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

Cécile Tran-Kiem[1,*], Miguel I. Paredes[1,2], Amanda C. Perofsky[3,4], Lauren A. Frisbie[5], Hong Xie[6], Kevin Kong[6], Amelia Weixler[6], Alexander L. Greninger[1,6], Pavitra Roychoudhury[1,6], JohnAric M. Peterson[5], Andrew Delgado[5], Holly Halstead[5], Drew MacKellar[5], Philip Dykema[5], Luis Gamboa[3], Chris D. Frazar[7], Erica Ryke[7], Jeremy Stone[3], David Reinhart[3], Lea Starita[3,7], Allison Thibodeau[5], Cory Yun[5], Frank Aragona[5], Allison Black[5], Cécile Viboud[4] & Trevor Bedford[1,8]

[1]Vaccine and Infectious Diseases Division, Fred Hutchinson Cancer Center, Seattle, WA, USA, [2]Department of Epidemiology, University of Washington, Seattle, WA, USA, [3]Brotman Baty Institute, University of Washington, Seattle, WA, USA, [4]Fogarty International Center, National Institutes of Health, Bethesda, MD, USA, [5]Washington State Department of Health, Shoreline, WA, USA, [6]Department of Laboratory Medicine and Pathology, University of Washington, Seattle, WA, USA, [7]Department of Genome Sciences, University of Washington, Seattle, WA, USA, [8]Howard Hughes Medical Institute, Seattle, WA, USA, *To whom correspondence should be addressed; E-mail: ctrankie@fredhutch.org

Abstract

Pathogen genomics can provide insights into disease transmission patterns, but new methods are needed to handle modern large-scale pathogen genome datasets. Genetically proximal viruses indicate epidemiological linkage and are informative about transmission events. Here, we leverage pairs of identical sequences using 114,298 SARS-CoV-2 genomes collected via sentinel surveillance from March 2021 to December 2022 in Washington State, USA, with linked age and residence information to characterize fine-scale transmission. The location of pairs of identical sequences is highly consistent with expectations from mobility and social contact data. Outliers in the relationship between genetic and mobility data can be explained by SARS-CoV-2 transmission between postal codes with male prisons, consistent with transmission between prison facilities. Transmission patterns between age groups vary across spatial scales. Finally, we use the timing of sequence collection to understand the age groups driving transmission. This work improves our ability to characterize transmission from large pathogen genome datasets.

Introduction

Pathogen transmission is impacted by a multiplicity of factors associated with individual, population and environmental characteristics. As exposure and transmission aren't directly observed, evaluating the contribution of these different factors to epidemic dynamics generally proves difficult. In order to anticipate the burden associated with epidemics and guide control policies, it is however pivotal to understand how these different elements shape transmission risk.

Sequence data can provide insights into the proximity of individuals in a transmission chain.

Phylogeographic approaches have helped characterize how pathogens spread between different geographical regions [1, 2] and demographic groups [3]. However, these methods currently face multiple limitations. First, they do not scale well past a few hundred or few thousand sequences due to difficulties in scaling phylogenetic tree inference. Second, conclusions can be highly biased when sequencing is uneven [4]. We thus critically need new methods to analyse large pathogen genome datasets, such as those produced during the COVID-19 pandemic, which number in the millions of genomes [5].

Here, we introduce a novel statistical framework describing the relative risk of observing genetically proximal sequences in specific subgroups of the population. Our metric of association accounts for heterogeneity in sequencing effort between sampled locations. We use this framework to investigate the spatial and social drivers of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) transmission in Washington state (WA) by analyzing 114,298 sequences (with associated age and home location information) collected through genomic sentinel surveillance in WA between March 2021 and December 2022.

Results and Discussion

Identical sequences are imprinted by SARS-CoV-2 spatial patterns of spread

As mutations accrue over time on pathogen sequences, individuals who are close together within a transmission chain are expected to be infected by genetically proximal viruses. For example, we expect that 64% of SARS-CoV-2 infected individuals are infected by a virus with the same consensus genome as their infector (Figure 1A). Identical sequences should hence be highly informative about SARS-CoV-2 transmission events as they are preferentially collected from the most epidemiologically linked individuals. In WA, we identify 17,231 clusters of identical sequences excluding singletons, corresponding to 59,660 sequences (Figure 1B). In some large clusters of identical sequences, we observe local spread prior to wider geographic expansion (Figure 1C). Using postal codes and collection dates, we estimated cluster radius in km. Across clusters, we find that the spatial expansion of clusters increases over time (Figure 1D) and is significantly lower than expected at random (Figure S1). The probability for a cluster to remain within the county and zip code where it was first identified decreases over time. These probabilities are significantly higher than expected at random (Figure S1). This confirms that clusters of identical sequences contain a strong spatial and temporal signature of spread.

Relative risk framework to look at the distribution of identical sequences

To quantify the association between subgroups of the population (such as geographical units or age groups) from genetic data, we introduce a measure of relative risk (RR) describing how the number of pairs of sequences separated by a fixed genetic distance observed in two subgroups differs from what we expect from the sequencing effort (Figure 1E). This RR can be interpreted as a measure of enrichment describing how the number of pairs shared by these two subgroups differs from what we expect from the overall number of pairs observed in these two subgroups.

Figure 1F depicts the relationship between the RR of observing sequences within the same county and the genetic distance between pairs. Among all counties, the median RR of observing identical sequences within the same county is equal to 4.7 (interquartile range: 2.4–21.2) across the time period. When considering a greater genetic distance between pairs, this signal decreases to plateau at 1. This confirms that the location of genetically close sequences (less than a couple

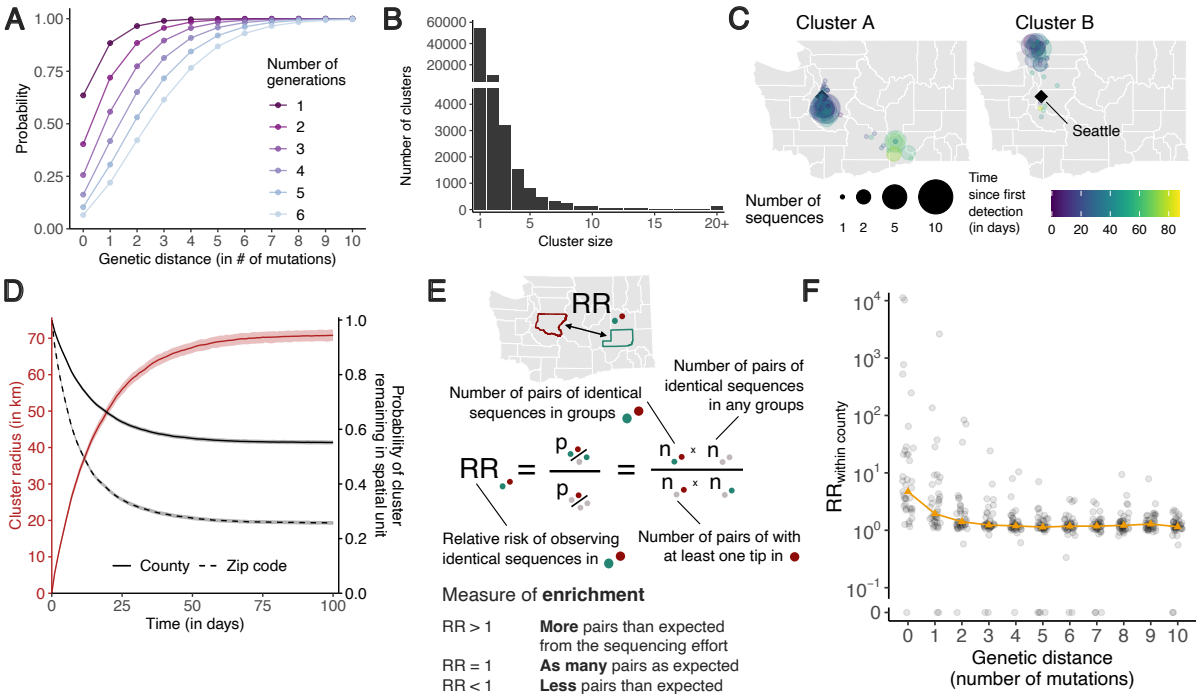


Figure 1. Temporal and spatial signature of spread in clusters of identical SARS-CoV-2 sequences. **A.** Probability for two individuals separated by a fixed number of transmission generations of being infected by viruses at a given genetic distance assuming a Poisson process for the occurrence of substitutions (at a rate $\mu = 8.98 \cdot 10^{-2}$ substitutions per day) and Gamma distributed generation time (of mean 5.9 days and standard deviation 4.8 days). **B.** Size distribution of clusters of identical sequences in the WA dataset. Clusters of size 1 correspond to singletons and are hence not included in the relative risk computations. **C.** Spatio-temporal dynamics of sequence collection in two large clusters of identical sequences. Black diamonds indicate the location of Seattle, the largest city in WA. **D.** Radius of clusters of identical sequences and probability for all sequences within a cluster of identical sequences of remaining in the same spatial units as a function of time since first sequence collection. In D, the cluster radius is computed as the mean spatial expansion of clusters of identical sequences. **E.** Definition of the relative risk of observing pairs of sequences in two subgroups as a measure of enrichment. **F.** Relative risk of observing pairs of sequences within the same county as a function of the genetic distance separating them. Grey points correspond to values for individual counties. Orange triangles correspond to the median across counties.

mutations away) and especially identical sequences is informative about local spread patterns, wherein infected individuals transmit more often within their home county.

We observe this trend across variants and periods (Figure S2). The magnitude of the absolute RR along with the speed at which it decays as a function of genetic distance vary. For example, during the period where the prevalence of Omicron rises and Delta declines, the RR of observing identical sequences in the same county is higher among Delta than Omicron sequences. This can be explained by differences in transmission intensity: a higher transmission rate results in larger clusters of identical sequences [6] that will tend to be more geographically widespread (Figure S2C). The spatial signal from genetically close sequences is hence weaker in periods characterized by a higher transmission intensity. Other factors, such as changes in mixing and travels patterns, can also impact the magnitude of the RR.

Sampling biases can considerably impact the results of phylogeographic inference [4]. Here, although the proportion of pairs of identical sequences observed in a county is highly correlated with the number of sequences observed in this county, we find that the RR is no longer correlated with sequencing effort (Figure S3). Using a simulation approach, we show that our RR metric captures the migration probability between population subgroups, including when sequencing effort is heterogeneous (Figure S4, Table S1). This contrasts with migration rates obtained from a discrete trait analysis (DTA) [7] which are poorly correlated with true migration rates when sequencing effort differs between regions (Figure S4, Table S1). These DTA results are obtained by inputting the exact simulated transmission tree. In practice, inferring the underlying tree will decrease accuracy due to phylogenetic uncertainty so that these DTA estimates represent an upper bound of DTA's potential performance. If we compare DTA accuracy between input phylogeny and phylogeny estimated from sequence data we find that Pearson correlation between true migration rates and estimated migration rates changes from 0.54 to 0.10 for unbiased sampling and changes from -0.22 to 0.15 for biased sampling (Table S1). Running the phylogenetic DTA analysis on simulated data with 1745 sequences requires 1 day when using the empirical tree and 24 days when jointly inferring the tree and the migration rates (see Methods). Running our RR analysis on the same sequence dataset takes 33 seconds. This result demonstrates that the RR framework constitutes an appropriate approach to study the determinants of SARS-CoV-2 transmission by explicitly accounting for sequencing effort and uneven sequencing between population subgroups.

Patterns of SARS-CoV-2 spread between WA counties

We explore geographic spread through analyzing patterns of occurrence of identical sequences in WA counties (Figure 2). The matrix of pairwise RRs between counties (Figure S5) is characterized by a strong diagonal, which is consistent with within-county transmission. To better understand the spatial patterns of SARS-CoV-2 spread between counties, we display these RRs on choropleth maps indicating RR for different focal counties (Figure 2A, Figure S6). These maps suggest that identical sequences have a higher risk of falling within counties that are geographically nearby. Across all pairs of counties, we find a geographic gradient in the RR of identical sequences, where the risk is highest within the same county, intermediate between adjacent counties, and lowest between non-adjacent counties (Figure 2B). The risk of observing identical sequences between counties also decays as a function of geographic distance (Figure 2C) and is no longer significant at distances greater than 177 km (95% confidence interval (CI): 137-241).

To assess whether global spatial structure is maintained, we implement a multidimensional scaling (MDS) algorithm by defining a similarity metric based on the RR of observing identical sequences between counties. MDS enables us to display the relatedness of observations based on a distance matrix. This MDS ordination shows country relationships that recapitulate the Western and Eastern WA regions, two regions separated by the Cascades mountain range (Figure 2D). Within Eastern and Western WA, we find a strong signal for local spread, with identical sequences having a higher risk of being observed between adjacent than between non-adjacent counties (Figure 2E). Across the Eastern / Western WA border, we no longer find that identical sequences have an increased risk of being observed in adjacent counties. Results are similar when analyzing pairs of identical sequences at the postal code level (Table S2). This lack of association is not affected by the low number of pairs of adjacent counties across the Eastern / Western WA border (Figure S7). This illustrates how heterogeneous physical landscape features can impact and distort patterns of disease spread and genetic diversity [8–11]. We also find that

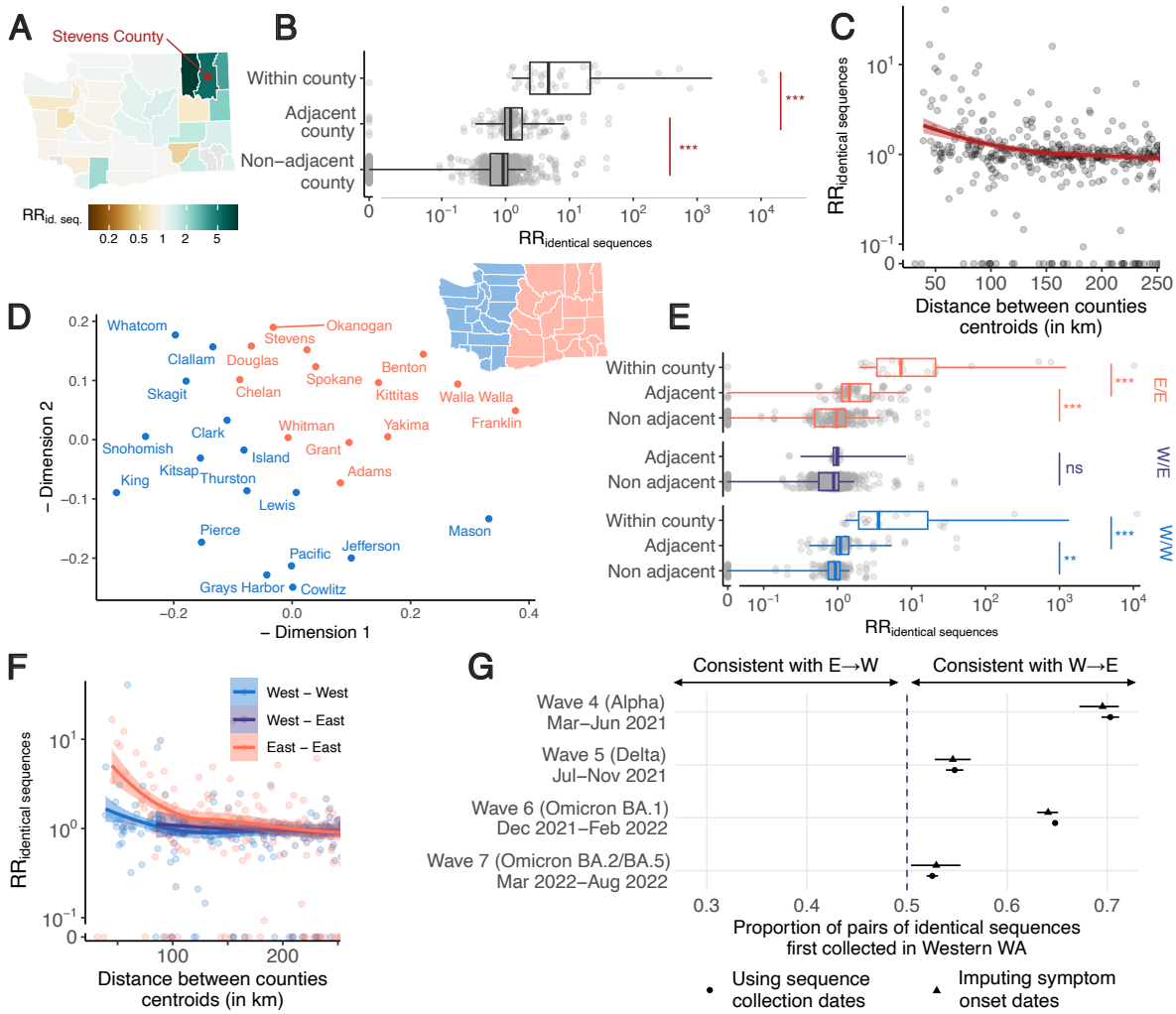


Figure 2. Identical sequences reveal patterns of spread between WA counties. **A.** Illustration of the pairwise relative risk of observing identical sequences between counties, using sequences shared between Stevens County (red point) and other counties in WA as an example. Similar maps for the other counties are depicted in Figure S6. **B.** Relative risk of observing pairs of identical sequences by counties' adjacency status. **C.** Relative risk of observing pairs of identical sequences as a function of the geographic distance between counties' centroids. **D.** Similarity between WA counties obtained from MDS based on the relative risk of observing pairs of identical sequences in two counties. Counties are colored by East / West region membership. **E.** Relative risk of observing pairs of identical sequences by counties' adjacency status stratified by counties East / West region membership. **F.** Relative risk of observing pairs of identical sequences as a function of the geographic distance between counties' centroids stratified by counties East / West region membership. **G.** Proportion of pairs of identical sequences observed in Eastern and Western WA that were first observed in Western WA. In C and F, the lines correspond to LOESS curves on the logarithmic scale. In B and E, p-values for Wilcoxon tests: *** < 0.0001, ** < 0.001, * < 0.05, ns ≥ 0.05.

the association between the RR of observing identical sequences in two counties is significant at greater distance within Eastern WA than within Western WA (Figure 2F). We do not find any association with distance across the W/E border, though this might be explained by the lack of counties with low distances across the Eastern / Western WA border.

Finally, we find that, across epidemic waves, pairs of identical sequences observed on both sides of the Cascades are consistently observed first in Western WA (Figure 2G, Figure S8). As testing behavior and access to healthcare are likely influenced by county demographic characteristics and how rural or urban a county is, we explore how this trend varied when using symptom onset dates instead of sequence collection dates, which provides similar trends (Figure 2G). This asymmetry suggests that identical sequence clusters tend to percolate from Western to Eastern WA more so than the reverse, indicating transmission generally flows from Western to Eastern WA. This trend is similar to the one reported in phylogeographic analyses of the first COVID-19 wave in WA, that concluded that more introductions occurred from Western to Eastern WA than from Eastern to Western WA [12].

Human mobility predicts the location of pairs of identical sequences

Next, we explore to which extent spatial transmission patterns inferred from identical sequences can be explained by human mobility indicators. We use aggregated mobile phone location data obtained from the Safegraph 'Weekly Patterns' dataset and pre-pandemic commuting data from the US Census Bureau [13] to compute the RR of movement between two counties or regions (see Methods). Despite commuting data being collected before the pandemic and mobile phone location data being collected during our study period, we find that these two mobility data sources are highly correlated (Figure S9). We assess how the RR of observing identical sequences in two counties relates to the RR of movement (Figure 3A, Figure S10, Table S3) by implementing a generalized additive model (GAM) that includes a single predictor of smoothed RR of movement between two counties as a covariate to predict RR of identical sequences between two counties. We use a GAM rather than linear regression as the functional form of the relationship appears non-linear (Figure 3A). When comparing RRs at the county level, we find that 60% of the variance in identical sequence data is explained by the mobile phone information (Table S3). For a subset of counties, the number of pairs of identical sequences or the number of trips reported in the mobility dataset are low. For these low counts, we expect RRs risks to be more noisy. To remove potential noise associated with these lower counts, we repeat this analysis at a larger spatial scale. Aggregating pairs at the regional level (9 regions in WA for 39 counties, Figure S11) increases the variance explained to 81% (Figure S12). We also find that pre-pandemic workflow data are highly informative of the spatial distribution of pairs of identical sequences with a similar strength of relationship as observed for mobile phone mobility data (Figure S10, S12, Table S3).

Non-pharmaceutical interventions along with behavioral changes have impacted human mobility patterns throughout the COVID-19 pandemic. We find that mobile phone derived mobility data explains a high percentage of variance in the RR of observing identical sequences between WA regions across epidemic waves (Figure S13) but not to a greater extent that over the entire study period. This can likely be explained by the high stability of the structure of the mobility network between WA counties across epidemic waves (Figure S14). This suggests that splitting the study periods in sub-periods tends to rather introduce noise than increase the spatial resolution, in line with a former analysis concluding to the high stability of between-county mobility patterns during the beginning of the pandemic in the United States [14].

Among counties located across the Eastern / Western WA border, the risk of movement across the border is lower than the risk of movement within the same region (Figure S15). This shows that human mobility is highly predictive of the location of pairs of identical sequences and explains some of the spatial patterns reported in Figure 2.

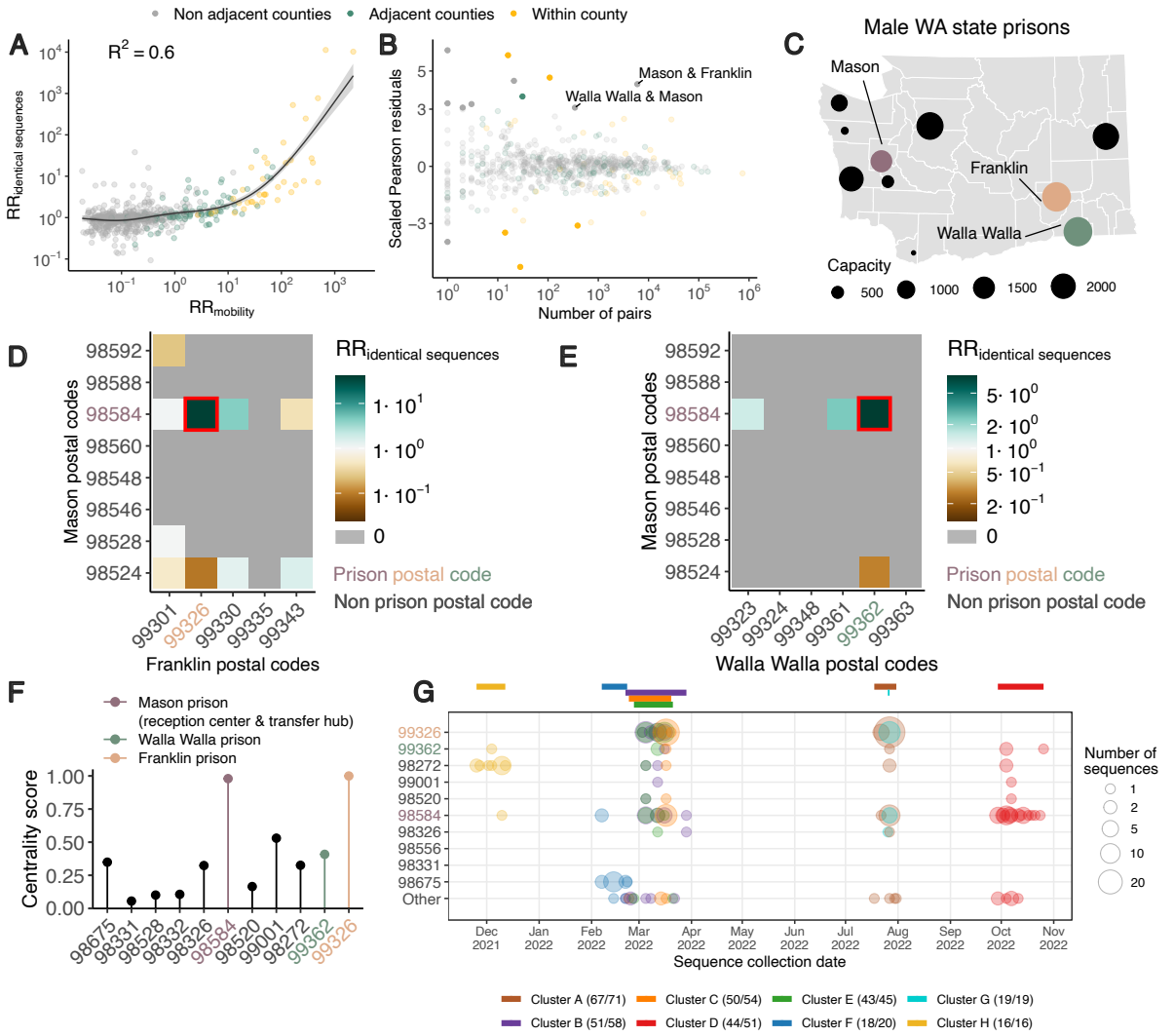


Figure 3. Comparison of the location of identical sequences with expectations from mobility data reveals spread between WA male prisons' postal codes. **A.** Relationship between the relative risk of observing identical sequences in two counties and the relative risk of movement between these counties as obtained from mobile phone mobility data. The trend line corresponds to predicted relative risk of observing identical sequences in two regions from a GAM. R^2 indicates the variance explained by the GAM. **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.** Map of male state prisons in WA. Mason, Walla Walla and Franklin male prisons are colored. **D.** Relative risk of observing identical sequence between Mason and Franklin County's postal codes. **E.** Relative risk of observing identical sequence between Mason and Walla Walla County's postal codes. **F.** Centrality score (eigenvector centrality) for each postal code that is the home of a male state prison. **G.** Timing of sequence collection of 8 large clusters of identical sequences identified in postal codes with WA male state prisons. In G, the top colored segments indicate the period during which each cluster was identified.

Outliers in the relationship between mobility and sequence data appear associated with male state prisons.

We identify unexpected patterns of transmission between counties from outliers in the relationship between mobility and genetic data (Figure 3A). We define outliers as pairs of counties for

which the absolute value of the scaled Pearson residuals from the GAM are greater than 3. As we expect RRs computed from a low number of pairs of identical sequences to be noisier, we focus on pairs of counties between which at least 100 pairs of identical sequences are observed. We find unexpected patterns of SARS-CoV-2 spread between two non-adjacent pairs of counties, Franklin/Mason and Walla Walla/Mason Counties, with more pairs of identical sequences observed than expected from mobility data (Figure 3B). The association between Franklin & Mason (RR of 13.4 (95% CI: 11.4-16.4)) and Walla Walla & Mason (RR of 5.9 (95% CI: 4.0-8.3)) is particularly surprising given that they are non-adjacent counties located on different sides of the Cascades. As no demographic or geographic factors provide a straightforward explanation for such an association, we hypothesize that such a pattern might arise from SARS-CoV-2 spread on a dissemination network that differs from the general community. We identify that these three counties are the home of male state correction centers (Figure 3C). We also find that identical sequences have a higher risk of being observed within Lincoln County and a lower risk of being observed within Pacific County that expected from mobile phone mobility data, without identifying any demographic factor explaining these associations.

To investigate whether the unexpected pattern of association between Franklin & Mason and Walla Walla & Mason Counties can be explained by transmission within the prison network, we look at patterns of association between Franklin & Mason and Walla Walla & Mason postal codes (Figure 3D-E). For most of these pairs of postal codes, we don't observe any pair of identical sequences throughout the study period. Interestingly, for each pair of counties, the genetic signal can be explained by a high RR of observing identical sequences between two postal codes, which correspond to the postal codes that are the home of the male correction centers we identified. The greater number of pairs of identical sequences observed between Mason & Franklin and Mason & Walla Walla counties than expected from mobile phone derived mobility data can hence be explained by a large number of pairs of identical sequences in specific postal codes with male correction centers.

We also investigate patterns of occurrence of pairs of identical sequences between the two counties (Mason and Pierce) that are the home of female prisons. At the county level, identical sequences don't have an increased risk of occurring between Mason & Pierce counties (RR of 0.59 (95% CI: 0.49-0.67)). At the postal code level however, we find that the RR of observing identical sequences is highest between the two postal codes with female prisons (Figure S16). This shows how our framework enables exploration of patterns of spread at different spatial scales: we don't find any signal at the county level, likely because Mason and Pierce are adjacent counties, but we can identify association at the postal code level.

It is interesting that the pairs of outliers we identified systematically involved Mason County (Figure 3B), which is the home of only the sixth (out of ten) most populated male prison in the state (Figure 3C). The prison in Mason County (Washington Corrections Center) plays a particular role in the WA prison network since it serves both as a reception center for anyone entering the WA prison system and as a transfer hub [15]. To understand whether the prison network structure can explain patterns of SARS-CoV-2 transmission, we conduct a centrality analysis. To do so, we analyse the network of postal codes with WA male prisons and we define the weight of each edge by the RR of observing identical sequences between these two postal codes. We find that the two nodes with the highest centrality scores are the postal codes that are the home of Washington Corrections Center (Figure 3F) and of the Franklin County prison (most populated prison). This shows that patterns of occurrence of identical sequences in WA are imprinted by the structure of the prison network.

Finally, we investigate whether large clusters of identical sequences are shared between postal codes with male state prisons, which we define as clusters with more than 15 sequences in male state prisons postal codes. Figure 3G depicts the timing of the large clusters we identify. Notably, the largest cluster (Cluster A) includes 71 sequences collected between 18 July and 31 July 2022, 67 of which came from postal codes with male state prisons. The second largest cluster (Cluster B) is composed of 58 sequences collected between 21 February and 29 March 2023, among which 51 came from 7 different prison postal codes. Interestingly, the postal code of Washington Corrections Center is the only one in which all these eight clusters were observed.

Populations who are incarcerated have been particularly affected by the COVID-19 pandemic [16, 17]. To mitigate the impact of the pandemic in these congregate settings, various interventions have been implemented. In WA, for example, testing followed by quarantine protocols were carried out in Washington Correction Center upon admission and before any transfer. Active screening of staff was also implemented throughout the pandemic. Individuals incarcerated diagnosed with COVID-19 however at times had to be transferred from Washington Correction Center to other WA prisons due to the finite capacity of the reception center. With vaccine mandates, staff also had to be relocated to cope with the departure of other employees. Our results reveal multiple SARS-CoV-2 introductions between WA prisons, that could be explained by the movements of both individuals incarcerated and staff.

This analysis showcases how identical sequences can help identify under recognized viral dissemination networks that differ from the general community. The counties we identified as outliers in the relationship between genetic and mobility data have a particularly high ratio between the prison population size and the county population size (between around 2% and 4%, Table S4). This likely explains why we were able to detect this signal at the county level but had to investigate transmission at the postal code level to study transmission between other prisons.

The spatial scale of spread impacts transmission patterns between age groups

Spatial and social factors (such as age) are key determinants of the spread of respiratory infections such as SARS-CoV-2 and influenza [18–21]. We expect movement patterns to differ between age groups (such as children, adults and elderly people), which can impact patterns of disease transmission [22–24]. There has however been limited empirical evidence of this phenomenon and data sources that can be leveraged to characterize this interaction are critically needed. Here, we show that we can combine pathogen sequence information with detailed metadata to investigate how age mixing patterns vary across spatial scales.

We first examine whether we can recover the expected age mixing signature from the sequence data before delving into the interaction between age and space. We find that the age groups in which identical sequences are observed are consistent with assortative mixing patterns and mixing between generations (Figure S17). Comparing this with expectations from synthetic social contact data for WA [25], we find that the signal obtained from identical sequences are highly correlated with that expected from age mixing matrices (Figure 4A) (90% of variance explained using a GAM ; Spearman $\rho = 0.86$, $p < 10^{-16}$). The signal for SARS-CoV-2 transmission between generations (such as the 0-9y and the 30-39y) fades out when considering pairs of sequences separated by a greater genetic distance (Figure S18). As sequences at a greater genetic distance come from individuals who are further apart within a transmission chain (Figure 1A), fine-scale patterns of spread might indeed not be apparent from sequences at more than a couple mutations away. This emphasizes the value of analyzing identical pathogen sequences to characterize subtle patterns of pathogen spread and population mixing, especially when population

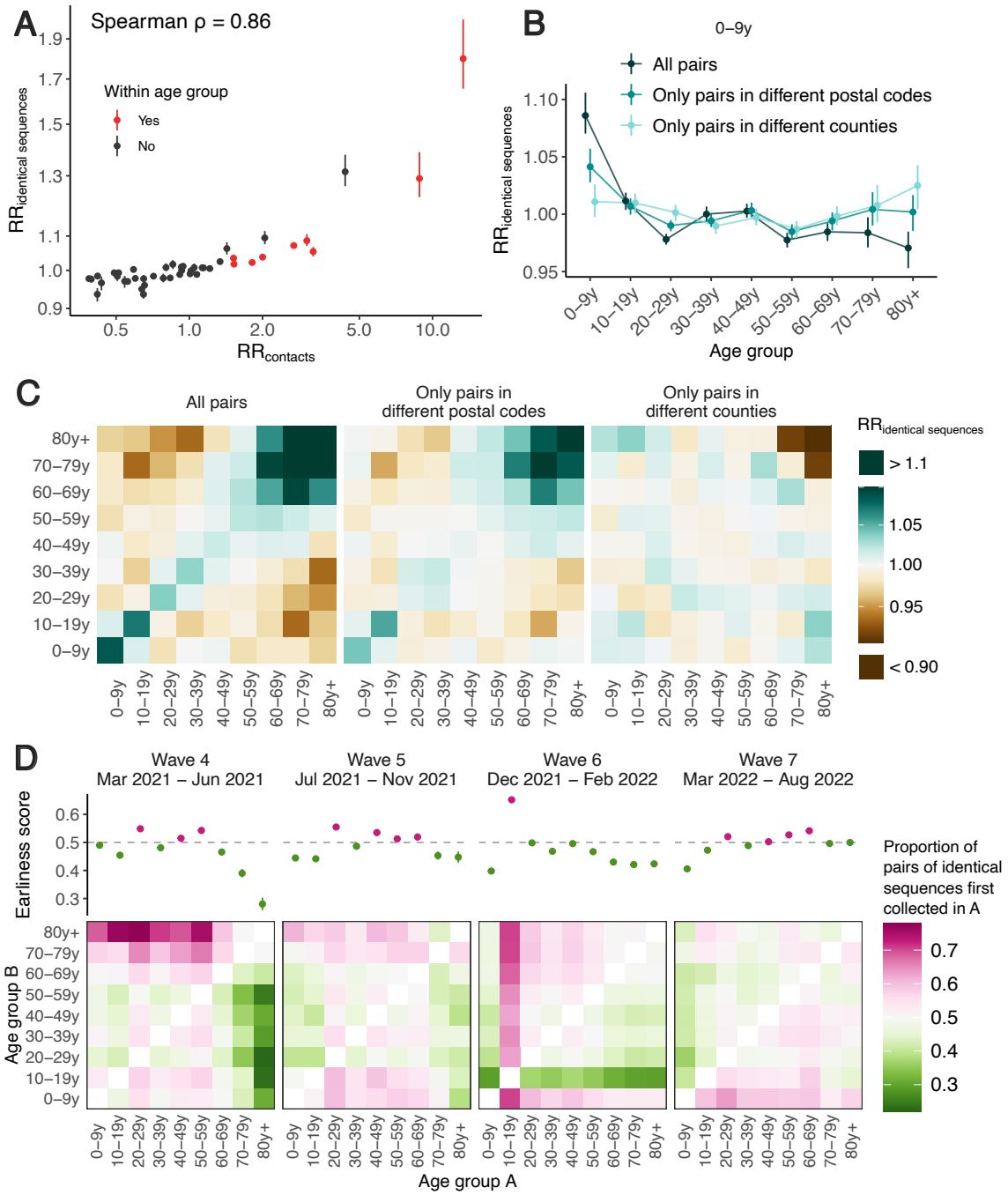


Figure 4. Patterns of SARS-CoV-2 transmission between age groups in WA. A. Relative risk of observing pairs of identical sequences in two age groups as a function of the relative risk of contact between these age groups. **B.** Impact of the spatial scale on the relative risk of observing pairs of identical sequences in the 0-9 y.o. and other age groups. We display similar plots for the other age groups in Figure S19. **C.** Relative risk of observing identical sequences between two age groups across all pairs of sequences, only pairs in different postal codes and only pairs in different counties. **D.** Proportion of pairs of identical sequences observed in age groups A and B that were first collected in age group A across different epidemic waves (heatmaps). The dot plots depict the earliness scores of age group A across epidemic waves. In A and B, vertical segments correspond to 95% subsampling confidence intervals. In D, vertical segments correspond to 95% binomial confidence intervals. In D, the heatmaps represent symmetric matrices $P = (p_{i,j})$ characterized by $p_{i,j} + p_{j,i} = 1$.

subgroups are very mixed.

Next, we compare the RR of observing identical sequences between two age groups by looking at either all pairs of sequences or only pairs of sequences from individuals living in different spatial units. We find that the spatial scale modulates patterns of disease transmission between age groups (Figure 4B-C, S19). We find that pairs of identical sequences coming from the same county and postal code are enriched in same-age pairs. This enrichment is particularly important in elderly groups. For example, the RR of observing identical sequences in individuals aged 80 and older drops from 1.80 (95% CI: 1.65-2.01) when considering all pairs to 0.79 (95% CI: 0.71-0.85) when considering pairs coming from individuals living in different counties (Figure S19). This shows that transmission to and from older age groups tends to occur close to their home location and suggests that elderly individuals' typical movement scale is lower than that of other age groups. Only considering pairs of sequences in 0-9 year-olds coming from different spatial units largely decreases the signal for SARS-CoV-2 transmission between children and adults aged 30-49y (Figure 4C). This is expected given that we anticipate most of these contacts to occur within the household [26]. Overall, we find that looking at patterns of occurrence of identical sequences at a greater geographic scale largely distorts the contact structure. For example, the location of identical sequences suggest that transmission to and from elderly individuals outside of their home counties tend to occur with younger age groups, including younger children (e.g. grandchildren).

Mixing patterns between age groups have been extensively studied [27, 28]. A social contact survey performed in southern China reported that elderly individuals' contacts occurred closer to their homes than younger individuals' contacts [29]. Overall, there has however been limited evidence to quantify the spatial distribution of these contacts. Spatial mixing is generally measured from mobility data sources which generally do not provide demographic information such as age. Because spatial and age mixing are reconstructed from different data sources, understanding their interplay has been difficult. Here, we show that we can directly leverage pathogen genome data with linked age and spatial information to understand where age-specific transmission is occurring. This suggests that the wider availability of sequencing data might provide an opportunity to directly infer how population groups interact in a way that is relevant for pathogen spread, without the need to implement laborious contact or mobility surveys.

The timing of identical sequence collection sheds light on the groups driving transmission

Finally, we use the timing of identical sequence collection to investigate the age groups driving SARS-CoV-2 transmission over the course of the pandemic in WA. We indeed expect pairs of identical sequences to have groups acting as source consistently observed before groups acting as sink (Figure S20). In Figure 4D, we display for every age group combination and across epidemic waves the proportion of pairs of identical sequences first collected in a given age group. During the fourth and fifth pandemic wave in WA (respectively mainly caused by the Alpha and Delta SARS-CoV-2 variants of concern), we find that pairs of identical sequences are consistently observed later in elderly groups even though the RR of observing identical sequences in elderly groups and younger groups is low (Figure S17). This is consistent with younger age groups acting as source of infections for elderly individuals. During the fourth pandemic wave, sequences from individuals aged 20-29y and 40-59y are systematically observed before any other groups within pairs of identical sequences and likewise during the fifth pandemic wave, sequences from 20-29y and 40-69y individuals are observed earlier than other age groups. This is consistent with these

groups acting as sources of infection for the other age groups. During the sixth wave, sequences from 10-19y tend to be observed first within pairs of identical sequences, which is consistent with them acting as sources for other ages groups and corresponds to the Omicron wave during a time when schools had recently returned to in-person instruction. From March 2022, the contribution to transmission is more distributed across age groups.

The role played by young children during the COVID-19 pandemic has been highly debated [21,30]. Here, we find that during the Alpha and Delta epidemics (waves 4 and 5), children aged 0-9y were acting as source of SARS-CoV-2 infections for elderly individuals but not for younger adults. This pattern disappears during the Omicron epidemic (wave 6) in which pairs of identical sequences tend to be observed first in other age groups before being collected in young children aged 0-9y. This could be explained by behavioral changes or by different immune profiles across age groups, resulting in different relative susceptibility to Omicron relative to Delta [31]. Overall, we do not find evidence for young school age children to act as major sources of SARS-CoV-2 transmission in the population, even after schools reopened.

Overall, our results highlight the porosity of SARS-CoV-2 transmission across age groups as well as the role played by lower risk groups in seeding infections in higher risk groups. We come to similar conclusions when looking at the timing of symptom onset dates (Figure S21), which suggests that our conclusions are robust to differences in testing behaviors across age groups. Our conclusions are in line with existing literature emphasizing the important role played by young adults and teenagers and the limited contribution of children and elderly individuals in driving SARS-CoV-2 spread [21, 32]. Analyzing the timing of identical sequence collection provides immediate insights into pathogen flow between population subgroups.

Here, we focus on understanding patterns of SARS-CoV-2 spread between age groups but this approach can be applied to investigate the spread of fast-evolving pathogens between various demographic groups, such as occupational and ethnic groups, behavioral and risk groups. For example, we find that identical sequences collected within an age group tend to be enriched in same-vaccine status pairs during the Alpha, Delta and Omicron BA.2/BA.5 waves in WA (Figure S22). Social clustering of unvaccinated individuals or generally individuals with different immune background can have important implications regarding the size and likelihood of infectious disease outbreaks [33, 34]. This suggests that our approach has potential to shed light on such a phenomenon and more generally on broad determinants of disease transmission

Caveats

Though our RR metric explicitly accounts for sampling intensity in locations in which pairs of identical sequences are collected, it cannot describe patterns of spread from non-sampled locations. Compared with existing phylogeographic methods, our approach however does not require including background sequences from outside locations and non-sampled locations don't impact the RR computation (Figure S23). Our approach could also overestimate RRs associated with transmission events that are over-represented in the sequencing data. For example, applying this analysis to sequences predominantly collected through household studies could overestimate the contribution of contacts within the household to the overall infection burden. In our case, patterns of occurrence of identical sequences in WA are potentially affected by intensive testing performed during outbreaks within WA prisons during the pandemic. Part of the signal we have detected might hence come from a higher sequencing rate in prisons compared with the general community. The very large clusters of identical sequences shared between multiple prison postal codes however confirm that SARS-CoV-2 considerably spread within the prison network. As the

magnitude of RRs is impacted by transmission intensity (Figure S2), more work is also required to quantify changes in mixing patterns over time.

Applicability beyond SARS-CoV-2

In this manuscript, we study patterns of SARS-CoV-2 transmission between geographies and age groups in WA using a particularly rich sequence dataset, both given the amount of sequences available and the quality of the associated metadata. This work can however readily be applied to other densely sampled pathogens.

The power of our method is determined by the number of pairs of identical sequences available which will be impacted by transmission intensity (higher reproduction numbers will tend to result in larger clusters of identical sequences [6]) and the relative timescale at which substitution and transmission events occur [6]. Overall, this approach is well tailored to study densely sampled outbreaks. Compared with phylogeny-based methods whose power comes from the number of unique haplotypes, our ability to characterize spread from identical sequences depends on the number of haplotypes with multiple observations. Whereas we expect diminishing returns of sequencing a greater proportion of cases for phylogeny-based methods with a decreasing number new haplotypes per additional sequence, the number of haplotypes with multiple occurrences will increase for each additional sequence included in the dataset (Figure S24). The number of population groups included in the analysis will also impact the amount of sequencing data required. In situations where the sample size results in a lower number of pairs of identical sequences, aggregating groups can be a valuable strategy to reduce uncertainty. Within our WA sequencing dataset, we find that assessing spread between 2 age groups requires around $10^2 - 10^3$ sequences whereas 9 age groups increases the number of sequences required to $10^4 - 10^5$ (Figure S25).

Perspectives

Large scale pathogen genome sequencing provides an incredible opportunity to understand where and how transmission is occurring. The computational cost of existing methods that rely on inferring the pathogen's phylogenetic tree has limited their ability to elucidate fine-scale transmission patterns. Here, we show that a simple count-based metric based on pairs of identical pathogen sequences with detailed linked metadata can provide unique insights into the determinants of SARS-CoV-2 transmission. Future work investigating how to better describe asymmetry in transmission between groups and how to infer group-level contributions to epidemic growth from such data are a promising research direction. This shows that relying on pairs of identical or nearly identical pathogen sequences along with fine-grained metadata is valuable to understand how and where transmission is happening. By providing scalable new tools to understand detailed pathogen spread patterns, we believe that this work represents an important development to guide future epidemic control efforts.

Methods

Data sources and preprocessing

Sequence data and metadata

We analyze 116,791 SARS-CoV-2 sequences from Washington state genomic sentinel surveillance system [35] sampled between 1 March 2021 and 31 December 2022. Sequence metadata are

collated by the Washington State Department of Health and include sample collection date, symptom onset date, de-identified patient ID, county of home location, postal code of home location, age group (0-9y, 10-19y, 20-29y, 30-39y, 40-49y, 50-59y, 60-69y, 70-79y and 80y+) and vaccination status upon positive test. For patients with multiple sequences in the database (2,309 out of 114,306 patients), we restrict our analysis to the earliest sequence collected. Among these 114,306 sequences, the age information is missing for 1 sequence, the county information for 659 sequences and the postal code information for 1011 sequences.

Consensus sequences are extracted from the GISAID EpiCoV database [36, 37] and curated using the Nextstrain nCoV ingest pipeline [38]. We discard sequences with undefined Nexstrain clade assignments (8 sequences out of 114,306). This leaves us with 114,298 sequences, with 114,297 sequences gathered from patients with known age, 113,639 sequences gathered from patients with known county of home location and 113,287 sequences gathered from patients with known postal code of home location. 96 % of sequences have coverage in greater than 90% of the genome. We match postal codes to zip code tabulation areas (ZCTAs). For postal codes that do not have a ZCTA with the same name, we manually match them by looking at ZCTA boundaries. All analyses at the postal code level use ZCTA metadata information. We extract the postal codes of WA prison facilities from [39].

Computing pairwise genetic distances from sequences

We compute pairwise genetic distances between Washington state sequences with the *ape* R package [40] using Hamming distances. To avoid unnecessary computational costs, we only compare sequences belonging to the same Nextstrain clade [41] and generate one distance matrix per clade.

Generating pairs of identical sequences from a sequence data file is the most computationally expensive step in this analysis. To provide context, generating a distance matrix from 1,000 sequences takes 33 seconds, while 10,000 sequences takes 1 hour 37 minutes, on 1 core of an Apple M2 chip. Generating the full distance matrix for the analysis set of 113,287 sequences took around 96 hours of compute time readily parallelized across a compute cluster. More efficient software tools can significantly bring that compute time down (e.g. 1.14 hour with pairsnp [42]).

Workflow data

We use data describing the daily number of commuters between each WA county from the American Community Survey (2016-2020) [13]. This dataset provides the number of directed commuting flows between residence and workplace counties. We use the number of commuting flows between counties to compute the RR of commute between two regions (see below).

Mobile phone location data

We obtain mobile device location data from SafeGraph (https://safegraph.com/), a data company that aggregates anonymized location data from 40 million devices, or approximately 10% of the United States population, to measure foot traffic to over 6 million physical places (points of interest) in the US. Following Perofsky et al. [43], we estimate movement within and between counties in Washington from January 2020 to June 2022, using SafeGraph's *Weekly Patterns* dataset, which provides weekly counts of the total number of unique devices visiting a point of interest (POI) from a particular home census block group. POIs are fixed locations, such as businesses or attractions. A *visit* indicates that a device entered the building or spatial perimeter

designated as a POI. The *home location* of a device is defined as its common nighttime (18:00-7:00) census block group (CBG) for the past 6 consecutive weeks. We restrict our dataset to POIs that have been tracked by SafeGraph since December 2019. To measure movement within and between counties, we extract the home CBG of devices visiting POIs in each week and limit the dataset to devices with home CBGs in the county of a given POI (within-county movement) or with home CBGs in counties outside of a given POI's county (between-county movement). To adjust for variation in SafeGraph's device panel size over time, we divide Washington's census population size by the number of devices in SafeGraph's panel with home locations in Washington state each month and multiply the number of weekly visitors by that value. For each mobility indicator, we sum adjusted weekly visits across POIs from March 2021 to June 2022. We use the number of visits between counties to compute the RR of movement from mobile phone data between two regions (see below).

Social contact data

We use synthetic social contact data for WA generated by Mistry et al. [25] based on reconstructing synthetic populations of interacting individuals using WA population demographics. They describe the per-capita probability for an individual of age i of interacting with someone of age j during a day.

Quantifying connectivity between different population groups

From genetic data: Relative risk for sequences separated by a given genetic distance of being in given subgroups of the population

Let $n^d_{A,B}$ denote the number of pairs of SARS-CoV-2 sequences separated by a genetic distance d that are collected in subgroups A and B of the population. Let $n^d_{A,\bullet}$ denote the number of pairs of sequences separated by a genetic distance d where at least one element lies within subgroup A. Let $n^d_{\bullet,\bullet}$ denote the total number of pairs of sequences separated by a genetic distance d.

We derive the relative risk $\mathrm{RR}^d_{A,B}$ for sequences separated by a genetic distance d of being observed in subgroups A and B compared to what is expected from the sequencing effort in the different subgroups of the population as:

$$RR^d_{A,B} = \frac{n^d_{A,B}/n^d_{A,\bullet}}{n^d_{B,\bullet}/n^d_{\bullet,\bullet}} = \frac{n^d_{A,B} \cdot n^d_{\bullet,\bullet}}{n^d_{A,\bullet} \cdot n^d_{B,\bullet}} \tag{1}$$

The numerator $n^d_{A,B}/n^d_{A,\bullet}$ corresponds to the proportion of pairs of sequences observed in group A that are occurring with the B group. The denominator $n^d_{B,\bullet}/n^d_{\bullet,\bullet}$ corresponds to the proportion of pairs of sequences observed in group B across all pairs of sequences. The ratio between these two quantities hence quantifies to which extent pairs of sequences observed in groups A and B are enriched compared to the number of sequences observed in these groups.

We use a subsampling strategy to compute confidence intervals around these RRs. Bootstrapping (random sampling with replacement) would result in comparing sequences with themselves and therefore lead to biased upwards RRs of observing identical sequences in the same group. To avoid this, we used a subsampling strategy (random sampling without replacement) with a 80% subsampling rate (1,000 replicate subsamples).

We provide the tools to compute this RR metric from user-provided sequence and metadata files in the GitHub repository associated with this manuscript [44].

From mobility data: Relative risk of movement between two geographical locations

Both the mobile phone and commuting mobility data provide directed flows between WA counties. Let $w_{A\to B}$ denote the number of commuters reported in the commuting data (respectively the number of visits for the mobile phone mobility data) whose home residence is in county A and who work in county B (respectively for which a visit in county B is reported). We compute the total movement flow between counties A and B as :

$$w_{A,B} = w_{A\to B} + w_{B\to A}$$

We then calculate the relative risk $\mathrm{RR}^{mobility}_{A,B}$ of movement between counties A and B as:

$$RR^{mobility}_{A,B} = \frac{w_{A,B} \cdot w_{\bullet,\bullet}}{w_{A,\bullet} \cdot w_{B,\bullet}}$$

where $w_{X,\bullet} = \sum_Y w_{X,Y}$ and $w_{\bullet,\bullet} = \sum_{X,Y} w_{X,Y}$.

We compute a similar statistic by aggregating counties at the regional level (Figure S11).

From social contact data: Relative risk of contact between two age groups

Mistry et al. estimated the average daily number of contacts $M_{i,j}$ that individuals of age i have with individuals of age j (considering one-year age bins) [25]. As we are interested in the age groups available in the sequence metadata, we reconstruct the average daily number of contacts $c_{A,B}$ that individuals within age group A have with individuals in age group B as:

$$c_{A,B} = \frac{\sum_{i\in A}\sum_{j\in B} M_{i,j} \cdot n_i}{\sum_{i\in A} n_i}$$

where n_i is the number of individuals of age i. We can then derive the total daily number of contacts between age groups A and B as $\Gamma_{A,B} = c_{A,B} \cdot N_A$ where N_A is the number of individuals in age group A. We then compute the relative risk $\mathrm{RR}^{contacts}_{A,B}$ for a contact of occurring between age groups A and B compared to what we expect if contacts were occurring at random in the population as:

$$RR^{contacts}_{A,B} = \frac{\Gamma_{A,B} \cdot \Gamma_{\bullet,\bullet}}{\Gamma_{A,\bullet} \cdot \Gamma_{B,\bullet}}$$

where $\Gamma_{A,\bullet}$ is the total daily number of contacts involving individuals within age group A and $\Gamma_{\bullet,\bullet}$ is the total daily number of contacts in the population.

Using the timing of sequences to understand directionality in transmission

From sequence collection dates

We introduce t_x as the time at which the sequence x was collected. Let $I_{A,B}$ denote the ensemble of pairs of identical sequences observed in groups A and B.

$$I_{A,B} = \big\{(a,b) \in I_{A,B} \mid |t_a - t_b| > 0\big\}$$

thus denotes the subset of these pairs with different sequence collection dates. We compute the proportion $p_{A\to B}$ of pairs consistent with the transmission direction $A \to B$ as:

$$p_{A\to B} = \frac{\#\big(\{(a,b) \in I_{A,B} \mid t_a < t_b\}\big)}{\#(I_{A,B})}$$

where $\#(X)$ is the cardinal of X.

We also report 95% binomial confidence intervals around these proportions.

From symptom onset dates

The delay between infection and sequence collection can be impacted by healthcare seeking behaviors and access to testing which might vary across age groups, geographical locations and time periods. If the distribution of the delay until testing differs between two subgroups A and B, the proportion of pairs of identical sequences $p_{A \to B}$ that are first collected in group A will both reflect the timing of infections and healthcare seeking behaviors. When available, symptom onset dates should be less impacted by healthcare seeking behaviors.

Among the 113,638 SARS-CoV-2 sequences with associated age group and county of home location information, symptom onset dates are available for 34,167 of them (30%). The availability of symptom onset information is susceptible to be impacted by individual demographic profiles (such as age), which could result in sequences with symptom onset information not being representative of all the sequences available. To avoid this, we impute missing symptom onset dates based on the empirical delay distribution between symptom onset and sequence collection (computed from individuals with known symptom onset dates) stratified by age group, time period and Eastern/Western WA region (Figure S26). Out of the sequences with known symptom onset dates, the absolute value of the delay between sequence collection and reported symptom onset is strictly greater than 30 days for 192 of them (<0.6%). We discarded these sequences in the computation of the symptom onset to sequence collection delay and considered that they were equivalent to sequences with missing symptom onset information (and hence imputed their symptom onset dates). We generate 1,000 imputed datasets. For each of these imputed datasets, we compute the proportion $p^{sympto}_{A \to B}$ of pairs with symptom onset dates occurring first in group A among pairs of identical sequences in groups A and B with distinct symptom onset dates. We then report the median across these 1,000 imputed datasets. We also generate a measure of uncertainty by computing on each of the imputed datasets 95% binomial confidence intervals around the proportion $p^{sympto}_{A \to B}$. We then report an uncertainty range around these proportion by using the minimum lower bound of the 95% CI and the maximum upper bound of the 95% CI across the imputed datasets.

Spatial analyses

Assessing the spatial extent of clusters of identical sequences

We reconstruct clusters of identical sequences from the pairwise genetic distance matrices [6]. To assess the spatial and temporal signal in clusters of identical sequences, we evaluate how the spatial extent of a cluster (summarised by its radius) evolves over time. For each cluster, we define primary sequences as the cluster's earliest collected sequence. We then define the cluster's primary ZIP Code Tabulation Areas (ZCTAs) as the ZCTAs of its primary sequences. We exclude clusters with ambiguous primary ZCTA (several primary ZCTAs) from this analysis. We define the radius of a cluster at a given time as the maximum distance between the primary ZCTA and the ZCTAs of the sequences collected by that time. We also compute the time required for sequences to be collected outside the primary ZCTA and primary county (using a similar definition as for the primary ZCTA). We report the mean cluster radius and the proportion of clusters remaining within the same geographical unit (ZCTA and county) as a function of the time since collection of the first sequence within a cluster. We generate 95% confidence intervals using a bootstrap approach with 1,000 replicates.

We compare the observed cluster radius and the observed proportion of clusters remaining within the same geographical unit to those expected from a null distribution assuming no spatial depen-

dency between sequences within a cluster of identical sequences. We simulate a null distribution by randomly permuting the geographical locations of the WA sequences and recomputing our statistics of interest (cluster radius, proportion of clusters within the same county and proportion of clusters within the same ZCTA).

Comparing the location of pairs identical sequences by counties' adjacency status

We compare the RR of observing identical sequences between two counties depending on counties' adjacency status (within the same county, between adjacent counties and between non-adjacent counties) by using Wilcoxon signed rank tests.

Assessing the relationship between the RR of observing identical sequences between two counties and the geographic distance separating them

We explore how the RR of observing identical sequences in two distinct counties compare with the geographic distance between counties' centroids. We summarise this trend by reporting the LOESS curves with 95% confidence intervals between the log RRs and the distance in kilometers.

Mapping the association between counties using multidimensional scaling

We evaluate the extent to which patterns of association obtained when looking at the location of pairs of identical sequences are consistent with global spatial structure. To do so, we perform Non-metric Multidimensional Scaling (NMDS) based on the matrix of RR of observing pairs of identical sequences between two counties. We restrict our analysis to the subset of counties for which there was always at least 5 pairs of identical sequences observed with the other counties in the subset. This is done to remove potential noise associated with low number of pairs observed. As the NMDS algorithm requires a measure of similarity between counties, we define the similarity $s_{A,B}$ between counties A and B as:

$$s_{A,B} = e^{-RR^0_{A,B}}$$

We perform 2-dimensional NMDS using the *vegan* R package [45].

Exploring directionality in transmission between Eastern and Western WA

We evaluate whether the timing of identical sequence collection is consistent with transmission rather occurring from Western to Eastern or Eastern to Western WA. We define 4 time periods corresponding the 4 epidemic waves experiences by WA during our study period (Figure S27):

- Wave 4: March 2021 - June 2021
- Wave 5: July 2021 - November 2021
- Wave 6: December 2021 - February 2022
- Wave 7: March 2022 - August 2022

For each of these time periods, we compute the proportion of pairs of identical sequences first collected in Eastern WA among pairs of identical sequences observed in both Eastern and Western WA that were not collected on the same day. We report 95% binomial proportion CI around these proportions.

To explore whether our conclusions could be explained by differences in testing behaviors between Eastern and Western WA, we conduct a sensitivity analysis by imputing the date of symptom onset.

Mobility analyses

Evaluating the relationship between genetic and mobility data

We compute the Spearman correlation coefficient between the RR of observing identical sequences between two counties and the RR of movement between two counties (both from mobile phone derived and commuting mobility data) as well as the geographic distance between counties' centroids. We determine the percentage of variance in the genetic data explained by the mobility data by fitting generalised additive models (GAMs) predicting the RR of observing identical sequences based on the RR of movement between two counties, both on a logarithmic scale, using a thin plate smoothing spline with 5 knots. For the GAM analyses, we remove pairs of counties for which the number of pairs of identical sequences or the total mobility flow is equal to 0, which ensures that both the RR of observing identical sequences and the RR of movement are strictly positive. We also fit a GAM between the RR of observing identical sequences between two counties (on a logarithmic scale) and the distance between counties centroids. We repeat these analyses at the regional level, instead of at the county one.

Identifying outliers in the relationship between genetic and mobility data

We define outliers in the relationship between genetic and mobility data as pairs of counties for which the absolute value of the scaled Pearson residuals of the GAM is greater than 3. As we expect RRs computed from a low number of pairs of identical sequences to be more noisy, we focus on pairs of counties for which at least 100 pairs of identical sequences are observed throughout the study period.

Characterizing spread between postal codes with male prisons

Centrality analysis

We characterize transmission between the 10 postal codes with male state prisons by performing a network centrality analysis. We consider a network with 10 nodes corresponding to these different postal codes. We define the weight of each edge as the RR of observing identical sequences between the two postal codes that define the nodes connected by this edge. This results in a non-directional fully connected network. For each node (postal code with a male state prison), we compute eigenvector centrality scores using the R *igraph* package.

Identifying large clusters of identical sequences shared between different male prison postal codes

We define large clusters of identical sequences in the prison networks as clusters of identical SARS-CoV-2 sequences (i) which are observed in at least two postal codes with male prisons and (ii) with at least 15 sequences in prison postal codes.

Age analyses

Evaluating the relationship between genetic and social contact data

We quantify the association between the relative risk $RR^0_{A,B}$ of observing identical sequences between two age groups A and B and the relative risk of contacts $RR^{contacts}_{A,B}$ between these two groups using a generalised additive model (GAM) on a logarithmic scale. We report the

percentage of variance in the RR of identical sequences explained by the RR of contact from the GAM. We also report the Spearman correlation coefficient between $RR^0_{A,B}$ and $RR^{contacts}_{A,B}$.

Investigating age-specific transmission patterns across spatial scales

To understand how age-specific transmission patterns vary across spatial scales, we compare the RR of observing identical sequences between age groups using all pairs of identical sequences, using only pairs of identical sequences in different postal codes and using only pairs of identical sequences in different counties.

Exploring typical transmission direction between age groups

We use sequence collection dates to explore transmission direction between age groups across 4 periods (March 2021 - June 2021, July 2021 - November 2021, December 2021 - February 2022 and March 2022 - August 2022). To facilitate the interpretation of these results, we introduce an earliness score that measures the contribution of a given age group to transmission to other age groups. For an age group A, this score is equal to the proportion of pairs of identical sequences first observed in age group A among all pairs of identical sequences observed in age groups A. We also report 95% binomial confidence interval around this score.

To explore whether our conclusions could be explained by differences in testing behaviors between age group, we conduct a sensitivity analysis by imputing the dates of symptom onset and using symptom onset dates instead of sequence collection ones (Figure S21). We also compute earliness scores on each of the 1,000 datasets with imputed symptom onset dates using the same definition as that based on sequence collection dates. We then report the median earliness score across all 1,000 datasets as well as an uncertainty range defined as the minimum lower bound and the maximum upper bound of the 95% binomial confidence interval around this score for each of the imputed dataset.

Relative risk of observing identical sequences between vaccination groups

Available matched patient information include details regarding the individuals' vaccination statuses upon positive test:

- No valid vaccination record (denoted Unvaccinated)
- Completed primary series (denoted Vaccinated)
- Completed primary series with an additional dose (denoted Boosted)

Here, we use this information to quantify mixing between groups characterized by their vaccination status. We focus on the mixing between vaccination groups within age groups to avoid biases coming from age group and vaccination status being correlated. Among sequences collected within each period (4 waves) and age groups in decade, we compute the RR of observing identical sequences between vaccination groups. We only include the Boosted vaccination group for wave 6 (Omicron BA.1 wave) for age groups older than 10, and from wave 7 for the 0-9y. We only include the 0-9y in our analysis from wave 6 (Omicron BA.1 wave) and the 10-19y from wave 5 (Delta wave).

To quantify the tendency of individuals of transmitting to individuals with the same vaccination status, we compute for each vaccination groups (V_1, V_2) the ratio $RR_{V_1,V_2}/RR_{V_1,V_1}$. Values lower than 1 indicate that the enrichment of pairs of identical sequences is greater within the same

vaccination group than between different vaccination groups. Such values suggest assortativity in mixing patterns between vaccination groups.

Deriving the distribution of the number of mutations conditional on the number of generations separating two infected individuals

In this section, we derive the probability distribution of the number of mutations M_{AB} separating the consensus genomes of two infected individuals A and B conditional on the number of transmission generations G_{AB} separating them.

Generation time distribution

We assume that the generation time (i.e. the average duration between the infection time of an infector and an infectee) follows a Gamma distribution of shape α and scale β. The time between g generations then follows a Gamma distribution of shape $g{\cdot}\alpha$ and scale β assuming independence of successive transmission events. Let $f_{\alpha,\beta}(\cdot)$ denote the probability density function of a Gamma distribution of shape α and scale β.

Mutations events

Let M_{AB} denote the number of mutations separating their infecting viruses. Let μ denote the mutation rate of the virus (in mutations per day). Let T_{AB}^{evo} denote the evolutionary time separating A and B (in days).

Assuming a Poisson process for the occurrence of mutations, we have:

$$M_{AB} \sim \mathcal{P}(\mu \cdot T_{AB}^{evo})$$

Distribution of the number of mutations conditional on the number of generations

Let G_{AB} denote the number of generations separating two infected individuals A and B belonging to the same transmission chain.

$$\begin{aligned}
P[M_{AB} = m \mid G_{AB} = g] &= \int_{t_{AB}^{evo}=0}^{+\infty} P[M_{AB} = m \mid G_{AB} = g, T_{AB}^{evo} = t_{AB}^{evo}] \cdot p(t_{AB}^{evo} \mid G_{AB} = g)\ \mathrm{dt}_{\mathrm{AB}}^{\mathrm{evo}} \\
&= \int_{t_{AB}^{evo}=0}^{+\infty} P[M_{AB} = m \mid T_{AB}^{evo} = t_{AB}^{evo}] \cdot f_{\alpha g,\beta}(t_{AB}^{evo})\ \mathrm{dt}_{\mathrm{AB}}^{\mathrm{evo}} \\
&= \int_{t_{AB}^{evo}=0}^{+\infty} \frac{(\mu\ t_{AB}^{evo})^m \cdot e^{-\mu t_{AB}^{evo}}}{m!} \cdot \frac{\beta^{\alpha g} \cdot (t_{AB}^{evo})^{\alpha g - 1} \cdot e^{-\beta t_{AB}^{evo}}}{\Gamma(\alpha g)}\ \mathrm{dt}_{\mathrm{AB}}^{\mathrm{evo}} \\
&= \frac{\mu^m\ \beta^{\alpha g}}{m!\ \Gamma(\alpha g)} \int_{t_{AB}^{evo}=0}^{+\infty} (t_{AB}^{evo})^{m+\alpha g-1} \cdot e^{-(\mu+\beta)\cdot t_{AB}^{evo}}\ \mathrm{dt}_{\mathrm{AB}}^{\mathrm{evo}} \\
&= \frac{\mu^m\ \beta^{\alpha g}\ \Gamma(m+\alpha g)}{m!\ \Gamma(\alpha g)\ (\mu+\beta)^{m+\alpha g}} \int_{t_{AB}^{evo}=0}^{+\infty} f_{m+\alpha g,\mu+\beta}(t_{AB}^{evo})\ \mathrm{dt}_{\mathrm{AB}}^{\mathrm{evo}} \\
&= \frac{\Gamma(m+\alpha g)}{m!\ \Gamma(\alpha g)} \cdot \Big(\frac{\beta}{\beta+\mu}\Big)^{\alpha g} \cdot \Big(\frac{\mu}{\beta+\mu}\Big)^{m}
\end{aligned}$$

which is the probability mass function of a negative binomial distribution of parameters:

$$r = \alpha g$$
$$p = \frac{\beta}{\beta + \mu}$$

Application to SARS-CoV-2

We compute these probabilities for SARS-CoV-2 considering a mutation rate $\mu = 8.98 \cdot 10^{-2}$ substitutions per day (32.76 substitutions per year) [46]. We assume that the generation time is Gamma distributed with a mean 5.9 days and standard deviation 4.8 days [47].

Simulation study

We conduct a simulation study to evaluate how our RR framework performs under different sequencing scenarios. We also compare the results obtained from a phylogeographic analysis.

Simulating synthetic sequence data

We use ReMASTER [48] to simulate an SEIR epidemic in a structured population with 5 demes, each populated with 100,000 inhabitants. We simulate an epidemic characterized by a basic reproduction number R_0 of 2 with a daily time-step. We consider a pathogen with a 3,000 kb genome evolving following Jukes-Cantor evolution model with a substitution rate of $3 \cdot 10^{-5}$ substitutions per site per day. Upon infection, infected individuals enter an Exposed (E) compartment during which they are not infectious yet and that they exit at a rate of 0.33 per day. They then enter an Infectious (I) compartment where they are infectious that they exit at a rate of 0.33 per day. Sequencing occurs upon exit of the I compartment. Given that our RR does not account for directionality in transmission, we considered a scenario with symmetric migration rates. We draw migration rates between demes from a log-Uniform distribution of parameters $(10^{-3}, 10^{-1})$.

We then explore two sequencing scenarios. In an unbiased scenario, we assume that each individual has the same probability of being sequenced in each deme. In a biased scenario, we assume that the sequencing probability varies between demes. We draw deme-specific relative sequencing probabilities from a log-Uniform distribution of parameters $(10^{-3}, 10^{-1})$. In the unbiased scenario, we fix sequencing probability to the mean of the sequencing probabilities across demes in the biased scenario. We explore different sequencing intensities by scaling these probabilities by different multiplicative factor (Table S1):

- a scaling factor of 0.1 resulting in a mean sequencing probability of 0.43% and a dataset of around 1700 sequences (used for the DTA analyses)
- a scaling factor of 0.5 resulting in a mean sequencing probability of 2.16% and a dataset of around 8600 sequences (used for the RR and the DTA analyses)
- a scaling factor of 2 resulting in a mean sequencing probability of 8.66% and a dataset of around 34,500 sequences (used for the RR analyses)

Discrete trait analysis

We conduct phylogeographic inference using symmetric discrete trait analysis (DTA) [7] using the Bayesian stochastic search variable selection (BSSVS) model implemented in BEAST 1.10.4 [49] applied to the synthetic data simulated in our two sequencing scenarios. In order to isolate

the accuracy and precision of the phylogeographic reconstruction, we run our discrete trait analysis using an empirical tree that is generated directly from ReMASTER simulations. This directly input tree is not possible in real-world scenarios where the genealogical tree must be (noisily) estimated from empirical sequence data. In this case, it serves a demonstration of the power of DTA when provided perfect geneaological signal. The empirical tree approach also requires substantially less computation and so allowed us to analyze datasets of thousands of sequences using DTA in acceptable time.

Two independent Markov chain Monte Carlo (MCMC) procedures are run for $2.5 \cdot 10^8$ iterations and sampled every 1,000 iterations. Resulting posterior distributions are combined after discarding initial 10% of sampled trees as burn-in from each of them. We use Tracer 1.7 [50] to assess convergence and to estimate effective sampling size (ESS), ensuring ESS values greater than 200 for each migration rate estimate. We adjust the estimated migration rates by the estimated rate scalar in order to calculate the per-day rates of transition between demes.

In practice, the genealogical tree has to be estimated from empirical sequence data. To evaluate how this may impact both results and computing times, we perform an additional phylogeographic analysis based on simulated but this time jointly inferring the genealogical tree and migration rates. We run this model for 24 days (corresponding to 475,733,000 MCMC steps), until each migration rate parameter has an ESS greater than 200.

Relative risk analysis

We compute the RR of observing identical sequences between two demes i and j and compare these RR to the daily probability $p_{i,j}$ of migration between these two demes which is computed as:

$$p_{i,j} = 1 - exp(-m_{i,j}) \quad j \neq i$$
$$p_{i,i} = 1 - \sum_{k \neq i} p_{i,k}$$

where $m_{i,j}$ is the migration rate between demes i and j. We generate 95% subsampling confidence intervals around the RRs using an 80% subsampling rate.

Downsampling approach to explore the genome dataset size required to compute relative risks

We implement a downsampling strategy to understand the amount of sequencing data required to compute RR estimates. We consider genome datasets of the following sizes:

$$\{1 \cdot 10^2, 2 \cdot 10^2, 3 \cdot 10^2, 4 \cdot 10^2, 5 \cdot 10^2, 6 \cdot 10^2, 7 \cdot 10^2, 8 \cdot 10^2, 9 \cdot 10^2,$$
$$1 \cdot 10^3, 2 \cdot 10^3, 3 \cdot 10^3, 4 \cdot 10^3, 5 \cdot 10^3, 6 \cdot 10^3, 7 \cdot 10^3, 8 \cdot 10^3, 9 \cdot 10^3,$$
$$1 \cdot 10^4, 2 \cdot 10^4, 3 \cdot 10^4, 4 \cdot 10^4, 5 \cdot 10^4, 6 \cdot 10^4, 7 \cdot 10^4, 8 \cdot 10^4, 9 \cdot 10^4, 1 \cdot 10^5\}$$

For each of these dataset sizes, we generated 100 downsampled datasets from our WA sequencing data. For each of these downsampled datasets, we compute the RR of observing identical sequences between age groups (Figure S28). To understand how the number of groups studied impacts the amount of data required, we also compute RR of observing identical sequences between aggregated age groups:

- 0-39y and over 40y for 2 age groups,

- 0-29y, 30-59y and over 60y for 3 age groups,
- 0-19y, 20-39y, 40-59y and over 60y for 4 age groups,
- 0-9y, 10-19y, 20-29y, 30-39y, 40-49y, 50-59y, 60-69y, 70-79y and over 80 (standard definition used throughout the paper) for 9 age groups.

We compute the error between the RR obtained from a subsampled dataset RR^d and the RR from the full dataset RR^f as:

$$\epsilon = \frac{RR^d - RR^f}{RR^f}$$

For each pair of age groups, we then compute the number of pairs of identical sequences required for the error to be below 10%.

Data and code accessibility

Code to reproduce our analyses can be found at `https://github.com/blab/phylo-kernel-public` [44]. All sequences referenced in this manuscript are publicly shared to GISAID and are publicly available with standard metadata (generally consisting of date of sample collection and sometimes county of sample collection). More detailed metadata curated by Washington State Department of Health (WA DOH) of county, postal code, age group and vaccination status were shared with the Fred Hutchinson Cancer Center under a Data Sharing Agreement for Confidential Data with an associated IRB. These more detailed metadata are not currently shared publicly while we seek clearance with WA DOH. GISAID accessions and a sequence-level acknowledgements table are provided in the GitHub repository associated with this manuscript [44].

Author contributions

CTK and TB conceived the study. CTK developed the methods. CTK, ACP and MIP performed the analyses with input from CV and TB. LF, HX, KK, AW, ALG, PR, JMP, AD, HH, DM, PD, LG, CDF, ER, JS, DR, AT, CY, FA and AB collected and curated the data. CTK, AP, MP and TB wrote the manuscript. All authors reviewed and edited the manuscript.

Acknowledgments

We would like to thank Sixtine Gurrey, Lara B. Strick and Michael A. Santoro for insightful discussions on the control of SARS-CoV-2 in WA Correction Centers. We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based.

Ethics approval

The Washington State and University of Washington Institutional Review Boards determined this project to be surveillance activity and exempt from review; the need for informed consent was waived through this determination. Under Washington State IRB Exempt Determination 2020-102, symptom onset date, age group, residence county, residence postal code and vaccination history was provided by the Washington Department of Health from the Washington Disease Reporting System for individuals with linked sequenced SARS-CoV-2 samples from March 1, 2021 through December 31, 2022. Sequencing and analysis of samples from the Seattle Flu Study

was approved by the Institutional Review Board (IRB) at the University of Washington (protocol STUDY00006181). Sequencing of remnant clinical specimens at UW Virology Lab was approved by the University of Washington Institutional Review Board (protocol STUDY00000408).

Funding

This work is supported by NIH NIGMS (R35 GM119774 to T.B.) and the CDC (CDC-RFA-CK22-2204, Pathogens Genomics Centers of Excellence, contract NU50CK000630). Sequencing of specimens by the Brotman Baty Institute of Precision Medicine was funded by Gates Ventures (Seattle Flu Study award), Howard Hughes Medical Institute (HHMI COVID-19 Collaboration Initiative award) and the CDC (contract number 200-2021-10982). Sequencing of specimens by UW Virology was funded by Fast Grants (award 2244), the CDC (contracts 75D30121C10540 and 75D30122C13720) and WA DOH (contract HED26002). Some of the analyses were completed using Fred Hutch Scientific Computing resources (NIH grants S10-OD-020069 and S10-OD-028685). TB is a Howard Hughes Medical Institute Investigator.

Competing interests

ALG reports contract testing from Abbott, Cepheid, Novavax, Pfizer, Janssen and Hologic, research support from Gilead, and salary and stock grants for LabCorp an immediate family member, outside of the described work. All other authors declare no competing interests.

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the US National Institutes of Health or Department of Health and Human Services.

References

[1] Russell CA, Jones TC, Barr IG, Cox NJ, Garten RJ, et al. (2008) The global circulation of seasonal influenza a (H3N2) viruses. Science 320: 340–346.

[2] Bedford T, Riley S, Barr IG, Broor S, Chadha M, et al. (2015) Global circulation patterns of seasonal influenza viruses vary with antigenic drift. Nature 523: 217–220.

[3] Moncla LH, Black A, DeBolt C, Lang M, Graff NR, et al. (2021) Repeated introductions and intensive community transmission fueled a mumps virus outbreak in washington state. Elife 10.

[4] Layan M, Müller NF, Dellicour S, De Maio N, Bourhy H, et al. (2023) Impact and mitigation of sampling bias to determine viral spread: Evaluating discrete phylogeography through CTMC modeling and structured coalescent model approximations. Virus Evol 9: vead010.

[5] Brito AF, Semenova E, Dudas G, Hassler GW, Kalinich CC, et al. (2022) Global disparities in SARS-CoV-2 genomic surveillance. Nat Commun 13: 7003.

[6] Tran-Kiem C, Bedford T (2023) Estimating the reproduction number and transmission heterogeneity from the size distribution of clusters of identical pathogen sequences .

[7] Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009) Bayesian phylogeography finds its roots. PLoS Comput Biol 5: e1000520.

[8] Real LA, Biek R (2007) Spatial dynamics and genetics of infectious diseases on heterogeneous landscapes. J R Soc Interface 4: 935–948.

[9] Smith DL, Lucey B, Waller LA, Childs JE, Real LA (2002) Predicting the spatial dynamics of rabies epidemics on heterogeneous landscapes. Proc Natl Acad Sci U S A 99: 3668–3672.

[10] Fountain-Jones NM, Kraberger S, Gagne RB, Trumbo DR, Salerno PE, et al. (2021) Host relatedness and landscape connectivity shape pathogen spread in the puma, a large secretive carnivore. Commun Biol 4: 12.

[11] Laenen L, Vergote V, Vanmechelen B, Tersago K, Baele G, et al. (2019) Identifying the patterns and drivers of puumala hantavirus enzootic dynamics using reservoir sampling. Virus Evol 5: vez009.

[12] Tordoff DM, Greninger AL, Roychoudhury P, Shrestha L, Xie H, et al. (2021) Phylogenetic estimates of SARS-CoV-2 introductions into washington state. Lancet Reg Health Am 1: 100018.

[13] United States Census Bureau. Commuting flows. `https://www.census.gov/topics/employment/commuting/guidance/flows.html`.

[14] Pullano G, Alvarez-Zuzek LG, Colizza V, Bansal S (2023) Characterizing US spatial connectivity: implications for geographical disease dynamics and metapopulation modeling. bioRxiv .

[15] Hardyman PL, Austin J, Peyton J (2004) Prisoner intake systems: Assessing needs and classifying prisoners. Technical report, National Institute of Corrections (NIC) (Washington DC).

[16] Marquez N, Ward JA, Parish K, Saloner B, Dolovich S (2021) COVID-19 incidence and mortality in federal and state prisons compared with the US population, april 5, 2020, to april 3, 2021. JAMA 326: 1865–1867.

[17] EMG Transmission Group (2021) COVID-19 transmission in prison settings. Technical report.

[18] Viboud C, Bjørnstad ON, Smith DL, Simonsen L, Miller MA, et al. (2006) Synchrony, waves, and spatial hierarchies in the spread of influenza. Science 312: 447–451.

[19] Cauchemez S, Carrat F, Viboud C, Valleron AJ, Boëlle PY (2004) A bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. Stat Med 23: 3469–3487.

[20] Rader B, Scarpino SV, Nande A, Hill AL, Adlam B, et al. (2020) Crowding and the shape of COVID-19 epidemics. Nat Med 26: 1829–1834.

[21] Tran Kiem C, Bosetti P, Paireau J, Crépey P, Salje H, et al. (2021) SARS-CoV-2 transmission across age groups in france and implications for control. Nat Commun 12: 6895.

[22] Haw DJ, Cummings DAT, Lessler J, Salje H, Read JM, et al. (2019) Differential mobility and local variation in infection attack rate. PLoS Comput Biol 15: e1006600.

[23] Apolloni A, Poletto C, Colizza V (2013) Age-specific contacts and travel patterns in the spatial spread of 2009 H1N1 influenza pandemic. BMC Infect Dis 13: 176.

[24] Gog JR, Ballesteros S, Viboud C, Simonsen L, Bjornstad ON, et al. (2014) Spatial transmission of 2009 pandemic influenza in the US. PLoS Comput Biol 10: e1003635.

[25] Mistry D, Litvinova M, Pastore Y Piontti A, Chinazzi M, Fumanelli L, et al. (2021) Inferring high-resolution human mixing patterns for disease modeling. Nat Commun 12: 323.

[26] Prem K, Cook AR, Jit M (2017) Projecting social contact matrices in 152 countries using contact surveys and demographic data. PLoS Comput Biol 13: e1005697.

[27] Mossong J, Hens N, Jit M, Beutels P, Auranen K, et al. (2008) Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med 5: e74.

[28] Hoang T, Coletti P, Melegaro A, Wallinga J, Grijalva CG, et al. (2019) A systematic review of social contact surveys to inform transmission models of close-contact infections. Epidemiology 30: 723–736.

[29] Read JM, Lessler J, Riley S, Wang S, Tan LJ, et al. (2014) Social mixing patterns in rural and urban areas of southern china. Proc Biol Sci 281: 20140268.

[30] Goldstein E, Lipsitch M, Cevik M (2021) On the effect of age on the transmission of SARS-CoV-2 in households, schools, and the community. J Infect Dis 223: 362–369.

[31] Sun K, Tempia S, Kleynhans J, von Gottberg A, McMorrow ML, et al. (2023) Rapidly shifting immunologic landscape and severity of SARS-CoV-2 in the omicron era in south africa. Nat Commun 14: 246.

[32] Monod M, Blenkinsop A, Xi X, Hebert D, Bershan S, et al. (2021) Age groups that sustain resurging COVID-19 epidemics in the united states. Science 371.

[33] Bosetti P, Poletti P, Stella M, Lepri B, Merler S, et al. (2020) Heterogeneity in social and epidemiological factors determines the risk of measles outbreaks. Proc Natl Acad Sci U S A 117: 30118–30125.

[34] Salathé M, Bonhoeffer S (2008) The effect of opinion clustering on disease outbreaks. J R Soc Interface 5: 1505–1508.

[35] Oltean HN, Allen KJ, Frisbie L, Lunn SM, Torres LM, et al. (2023) Sentinel surveillance system implementation and evaluation for SARS-CoV-2 genomic data, washington, USA, 2020-2021. Emerg Infect Dis 29: 242–251.

[36] Shu Y, McCauley J (2017) GISAID: Global initiative on sharing all influenza data – from vision to reality. Euro Surveill 22.

[37] GISAID - gisaid.org. `https://gisaid.org/`. Accessed: 2023-2-9.

[38] ncov-ingest: A pipeline that ingests SARS-CoV-2 (i.e. nCoV) data from GISAID and genbank, transforms it, stores it on S3, and triggers nextstrain nCoV rebuilds. `https://github.com/nextstrain/ncov-ingest`. Accessed: 2023-2-9.

[39] Department of Corrections, WA. Prison facilities in WA. `https://www.doc.wa.gov/corrections/incarceration/default.htm`. Accessed: 2024-2-20.

[40] Paradis E, Schliep K (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35: 526–528.

[41] Aksamentov I, Roemer C, Hodcroft E, Neher R (2021) Nextclade: clade assignment, mutation calling and quality control for viral genomes. J Open Source Softw 6: 3773.

[42] Tonkin-Hill G. pairsnp: A set of scripts for calculating pairwise SNP distance.

[43] Perofsky AC, Hansen CL, Burstein R, Boyle S, Prentice R, et al. (2024) Impacts of human mobility on the citywide transmission dynamics of 18 respiratory viruses in pre- and post-COVID-19 pandemic years. Nat Commun 15: 4164.

[44] Tran-Kiem C. Code and data for: "fine-scale spatial and social patterns of SARS-CoV-2 transmission from identical pathogen sequences". `https://github.com/blab/phylo-kernel-public`.

[45] Oksanen J (2010) vegan : Community ecology package. http://CRANR-projectorg/package=vegan .

[46] nextstrainorg. Genomic epidemiology of SARS-CoV-2 with subsampling focused globally over the past 6 months. `https://nextstrain.org/ncov/gisaid/global/6m?l=clock`. Accessed: 2024-3-28.

[47] Hart WS, Abbott S, Endo A, Hellewell J, Miller E, et al. (2022) Inference of the SARS-CoV-2 generation time using UK household data. Elife 11.

[48] Vaughan TG (2024) ReMASTER: improved phylodynamic simulation for BEAST 2.7. Bioinformatics 40.

[49] Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, et al. (2018) Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol 4: vey016.

[50] Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA (2018) Posterior summarization in bayesian phylogenetics using tracer 1.7. Syst Biol 67: 901–904.

Supplementary materials for:

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

List of supplementary figures

S1 Permutation test to explore the spatio-temporal signal in clusters of identical SARS-CoV-2 sequences. . . . 31
S2 The magnitude of the relative risk of observing sequences at a given genetic distance within the same county is impacted by transmission intensity. . . . 32
S3 Our measure of relative risk corrects for uneven sequencing between regions. . . . 33
S4 Simulation study exploring the impact of sequencing bias on results from a discrete trait analysis and from our RR framework. . . . 34
S5 Relative risk of observing identical sequences in two counties. . . . 35
S6 Relative risk of observing pairs of identical sequences between counties. . . . 37
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. 38
S8 Distribution of the delay between sequence collection within pairs of identical sequences collected between Eastern and Western WA and across epidemic waves. 39
S9 Comparison between the number of directed commuting flows and the number of directed visits between two counties. . . . 39
S10 Comparison between the relative risk of observing identical sequences and the relative risk of movement at the county level. . . . 40
S11 Geographical regions used to aggregate counties. . . . 41
S12 Comparison between the relative risk of observing identical sequences and the relative risk of movement at the region level. . . . 42
S13 Relationship between the relative risk of observing identical sequences in two regions and the relative risk of movement between these regionsobtained from mobile phone mobility data across epidemic waves . . . 43
S14 Comparison of the total number of visits between pairs of WA counties across the 4 epidemic waves during our study period. . . . 44
S15 Comparison of connectivity metrics across the Eastern / Western WA border among counties located on the border. . . . 45
S16 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. . . . 46
S17 Relative risk for pairs of identical sequences of being observed between two age groups. . . . 47
S18 Relative risk for pairs sequences of being observed between two age groups depending on their genetic distance. . . . 48
S19 Impact of the spatial scale on the relative risk for pairs sequences of being observed between two age groups. . . . 49
S20 Median delay between the dates of sequence collection within pairs of identical sequences . . . 50
S21 Sensitivity analysis on the timing of pairs identical sequences between age groups using symptom onset dates . . . 51

S22 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). 51
S23 Impact of non-sampled locations on the computation of the RR. 52
S24 Impact of dataset size on the number of clusters of identical sequences and the number of sequences with another identical sequence in the dataset. 52
S25 Impact of the number of groups included in the analysis on the dataset size required for the error relative risk of observing identical sequences to be lower than 10%. 53
S26 Empirical distribution of the delay between symptom onset and sequence collection by age (rows), period (columns) and geographic region (colours). 54
S27 Time-series of COVID-19 cases in WA over the study periods. 55
S28 Illustration of the downsampling strategy used to quantify the amount of data required to compute relative risks. 55

List of supplementary tables

S1 Performance of Discrete Trait Analysis (DTA) and our relative risk metric (RR) in quantifying migration patterns. 56
S2 Comparison of the relative of risk of observing identical sequences at the ZCTA level by adjacency level. 56
S3 Comparison between the relative risk of observing identical sequences between two geographic regions and the risk of movement between different geographies. . 57
S4 Characteristics of WA male prisons. 57

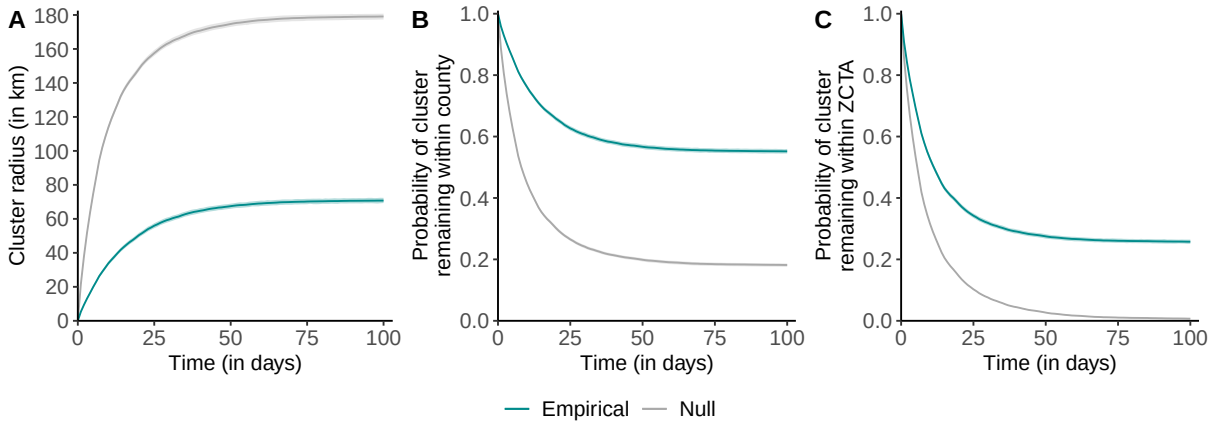


Figure S1. Permutation test to explore the spatio-temporal signal in clusters of identical SARS-CoV-2 sequences. A. Radius of clusters of identical sequences as a function of time since first sequence collection. **B.** Probability for all sequences within a cluster of identical sequences of remaining 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.

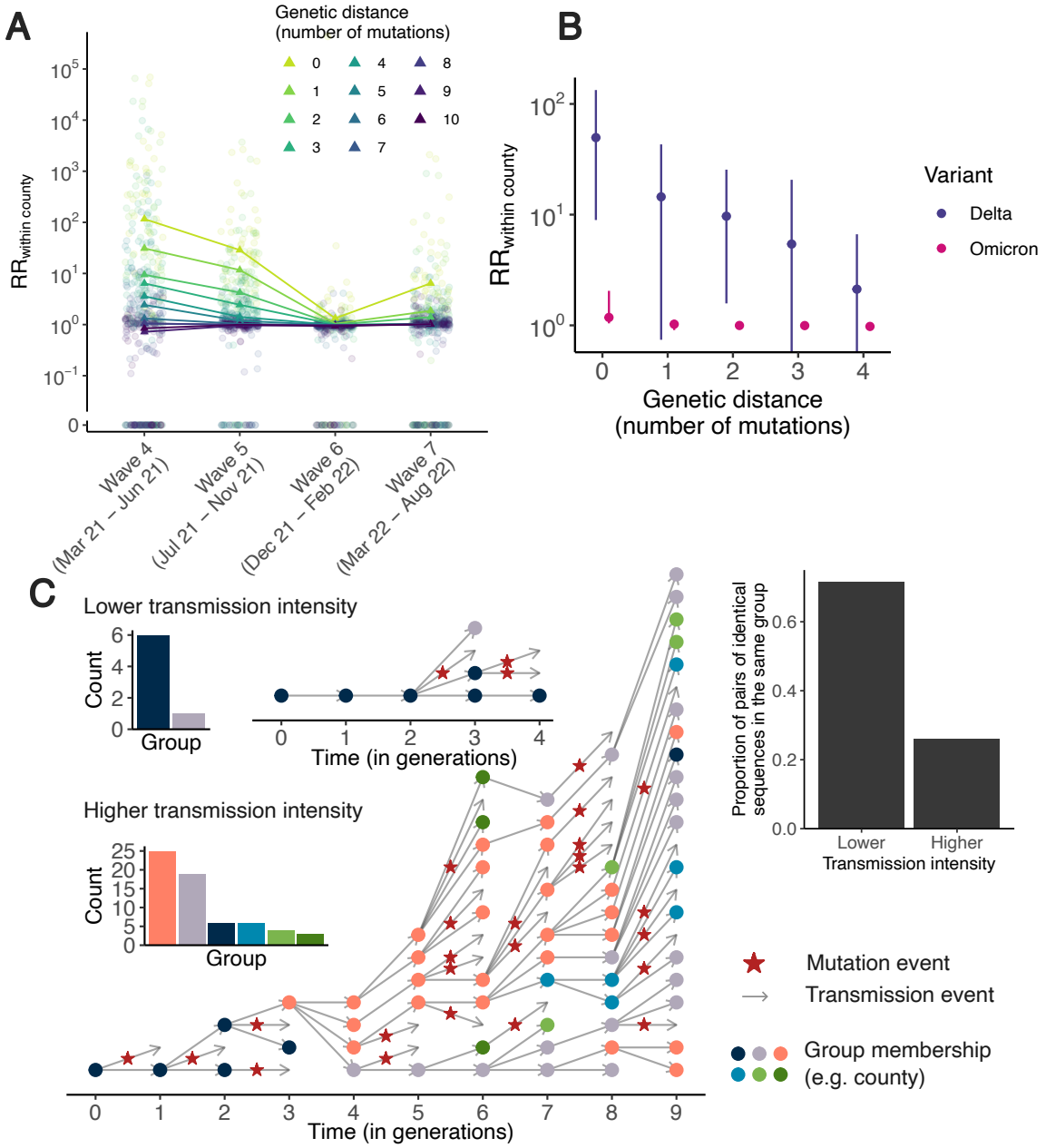


Figure S2. The magnitude of the relative risk of observing sequences at a given genetic distance within the same county is impacted by transmission intensity. A. Relative risk of observing sequences at a given genetic distance within the same county across multiple epidemic waves. We defined waves as: 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 [6] 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.

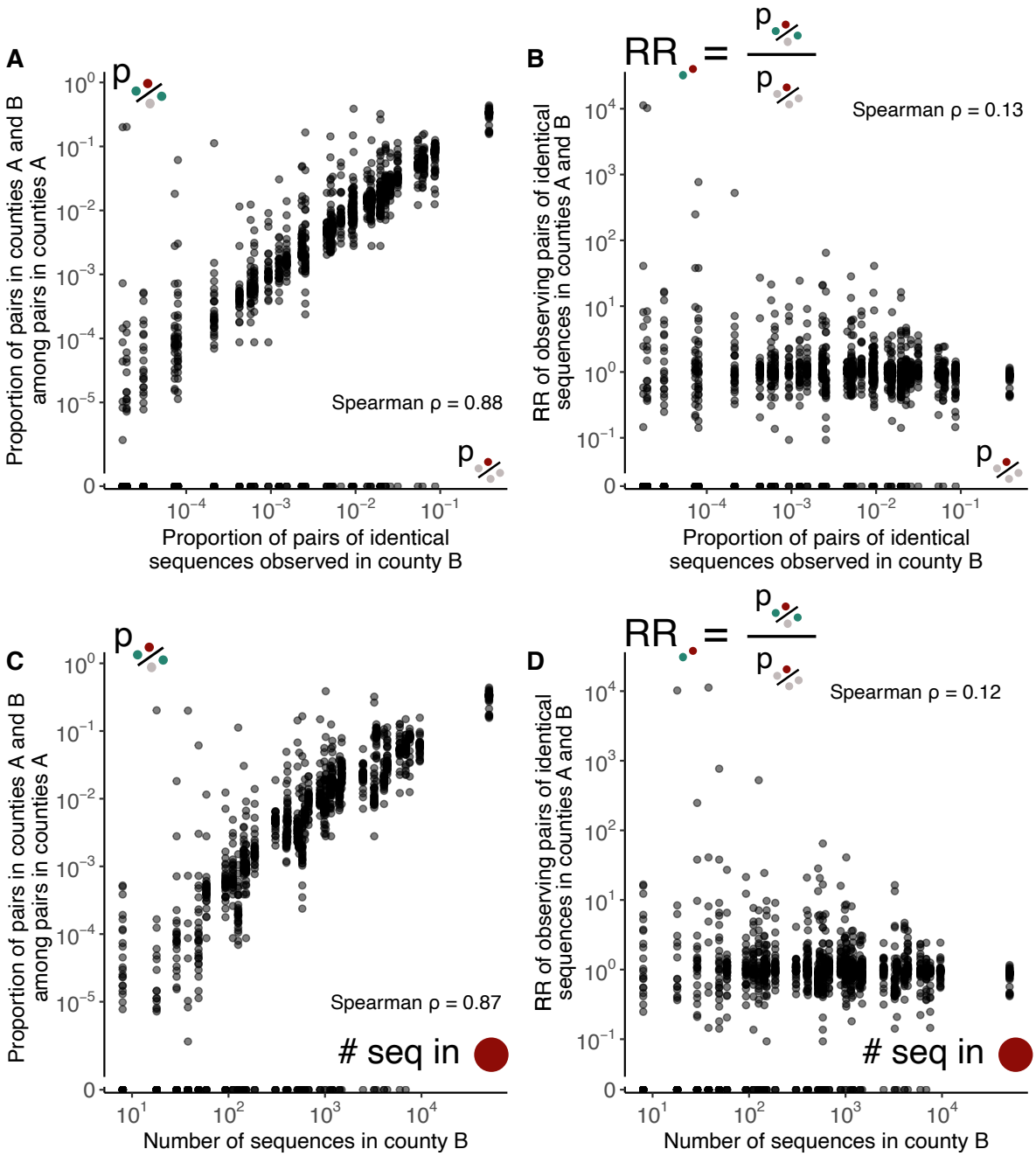


Figure S3. Our measure of relative risk corrects for uneven sequencing between regions. **A.** Proportion of pairs of identical sequences shared between counties A and B among pairs observed in county A as a function of the proportion of pairs of identical sequences observed in county B. **B.** Relative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

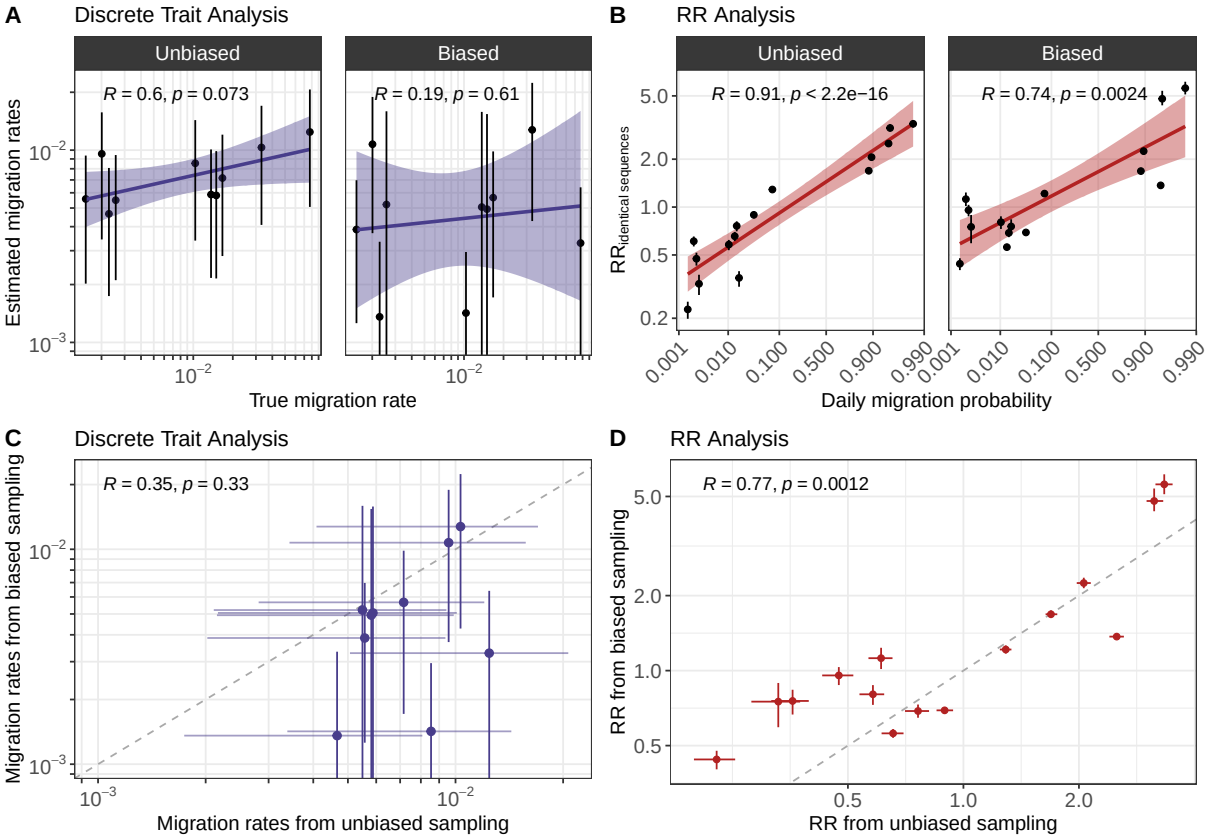


Figure S4. Simulation study exploring the impact of sequencing bias on results from a discrete trait analysis and from our RR framework. **A.** Comparison between migration rates estimated from a discrete trait analysis and the true migration rates used to simulate the sequence data. **B.** Comparison 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).

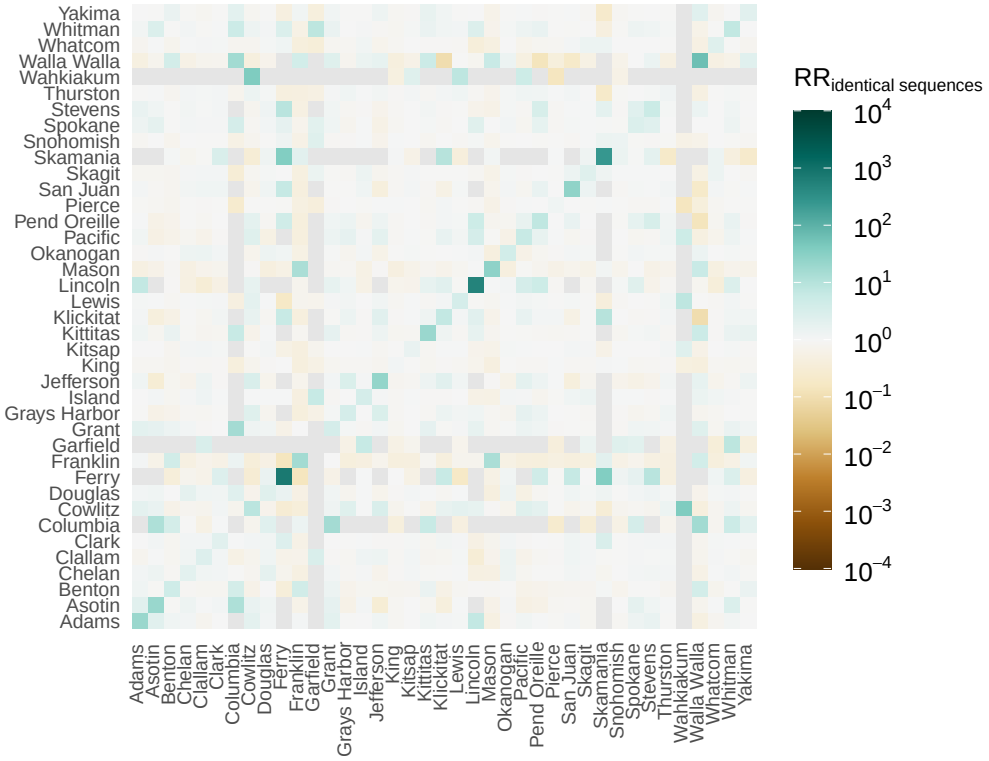
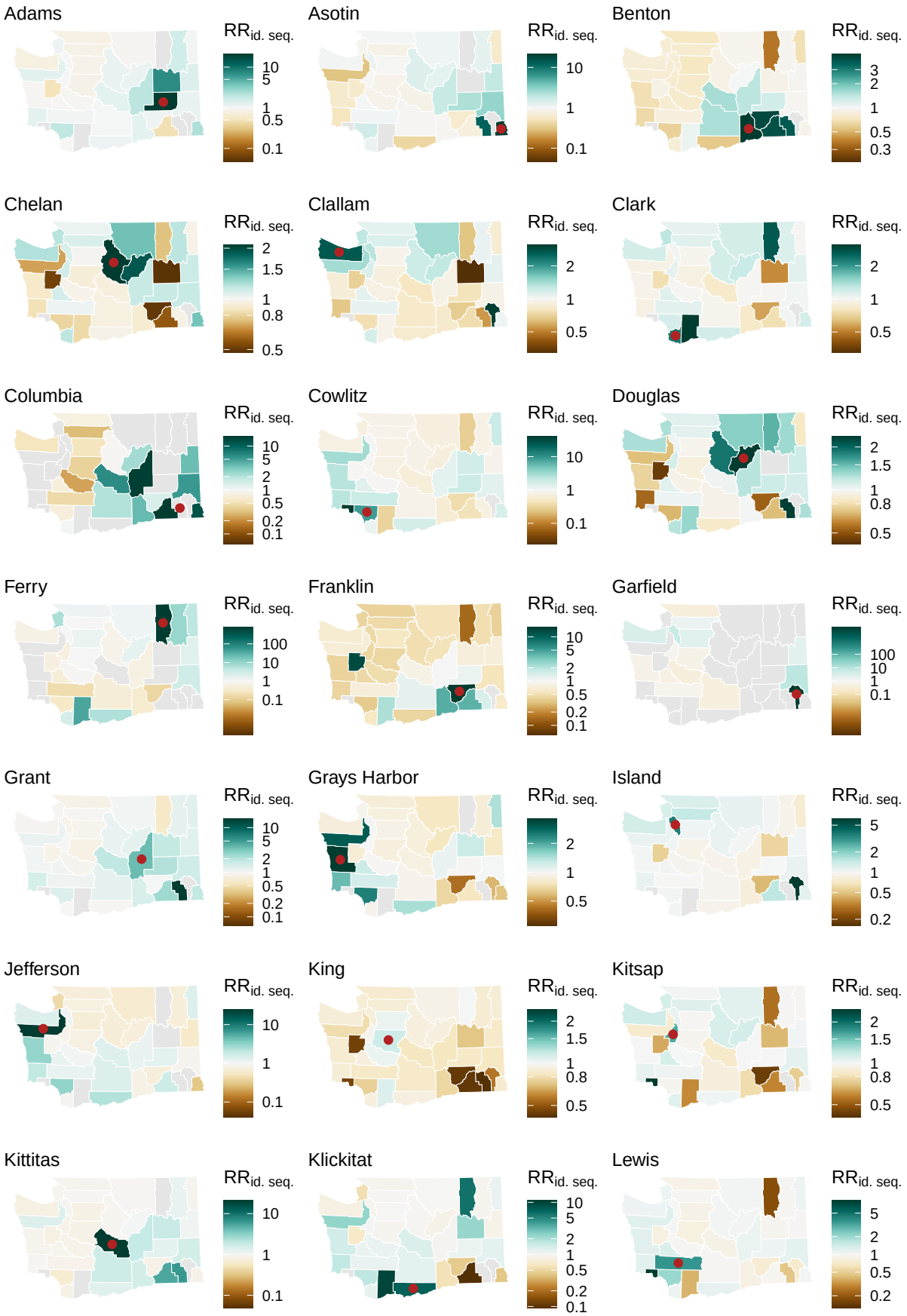


Figure S5. Relative risk of observing identical sequences in two counties. Grey squares correspond to pairs of counties between which no pairs of identical sequences were observed during the study period.


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

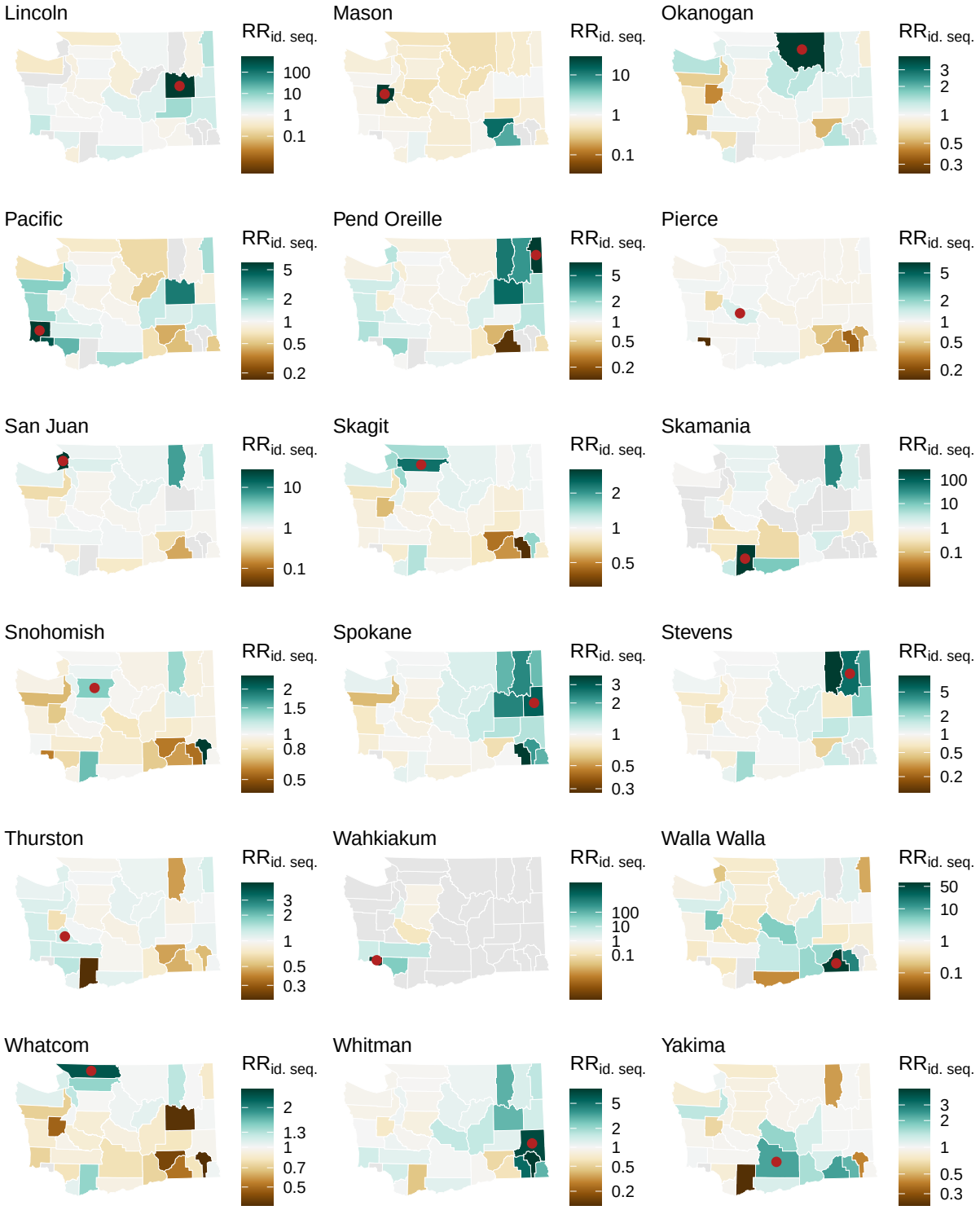


Figure S6. Relative risk of observing pairs of identical sequences between counties. On each map, we represent the relative risk of observing pairs of identical sequences in the county indicated by a red point (map title) and all the other counties in Washington state. Areas are colored in grey when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

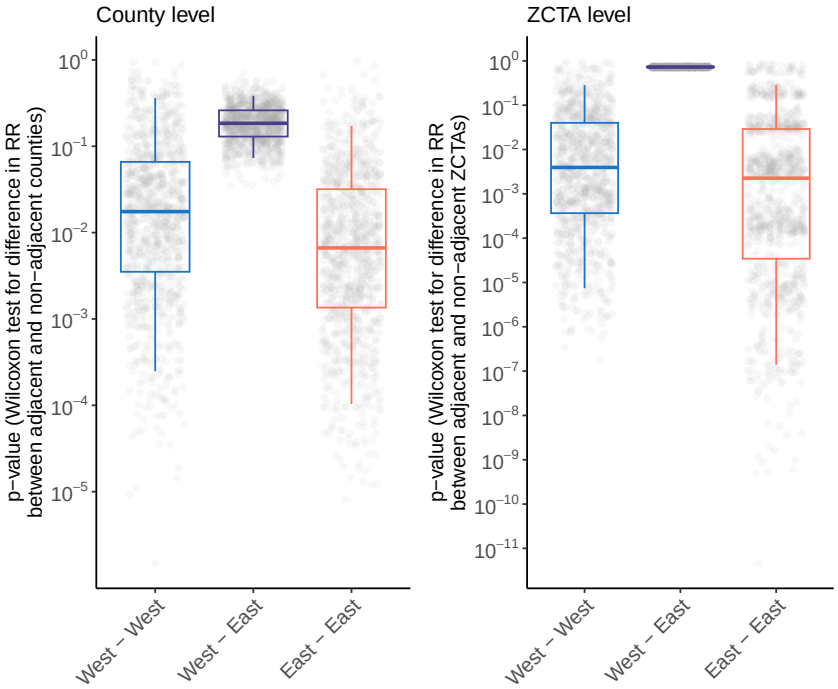


Figure S7. Impact of subsampling on the significance of the association between the relative risk of identical sequences and whether counties and ZCTAs are adjacent or not. We investigate whether our conclusions regarding the significance of the association between the relative risk of identical sequences 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.

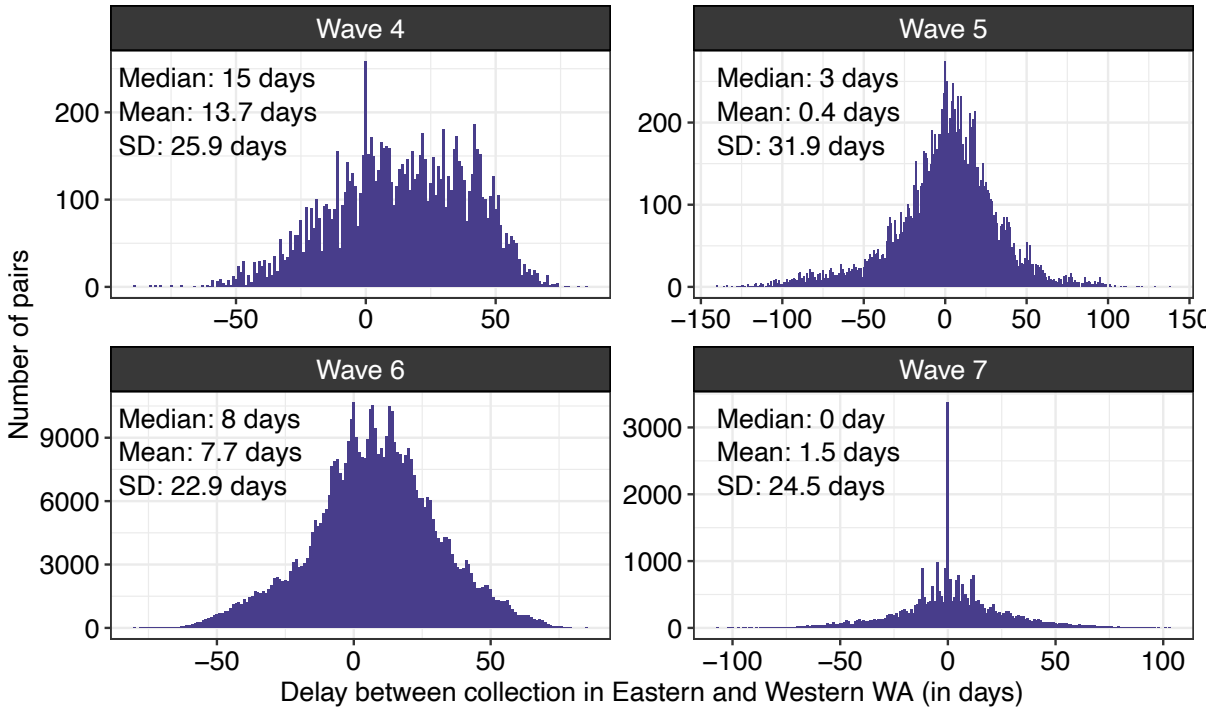


Figure S8. Distribution of the delay between sequence collection within pairs of identical sequences collected between Eastern and Western WA and across epidemic waves.

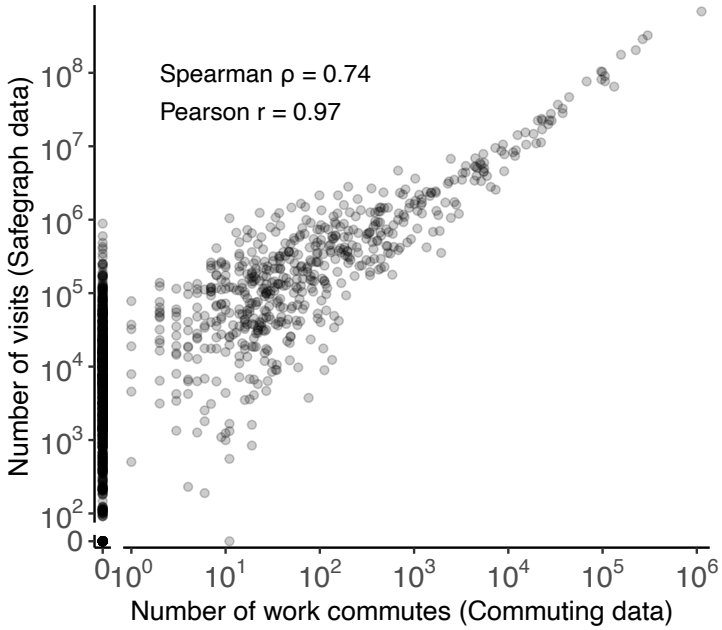


Figure S9. Comparison between the number of directed commuting flows and the number of directed visits between two counties. The number of work commutes is extracted from [13]. The number of visits is estimated using Safegraph *Weekly patterns* mobility data. The comparison is done by matching the origin 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.

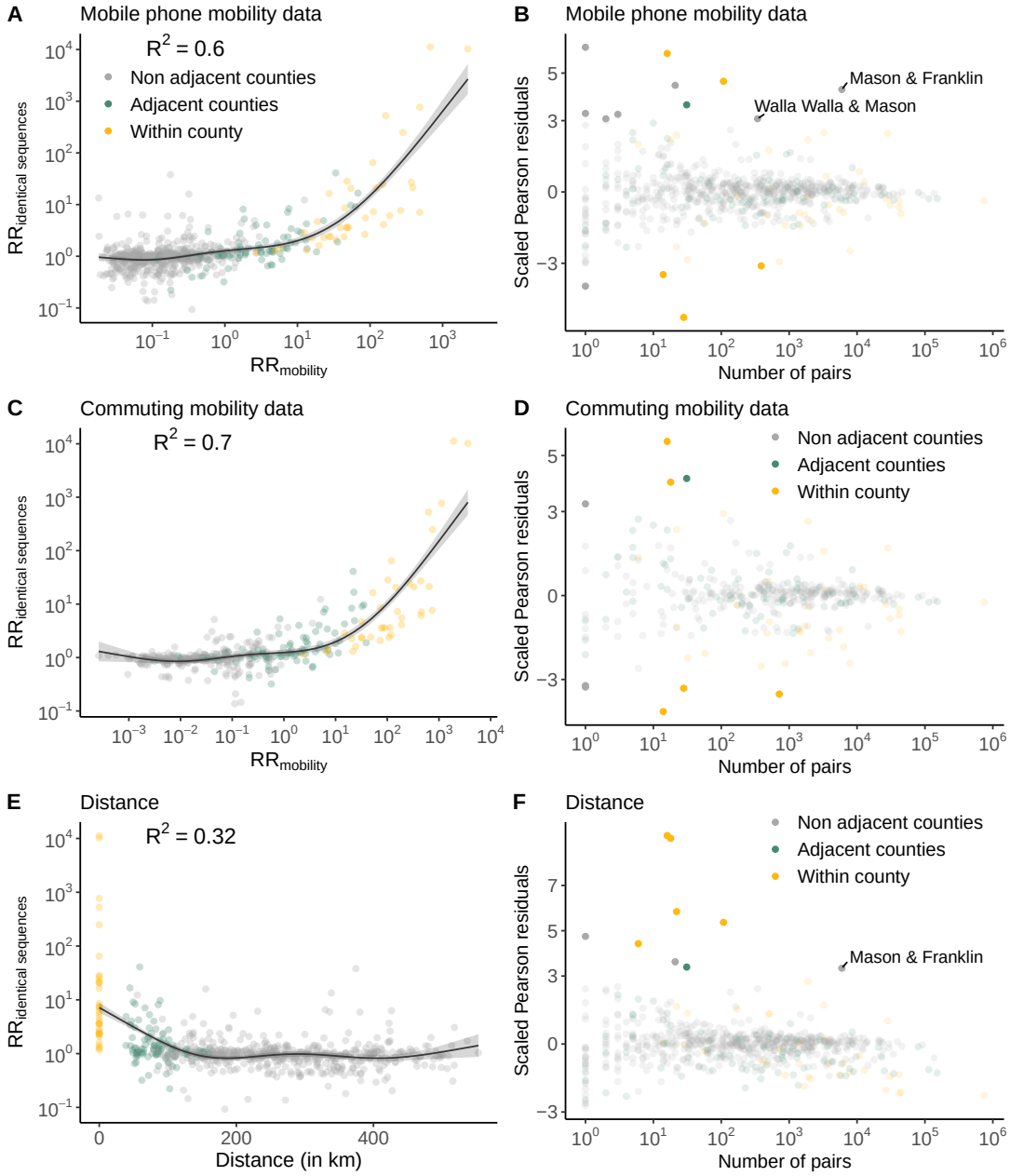


Figure S10. Comparison between the relative risk of observing identical sequences and the relative risk of movement at the county level. **A.** Relationship between the relative risk of observing identical sequences in two counties and the relative risk of movement between these counties as obtained 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.

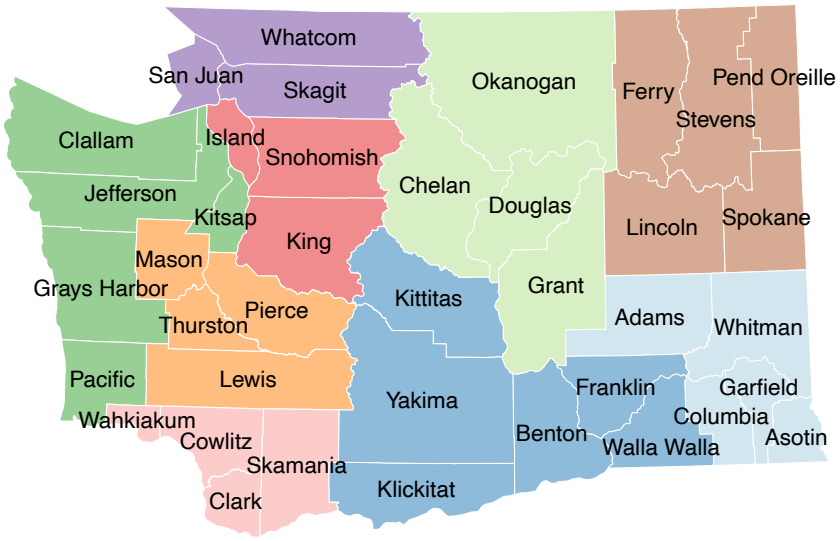


Figure S11. Geographical regions used to aggregate counties. Dark green: Peninsula/Coastal. Orange: South Puget Sound. Purple: Northwest. Red: North Puget Sound. Pink: Southwest. Dark blue: South Central. Light blue: Southeast.Brown: Northeast. Light green: North Central.

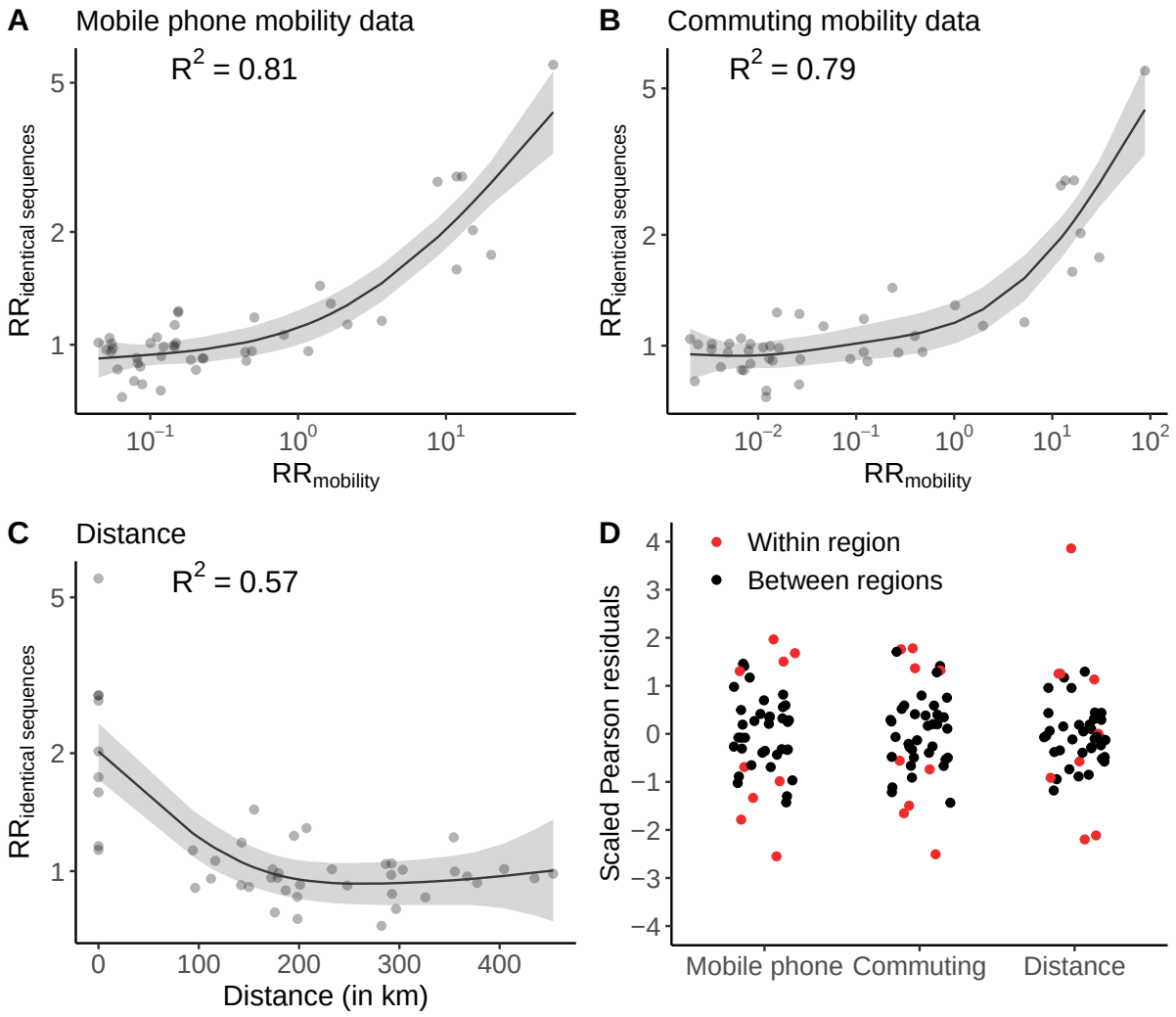


Figure S12. Comparison between the relative risk of observing identical sequences and the relative risk of movement at the region level. **A.** Relationship between the relative risk of observing identical sequences in two regions and the relative risk of movement between these regions as obtained from 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.

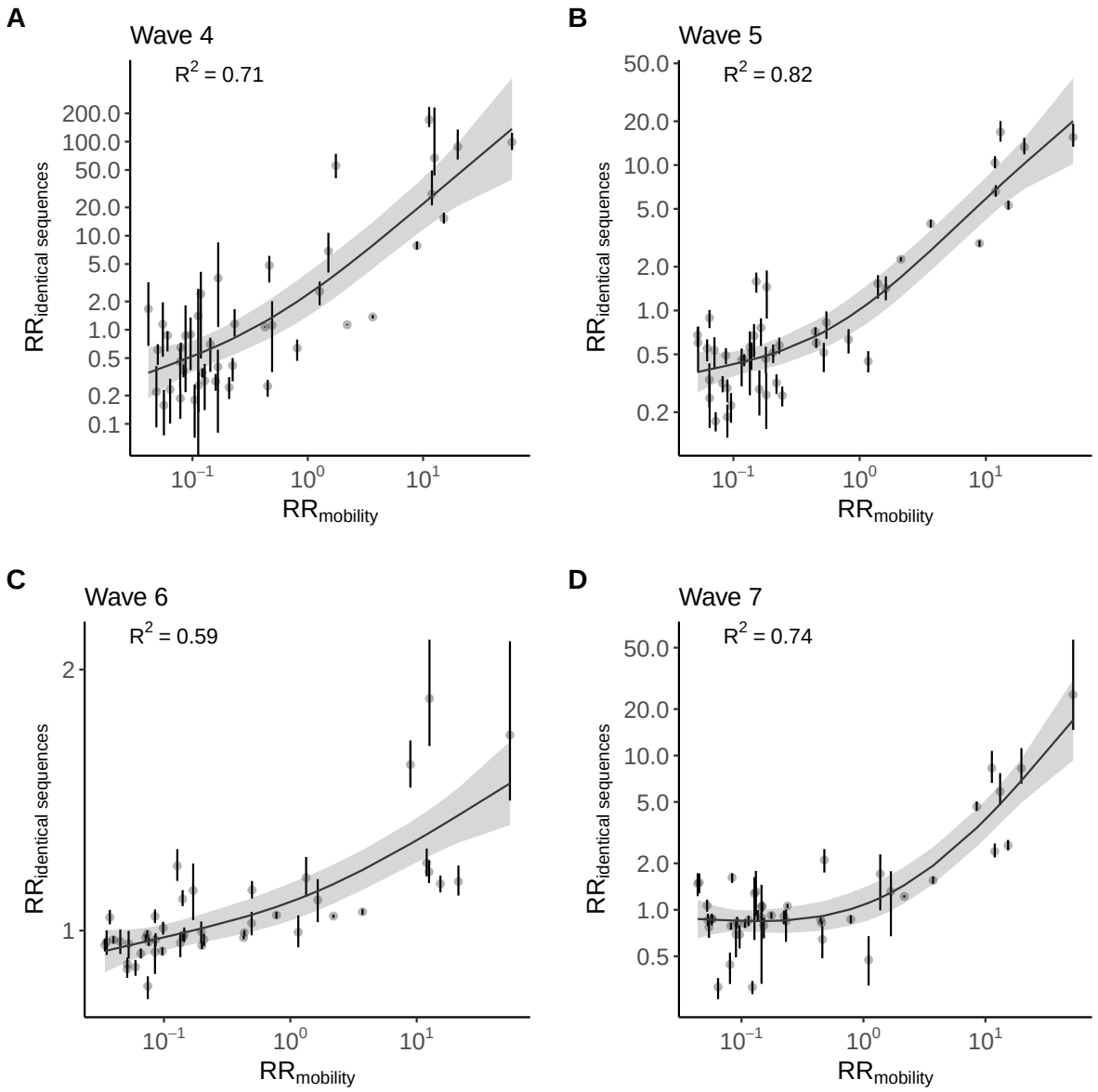


Figure S13. Relationship between the relative risk of observing identical sequences in two regions and the relative risk of movement between these regionsobtained from mobile phone mobility data across epidemic waves **A.** Wave 4. **B.** Wave 5. **C.** Wave 6. **D.** Wave 7. Vertical segments indicate 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.

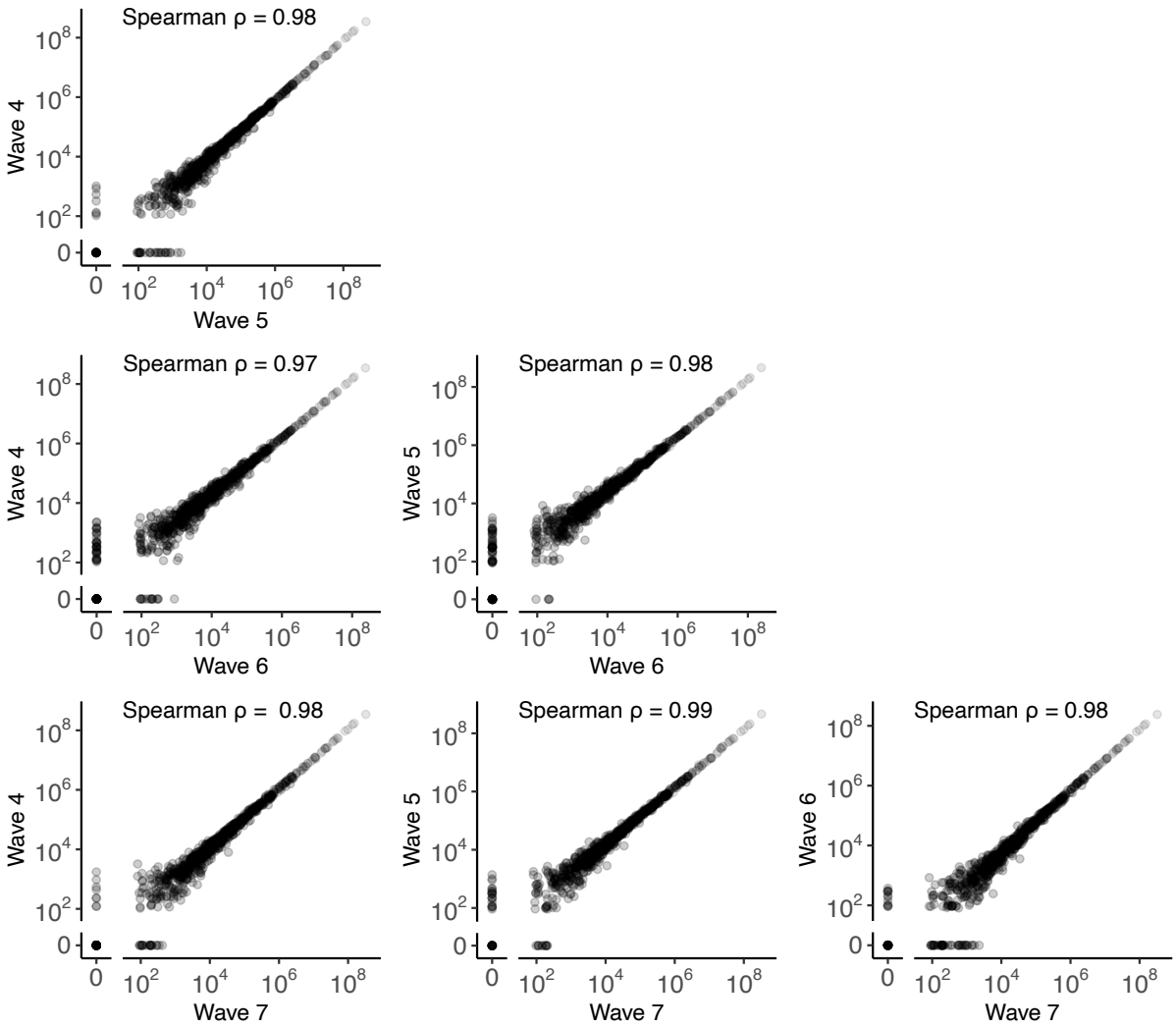


Figure S14. Comparison of the total number of visits between pairs of WA counties across the 4 epidemic waves during our study period. Points indicate the total number of visits between pairs of counties over the study periods labeled on the plot axes.

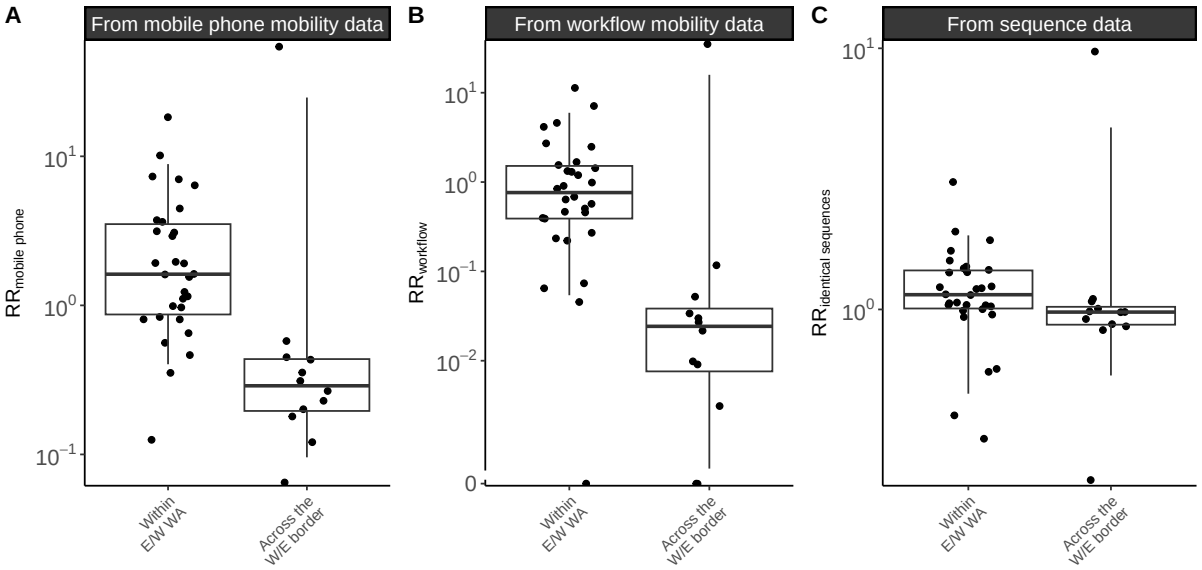


Figure S15. Comparison of connectivity metrics across the Eastern / Western WA border among counties located on the border. **A.** Relative risk of movement from mobile phone data across the border or within Eastern / Western WA (p-value for Wilcoxon rank sum test of $6.1 \cdot 10^{-5}$). **B.** Relative 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.

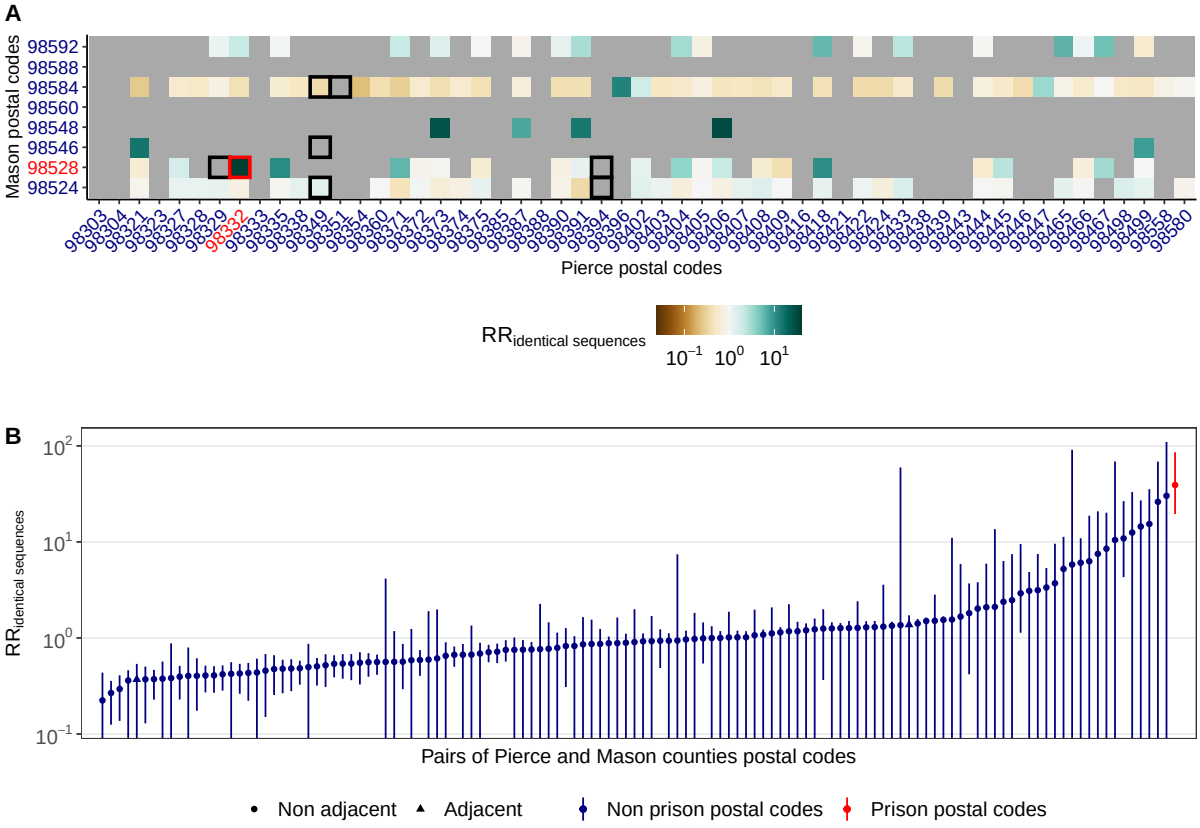


Figure S16. Patterns of occurrence of pairs of identical sequences between ZCTAs in Pierce and Mason counties, the two counties that are home of WA female prisons. A. Relative risk of observing identical sequences between ZCTAs in Mason and Pierce counties. Black squares indicate adjacent ZCTAs. ZCTAs 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.

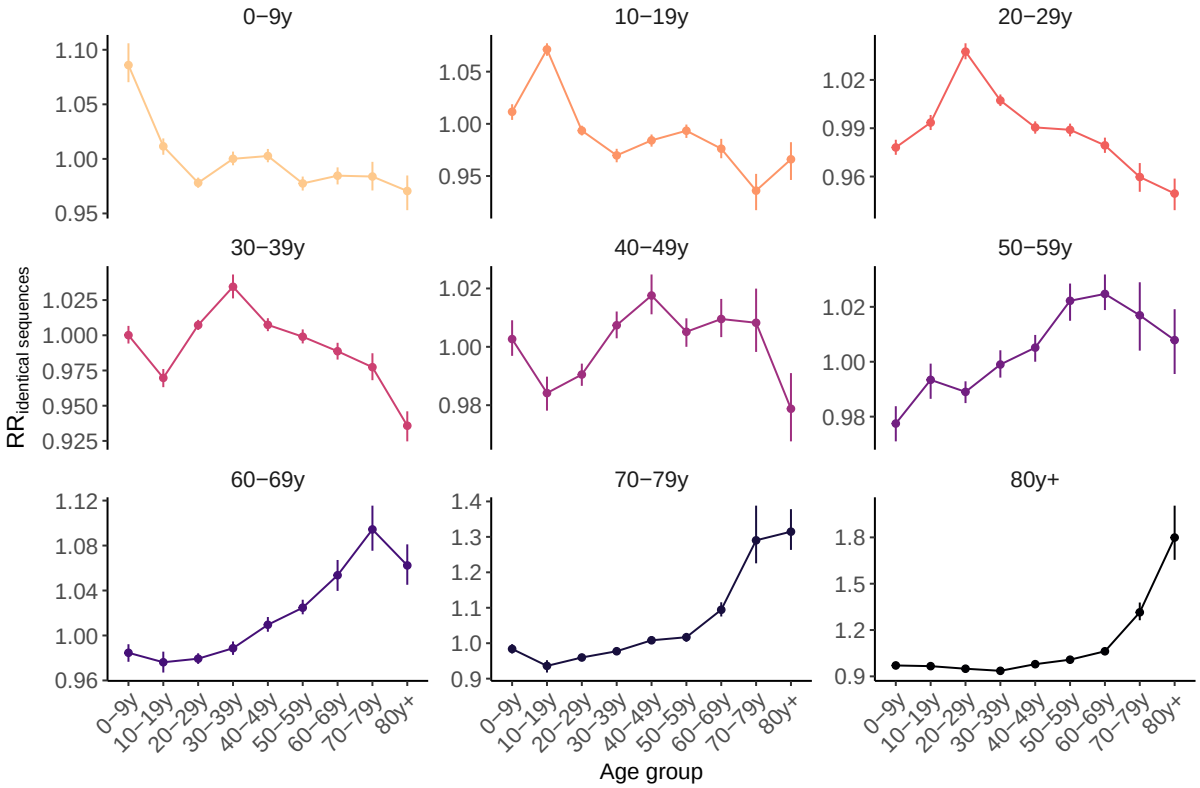


Figure S17. Relative risk for pairs of identical sequences of being observed between two age groups. Vertical segments correspond to 95% confidence intervals obtained through subsampling.

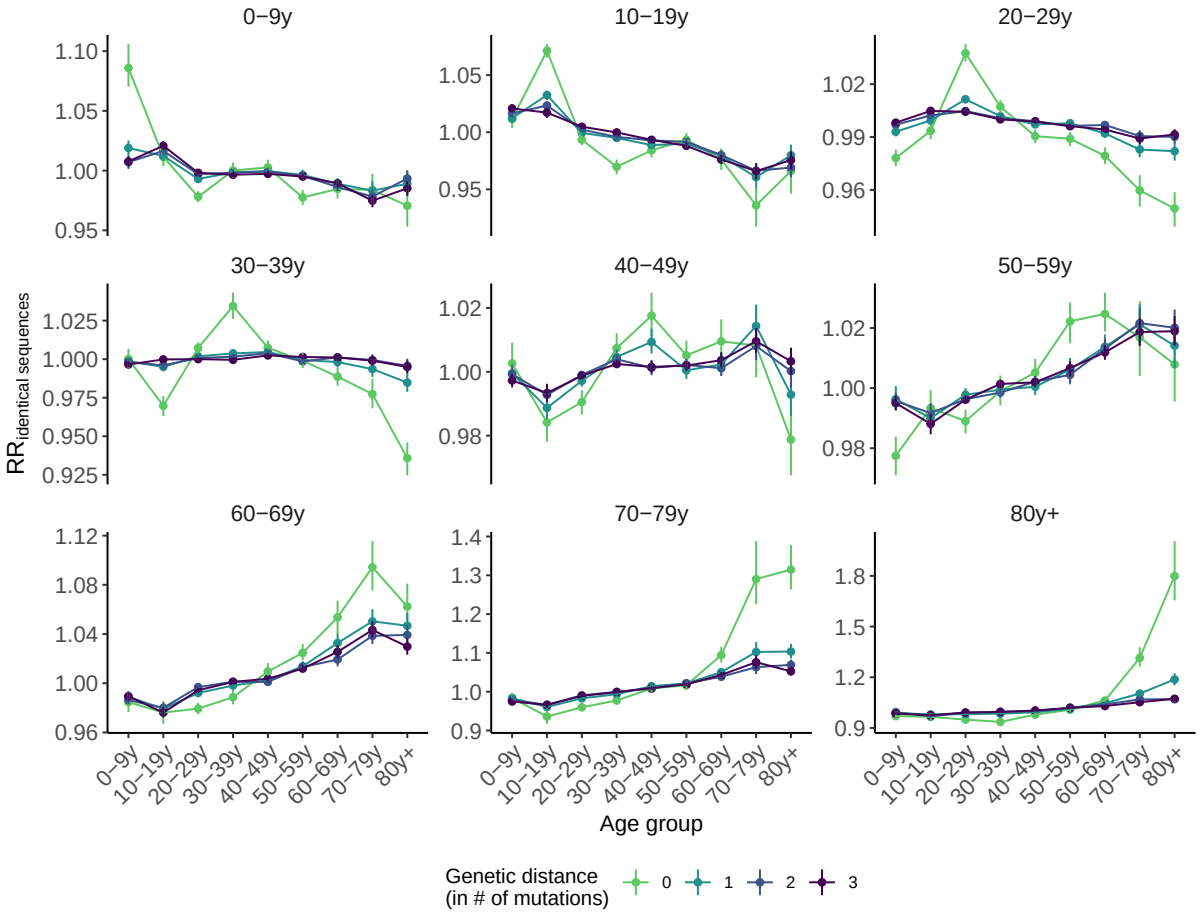


Figure S18. Relative risk for pairs sequences of being observed between two age groups depending on their genetic distance. Vertical segments correspond to 95% confidence intervals obtained through subsampling.

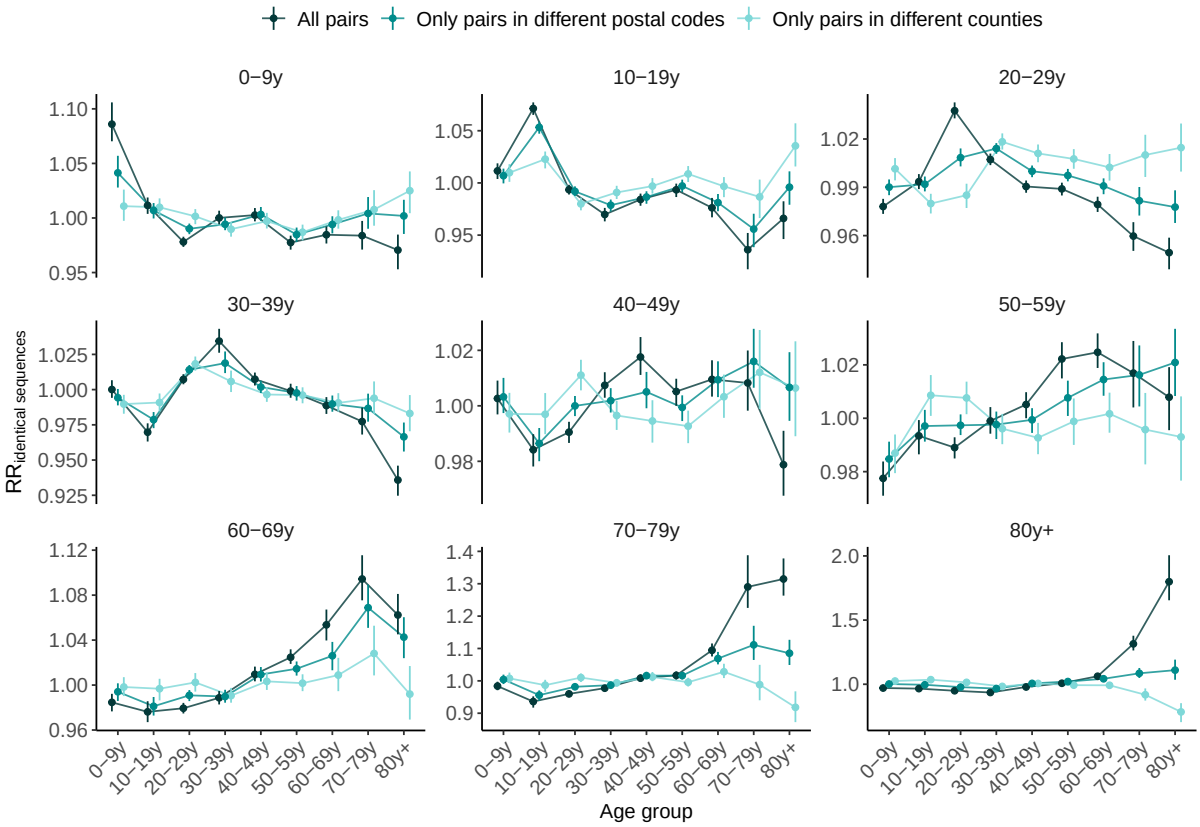


Figure S19. Impact of the spatial scale on the relative risk for pairs sequences of being observed between two age groups. Vertical segments correspond to 95% confidence intervals obtained through subsampling.

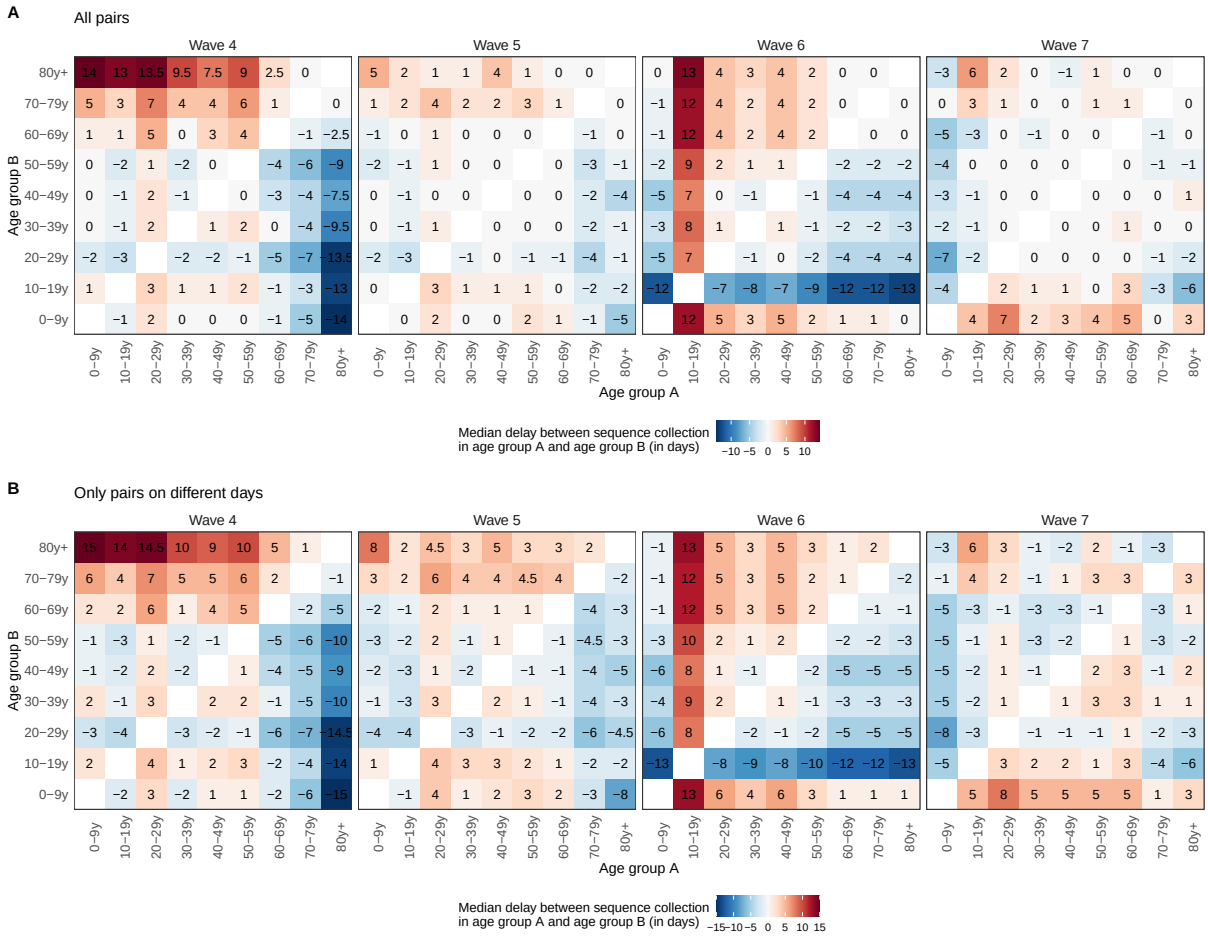


Figure S20. Median delay between the dates of sequence collection within pairs of identical sequences A. considering all pairs of identical sequences collected in two age groups and **B.** considering only pairs of identical sequences collected on different days in two age groups.

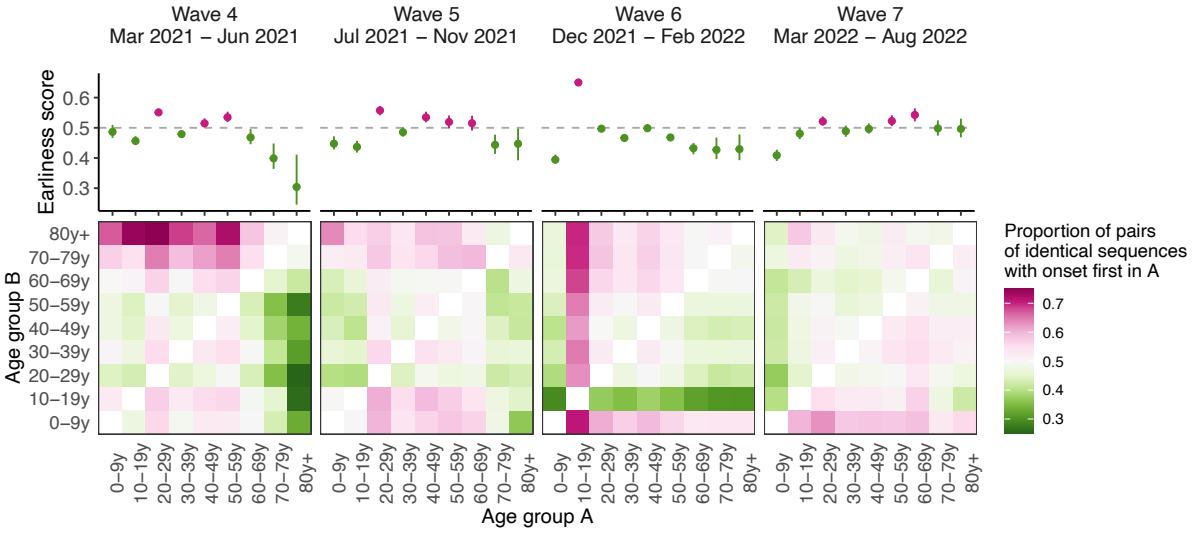


Figure S21. Sensitivity analysis on the timing of pairs identical sequences between age groups using symptom onset dates Median proportion of pairs of identical sequences with onset dates in age groups A before age group B across different epidemic waves from 1,000 imputed datasets (heatmaps). The dots 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).

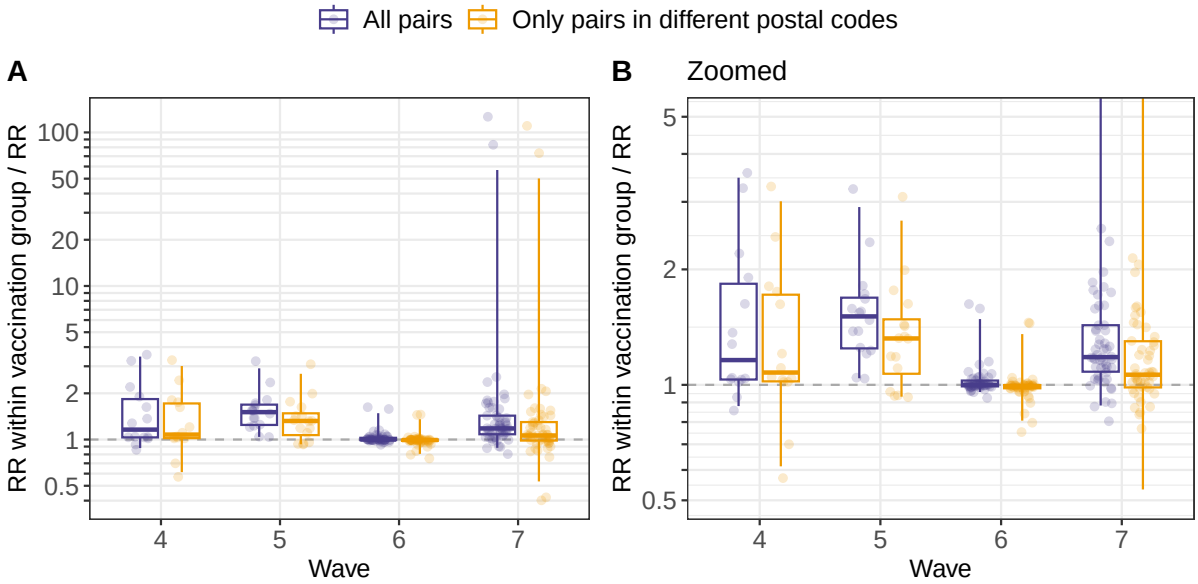


Figure S22. Ratio between the relative risk of observing identical sequences within a given vaccination group (denoted V_1) and between two vaccination groups (denoted V_1 and V_2). Values above 1 indicate that pairs of identical sequences tend to be enriched in pairs observed within the same vaccination 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.

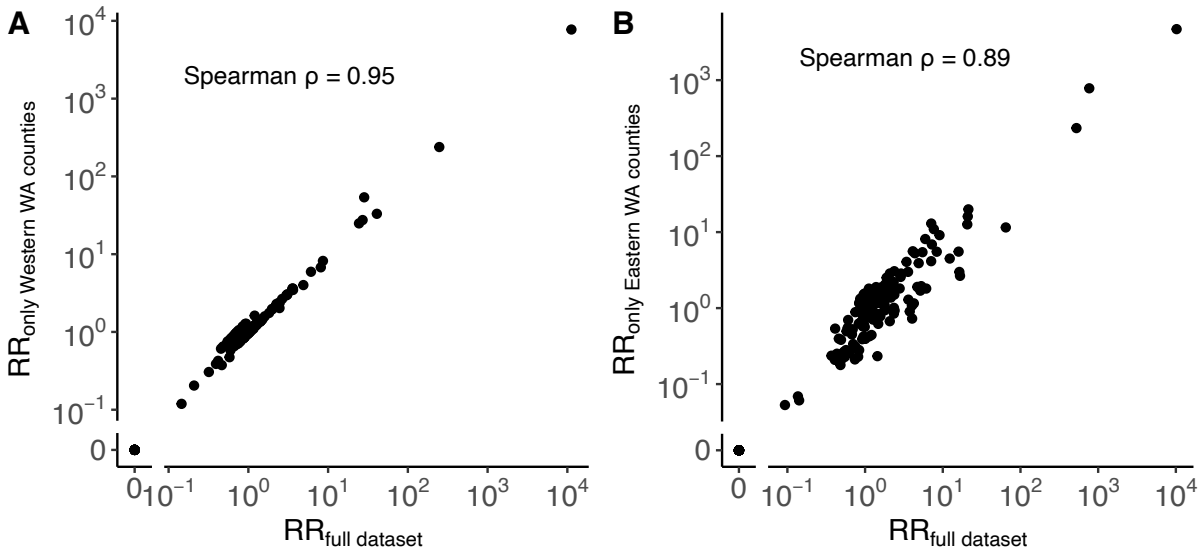


Figure S23. Impact of non-sampled locations on the computation of the RR. A. Comparison between the relative risk of observing identical sequences between Western WA counties using only sequence in Western WA counties or the entire sequence dataset. **B.** Comparison between the relative risk of observing identical sequences between Eastern WA counties using only sequence in Eastern WA counties or the entire sequence dataset.

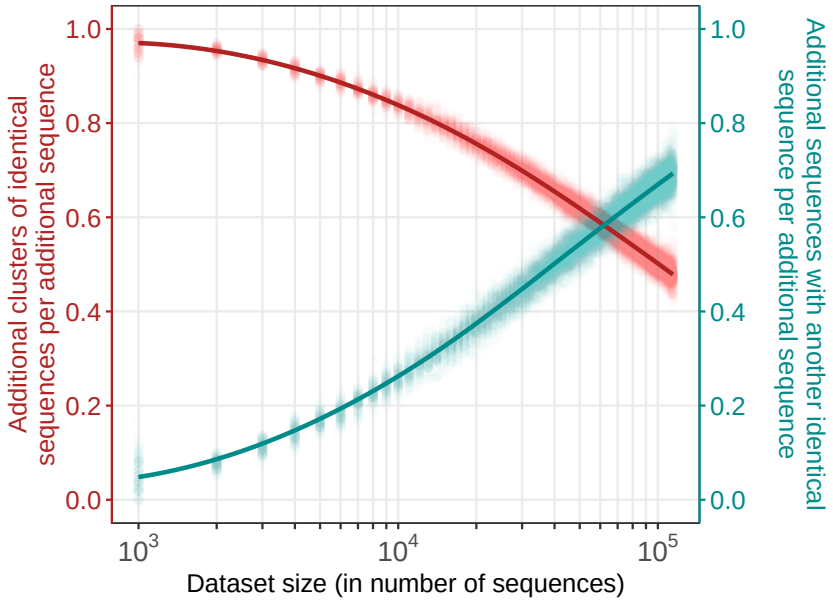


Figure S24. Impact of dataset size on the number of clusters of identical sequences and the number of sequences with another identical sequence in the dataset. We generate this figure by considering sequence datasets of increasing sizes, ranging between 10^2 and the 114,298 (the size of our WA dataset) 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.

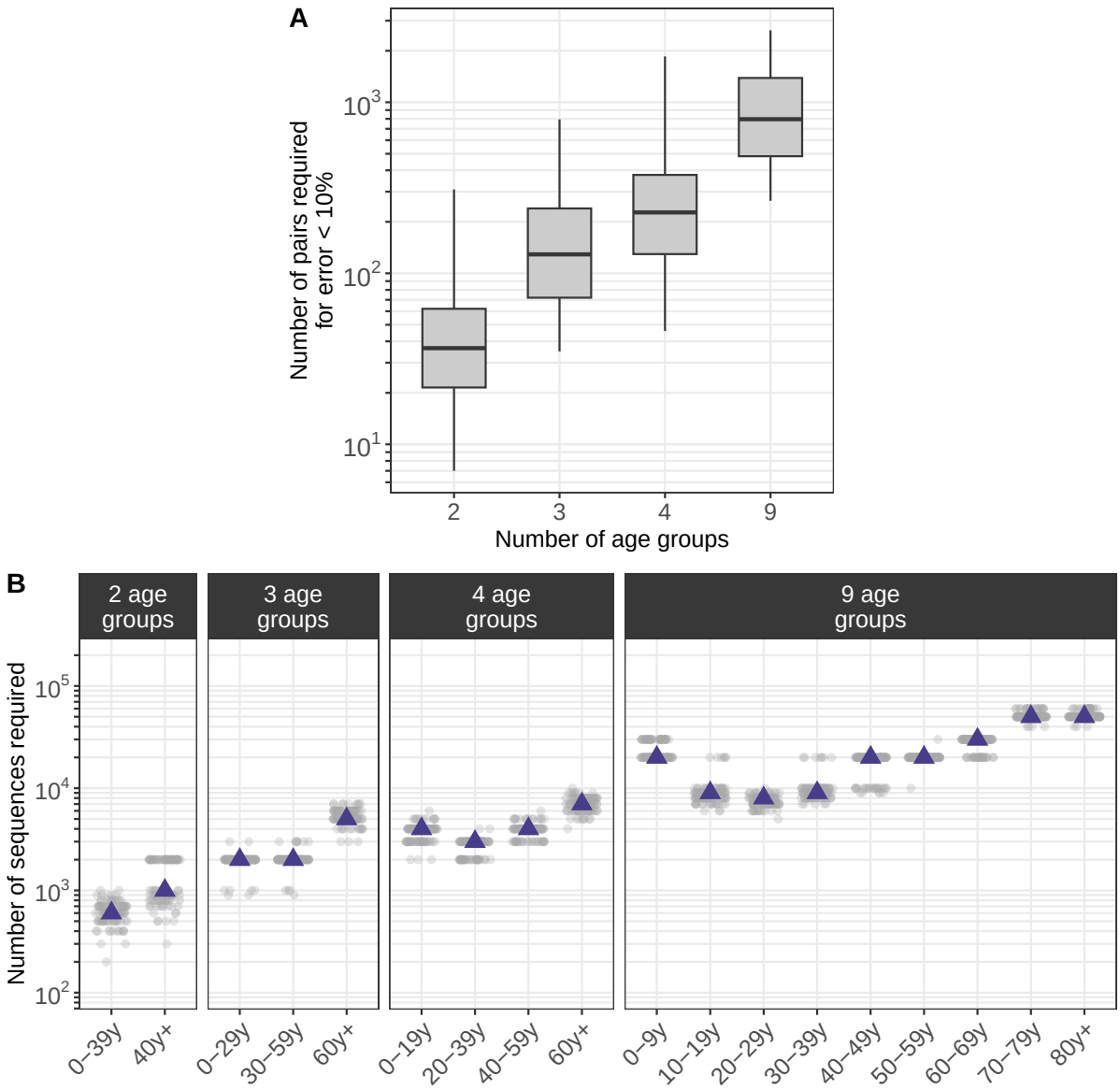


Figure S25. Impact of the number of groups included in the analysis on the dataset size required for the error relative risk of observing identical sequences to be lower than 10%. A. Number of pairs of identical sequences required for the error relative risk of observing identical sequences to be lower 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.

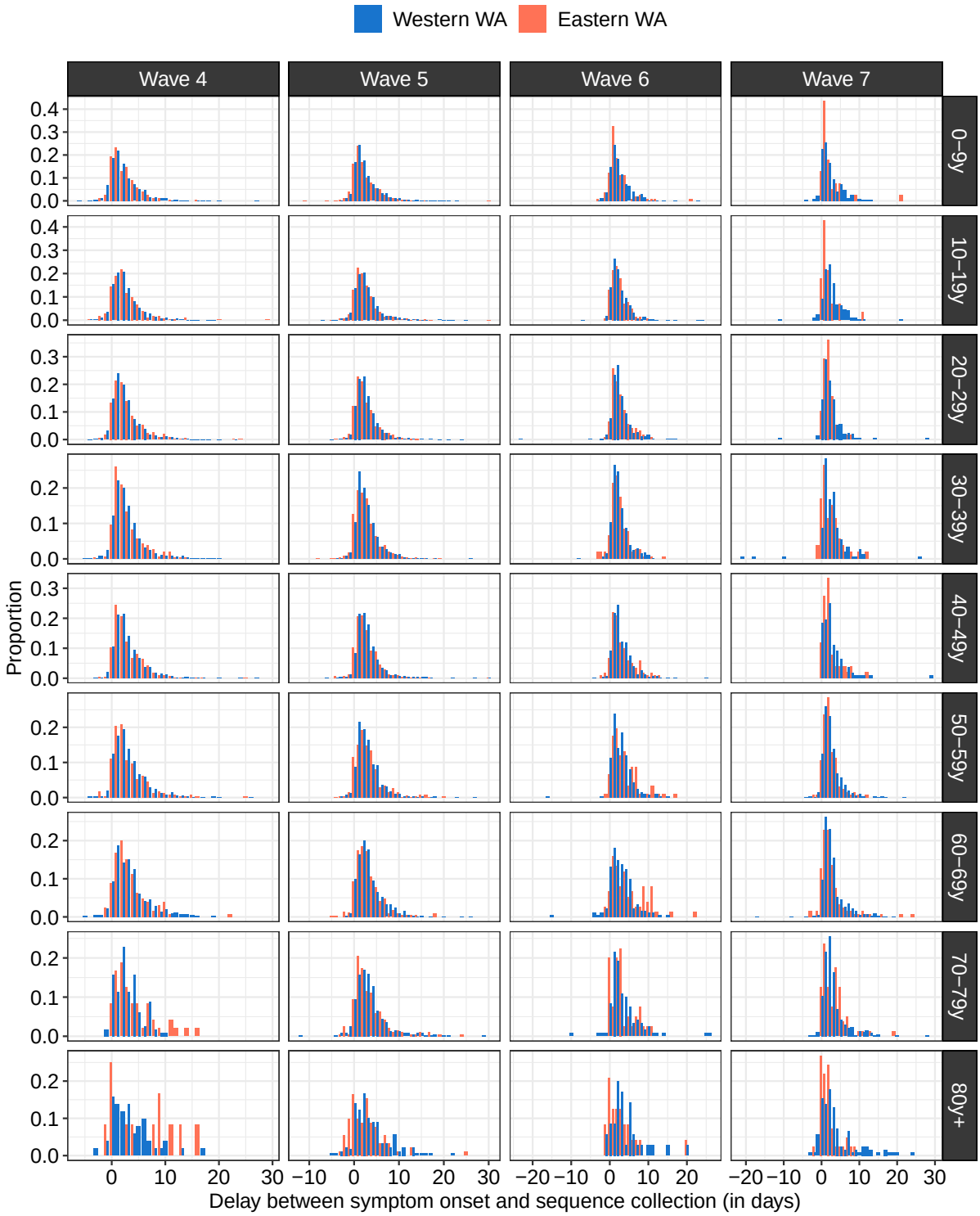


Figure S26. Empirical distribution of the delay between symptom onset and sequence collection by age (rows), period (columns) and geographic region (colours).

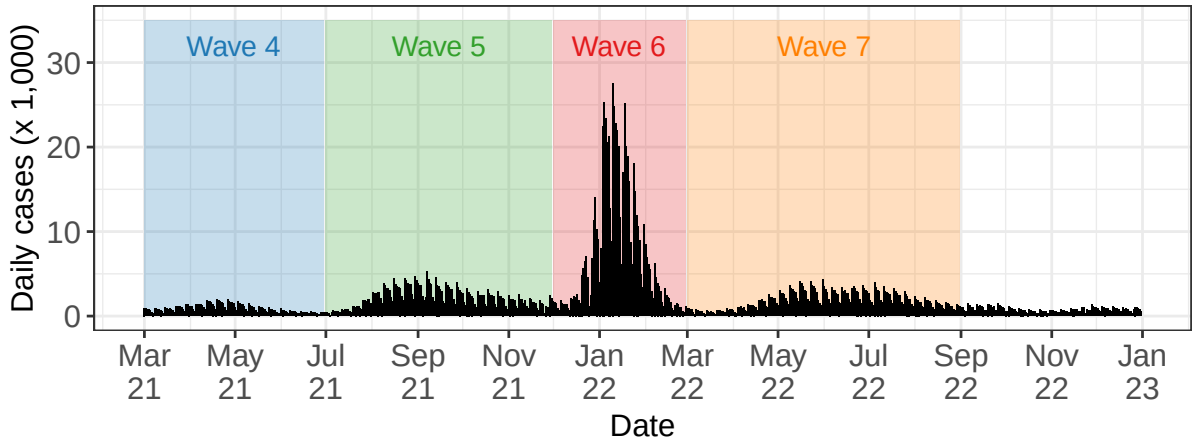


Figure S27. Time-series of COVID-19 cases in WA over the study periods. Shaded rectangles indicate the periods used to define the successive epidemic waves.

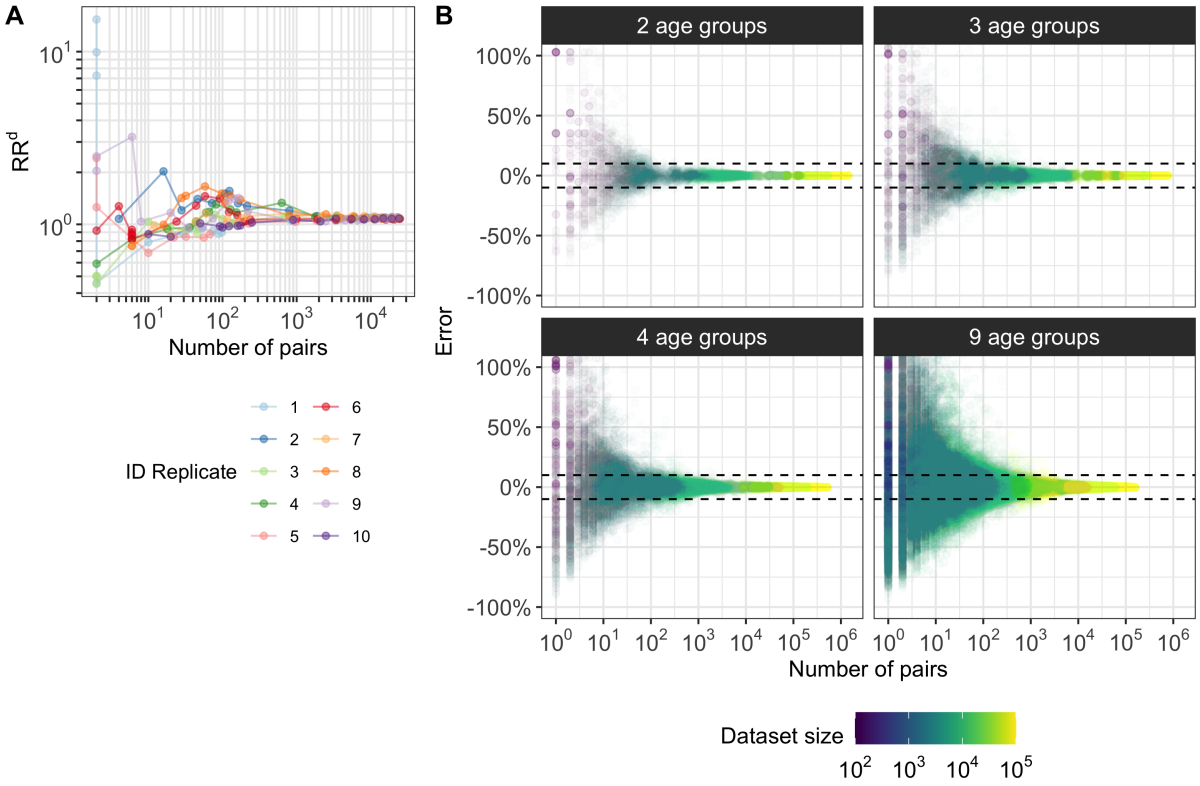


Figure S28. Illustration of the downsampling strategy used to quantify the amount of data required to compute relative risks. **A.** Relative risk RR^d of identical sequences being shared between the 0-9y computed on 10 downsampled datasets as a function of the number of pairs of identical sequences 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 Fixing tree	0.43 %	1	1745	1744	0.54	-0.22	0.10
DTA Fixing tree	0.43 %	2	1714	1805	0.60	0.19	0.39
DTA Fixing tree	2.16 %	1	8723	8709	0.61	0.39	0.84
DTA Fixing tree	2.16 %	2	8551	8675	0.77	0.25	0.76
DTA Inferring tree	0.43 %	1	1745	1744	0.10	0.15	-0.13
RR	2.16 %	1	8723	8709	0.94	0.80	0.79
RR	2.16 %	2	8551	8675	0.80	0.75	0.86
RR	8.66 %	1	34338	35304	0.91	0.74	0.77
RR	8.66 %	2	34736	35123	0.93	0.84	0.79

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.