Supplementary materials for:

Fine-scale spatial and social patterns of SARS-CoV-2 transmission from identical pathogen sequences

List of supplementary figures

List of supplementary tables

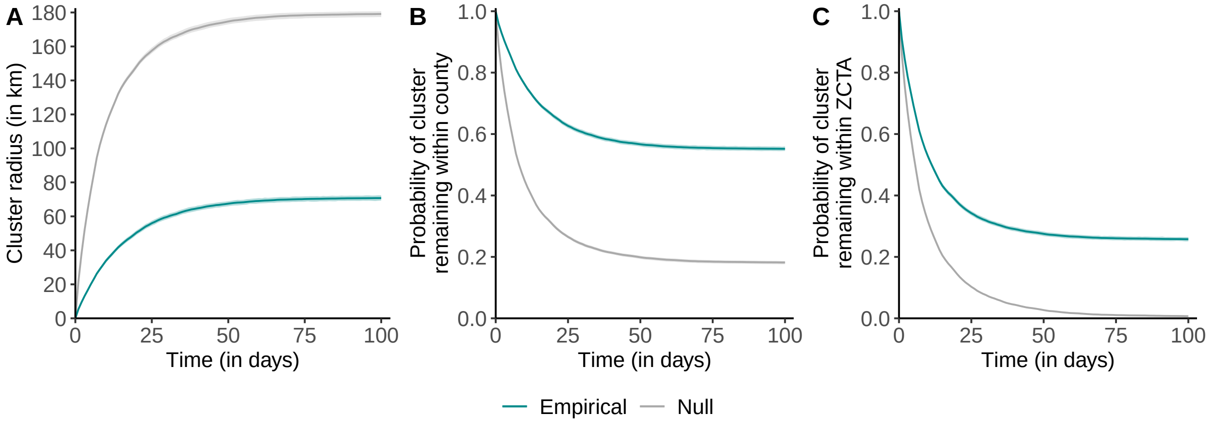


Figure S1. Permutation test to explore the spatio-temporal signal in clusters of identical SARS-CoV-2 sequences. A. Radius of clusters of identical sequences as a function of time since first sequence collection. **B.** Probability for all sequences within a cluster of identical sequences of remaining in the same county as a function of time since first sequence collection. **C.** Probability for all sequences within a cluster of identical sequences of remaining in the same ZCTA as a function of time since first sequence collection. The grey shaded areas correspond to 95% confidence intervals of a null distribution generated from 100 simulation where the geographical location of sequences from WA Sentinel surveillance are permuted. The grey lines correspond to the medians the simulated null distributions.

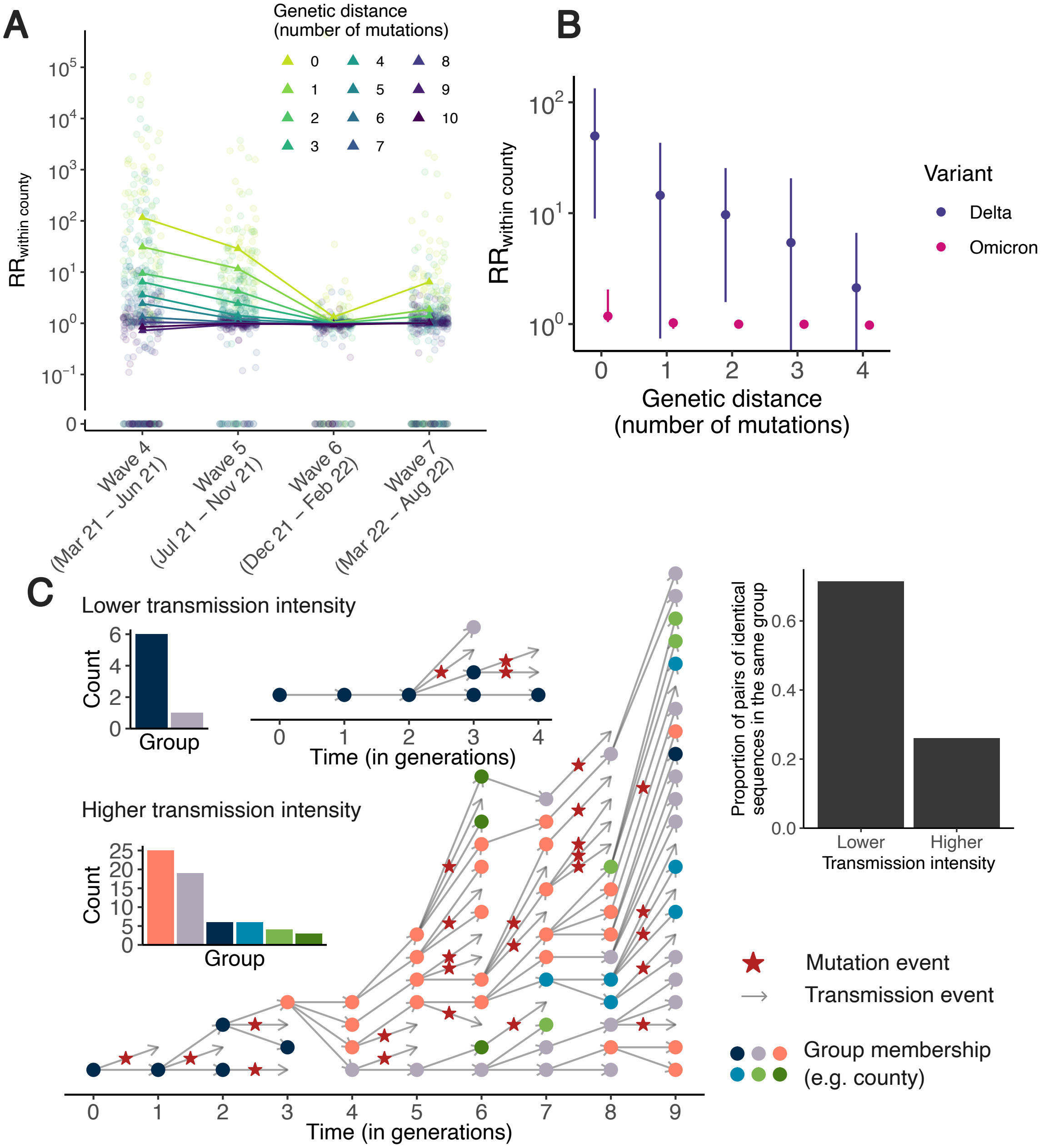


Figure S2. The magnitude of the relative risk of observing sequences at a given genetic distance within the same county is impacted by transmission intensity. A. Relative risk of observing sequences at a given genetic distance within the same county across multiple epidemic waves. We defined waves as: March 2021-June 2021 (Wave 4), July 2021-November 2021 (Wave 5), December 2021-February 2022 (Wave 6) and March 2022-August 2022 (Wave 7). In A, circular points correspond to individuals counties and triangles correspond to the median across counties. **B.** Median relative risk of observing pairs sequences within the same county (with IQR) as a function of genetic distance stratified by variant during Wave 6. **C.** A higher transmission intensity results in larger clusters of identical sequences that tend to be more mixed across groups. In C, the two clusters are simulated using a branching process with mutation [8] by assuming the probability for an infector and an infectee to have the same consensus sequence equal to 0.69 and a probability for an infectee of being in the same groups as its infector of 0.7. We consider a reproduction number of 1.2 for the lower transmission intensity scenario and of 2.0 in the higher transmission intensity scenario.

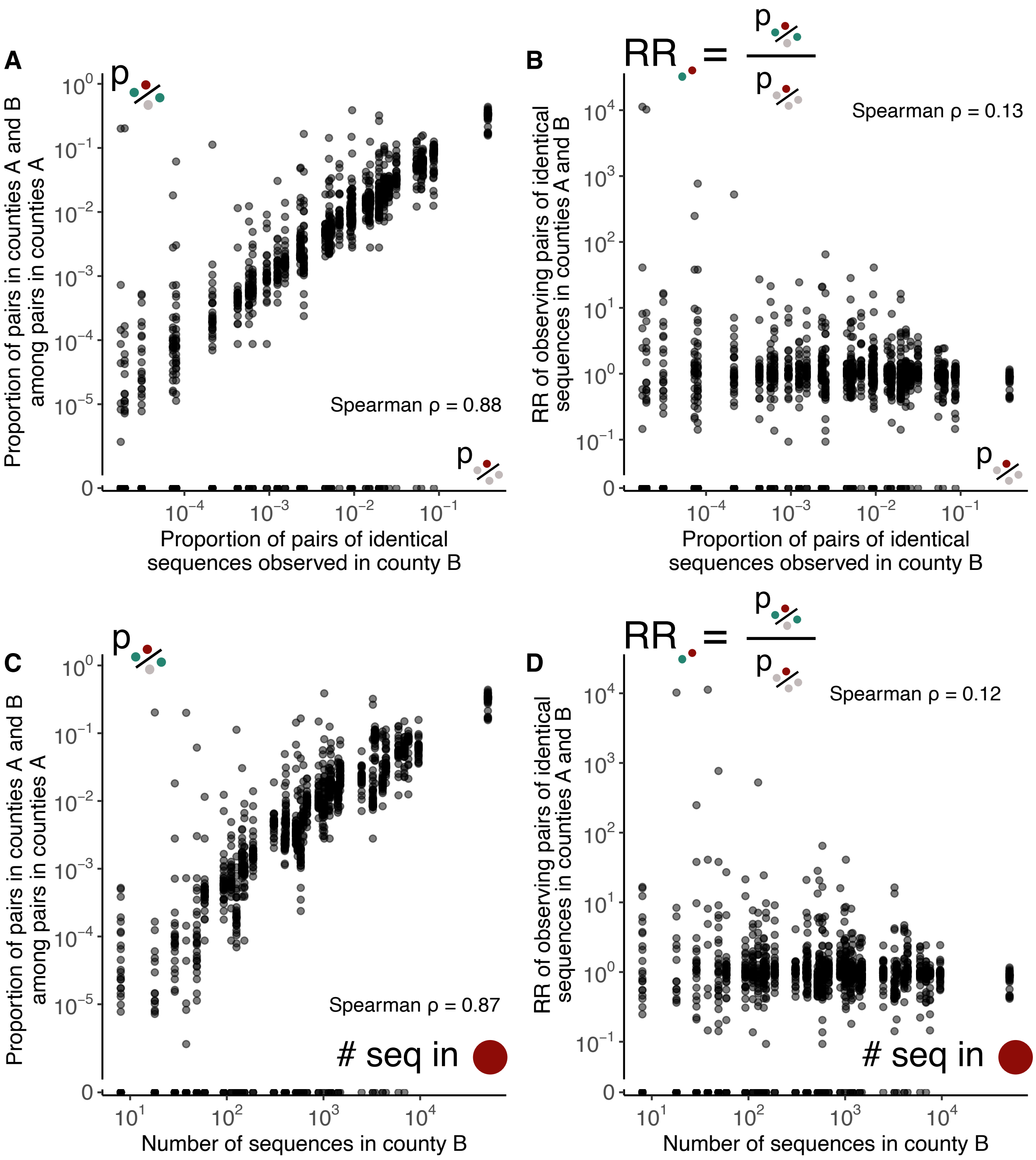


Figure S3. Our measure of relative risk corrects for uneven sequencing between regions. **A.** Proportion of pairs of identical sequences shared between counties A and B among pairs observed in county A as a function of the proportion of pairs of identical sequences observed in county B. **B.** Relative risk for pairs of identical sequences of being observed in counties A and B as a function of the proportion of pairs of identical sequences observed in county B. **C.** Proportion of pairs of identical sequences shared between counties A and B as a function of the number of sequences available in county B. **D.** Relative risk for pairs of identical sequences of being observed in counties A and B as a function of the number of sequences available in county B.

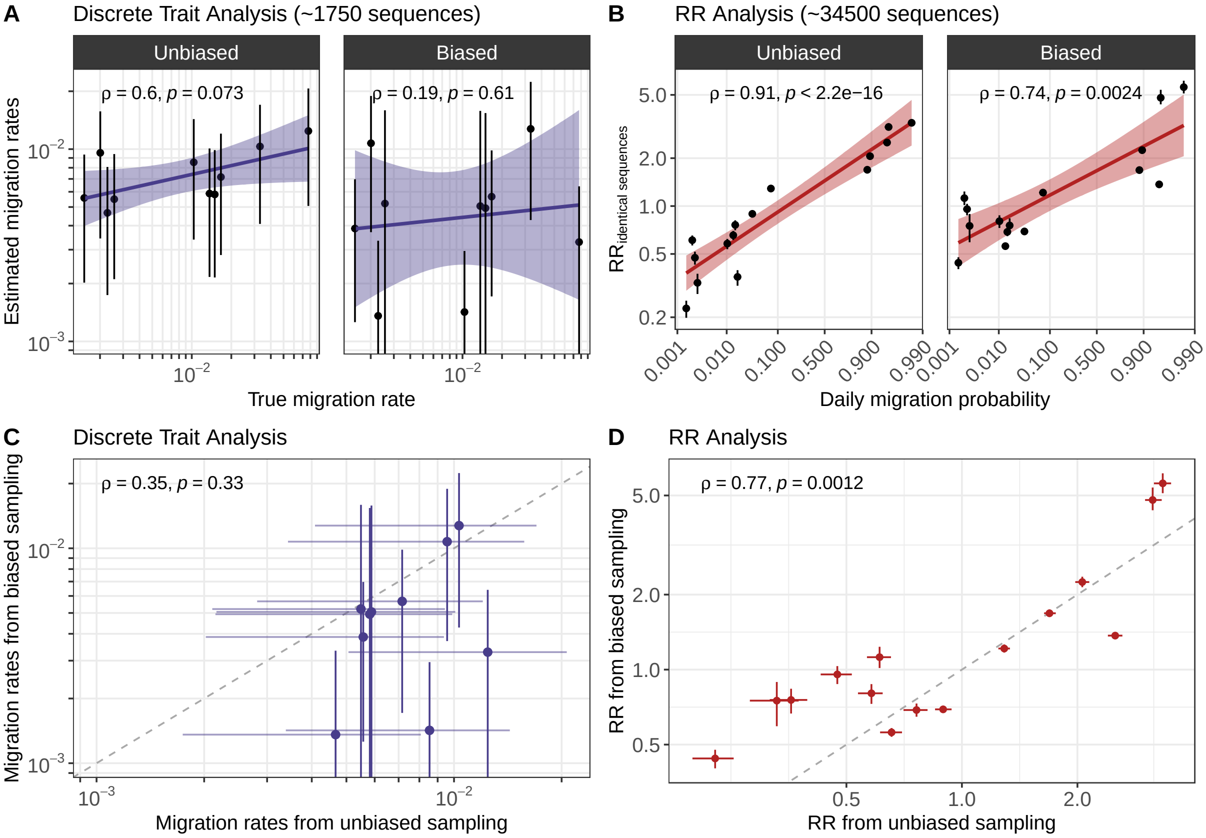


Figure S4. Simulation study exploring the impact of sequencing bias on results from a discrete trait analysis and from our RR framework. **A.** Comparison between migration rates estimated from a discrete trait analysis and the true migration rates used to simulate the sequence data. **B.** Comparison between the relative risk of observing identical sequences between two demes and the weekly migration probability between demes. **C.** Comparison between migration rates inferred from a sequence dataset generated in a biased sampling and an unbiased sampling scenario. **D.** Comparison between the relative risk of observing identical sequences in two groups from a sequence dataset generated in a biased sampling and an unbiased sampling scenario. For the RR, segments indicate 95% subsampling confidence intervals. For the migration rates, segments indicate 95% highest posterior density intervals. For each plot, we indicate the Spearman correlation coefficient (and the associated p-value).

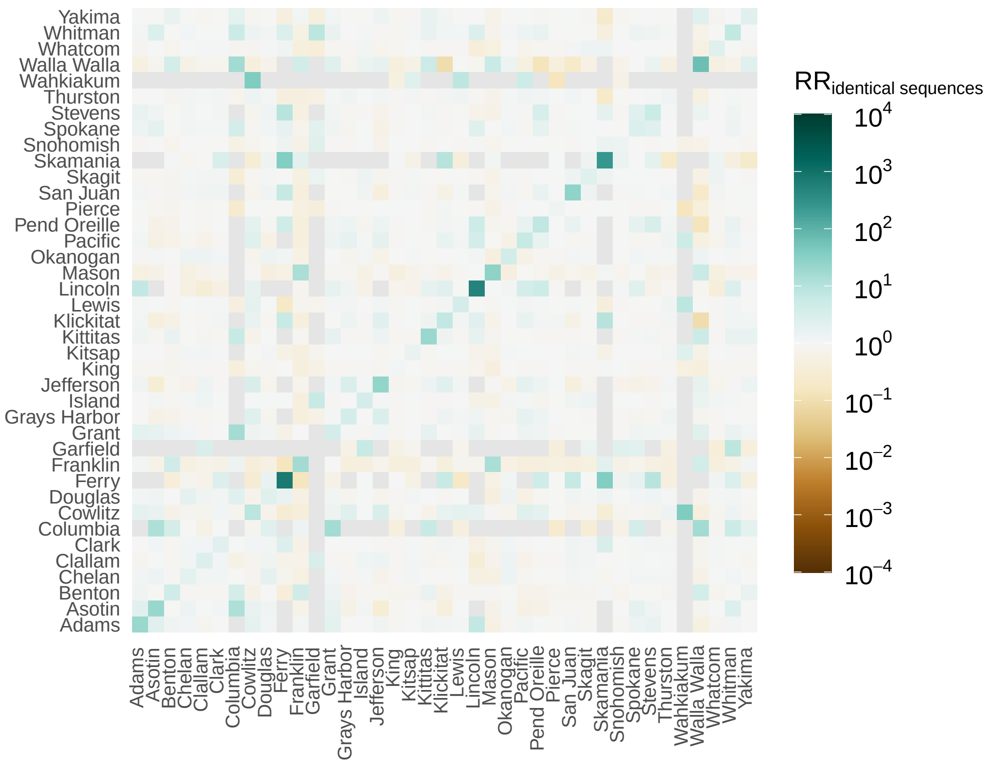
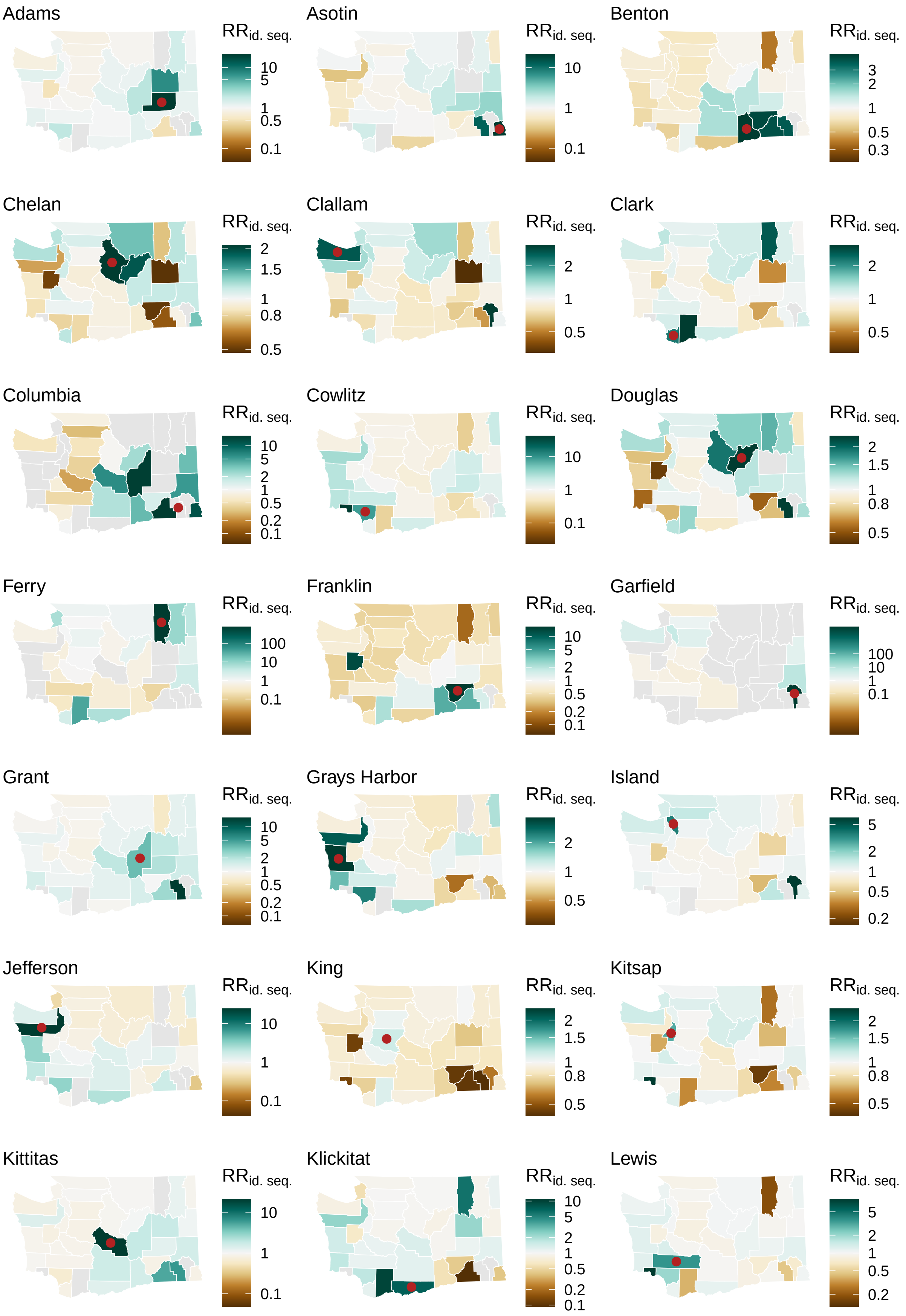


Figure S5. Relative risk of observing identical sequences in two counties. Grey squares correspond to pairs of counties between which no pairs of identical sequences were observed during the study period.


Adams
RRid. seq.
10
5
1
0.5
0.1
Asotin
RRid. seq.
10
1
0.1
Benton
RRid. seq.
3
2
1
0.5
0.3
Chelan
RRid. seq.
2
1.5
1
0.8
0.5
Clallam
RRid. seq.
2
1
0.5
Clark
RRid. seq.
2
1
0.5
Columbia
RRid. seq.
10
5
2
1
0.5
0.2
0.1
Cowlitz
RRid. seq.
10
1
0.1
Douglas
RRid. seq.
2
1.5
1
0.8
0.5
Ferry
RRid. seq.
100
10
1
0.1
Franklin
RRid. seq.
10
5
2
1
0.5
0.2
0.1
Garfield
RRid. seq.
100
10
1
0.1
Grant
RRid. seq.
10
5
2
1
0.5
0.2
0.1
Grays Harbor
RRid. seq.
2
1
0.5
Island
RRid. seq.
5
2
1
0.5
0.2
Jefferson
RRid. seq.
10
1
0.1
King
RRid. seq.
2
1.5
1
0.8
0.5
Kitsap
RRid. seq.
2
1.5
1
0.8
0.5
Kittitas
RRid. seq.
10
1
0.1
Klickitat
RRid. seq.
10
5
2
1
0.5
0.2
0.1
Lewis
RRid. seq.
5
2
1
0.5
0.2

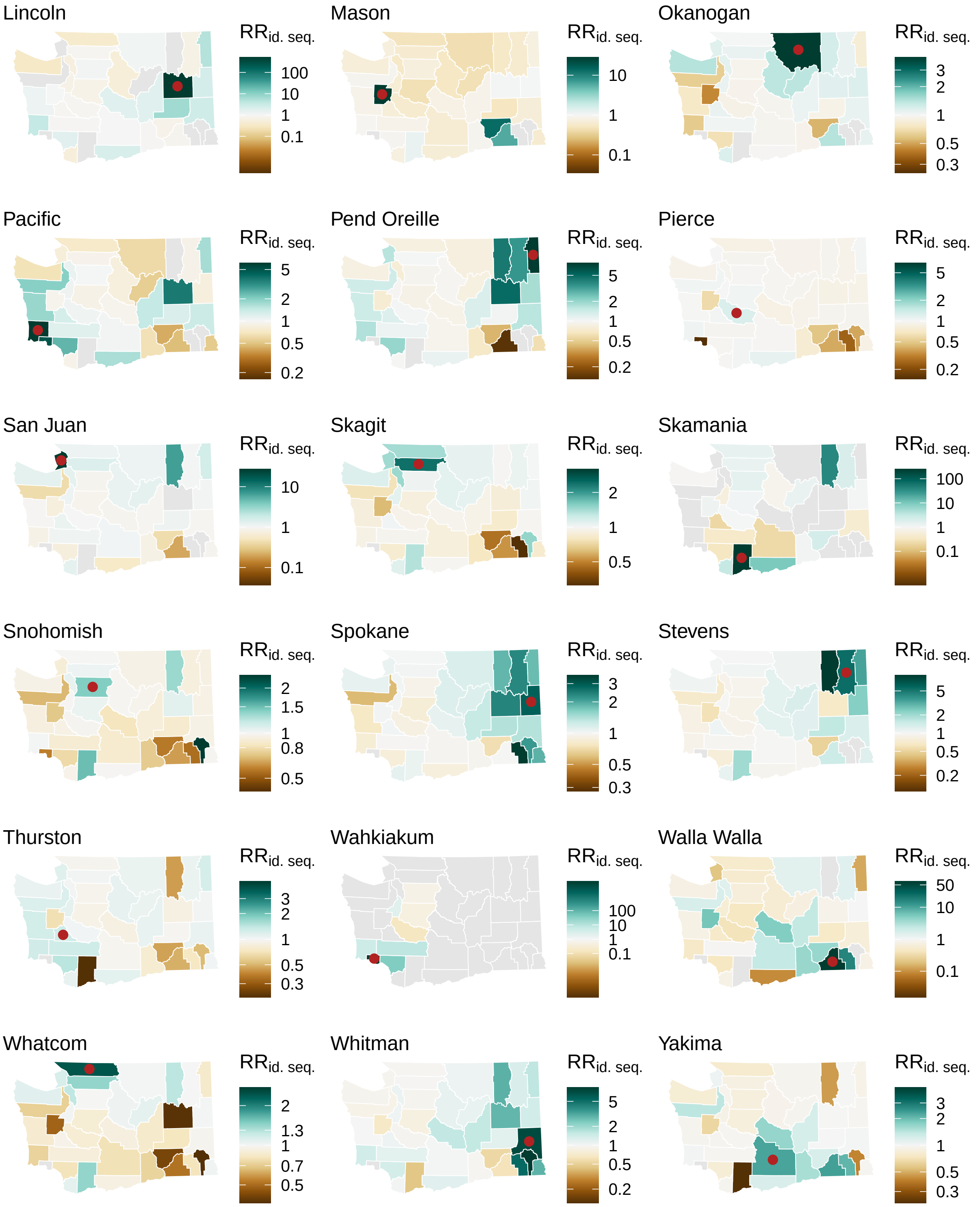


Figure S6. Relative risk of observing pairs of identical sequences between counties. On each map, we represent the relative risk of observing pairs of identical sequences in the county indicated by a red point (map title) and all the other counties in Washington state. Areas are coloured in grey when no pairs of identical sequences are observed. To increase readability, each map has its own colour scale.

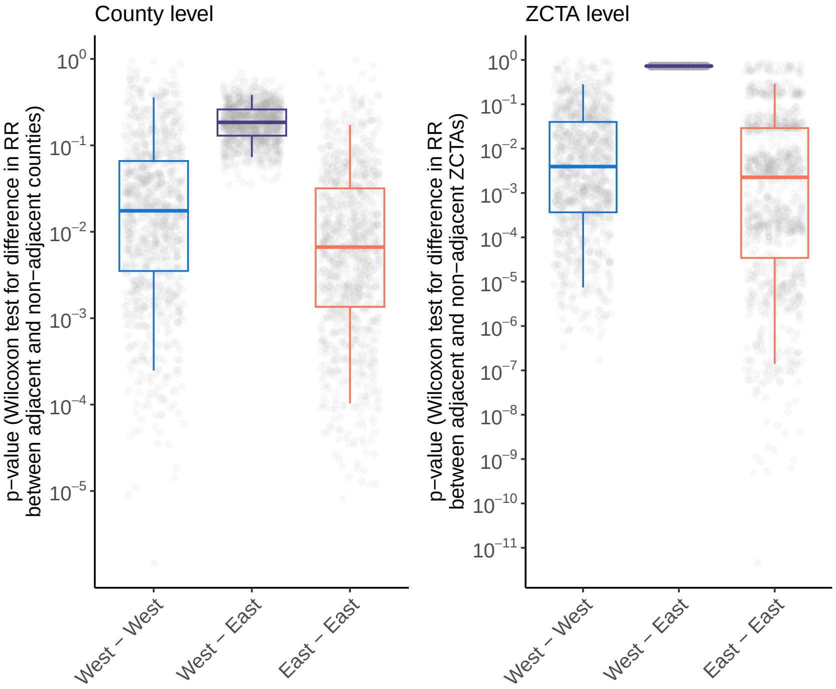


Figure S7. Impact of subsampling on the significance of the association between the relative risk of identical sequences and whether counties and ZCTAs are adjacent or not. We investigate whether our conclusions regarding the significance of the association between the relative risk of identical sequences falling in two distinct counties / ZCTAs and their adjacency status (adjacent / non-adjacent) can be impacted by the number of pairs of counties involved in the comparison (within Eastern WA, within Western WA and between Eastern and Western WA). At the county level, we subsample the pairs of counties involved in these 3 comparisons to 12 adjacent pairs of counties (number of pairs of adjacent counties between Eastern and Western WA) and 132 non-adjacent pairs of counties (number of pairs of non-adjacent counties within Western WA). This ensures that all comparisons are performed on the same number of pairs of counties. On each subsampled dataset, we compute the p-value from a Wilcoxon test evaluating differences between the relative risk of observing identical sequences in adjacent and non-adjacent counties. This is done for 1,000 subsampled datasets. Boxplots indicate the p-values obtained across these different subsampling iterations (5%, 25%, 50%, 75% and 95% quantiles). We do a similar analysis at the ZCTA level.

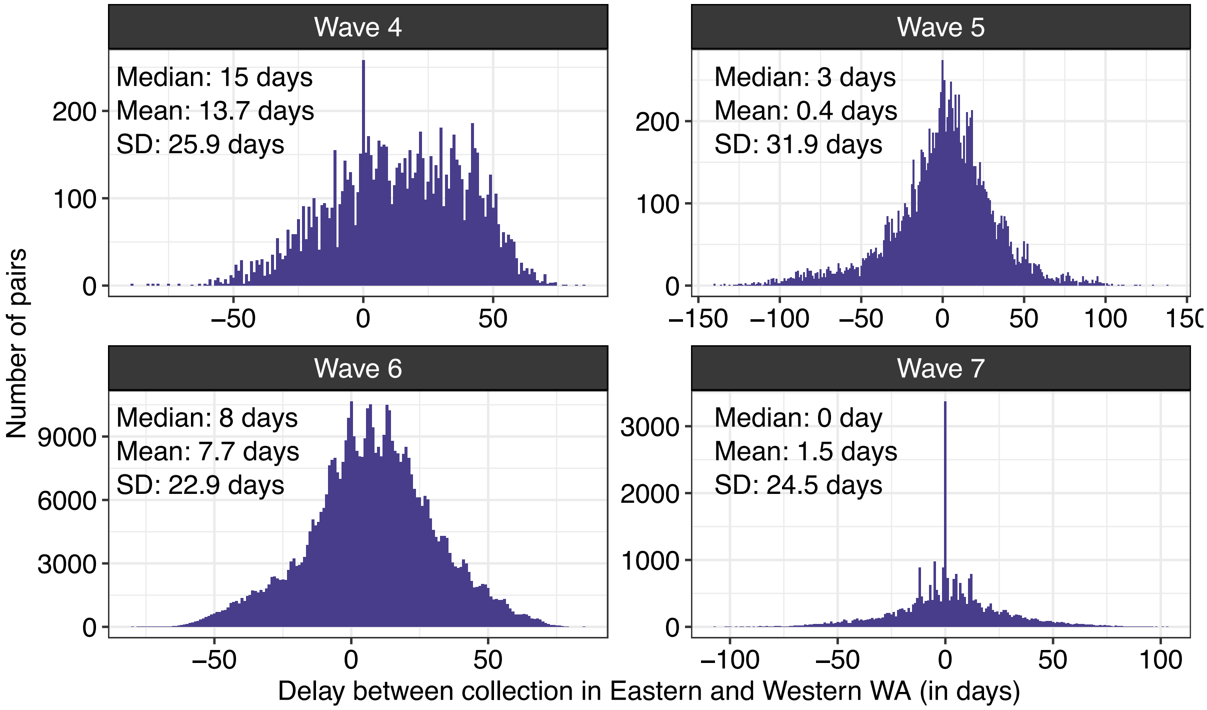


Figure S8. Distribution of the delay between sequence collection within pairs of identical sequences collected between Eastern and Western WA and across epidemic waves.

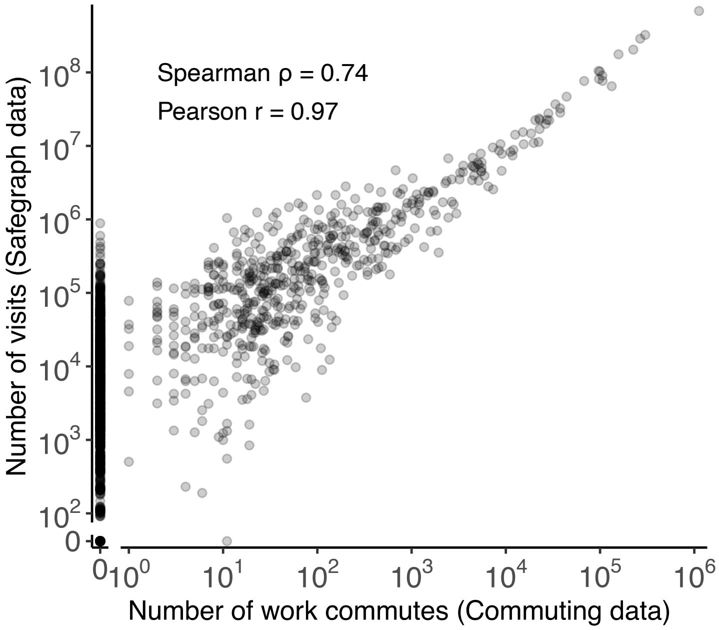


Figure S9. Comparison between the number of directed commuting flows and the number of directed visits between two counties. The number of work commutes is extracted from [17]. The number of visits is estimated using Safegraph *Weekly patterns* mobility data. The comparison is done by matching the origin county in the mobile phone data to the residence county in the workflow data and the destination county in the mobile phone data to the workplace county in the workflow data.

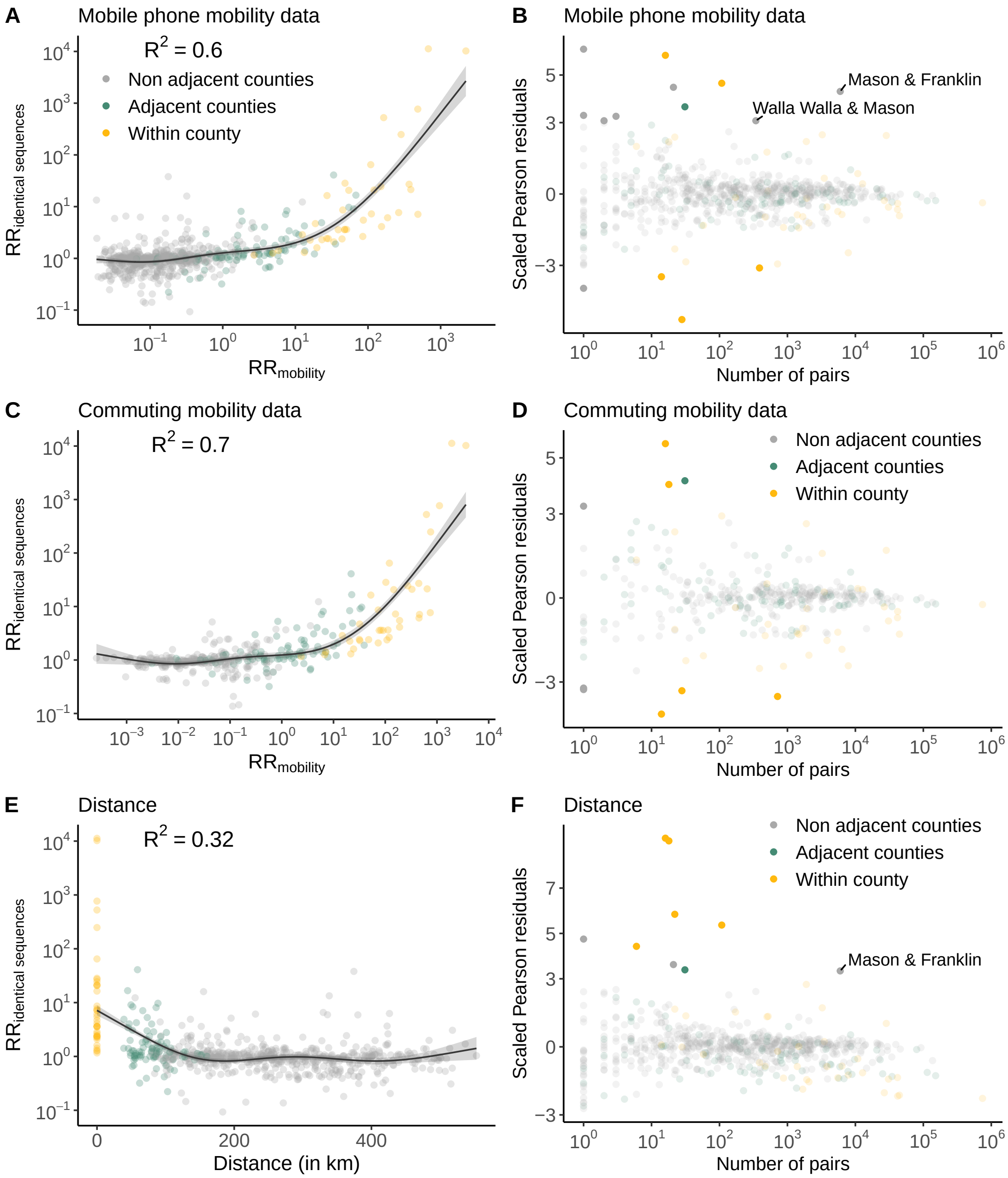


Figure S10. Comparison between the relative risk of observing identical sequences and the relative risk of movement at the county level. **A.** Relationship between the relative risk of observing identical sequences in two counties and the relative risk of movement between these counties as obtained from mobile phone mobility data. **B.** Scaled Pearson residuals of the GAM plotted in A as a function of the number of pairs of identical sequences observed in pairs of counties. **C.** Relationship between the relative risk of observing identical sequences in two counties and the relative risk of movement between these counties as obtained from workflow mobility data. **D.** Scaled Pearson residuals of the GAM plotted in C as a function of the number of pairs of identical sequences observed in pairs of counties. **E.** Relationship between the relative risk of observing identical sequences in two counties and the Euclidean distance between counties centroids. **F.** Scaled Pearson residuals of the GAM plotted in E as a function of the number of pairs of identical sequences observed in pairs of counties. In B, D and F, we label pairs of non-adjacent counties sharing at least 100 pairs of identical sequences and for which the absolute value of the Scaled Pearson residual is greater than 3. The trend lines correspond to predicted relative risk of observing identical sequences in two regions from each GAM. R^2 indicate the variance explained by each GAM.

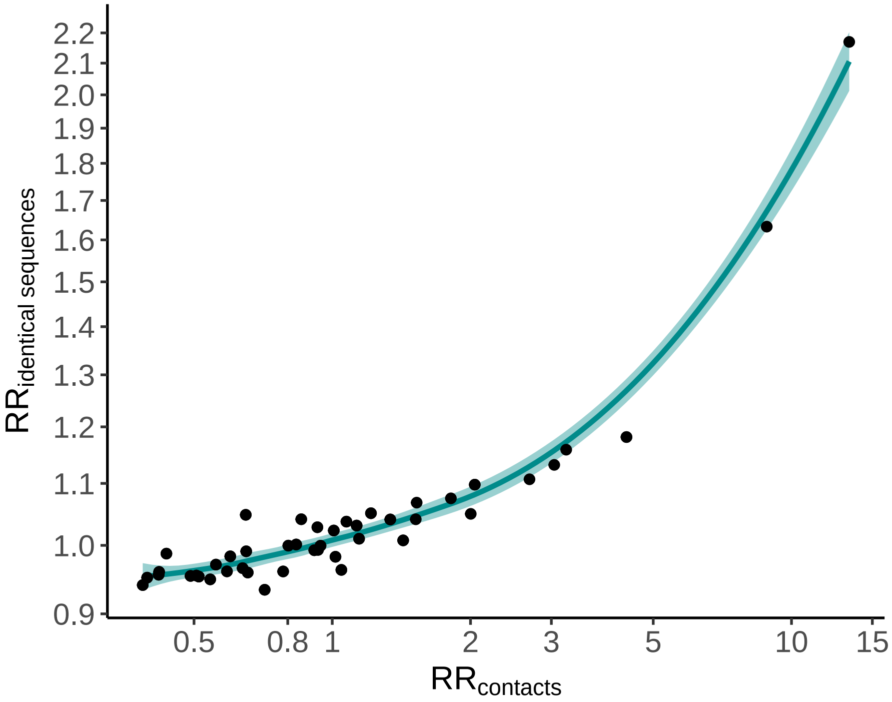


Figure S11. Expected relationship between the RR of observing identical sequences in two age groups and the RR of contacts between these age groups. These results were obtained by simulating 10^5 clusters of identical sequences from a branching process with mutations [8] using a Poisson offspring distribution. The simulation was parametrised by a reproduction number of 1.2, a probability that an infector and an infectee have the same consensus sequence of 0.7 and a sequencing fraction of 10%.

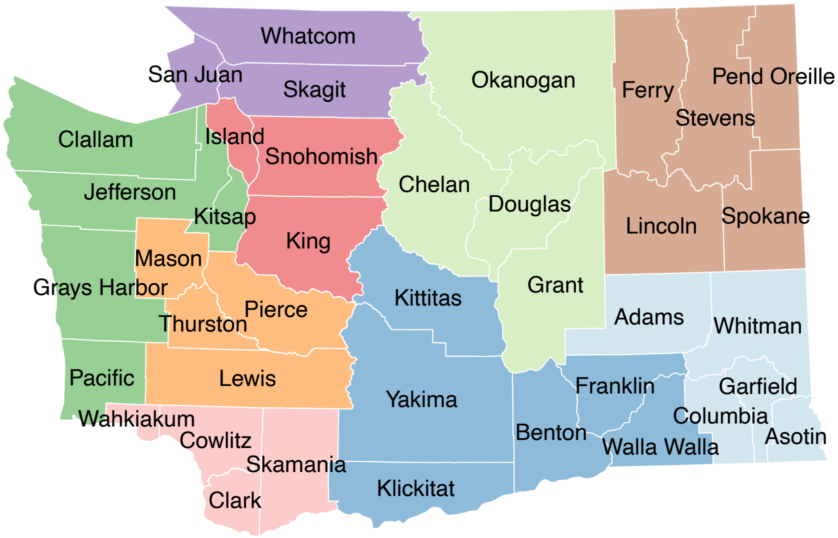


Figure S12. Geographical regions used to aggregate counties. Dark green: Peninsula/Coastal. Orange: South Puget Sound. Purple: Northwest. Red: North Puget Sound. Pink: Southwest. Dark blue: South Central. Light blue: Southeast.Brown: Northeast. Light green: North Central.

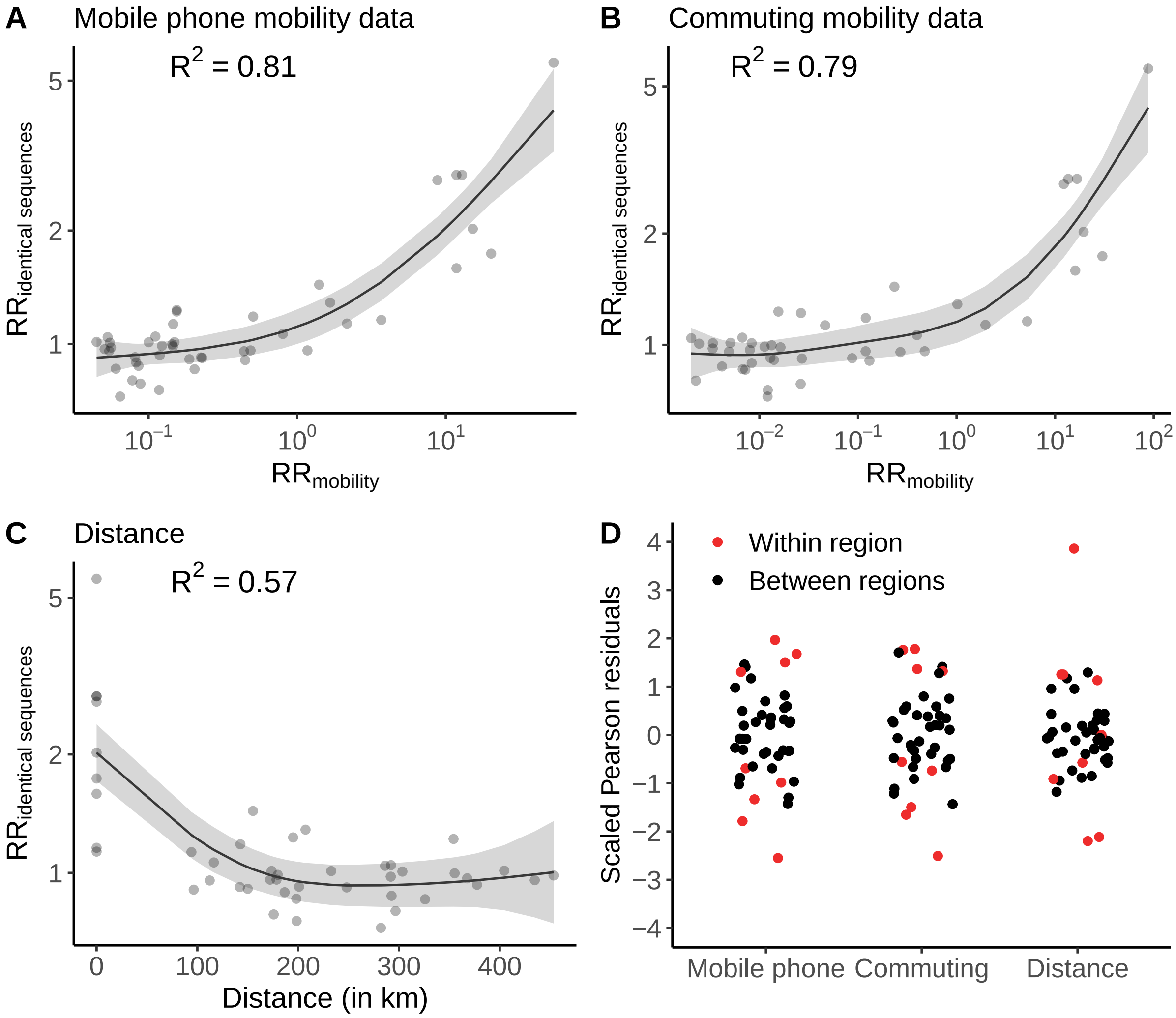


Figure S13. Comparison between the relative risk of observing identical sequences and the relative risk of movement at the region level. **A.** Relationship between the relative risk of observing identical sequences in two regions and the relative risk of movement between these regions as obtained from mobile phone mobility data. **B.** Relationship between the relative risk of observing identical sequences in two regions and the relative risk of movement between these regions as obtained from workflow mobility data. **C.** Relationship between the relative risk of observing identical sequences in two regions and the euclidean distance between region centroids. **D.** Scaled Pearson residuals of the GAM between the relative risk of observing identical sequences in two regions and (i) the relative risk of movement from commuting data, (ii) the relative risk of movement from mobile phone data and (iii) the geographic distance between regions' centroids. The trend lines correspond to predicted relative risk of observing identical sequences in two regions from each GAM. R^2 indicate the variance explained by each GAM.

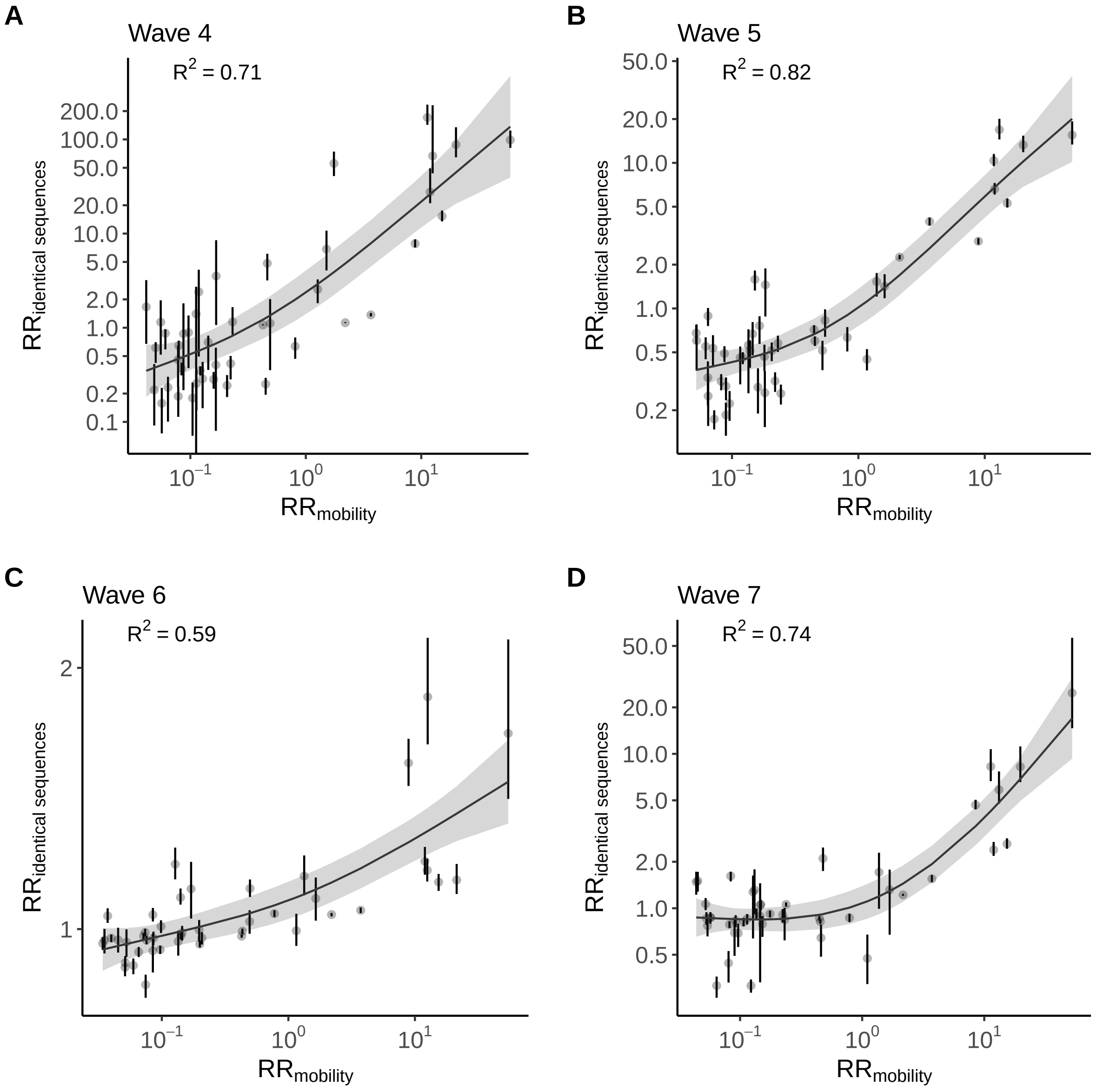


Figure S14. Relationship between the relative risk of observing identical sequences in two regions and the relative risk of movement between these regions obtained from mobile phone mobility data across epidemic waves **A.** Wave 4. **B.** Wave 5. **C.** Wave 6. **D.** Wave 7. Vertical segments indicate 95% subsampling confidence intervals. The trend line correspond to predicted relative risk of observing identical sequences in two regions from a GAM. R^2 indicate the variance explained by each GAM.

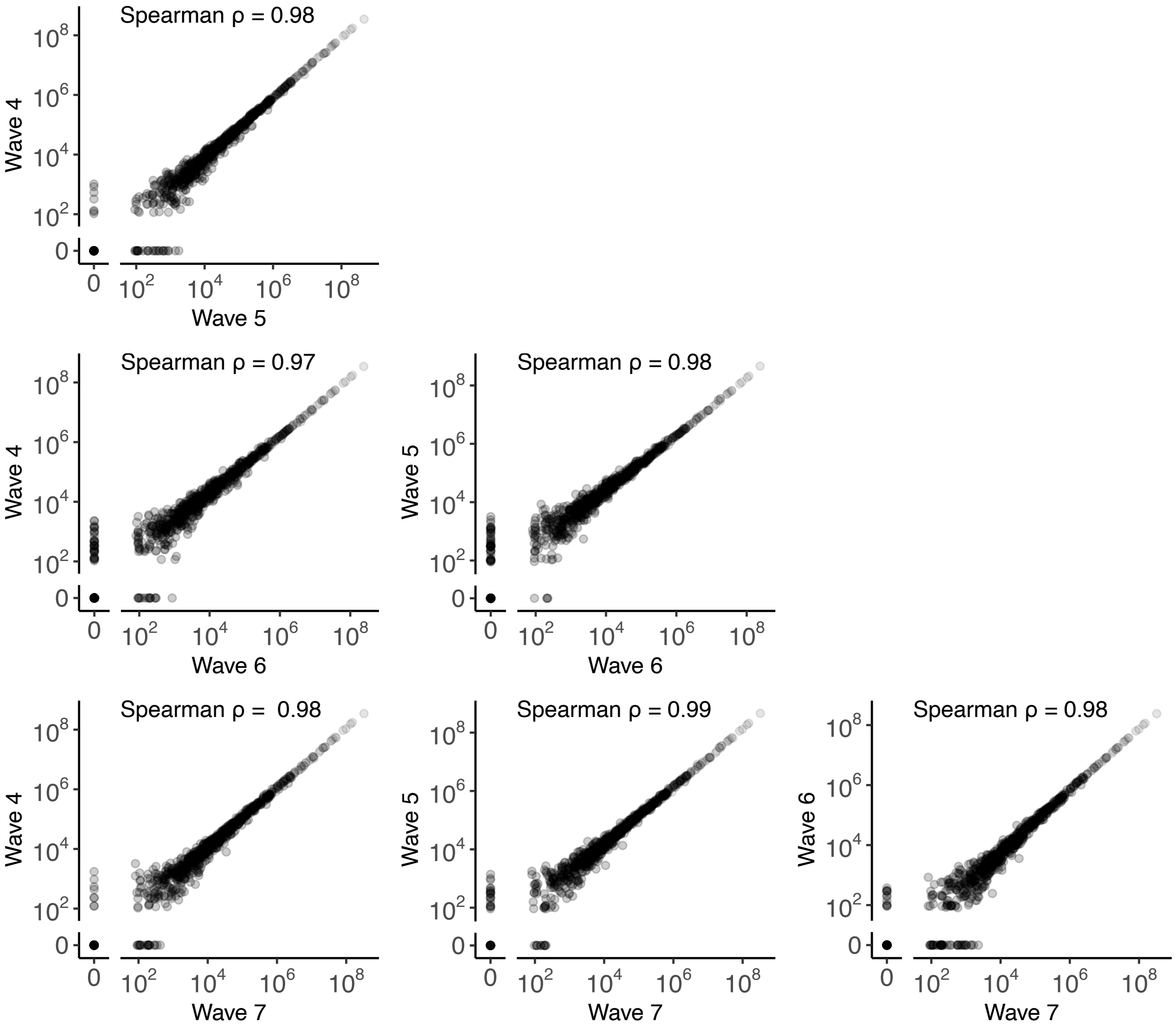


Figure S15. Comparison of the total number of visits between pairs of WA counties across the 4 epidemic waves during our study period. Points indicate the total number of visits between pairs of counties over the study periods labeled on the plot axes.

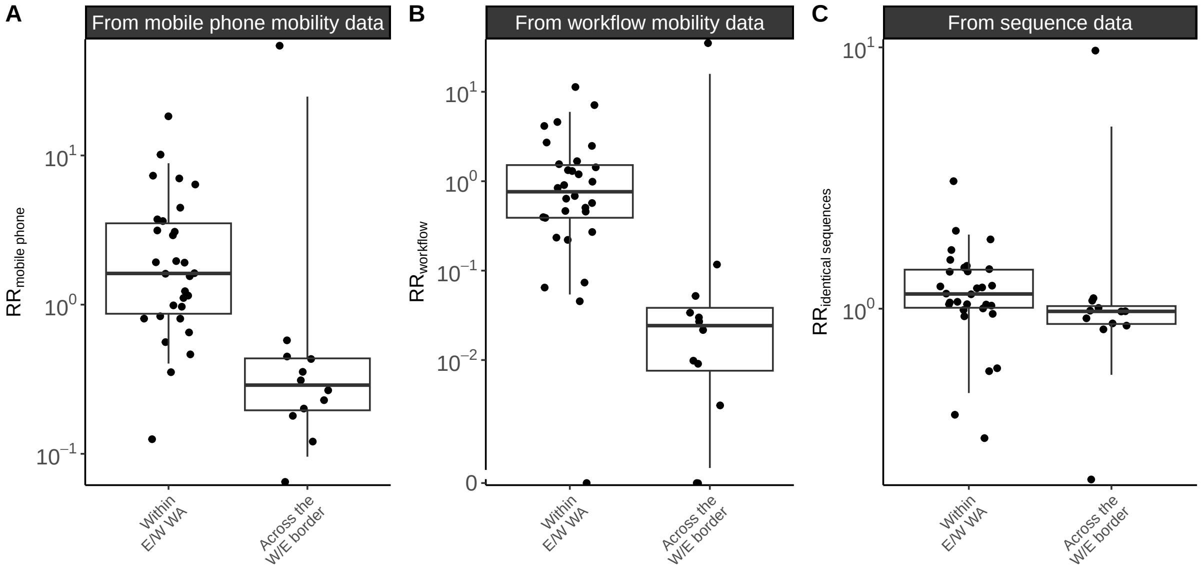


Figure S16. Comparison of connectivity metrics across the Eastern / Western WA border among counties located on the border. **A.** Relative risk of movement from mobile phone data across the border or within Eastern / Western WA (p-value for Wilcoxon rank sum test of $6.1 \cdot 10^{-5}$). **B.** Relative risk of movement from commuting data across the border or within Eastern / Western WA (p-value for Wilcoxon rank sum test of $1.6 \cdot 10^{-4}$). **C.** Relative risk of observing identical sequences across the border or within Eastern / Western WA (p-value for Wilcoxon rank sum test of $2.5 \cdot 10^{-2}$). In this analysis, we only consider WA counties along the W/E border.

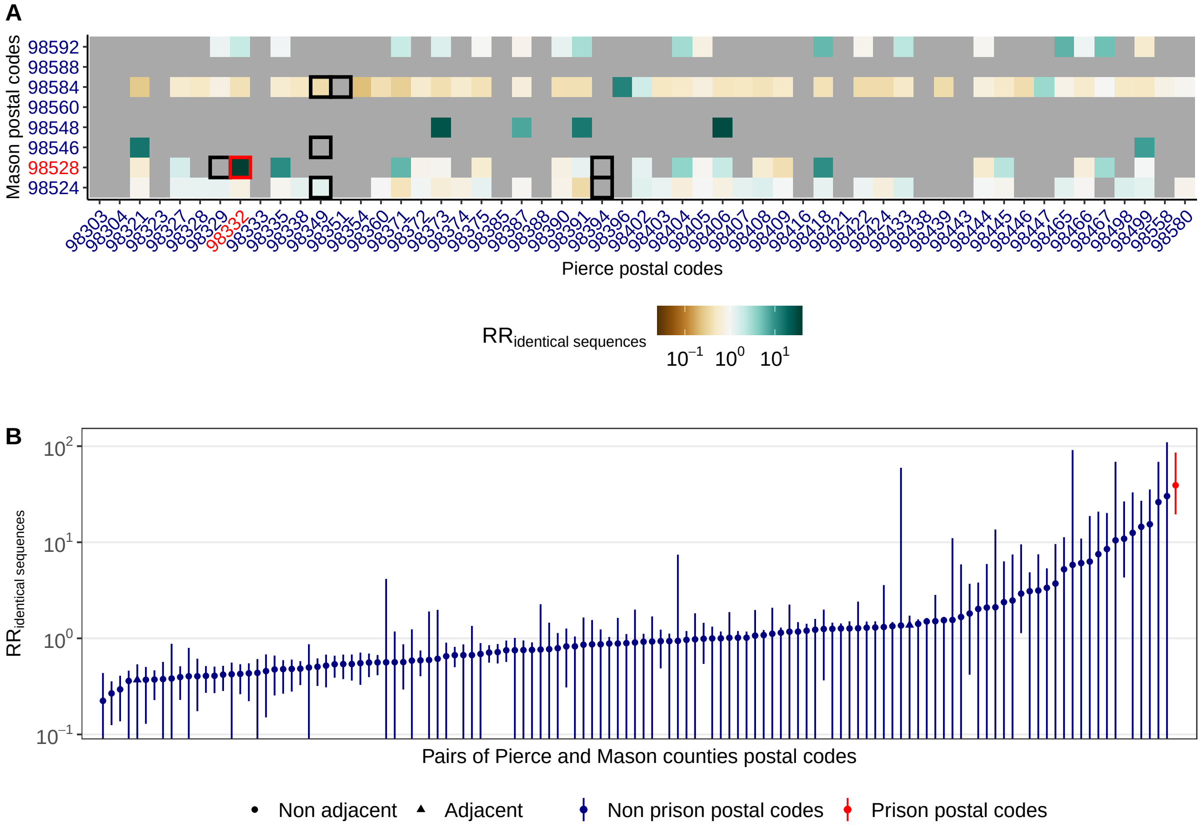


Figure S17. Patterns of occurrence of pairs of identical sequences between ZCTAs in Pierce and Mason counties, the two counties that are home of WA female prisons. A. Relative risk of observing identical sequences between ZCTAs in Mason and Pierce counties. Black squares indicate adjacent ZCTAs. ZCTAs in red correspond to postal codes that are the home of female prisons. **B.** Relative risk of observing identical sequences between Mason and Pierce counties ZCTAs ordered by increasing values. Vertical segments correspond to 95% subsampling confidence intervals.

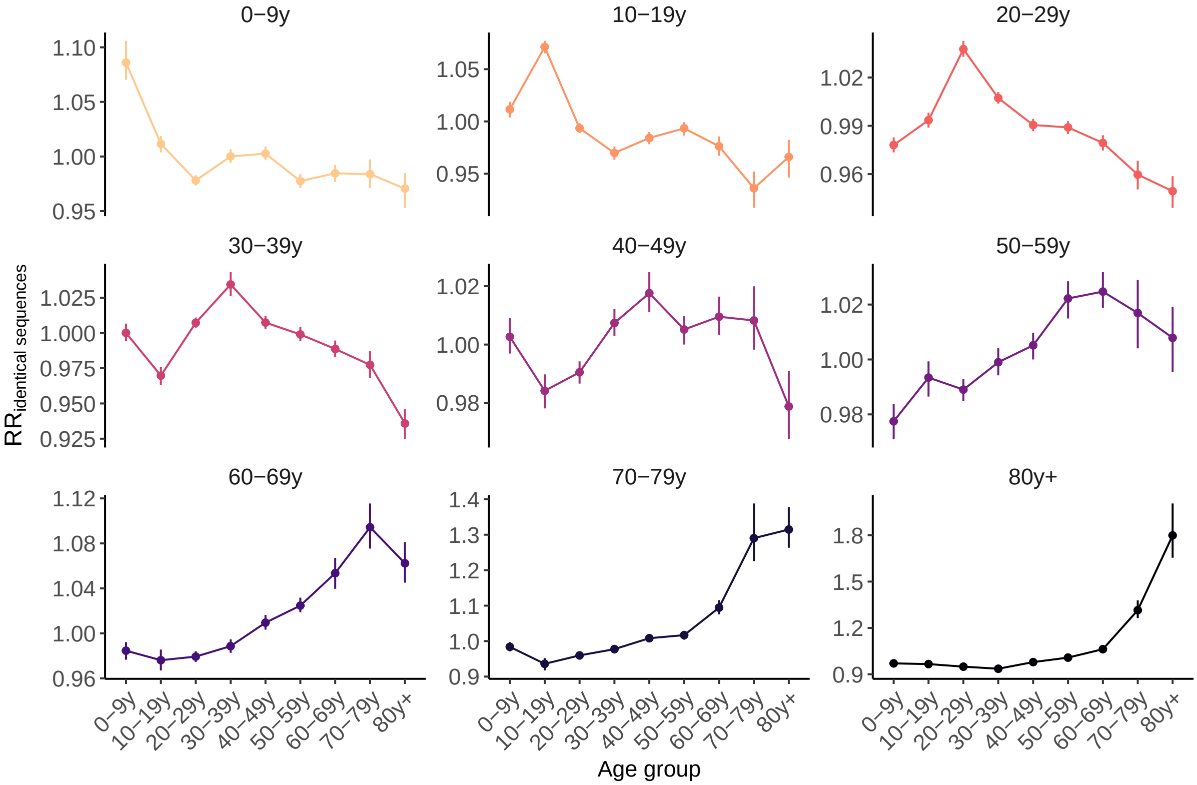


Figure S18. Relative risk for pairs of identical sequences of being observed between two age groups. Vertical segments correspond to 95% confidence intervals obtained through subsampling.

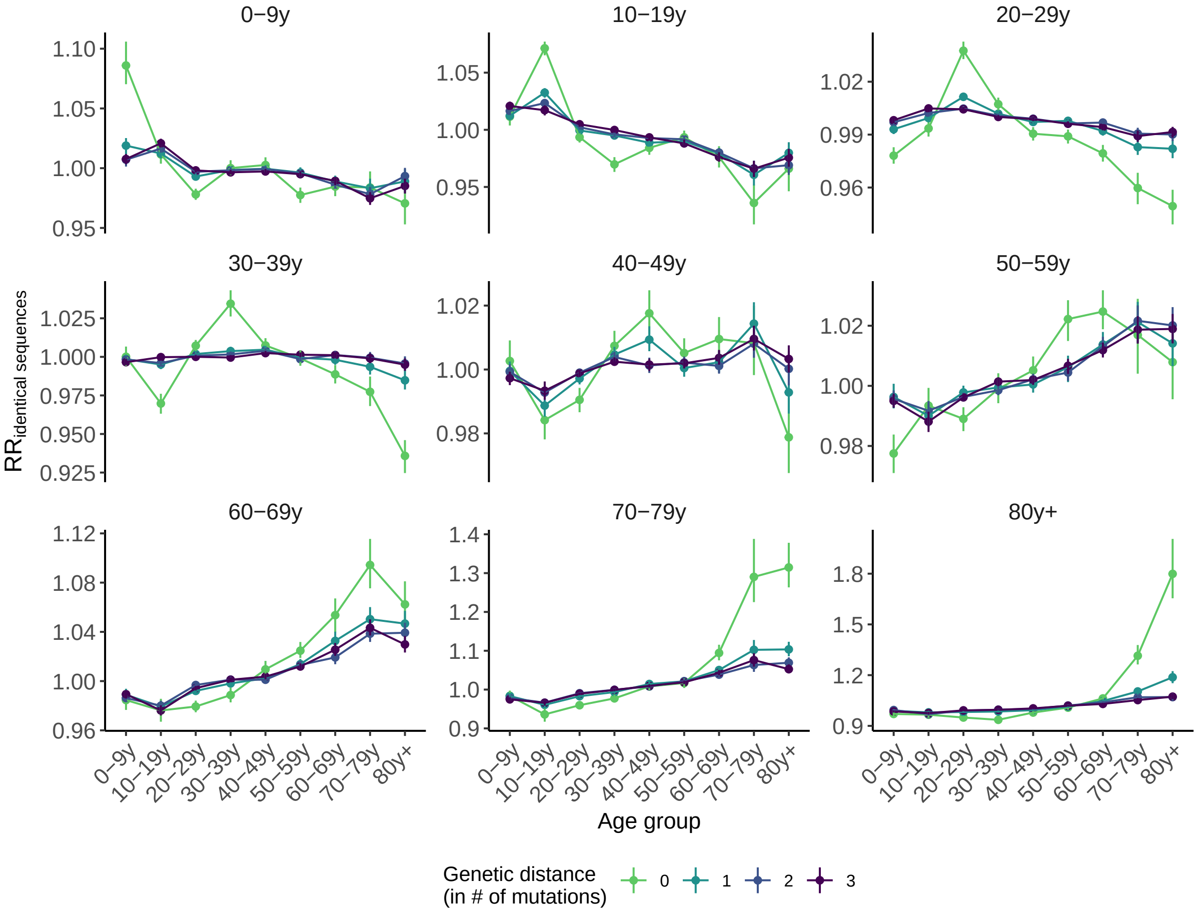


Figure S19. Relative risk for pairs sequences of being observed between two age groups depending on their genetic distance. Vertical segments correspond to 95% confidence intervals obtained through subsampling.

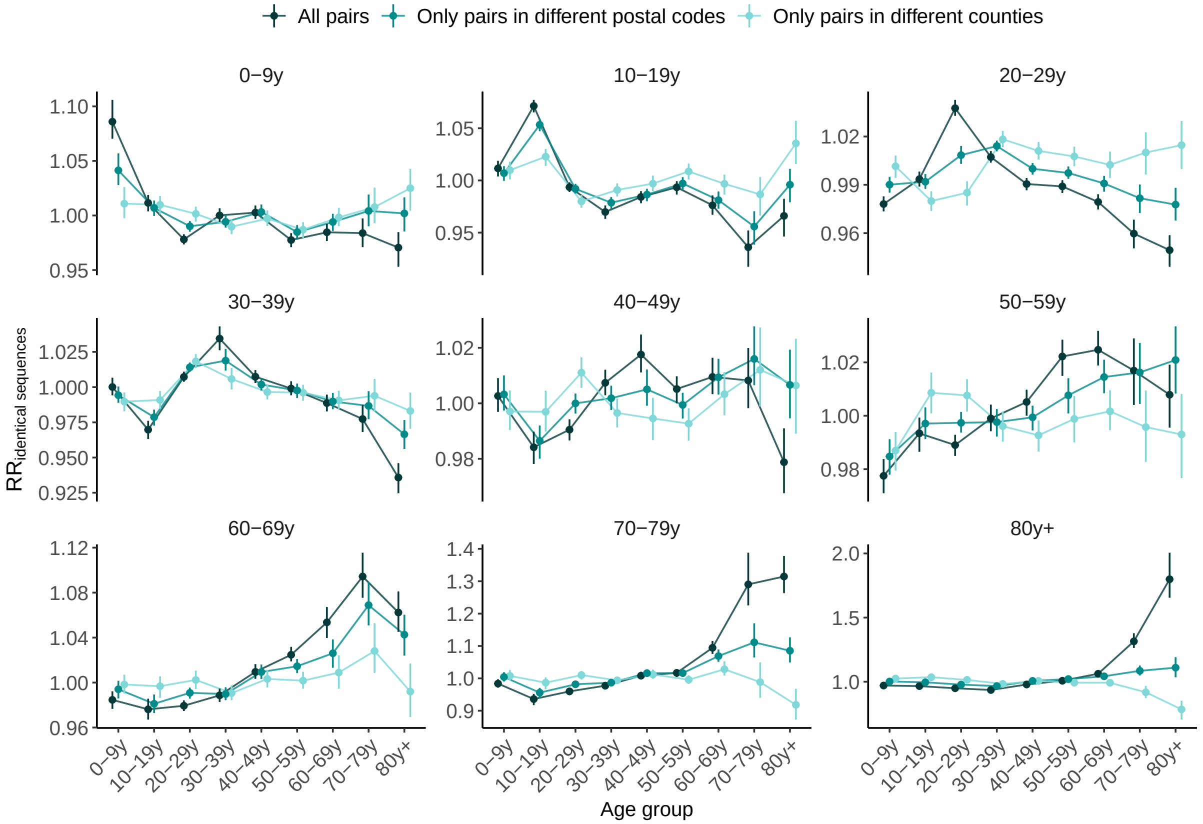


Figure S20. Impact of the spatial scale on the relative risk for pairs sequences of being observed between two age groups. Vertical segments correspond to 95% confidence intervals obtained through subsampling.

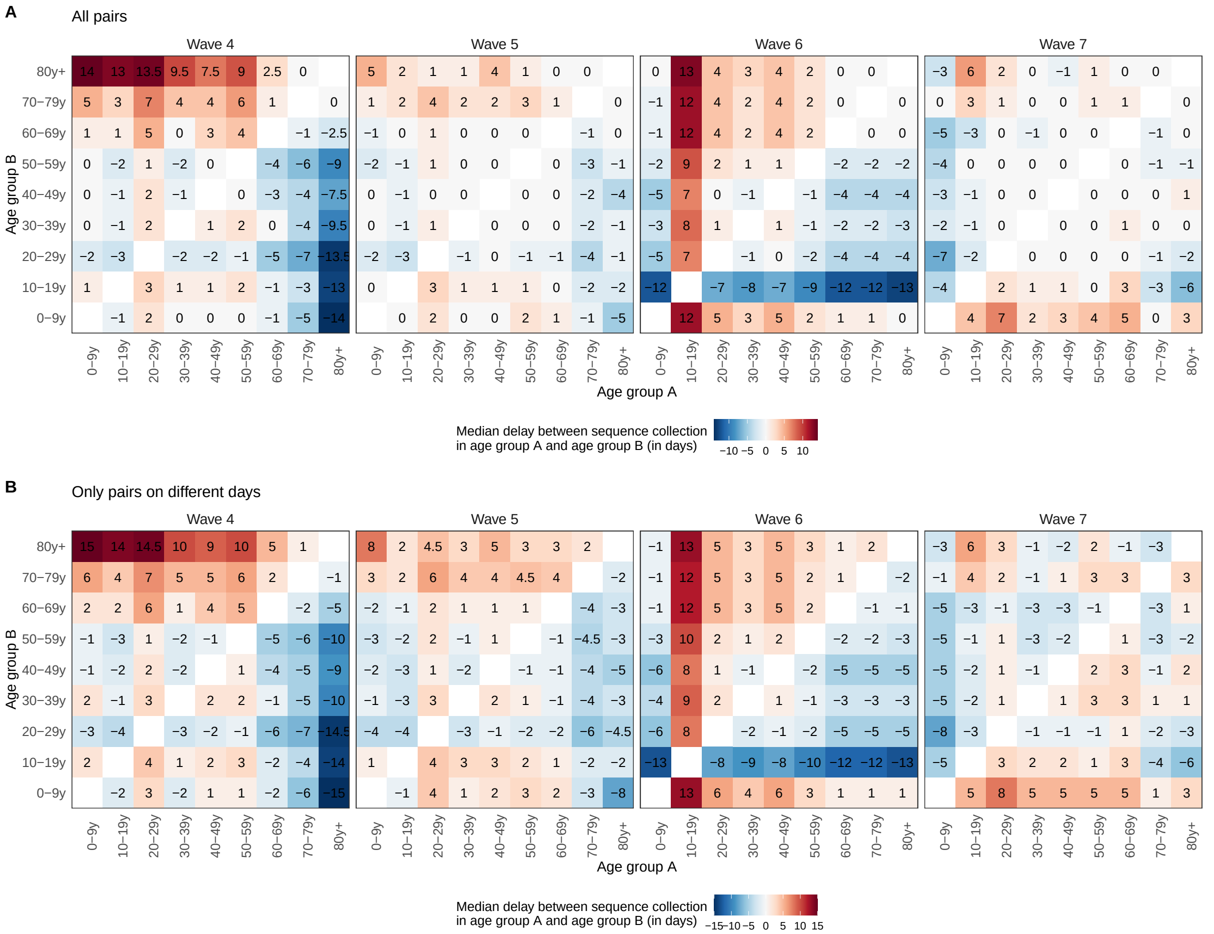


Figure S21. Median delay between the dates of sequence collection within pairs of identical sequences A. considering all pairs of identical sequences collected in two age groups and **B.** considering only pairs of identical sequences collected on different days in two age groups.

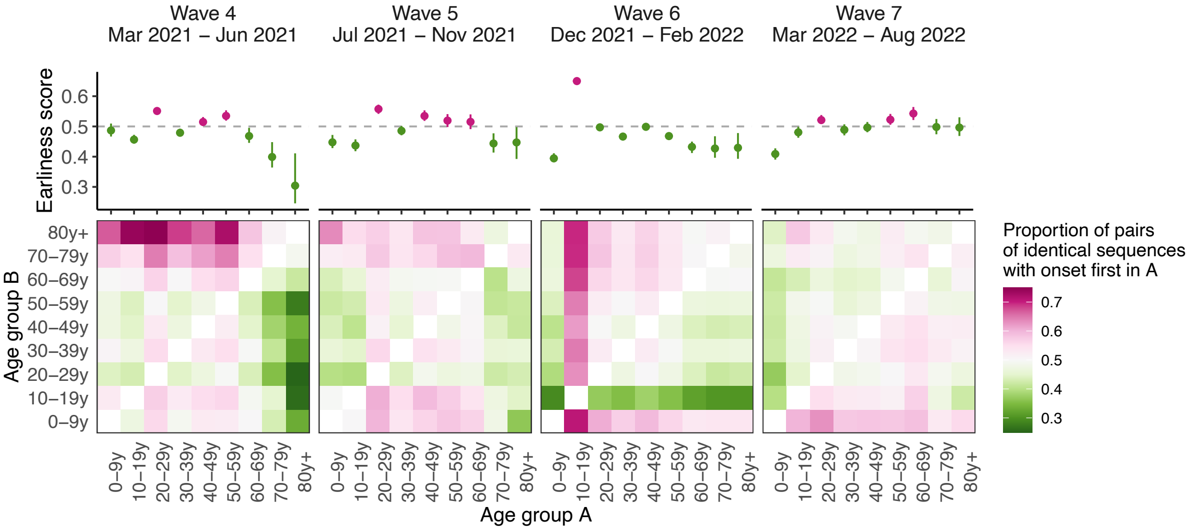


Figure S22. Sensitivity analysis on the timing of pairs identical sequences between age groups using symptom onset dates Median proportion of pairs of identical sequences with onset dates in age groups A before age group B across different epidemic waves from 1,000 imputed datasets (heatmaps). The dots plots depict the median earliness scores of age group A across 1,000 imputed datasets for the different epidemic waves. Vertical segments indicate uncertainty range around earliness score (see Methods).

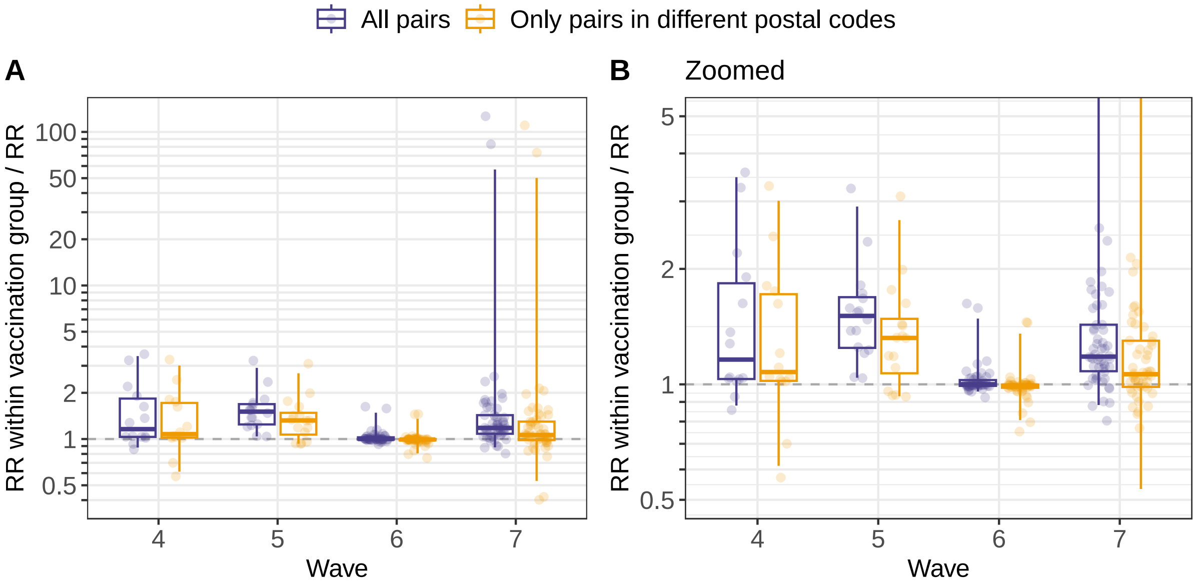


Figure S23. Ratio between the relative risk of observing identical sequences within a given vaccination group (denoted V_1) and between two vaccination groups (denoted V_1 and V_2). Values above 1 indicate that pairs of identical sequences tend to be enriched in pairs observed within the same vaccination group. The analysis is restricted to pairs observed within the same age group. Each point correspond to the ratio computed for a given pair of vaccination status (V_1, V_2) and age group. Boxplots indicate the 2.5%, 25%, 50%, 75% and 97.5% percentiles.

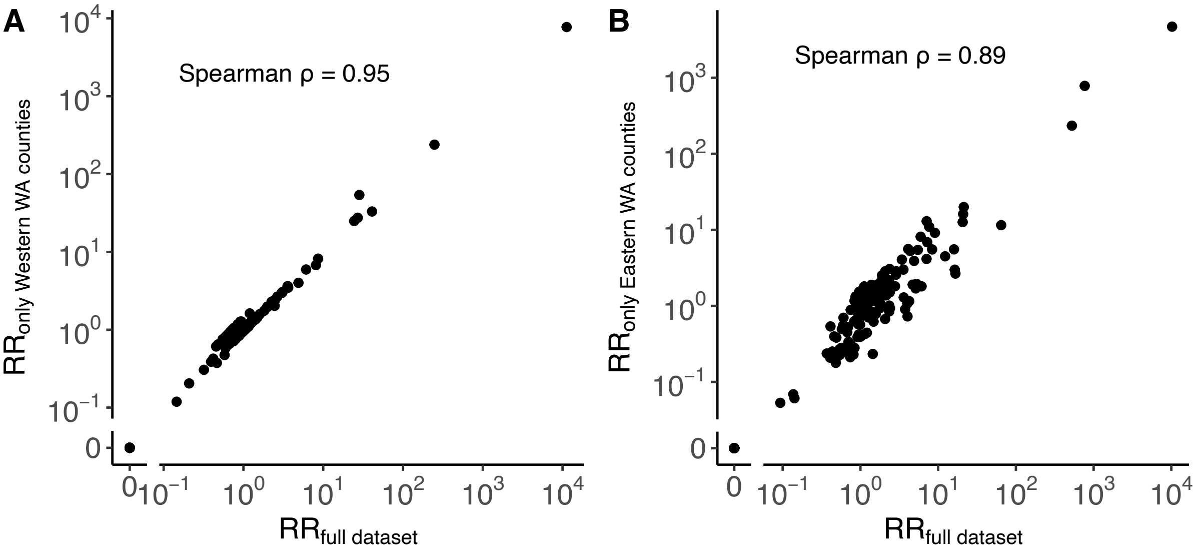


Figure S24. Impact of non-sampled locations on the computation of the RR. A. Comparison between the relative risk of observing identical sequences between Western WA counties using only sequence in Western WA counties or the entire sequence dataset. **B.** Comparison between the relative risk of observing identical sequences between Eastern WA counties using only sequence in Eastern WA counties or the entire sequence dataset.

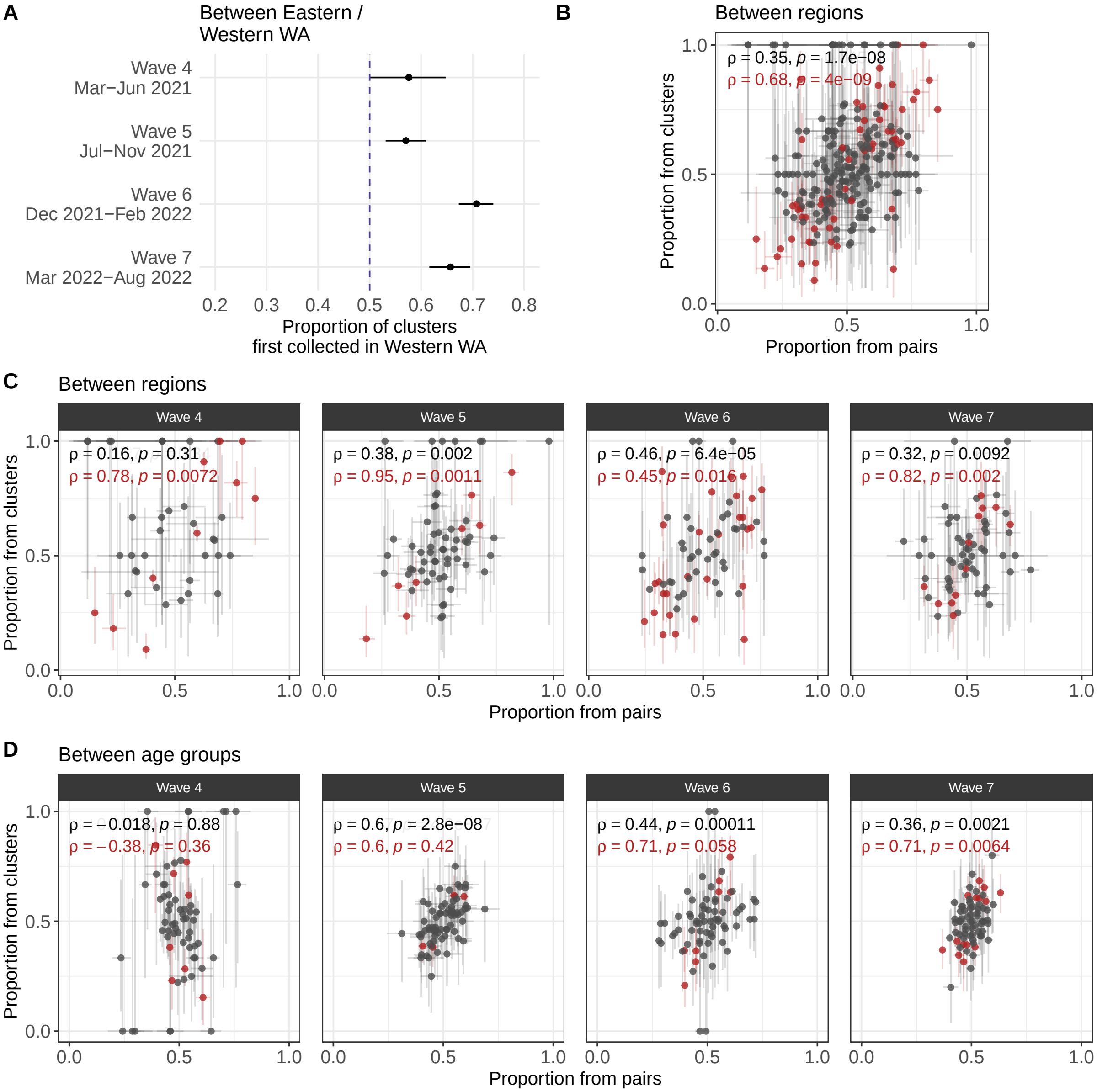


Figure S25. Sensitivity analysis for our transmission direction analysis relying on clusters of identical sequences observed only in two groups. **A.** Proportion of clusters first collected in Western WA among clusters observed in Eastern and Western WA and across pandemic waves. Like in Figure 2G, these proportions are all greater than 0.5. **B.** Sensitivity analysis at the regional level (Figure S12) comparing the proportion from pairs and the proportion from clusters. **C.** Sensitivity analysis at the regional level comparing the proportion from pairs and the proportion from clusters across waves. **D.** Sensitivity analysis at the age level comparing the proportion from pairs and the proportion from clusters across waves. For wave 4, the cluster based analysis relies on less than 10 clusters in 13 out of 36 pairs of age groups, which could explain the poor correlation. Segments indicate 95% CIs around proportions. In B, C and D, the colour red depicts points for which the CIs don't cross 0.5 for both the proportion from clusters and the proportion from clusters. We report in red the Spearman correlation coefficient with p-values for these red points and in black for all the points.

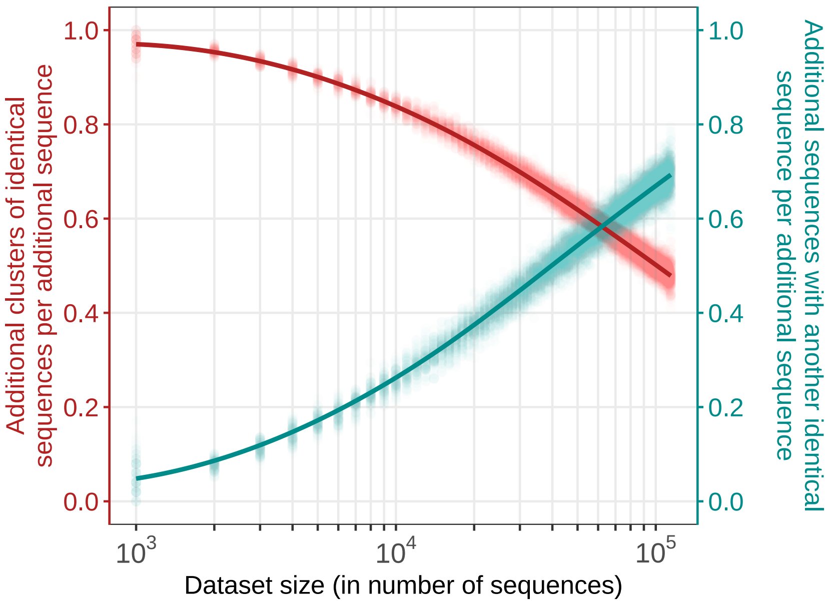


Figure S26. Impact of dataset size on the number of clusters of identical sequences and the number of sequences with another identical sequence in the dataset. We generate this figure by considering sequence datasets of increasing sizes, ranging between 10^2 and the 114,298 (the size of our WA dataset) with an increment of 10^2 between 10^2 and 10^3 and an increment of 10^3 above 10^3. We run 100 simulations where we first downsample 10^2 sequences from our full dataset and then incrementally include more sequences (drawn from the total remaining sequences not yet included). At every step, we compute the additional number of clusters of identical sequences per additional sequences (red) as well as the additional number of sequences with another identical sequence in the dataset per additional sequence (cyan). Points indicate the results from individual simulations and lines the LOESS curves.

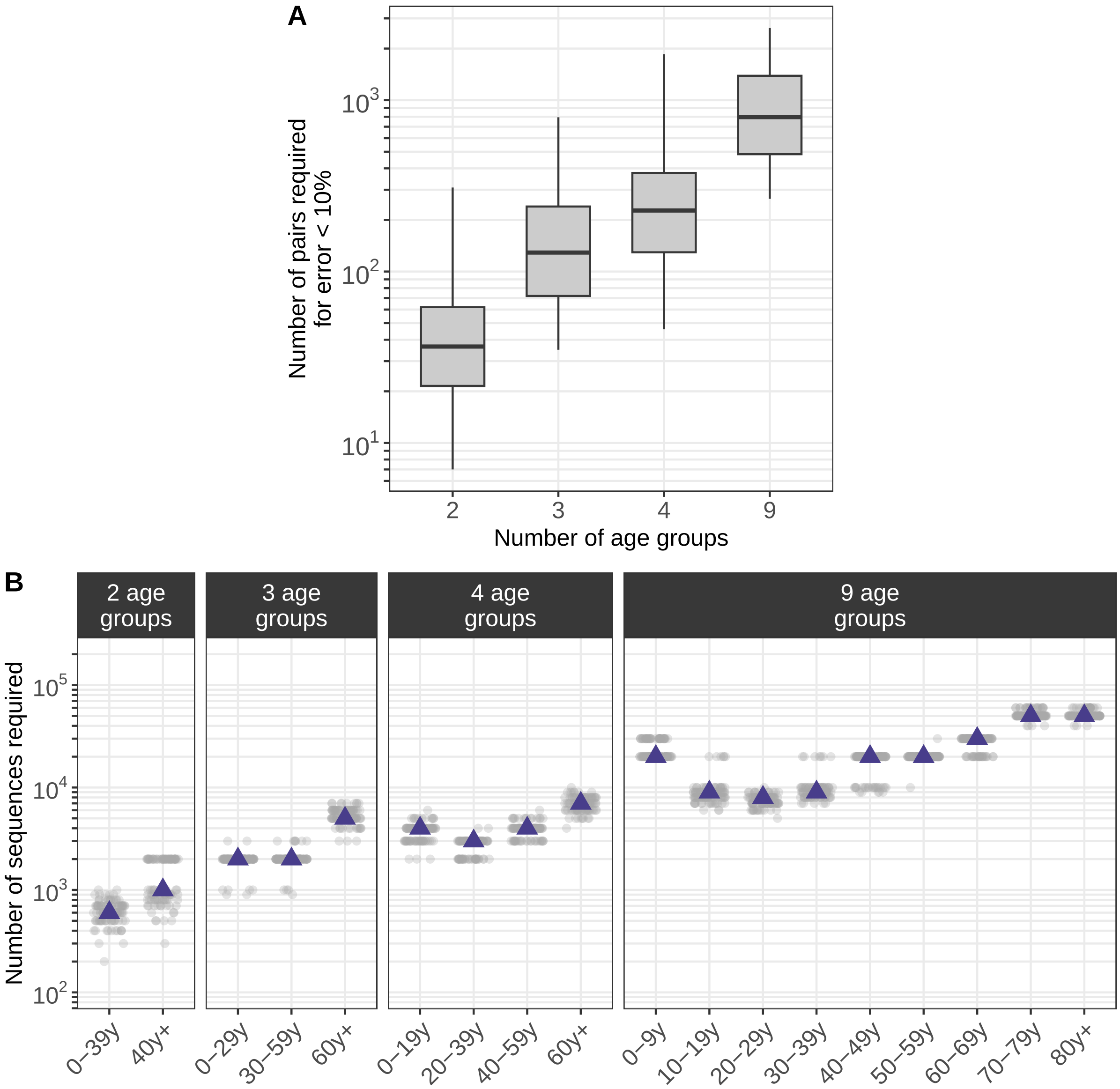


Figure S27. Impact of the number of groups included in the analysis on the dataset size required for the error in the relative risk of observing identical sequences to be lower than 10%. **A.** Number of pairs of identical sequences required for the error relative risk of observing identical sequences to be lower than 10%. Boxplots indicate the 2.5%, 25%, 50%, 75% and 97.5% percentiles. See Methods for a description of the downsampling strategy. **B.** Number of sequences required for the number of pairs of identical sequences observed within the age group on the x-axis to reach the median depicted in A. Each point corresponds to a subsampled dataset. Purple triangles indicate the median.

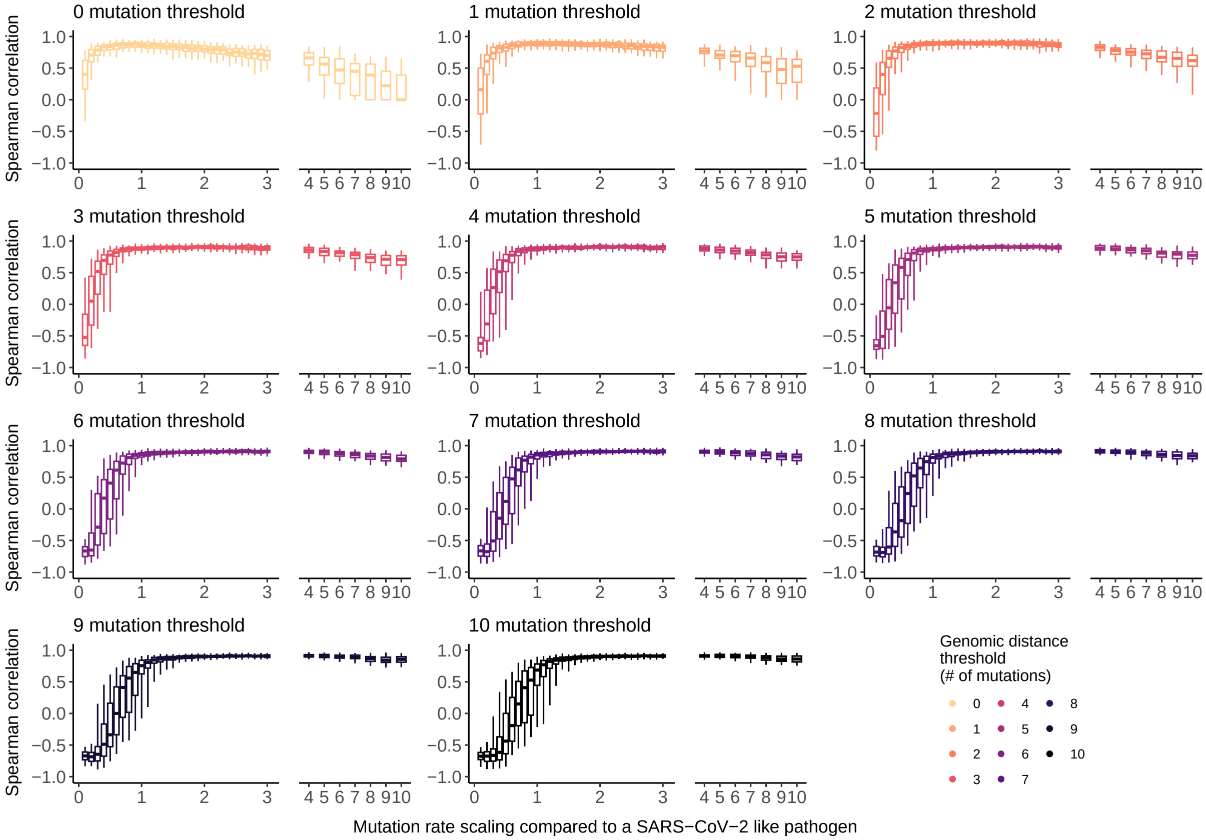


Figure S28. Impact of the pathogen's mutation rate on the optimal Hamming distance threshold to apply our RR framework. Boxplots indicate Spearman correlation coefficient between the relative risk of pairs of sequences below the genomic distance threshold of being observed in two regions and the daily migration probability between these two regions. Boxplots indicate the 2.5%, 25%, 50%, 75% and 97.5% percentiles. See Methods for a description of the simulation approach.

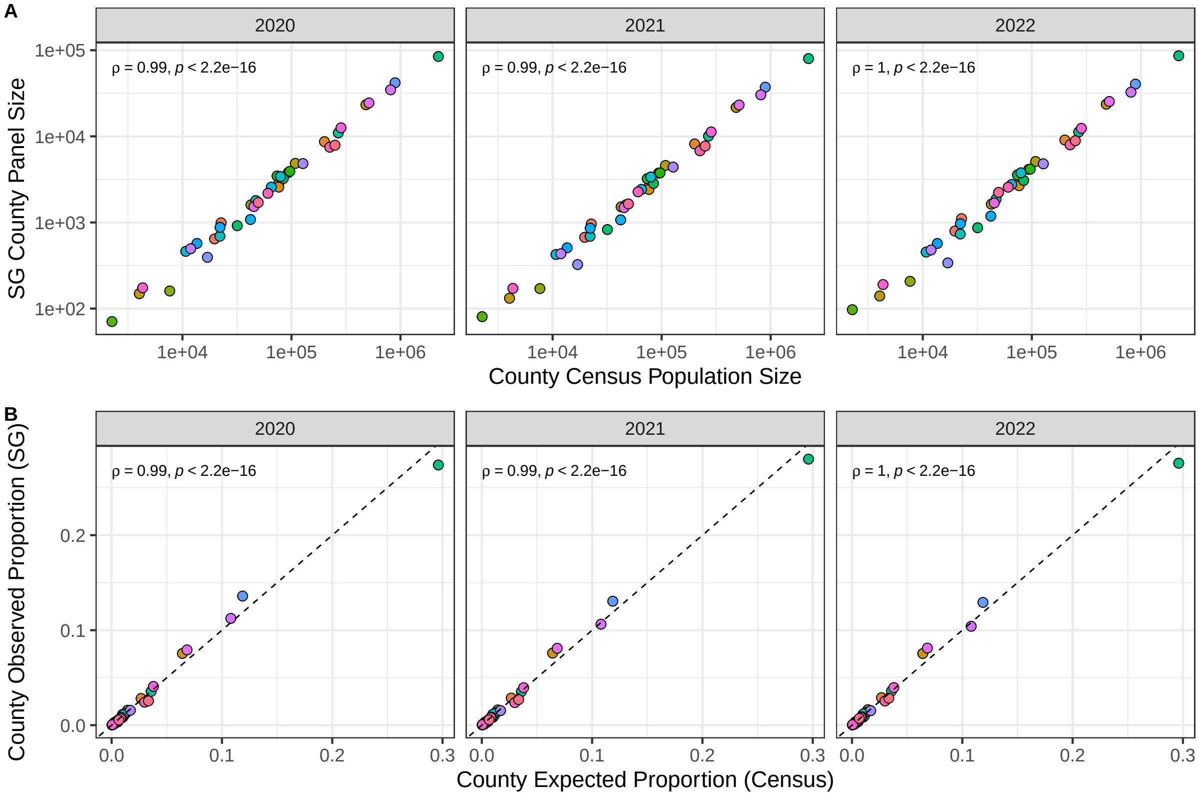


Figure S29. Comparisons between county census population sizes and SafeGraph panel sizes in Washington state, 2020 − 2022. A. County census population sizes strongly correlate with the mean number of devices tracked by SafeGraph ("SG") in each year. **B.** Expected proportions of devices based on county and state census population sizes strongly correlate with the observed proportion of devices tracked by SafeGraph ("SG") in each year. Points represent individual counties in WA state. In B, the black dashed line indicates the expected relationship for a true random sample of devices.

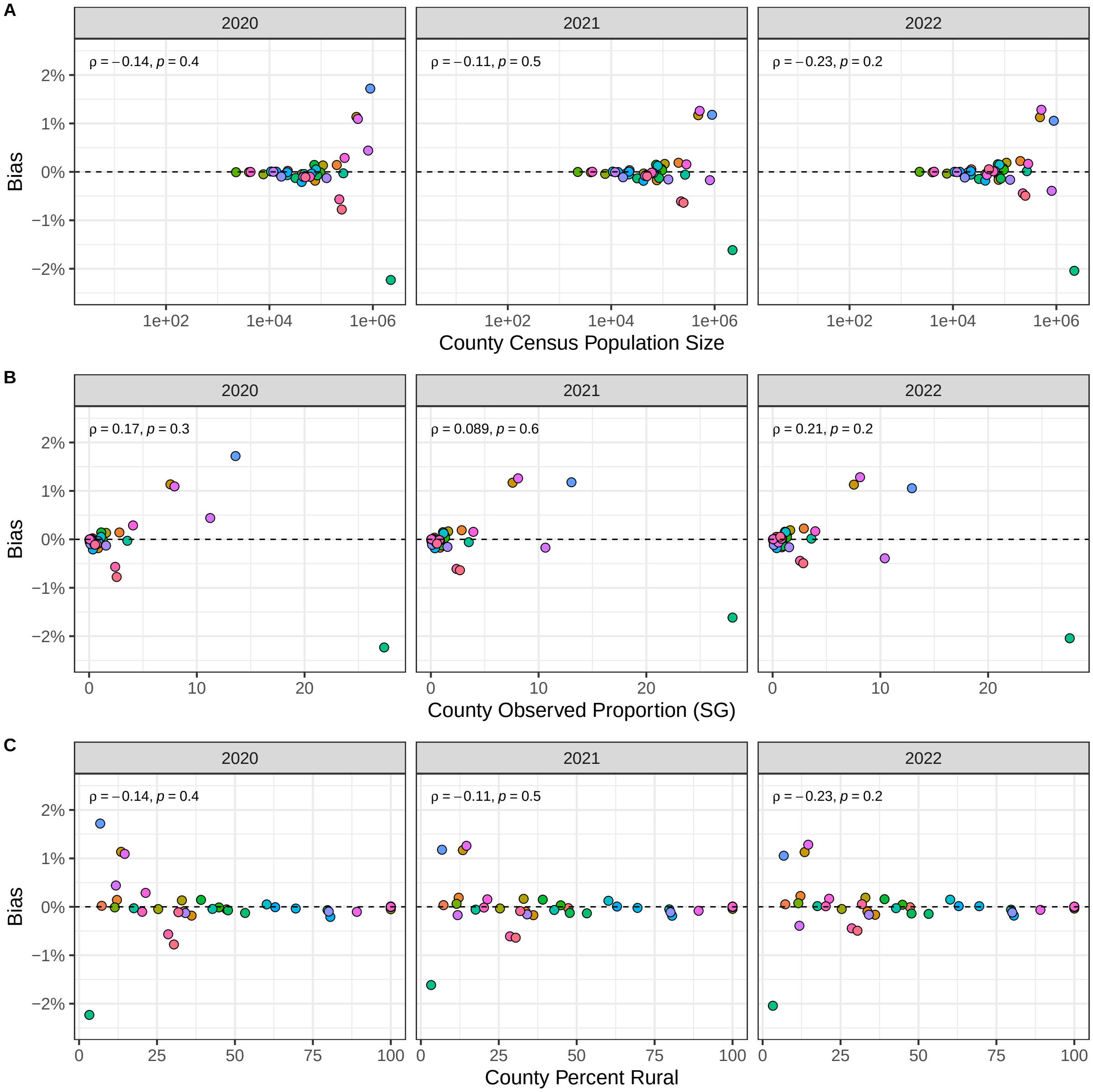


Figure S30. County-level bias of SafeGraph data in Washington state does not correlate with A. census population size, B. SafeGraph panel sizes in individual counties relative to WA state ("county observed proportion"), or C. census urban-rural classification, 2020 – 2022. Points represent individual counties in WA state. Bias is estimated as the "observed proportion" of devices tracked by SafeGraph in individual counties relative to WA state minus the "expected proportion" of devices based on census population sizes. Negative values indicate under-represented counties, and positive values indicate over-represented counties. The black dashed line (y = 0) indicates no bias in the SafeGraph panel of devices.

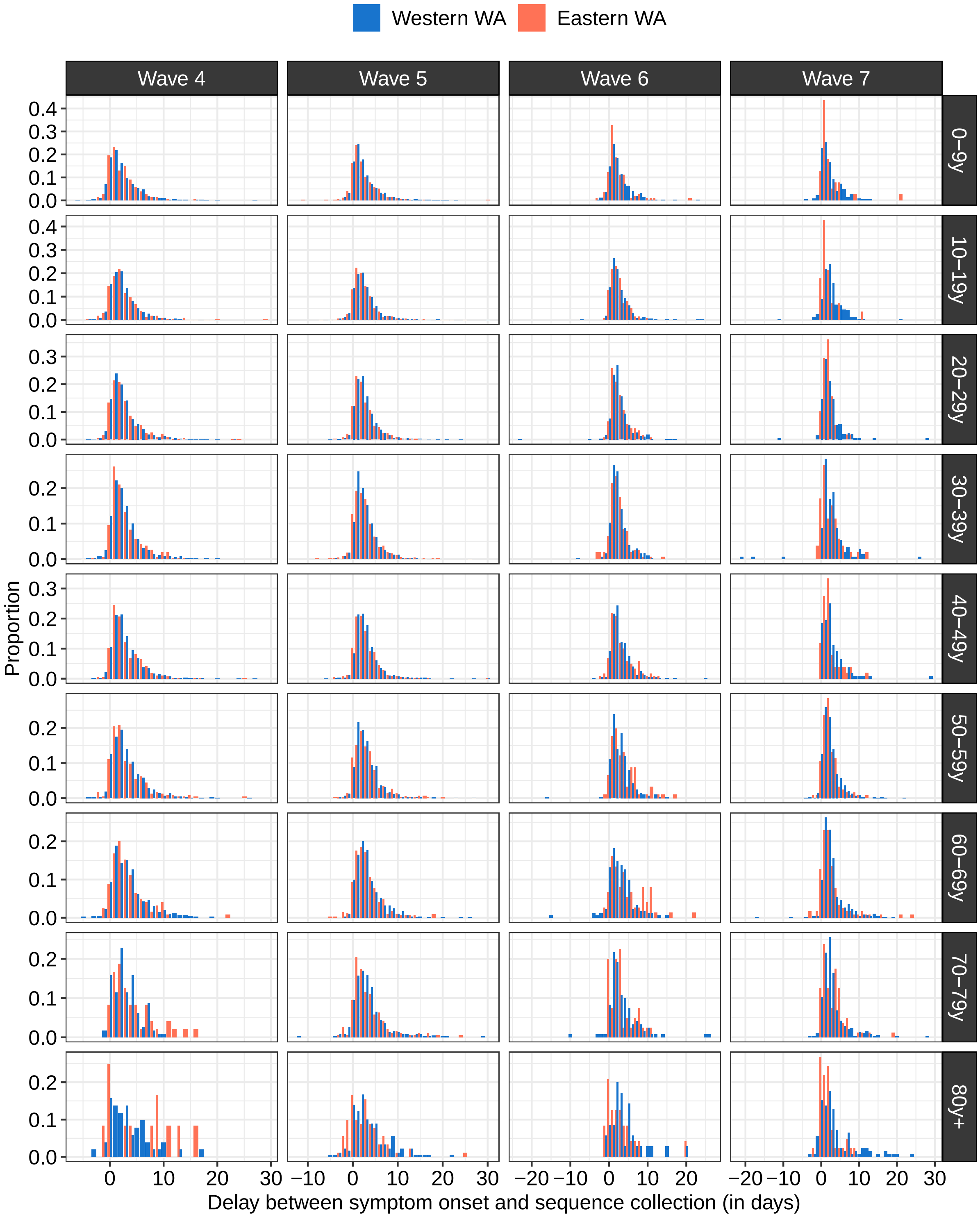


Figure S31. Empirical distribution of the delay between symptom onset and sequence collection by age (rows), period (columns) and geographic region (colours).

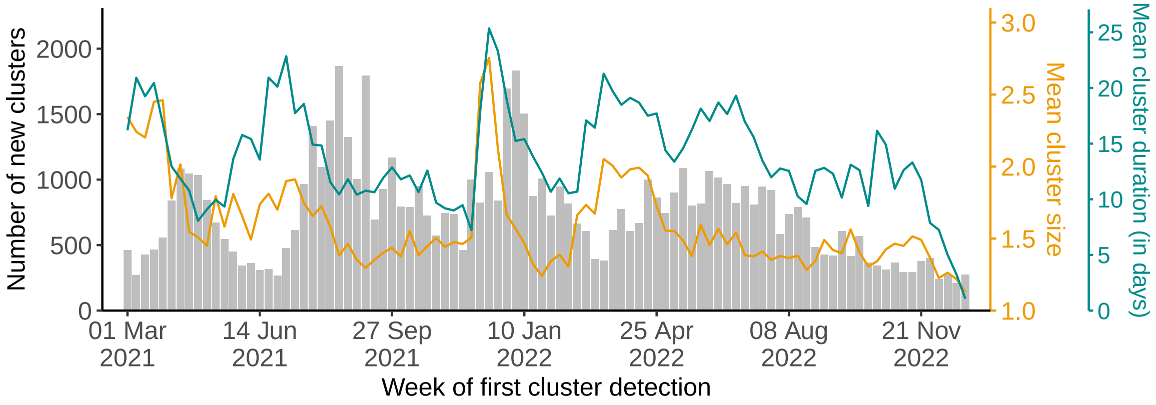


Figure S32. Characteristics of clusters of identical sequences across the study period. Grey bars (left y-axis) indicate the number of clusters by week of first cluster detection (defined as the week where the sequence with the earliest collection date was collected). The orange line (right orange y-axis) depicts the mean size of clusters of identical sequences by week of first cluster detection. The cyan line (right cyan y-axis) depicts the mean cluster duration by week of first cluster detection.

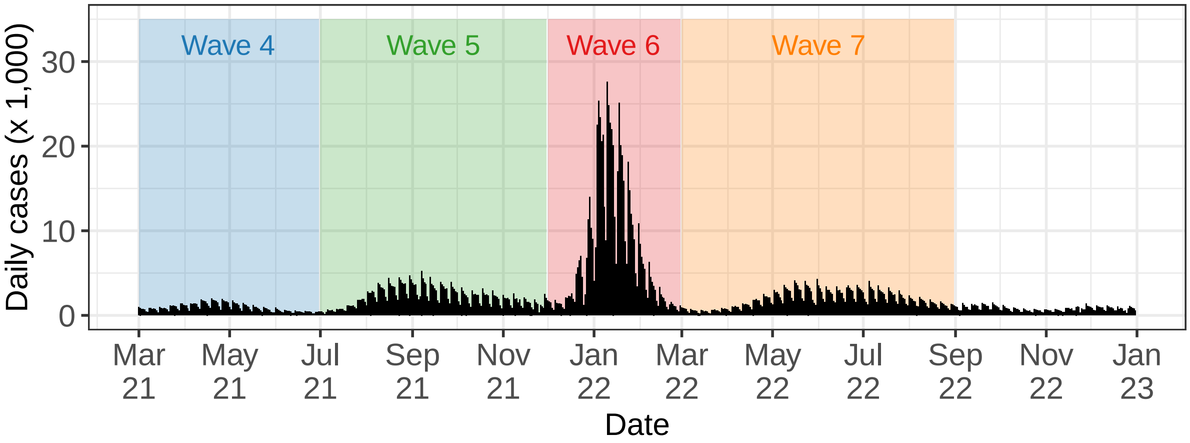


Figure S33. Time-series of COVID-19 cases in WA over the study periods. Shaded rectangles indicate the periods used to define the successive epidemic waves.

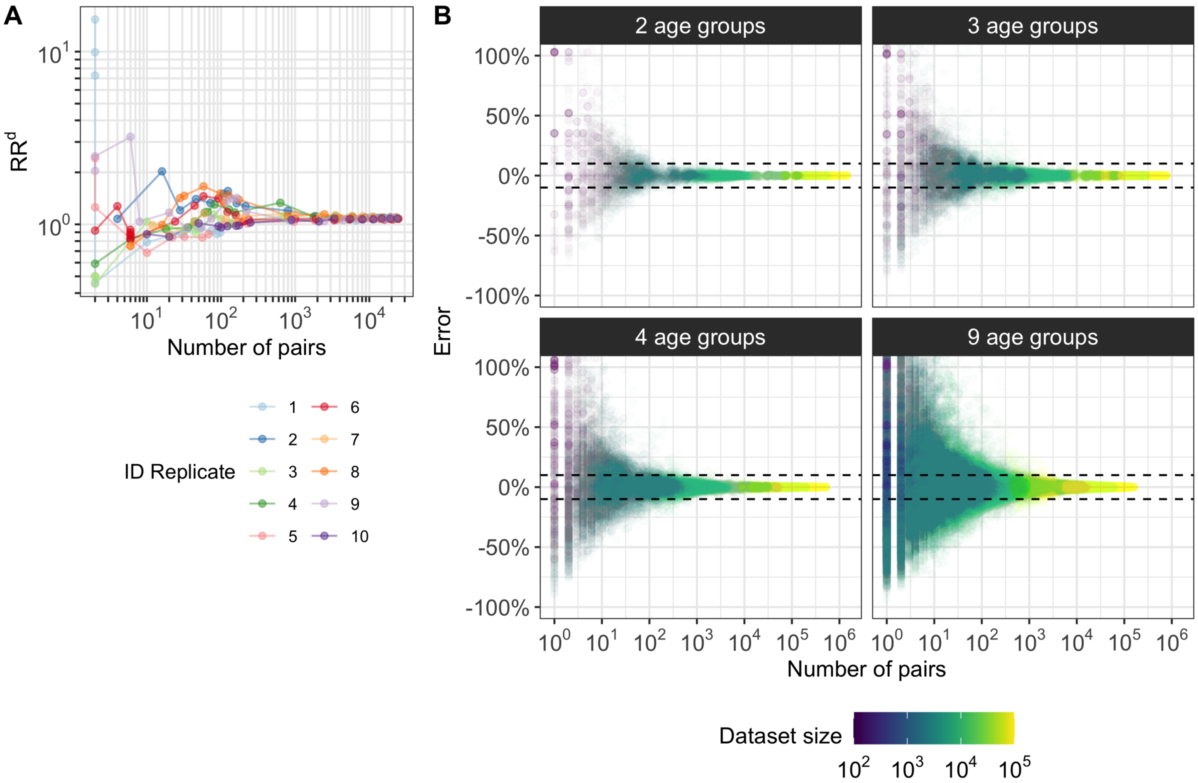


Figure S34. Illustration of the downsampling strategy used to quantify the amount of data required to compute relative risks. **A.** Relative risk RR^d of identical sequences being shared between the 0-9y computed on 10 downsampled datasets as a function of the number of pairs of identical sequences shared between the 0-9y. **B.** Error ϵ on the relative risk of observing identical sequences in downsampled datasets as a function of the number of pairs of identical sequences present for a pair of age groups in the downsampled datasets.

Method	Average sequencing probability	Replicate	Sample size (unbiased)	Sample size (biased)	Corr. with sim (unbiased)	Corr. with sim (biased)	Corr. biased / unbiased
DTA Fix tree	0.43 %	1	1745	1744	0.54	-0.22	-0.26
DTA Fix tree	0.43 %	2	1714	1805	0.60	0.19	0.35
DTA Fix tree	2.16 %	1	8723	8709	0.61	0.39	0.87
DTA Fix tree	2.16 %	2	8551	8675	0.77	0.25	0.70
DTA Inferring tree	0.43 %	1	1745	1744	0.10	0.15	-0.13
RR	8.66 %	1	34338	35304	0.91	0.74	0.77
RR	8.66 %	2	34736	35123	0.93	0.84	0.79
RR	2.16 %	1	8723	8709	0.94	0.80	0.79
RR	2.16 %	2	8551	8675	0.80	0.75	0.86

Table S1. Performance of Discrete Trait Analysis (DTA) and our relative risk metric (RR) in quantifying migration patterns. The sample sizes correspond to the number of sequences on which the inference is performed. All correlation coefficients reported are Spearman rank correlation coefficients. In the DTA analysis, we report the correlation between estimated and true migration rates (both for the biased and unbiased sequencing scenarios) and the correlation between the migration rates estimated on the biased and unbiased datasets. In the RR analysis, we report the correlation between the RR and the migration probability between demes (both for the biased and unbiased sequencing scenarios) as well as the correlation between the RR estimated on the biased and unbiased datasets.

Region	Adjacency status compared	p-value (Wilcoxon test)	p-value (Wilcoxon test) without 0
East-East	Within ZCTA & Adjacent ZCTAs	$5 \cdot 10^{-10}$	$3 \cdot 10^{-12}$
	Adjacent ZCTAs & Non-adjacent ZCTAs	$< 10^{-16}$	$< 10^{-16}$
East-West	Adjacent ZCTAs & Non-adjacent ZCTAs	0.73	0.39
West-West	Within ZCTA & Adjacent ZCTAs	$< 10^{-16}$	$< 10^{-16}$
	Adjacent ZCTAs & Non-adjacent ZCTAs	$< 10^{-16}$	$< 10^{-16}$

Table S2. Comparison of the relative of risk of observing identical sequences at the ZCTA level by adjacency level. We report the p-values of Wilcoxon rank sum test using either all pairs of ZCTAs or only pairs of ZCTAs for which pairs of identical sequences are collected (column ”without 0”).

	At the county level	At the region level
Spearman correlation ρ		
Mobile phone mobility	35%	61%
Workflow mobility data	40%	59%
Geographic distance	-35%	-48%
Spearman correlation ρ (without 0)		
Mobile phone mobility	43%	61%
Workflow mobility data	56%	59%
Geographic distance	-36%	-48%
Variance explained (GAM)		
Mobile phone mobility	60%	81%
Workflow mobility data	70%	79%
Geographic distance	32%	57%

Table S3. Comparison between the relative risk of observing identical sequences between two geographic regions and the risk of movement between different geographies. We consider three data sources to inform the relative risk of movement between geographies: the relative risk for a visit to occur between two geographies (from mobile phone data), the relative risk for a work commute to occur between two geographies (from workflow data) and the geographic distance between geographies' centroids.

Facility name	County	Prison capacity	County population size	Ratio prison capacity / county population
Washington State Penitentiary	Walla Walla	2439	62584	$3.90 \cdot 10^{-2}$
Stafford Creek Corrections Center	Grays Harbor	1936	75636	$2.56 \cdot 10^{-2}$
Coyote Ridge Corrections Center	Franklin	2468	96749	$2.55 \cdot 10^{-2}$
Washington Corrections Center	Mason	1268	65726	$1.93 \cdot 10^{-2}$
Clallam Bay Corrections Center	Clallam	858	77155	$1.11 \cdot 10^{-2}$
Airway Heights Corrections Center	Spokane	2258	539339	$4.19 \cdot 10^{-3}$
Olympic Corrections Center	Clallam	272	77155	$3.53 \cdot 10^{-3}$
Monroe Correctional Complex	Snohomish	2400	827957	$2.90 \cdot 10^{-3}$
Cedar Creek Corrections Center	Thurston	480	294793	$1.63 \cdot 10^{-3}$
Larch Corrections Center	Clark	240	503311	$4.77 \cdot 10^{-4}$

Table S4. Characteristics of WA male prisons.

Supplementary text 1: Relationship between the number of transmission pairs and the number of pairs with timing consistent with a transmission direction

Notations

Let $N_{X\to Y}$ denote the number of transmission pairs where the infector is in subgroup X and the infectee in subgroup Y. Let $N_{X<Y}$ denote the number of transmission pairs between subgroups X and Y (regardless of the transmission direction) where the timing of symptom onset is earlier in X than in Y. Let N_{XY} denote the total number of transmission pairs between subgroups X and Y (regardless of the transmission direction). Let $p_{X\to Y} = N_{X\to Y}/N_{XY}$ denote the proportion of transmission pairs between X and Y that were in the direction $X \to Y$. Let $p_{X<Y} = N_{X<Y}/N_{XY}$ denote the proportion of transmission pairs between X and Y where the timing of symptom onset is earlier in X than in Y.

We introduce p_0 as the proportion of transmission events with positive serial intervals (defined by the delay between the onset of symptom of the infectee and the infector).

Relationship between $p_{X<Y}$ and $p_{X\to Y}$

Here, we demonstrate that comparing $N_{X\to Y}$ and $N_{Y\to X}$ is equivalent to comparing $N_{X<Y}$ and $N_{Y<X}$ as long as p_0 is greater than 50%.

$$N_{X<Y} = N_{X\to Y} \cdot p_0 + N_{Y\to X} \cdot (1-p_0) = N_{X\to Y} \cdot (2p_0 - 1) + (1-p_0)$$

By diving the two sides of this equation by N_{XY}, we have:

$$p_{X<Y} = p_{X\to Y} \cdot (2p_0 - 1) + (1-p_0)$$

Therefore, if $p_0 > 0.5$,

$$\mathbf{p_{X<Y} > 0.5} \;\Leftrightarrow\; p_{X\to Y} \cdot (2p_0 - 1) + (1-p_0) > 0.5 \;\Leftrightarrow\; p_{X\to Y} > \frac{p_0 - 0.5}{2p_0 - 1} \;\Leftrightarrow\; \mathbf{p_{X\to Y} > 0.5}$$

This means that as long as p_0 is greater than 0.5, comparing the number of transmission pairs between X and Y with symptom onset dates first occurring in X to the transmission pairs between X and Y with symptom onset dates first occurring in Y provides direct insights into the proportion of transmission pairs between X and Y happening in the $X \to Y$ direction.

Estimation of p_0 for SARS-CoV-2

For a given pathogen, p_0 can directly be estimated from a known serial interval distribution or from transmission pair data. For SARS-CoV-2, Geismar et al. [15] estimated this from reconstructed SARS-CoV-2 transmission events across a range of variants of concerns. In all analyses, their results show values greater of p_0 greater than 50 %.

Here, we focus on the timing of symptom onset within transmission pairs. However, this argument is directly transposable to other timing definitions, such as the timing of sequence collection dates by replacing p_0 by the proportion of transmission pairs where the sequence collection date of the infectee occurs after the sequence collection date of the infector.