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. The numbers in parentheses correspond to uncertainty ranges.

Pathogen	Generation time in days		Mutation rate in subs/site/ year (uncertainty range)	Genome length	Ref. for the generation time	Ref. for the mutation rate
	Mean (uncertainty range)	Standard deviation (days)				
MERS-CoV	6.8 (6.0-7.8)	6.3	$4.81\cdot10^{-4}$ ($2.74\cdot10^{-4}$-$6.88\cdot10^{-4}$) [A]	30130	(1)	(2)
Measles virus	11.7 (9.9-13.8) [B]	1.8	$5.13\cdot10^{-4}$ ($4.84\cdot10^{-4}$-$5.13\cdot10^{-4}$) [C]	15894	(3)	(4)
Ebola virus	14.2 (13.1-15.5)	7.1	$9.82\cdot10^{-4}$ ($9.01\cdot10^{-4}$–$10.6\cdot10^{-4}$)	18958	(5)	(6)
Zika virus	20.0 (15.6-25.6)	7.4	$1.12\cdot10^{-3}$ ($0.97\cdot10^{-3}$-$1.27\cdot10^{-3}$)	10274	(7)	(8)
Mpox virus (2022-2023 outbreak)	12.5 (7.5-17.3)	5.7	$8.41\cdot10^{-5}$ ($7.71\cdot10^{-5}$ -$9.10\cdot10^{-5}$)	197209	(9)	(10)
Influenza A (H1N1)	2.6 (2.2-3.5)	1.3	$3.41\cdot10^{-3}$ ($3.15\cdot10^{-3}$-$3.67\cdot10^{-3}$)	13154	(11)	(12)
Influenza A (H3N2)	2.8 (3.1-4.6)	2.0	$1.43\cdot10^{-3}$ ($1.41\cdot10^{-3}$-$1.44\cdot10^{-3}$)	13486	(13)	Obtained from the slope of a linear regression of number of mutations accumulated vs time
Mumps virus	18.0 (17.4-18.6)	3.5	$8.6\cdot10^{-4}$ ($5.06\cdot10^{-4}$-$12.7\cdot10^{-4}$)	15384	(3)	(6)
RSV-A	7.5 (7.0-8.1)	2.1	$6.47\cdot10^{-4}$ ($5.56\cdot10^{-4}$ - $7.38\cdot10^{-4}$)	15200	(3)	(14)
SARS-CoV	8.7 [D]	3.6	$2.08\cdot10^{-3}$ ($0.8\cdot10^{-3}$-$2.38\cdot10^{-3}$)	29714	(15)	(16)
SARS-CoV-2 (pre-Omicron)	5.9 (5.2-7.0)	4.8	$1.10\cdot10^{-3}$ ($7.03\cdot10^{-4}$ -$1.50\cdot10^{-3}$)	29500	(17)	(18)
SARS-CoV-2 (Omicron)	4.9 (4.2-6.0)[E]	4.8			(19, 20)	

[A] The uncertainty range for the MERS-CoV mutation rate was obtained by subtracting and adding the standard deviation reported in (2) to the central estimate.

[B] The uncertainty range for the measles generation time was obtained by considering the range of values reported for the mean measles generation time in (3).

[C] The uncertainty range for the measles' mutation rate was obtained by subtracting and adding the standard deviation obtained with the Nextstrain measles workflow (4).

[D] We did not explore any uncertainty around the SARS mean generation time (no estimates found).

[E] For Omicron, we assumed that the mean generation time was one day shorter than for pre-Omicron variants (19, 20) and considered that it was characterized by the same standard deviation.

Table S2: Estimates of the probability that transmission occurs before mutation for different pathogens.

Pathogen	Values used to inform the generation time / the mutation rate					Central estimate (uncertainty range)
	Central / Central	Lower/ Central	Upper / Central	Central / Lower	Central / Upper	
MERS-CoV	0.78	0.81	0.75	**0.87**	**0.72**	0.78 (0.72-0.87)
Measles virus	0.77	**0.80**	**0.74**	0.78	0.77	0.77 (0.74-0.80)
Ebola virus	0.51	**0.54**	**0.48**	0.54	0.49	0.51 (0.48-0.54)
Zika virus	0.55	**0.63**	**0.46**	0.59	0.51	0.55 (0.46-0.63)
Mpox virus (2022-2023 outbreak)	0.59	**0.73**	**0.47**	0.61	0.56	0.59 (0.47-0.73)
Influenza A (H1N1)	0.74	**0.77**	**0.66**	0.75	0.72	0.74 (0.66-0.77)
Influenza A (H3N2)	0.82	**0.85**	**0.79**	0.83	0.82	0.82 (0.79-0.85)
Mumps virus	0.53	0.54	0.51	**0.68**	**0.39**	0.53 (0.39-0.68)
RSV-A	0.82	0.83	0.81	**0.84**	**0.80**	0.82 (0.80-0.84)
SARS-CoV	0.27	-	-	**0.58**	**0.23**	0.27 (0.23-0.58)
SARS-CoV-2 (pre-Omicron)	0.64	0.68	0.58	**0.74**	**0.56**	0.64 (0.56-0.74)
SARS-CoV-2 (Omicron)	0.69	0.74	0.63	**0.78**	**0.63**	0.69 (0.63-0.78)

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

Estimate used for the probability that transmission occurs before mutation	Proportion of infections detected as cases	Reproduction number *R* estimate	Dispersion parameter *k* estimate
Central	1.0	0.57 50%CI: (0.54-0.61) 95%CI: (0.46-0.70)	0.14 50%CI: (0.09-0.20) 95%CI: (0.04-0.46)
	0.5	0.65 50%CI: (0.61-0.68) 95%CI: (0.54-0.77)	0.09 50%CI: (0.07-0.13) 95%CI: (0.03-0.26)
Lower bound from uncertainty range	1.0	0.63 50%CI: (0.59-0.67) 95%CI: (0.50-0.77)	0.14 50%CI: (0.09-0.20) 95%CI: (0.04-0.46)
	0.5	0.71 50%CI: (0.67-0.75) 95%CI: (0.59-0.84)	0.09 50%CI: (0.07-0.13) 95%CI: (0.03-0.26)
Upper bound from uncertainty range	1.0	0.52 50%CI: (0.46-0.56) 95%CI: (0.42-0.64)	0.14 50%CI: (0.09-0.20) 95%CI: (0.04-0.46)
	0.5	0.59 50%CI: (0.55-0.62) 95%CI: (0.49-0.70)	0.09 50%CI: (0.07-0.13) 95%CI: (0.03-0.26)

Table S4: Parameter estimates for measles. Maximum likelihood estimates are reported along 50% and 95% confidence intervals (CI).

Estimate used for the probability that transmission occurs before mutation	Proportion of infections detected as cases	Reproduction number *R* estimate	Dispersion parameter *k* estimate
Central	1.0	0.58 50%CI: (0.47-0.73) 95%CI: (0.29-1.18)	0.04 50%CI: (0.016-0.092) 95%CI: (0.003-0.45)
	0.5	0.62 50%CI: (0.50-0.76) 95%CI: (0.32-1.17)	0.02 50%CI: (0.009-0.05) 95%CI: (0.002-0.19)
Lower bound from uncertainty range	1.0	0.61 50%CI: (0.49-0.77) 95%CI: (0.31-1.23)	0.04 50%CI: (0.016-0.092) 95%CI: (0.003-0.45)
	0.5	0.65 50%CI: (0.52-0.80) 95%CI: (0.34-1.23)	0.02 50%CI: (0.009-0.05) 95%CI: (0.002-0.19)
Upper bound from uncertainty range	1.0	0.56 50%CI: (0.45-0.71) 95%CI: (0.28-1.13)	0.04 50%CI: (0.016-0.092) 95%CI: (0.003-0.45)
	0.5	0.59 50%CI: (0.31-1.13) 95%CI: (0.48-0.73)	0.02 50%CI: (0.009-0.05) 95%CI: (0.002-0.19)

Table S5: Parameter estimates for SARS-CoV-2 in New Zealand under our central estimate for the probability *p* that transmission occurs before mutation. Maximum likelihood estimates are reported along 50% and 95% confidence intervals (CI).

Proportion of infections detected as cases	Period	Reproduction number *R* estimate	Dispersion parameter *k* estimate
1.0	April – May 2020	0.82 95%CI: (0.65-1.01) 50%CI: (0.76-0.88)	0.63 95%CI: (0.34-1.56) 50%CI: (0.5-0.82)
	June – December 2020	0.87 95%CI: (0.75-1.01) 50%CI: (0.83-0.91)	
	January – April 2021	0.74 95%CI: (0.61-0.90) 50%CI: (0.70-0.79)	
	May – July 2021	0.66 95%CI: (0.48-0.87) 50%CI: (0.60-0.72)	
0.8	April – May 2020	0.87 95%CI: (0.70-1.06) 50%CI: (0.82-0.93)	0.62 95%CI: (0.33-1.54) 50%CI: (0.49-0.81)
	June – December 2020	0.92 95%CI: (0.80-1.05) 50%CI: (0.88-0.96)	
	January – April 2021	0.80 95%CI: (0.66-0.95) 50%CI: (0.75-0.84)	
	May – July 2021	0.71 95%CI: (0.54-0.92) 50%CI: (0.65-0.78)	
0.5	April – May 2020	0.98 95%CI: (0.82-1.14) 50%CI: (0.92-1.03)	0.59 95%CI: (0.31-1.43) 50%CI: (0.47-0.77)
	June – December 2020	1.02 95%CI: (0.91-1.13) 50%CI: (0.98-1.05)	
	January – April 2021	0.90 95%CI: (0.78-1.04) 50%CI: (0.86-0.94)	
	May – July 2021	0.82 95%CI: (0.65-1.02) 50%CI: (0.77-0.88)	

Table S6: Parameter estimates for SARS-CoV-2 in New Zealand under our lower bound estimate for the probability *p* that transmission occurs before mutation. Maximum likelihood estimates are reported along 50% and 95% confidence intervals (CI).

Proportion of infections detected as cases	Period	Reproduction number *R* estimate	Dispersion parameter *k* estimate
1.0	April – May 2020	0.94 95%CI: (0.74-1.16) 50%CI: (0.87-1.01)	0.63 95%CI: (0.34-1.53) 50%CI: (0.50-0.82)
	June – December 2020	1.00 95%CI: (0.86-1.15) 50%CI: (0.95-1.05)	
	January – April 2021	0.85 95%CI: (0.69-1.03) 50%CI: (0.79-0.90)	
	May – July 2021	0.75 95%CI: (0.55-1.00) 50%CI: (0.68-0.83)	
0.8	April – May 2020	1.00 95%CI: (0.80-1.21) 50%CI: (0.93-1.07)	0.62 95%CI: (0.32-1.51) 50%CI: (0.49-0.80)
	June – December 2020	1.05 95%CI: (0.92-1.20) 50%CI: (1.01-1.10)	
	January – April 2021	0.91 95%CI: (0.75-1.08) 50%CI: (0.86-0.96)	
	May – July 2021	0.81 95%CI: (0.61-1.05) 50%CI: (0.74-0.89)	
0.5	April – May 2020	1.11 95%CI: (0.93-1.30) 50%CI: (1.06-1.18)	0.58 95%CI: (0.30-1.39) 50%CI: (0.46-0.76)
	June – December 2020	1.16 95%CI: (1.04-1.30) 50%CI: (1.12-1.21)	
	January – April 2021	1.03 95%CI: (0.88-1.19) 50%CI: (0.98-1.08)	
	May – July 2021	0.94 95%CI: (0.74-1.16) 50%CI: (0.87-1.01)	

Table S7: Parameter estimates for SARS-CoV-2 in New Zealand under our upper bound estimate for the probability *p* that transmission occurs before mutation. Maximum likelihood estimates are reported along 50% and 95% confidence intervals (CI).

Proportion of infections detected as cases	Period	Reproduction number *R* estimate	Dispersion parameter *k* estimate
1.0	April – May 2020	0.71 95%CI: (0.56-0.87) 50%CI: (0.66-0.76)	0.64 95%CI: (0.34-1.58) 50%CI: (0.51-0.83)
	June – December 2020	0.75 95%CI: (0.65-0.87) 50%CI: (0.72-0.79)	
	January – April 2021	0.64 95%CI: (0.53-0.78) 50%CI: (0.60-0.68)	
	May – July 2021	0.57 95%CI: (0.42-0.75) 50%CI: (0.52-0.62)	
0.8	April – May 2020	0.75 95%CI: (0.61-0.91) 50%CI: (0.71-0.80)	0.63 95%CI: (0.33-1.57) 50%CI: (0.50-0.82)
	June – December 2020	0.80 95%CI: (0.70-0.91) 50%CI: (0.76-0.83)	
	January – April 2021	0.69 95%CI: (0.57-0.82) 50%CI: (0.65-0.73)	
	May – July 2021	0.62 95%CI: (0.46-0.80) 50%CI: (0.56-0.67)	
0.5	April – May 2020	0.84 95%CI: (0.71-0.98) 50%CI: (0.80-0.89)	0.60 95%CI: (0.31-1.47) 50%CI: (0.47-0.79)
	June – December 2020	0.88 95%CI: (0.79-0.98) 50%CI: (0.85-0.91)	
	January – April 2021	0.78 95%CI: (0.67-0.90) 50%CI: (0.75-0.82)	
	May – July 2021	0.71 95%CI: (0.56-0.88) 50%CI: (0.66-0.76)	

Table S8: Definitions of the study periods for the Washington state SARS-CoV-2 analysis. Dates are reported in a YYYY-MM-DD format.

Variant under study	Date of first collection of the variant	Date from which at least 10 variant sequences were collected (cumulative)	Corresponding Nextstrain clades
D614G	2020-02-22	2020-03-10	19A, 19B, 19C
Epsilon	2020-11-17	2020-12-13	21C (Epsilon)
Alpha	2020-11-23	2021-01-18	20I (Alpha, V1)
Delta	2021-04-03	2021-04-12	21A (Delta), 21I (Delta), 21J (Delta)
Omicron (BA.1)	2021-09-29	2021-12-01	21K (Omicron)
Omicron (BA.2)	2022-01-03	2022-01-12	21L (Omicron)
Omicron (BA.4, BA.5)	2022-04-15	2022-05-08	22A (Omicron), 22B (Omicron)

Table S9: Genbank accession numbers for measles sequences used in the analysis. All sequences were obtained from Pacenti et al. using the Nextstrain measles workflow (21, 22).

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

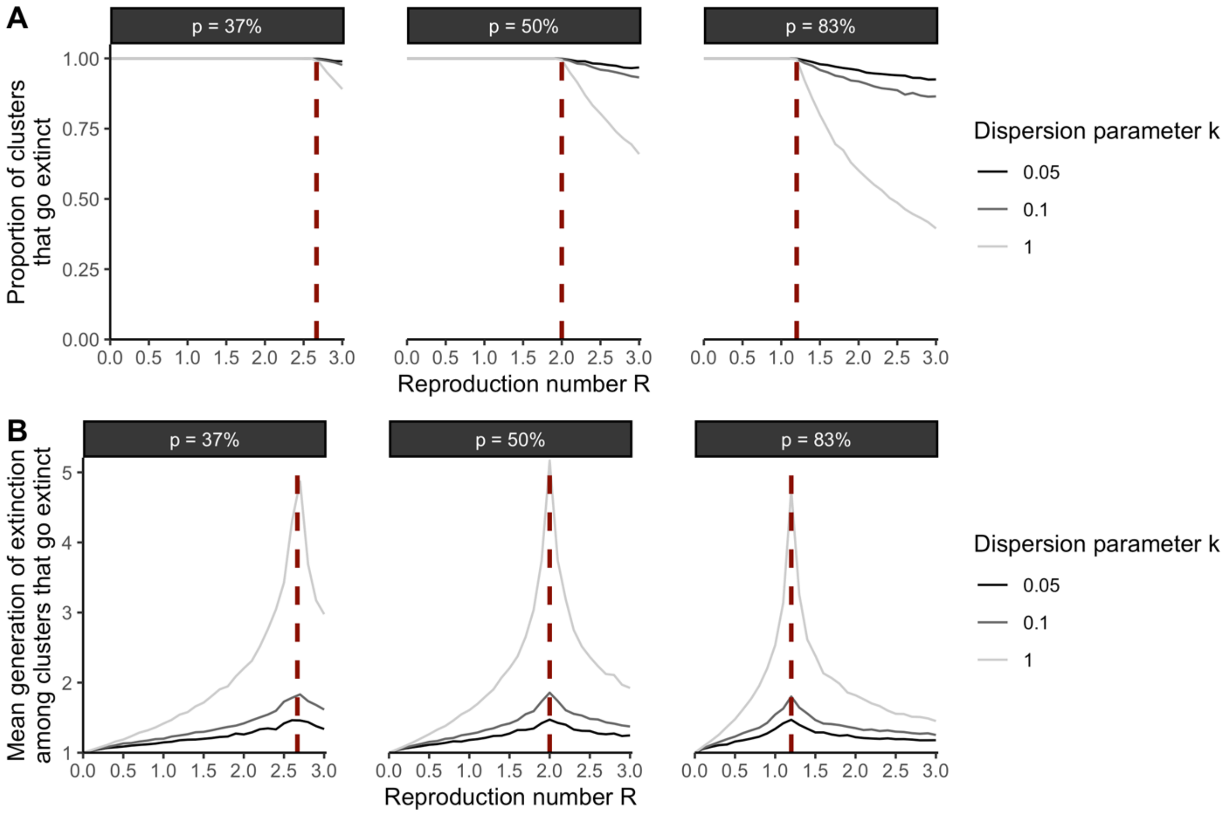

Figure S1: Dynamics of extinction for clusters of identical pathogen sequences. A. Proportion of clusters of identical sequences that go extinct as a function of the reproduction number *R* (x-axis) exploring different assumptions regarding the dispersion parameter *k* (colored lines) and the probability 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.

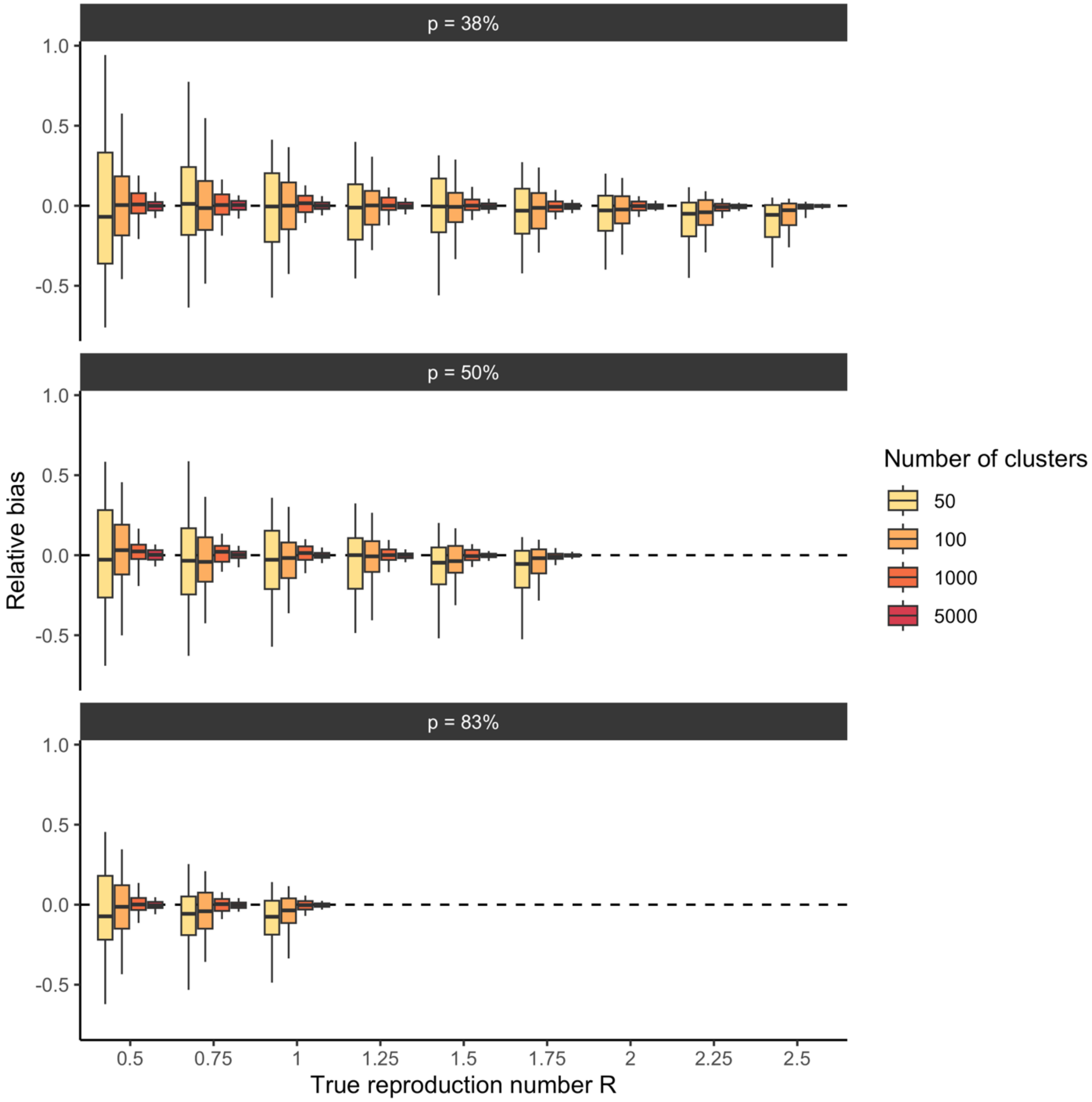

Figure S2: Relative bias on the reproduction number *R* estimate when the reproduction number lies below the threshold of *1/p*. For each true value of the reproduction number *R* (x-axis) and value of the probability p that transmission occurs before mutation, the boxplot depicts the distribution of 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.

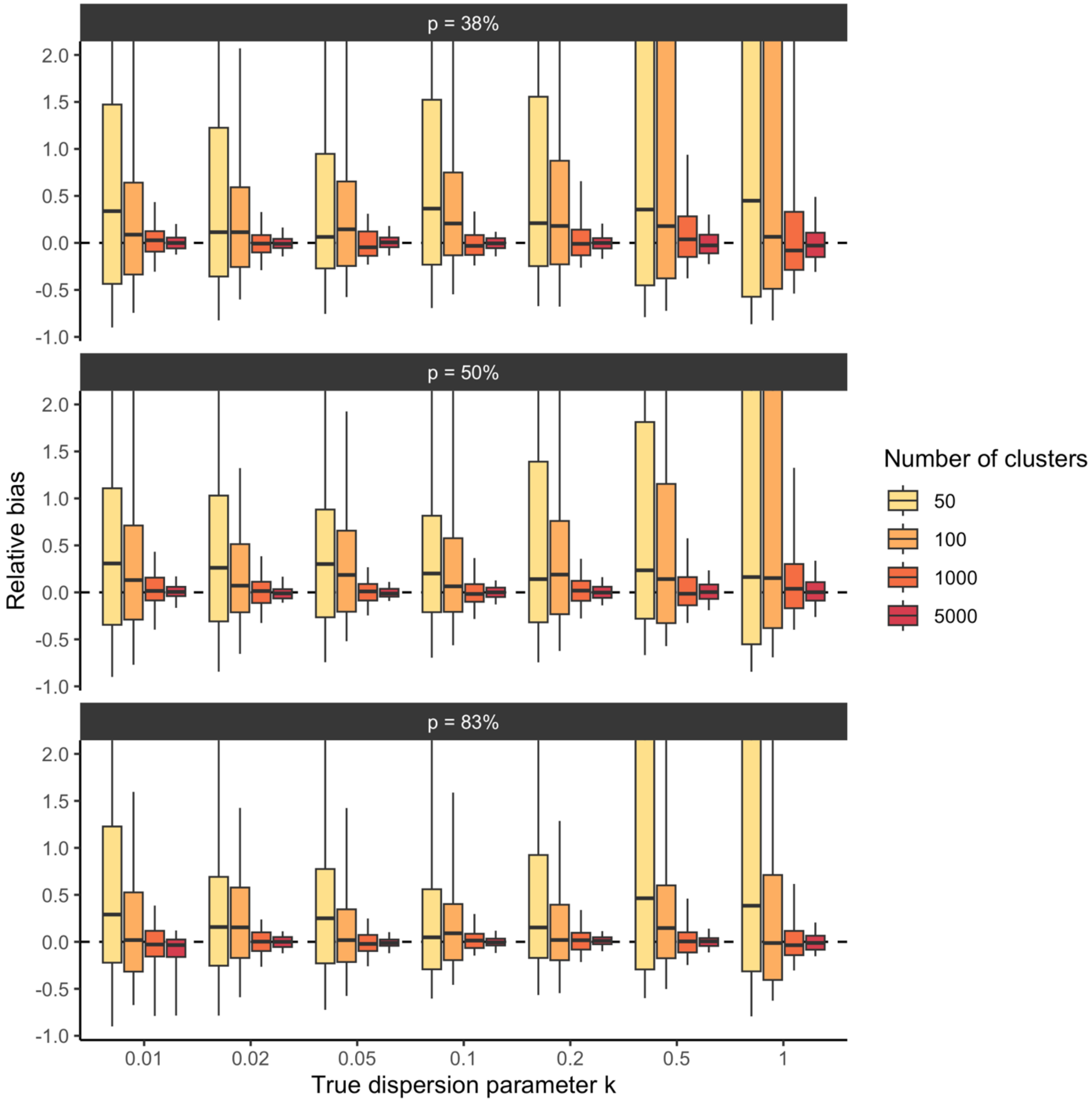

Figure S3: Relative bias on the dispersion parameter *k* estimate when the reproduction number lies below the threshold of *1/p*. For each true value of the dispersion parameter *k* (x-axis) and value of the probability *p* that transmission occurs before mutation, the boxplot depicts the distribution 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.

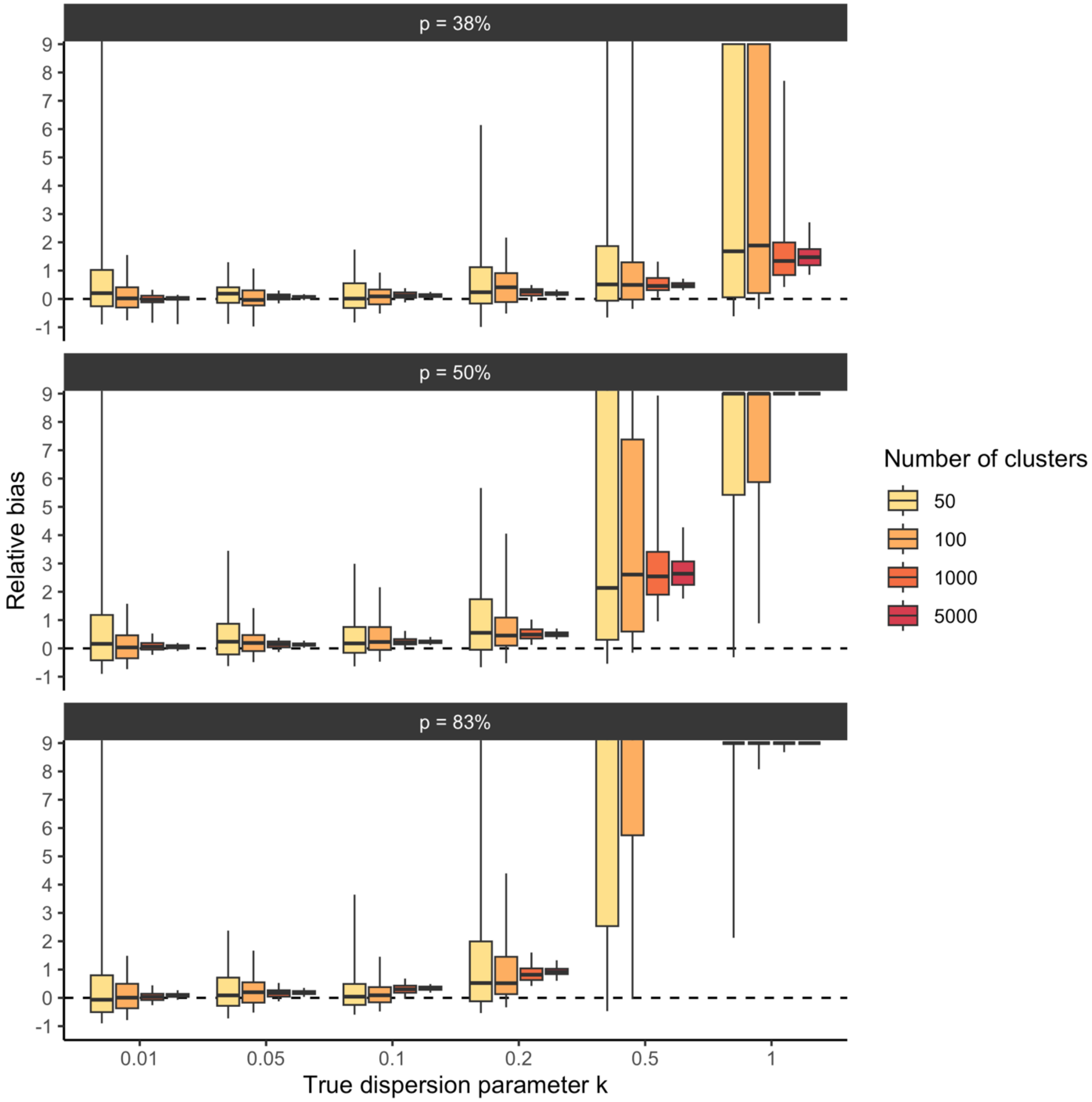

Figure S4: Relative bias on the dispersion parameter *k* estimate when the reproduction number lies above the threshold of *1/p*. For each true value of the dispersion parameter *k* (x-axis) and value of the probability *p* that transmission occurs before mutation, the boxplot depicts the distribution 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.

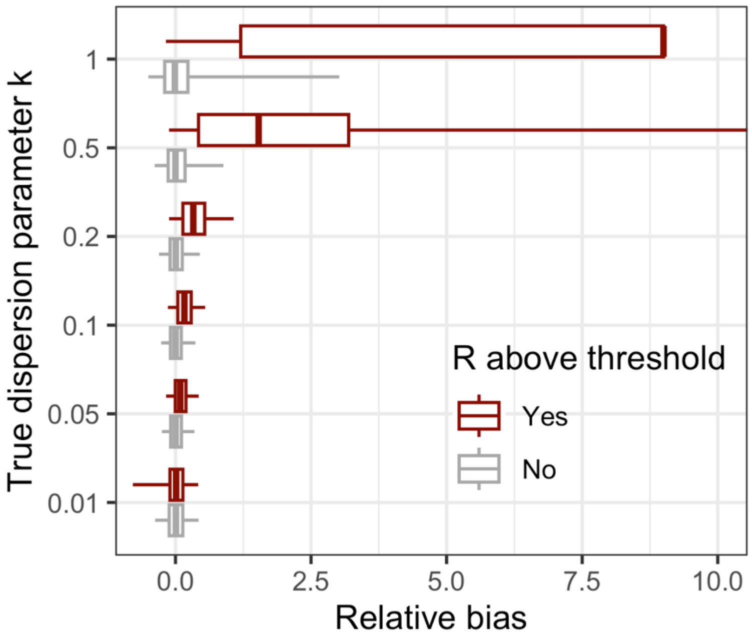

Figure S5: Impact of reaching the reproduction number threshold on dispersion parameter estimates. The relative bias is defined as $(k^{MLE} - k^{true})/k^{true}$ where k^{true} is the true dispersion parameter used to generate synthetic cluster data and k^{MLE} our maximum likelihood estimate. The boxplots 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.

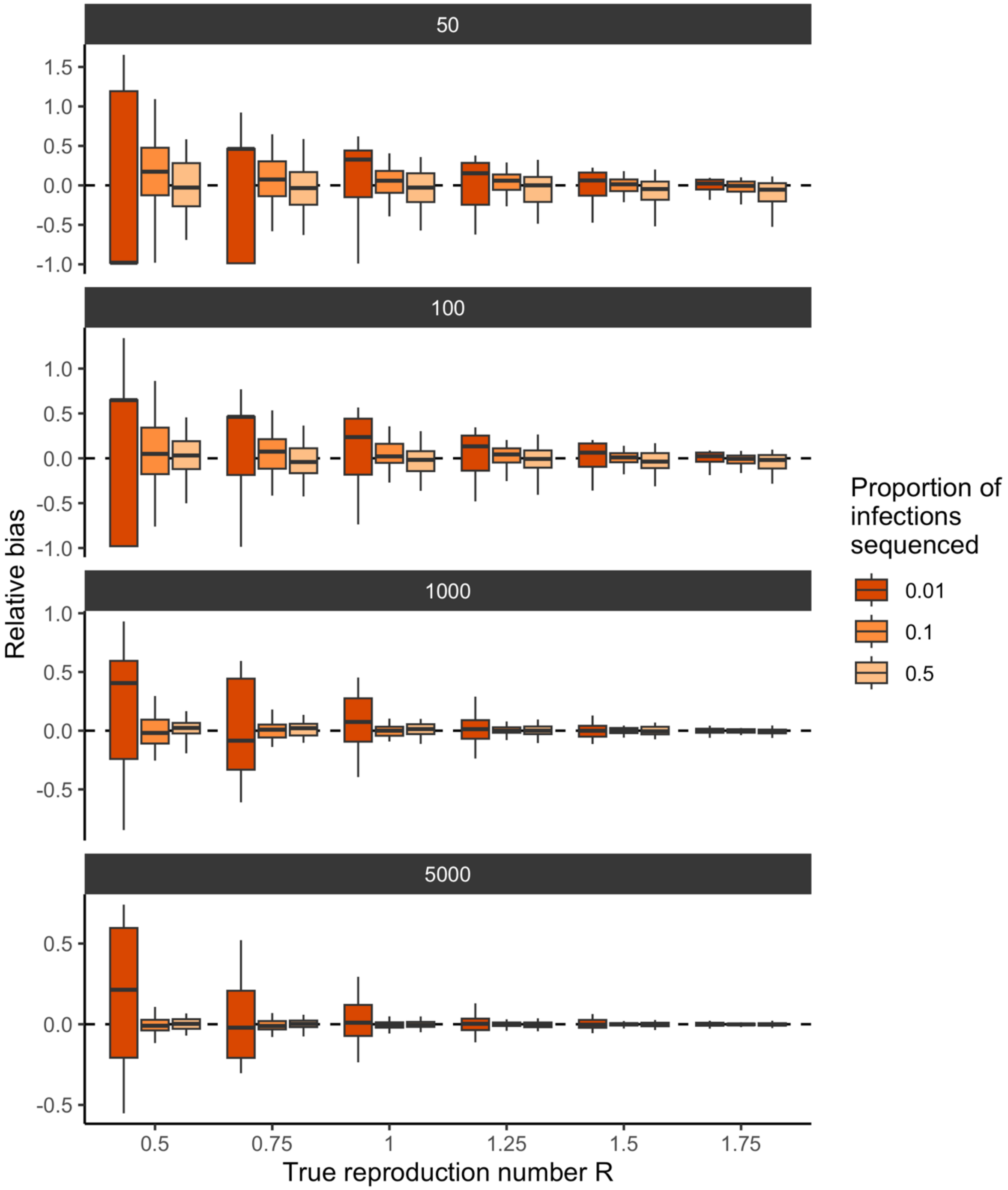

Figure S6: Impact of the proportion of infections sequenced on the relative bias on the reproduction number *R* estimate. Results are reported for a probability that transmission occurs before mutation of 50% and a dispersion parameter value of 0.1. For each true value of the reproduction number *R* (x-axis) and different dataset sizes (different subplots), the boxplots depict the distribution of the relative bias across 100 simulations for different proportion of infections sequenced (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.

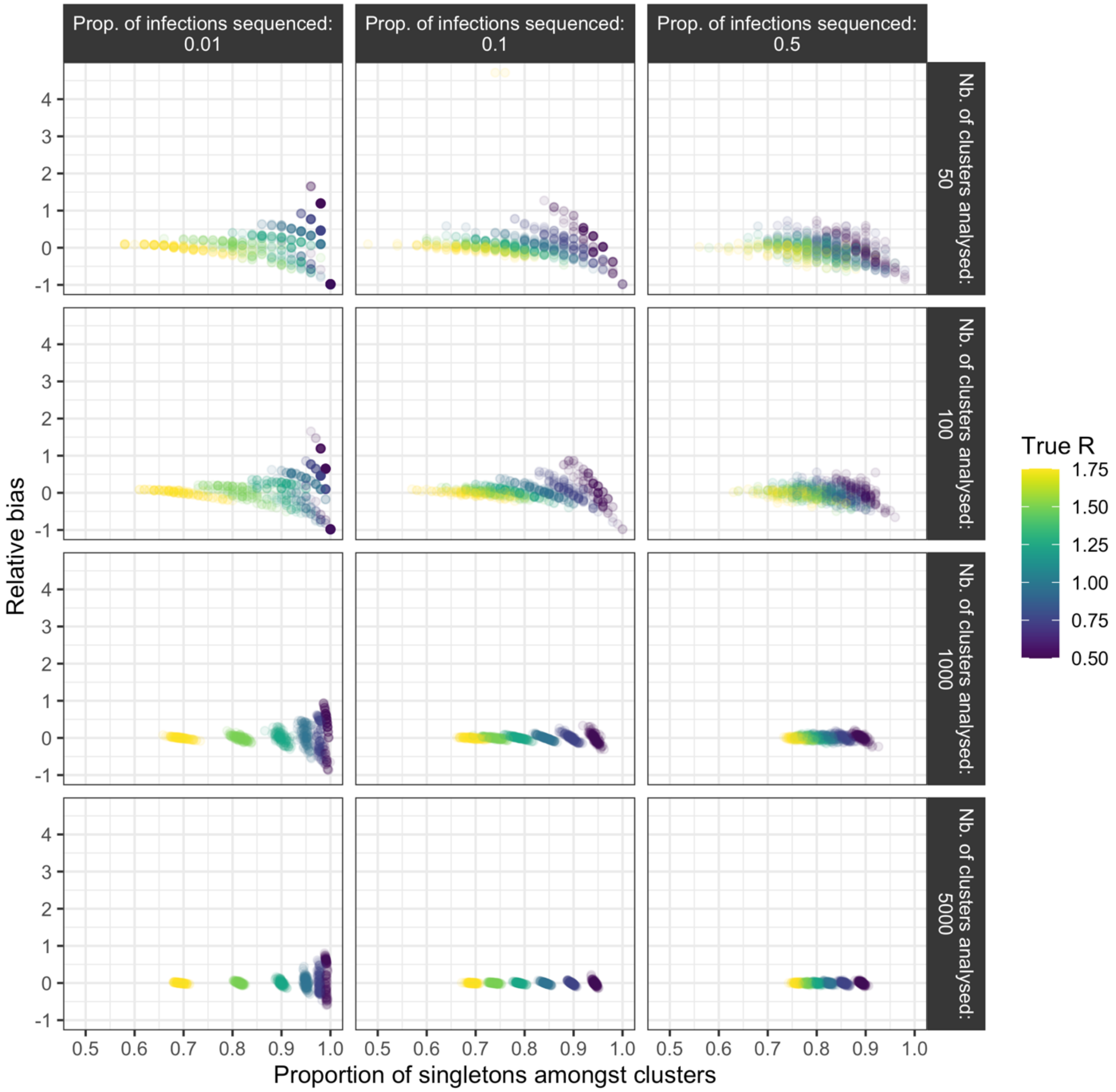

Figure S7: Relationship between the proportion of singletons among clusters analyzed and the relative bias on the reproduction number *R* estimate. Different assumptions regarding the proportion of infections sequenced (columns) and the size of the dataset on which the inference was run (rows) are explored. Points are coloured by true reproduction number value. Results are reported for a probability that transmission occurs before mutation of 50% and a dispersion parameter value of 0.1. 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.

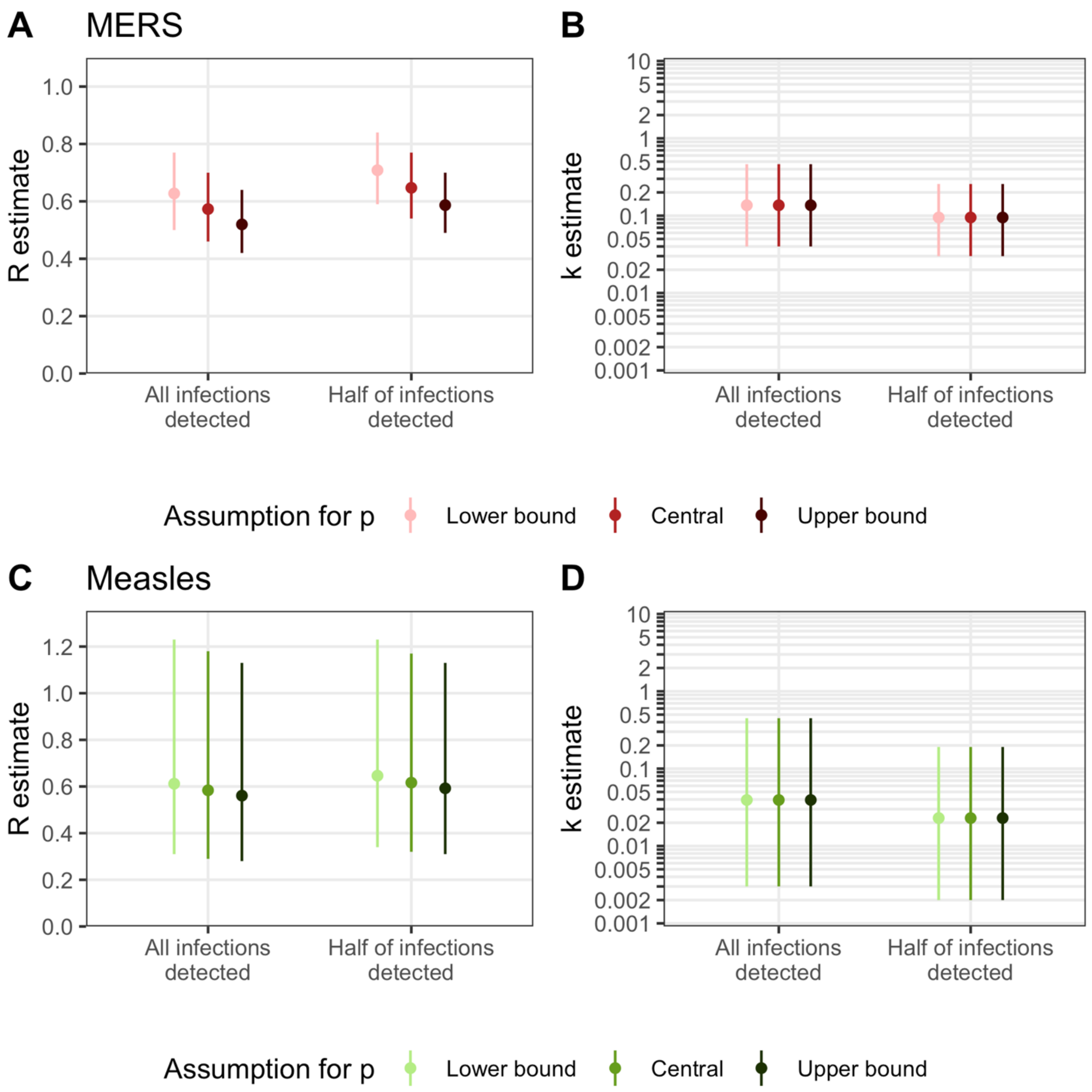

Figure S8: Sensitivity analysis exploring how *R* and *k* estimates for MERS and measles are impacted by assumptions regarding the probability *p* that transmission occurs before mutation. Estimates of **A.** the reproduction number R and **B.** the dispersion parameter *k* for MERS. Estimates of **C.** the reproduction number *R* and **D.** the dispersion parameter *k* for measles during the 2017-2018 Italy outbreak.

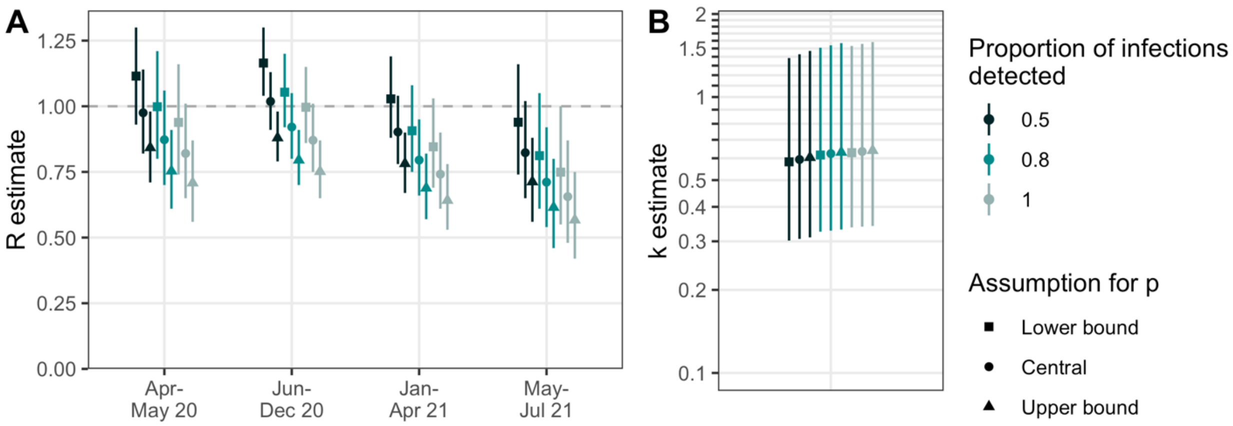

Figure S9: Sensitivity analysis exploring how *R* and *k* estimates for SARS-CoV-2 in New Zealand are impacted by assumptions regarding the probability p that transmission occurs before mutation. Estimates of **A.** the reproduction number *R* and **B.** the dispersion parameter *k* for SARS-CoV-2 in New Zealand.

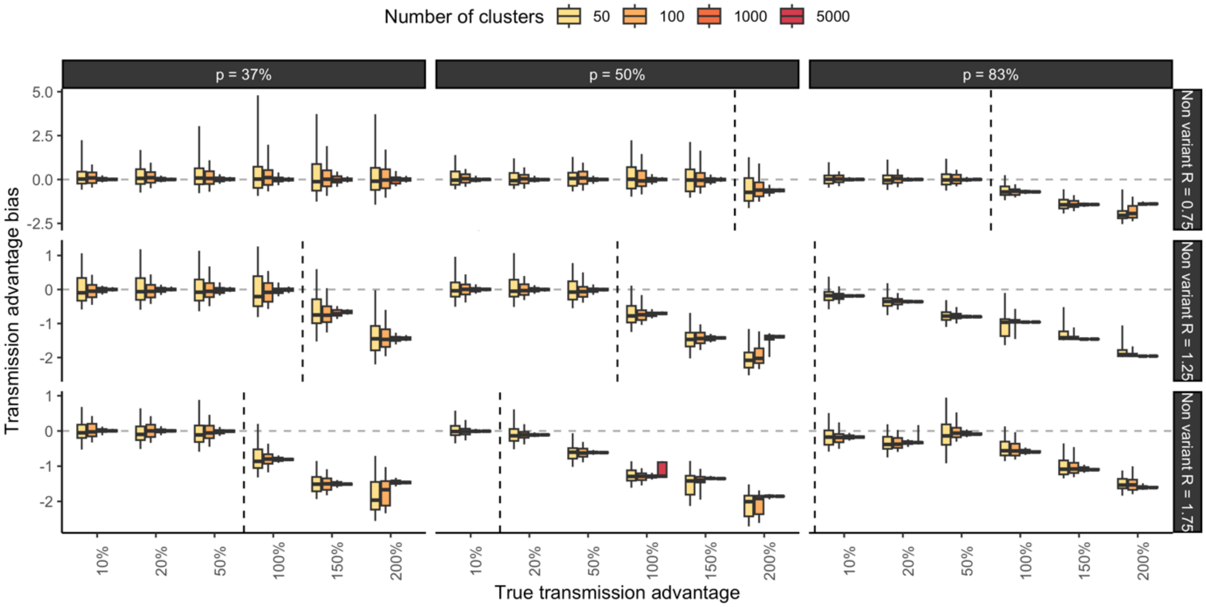

Figure S10: Transmission advantage bias as a function of the true transmission advantage and varying the probability that transmission occurs before mutation (rows) and the reproduction number of the non-variant R_{NV} (columns). Each subplot corresponds to a given assumption regarding the probability 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*.

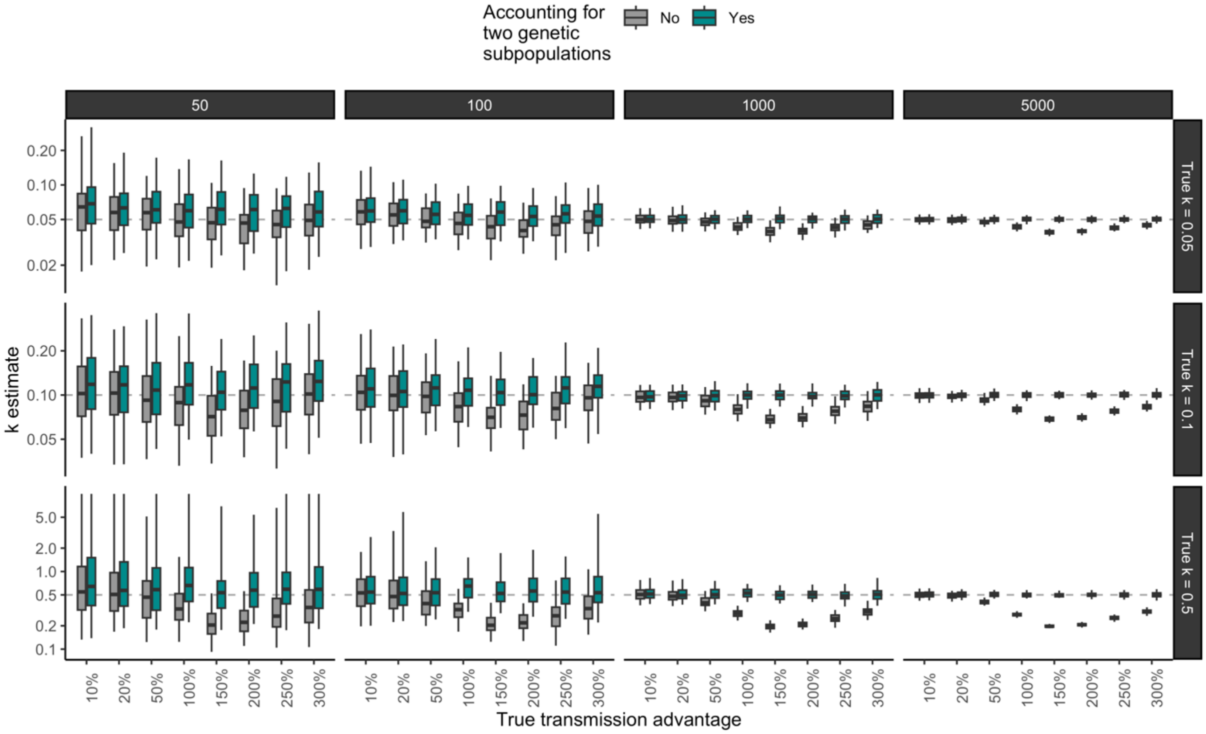

Figure S11: Impact of accounting for different genetic subpopulations on estimates of the dispersion parameter for different assumptions regarding the true dispersion parameter (rows) and different dataset sizes (columns) In each subplot, the horizontal dashed grey line corresponds to the true dispersion 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.

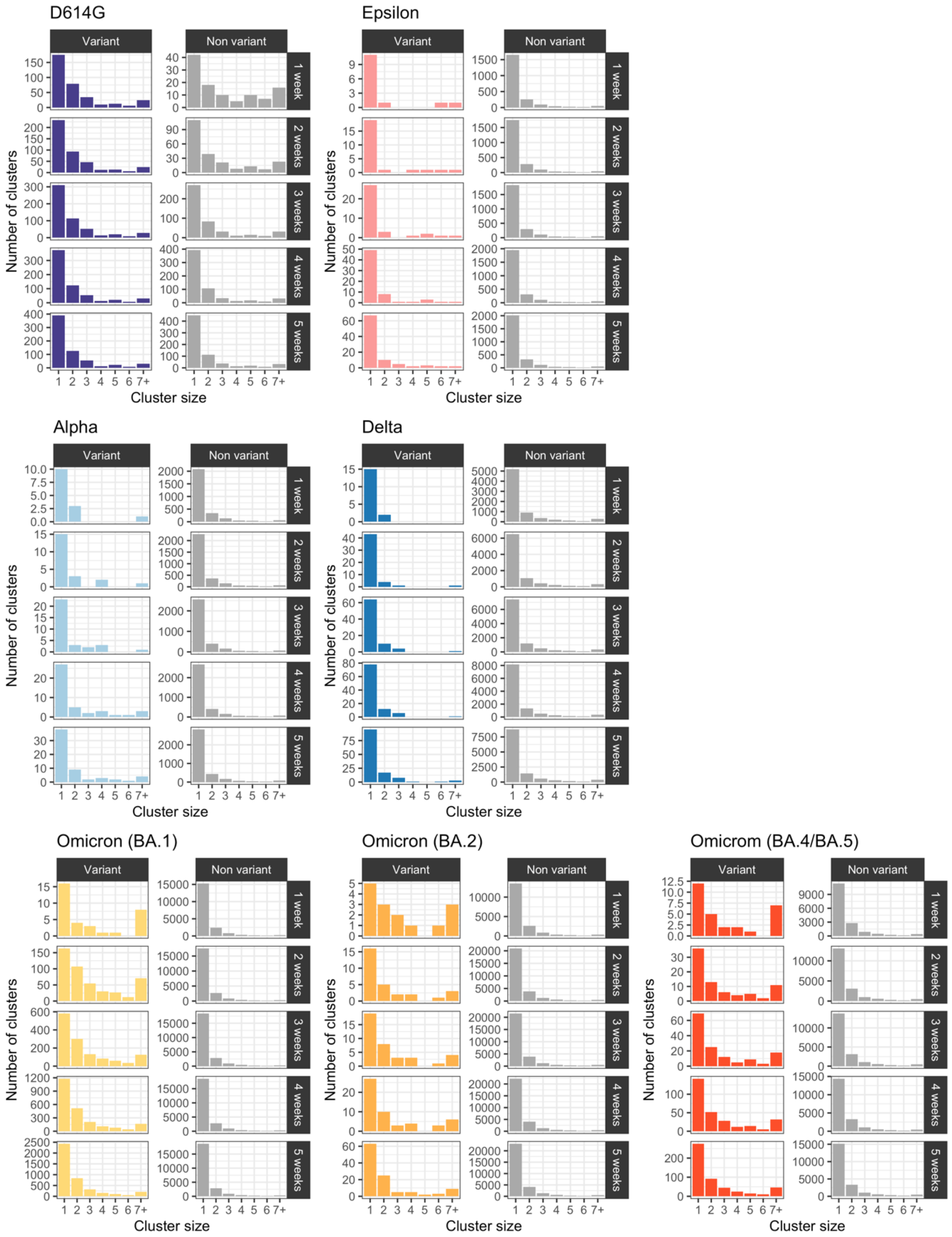

Figure S12: Size distribution of clusters of identical SARS-CoV-2 sequences in Washington state split by variant of interest. For each variant that we studied, we displayed the distribution of cluster sizes for the variant and the non-variant considering different time window length (1 to 5 weeks). The time windows begin when the cumulative number of collected in Washington state variant sequences reached 10 (See Table S8). We considered that clusters of identical sequences fell into the time-window if they were first detected during that time window.

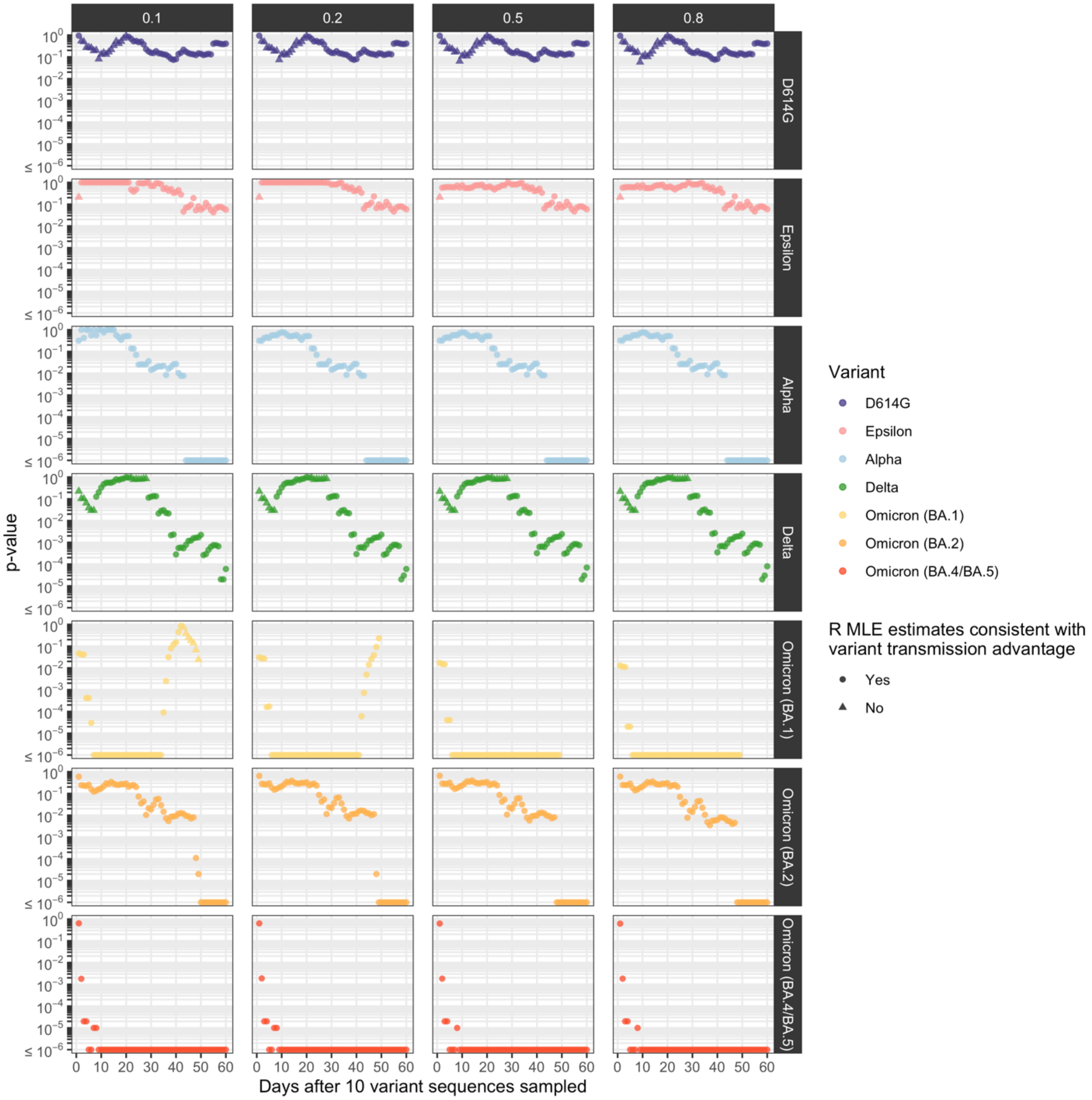

Figure S13: Sensitivity analysis varying our assumption regarding the fraction of infection detected as cases (different panels) on the p-values for variant transmission advantage in WA state. P-values over time since collection of 10 variant sequences for different SARS-CoV-2 variants during the COVID-19 pandemic in Washington state exploring different assumptions regarding the fraction of infections detected (columns). We considered maximum likelihood estimated (MLE) to be consistent with a variant transmission advantage if the estimated reproduction number of the variant was higher than that of the non-variant.

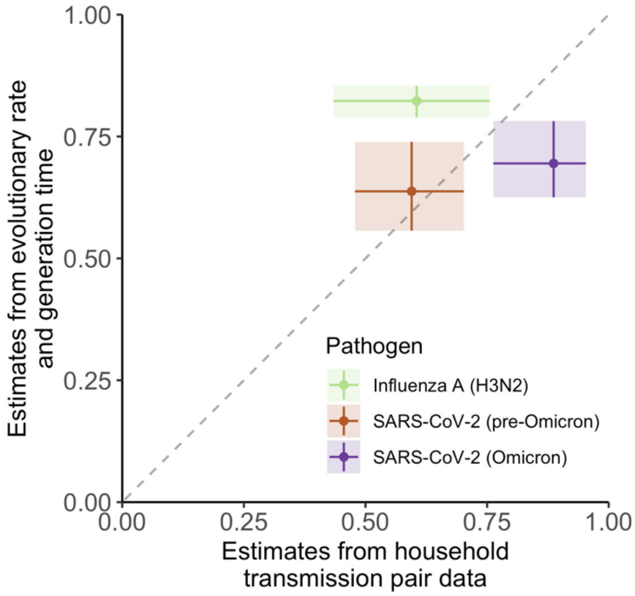

Figure S14: Comparison between estimates of *p* obtained from household transmission pair data and from assumptions regarding the evolutionary rate and the generation time. Vertical segments correspond to our uncertainty ranges around *p* estimates. Horizontal segments correspond to 95% binomial confidence intervals around the proportions obtained from transmission pair data.

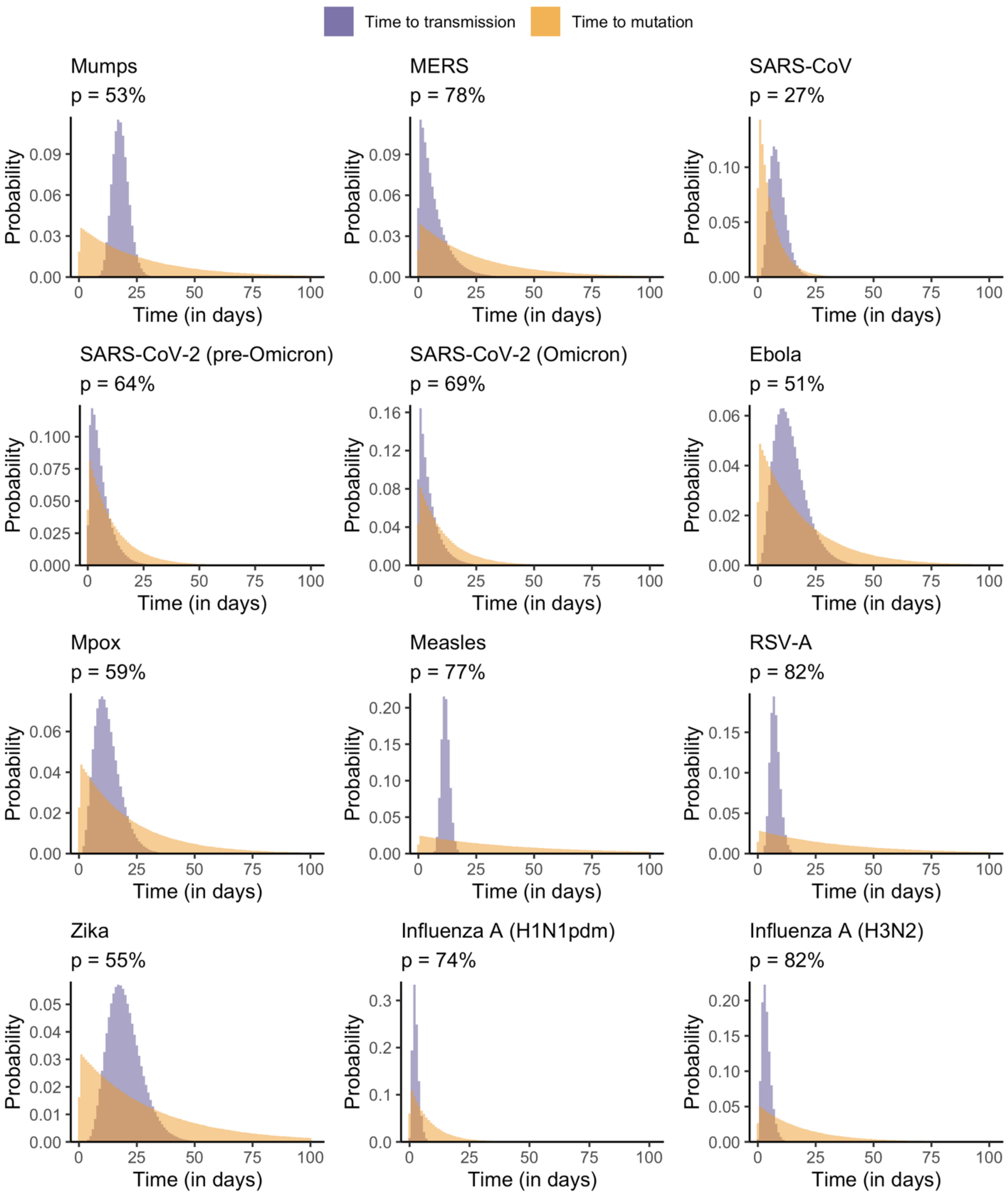

Figure S15: Comparison of the distribution of the time to occurrence of a first mutation and the time to transmission for different pathogens. For each pathogen, we additionally report the estimated probability *p* that transmission occurs before mutation. Here, we depict the simulations corresponding to the central estimate for the mutation rate and the generation time.

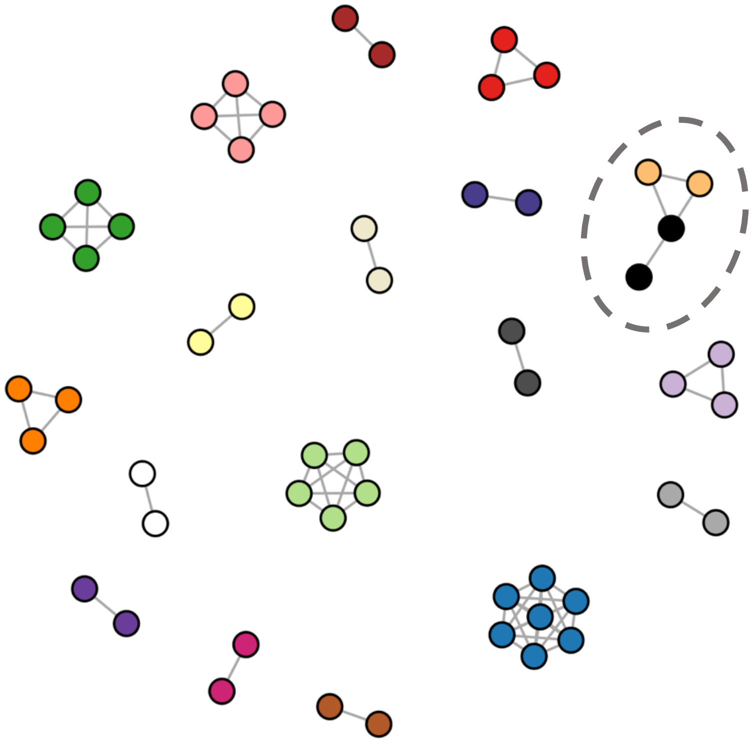

Figure S16: Difference between identical sequences obtained from the distance matrix and the reconstructed clusters of identical sequences for MERS-CoV sequences. Each vertex corresponds to a MERS-CoV sequence. Vertices are connected if their pairwise distance is equal to 0. Vertices have the same colour 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.

Supplementary text A – Impact of infectious duration and transmission bottleneck size on the proportion of transmission pairs with identical consensus sequences

In this manuscript, we assumed that the proportion of infectees that have the same consensus sequence as their infector could be approximated by the probability that a transmission event occurs before a mutation event. In this section, we report a simulation exercise that was conducted to evaluate the relevance of this assumption when varying the transmission bottleneck size and the duration of infectiousness.

Description of the simulations

We used the SEEDY R package developed by Worby and Read (23) to simulate deep-sequencing in transmission pairs under different assumptions regarding the transmission bottleneck size (1, 2, and 10) and the generation time (6, 12, 100 or 200 days). We used these transmission pairs to compute the fraction of pairs that had the same consensus genome and compared this with the probability that a transmission event occurs before a mutation one.

We considered the spread of a pathogen with a genome of length 10,000 kb. We assumed a pathogen equilibrium within-host population size of 1000 and that the pathogen would undergo 10 generations per day. We assumed that mutations occur on average every 12 days. We assumed that individuals were sequenced at a random time between when they were infected and when they infect the recipient. For each scenario, we generated 1200 transmission pairs.

Relationship between frequency of a mutant in the donor (infector) and the recipient (infectee)

Figure S17 depicts the relationship between the frequency of a mutant allele in the donor of a transmission pair and the recipient. As transmission bottleneck size increases, we observe more points inside the unity square (square between (0,0) (0,1) (1,1) and (1,0)). This is consistent with more intra-host variant being passed from the donor to the recipient.

For each of these scenarios, we computed the fraction of transmission pairs with identical consensus sequences and compared this to the theoretical probability that a transmission event occurs before a mutation one (dashed horizontal lines in Figure S18). For narrow transmission bottlenecks of size 1, the theoretical probability value approximates well the proportion of pairs with identical consensus sequences. As transmission bottleneck size widens, this theoretical probability no longer approximates well the proportion of pairs with identical consensus sequences. When simulating the spread and evolution of a pathogen characterized by a longer generation time (200 days), we found that the proportion of transmission pairs with identical consensus genomes differed slightly from that expected from the theoretical probability that a transmission event occurs before a mutation event.

To conclude, this simulation study shows that the probability that transmission occurs before mutation is a good proxy for the proportion of pairs with identical consensus genomes for pathogens characterized by narrow transmission bottlenecks, relatively short infectious durations and limited within host-diversity.

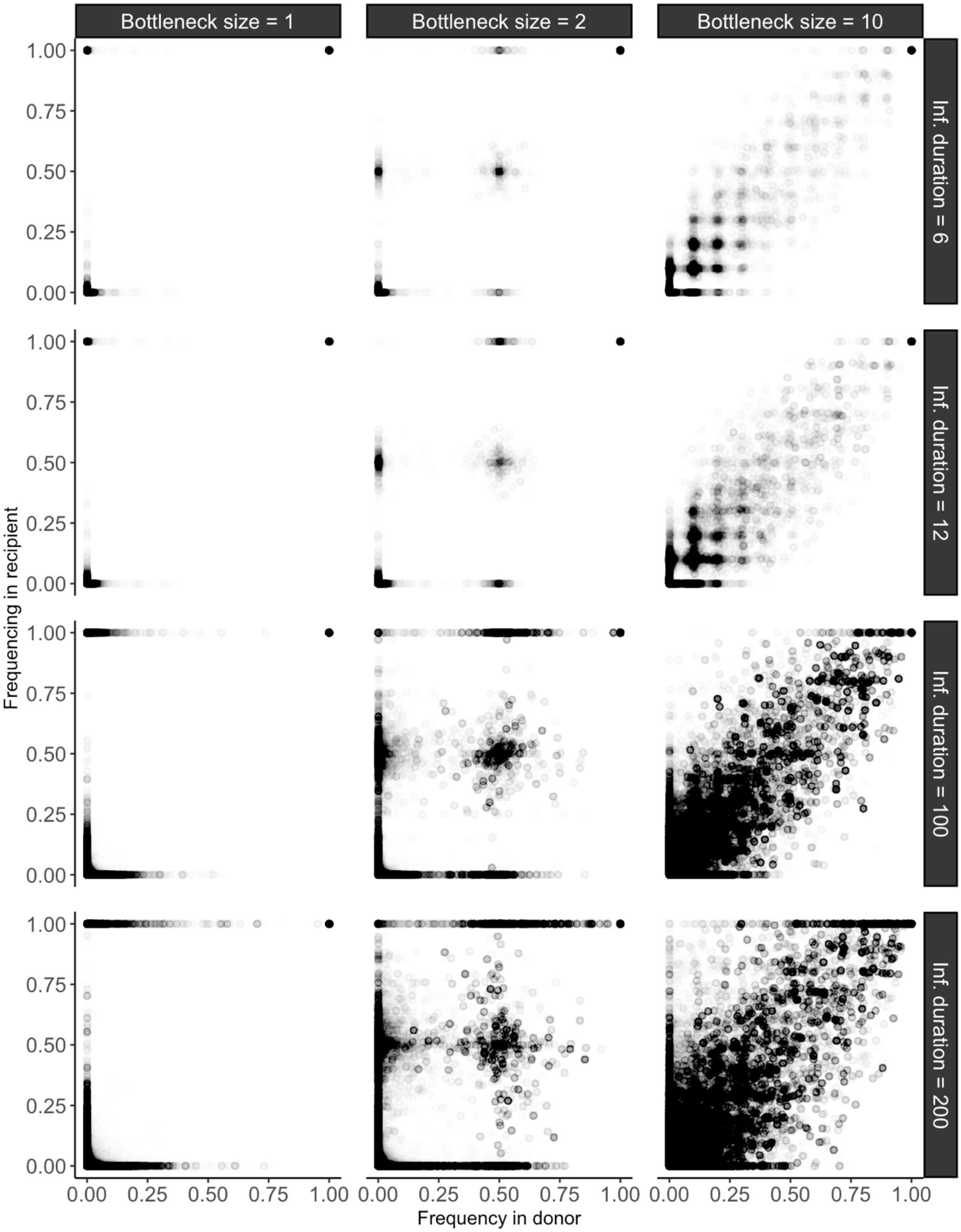

Figure S17: Relationship between the frequency of a mutant in the donor (infector) and the recipient (infectee) of a transmission pair. Different assumptions regarding the transmission bottleneck size (different columns) and the generation time (in days – different rows) are explored.

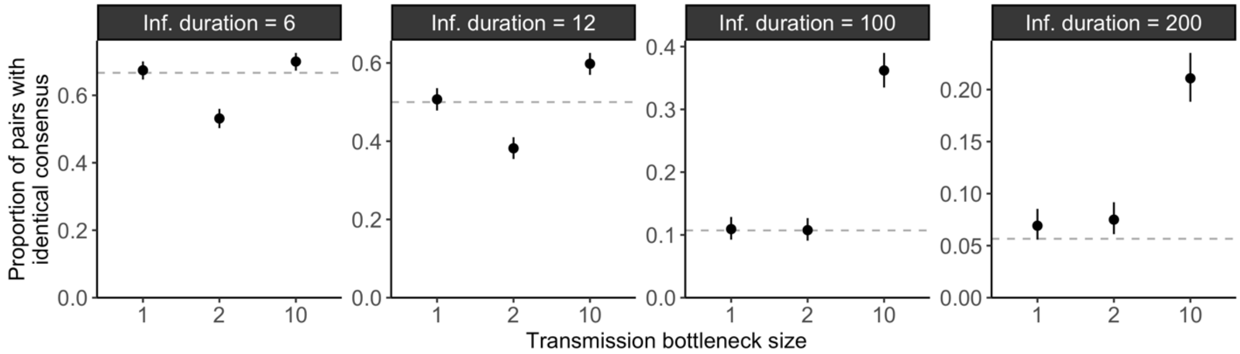

Figure S18: Proportion of transmission pairs with identical consensus sequences exploring different assumptions regarding the transmission bottleneck size and the disease generation time (“inf. duration” in days). Vertical segments correspond to 95% confidence intervals. Each point was obtained by generating 1,200 transmission pairs. The horizontal dotted lines correspond to the probability that a transmission event occurs before a mutation event.

Supplementary text B - Inference of transmission parameters conditional on cluster extinction

In the main text, we showed that our inference framework provides unbiased estimates of both R and k when the mean number of offspring with identical sequences is lower than 1 ($R \cdot p < 1$), corresponding to situations where cluster extinction is almost certain. Our method however becomes unreliable when the probability of cluster extinction is strictly lower than 1. This was done relying on the full distribution of clusters of identical sequences, including those that did not go extinct (which were then set to an arbitrary high threshold value). An alternative approach would consist in looking at cluster sizes conditional on extinction (24, 25). Previous theoretical work has indeed shown that a supercritical epidemic process where extinction is uncertain (characterized by *R > 1*) can be mapped to a subcritical counterpart characterized by a mean number of offspring lower than 1 (*R < 1*) and the same dispersion parameter (24).

In the following paragraphs, we show how our inference framework could be adapted to look at the size of clusters of identical sequences conditional on them having gone extinct. We then evaluate the performance of this adapted statistical framework and highlight some remaining challenges for real-world applications.

Distribution of the size of clusters of identical sequences conditional on extinction

We used the formalism introduced by Waxman and Nouvellet (24) to describe the size distribution of clusters of identical pathogen sequences conditional on extinction. They showed that, conditional on extinction, supercritical and subcritical dynamics (respectively characterized by a reproduction number below and above 1) cannot be distinguished. Waxman and Nouvellet had characterized the size of finite disease outbreaks. Here, we instead consider finite *mutation-less* outbreaks, *i.e.* clusters of infected individuals characterized by the same pathogen sequence.

Let ϵ denote the probability of extinction for a cluster of identical pathogen sequences. If the mean number of offspring with identical sequences is lower than 1, ϵ is equal to 1. Otherwise, ϵ is lower than 1. Following Waxman and Nouvellet, we introduce R_s, as the reproduction number associated with clusters of identical sequences that got extinct. In the following, we refer to R_s as the subcritical reproduction number. We have the following relationship between the reproduction number R and R_s:

$$R_s = R \cdot \epsilon^{1+\frac{1}{k}} \qquad (*)$$

where k is the dispersion parameter of the offspring distribution. Figure S19 depicts how the subcritical reproduction number R_s is impacted by the reproduction number R, the dispersion parameter k and the probability that transmission occurs before mutation. As Waxman and Nouvellet note, the subcritical reproduction number R_s mirrors the supercritical one R. This means that inferring the subcritical reproduction number R_s enables to infer the reproduction number R as there is a direct relationship between the two (Figure S19).

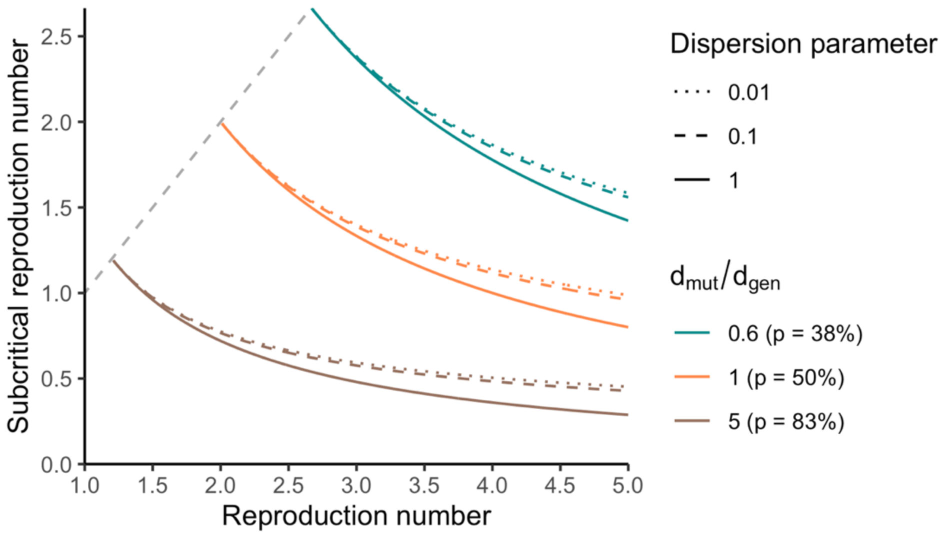

Figure S19: Relationship between the reproduction number *R* and the subcritical reproduction number *R*$_s$ for different probabilities *p* that transmission occurs before mutation and different values of the dispersion parameter *k*. Colored lines correspond to reproduction numbers lying above the 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).

The probability $r_j^{extinct}$ for a cluster of identical sequences to be of size j conditional on extinction is equal to (24, 25):

$$r_j^{extinct} = \frac{\Gamma(kj + j - 1)}{\Gamma(kj) \cdot \Gamma(j + 1)} \cdot \frac{\left(\frac{pR_s}{k}\right)^{j-1}}{\left(1 + \frac{pR_s}{k}\right)^{kj+j-1}}$$

and the probability r_j for a cluster of identical sequences of being of size j is thus equal to:

$$r_j = r_j^{extinct} \cdot \epsilon \qquad (**)$$

More specifically, we note that $R_s < 1/p$. In the specific situation where $R < 1/p$, we have $R = R_s$ and $r_j^{extinct} = r_j$.

Inference from the size of clusters of identical sequences conditional on extinction

Assuming we have a dataset comprised of the size of clusters of identical sequences that got extinct, we can hence infer the value of the subcritical reproduction number R_s and the dispersion parameter by using the updated formula $(**)$ for the probability of cluster of identical sequences of being of size j in the derivation of the likelihood. Implementing this updated framework, we then obtained maximum likelihood estimates of R_s and k by imposing values of the reproduction number ranging between 0.01 and $1/p$ and values of the dispersion parameter ranging between 0.001 and 10.0.

We evaluated our inference framework on synthetic cluster data generated using a branching process with mutation (see main text). Clusters were simulated until reaching a maximum size of

10,000. We then considered that clusters who reached 10,000 had not gone extinct and applied our inference framework to the subset that got extinct (size < 10,000). Figure S20-S22 shows that we were able to recover the expected value of the dispersion parameter and the subcritical reproduction number R_s.

Assuming prior knowledge on whether the reproduction number lies above or below the threshold of $1/p$, the estimated subcritical reproduction number can either directly be interpreted as the reproduction number of the outbreak (below the threshold) or mapped to a corresponding reproduction number higher than $1/p$ (using equation $(*)$, Figure S19)

Challenges for the application to real-world data

In the previous paragraphs, we introduced an alternative approach to characterize the disease offspring distribution when the mean number of offspring with identical genomes is higher than 1. By restricting the analysis to clusters of identical sequences that got extinct, we showed that we could accurately infer the reproduction number and the dispersion parameter.

However, we acknowledge that determining in practice whether a cluster of identical sequences has become extinct may be challenging. Furthermore, we assumed here that the epidemiological process under study was stationary (*i.e.* that the reproduction number and the dispersion parameter are constant throughout the study period). In practice, behaviour changes, the implementation of interventions or the depletion of the susceptible population as the epidemic progresses can modify the effective reproduction number. This is likely especially problematic for reproduction numbers greater than 1 (and by extension above the threshold of $1/p$). Overall, further work is warranted to estimate the offspring distribution's parameters above the threshold of $1/p$ from real-world data describing the size of clusters of identical sequences.

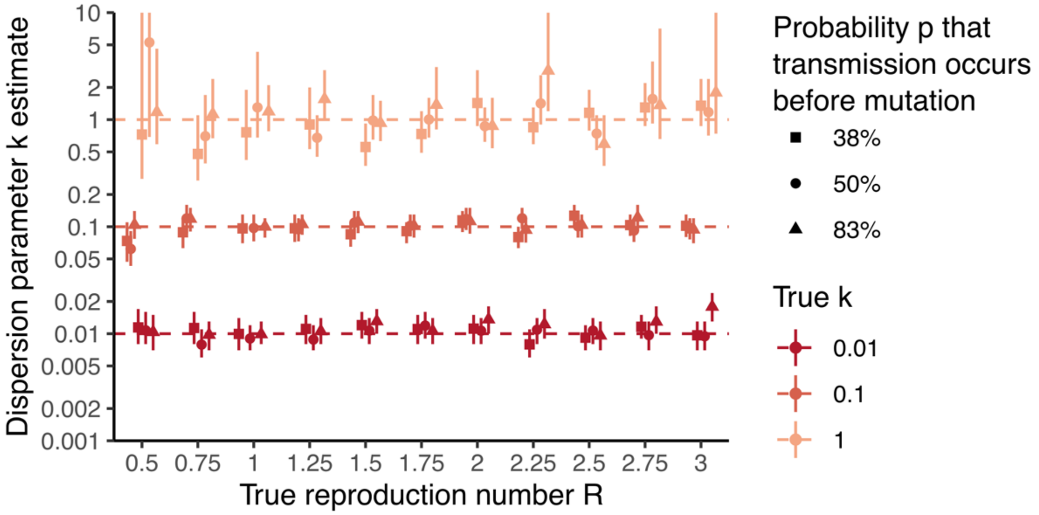

Figure S20: Estimates of the dispersion parameter *k* using the size of clusters that got extinct. Estimates are reported as a function of the true reproduction number *R* used to generate synthetic clusters. Point estimates correspond to maximum-likelihood estimates and vertical segments to 95% likelihood profile confidence intervals obtained from analyzing 1000 synthetic clusters of identical sequences.

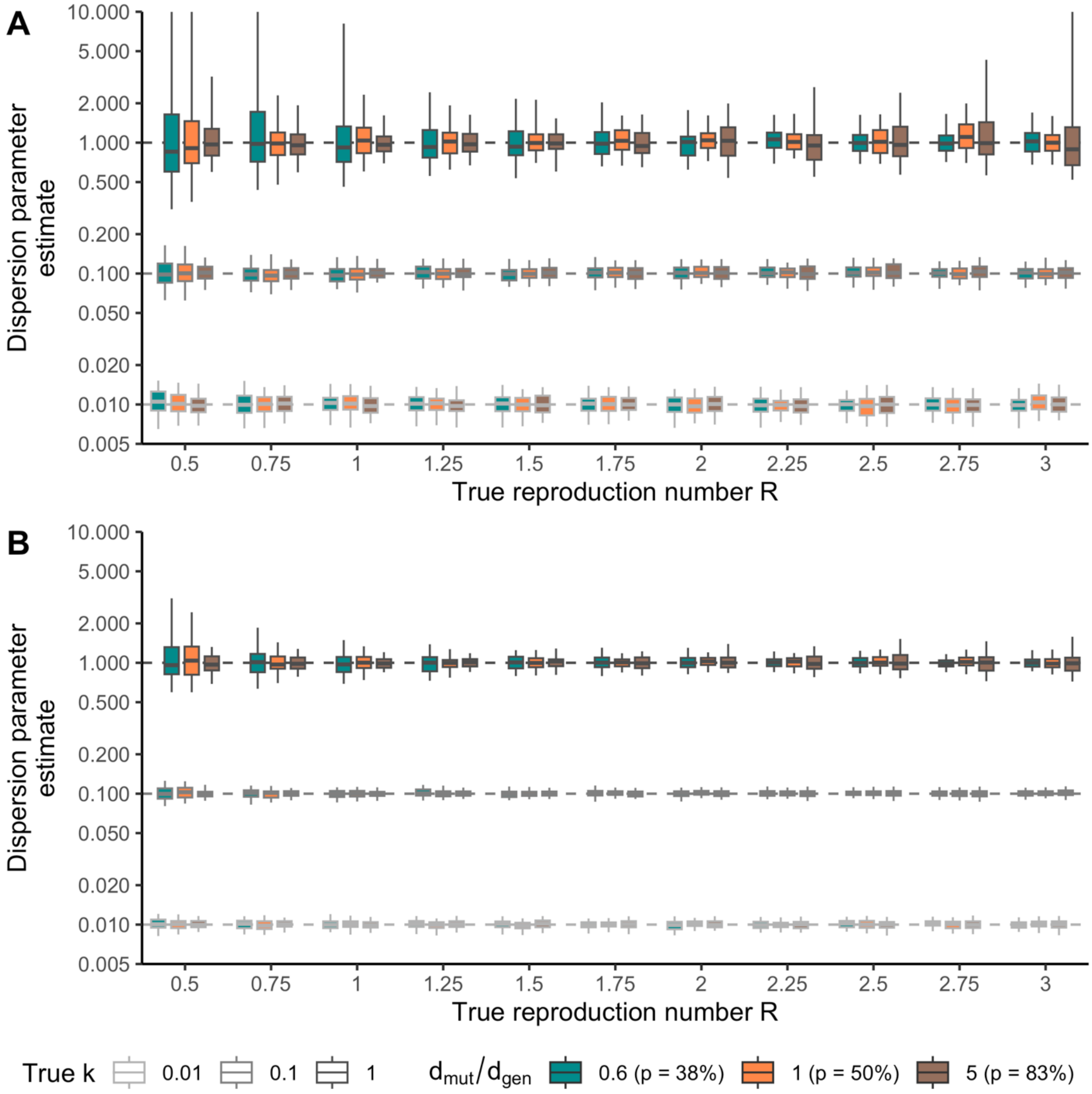

Figure S21: Dispersion parameter estimates as a function of the true reproduction number when using the inference framework relying on cluster size distribution conditional on cluster extinction. A. Using a dataset comprised of 1,000 clusters of identical sequences. **B.** Using a dataset comprised of 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.

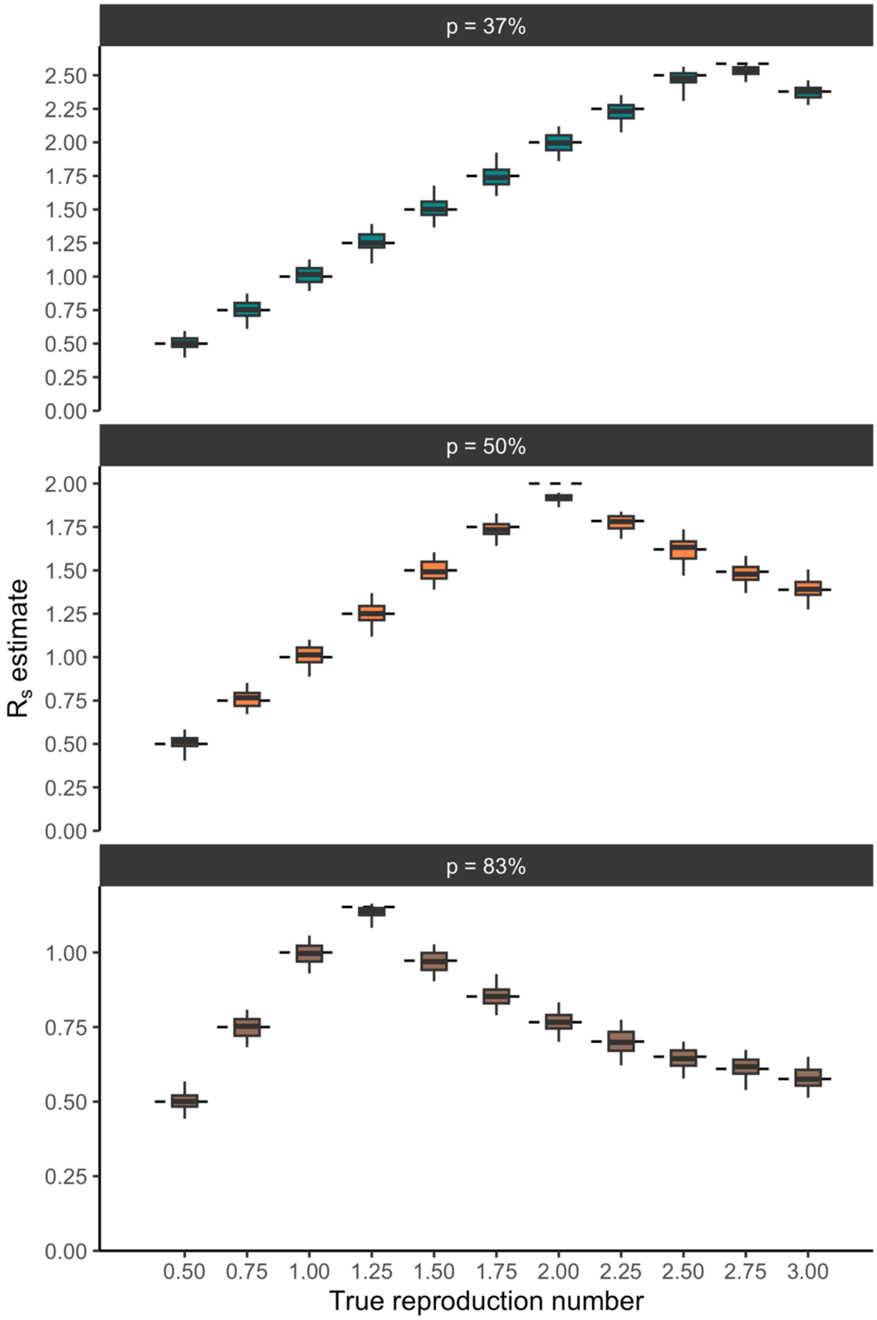

Figure S22: Subcritical reproduction number *R_s* estimates as a function of the true reproduction number when using the inference framework relying on cluster size distribution conditional on cluster extinction. Each boxplot represents the distribution of *R_s* maximum likelihood estimates across 100 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).

References

1. S. Cauchemez, *et al.*, Unraveling the drivers of MERS-CoV transmission. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 9081–9086 (2016).

2. Z. Zhang, L. Shen, X. Gu, Evolutionary dynamics of MERS-CoV: Potential recombination, positive selection and transmission. *Sci. Rep.* **6** (2016).

3. M. A. Vink, M. C. J. Bootsma, J. Wallinga, Serial intervals of respiratory infectious diseases: A systematic review and analysis. *Am. J. Epidemiol.* **180**, 865–875 (2014).

4. , Nextstrain - Real-time tracking of measles virus evolution.

5. O. Faye, *et al.*, Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. *Lancet Infect. Dis.* **15**, 320–326 (2015).

6. G. Dudas, T. Bedford, The ability of single genes vs full genomes to resolve time and space in outbreak analysis. *BMC Evol. Biol.* **19**, 232 (2019).

7. N. M. Ferguson, *et al.*, EPIDEMIOLOGY. Countering the Zika epidemic in Latin America. *Science* **353**, 353–354 (2016).

8. N. R. Faria, *et al.*, Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* **546**, 406–410 (2017).

9. G. Guzzetta, *et al.*, Early estimates of Monkeypox incubation period, generation time, and reproduction number, Italy, may-June 2022. *Emerg. Infect. Dis.* **28**, 2078–2081 (2022).

10. M. I. Paredes, *et al.*, Early underdetected dissemination followed by extensive local transmission propelled the 2022 mpox epidemic and limited impact of vaccination. *bioRxiv* (2023) https:/doi.org/10.1101/2023.07.27.23293266.

11. S. Cauchemez, *et al.*, Household transmission of 2009 pandemic influenza A (H1N1) virus in the United States. *N. Engl. J. Med.* **361**, 2619–2627 (2009).

12. J. Hedge, S. J. Lycett, A. Rambaut, Real-time characterization of the molecular epidemiology of an influenza pandemic. *Biol. Lett.* **9**, 20130331 (2013).

13. S. Cauchemez, F. Carrat, C. Viboud, A. J. Valleron, P. Y. Boëlle, A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Stat. Med.* **23**, 3469–3487 (2004).

14. L. Tan, *et al.*, The comparative genomics of human respiratory syncytial virus subgroups A and B: genetic variability and molecular evolutionary dynamics. *J. Virol.* **87**, 8213–8226 (2013).

15. F. Campbell, C. Strang, N. Ferguson, A. Cori, T. Jombart, When are pathogen genome sequences informative of transmission events? *PLoS Pathog.* **14**, e1006885 (2018).

16. Z. Zhao, *et al.BMC Evol. Biol.* **4**, 21 (2004).

17. W. S. Hart, *et al.*, Inference of the SARS-CoV-2 generation time using UK household data. *Elife* **11** (2022).

18. S. Duchene, *et al.*, Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evol.* **6**, veaa061 (2020).

19. S. Abbott, K. Sherratt, M. Gerstung, S. Funk, Estimation of the test to test distribution as a proxy for generation interval distribution for the Omicron variant in England. *bioRxiv* (2022) https:/doi.org/10.1101/2022.01.08.22268920.

20. K. Ito, C. Piantham, H. Nishiura, Estimating relative generation times and reproduction numbers of Omicron BA.1 and BA.2 with respect to Delta variant in Denmark. *Math. Biosci. Eng.* **19**, 9005–9017 (2022).

21. ,nextstrain.org/measles (16 December 2022).

22. M. Pacenti, *et al.*, Measles virus infection and immunity in a suboptimal vaccination coverage setting. *Vaccines (Basel)* **7**, 199 (2019).

23. C. J. Worby, T. D. Read, 'SEEDY' (simulation of evolutionary and epidemiological dynamics): An R package to follow accumulation of within-host mutation in pathogens. *PLoS One* **10**, e0129745 (2015).

24. D. Waxman, P. Nouvellet, Sub- or supercritical transmissibilities in a finite disease outbreak: Symmetry in outbreak properties of a disease conditioned on extinction. *J. Theor. Biol.* **467**, 80–86 (2019).

25. A. Cori, *et al.*, A graph-based evidence synthesis approach to detecting outbreak clusters: An application to dog rabies. *PLoS Comput. Biol.* **14**, e1006554 (2018).