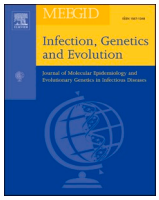
Infection, Genetics and Evolution 113 (2023) 105484

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid



The utility of SARS-CoV-2 genomic data for informative clustering under different epidemiological scenarios and sampling

Benjamin Sobkowiak [a,*,1], Pouya Haghmaram [a], Natalie Prystajecky [b,c], James E.A. Zlosnik [b], John Tyson [b], Linda M.N. Hoang [b,c], Caroline Colijn [a]

[a] Department of Mathematics, Simon Fraser University, Burnaby, Canada
[b] BC Centre for Disease Control Public Health Laboratory, BC Centre for Disease Control, Vancouver, Canada
[c] Department of Pathology and Laboratory Medicine, Faculty of Medicine, University of British Columbia, Canada

ARTICLE INFO

Keywords:
Bioinformatics
Epidemiology
SARS-CoV-2
Phylogenetics
Mathematical modelling
Infectious diseases

ABSTRACT

Objectives: Clustering pathogen sequence data is a common practice in epidemiology to gain insights into the genetic diversity and evolutionary relationships among pathogens. We can find groups of cases with a shared transmission history and common origin, as well as identifying transmission hotspots. Motivated by the experience of clustering SARS-CoV-2 cases using whole genome sequence data during the COVID-19 pandemic to aid with public health investigation, we investigated how differences in epidemiology and sampling can influence the composition of clusters that are identified.

Methods: We performed genomic clustering on simulated SARS-CoV-2 outbreaks produced with different transmission rates and levels of genomic diversity, along with varying the proportion of cases sampled.

Results: In single outbreaks with a low transmission rate, decreasing the sampling fraction resulted in multiple, separate clusters being identified where intermediate cases in transmission chains are missed. Outbreaks simulated with a high transmission rate were more robust to changes in the sampling fraction and largely resulted in a single cluster that included all sampled outbreak cases. When considering multiple outbreaks in a sampled jurisdiction seeded by different introductions, low genomic diversity between introduced cases caused outbreaks to be merged into large clusters. If the transmission and sampling fraction, and diversity between introductions was low, a combination of the spurious break-up of outbreaks and the linking of closely related cases in different outbreaks resulted in clusters that may appear informative, but these did not reflect the true underlying population structure. Conversely, genomic clusters matched the true population structure when there was relatively high diversity between introductions and a high transmission rate.

Conclusion: Differences in epidemiology and sampling can impact our ability to identify genomic clusters that describe the underlying population structure. These findings can help to guide recommendations for the use of pathogen clustering in public health investigations.

1. Introduction

The evaluation of pathogens for genomic similarities, or clusters, to identify common origins and patterns of transmission can be an important step in the surveillance and investigation of disease outbreaks. Whole genome sequence (WGS) data is a valuable tool in complement with classical epidemiological data sources for clustering and has been increasingly used to link cases by the distance between genomic sequences, notably during the COVID-19 pandemic (Seemann et al., 2020; Geoghegan et al., 2020). The definition of a 'cluster' can vary depending on the setting and purpose of the investigation. In outbreak investigations, clusters of closely related pathogens can indicate recent transmission between hosts (Campbell et al., 2018). When coupled with epidemiological information, genomic clusters can also indicate transmission hotspots within a particular setting to guide public health investigation (Poon et al., 2016). In practice, clustering serves to connect cases with a shared transmission history that are distinct from other samples or groups of infections in a population.

* Corresponding author.
E-mail address: bs2259@yale.edu (B. Sobkowiak).
[1] Current address - Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, Connecticut, USA.

https://doi.org/10.1016/j.meegid.2023.105484
Received 18 May 2023; Received in revised form 25 July 2023; Accepted 30 July 2023
Available online 31 July 2023

All hosts infected with a pathogen that spread only through person-to-person transmission are descended from an infected host. Therefore, if transmission is occurring locally and all infected individuals are sampled then, in truth, there will be a single network connecting all cases. We often identify genomic clusters based on linking cases under a given genomic or phylogenetic distance threshold (Stimson et al., 2019). With these methods, the complete transmission network may delineate into smaller groups when disease and epidemiological characteristics result in heterogeneity in the genomic distance between hosts in transmission events. For example, this can occur in pathogens with variable periods of disease latency or where multiple, rapid transmission events can descend from a host through superspreading.

However, clustering may be erroneous when true transmission networks are fragmented or multiple, distinct networks are joined together in a single cluster. One or more transmission links may be missed where sampling, and by extension sequencing, is incomplete, causing true clusters to break up into smaller groups of sampled cases (Glynn et al., 1999). This phenomenon has been explored in other pathogens, including HIV (Grabowski et al., 2018). The fraction of hosts that can be sequenced will be driven by the local sampling approaches and sequencing capacity, the proportion of subclinical infection, and movement of individuals between settings. Furthermore, transmission can occur both locally and from infected individuals that are introduced from outside sources, which may seed outbreaks inside a sampled jurisdiction. Whether all cases in these separate outbreaks are identified as distinct clusters will depend on the genomic divergence of the introduced cases, as well as the diversity within the local disease population. These factors can be influenced by the rate of introduction and the transmission rate within the sampled population.

Genomic clustering of SARS-CoV-2 sequences during the COVID-19 pandemic presented a challenge due to rapid and extensive transmission across multiple jurisdictions of strains with limited genomic diversity. Additionally, there were varying levels of strain diversity throughout the pandemic due to the emergence of different variants, which resulted in periods where multiple distinct strains were co-circulating (Stefanelli et al., 2022), along with very few periods where cases were almost exclusively caused by a single SARS-CoV-2 variant (Davies et al., 2021). These factors complicated the ability to identify a consistent but useful method for clustering SARS-CoV-2 cases and limited the utility of sequence-based genomic clustering (Bendall et al., 2022).

Here, we describe the utility and limitations of genomic clustering using sequence data under different epidemiological scenarios in pathogen outbreaks characterised by rapid transmission and relatively short infection time, such as COVID-19. Motivated by our experience of identifying clusters for public health surveillance during the COVID-19 pandemic, we describe the diversity and structure of clusters found in SARS-CoV-2 WGS strains collected in British Columbia, Canada, using our recently developed method for genomic clustering, cov2clusters (Sobkowiak et al., 2022). We compare cov2clusters to standard SNP-based methods, with reference to simulated benchmarking data. Finally, we use simulated SARS-CoV-2 outbreaks to determine how clustering is shaped by the transmission rate and sampling fraction, including scenarios where outbreaks are seeded from introductions from outside pathogen populations with different levels of diversity at varying rates. The principles established through this work can be used to guide clustering methods and interpretation for surveillance and outbreak investigation for pathogens of public health importance.

2. Methods

2.1. SARS-CoV-2 whole genome sequence data

Whole genome sequence data were obtained from SARS-CoV-2 samples from 31,115 individuals in British Columbia, Canada collected between 15th March 2021 and 13th August 2021. These sequences are available through the GISAID database (Shu and McCauley, 2017). Full details of the DNA extraction, genomic sequencing and sequencing data analysis can be found in (Sobkowiak et al., 2022). Lineage calling was performed using the Pangolin lineage assignment tool (v.4.1.2) (Rambaut et al., 2020).

2.2. The outbreak simulation model

We used a stochastic susceptible-infectious-removed (SIR) model to simulate pathogen outbreaks based on the 'simulateoutbreak' function in the SEEDY package (v.1.3) (Worby and Read, 2015). This original function, written in R, simulates a complete outbreak in a single population, beginning with one or more infected individuals introduced into a fully susceptible population. In the context, 'outbreak' refers to all cases in the population that arise from the initial infection, rather than other definitions such as a high concentration of cases over a given time. The user defines the size of the initial susceptible population (*init.sus*), the transmission rate in units of the average number of susceptible individuals infected per day (*inf.rate*), and recovery rate in units of the average number of infected individuals recovered per day (*rem.rate*). The outbreak will finish when no further individuals are infected. A minimum final outbreak size can be specified a priori, and the simulation will repeat until this size is reached. The resulting output from the simulations includes a matrix identifying all infected individuals with their source of infection and a vector of nucleotide changes in the pathogen sequences of infected hosts. These outputs can be used to construct the full transmission tree and calculate pairwise SNP distances between infected hosts.

We extended the functionality of this simulation tool to explore more epidemiological scenarios and to describe SARS-CoV-2 diversification. We modified the original source code of 'simulateoutbreak' to implement the following changes:

We have two populations, named 'inside' and 'outside', to simulate transmission in multiple jurisdictions with different epidemiology. After the disease starts transmitting in the 'outside' population, infected cases can enter the 'inside' jurisdiction at a given rate. These cases can now transmit within the 'inside' population to seed new outbreaks. The parameters for this feature are

- ***init.sus.in***: the initial number of susceptible individuals inside.
- ***min.cases.in***: the desired minimum final size of the inside outbreak.
- ***intr.rate***: the probability of a new introduction at each time point.
- ***inf.rate.in*** and ***inf.rate.out***: the infection rate in the inside and outside populations, replacing the original parameter 'inf.rate'.
- ***mut.prob.site.in*** and ***mut.prob.site.out***: the mutation probability per site per day of infected hosts in the inside and outside populations.
- ***time.lag.outside***: a specified time lag in the outside population before new introductions can enter the inside population.
- min***.perc.outside.inf***: the minimum percentage of the outside population that must have been infected before new introductions can enter the inside population.

1. We have a parameter to simulate diversity between the initially infected individuals when the number of initial infected individuals (***init.inf***) is >1. ***min.init.dist*** and max.***init.dist*** require integer value for the SNP distance between an initially infected individual and the reference sequence (default = 0). If ***min.init.dist*** equals ***max.init.dist*** then this is a fixed value, otherwise the distance is randomly sampled between the minimum and maximum value.
2. We have included two mutation rate parameters, ***mut.rate.in*** and ***mut.rate.out*** to allow for different rates in the inside and outside populations. This value represents the mutation rate in units of SNPs per site per day. Multiple SNPs can evolve between source and infection sequences at each transmission event. There is no within-host variation, and the pathogen genome is fixed in the host after infection.

3. When a susceptible individual passes to the infected compartment at time *t*, its source case is chosen at random from a pool of hosts in the infected compartment at time *t* - 1 that do not recover at time *t* and whose infection dates correspond to the date sampled from a gamma probability distribution, with user-defined shape (a) (***shape.infect***) and rate (l) (***rate.infect***) parameters. This parameter simulates a generation time distribution. In contrast, the original 'simulateoutbreak' function in SEEDY chose the source case from all hosts in the infected compartment at time *t* uniformly at random.

The function 'simulate_outbreak' and associated documentation can be found at github.com/bensobkowiak/Clustering_simulations.

2.3. Genomic clustering of SARS-CoV-2 outbreaks

Motivated in part by the need to refine very large clusters using additional information, we developed a clustering approach ('cov2-clusters') that incorporates the genetic distance between two viral isolates as well as the difference in collection times. The approach, based on a logit model, readily incorporates additional data if needed (these could include geographic location, known contact, exposure site or other variables).

We used cov2clusters to produce genomic clusters from sequence data in both the real-world and simulated outbreaks (Sobkowiak et al., 2022). This tool incorporates genomic distance and sampling times to estimate the probability of cases belonging to the same cluster using a logit model. We used the same genomic distance and date coefficients as used previously for SARS-CoV-2 clustering ($\beta1 = 0.66$, $\beta2 = 0.075$). We validated clustering using this model in simulated outbreaks against a logit model with the same genomic distance but no date information, and clusters produced using only pairwise SNP distance thresholds of 1, 2, and 3 SNPs. All analysis was conducted in R (scripts available at github.com/bensobkowiak/Clustering_simulations) and network plots illustrating genomic clusters in outbreaks were produced using 'igraph' (Csardi and Nepusz, 2006).

2.4. Simulated SARS-CoV-2 outbreaks

We used our pathogen outbreak simulation model to generate SARS-CoV-2 outbreaks under different epidemiological scenarios with varying sampling fractions and infection rates. Outbreaks were simulated with initial susceptible population sizes (*init.sus*) that produced final outbreaks of 200 to 300 infected individuals.

To validate the ability of cov2clusters to correctly cluster known sequences that are linked in transmission networks, outbreaks were simulated as single outbreaks in the 'outside' population, with varying infection and mutation rates, and sampling fractions. To simulate single 'outside' outbreaks, the introduction rate and minimum inside outbreak size (*intr.rate* and *min.cases.in)* were set to 0. The infection rate (*inf.rate.out)* parameter was set to 0.6 and 0.25 with the same removal rate (*rem.rate*) of 0.2, equal to a reproduction number (R) of 3 and 1.25. We tested two values for the mutation rate (*mut.rate.out*) parameter of 1×10^{-6} and 5×10^{-6} substitutions per site per day, which correspond to upper and lower estimates in SARS-CoV-2 (Li et al., 2022). Sampling fractions were changed by varying the probability of sampling an infected individuals using the binomial distribution, testing the sampling fractions of 0.1, 0.25, 0.5, and 1.

Next, to assess the utility of genomic clustering to reflect the true underlying population structure under different outbreak scenarios, we first simulated single outbreaks, varying the infection rate and sampling fraction as before. We used a single mutation probability of 3×10^{-6} substitutions per site per day in both the inside and outside populations, which corresponds to the previously reported mutation rate of SARS-CoV-2 (Li et al., 2022). Genomic clustering was performed using a logit probability threshold of 0.8 and 100 simulated outbreaks per epidemiological scenario were used to calculate the clustering properties (the proportion of un-clustered sequences, the proportion of sequences in the largest cluster, and the number of proposed clusters).

Finally, we tested simulations with both 'inside' and 'outside' populations, where one or more infected cases circulating in the 'outside' population may give rise to an outbreak in the 'inside' population through introduction events. We looked at the impact of genomic clustering in outbreaks within the 'inside' population at high and low infection rates (*inf.rate.in* = 0.6 and 0.25) and different levels of genomic diversity in the 'outside' population. To achieve varying levels of diversity in the 'outside' population, we initiated simulations with 20 initially infected cases separated by either 0 SNPs for low diversity settings or with a pairwise divergence of between 15 and 25 SNPs for high diversity settings. We fixed the infection rate outside at 0.4 and ran the 'outside' simulation allowing for introduction events to the 'inside' population after 20% of outside susceptible individuals had been infected. We also investigated varying the rate of introduction of infected cases from the 'outside' population to the 'inside' (*intr.rate* = 0.5 and 0.1). Initial population sizes (*init.sus* and *init.sus.in*) were specified to produce outbreaks of between 200 and 300 infected individuals 'inside', and between 150 and 250 infected individuals in the 'outside' population. We again explored clustering under different sampling probabilities of 0.1, 0.25, 0.5, and 1 and repeated all simulations for 100 replicates to calculate cluster statistics. All scripts to produce simulated outbreaks are written in R and can be found at github.com/bensobkowiak/Clustering_simulations.

3. Results

Fig. 1 illustrates the changing epidemiology over the study period, with rises and falls in reported COVID-19 cases, and commensurate rises and falls in the number of sequenced cases. The sequencing fraction (the proportion of confirmed cases for which there were sequence data) was relatively high, particularly in the summer of 2021 where there were sequences for most confirmed cases (Fig. 1A). Seroprevalence data also suggests that the ascertainment fraction (the number of cases reported vs true number of infections) was also likely to be high in BC during the study period (COVID-19 Immunity Task Force, 2021). In the first third of the time period, both B.1.1.7 (the Alpha variant) and P.1 (Gamma) were rising in BC. Following the introduction of public health restrictions (non-pharmaceutical interventions) including limiting social interaction to households only and a ban on non-essential travel within the province, the number of cases began to fall. Even with the easing of some of the toughest restrictions (Government of British Columbia, 2021), this decline lasted until the introduction and rapid rise of the Delta variant (B.1.617.2 and AY.25) in the late summer of 2021. Both Gamma and Delta exhibited rapid rises in BC, with the rises of both Gamma and Delta AY.25 occurring shortly in BC after their first identification globally. In contrast, when Delta's B.1.617.2 sub-lineage rose in BC, this was several months after its dramatic rise in India in February 2021 (Kirola, 2021). This was similar in Alpha, which emerged globally months before its substantial rise in BC (Murall et al., 2021).

The global diversity of each variant, and the epidemiology in BC as it arose, combined to shape the variants' SNP distance and clustering patterns. Fig. 1C shows the pairwise SNP distances among samples of each variant in each week. The Alpha variant had a relatively high pairwise SNP distance at the beginning of 2021, as did the B.1.617.2 Delta types. In contrast, both P.1 (Gamma) and AY.25 (Delta) had very low pairwise SNP diversity in BC, consistent with very recent emergence globally at the relevant time. This difference is reflected in the clustering experience (Fig. 1D), with P.1 and AY.25 having high fractions of the sequences in one large cluster (the largest). Naturally, if a large part of the utility of sequencing is to group cases into small, well-distinguished

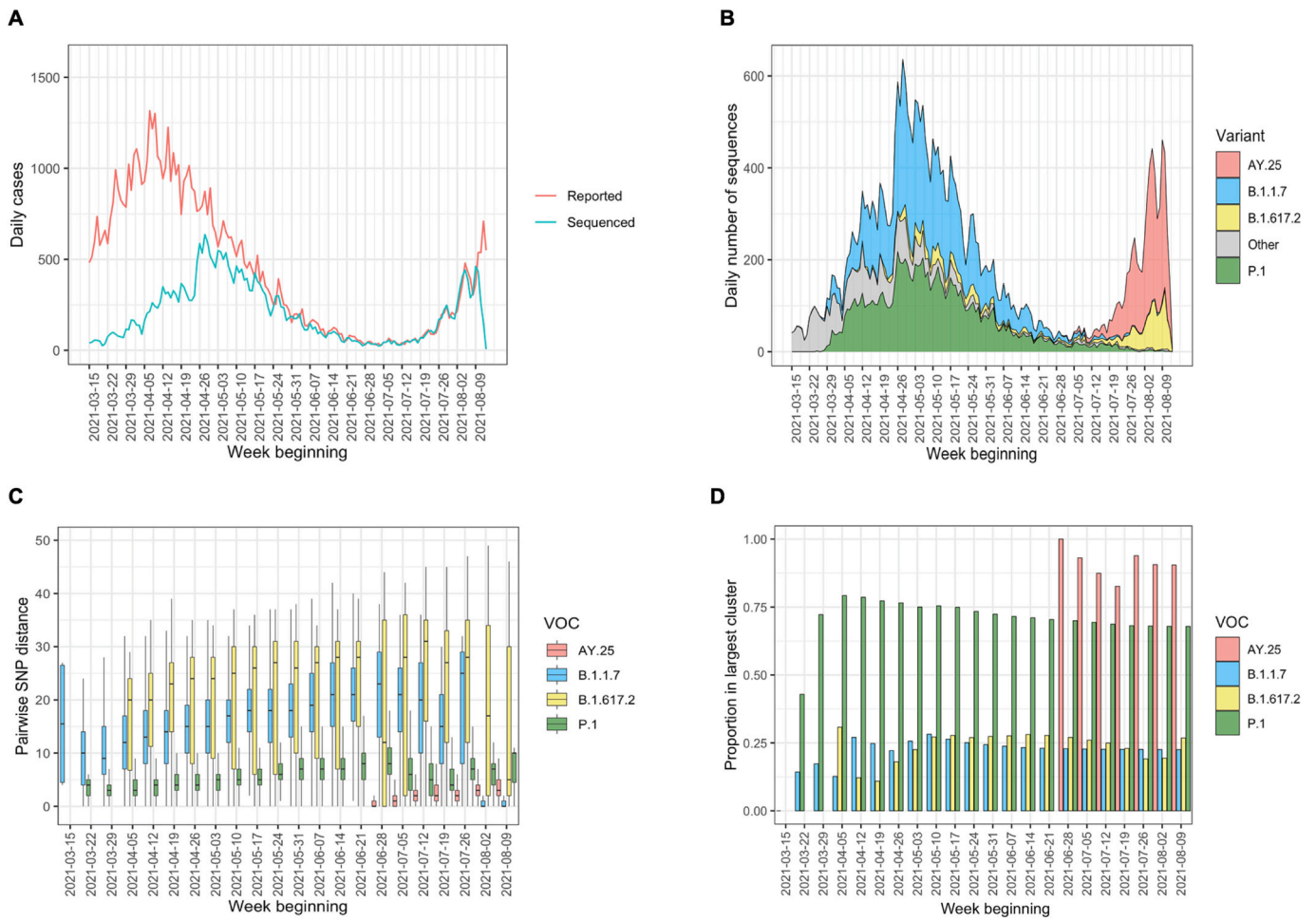


Fig. 1. BC SARS-CoV-2 sequence data for samples collected between 3rd March 2021 and 13th August 2021. A) Daily cases reported, and the daily number of sequences collected over the study period. B) The daily number of sequenced cases by the four major variants of concern (VOCs). C) The weekly pairwise SNP distance by VOC. D) The proportion of sequences in the largest VOC cluster by week, identified by cov2clusters.

and meaningful clusters to guide public health action, having a high fraction of cases in one very large cluster is not informative.

Next, we compare cov2clusters and SNP thresholds' ability to classify whether a direct transmission pair in a simulated outbreak is linked. Clustering with both SNP thresholds and cov2clusters work with pairs as their operational unit, first asking whether a pair of individuals should be linked (based on virus sequences and timing, or based only on sequences, respectively), and then creating clusters using connected components of the graph in which edges are created among all pairs that meet the criteria for being linked. Fig. 2 shows the results; in particular, we find that logit clustering achieves a higher model performance than SNP thresholds under all conditions (AUC 0.84–0.92 compared to the best performing SNP threshold of 1 SNP with AUC 0.55–0.83; Supplementary Table 1). This performance is good in the context of correctly clustering transmission pairs in a pathogen characterised by low levels of diversity and incomplete sampling.

The higher model performance of logit clustering compared to using only SNP distance is in part due to the discrete nature of SNP thresholds (1, 2, or 3 SNPs rather than a continuous distance), as well as to the inclusion of date information. Date information helps the most when the mutation rate is low (columns 1 and 3, compare the solid and dotted red lines), regardless of R. This is because under a low mutation rate, more sequences are genetically very similar, and adding date information reduces false positives (classifying a pair as a transmission pair when it is not). In contrast, when the mutation rate is higher, fewer sequences are very similar and true transmission links are likely to be the (rarer) pairs with very small genetic distance. We found that the sampling fraction does not have much effect on either methods' performance: under lower sampling there are fewer pairs overall, and so fewer false and true positives.

We simulated outbreaks under different assumptions about sampling and the rate of growth and determined how this impacts the cluster structure. This helps to interpret the clustering patterns. We fixed a set mutation rate that corresponded to a realistic amount of genomic diversity over time in SARS-CoV-2 outbreaks of around 2 SNPs per month (Supplementary Fig. 1). Fig. 3 shows the results arising from single outbreaks simulated in one jurisdiction (or one importation event per outbreak). In this context, the "truth" is that there is one large cluster: every case (except the index) results from transmission from another individual in the jurisdiction, who could have been sampled. We find that a lower sampling fraction gives rise to apparent cluster structure in the data, due to missing intermediate cases (Fig. 3A). Thus, while the truth is that all the cases are linked (and there are no subpopulations with shared exposure sites, high-risk settings, households, workplaces and so on), cases appear to be grouped into small clusters. This effect is particularly pronounced at lower values of *R*, when transmission chains are more chain-like (one person to the next to the next) and is less pronounced when *R* is higher (individuals infect three others on average). We note that we simulated until we reached a fixed outbreak size range, so the outbreak durations differ. The spurious break-up of the one true cluster is summarized in the measures shown in Figs. 3B-E. The portion of cases not in a cluster is significantly higher at lower R than

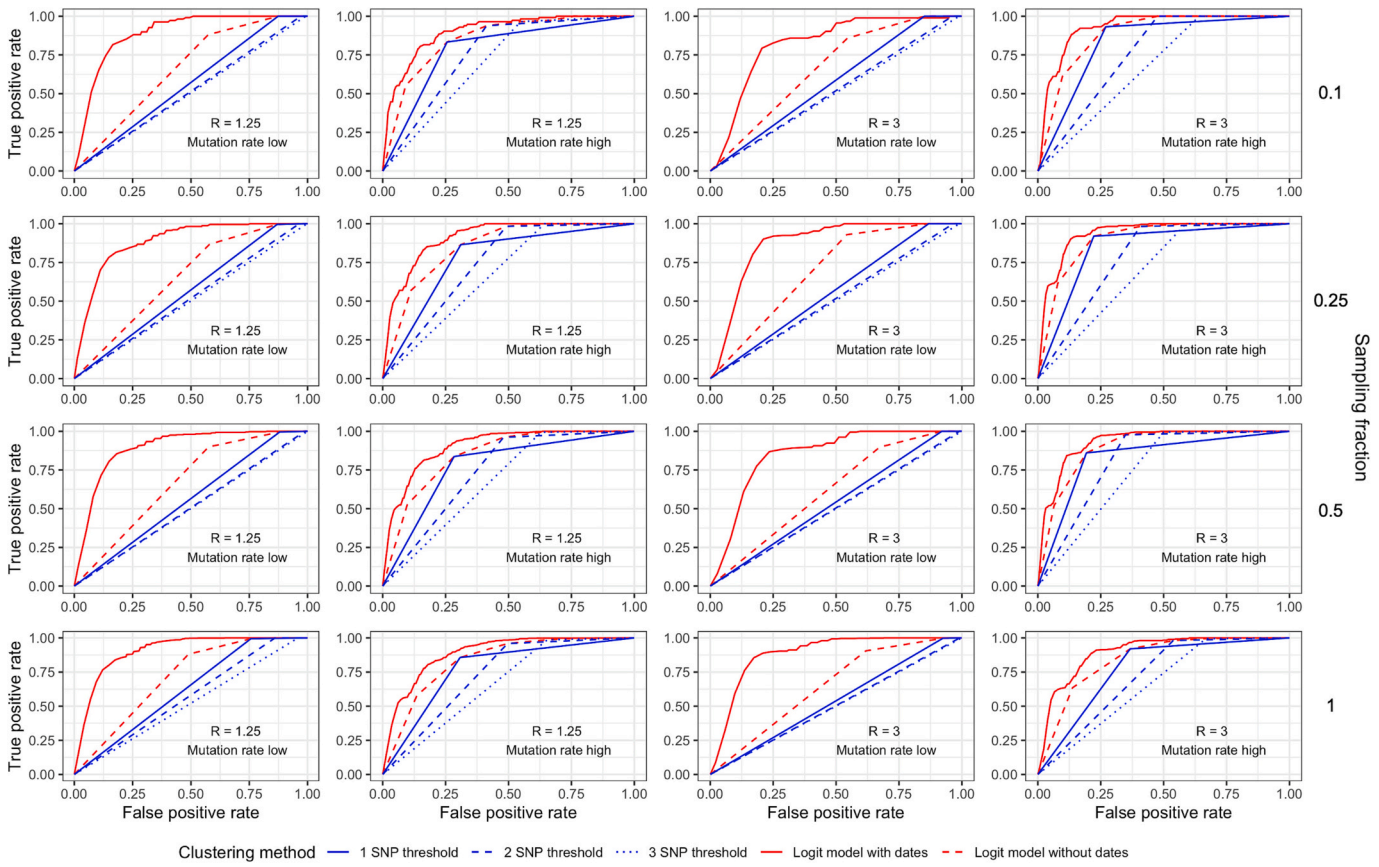


Fig. 2. Validation of the cov2clusters tool. ROC curves for clustering in simulated outbreaks, comparing logit clustering with only genomic distance, genomic distance and dates, and clustering by pairwise SNP distance at 1, 2, and 3 SNP thresholds. High and low R and mutation rate was tested, and the sampling fraction increased from 0.1, 0.25, 0.50, to 1. The first 2 columns are from simulations with low R (1.25), with low (1st column) and high (2nd column) mutation probability per site. In the rightmost two columns, R was high (3), with low (3rd column) and high (4th column) mutation probability per site.

higher R at all sampling fractions except 0.5 (Fig. 3B; two sample *t*-test $P < 0.05$), and the proportion of sequences in the largest cluster is lower at lower R at all sampling fractions (Fig. 3B; two sample t-test $P < 0.05$). We also find more instances where clustering predicts two or more clusters in simulations with a low R than higher R (Fig. 3E).

The size and membership of clusters was sensitive to changes in the logit probability threshold to link cases (Supplementary Fig. 2). As would be expected, the true cluster broke up into a higher number of smaller genomic clusters and an increased number of unclustered cases when we used a higher probability threshold (0.9). Notably, outbreaks simulated with a lower R were more sensitive to changes in the probability threshold than outbreaks simulated at higher R, with a far greater difference in the proportion of unclustered cases and cases in the largest cluster when the probability threshold changed from 0.5 to 0.9. The differences in clustering at higher and lower thresholds were also most pronounced at lower sampling fractions, with far fewer cases assigned to the largest genomic cluster when using a high probability threshold compared to a low threshold at lower sampling fractions. This suggests that the fragility of clustering patterns to the choice of threshold may be informative as to whether cluster structure is at least in part an artefact of incomplete sampling.

Finally, we carried out simulations of outbreaks that co-occur inside a sampled jurisdiction, after being seeded by introduced cases from an outside circulating pathogen population. Fig. 4 shows the impact of varying *R* in the sampled jurisdiction, the genomic diversity in the outside population, and the sampling fraction, in different scenarios with multiple introductions. When genomic diversity is high in the population from which cases are being introduced into the sampled population, we identified a high number of distinct genomic clusters from separate introductions with complete sampling, irrespective of R. In this context, clustering has high utility for distinguishing among chains of transmission; the Alpha variant in 2021 is an example (Fig. 1): whether cases were rising or falling, high diversity in Canada and internationally would give discriminating power to clustering tools. This scenario also reflects a scenario in which sampling is done primarily in high-risk settings *within* a jurisdiction, for example, high diversity of a virus in the general community ("outside"; rarely sampled) with outbreaks in high-risk settings ("inside"; well-sampled).

Both high transmission and low diversity in the global pool from which cases are introduced lead to the challenge that sometimes, many or even most cases are grouped into one very large cluster. In British Columbia, this occurred in the P.1 (Gamma) and AY.25 (Delta) outbreaks, which had low global diversity at the time and were highly transmissible. In this context, there is truth in the "one very large cluster" - with rapid transmission of an obligate human pathogen, and high sampling, many cases may genuinely be linked, via rapid (and potentially long) transmission chains in the jurisdiction. There is also likely some spurious grouping of distinct introductions, due to low global diversity. Reducing sampling or changing the clustering threshold did not separate the isolates into well-resolved smaller clusters, likely due to the lack of genuine cluster structure in the population.

The scenario in the top row of Fig. 4B, with high "outside" diversity and high transmission, depicts the context for SARS-CoV-2 clustering since the pandemic has moved towards an endemic phase. Sampling is taking place primarily in high-risk settings (a high transmission "inside" jurisdiction), while both globally and in the community (which is much less sampled), diversity is high. Accordingly, outbreaks are well-separated, and clustering performs well. In contrast, when introduced

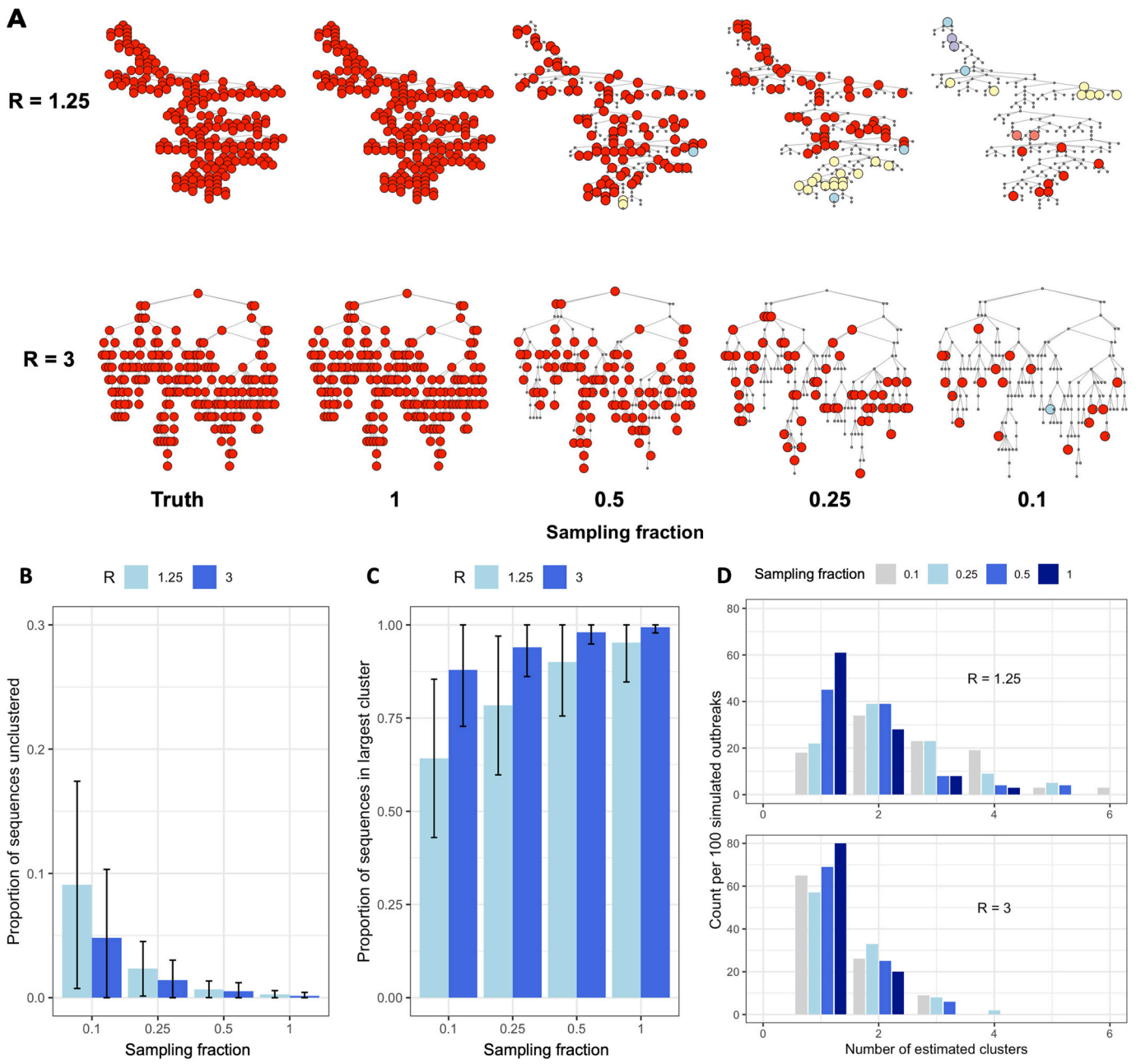


Fig. 3. Simulated single outbreaks of infected individuals at low and high R, with 100 replicates per R to calculate clustering properties. A) Graph plots of example simulated outbreaks at low (1.25) and high (3) R and varying sampling fraction. Nodes are coloured by genomic clusters, with nodes belonging to the largest cluster coloured red and non-clustered nodes coloured light blue. All other colours represent smaller clusters. Unsampled cases are small, grey nodes. B) The mean proportion of sequences that were assigned as un-clustered, C) the mean proportion of sequences assigned to the largest cluster, and D) the number of clusters predicted, in 100 replicate outbreaks. Final outbreaks are simulated to be between 200 and 300 infected individuals. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

cases were closely related, some outbreaks seeded from different introductions were still classed as the same genomic cluster. Lowering the sampling fraction had a greater impact on the number and completeness of clusters when R was low, with a greater increase in the proportion of cases classified as un-clustered at the lower sampling fraction with both high and low outside diversity (Fig. 4C). This is in line with the results shown in Fig. 3, where transmission networks are more chain-like in low R so not sampling some intermediate cases can give rise to spurious cluster structure. We have summarized the clustering patterns under our scenarios, along with suggestions for appropriate actions, in Table 1.

Overall, the same clustering patterns were found when the rate of introduced cases into the sampled jurisdiction was lowered from a probability of 0.5 to 0.1 introductions per day, though with a lower proportion of unclustered cases and a higher majority of cases in one, large cluster (Supplementary Fig. 3). With fewer introductions, transmission is more likely to occur from infected cases in existing outbreaks than by new introductions or cases from new outbreaks. The biggest difference observed with a lower introduction rate was that a higher number of clusters were identified at low R than high R when the sampling fraction was low. This was due to a greater effect of not sampling the intermediate cases in chain-like outbreaks (supplementary Fig. 4C).

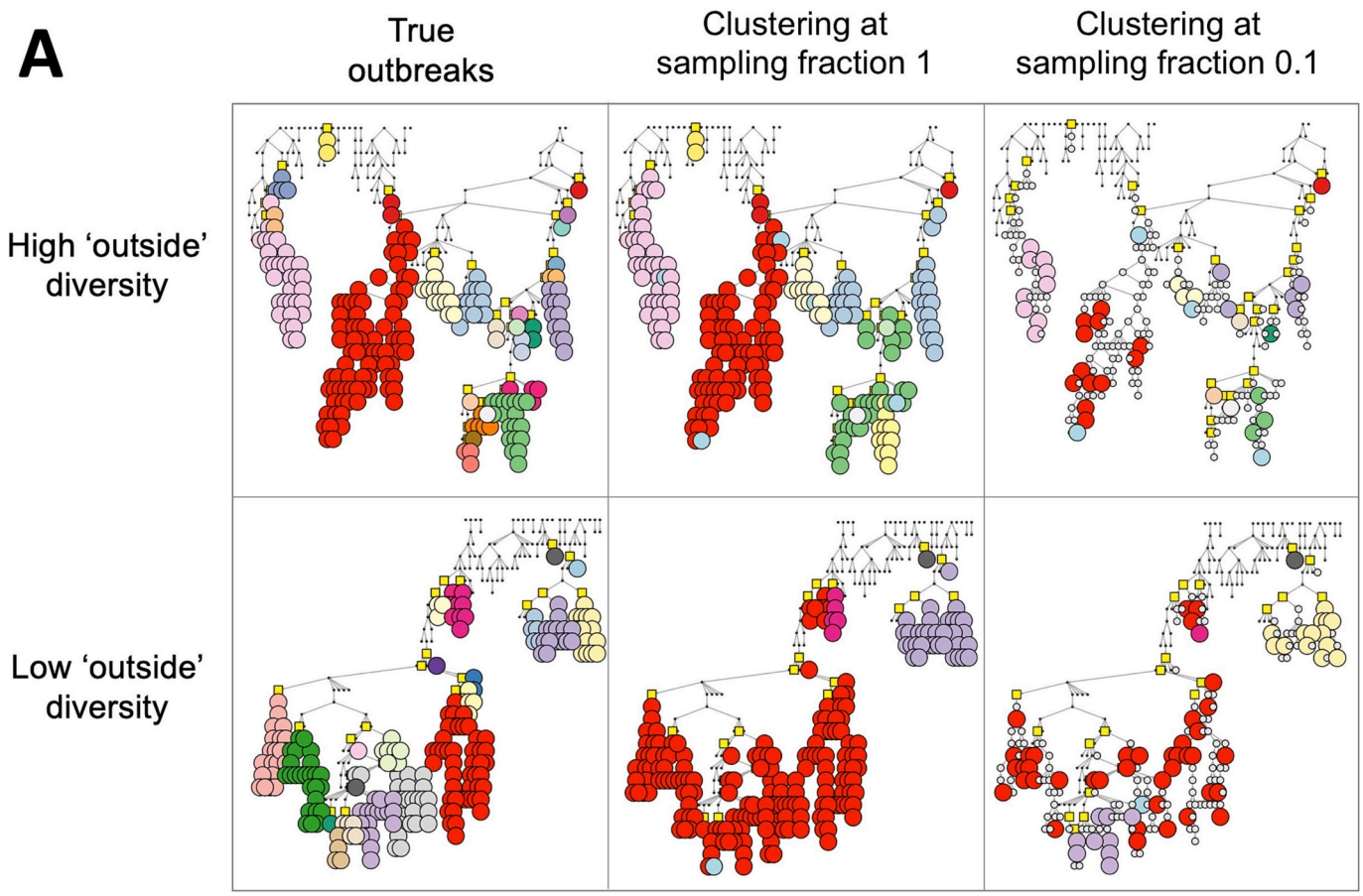


Fig. 4. Graph plots of clustering in example simulated outbreaks 'inside' a sampled jurisdiction seeded from introduced infected cases from an 'outside' population with high and low genomic diversity and a high introduction rate of 0.5. Simulations were run with A) low and B) high infection rate 'inside' ($R = 1.25$ and $R = 3$), and clustering at sampling fraction of 1 and 0.25 was compared to true outbreaks. Nodes are coloured by genomic clusters, with nodes belonging to the largest cluster coloured red and non-clustered nodes coloured light blue. All other colours represent smaller clusters. Unsampled cases are small, grey nodes. C) The mean proportion of sequences that were assigned as un-clustered, D) the mean proportion of sequences assigned to the largest cluster, and E) the number of clusters predicted per introduced outbreak (cases introduced into the inside population that transmitted to at least one susceptible individual), in 100 replicate outbreaks. Final outbreaks are simulated to be between 200 and 300 infected individuals. Final outbreaks are simulated to be between 200 and 300 infected individuals in the 'inside' population, and 100 replicates per outbreak scenario to calculate clustering properties. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Discussion

Here we used simulations to explore clustering using pathogen whole genome sequence data under various epidemiological and sampling scenarios in the context of SARS-CoV-2. The behaviour and apparent utility of clustering as a tool in genomic epidemiology depends on the contextual factors we have explored: the epidemiology of the pathogen, its genomic diversity, the sampling level, and the relationship between the jurisdiction in which sampling and analysis is done and the rest of the world. These exhibit a complex interplay that acts to shape the diversity, cluster sizes, and the extent to which the apparent cluster structure mirrors the true population structure.

For genomic clustering to be of most use for epidemiological investigation, cluster sizes would be small to intermediate, and each cluster would not represent a high portion of the total number of sequences, unless the number of sequences is small. Additionally, predicted clusters should reflect the true population structure, with clusters delineated into outbreaks or transmission clusters seeded from different index cases or that are significantly different in time and geography. We have found that the underlying epidemiology and evolution of the pathogen, along with the proportion of cases that were sampled and sequenced, can impact whether the clusters we identify are reflective of the disease population. Several mechanisms lead to clustering that does not reflect the population structure; spurious break-up of clusters due to missing intermediate cases (affecting lower-transmission contexts most) and spuriously joined clusters (significant in high-transmission contexts, especially when global diversity is low). In addition, it can be the case that there is no underlying clustering structure to find and there genuinely is one very large cluster.

The key phenomena that we observed in British Columbia during the COVID-19 pandemic are mirrored here: rapid transmission, particularly of a new variant such as P1 (Gamma), resulted in low genetic diversity per transmission event, and therefore, very large clusters. These large clusters likely include multiple true outbreaks comprising different introduction events and transmission settings. This complicated the ability to identify clusters at a resolution that would be helpful for epidemiological investigation. The "one very large cluster" problem may be worse than it appears, because with incomplete sampling, some links to a cluster are lost, and clusters can be spuriously broken up. Conversely, seemingly more useful, smaller clusters can be found in contexts when they do not reflect the true population structure. In this case, it may not be appropriate to use clusters and their membership as a basis for onward actions. In Table 1, we present some recommendations for when clustering may be useful under different epidemiological and sampling scenarios.

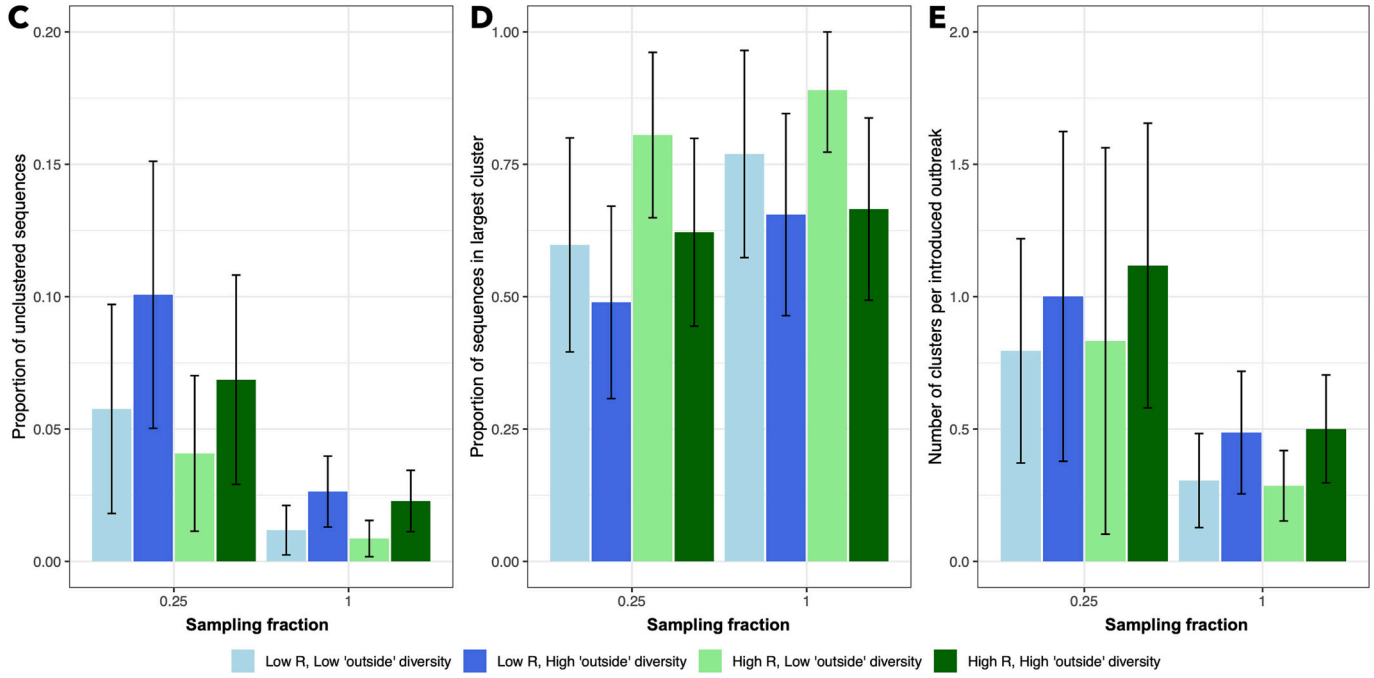


Fig. 4. (*continued*).

In our simulations we consider differences in R and sampling but keep a fixed mutation rate across our population through time, which may not be reflective of the true evolutionary dynamics. There is some evidence of variation in the non-synonymous mutation rate through time (Neher, 2022), though the overall SARS-CoV-2 mutation rate is similar between VOCs (Markov et al., 2023). Clustering patterns may be affected by mutation rate differences between strains (e.g., differences in mutation rates between TB lineages (Ford et al., 2013)) or where there may be heterogeneity in the mutation rate (e.g., adaptive response to environmental pressure (Matic, 2019). As genomic clustering requires either a distance or probability threshold, differences in mutation rate across the tested population may necessitate some informed decision on how to set appropriate thresholds in these scenarios or the testing of multiple thresholds. A limitation of our model is that we set a single value for R and sampling fraction throughout the outbreak simulations, which is not reflective of the true epidemiology and sampling through the COVID-19 pandemic. Future work to extend the model and investigate dynamic scenarios would allow for further exploration of clustering in different epidemiological contexts, including periods where R is lower than 1 and the disease is declining.

While we have simulated outbreaks based on the epidemiology of SARS-CoV-2, the findings presented here could be interpreted in respect to outbreaks of other pathogens that primarily spread through human-to-human transmission. Clustering from genomic sequence data has been used extensively to characterize tuberculosis (TB) outbreaks and detect signatures of recent transmission between cases (Nikolayevskyy et al., 2019). The long evolutionary history of TB has given rise to divergent lineages and sub-lineages (Napier et al., 2020), which can result in outbreaks of distinct genomically identical or near-identical clusters in periods of high transmission (Casali et al., 2016). This mirrors our simulations where inside outbreaks were seeded by diverse introductions, which allowed for clustering to capture the true population structure. Similarly, some viral pathogens, such as human immunodeficiency virus (HIV), evolve rapidly and so clustering can be used to detect periods of elevated transmission where lower diversity clusters are detected compared to the expected divergence over time (Chato et al., 2022), again reflecting our simulations with high outside diversity and high R inside the sampled jurisdiction. In contrast, factors such as a low mutation rate and population bottlenecks can reduce the diversity in some pathogens (e.g., human herpesviruses (López-Muñoz et al., 2021) and mpox (Isidro et al., 2022)). Thus, clustering may not capture the true population structure, and spurious clustering may occur where the proportion of sampled cases is low.

Different pathogens will have different molecular clock rates and levels of diversity as measured in SNP distances (or SNPs per site in the genome); mpox, for example, has low genetic diversity for its genome length, and low transmission. For our purposes, “low diversity” has meant that two outbreaks that one might wish to distinguish with genomic tools are similar enough to each other that there is a reasonable probability that they would be grouped together as one cluster. This of course depends on the cluster threshold, but also on whether the global - local epidemiology is such that multiple outbreaks are drawn from a very similar pathogen pool, and on whether the diversification rate and bioinformatic ability to detect variation is sufficient. It also depends on the level of resolution at which public health investigators need to distinguish outbreaks: it may be very feasible to classify an outbreak by pathogen lineage or clade, but not feasible to distinguish introductions from a nearby jurisdiction where the same lineages are spreading. On a similar note, the sampling fraction will vary over time and by pathogen, and the number of unknown cases may be uncertain.

5. Conclusion

We have shown how epidemiological and evolutionary factors, along with the proportion of cases sampled, can influence how representative genomic clustering patterns are of the true population structure. Large clusters with little utility for informing epidemiological investigation of disease outbreaks can result from populations experiencing high transmission rates where pathogen diversity is relatively low, as demonstrated here with simulations of SARS-CoV-2 transmission clusters. Furthermore, the spurious breakup of clusters into smaller groups that do not reflect true differences in the underlying population structure can occur when the transmission rate is low, and cases are missed from incomplete sampling. In contrast, when standing diversity is high, distinct outbreaks can be effectively identified using genomic clustering, and these clusters are particularly robust to changes in the sampling

Table 1
Recommended guidelines and actions given the likely clustering patterns identified under a range of epidemiological and sampling scenarios.

Transmission inside jurisdiction (Example figure)	Outside diversity and introductions	Sampling	Clustering properties	Does clustering appear to be useful?	Does the clustering reflect the true population structure?	Action
Low general transmission (Fig. 3A, top row)	Single jurisdiction (one introduction)	High	One large cluster.	No: one large cluster	Yes	Use methods incorporating data on individuals or exposure sites, not just clusters.
		Low	Multiple distinct clusters of high apparent utility. Spurious break-up: some apparently unclustered cases should be clustered, but links are missing.	Yes	No	Additional case-finding and sampling where warranted. Explore impact of cluster thresholds and sampling
High general transmission (Fig. 3A, bottom row)	Single jurisdiction (one introduction)	High	Very large clusters	No	Yes	Use methods incorporating data on individuals or exposure sites, not just clusters.
		Low	Large clusters persist due to low diversity per transmission event	No	Yes	
Low transmission in sampled jurisdiction (Fig. 4A, top row)	High global pool of diversity, multiple introductions.	High	Distinct clusters for distinct (and diverse) introductions; some spurious break-up of clusters	Yes	Yes	Proceed with clustering using genomic data.
		Low	Distinct clusters for distinct introductions, more spurious break-up	Yes	Mostly	Additional case-finding and sampling where warranted. Explore impact of cluster thresholds and sampling
Low transmission in sampled jurisdiction (Fig. 4A, bottom row)	Low global pool of diversity, multiple introductions.	High	Large clusters result from erroneous merging of distinct introductions. Some distinct clusters are still detectable.	Partly	Partly	Use methods incorporating data on individuals or exposure sites, not just clusters.
		Low	Two things can go wrong: (1) large clusters due to erroneous merging (low-diversity introduction) and (2) spurious cluster break-up.	Yes	No	Determine whether clusters reflect transmission links, or spurious joining and breakup in combination: combine using additional data with additional case-finding and impact of cluster thresholds.
High transmission in sampled jurisdiction (Fig. 4B, top row) E.g., ongoing unsampled community transmission, sampling in high-transmission settings.[a]	High global pool of diversity, multiple introductions.	High	Multiple distinct clusters associated with distinct introductions. Community level: there may be very large clusters due to rapid transmission in the jurisdiction.	Yes	Yes	Proceed with clustering using genomic data.Use methods incorporating data on individuals or exposure sites, not just clusters.
		Low	Multiple clusters associated with distinct introductions; reduced sampling does not sufficiently break up large clusters.	Yes	Yes	Additional case-finding and sampling where warranted. Explore impact of cluster thresholds and sampling
High transmission in sampled jurisdiction[b] (Fig. 4B, bottom row)	Low global pool of diversity, multiple introductions.	High	Likely to see large clusters.	No	No	Use methods incorporating data on individuals or exposure sites, not just clusters.
		Low	Low global diversity can lead to erroneous merging of distinct introductions	No	No	

[a] This scenario represents the COVID-19 in British Columbia when the situation has moved towards the endemic phase, with low community sampling, a diverse circulating viral pool, and higher sampling in high-transmission, high-risk settings.
[b] This scenario reflects the experience with P.1 and AY.25, which had low (local and global) diversity, rapid transmission and very large clusters that were not divided under a different choice of threshold.

fraction and clustering thresholds when the transmission rate is high. These findings, along with our recommendations, can provide guidelines for carrying out genomic clustering and interpreting the results in the context of the true epidemiology under different scenarios.

Statement of funding

This work was supported by funding from Michael Smith Foundation for Health Research and the Federal Government of Canada's Canada 150 Research Chair program. This work was supported by the Canadian Covid genomics network (CanCoGen).

Ethical approval

This research was conducted in accordance with the Declaration of Helsinki and ethical approval for the study was obtained from the University of British Columbia Ethics Board (#H20–02285). Informed consent to participate was not required as patient data were collected under a surveillance mandate, authorized by the British Columbia Provincial Health Officer under the Public Health Act.

CRediT authorship contribution statement

Benjamin Sobkowiak: Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. **Pouya Haghmaram:** Methodology. **Natalie Prystajecky:** Conceptualization, Writing – review & editing. **James E.A. Zlosnik:** Data curation, Writing – review & editing. **John Tyson:** Data curation, Writing – review & editing. **Linda M.N. Hoang:** Writing – review & editing. **Caroline Colijn:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare they have no competing interests.

Data availability

Whole genome sequence data included in this study are deposited in the GISAID repository https://www.gisaid.org. All code used in this study is available at: https://github.com/bensobkowiak/Clustering_simulations.

Acknowledgements

We would like to acknowledge the work of the Public Health Laboratory at the British Columbia Centre for Disease Control (BCCDC), for providing the SARS-CoV-2 isolates used in this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2023.105484.

References

Bendall, Emily E., Paz-Bailey, Gabriela, Santiago, Gilberto A., Porucznik, Christina A., Stanford, Joseph B., Stockwell, Melissa S., Duque, Jazmin, et al., 2022. SARS-CoV-2 genomic diversity in households highlights the challenges of sequence-based transmission inference. mSphere 7 (6), e0040022.

Campbell, Finlay, Strang, Camilla, Ferguson, Neil, Cori, Anne, Jombart, Thibaut, 2018. When are pathogen genome sequences informative of transmission events? PLoS Pathog. 14 (2), e1006885.

Casali, Nicola, Broda, Agnieszka, Harris, Simon R., Parkhill, Julian, Brown, Timothy, Drobniewski, Francis, 2016. Whole genome sequence analysis of a large isoniazid-resistant tuberculosis outbreak in London: A retrospective observational study. PLoS Med. 13 (10), e1002137.

Chato, Connor, Feng, Yi, Ruan, Yuhua, Xing, Hui, Herbeck, Joshua, Kalish, Marcia, Poon, Art F.Y., 2022. Optimized phylogenetic clustering of HIV-1 sequence data for public health applications. PLoS Comput. Biol. 18 (11), e1010745.

COVID-19 Immunity Task Force, 2021. URL: https://www.covid19immunitytaskforce.ca/seroprevalence-in-canada/#:~:text=Western%20Canada's%20estimated%20mean%20seropositivity,78.7%20to%2084.8)%20in%20Alberta (Accessed 19th July 2023).

Csardi, Gabor, Nepusz, Tamas, 2006. The igraph software package for complex network research. InterJournal, complex systems 1695 (5), 1–9.

Davies, Nicholas G., Abbott, Sam, Barnard, Rosanna C., Jarvis, Christopher I., Kucharski, Adam J., Munday, James D., Pearson, Carl A.B., et al., 2021. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science 372 (6538). https://doi.org/10.1126/science.abg3055.

Ford, Christopher B., Shah, Rupal R., Maeda, Midori Kato, Gagneux, Sebastien, Murray, Megan B., Cohen, Ted, Johnston, James C., Gardy, Jennifer, Lipsitch, Marc, Fortune, Sarah M., 2013. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nat. Genet. 45 (7), 784–790.

Geoghegan, Jemma L., Ren, Xiaoyun, Storey, Matthew, Hadfield, James, Jelley, Lauren, Jefferies, Sarah, Sherwood, Jill, et al., 2020. Genomic epidemiology reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand. Nat. Commun. 11 (1), 6351.

Glynn, J.R., Vynnycky, E., Fine, P.E., 1999. Influence of sampling on estimates of clustering and recent transmission of mycobacterium tuberculosis derived from DNA fingerprinting techniques. Am. J. Epidemiol. 149 (4), 366–371.

Government of British Columbia, 2021. URL: https://news.gov.bc.ca/releases/2021PREM0041-001155 (Accessed 15th February 2023).

Grabowski, M.K., Herbeck, J.T., Poon, A.F.Y., 2018. Genetic cluster analysis for HIV prevention. Curr HIV/AIDS Rep. 15 (2), 182–189. https://doi.org/10.1007/s11904-018-0384-1.

Isidro, Joana, Borges, Vítor, Pinto, Miguel, Sobral, Daniel, Santos, João Dourado, Nunes, Alexandra, Mixão, Verónica, et al., 2022. Addendum: Phylogenomic characterization and signs of microevolution in the 2022 multi-country outbreak of Monkeypox virus. Nat. Med. 28 (10), 2220–2221.

Kirola, L., 2021. Genetic emergence of B.1.617.2 in COVID-19. New Microbes and New Infections 43 (September), 100929.

Li, Yinhu, Jiang, Yiqi, Li, Zhengtu, Yonghan, Yu, Chen, Jiaxing, Jia, Wenlong, Ng, Yen Kaow, Ye, Feng, Li, Shuai Cheng, Shen, Bairong, 2022. Both simulation and sequencing data reveal coinfections with multiple SARS-CoV-2 variants in the COVID-19 pandemic. Computational and Structural Biotechnology Journal 20, 1389.

López-Muñoz, Alberto Domingo, Rastrojo, Alberto, Martín, Rocío, Alcamí, Antonio, 2021. Herpes simplex virus 2 (HSV-2) evolves faster in cell culture than HSV-1 by generating greater genetic diversity. PLoS Pathog. 17 (8), e1009541.

Markov, Peter V., Ghafari, Mahan, Beer, Martin, Lythgoe, Katrina, Simmonds, Peter, Stilianakis, Nikolaos I., Katzourakis, Aris, 2023. The evolution of SARS-CoV-2. Nat. Rev. Microbiol. April, 1–19.

Matic, Ivan, 2019. Mutation rate heterogeneity increases odds of survival in unpredictable environments. Mol. Cell 75 (3), 421–425.

Murall, Carmen Lia, Poujol, Raphael, Petkau, Aaron, Sobkowiak, Benjamin, Zetner, Adrian, Susanne A. Kraemer, et al., 2021. Monitoring the evolution and spread of Delta sublineages AY.25 and AY.27 in Canada. URL. https://virological.org/t/monitoring-the-evolution-and-spread-of-delta-sublineages-ay-25-and-ay-27-in-canada/767 (Accessed 15th February 2023).

Napier, Gary, Campino, Susana, Merid, Yared, Abebe, Markos, Woldeamanuel, Yimtubezinash, Aseffa, Abraham, Hibberd, Martin L., Phelan, Jody, Clark, Taane G., 2020. Robust barcoding and identification of mycobacterium tuberculosis lineages for epidemiological and clinical studies. Genome Medicine 12 (1), 114.

Neher, Richard A., 2022. Contributions of adaptation and purifying selection to SARS-CoV-2 evolution. Virus Evolution 8 (2), veac113.

Nikolayevskyy, V., Niemann, S., Anthony, R., van Soolingen, D., Tagliani, E., Ködmön, C., van der Werf, M.J., Cirillo, D.M., 2019. Role and value of whole genome sequencing in studying tuberculosis transmission. Clinical Microbiology and Infection 25 (11), 1377–1382.

Poon, Art F.Y., Gustafson, Réka, Daly, Patricia, Laura Zerr, S., Demlow, Ellen, Wong, Jason, Woods, Conan K., et al., 2016. Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. The Lancet. HIV 3 (5), e231–e238.

Rambaut, Andrew, Holmes, Edward C., O'Toole, Áine, Hill, Verity, McCrone, John T., Ruis, Christopher, du Plessis, Louis, Pybus, Oliver G., 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. 5 (11), 1403–1407.

Seemann, Torsten, Lane, Courtney R., Sherry, Norelle L., Duchene, Sebastian, Gonçalves, Anders, da Silva, Leon, Caly, Michelle Sait, et al., 2020. Tracking the COVID-19 pandemic in Australia using genomics. Nat. Commun. 11 (1), 4376.

Shu, Y., McCauley, J., 2017. GISAID: global initiative on sharing all influenza data – from vision to reality. Eurosurveillance 22, 2–4.

Sobkowiak, Benjamin, Kamelian, Kimia, Zlosnik, James E.A., Tyson, John, Gonçalves, Anders, da Silva, Linda M.N., Hoang, Natalie Prystajecky, Colijn, Caroline, 2022. Cov2clusters: genomic clustering of SARS-CoV-2 sequences. BMC Genomics 23 (1), 710.

Stefanelli, Paola, Trentini, Filippo, Guzzetta, Giorgio, Marziano, Valentina, Mammone, Alessia, Schepisi, Monica Sane, Poletti, Piero, et al., 2022. Co-circulation of SARS-CoV-2 alpha and gamma variants in Italy, February and March 2021. European Communicable Disease Bulletin 27 (5). https://doi.org/10.2807/1560-7917.ES.2022.27.5.2100429.

Stimson, James, Gardy, Jennifer, Mathema, Barun, Crudu, Valeriu, Cohen, Ted, Colijn, Caroline, 2019. Beyond the SNP threshold: identifying outbreak clusters using inferred transmissions. Mol. Biol. Evol. 36 (3), 587–603.

Worby, Colin J., Read, Timothy D., 2015. 'SEEDY' (simulation of evolutionary and epidemiological dynamics): an R package to follow accumulation of within-host mutation in pathogens. PLoS One 10 (6), e0129745.