Supplementary information for:

Estimating the reproduction number and transmission heterogeneity from the size distribution of clusters of identical pathogen sequences

Cécile Tran-Kiem[1], Trevor Bedford[1,2]

1. Vaccine and Infectious Diseases Division, Fred Hutchinson Cancer Center, Seattle, WA, USA
2. Howard Hugues Medical Institute, Seattle, WA, USA

Table S1: Estimates of the probability that transmission occurs before mutation for different pathogens along assumptions for the generation time distribution and the mutation rate used for the estimation.

Pathogen	Generation time		Mutation rate	Genome length (if relevant)	Probability that transmission occurs before mutation (estimated)
	Mean (days)	Standard deviation (days)			
MERS-CoV	6.8 (1)	6.3 (1)	$4.59 \cdot 10^{-4}$ subs/site/year (2)	30130	0.79
Measles virus	11.2 (3)	1.8 (3)	$4.97 \cdot 10^{-4}$ subs/site/year (5)	15894	0.79
Ebola virus	14.4 (5)	8.9 (5)	$3.10 \cdot 10^{-6}$ subs/site/day (5)	18958	0.48
Zika virus	20.0 (6)	7.4 (6)	$1.12 \cdot 10^{-3}$ subs/site/year (7)	10274	0.55
Mpox virus (2022-2023 outbreak)	12.5 (8)	5.7 (8)	$6.38 \cdot 10^{-5}$ subs/site/year (9)	197209	0.66
Influenza A (H1N1)	3.0 (9)	1.5 (5)	$3.41 \cdot 10^{-3}$ subs/site/year (10)	13152	0.70
Mumps virus	21.0	15.7	$4.35 \cdot 10^{-4}$ subs/site/year (11)	15384	0.71
	Assuming an exponentially distributed incubation period of 14 days followed by an exponentially distributed infectious duration of 7 days (12,13)				
RSV-A	13.0	9.8	$6.47 \cdot 10^{-4}$ subs/site/year (14,15)	15200	0.73
	Using the parametrization used in (13,16)				
Dengue virus (I)	18.2 (17)	6.1 (17)	$6.21 \cdot 10^{-4}$ subs/site/year (18)	11000	0.72
SARS-CoV	8.7 (17)	3.6 (18)	$1.14 \cdot 10^{-5}$ subs/site/day (18)	29714	0.09
SARS-CoV-2 (pre-Omicron)	5.9 (19)	4.8 (19)	26.6 subs/year (20)	-	0.69
SARS-CoV-2 (Omicron)	4.9 (1 day shorter than pre-Omicron) (21,22)	4.8			0.74

Table S2: Parameter estimates for MERS. Maximum likelihood estimates are reported along 50% and 95% confidence intervals (CI).

Proportion of infections detected	Reproduction number R estimate	Dispersion parameter k estimate
1.0	0.57 50%CI: (0.53-0.61) 95%CI: (0.45-0.70)	0.14 50%CI: (0.09-0.20) 95%CI: (0.04-0.46)
0.5	0.61 50%CI: (0.60-0.68) 95%CI: (0.53-0.76)	0.09 50%CI: (0.07-0.13) 95%CI: (0.03-26)

Table S3: Parameter estimates for measles. Maximum likelihood estimates are reported along 50% and 95% confidence intervals (CI).

Proportion of infections detected	Reproduction number R estimate	Dispersion parameter k estimate
1.0	0.57 50%CI: (0.46-0.72) 95%CI: (0.29-1.15)	0.04 50%CI: (0.016-0.092) 95%CI: (0.003-0.45)
0.5	0.61 50%CI: (0.49-0.75) 95%CI: (0.32-1.15)	0.02 50%CI: (0.009-0.05) 95%CI: (0.002-0.19)

Table S4: Parameter estimates for SARS-CoV-2 in New Zealand. Maximum likelihood estimates are reported along 50% and 95% confidence intervals (CI).

Proportion of infections detected	Period	Reproduction number R estimate	Dispersion parameter k estimate
1.0	April – May 2020	0.76 95%CI: (0.60-0.94) 50%CI: (0.71-0.82)	0.64 95%CI: (0.34-1.57) 50%CI: (0.51-0.83)
	June – December 2020	0.81 95%CI: (0.70-0.94) 50%CI: (0.77-0.85)	
	January – April 2021	0.69 95%CI: (0.57-0.84) 50%CI: (0.65-0.74)	
	May – July 2021	0.61 95%CI: (0.45-0.81) 50%CI: (0.55-0.67)	
0.8	April – May 2020	0.81 95%CI: (0.66-0.98) 50%CI: (0.76-0.87)	0.63 95%CI: (0.33-1.56) 50%CI: (0.49-0.82)
	June – December 2020	0.86 95%CI: (0.75-0.98) 50%CI: (0.82-0.89)	
	January – April 2021	0.74 95%CI: (0.62-0.88) 50%CI: (0.70-0.78)	
	May – July 2021	0.66 95%CI: (0.50-0.86) 50%CI: (0.61-0.72)	
0.5	April – May 2020	0.91 95%CI: (0.76-1.06) 50%CI: (0.86-0.96)	0.60 95%CI: (0.31-1.45) 50%CI: (0.47-0.78)
	June – December 2020	0.95 95%CI: (0.85-1.05) 50%CI: (0.92-0.98)	
	January – April 2021	0.84 95%CI: (0.72-0.97) 50%CI: (0.80-0.88)	
	May – July 2021	0.77 95%CI: (0.61-0.95) 50%CI: (0.71-0.82)	

Table S5: Definitions of the study periods for the Washington state SARS-CoV-2 analysis.

Variant under study	Start of the time window of interest	End of the time window of interest	Corresponding Nextstrain clades
D614G	April 1st, 2020	June 1st, 2020	19A, 19B, 19C
Epsilon	January 1st, 2021	February 15th, 2021	21C (Epsilon)
Alpha	March 1st, 2021	April 1st, 2021	20I (Alpha, V1)
Delta	June 1st, 2021	July 1st, 2021	21A (Delta), 21I (Delta), 21J (Delta)
Omicron (BA.1)	November 15th, 2021	December 15th, 2021	21K (Omicron)
Omicron (BA.2)	February 1st, 2022	March 1st, 2022	21L (Omicron)
Omicron (BA.4, BA.5)	May 1st, 2022	May 15th, 2022	22A (Omicron), 22B (Omicron)

Table S6: Genbank accession numbers for measles sequences used in the analysis. All sequences were obtained from Pacenti et al. using the Nextstrain measles workflow (23,24).

Strain name	Accession number	URL
Padova.ITA/13.17/1/D8	MK513623	https://www.ncbi.nlm.nih.gov/nuccore/MK5136223
Padova.ITA/14.17/3/D8	MK513625	https://www.ncbi.nlm.nih.gov/nuccore/MK513625
Padova.ITA/16.17/4/B3	MK513613	https://www.ncbi.nlm.nih.gov/nuccore/MK513613
Padova.ITA/14.17/7/B3	MK513607	https://www.ncbi.nlm.nih.gov/nuccore/MK513607
Padova.ITA/16.17/2/B3	MK513611	https://www.ncbi.nlm.nih.gov/nuccore/MK513611
Padova.ITA/16.17/3/B3	MK513612	https://www.ncbi.nlm.nih.gov/nuccore/MK513612
Padova.ITA/14.17/2/D8	MK513624	https://www.ncbi.nlm.nih.gov/nuccore/MK513624
Padova.ITA/16.17/6/B3	MK513615	https://www.ncbi.nlm.nih.gov/nuccore/MK513615
Padova.ITA/20.17/1/B3	MK513619	https://www.ncbi.nlm.nih.gov/nuccore/MK513619
Padova.ITA/14.17/4/B3	MK513605	https://www.ncbi.nlm.nih.gov/nuccore/MK513605
Padova.ITA/13.17/1/B3	MK513600	https://www.ncbi.nlm.nih.gov/nuccore/MK513600
Padova.ITA/21.17/1/B3	MK513620	https://www.ncbi.nlm.nih.gov/nuccore/MK513620
Padova.ITA/19.17/1/B3	MK513617	https://www.ncbi.nlm.nih.gov/nuccore/MK513617
Padova.ITA/14.17/2/B3	MK513603	https://www.ncbi.nlm.nih.gov/nuccore/MK513603
Padova.ITA/14.17/5/B3	MK513606	https://www.ncbi.nlm.nih.gov/nuccore/MK513606
Padova.ITA/16.17/1/B3	MK513610	https://www.ncbi.nlm.nih.gov/nuccore/MK513610
Padova.ITA/13.17/2/B3	MK513601	https://www.ncbi.nlm.nih.gov/nuccore/MK513601
Venezia.ITA/22.17/3/D8	MK513627	https://www.ncbi.nlm.nih.gov/nuccore/MK513627
Padova.ITA/19.17/2/B3	MK513618	https://www.ncbi.nlm.nih.gov/nuccore/MK513618
Padova.ITA/11.17/1/B3	MK513598	https://www.ncbi.nlm.nih.gov/nuccore/MK513598
Padova.ITA/24.17/1/B3	MK513622	https://www.ncbi.nlm.nih.gov/nuccore/MK513622
Padova.ITA/15.17/1/B3	MK513608	https://www.ncbi.nlm.nih.gov/nuccore/MK513608
Padova.ITA/21.17/2/B3	MK513621	https://www.ncbi.nlm.nih.gov/nuccore/MK513621
Padova.ITA/14.17/1/B3	MK513602	https://www.ncbi.nlm.nih.gov/nuccore/MK513602
Padova.ITA/15.17/2/B3	MK513609	https://www.ncbi.nlm.nih.gov/nuccore/MK513609
Padova.ITA/14.17/3/B3	MK513604	https://www.ncbi.nlm.nih.gov/nuccore/MK513604
Padova.ITA/17.17/3/B3	MK513616	https://www.ncbi.nlm.nih.gov/nuccore/MK513616
Verona.ITA/19.17/2/D8	MK513626	https://www.ncbi.nlm.nih.gov/nuccore/MK513626
Padova.ITA/12.17/1/B3	MK513599	https://www.ncbi.nlm.nih.gov/nuccore/MK513599
Padova.ITA/16.17/5/B3	MK513614	https://www.ncbi.nlm.nih.gov/nuccore/MK513614

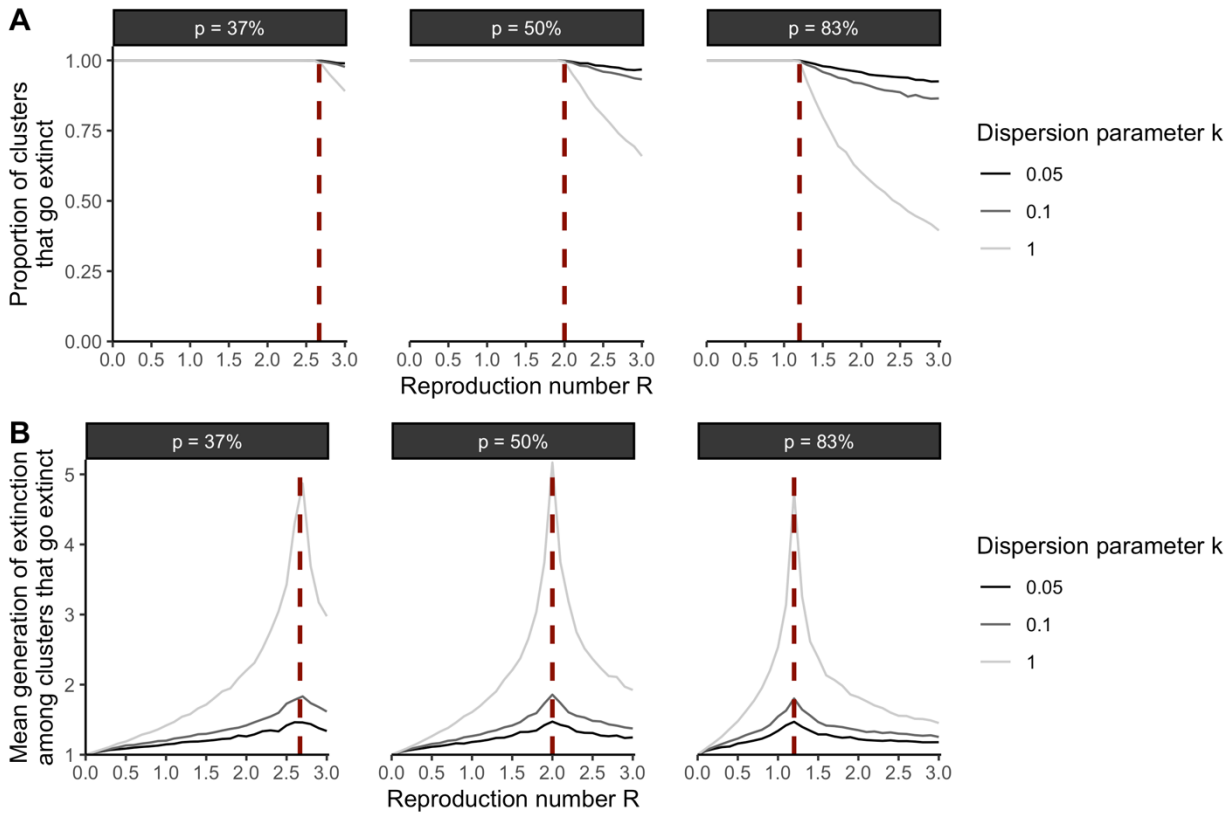

Figure S1: Dynamics of extinction for clusters of identical pathogen sequences. A. Proportion of clusters of identical sequences that go extinct as a function of the reproduction number R (x-axis) exploring different assumptions regarding the dispersion parameter k (colored lines) and the probability p that transmission occurs before mutation. **B.** Mean number of generations until cluster extinction (among clusters that go extinct) extinct as a function of the reproduction number R (x-axis) exploring different assumptions regarding the dispersion parameter k (colored lines) and the probability p that transmission occurs before mutation. The vertical red dashed lines correspond to the inverse of the probability p that transmission occurs before mutation.

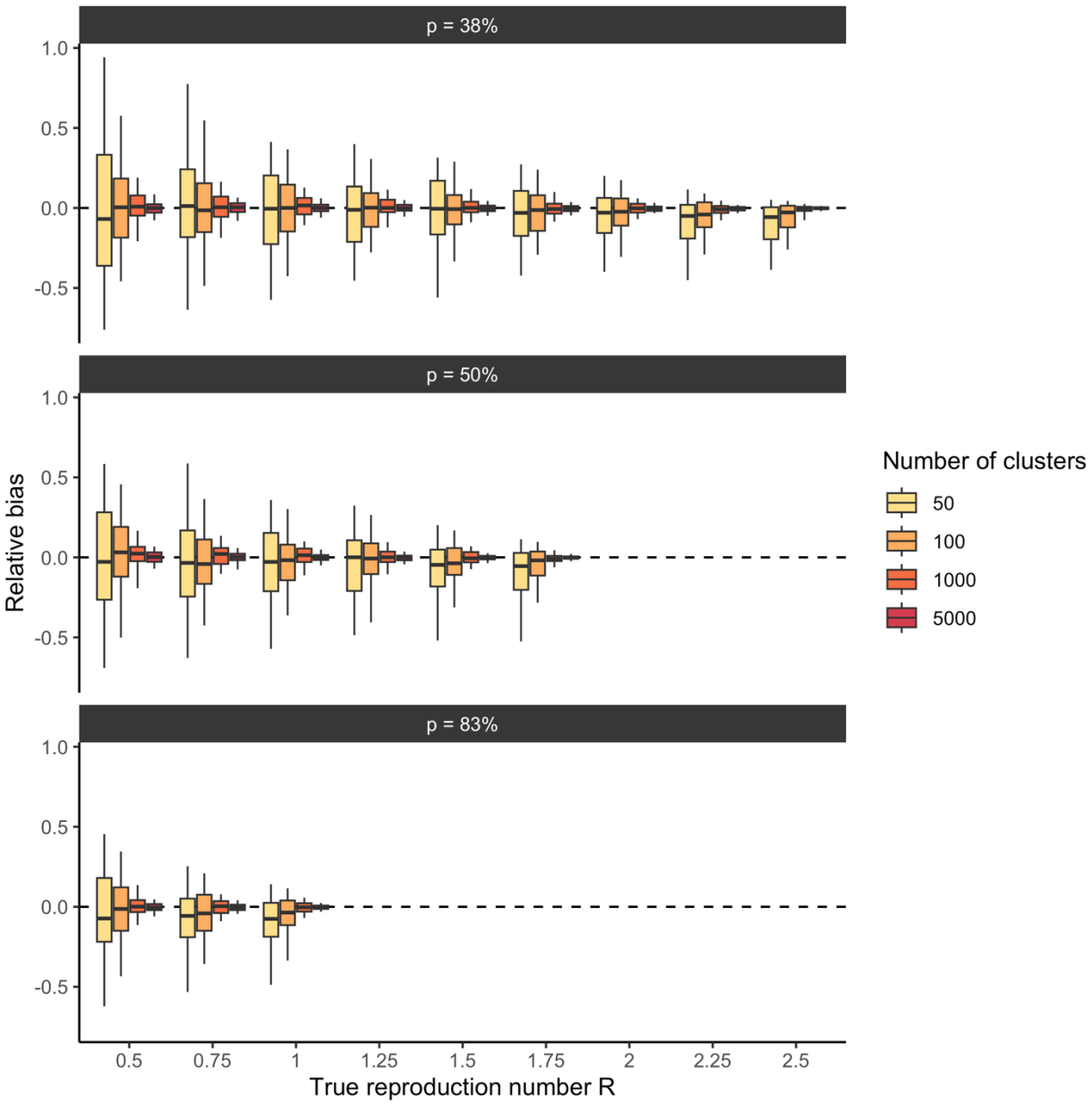

Figure S2: Relative bias on the reproduction number R estimate when the reproduction number lies below the threshold of 1/p. For each true value of the reproduction number R (x-axis) and value of the probability p that transmission occurs before mutation, the boxplot depicts the distribution of the relative bias across 100 simulations for different dataset sizes (colours). The relative bias is defined as $(R^{MLE} - R^{true})/R^{true}$ where R^{true} is the true reproduction number used to generate synthetic cluster data and R^{MLE} our maximum likelihood estimates. The simulations were run assuming that 50% of infections were sequenced. The boxplots represent the 2.5%, 25%, 50%, 75% and 97.5% percentiles.

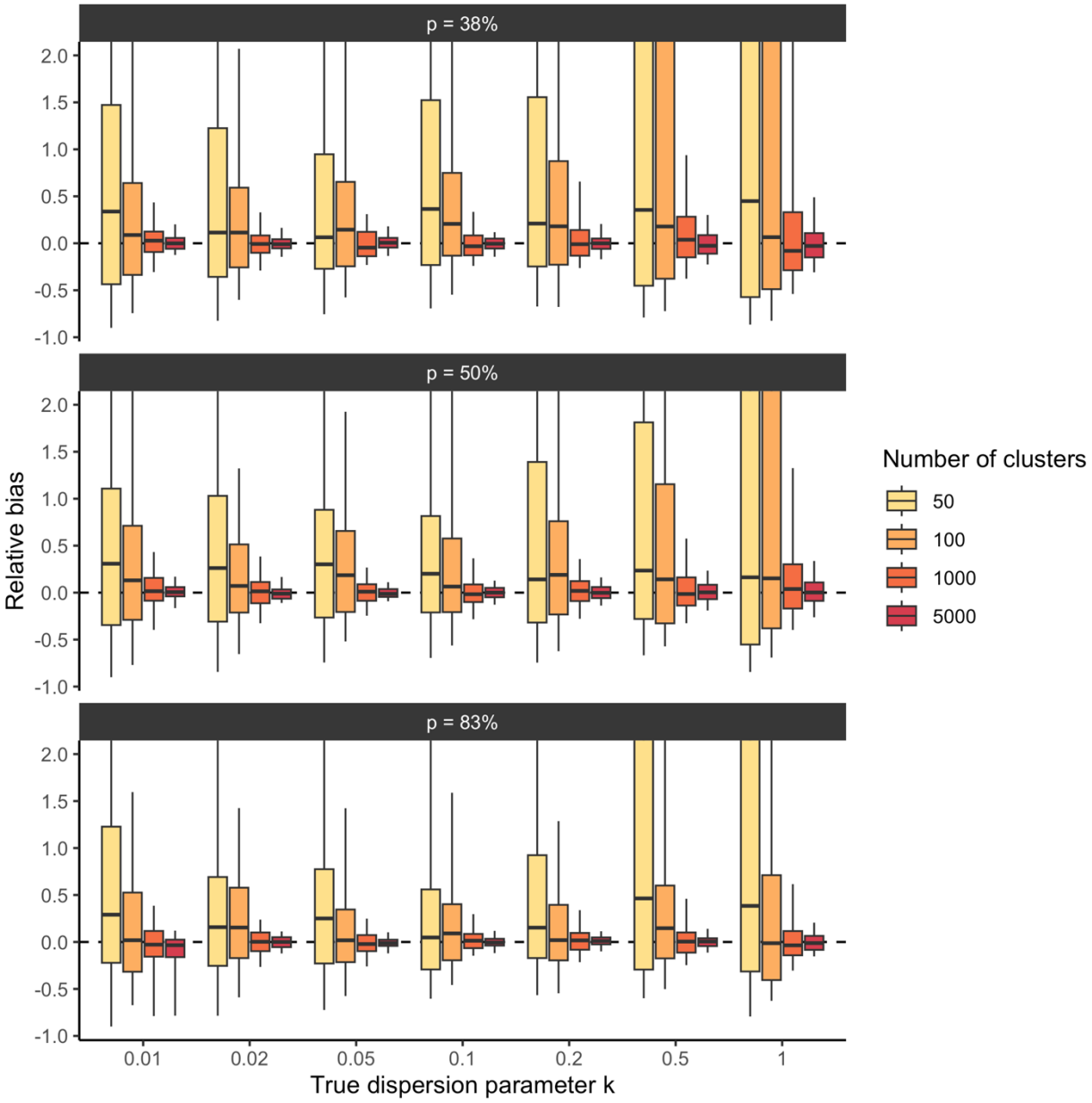

Figure S3: Relative bias on the dispersion parameter k estimate when the reproduction number lies below the threshold of 1/p. For each true value of the dispersion parameter k (x-axis) and value of the probability p that transmission occurs before mutation, the boxplot depicts the distribution of the relative bias across 100 simulations for different dataset sizes (colours). The relative bias is defined as $(k^{MLE} - k^{true})/k^{true}$ where k^{true} is the true dispersion parameter used to generate synthetic cluster data and k^{MLE} our maximum likelihood estimate. The simulations were run assuming that 50% of infections were sequenced and for a true reproduction number of 1.0. The y-axis was cropped at 2 to increase readability. The boxplots represent the 2.5%, 25%, 50%, 75% and 97.5% percentiles.

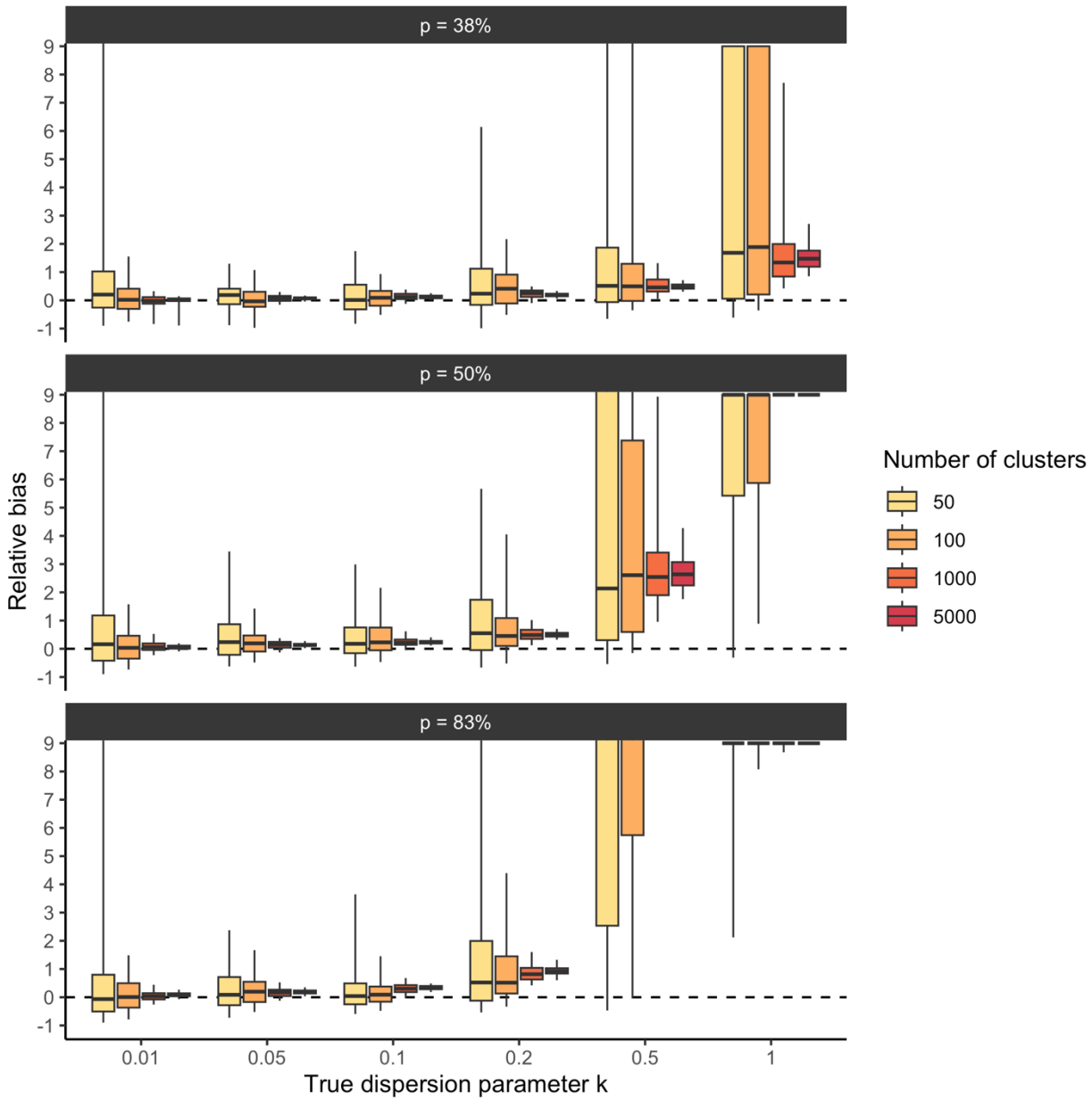

Figure S4: Relative bias on the dispersion parameter k estimate when the reproduction number lies above the threshold of 1/p. For each true value of the dispersion parameter k (x-axis) and value of the probability p that transmission occurs before mutation, the boxplot depicts the distribution of the relative bias across 100 simulations for different dataset sizes (colours). The relative bias is defined as $(k^{MLE} - k^{true})/k^{true}$ where k^{true} is the true dispersion parameter used to generate synthetic cluster data and k^{MLE} our maximum likelihood estimate. The simulations were run assuming that 50% of infections were sequenced and for a true reproduction number of 3.0. The y-axis was cropped at 9 to increase readability. The boxplots represent the 2.5%, 25%, 50%, 75% and 97.5% percentiles.

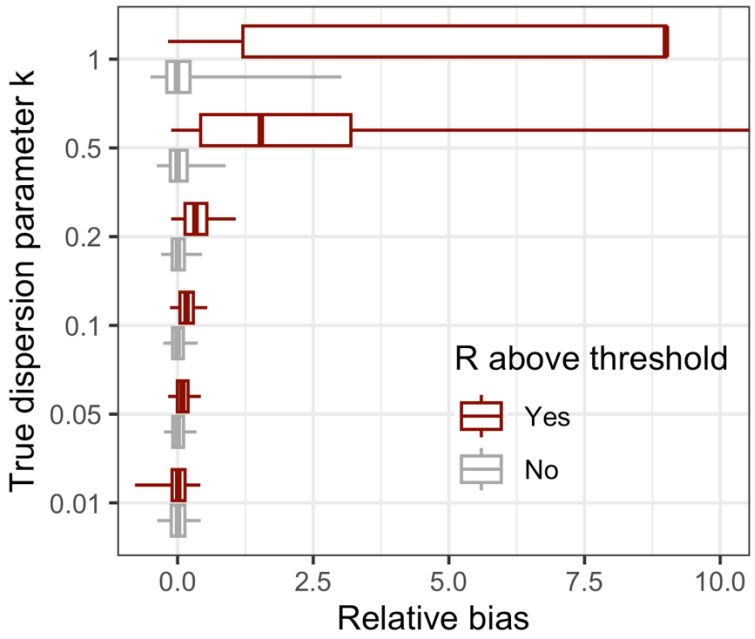

Figure S5: Impact of reaching the reproduction number threshold on dispersion parameter estimates. The relative bias is defined as $(k^{MLE} - k^{true})/k^{true}$ where k^{true} is the true dispersion parameter used to generate synthetic cluster data and k^{MLE} our maximum likelihood estimate. The boxplots depict the 2.5%, 25%, 50%, 75% and 97.5% percentiles of relative bias obtained across all the simulations we performed and that are detailed in the methods section.

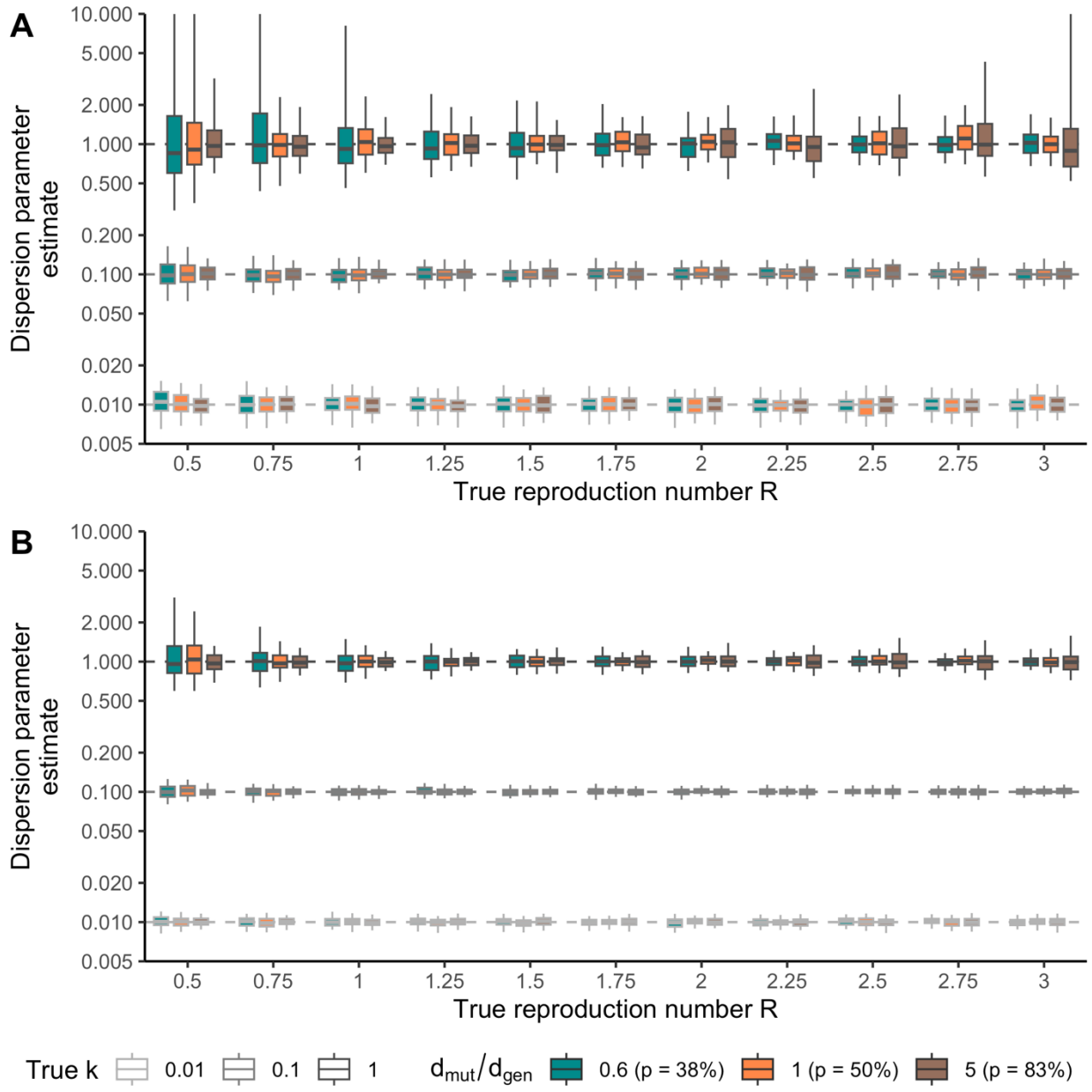

Figure S6: Dispersion parameter estimates as a function of the true reproduction number when using the inference framework relying on cluster size distribution conditional on cluster extinction. A. Using a dataset comprised of 1,000 clusters of identical sequences. **B.** Using a dataset comprised of 5,000 clusters of identical sequences. Each boxplot represents the distribution of k maximum likelihood estimates across 100 simulations (2.5%, 25%, 50%, 75% and 97.5% percentiles). We explored different values of the true dispersion parameter k (boxplot contour colours) and different values for the probability p that transmission occurs before transmission (boxplot filling). The fraction d_{mut}/d_{gen} corresponds to the ratio between the mean duration before the appearance of a mutation and the mean generation time. The correspondence between values of this fraction and of p is established assuming the generation time is exponentially distributed.

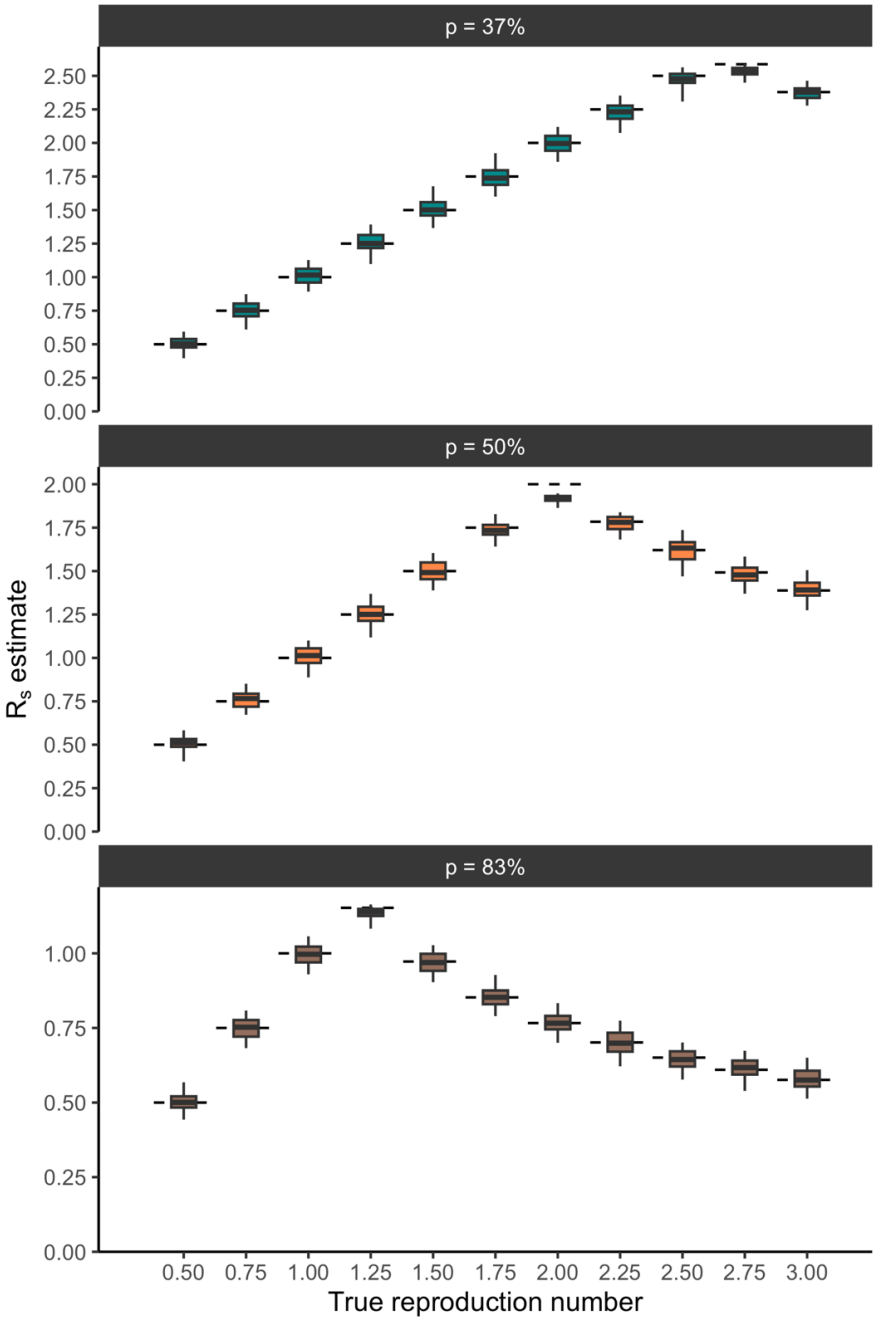

Figure S7: Subcritical reproduction number R_s estimates as a function of the true reproduction number when using the inference framework relying on cluster size distribution conditional on cluster extinction. Each boxplot represents the distribution of R_s maximum likelihood estimates across 100 simulations (2.5%, 25%, 50%, 75% and 97.5% percentiles). Results are displayed for a true dispersion parameter of 0.1 and running the inference on 1,000 clusters of identical sequences. Each panel corresponds to a different assumption regarding the probability p that transmission occurs before mutation. The horizontal dashed segments correspond to the true value of R_s (associated with the true reproduction number and the true dispersion parameter).

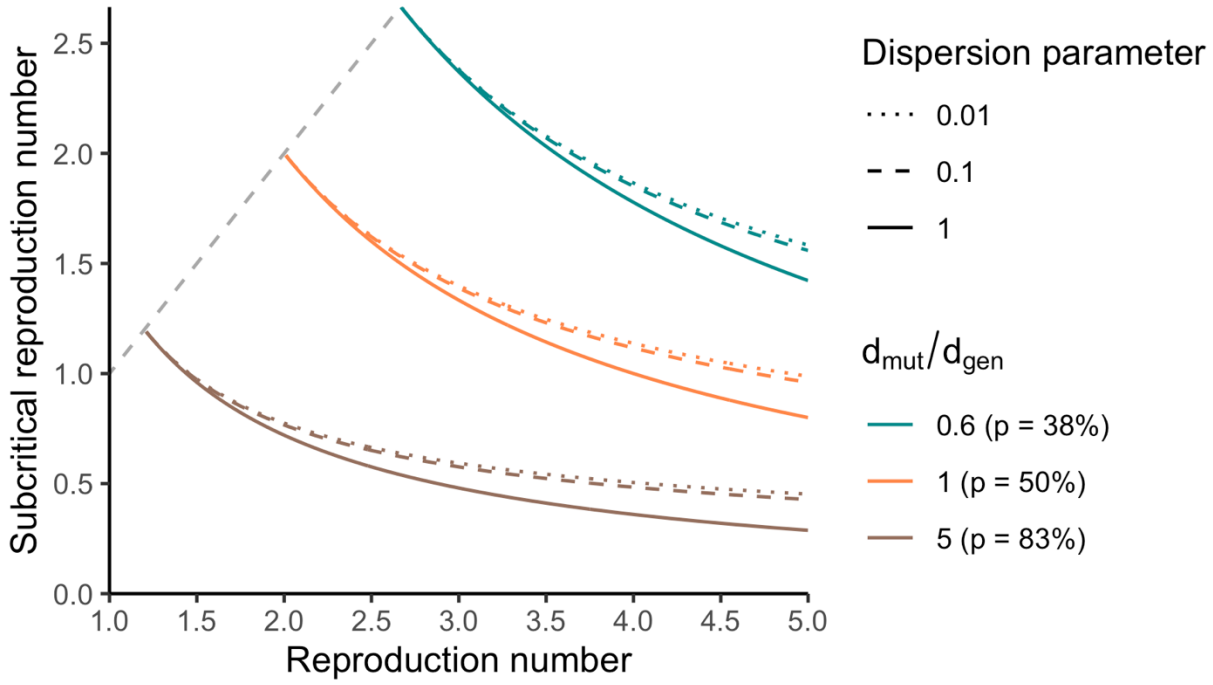

Figure S8: Relationship between the reproduction number and the subcritical reproduction number R_s for different probabilities p that transmission occurs before mutation and different values of the dispersion parameter k. Colored lines correspond to reproduction numbers lying above the threshold of 1/p. The dashed grey lines correspond to reproduction numbers lying below the reproduction number threshold (for which the reproduction number is equal to the subcritical reproduction number).

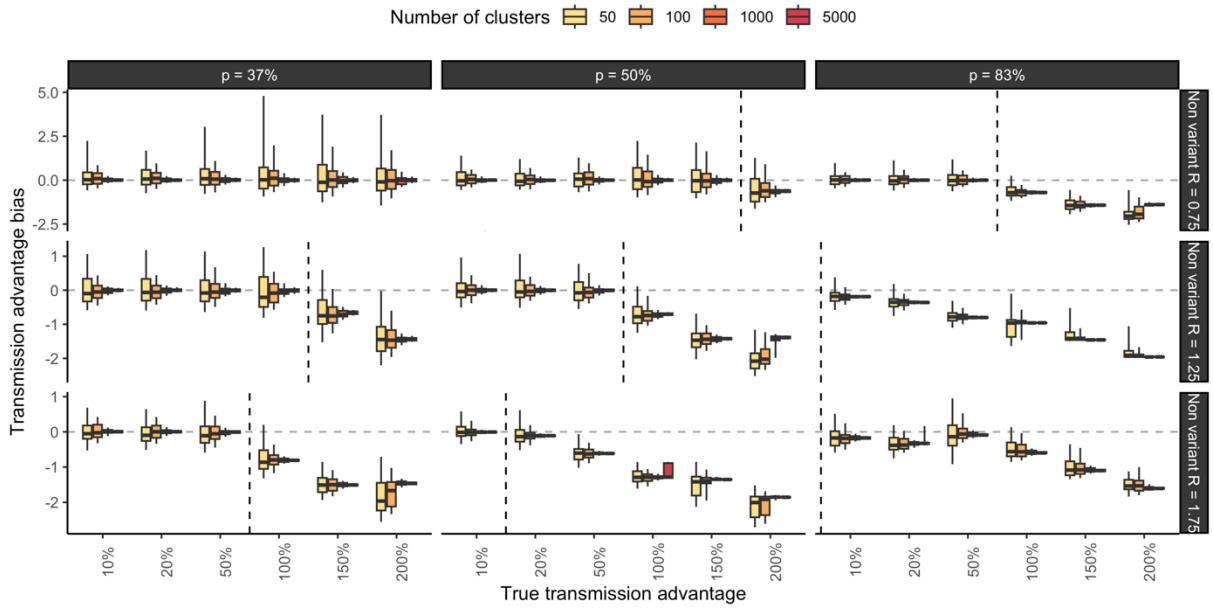

Figure S9: Transmission advantage bias as a function of the true transmission advantage and varying the probability that transmission occurs before mutation (rows) and the reproduction number of the non-variant R_{NV} (columns). Each subplot corresponds to a given assumption regarding the probability that transmission occurs before mutation and the reproduction number of the non-variant. In each subplot, the vertical dashed line corresponds to the limit from which the reproduction number of the variant R_V reaches the threshold of 1/p. Vertical dashed lines before the 10% x-axis tick correspond to situations where the reproduction number of the non-variant R_{NV} is also above the threshold of 1/p.

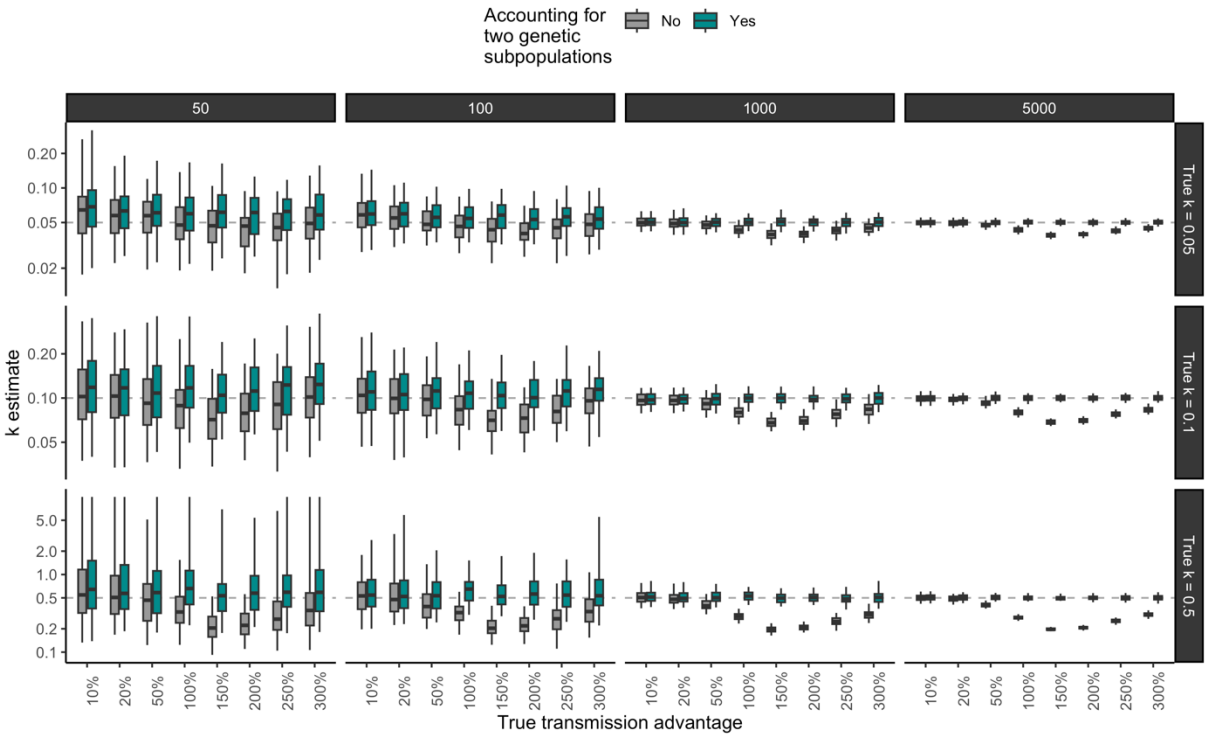

Figure S10: Impact of accounting for different genetic subpopulations on estimates of the dispersion parameter for different assumptions regarding the true dispersion parameter (rows) and different dataset sizes (columns) In each subplot, the horizontal dashed grey line corresponds to the true dispersion parameter value used to generate synthetic clusters of identical sequences. The boxplots summarize the 2.5%, 25%, 50%, 75% and 97.5% percentile of maximum-likelihood estimates obtained across 100 simulated datasets.

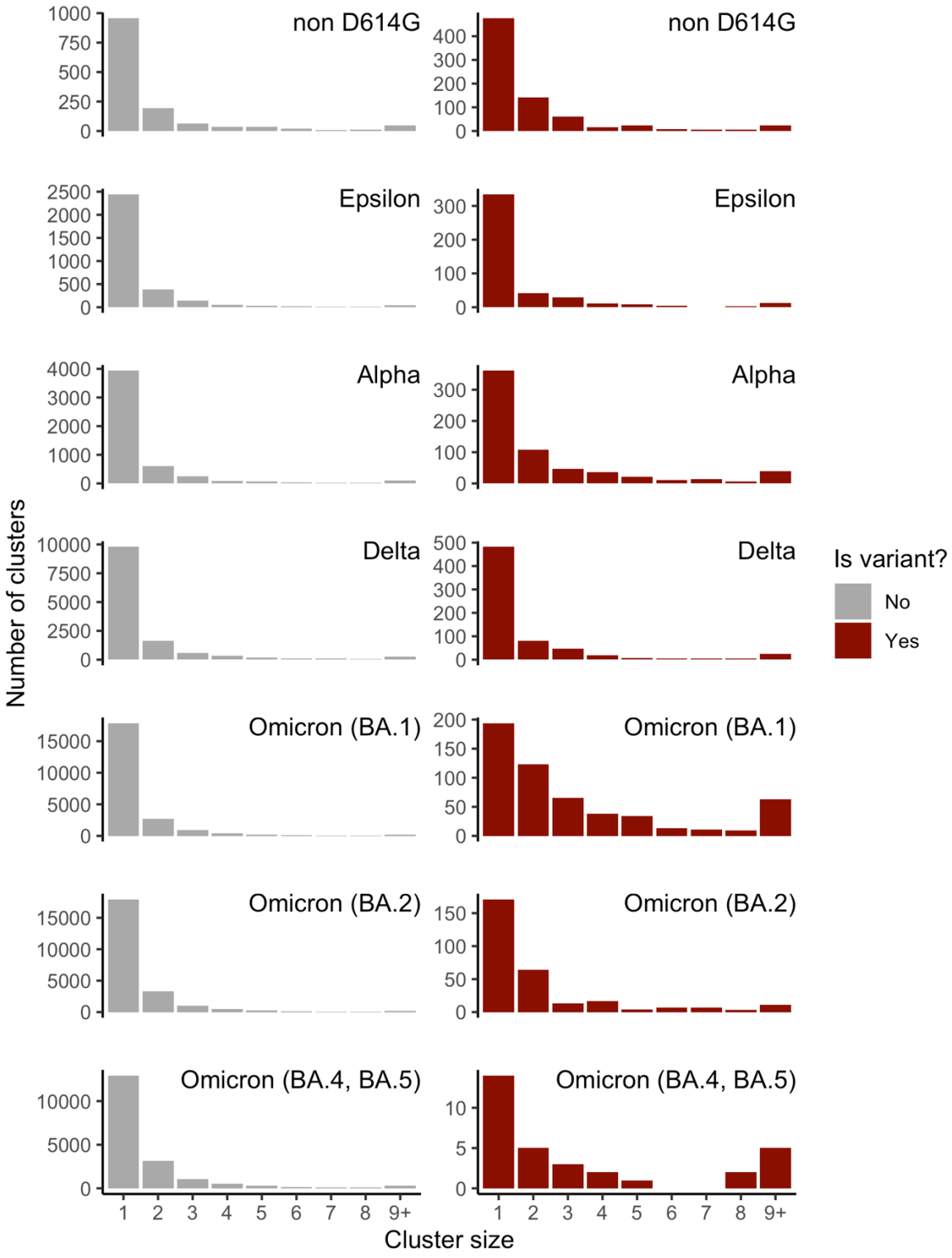

Figure S11: Size distribution of clusters of identical SARS-CoV-2 sequences in Washington state split by variant of interest. For each variant that we studied (rows), we displayed the distribution of cluster sizes for the variant and the non-variant. The time windows used to define these clusters are detailed in Table S5.

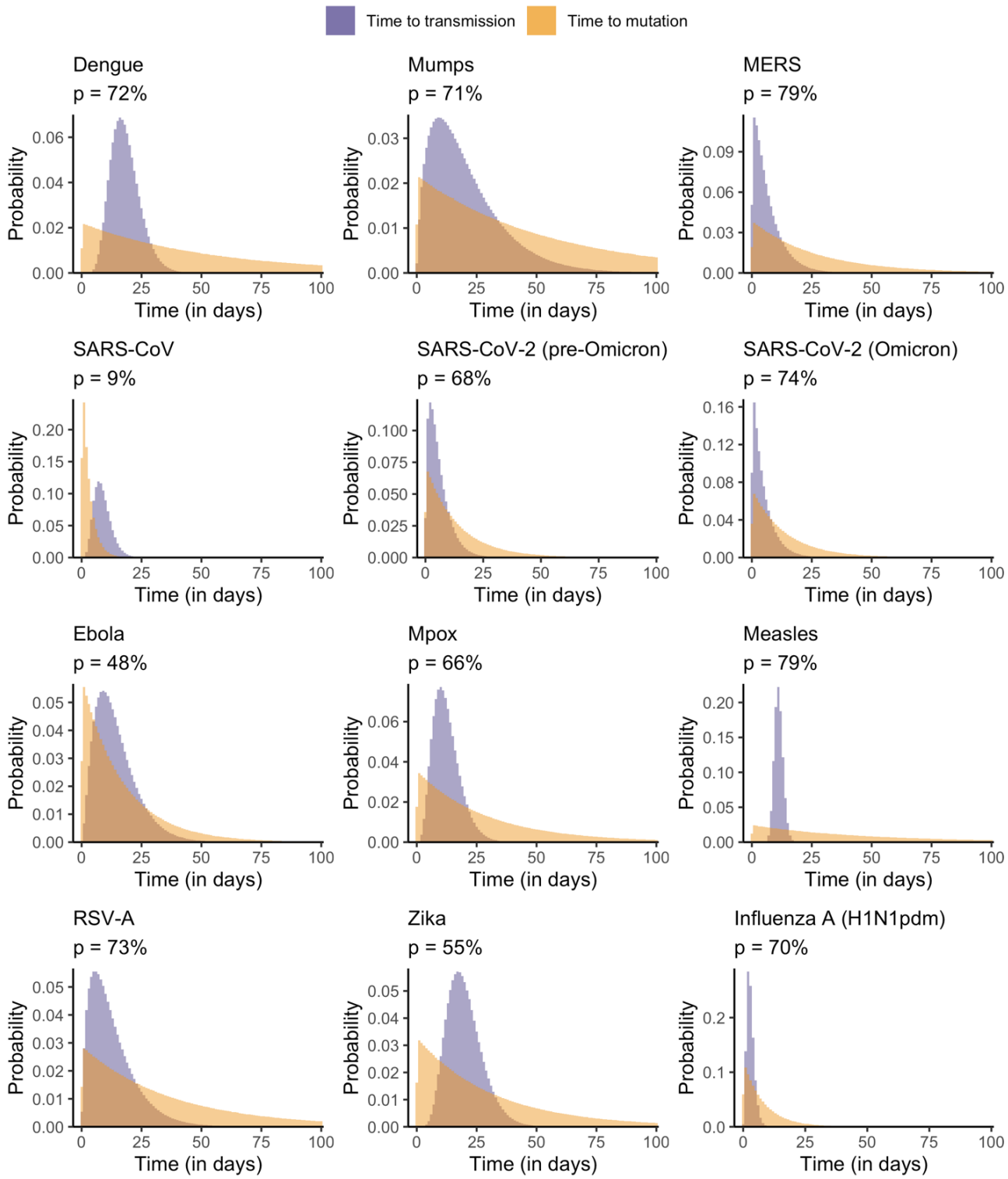

Figure S12: Comparison of the distribution of the time to occurrence of a first mutation and the time to transmission for different pathogens. For each pathogen, we additionally report the estimated probability p that transmission occurs before mutation.

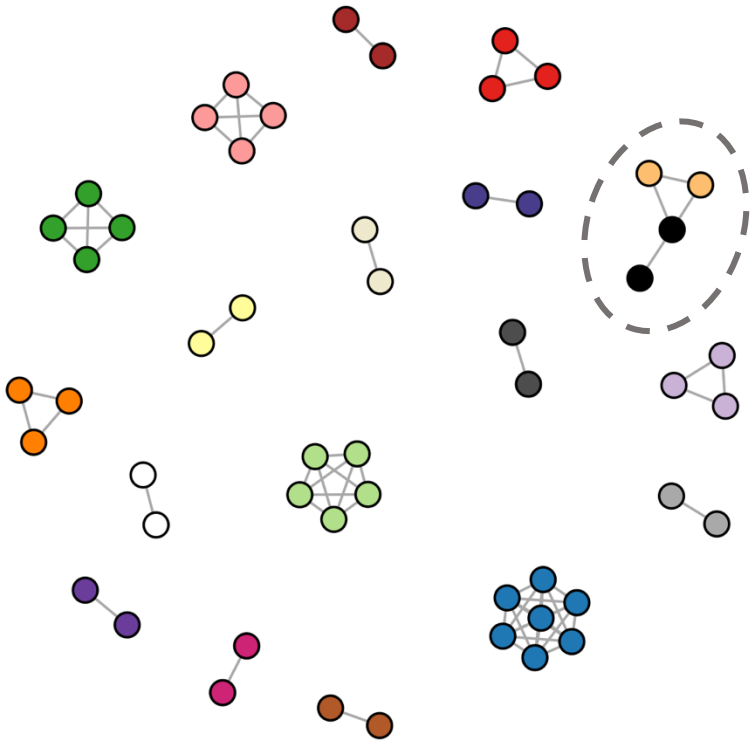

Figure S13: Difference between identical sequences obtained from the distance matrix and the reconstructed clusters of identical sequences for MERS-CoV sequences. Each vertex corresponds to a MERS-CoV sequence. Vertices are connected if their pairwise distance is equal to 0. Vertices have the same colour if they were allocated to the same cluster of identical sequences. The clusters for which there is a disagreement between the distance matrix and the cluster allocation (i.e. when some identical sequences are not in the same cluster) are circled. For clarity, we only displayed sequences with at least one other identical sequence in the pairwise distance matrix.

References

1. Cauchemez S, Nouvellet P, Cori A, Jombart T, Garske T, Clapham H, et al. Unraveling the drivers of MERS-CoV transmission. Proc Natl Acad Sci U S A. 2016 Aug 9;113(32):9081–6.
2. Dudas G, Carvalho LM, Rambaut A, Bedford T. MERS-CoV spillover at the camel-human interface. Elife [Internet]. 2018 Jan 16;7. Available from: http://dx.doi.org/10.7554/elife.31257
3. Klinkenberg D, Nishiura H. The correlation between infectivity and incubation period of measles, estimated from households with two cases. J Theor Biol. 2011 Sep 7;284(1):52–60.
4. Nextstrain - Real-time tracking of measles virus evolution [Internet]. Available from: https://nextstrain.org/measles?l=clock
5. Campbell F, Strang C, Ferguson N, Cori A, Jombart T. When are pathogen genome sequences informative of transmission events? PLoS Pathog. 2018 Feb;14(2):e1006885.
6. Ferguson NM, Cucunubá ZM, Dorigatti I, Nedjati-Gilani GL, Donnelly CA, Basáñez M-G, et al. EPIDEMIOLOGY. Countering the Zika epidemic in Latin America. Science. 2016 Jul 22;353(6297):353–4.
7. Faria NR, Quick J, Claro IM, Thézé J, de Jesus JG, Giovanetti M, et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature. 2017 Jun 15;546(7658):406–10.
8. Guzzetta G, Mammone A, Ferraro F, Caraglia A, Rapiti A, Marziano V, et al. Early estimates of Monkeypox incubation period, generation time, and reproduction number, Italy, may-June 2022. Emerg Infect Dis. 2022 Oct;28(10):2078–81.
9. Nextstrain - Genomic epidemiology of monkeypox virus [Internet]. Available from: https://nextstrain.org/monkeypox/hmpxv1?dmin=2022-01-01&l=clock
10. Hedge J, Lycett SJ, Rambaut A. Real-time characterization of the molecular epidemiology of an influenza pandemic. Biol Lett. 2013 Oct 23;9(5):20130331.
11. Nextstrain - mumps [Internet]. https://nextstrain.org/mumps/na?l=clock. Available from: https://nextstrain.org/mumps/na?l=clock
12. CDC. Mumps [Internet]. Centers for Disease Control and Prevention. 2022 [cited 2023 Feb 7]. Available from: https://www.cdc.gov/mumps/hcp.html
13. Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proc Biol Sci. 2007 Feb 22;274(1609):599–604.
14. Tan L, Coenjaerts FEJ, Houspie L, Viveen MC, van Bleek GM, Wiertz EJHJ, et al. The comparative genomics of human respiratory syncytial virus subgroups A and B: genetic variability and molecular evolutionary dynamics. J Virol. 2013 Jul;87(14):8213–26.
15. Dudas G, Bedford T. The ability of single genes vs full genomes to resolve time and space in outbreak analysis. BMC Evol Biol. 2019 Dec 26;19(1):232.
16. Hogan AB, Glass K, Moore HC, Anderssen RS. Exploring the dynamics of respiratory syncytial virus (RSV) transmission in children. Theor Popul Biol. 2016 Aug 1;110:78–85.
17. Salje H, Wesolowski A, Brown TS, Kiang MV, Berry IM, Lefrancq N, et al. Reconstructing unseen transmission events to infer dengue dynamics from viral sequences. Nat Commun. 2021 Mar 22;12(1):1810.
18. Nextstrain - dengue [Internet]. https://nextstrain.org/dengue/denv1?l=clock. Available from: https://nextstrain.org/dengue/denv1?l=clock
19. Hart WS, Abbott S, Endo A, Hellewell J, Miller E, Andrews N, et al. Inference of the SARS-CoV-2 generation time using UK household data. Elife [Internet]. 2022 Feb 9;11. Available from: http://dx.doi.org/10.7554/eLife.70767
20. Nextstrain - Genomic epidemiology of SARS-CoV-2 with subsampling focused globally since pandemic start [Internet]. Available from: https://nextstrain.org/ncov/gisaid/global/all-time?l=clock

21. Abbott S, Sherratt K, Gerstung M, Funk S. Estimation of the test to test distribution as a proxy for generation interval distribution for the Omicron variant in England [Internet]. bioRxiv. 2022. Available from: http://dx.doi.org/10.1101/2022.01.08.22268920
22. Ito K, Piantham C, Nishiura H. Estimating relative generation times and reproduction numbers of Omicron BA.1 and BA.2 with respect to Delta variant in Denmark. Math Biosci Eng. 2022 Jun 21;19(9):9005–17.
23. nextstrain.org/measles [Internet]. [cited 2022 Dec 16]. Available from: https://github.com/nextstrain/measles
24. Pacenti M, Maione N, Lavezzo E, Franchin E, Dal Bello F, Gottardello L, et al. Measles virus infection and immunity in a suboptimal vaccination coverage setting. Vaccines (Basel). 2019 Nov 28;7(4):199.