laboratory-confirmed cases as well, estimating k has historically been considerably more challenging since it requires either direct information about transmission chains or clusters. With the novel method presented here, this challenge can be overcome through making use of the large number of sequences available for SARS-CoV-2. In addition, our method uses a computationally efficient way to identify clusters of identical sequences without the need to reconstruct very large phylogenies. Finally, the Bayesian inference model allows a full propagation of the uncertainty in the

medRxiv preprint doi: https://doi.org/10.1101/2024.05.26.24307940; this version posted May 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

viral transmission parameters, the mutation rate, and the incomplete case-detection.

Our study comes with a number of limitations. First, the presented method requires a large amount of sequence data and a relatively high case ascertainment to obtain reliable estimates of epidemiological parameters. We expect such data sets to become more common with further technological developments in diagnostics and sequencing, in particular for outbreaks and epidemics of emerging infectious diseases. Second, miscalls in sequences, such as reversions to a reference sequence, might have resulted in some clusters that are artificially separated or combined. We countered this problem to a certain extent by replacing variant-defining sequences with the variant name. For the remaining sequences, we expect that miscalls would have only little impact on the overall cluster size distribution and the parameter estimates. Third, we assumed an independent observation process, i.e., infected cases are detected randomly and tested cases are sequenced at random. However, testing and sequencing uptake might cluster in certain settings, such as during contact tracing and outbreak investigations. Lastly, we did not consider changes in the mutation rate and the generation time during the study period, which could have led to different ratios between R_e and R_g, especially for different SARS-CoV-2 variants.

Our estimates of R_e were not able to fully capture the observed changes when R_e is estimated from laboratory-confirmed cases. This could be a result of our assumption to assume a fixed time window (months) for which we estimated averaged epidemiological parameters during an ongoing epidemic. This introduces potential biases due to left and right censoring. On the one hand, we partly circumvented this problem by assigning identical sequence clusters to a given month if at least one sequence was sampled during that month. On the other hand, this has the disadvantage that we included clusters in our analysis that can span much longer time periods than one month and that cover different epidemic phases with widely different values of R_e. Furthermore, the transition periods from one SARS-CoV-2 variant to another pose an additional challenge to our inference method, as by definition sequence clusters cannot span over these transitions. Thus, it is not surprising that our method cannot capture the rapid changes of R_e and its rise after the arrival of new variants, such as the growth of Delta and Omicron in June and December 2021, respectively. The issue with left and right censoring also means that our method cannot reliably estimate R_e when the genomic reproduction number R_g is above 1, i.e., when some clusters can grow - in theory - indefinitely. Together, these limitations can explain why our estimates of R_e seemed to fluctuate around the epidemic threshold of 1, which corresponds to the long-term average of an ongoing epidemic.

In contrast, the estimates of k (range: 0.15-0.38) that we obtained with our new method fit well within the range of previous estimates for SARS-CoV-2 [18–32]. Interestingly, k was considerably higher in Denmark compared to Switzerland and Germany, indicating fewer superspreading events. While Switzerland and Germany had similar levels of testing uptake and laboratory-confirmed cases during most of 2021, the testing uptake in Denmark was considerably higher while the number of laboratory-confirmed cases was similar to the other countries. This suggests a higher case ascertainment and a better control of SARS-CoV-2 transmission in Denmark, which could have led to fewer opportunities for superspreading events and, as a consequence, higher values of k. Using a similar methodology, Tran-Kiem and Bedford [39] estimated k at 0.63 (95% confidence interval: 0.34-1.56) in New Zealand, which is somewhat higher than our estimates. Again, this can be explained by the fact that they analyzed sequence clusters during a period when there was a high level of transmission control with little opportunities for superspreading events. Lastly, it remains unclear how our estimates of k were affected by the left and right censoring of clusters, but the relatively low estimates in January and December 2021 could be

medRxiv preprint doi: https://doi.org/10.1101/2024.05.26.24307940; this version posted May 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

impacted by the early spread of Alpha and Omicron. Another potential explanation is that superspreading events may favoured by indoor gathering during winter periods.

The vast amount of genomic data that was generated during the pandemic provides novel opportunities to characterize and track the transmission dynamics of SARS-CoV-2. Our newly developed Bayesian model to infer epidemiological parameters from the size distribution of identical sequence clusters can reliably estimate the level of superspreading through the dispersion parameter k and has the potential to inform about R_e, albeit with some practical limitations for the latter. With the increasing affordability and ease of sequencing, we expect that large volumes of sequence data will become more readily available in the future, making it possible to adapt this method to other pathogens and to estimate the transmission heterogeneity in different countries, for different variants, and over time.

Supporting information

S1 Text. Additional results, figures and tables.

S1 Table. GISAID EPI_SET Table 1 - Switzerland

S2 Table. GISAID EPI_SET Table 2 - Denmark

S3 Table. GISAID EPI_SET Table 3 - Germany

Acknowledgments

We gratefully acknowledge all data contributors, i.e., the authors and their originating laboratories responsible for obtaining the specimens, and their submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. We also gratefully acknowledge and thank all labs around the world that have collected and shared SARS-CoV-2 sequences we used in our study, and in particular the labs in Switzerland, Germany, and Denmark, where the sequence data for this study originated. A complete list of the labs that generated the data we used from GISAID can be found in at the EPI_SETs.

We additionally thank Pierre-Yves Boëlle for valuable feed-back on the project and for the suggestion to validate the compatibility of the simulation of identical sequences with cluster data from Switzerland, Denmark and Germany by a posterior predictive check. We also thank Eugenio Valdano for a stimulating discussion of the project and his advice on how to assign clusters of identical sequences to months. Calculations were performed on UBELIX (www.id.unibe.ch/hpc), the HPC cluster at the University of Bern.

Funding

EH, RH, and CA were supported or received funding by the Swiss National Science Foundation (No 196046). MW, JR, and CA were supported or received funding by the Multidisciplinary Center for Infectious Diseases, University of Bern, Bern, Switzerland. JR was supported by the Swiss National Science Foundation (No 189498). CA received funding from the European Union's Horizon 2020 research and innovation program - project EpiPose (No 101003688). This project was supported by the ESCAPE project (101095619), funded by the European Union. Views and opinions expressed are however

medRxiv preprint doi: https://doi.org/10.1101/2024.05.26.24307940; this version posted May 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them. This work was funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (grant number 10051037). This work has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00482. EH was supported by a Swiss National Science Foundation Starting Grant (TMSGI3_211225).

Data and Code Availability

Sequence data are available via GISAID after registration, and are available in EPI_SET_240326pm (doi.org/10.55876/gis8.240326pm) for Switzerland, EPI_SET_240326mz (doi.org/10.55876/gis8.240326mz) for Denmark, and EPI_SET_240326uh (doi.org/10.55876/gis8.240326uh) for Germany; see also the supplementary tables 1-3.

Code to generate identical sequence clusters from the starting alignments is available via github.com/emmahodcroft/sc2_rk_public. Functions for estimation of parameters and simulation of identical sequence clusters are available via the R package estRodis github.com/mwohlfender/estRodis. Code used for the analysis of data and results as well as the creation of plots and tables is available via github.com/mwohlfender/R_overdispersion_cluster_size

References

1. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance. 2017;22(13):30494. doi:10.2807/1560-7917.ES.2017.22.13.30494.

2. Pan A, Liu L, Wang C, Guo H, Hao X, Wang Q, et al. Association of Public Health Interventions With the Epidemiology of the COVID-19 Outbreak in Wuhan, China. JAMA. 2020;323(19):1915–1923. doi:10.1001/jama.2020.6130.

3. D'Arienzo M, Coniglio A. Assessment of the SARS-CoV-2 basic reproduction number, $R0$, based on the early phase of COVID-19 outbreak in Italy. Biosafety and Health. 2020;2(2):57–59. doi:10.1016/j.bsheal.2020.03.004.

4. Abbott S, Hellewell J, Thompson R, Sherratt K, Gibbs H, Bosse N, et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts [version 2; peer review: 1 approved, 1 approved with reservations]. Wellcome Open Research. 2020;5(112). doi:10.12688/wellcomeopenres.16006.2.

5. Miller D, Martin MA, Harel N, Tirosh O, Kustin T, Meir M, et al. Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. Nature Communications. 2020;11(1):5518. doi:10.1038/s41467-020-19248-0.

6. Lai A, Bergna A, Acciarri C, Galli M, Zehender G. Early phylogenetic estimate of the effective reproduction number of SARS-CoV-2. Journal of Medical Virology. 2020;92(6):675–679. doi:10.1002/jmv.25723.

7. Huisman JS, Scire J, Angst DC, Li J, Neher RA, Maathuis MH, et al. Estimation and worldwide monitoring of the effective reproductive number of SARS-CoV-2. eLife. 2022;11:e71345. doi:10.7554/eLife.71345.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.26.24307940; this version posted May 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

8. Nishiura H, Yan P, Sleeman CK, Mode CJ. Estimating the transmission potential of supercritical processes based on the final size distribution of minor outbreaks. Journal of Theoretical Biology. 2012;294:48–55. doi:10.1016/j.jtbi.2011.10.039.

9. Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proceedings of the National Academy of Sciences. 2013;110(1):228–233. doi:10.1073/pnas.1207965110.

10. Volz E, Pond S. Phylodynamic Analysis of Ebola Virus in the 2014 Sierra Leone Epidemic. PLoS Currents. 2014;6:ecurrents.outbreaks.6f7025f1271821d4c815385b08f5f80e. doi:10.1371/currents.outbreaks.6f7025f1271821d4c815385b08f5f80e.

11. Turk T, Bachmann N, Kadelka C, Böni J, Yerly S, Aubert V, et al. Assessing the danger of self-sustained HIV epidemics in heterosexuals by population based phylogenetic cluster analysis. eLife. 2017;6:e28721. doi:10.7554/eLife.28721.

12. Wegehaupt O, Endo A, Vassall A. Superspreading, overdispersion and their implications in the SARS-CoV-2 (COVID-19) pandemic: a systematic review and meta-analysis of the literature. BMC Public Health. 2023;23(1):1003. doi:10.1186/s12889-023-15915-1.

13. Goyal A, Reeves DB, Cardozo-Ojeda EF, Schiffer JT, Mayer BT. Viral load and contact heterogeneity predict SARS-CoV-2 transmission and super-spreading events. eLife. 2021;10:e63537. doi:10.7554/eLife.63537.

14. Chen PZ, Koopmans M, Fisman DN, Gu FX. Understanding why superspreading drives the COVID-19 pandemic but not the H1N1 pandemic. The Lancet Infectious Diseases. 2021;21(9):1203–1204. doi:10.1016/S1473-3099(21)00406-0.

15. Chen PZ, Bobrovitz N, Premji Z, Koopmans M, Fisman DN, Gu FX. Heterogeneity in transmissibility and shedding SARS-CoV-2 via droplets and aerosols. eLife. 2021;10:e65774. doi:10.7554/eLife.65774.

16. Wang J, Chen X, Guo Z, Zhao S, Huang Z, Zhuang Z, et al. Superspreading and heterogeneity in transmission of SARS, MERS, and COVID-19: A systematic review. Computational and Structural Biotechnology Journal. 2021;19:5039–5046. doi:10.1016/j.csbj.2021.08.045.

17. Gómez-Carballa A, Pardo-Seco J, Bello X, Martinón-Torres F, Salas A. Superspreading in the emergence of COVID-19 variants. Trends in Genetics. 2021;37(12):1069–1080. doi:10.1016/j.tig.2021.09.003.

18. Sun K, Wang W, Gao L, Wang Y, Luo K, Ren L, et al. Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. Science. 2021;371(6526):eabe2424. doi:10.1126/science.abe2424.

19. Adam DC, Wu P, Wong JY, Lau EHY, Tsang TK, Cauchemez S, et al. Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. Nature Medicine. 2020;26(11):1714–1719. doi:10.1038/s41591-020-1092-0.

20. Bi Q, Wu Y, Mei S, Ye C, Zou X, Zhang Z, et al. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. The Lancet Infectious Diseases. 2020;20(8):911–919. doi:10.1016/S1473-3099(20)30287-5.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.26.24307940; this version posted May 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

21. He D, Zhao S, Xu X, Lin Q, Zhuang Z, Cao P, et al. Low dispersion in the infectiousness of COVID-19 cases implies difficulty in control. BMC Public Health. 2020;20(1):1558. doi:10.1186/s12889-020-09624-2.

22. Hasan A, Susanto H, Kasim MF, Nuraini N, Lestari B, Triany D, et al. Superspreading in early transmissions of COVID-19 in Indonesia. Scientific Reports. 2020;10(1):22386. doi:10.1038/s41598-020-79352-5.

23. Kwok KO, Chan HHH, Huang Y, Hui DSC, Tambyah PA, Wei WI, et al. Inferring super-spreading from transmission clusters of COVID-19 in Hong Kong, Japan, and Singapore. Journal of Hospital Infection. 2020;105(4):682–685. doi:10.1016/j.jhin.2020.05.027.

24. Lau MSY, Grenfell B, Thomas M, Bryan M, Nelson K, Lopman B. Characterizing superspreading events and age-specific infectiousness of SARS-CoV-2 transmission in Georgia, USA. Proceedings of the National Academy of Sciences. 2020;117(36):22430–22435. doi:10.1073/pnas.2011802117.

25. Tariq A, Lee Y, Roosa K, Blumberg S, Yan P, Ma S, et al. Real-time monitoring the transmission potential of COVID-19 in Singapore, March 2020. BMC Medicine. 2020;18(1):166. doi:10.1186/s12916-020-01615-9.

26. Zhao S, Shen M, Musa SS, Guo Z, Ran J, Peng Z, et al. Inferencing superspreading potential using zero-truncated negative binomial model: exemplification with COVID-19. BMC Medical Research Methodology. 2021;21(1):30. doi:10.1186/s12874-021-01225-w.

27. Zhang Y, Li Y, Wang L, Li M, Zhou X. Evaluating Transmission Heterogeneity and Super-Spreading Event of COVID-19 in a Metropolis of China. International Journal of Environmental Research and Public Health. 2020;17(10):3705. doi:10.3390/ijerph17103705.

28. Shi Q, Hu Y, Peng B, Tang XJ, Wang W, Su K, et al. Effective control of SARS-CoV-2 transmission in Wanzhou, China. Nature Medicine. 2021;27(1):86–93. doi:10.1038/s41591-020-01178-5.

29. James A, Plank MJ, Hendy S, Binny RN, Lustig A, Steyn N. Model-free estimation of COVID-19 transmission dynamics from a complete outbreak. PLOS ONE. 2021;16(3):e0238800. doi:10.1371/journal.pone.0238800.

30. Kremer C, Torneri A, Boesmans S, Meuwissen H, Verdonschot S, Vanden Driessche K, et al. Quantifying superspreading for COVID-19 using Poisson mixture distributions. Scientific Reports. 2021;11(1):14107. doi:10.1038/s41598-021-93578-x.

31. Endo A, Abbott S, Kucharski AJ, Funk S. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. Wellcome Open Research. 2020;5:67. doi:10.12688/wellcomeopenres.15842.3.

32. Riou J, Althaus CL. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. Eurosurveillance. 2020;25(4):2000058. doi:10.2807/1560-7917.ES.2020.25.4.2000058.

33. Susswein Z, Bansal S. Characterizing superspreading of SARS-CoV-2 : from mechanism to measurement; 2020. Available from: `https://www.medrxiv.org/content/10.1101/2020.12.08.20246082v1`.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.26.24307940; this version posted May 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

34. Blumberg S, Lloyd-Smith JO. Inference of R0 and Transmission Heterogeneity from the Size Distribution of Stuttering Chains. PLOS Computational Biology. 2013;9(5):e1002993. doi:10.1371/journal.pcbi.1002993.

35. Breban R, Riou J, Fontanet A. Interhuman transmissibility of Middle East respiratory syndrome coronavirus: estimation of pandemic risk. The Lancet. 2013;382(9893):694–699. doi:10.1016/S0140-6736(13)61492-0.

36. Kucharski AJ, Althaus CL. The role of superspreading in Middle East respiratory syndrome coronavirus (MERS-CoV) transmission. Eurosurveillance. 2015;20(25):21167. doi:10.2807/1560-7917.ES2015.20.25.21167.

37. Cauchemez S, Nouvellet P, Cori A, Jombart T, Garske T, Clapham H, et al. Unraveling the drivers of MERS-CoV transmission. Proceedings of the National Academy of Sciences. 2016;113(32):9081–9086. doi:10.1073/pnas.1519235113.

38. Blumberg S, Lloyd-Smith JO. Comparing methods for estimating $R0$ from the size distribution of subcritical transmission chains. Epidemics. 2013;5(3):131–145. doi:10.1016/j.epidem.2013.05.002.

39. Tran-Kiem C, Bedford T. Estimating the reproduction number and transmission heterogeneity from the size distribution of clusters of identical pathogen sequences. Proceedings of the National Academy of Sciences. 2024;121(15):e2305299121. doi:10.1073/pnas.2305299121.

40. Nextstrain Team. ncov-ingest pipeline; 2024. Available from: `https://github.com/nextstrain/ncov-ingest`.

41. Aksamentov I, Roemer C, Hodcroft EB, Neher RA. Nextclade: clade assignment, mutation calling and quality control for viral genomes. Journal of Open Source Software. 2021;6(67):3773. doi:10.21105/joss.03773.

42. Hodcroft EB, Aksamentov I, Zuber M, Neher RA. CoVariants; 2024. Available from: `https://covariants.org/`.

43. Federal Department of Home Affairs FDHA, Federal Office of Public Health, Prevention and Public Health Services Directorate, Communicable Diseases Division. Infectious Diseases Dashboard (IDD); 2024. Available from: `https://idd.bag.admin.ch/portal-data`.

44. Statens Serum Institut. Covid-19 - Danmark (region); 2024. Available from: `https://experience.arcgis.com/experience/220fef27d07d438889d651cc2e00076c/page/Covid-19-Regionalt/`.

45. Robert Koch-Institut. 7-Tage-Inzidenz der COVID-19-Fälle in Deutschland; 2024. Available from: `https://robert-koch-institut.github.io/COVID-19_7-Tage-Inzidenz_in_Deutschland/`.

46. Neher RA. Contributions of adaptation and purifying selection to SARS-CoV-2 evolution. Virus Evolution. 2022;8(2):veac113. doi:10.1093/ve/veac113.

47. Ganyani T, Kremer C, Chen D, Torneri A, Faes C, Wallinga J, et al. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. Eurosurveillance. 2020;25(17). doi:https://doi.org/10.2807/1560-7917.ES.2020.25.17.2000257.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.26.24307940; this version posted May 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

48. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A Probabilistic Programming Language. Journal of statistical software. 2017;76:1. doi:10.18637/jss.v076.i01.

49. Eales O, Haw D, Wang H, Atchison C, Ashby D, Cooke GS, et al. Dynamics of SARS-CoV-2 infection hospitalisation and infection fatality ratios over 23 months in England. PLOS Biology. 2023;21(5):e3002118. doi:10.1371/journal.pbio.3002118.