Public health in genetic spaces: a statistical framework to optimize cluster-based outbreak detection

Connor Chato[1] **and Art F. Y. Poon**[1,2,3]

[1]Department of Pathology and Laboratory Medicine, Western University, London, Canada;

[2]Department of Applied Mathematics Medicine, Western University, London, Canada;

[3]Department of Microbiology and Immunology, Western University, London, Canada

Abstract

In infectious disease epidemiology, clustering cases of infection in space and time is a standard method to identify and characterize outbreaks. Clustering cases by genetic similarity is analogous to spatial clustering, and may be more effective for pathogens transmitted at a relatively low rate by intimate contact. However, the statistical properties of genetic clustering in the context of outbreak detection are not well characterized and cluster-defining criteria are generally set to arbitrary values. We describe a new method to optimize the predictive value of a clustering method by optimizing its parameters to maximize the difference in the Akaike information criterion (AIC) between individual-weighted and null models of cluster growth. This approach mirrors solutions to the modifiable areal unit problem (MAUP): the statistical association between covariates and an outcome is contingent on how their spatial distribution is partitioned into units of observation. To evaluate our method, we analyzed the distributions of pairwise Tamura-Nei (TN93) genetic distances from two published sets of anonymized HIV-1 subtype B *pol* sequence data stratified by collection year. We generated 46 different graphs by varying the pairwise threshold, where an edge in a graph indicates that the TN93 distance between the respective cases is below the corresponding threshold. For each graph, we generated predictions of cluster growth (numbers of new cases with edges to clusters of known cases) under two different Poisson regression models: a null model in which growth is only proportional to cluster size (*i.e.*, no variation among individuals);

and a weighted model where the variation associated with individual-level covariates are summed by cluster. Next, we calculated the AIC for each model on the distributions of observed cluster growth in two published HIV-1 *pol* data sets from Seattle, USA ($n = 1,653$) and Alberta, Canada ($n = 809$). Based on the difference in AICs, we obtained different optimized TN93 thresholds for these data sets (0.014 and 0.011, respectively). We show that selection of this threshold parameter can substantially limit the utility of genetic clusters for public health, and that the optimal parameter for one population can misdirect prevention efforts in another. This statistical framework can potentially be used to optimize any clustering method, and to evaluate it against other methods including those that do not use genetic information.

Introduction

Spatiotemporal clustering is a fundamental public health methodology for the detection of disease outbreaks [1]. The colocalization of cases in space and time implies a common source. For example, an automated space-time clustering method [2] was demonstrated to retrospectively detect outbreaks of nosocomial bacterial infection in a US-based hospital, including the outbreaks that were detected contemporaneously by the hospital's pre-existing infection control program [3]. At a broader spatial scale, the same clustering method was recently used to identify outbreaks of severe acute respiratory infections over a five year period, using case data from a network of hospitals in Uganda [4]. Early detection of a cluster represents a potential opportunity for a targeted public health response to prevent additional cases. Space-time clustering may be less effective, however, for pathogens that can establish a chronic infection with a long asymptomatic period (*e.g.*, *Mycobacterium tuberculosis*, hepatitis C virus, or human immunodeficiency virus type 1; HIV-1) where the precise time and location of a transmission event is usually unknown and difficult to reconstruct. Furthermore, pathogens with a low per-act transmission rate present difficulties for space-time clustering because a single exposure in a specific location is unlikely to result in transmission. Under these circumstances, the spread of an epidemic is more likely to be shaped by a network of repeated contacts between individuals, rather than shared venues.

For many infectious diseases, the molecular evolution of the pathogen is sufficiently rapid that genetic differences can accumulate between related infections on a similar time scale as disease transmission. Consequently, it can be effective to cluster cases in a high-dimensional *genetic space* in addition to clustering in physical space and time. In these studies, a case of infection is represented by a pathogen-derived molecular sequence that maps to some point in genetic space, and it may be associated with subject-level metadata such as the diagnosis date or treatment history. Clustering infections by their evolutionary relatedness is a popular method to identify and characterize potential transmission 'hotspots'. For example, pairs of sequences can be clustered if the number of genetic differences between them falls below some threshold. The resulting clusters are often visualized as a network or undirected graph, where each vertex represents an individual case of infection, and each edge connecting vertices indicate that the sequences of the corresponding cases are within a threshold genetic distance of each other. Sampling a group of cases that are nearly genetically identical implies that they are related through an unknown number of recent and rapid transmission events. A substantial number of genetic clustering studies have focused on the molecular epidemiology of HIV-1 [5–8]. Under current global treatment and prevention guidelines [9], greater proportions of HIV cases are being diagnosed, and new diagnoses are more frequently screened for drug resistance by genetic sequencing prior to initiating antiretroviral treatment. As a result, public health organizations are beginning to use genetic clustering methods in 'near real-time' to identify ongoing HIV-1 outbreaks [10, 11], to reconstruct the risk factors and etiology, and to prioritize groups for prevention initiatives such as access to pre-exposure prophylaxis (PrEP) [12].

A significant and often overlooked challenge in the use of genetic clustering to identify potential outbreaks is that these methods usually require the specification of one or more clustering criteria [8, 13]. For instance, HIV-1 studies that employ pairwise genetic clustering tend to use a threshold of 1.5% expected nucleotide substitutions per site [6, 14–16], a measure that adjusts for the occurrence of multiple substitutions at the same nucleotide. In contrast, the United States Centers for Disease Control and Prevention (US-CDC) currently mandates a stricter pairwise distance threshold of 0.5% [17]. In some cases the selected threshold is informed by the expected divergence between HIV-1 sequences sampled longitudinally from the same patient [18, 19] — however, this empirical distribution can vary substantially with the extent of clinical follow-up. Population studies from other regions such as Botswana [20] and South Africa [21] have used substantially higher distance thresholds (≥4.5%) that vary among HIV-1 subtypes [22, 23]. Simulation-based studies [8, 20, 24] have demonstrated that clustering is highly sensitive to the sampled proportion of the infected population. Given the significant global disparities in HIV-1 prevalence and access to testing and treatment, it is exceedingly unlikely that a meaningful 'gold standard' clustering criterion can exist.

Here we propose that the most useful approach to select clustering criteria is to base this decision on our ability to predict where the next cases will occur. A high, permissive clustering threshold tends to result in a single cluster that comprises the majority of known cases. The next cases are proportionately more likely to connect to this large cluster, but its size will also average out the individual- and group-level attributes that are informative for predicting the next cases. Put another way, a single large cluster is not likely to confer a public health benefit as an alternative to working with the entire population database. Conversely, setting a low, strict clustering threshold results in a large number of small clusters. This increases the variation of attributes among clusters, resolving greater information, but the association of new cases with clusters also becomes increasingly stochastic. This trade-off is analogous to the modifiable areal unit problem (MAUP), a concept in spatial statistics first fully conceptualized by Openshaw and Taylor in 1979 [25]. Areal units are derived from a partition of a geographic range by drawing boundaries that separate households or neighbourhoods. The MAUP formally recognizes the inconsistency of statistical associations with changing boundaries. For example, aggregating units into larger spatial units, such as cities or countries, can prevent an investigator from detecting a strong association between water quality and gastrointestinal illness [26]. To address the MAUP in the context of genetic clustering and public health, we adapt an information criterion-based approach described by Nakaya [27] to select an aggregation level for count data, such that distribution of cases in genetic space is partitioned in a way that maximizes the information content of clusters for forecasting where the next cases will occur.

Methods

Data collection and processing

From the public GenBank database (https://www.ncbi.nlm.nih.gov/genbank), we obtained $n = 809$ anonymized HIV-1 *pol* sequences that were sampled in Northern Alberta, Canada, between 2007 and 2013 [28]; as well as $n = 1653$ sequences collected in Seattle, USA, between 2000 and 2013 [29]. Each data set was manually screened to remove all sequences corresponding to HIV-1 subtypes other than subtype B, and to remove repeated samples from the same individual. Given the relatively small number of sequences collected in 2013 for the Seattle dataset ($n = 35$, Figure 1), we excluded this year to maintain a consistent sampling rate. We retrieved the sample collection dates for each sequence by querying GenBank with the respective accession numbers and extracting this information from the XML stream returned from the server using the BioPython module [30] in Python. Next, we used an open-source implementation of the Tamura-Nei [31] genetic distance in C++ (TN93 version 1.0.6, https://github.com/veg/tn93) for each data-set to compute these distances between all pairs of sequences. We set the maximum reporting distance to 0.05 to limit the number of pairs written to the comma-separated values (CSV) output file, which excluded an additional 20 cases from the Seattle dataset and 6 from Northern Alberta. All other options for the TN93 analyses were set to the default values.

Defining clusters

For each data set, we imported the TN93 output from the CSV file into R and generated an undirected graph $G = (V,E)$ using the igraph package [32]. The set of vertices V represents the individual cases in the data set, each uniquely labeled with an anonymized subject identifier. Every edge $e(v,u) \in E$ between vertices v and u was weighted with the corresponding TN93 distance between their sequences, which we denote by $w(v,u)$. In practice, we store a larger range of distances as a densely connected graph makes it more efficient to obtain the spanning subgraph that results from filtering the edge weights by some threshold $w_{\max}$ — hence for a given threshold, $E = \{e(v,u) : w(v,u) \leq w_{\max}\}$. In addition, each vertex $v \in V$ carries an attribute $t(v)$, which represents the sample collection year of the corresponding sequence. We note that our framework is not limited to analyzing collection dates at the level of years and can applied to more precise time intervals, *e.g.*, quarters or months — however, years are most frequently the precision at which this sampling information is released into the public domain. The subset of sequences with the most recent collection year was specified as $U = \{v \in V : t(v) = t_{\max})$ such that the total number of new cases is $|U|$. In other words, sampling time cuts the graph G into disjoint vertex sets V^c and U such that $V^c \cup U = V$. Later it will be useful to refer to the subset of edges in E that connect a vertex in V^c and a vertex in U, which we denote as $E_U = \{e(v,u) : e \in E, v \in V^c, u \in U\}$.

Clusters were defined as the connected components within the set of vertices representing known cases, $V^c = \{v \in V : t(v) < t_{\max}\}$. A clustering method defines a partition on V^c into a set of clusters $\{C_1, C_2, \ldots, C_n\}$ such that $C_i \cap C_j = \varnothing$ for all $i \neq j$ and $1 \leq i, j \leq n$; and such that the union of all clusters recovers the entire set: $\bigcup_{i=1}^{n} C_i = V^c$. Any pair of vertices within the same cluster ($v, u \in C_i$) are connected by at least one sequence of edges (path), and any pair of vertices in different clusters are not connected by any path. Under this definition, a cluster can comprise a single known case. Note that this definition does not strictly require the existence of edges, which we use to represent genetic similarity, but can be adapted to any clustering method that defines a partition on the database of known cases.

Modeling growth

We define total cluster growth R as the number of vertices in U adjacent (connected by edge) to any vertex in V^c, where $R \leq |U|$. The number of new cases adjacent to a specific cluster $C_i \subset V^c$ is defined as:

$$R(C_i) = |\{u : u \in U,\ v \in C_i,\ e(v,u) \in E_U\}| \tag{1}$$

To resolve the event that a new vertex in U is adjacent to vertices in more than one cluster, we reduced the subset of edges between U and its compliment, maintaining only the edges with minimum weight per vertex in U. If more than one edge to a given vertex $u \in U$ had exactly the same minimum weight, then we selected one edge at random:

$$E'_U = \left\{e : e(v,u) \in E_U, v \notin U, v \in \text{argmin}^1_v w(v,u)\ \forall u \in U\right\} \tag{2}$$

where $\text{argmin}_x f$ returns the set of values x that minimize the function f, and the superscript ^{1} is used to indicate that only a single value is being returned.

We formulated two predictive models to generate estimates of growth for the i-th cluster C_i, which we denote by $\hat{R}(C_i)$ and $\hat{R}_0(C_i)$, respectively. $\hat{R}_0$ requires less information than $\hat{R}$ by postulating that each cluster is expected to grow in proportion to its current size, prior to the addition of new cases, as a fraction of the entire population of known cases:

$$\hat{R}_0(C_i) = \exp\left(\frac{|C_i|}{|V^c|}R\right). \tag{3}$$

For example, a cluster that comprises half of the known cases is predicted to accumulate half of new cases that are adjacent to any cluster. Thus $\hat{R}_0$ does not use any individual-level attributes to predict cluster growth — it is a naive model that assumes that the allocation of new cases in R (those adjacent to clusters) is not influenced by any characteristics of those clusters other than the 'space' they occupy. This model assumes that the observed numbers of edges connecting vertices in C_i to vertices in U is a Poisson-distributed outcome with mean $\hat{R}_0$.

$\hat{R}$ assigns an individual-level weight $\rho(v)$ to every vertex in $v \in C_i$, proposing that the growth of cluster C_i is proportional to the exponential of the combined weights of the vertices in C_i:

$$\hat{R}(C_i) = \exp\left(\alpha + \beta \sum_{v \in C_i} \rho(v)\right) \tag{4}$$

where α and β are quantities to be estimated by regression. We use an exponential function because the numbers of new cases are counts that will be modelled as outcomes of a Poisson log-linear model. The weight $\rho(v)$ of a given vertex in V^c is based on the expected rate of adjacency to cases in U for a known case that is Δt years behind the cases in U. We quantified this expected rate of adjacency by the densities of edges between two sets of cases from different time points separated by a time lag of Δt years. Specifically, the edge density for years i and j is calculated from the bipartite graph $K_{i,j}$ between two sets of vertices: $V_i = \{v \in V^c : t(v) = i\}$ and $V_j = \{v \in V^c : t(v) = j\}$. Note that no information about edges to vertices in the most recent time period, U, is being used to train our model (Eqn. 4), since we are reserving these cases to evaluate the effectiveness of this model. For a given Δt, there are a total of $(t_{\max} - t_{\min} - \Delta t)$ different bipartite graphs, where $t_{\max}$ is the time point prior to the most recent time. We refer to this subset of bipartite graphs as $K(\Delta t) = \{K_{i,j} : j > i, j - i = \Delta t\}$.

Next, we define the following indicator function for $v \in V_i$ and $u \in V_j$:

$$\mathbf{1}(v,u) = \begin{cases} 1 & \text{if } e(v,u) \in E \text{ and } v = \operatorname{argmin}^1_x w(x,u)\ \forall x \in V_i \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

where we remove edges in the bipartite graph until the edge with the minimum weight for each given u in V_j remains. Given this indicator function, we calculate the observed edge density for

$K_{i,j} = (V_i, V_j)$ as:

$$\rho(K_{i,j}) = \sum_{v \in V_i} \sum_{u \in V_j} \mathbf{1}(v,u) \tag{6}$$

The log likelihood for the observed edge densities for all bipartite graphs in the set $K(\Delta t)$ is obtained from the Bernoulli distribution:

$$\log L(\hat{\rho}|\Delta t) = \sum_{V_i, V_j \in K(\Delta t)} \left(\sum_{v \in V_i} \sum_{u \in V_j} \hat{\rho}(\Delta t)\mathbf{1}(v,u) + (1-\hat{\rho}(\Delta t))(1-\mathbf{1}(v,u)) \right) \tag{7}$$

where $\hat{\rho}(\Delta t)$ is the expected edge density for a given time lag Δt. Finally, we fit a binomial regression model to observed decay of edge densities with increasing Δt to estimate $\hat{\rho}$:

$$\log\left(\frac{\hat{\rho}}{1-\hat{\rho}}\right) = \alpha + \beta \Delta t \tag{8}$$

where the left hand side of the equation is the log odds of an edge from v to u over a lag Δt. Our model assumes that the number of vertices in U adjacent to vertices in C_i is a Poisson-distributed outcome with mean parameter $\lambda = \hat{R}(C_i)$, which is calculated by combining Equations 4 and 8.

For the following demonstration, we use the simplest model where the probability that a given vertex v is adjacent to a new case is dependent only on the age of that vertex (Eqn. 8). The weight of a given vertex is thus $\rho = \hat{\rho}\,(t(v))$. However, Equation 8 can be easily extended to incorporate M additional individual-level attributes (*e.g.*, plasma viral load) into a linear predictor of a vertex's adjacency to new cases:

$$\log\left(\frac{\hat{\rho}}{1-\hat{\rho}}\right) = \alpha + \beta_0\,(t(v) - t_{\max}) + \beta_1 a_1(v) + \ldots + \beta_M a_M(v) \tag{9}$$

where we have replaced the lag Δt defining bipartite graphs with the vertex-specific age relative to the most recent time point, $t_{\max}$, and $a_m(v)$ is a function that extracts the m-th attribute from the vertex v.

Evaluating cluster thresholds

For both data sets, we segregated all HIV-1 sequences that were sampled in the most recent year as new cases comprising the set U. Next, we extracted the observed cluster growth data $R(C_i)$ and edge densities ρ at 46 different cluster-defining distance thresholds, ranging from $d = 0.005$ to $d = 0.05$ in steps of 0.001. We extracted the graph for a given cutoff d by filtering edges in E so that $w(e) < d$. Next, we fit the generalized linear models described by Equations (3) and (8), which correspond to $\hat{R}_0$ and $\hat{R}$ respectively, to the observed growth distribution among clusters $R(C_i)$. The resulting Akaike information criterion (AIC) for each model was recorded as a measurement of fit [33]. Nakaya [27] describes a “generalized AIC” (GAIC) as the difference in AIC between models; applying this framework to our analysis, we obtain the following criterion:

$$\begin{aligned} \text{GAIC} &= \text{AIC}(\hat{R}) - \text{AIC}(\hat{R}_0) \\ &= 2(k - k_0) - 2\left(\log L(\hat{R}) - \log L(\hat{R}_0)\right) \end{aligned} \tag{10}$$

where k is the degrees of freedom. Cutoffs with a negative GAIC indicate that the weighted model $\hat{R}$ explains the data more effectively than the naive model $\hat{R}_0$.

Results

Study populations

A total of $n = 1,591$ and $n = 803$ HIV-1 sequences were obtained from published studies in Seattle [29] and Northern Alberta [28], respectively. The distributions of sample collection dates are summarized in Figure 1. Although the sampling frame was shorter for Northern Alberta, similar numbers of cases were sampled per year at both locations, averaging 122.4 and 114.7 per year for Seattle and Northern Alberta, respectively. The most recent years of sampling (2012 and 2013, respectively) were withheld as the sets of new cases (U) for all subsequent analyses. Coincidentally, the numbers of new cases were the same for each location ($|U| = 110$).

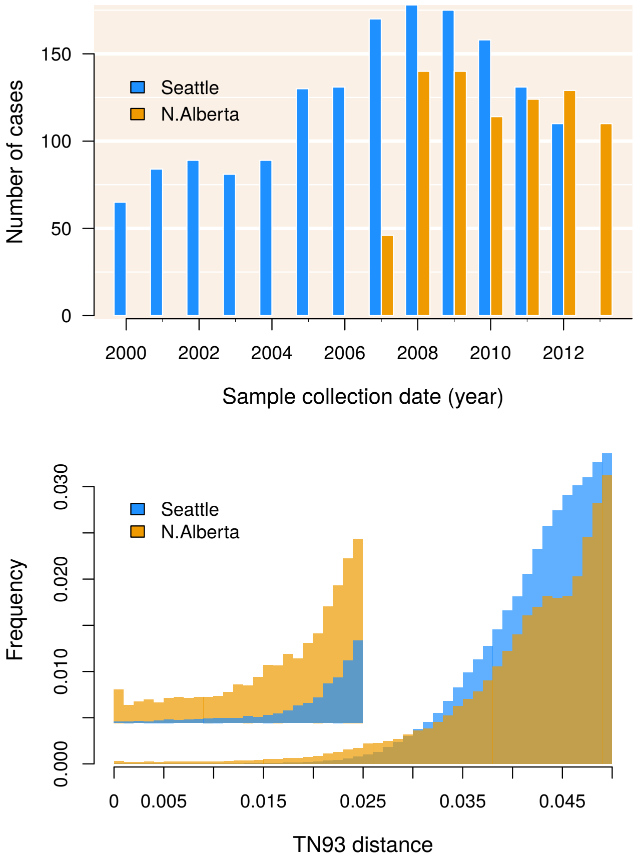

Figure 1: (top) Distribution of sample collection years for the Seattle (blue) and North Alberta (orange) data sets. Absent bars indicate that no sampling was carried out in the respective years, and does not reflect an absence of cases. (bottom) Histograms of Tamura-Nei (TN93) genetic distances among pairwise comparisons of HIV-1 sequences from Seattle and North Alberta. The height of each bin has been rescaled to reflect the total number of pairwise comparisons, for which the majority were censored from the data. A magnified section of the histograms is provided as an inset figure to clarify the distributions in the range (0, 0.025).

The truncated distributions of TN93 distances (below a limit of 0.05) were generally similar between the two locations (Figure 1). The filtered graphs from Seattle and North excluded 27 and 6 cases respectively as these cases would have been disconnected singletons even at the highest cutoff of 0.05. Although the Northern Alberta data set comprised a smaller number of sequences, the lower tail of its TN93 distribution contained relatively higher numbers of pairs than the Seattle data set (Figure 1). Similarly, we observed substantially lower frequencies of pairs with high

TN93 distances (≥ 0.03) in the Northern Alberta data. Thus, these distributions are significantly different (Kolmogorov-Smirnov test, $P < 10^{-4}$) with a slightly lower mean distance in Northern Alberta (0.0414) than Seattle (0.0426).

Adjacency of cases decays with time lag

We generated graphs from each HIV-1 data set by computing for every pair of sequences the Tamura-Nei (TN93) genetic distance, which adjusts for differences in the mean rates among nucleotide transversions and the two types of transitions [31]. This approach has been popularized by the software HIV-TRACE [34], which is employed by the U.S. Centers for Disease Control and Prevention [35]. In the following sections, we will refer to vertices in the graph derived from the pairwise genetic distances as *cases*. Vertices in the most recent year U will be referred to as new cases, whereas vertices in the rest of the graph V_c are known cases. Applying a cutoff $w_{\max}$ to select edges based on their weights (pairwise distances) yields a different partition of the known cases into clusters (connected network components). Two cases are *adjacent* when they are connected by an edge. A cluster may comprise a single known case. We expect that the probability of an edge between cases v and u should be influenced by the number of years that separate the respective sample collection dates (*e.g.*, Δt). To quantify this effect, we plotted the observed edge densities for bipartite graphs with a given lag Δt as a declining function of Δt for a distance cutoff $w_{\max} = 0.05$ (Figure 2). Fitting binomial regression models (Equation 8) to each data set obtained similar trends, despite the reduced number of bipartite graphs for the Northern Alberta data set due to a narrower range of sample collection dates. Specifically, the estimated effect of Δt on the log-odds of a bipartite edge was -0.35 (95% C.I. $= -0.38, -0.32$) year^{-1} for the Seattle data, and -0.45 $(-0.54, -0.37)$ year^{-1} for Northern Alberta. The coefficient of determination for the respective models was $R^2 = 0.48$ and 0.28. Given that the observed edge densities were low ($< 2\%$) for the range of distance cutoffs assessed here, the predicted trend could be well approximated by an exponential decay function, *i.e.*, Poisson approximation to the binomial distribution. Lowering the cutoff $w_{\max}$ reduced the observed bipartite edge densities as fewer edge weights passed the threshold. Nevertheless, the negative association between Δt and the log-odds of bipartite edges was robust to varying the cutoff (Supplementary Figure S1). This analysis supports the use of the difference in sample collection dates, or 'case recency' (Δt), as an individual-level predictor of a new case joining a cluster of known cases.

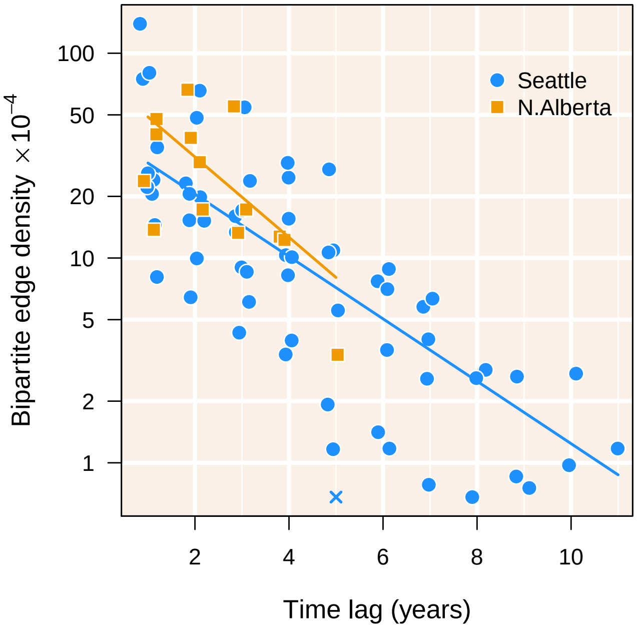

Figure 2: Decay in bipartite edge density with increasing time lag. Each point represents a bipartite graph at a given time lag (difference in years between time points, *x*-axis). We added random noise to the time lag associated with each point to make it easier to distinguish overlapping points. The *y*-axis represents the log-transformed bipartite edge density, *i.e.*, the frequency of edges between cases in different time points out of all possible edges (Equation 6). A × symbol is used to indicate the approximate position of an empty bipartite graph in the Seattle data set with a time lag of 5 years. Trend lines represent the predicted edge densities from the binomial regression described by Equation 8.

Trade-off between case coverage and cluster information

Figure 3 illustrates the effect of relaxing the cutoff $w_{\max}$ on the number of new cases in U that are adjacent to known cases in V^c, which we denote as R. When R approaches the total number of new cases, $|U|$, we say that the clusters have a high *case coverage*. As expected, decreasing the cutoff reduced R as a progressively greater number of edges between $v \in V^c$ and $u \in U$ were excluded by the clustering criterion $w(v,u) \leq w_{\max}$. The rate of decline in R with decreasing $w_{\max}$ was visibly faster for the Seattle data set than Northern Alberta (Figure 3), which was consistent with the greater proportion of short pairwise distances in the latter data set (Figure 1). These trends illustrate one of the expected disadvantages of applying a more stringent cutoff to a pairwise clustering analysis, as it limits the ability of the investigator to predict new cases.

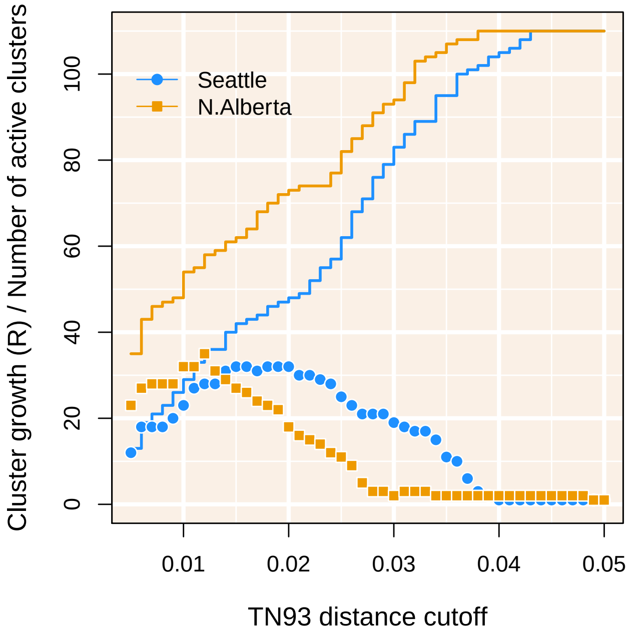

Figure 3: Distribution of new cases in among clusters as a function of distance cutoff. The solid lines represent the total number (R) of new cases in U adjacent to clusters of known cases in V^c. The points correspond to the number of clusters with edges to new cases, which we refer to as 'active' clusters.

In addition, Figure 3 summarizes how the number of clusters with at least one edge to U declines towards 1 with an increasing distance cutoff. We refer to any cluster C_i that meets this criterion ($R(C_i) > 0$) as an *active cluster*. At the upper limit of $w_{\max} = 0.05$, all new cases in U were connected to a single giant cluster of known cases. This outcome does not represent actionable information for public health because the giant cluster is indistinguishable from the entire population of known cases. Put another way, relaxing the cutoff to increase case coverage comes at the cost of lost predictive power. Decreasing the distance cutoff causes this giant cluster to be broken up into smaller clusters of which a subset have edges to new cases, such that the number of active clusters increases. As the cutoff continues to decline, however, the number of active clusters starts to decline again — despite their continued break-up into smaller clusters — because shorter cutoffs limit the adjacency of new cases to clusters. In other words, R sets an upper limit to the number of active clusters; these numbers can only be equal if each new case is uniquely adjacent to its own cluster, as we observed for the Seattle data set for cutoffs below $w_{\max} = 0.007$ (Figure 3).

Obtaining GAIC

The results in the preceding sections imply that there exists an intermediate distance cutoff that optimizes the trade-off between case coverage and the number of active clusters, both quantities having a significant impact on the information content of clusters for public health. We propose to use our ability to predict where the next cases will occur as the framework for evaluating distance cutoffs that define different partitions of the known case population into clusters. Specifically, we adapted the generalized Akaike information criterion (GAIC), which was developed by Nakaka [27] to select the granularity of districts in Tokyo, Japan, that best explained variation in death rates among elderly males in relation to socioeconomic and demographic factors. Our implementation of the GAIC is a comparison between two Poisson regression models, where the outcomes are the numbers of new cases adjacent to clusters, $R(C_i)$. In the first model, we assume that the expected number of new cases adjacent to the i-th cluster, $\hat{R}_0(C_i)$, is Poisson distributed with a rate proportional to the size of that cluster relative to the entire population of known cases. In other words, we would expect a large cluster to be adjacent to more new cases than a small cluster simply because it is large. In the second model, $\hat{R}(C_i)$ is Poisson distributed with a rate proportional to the total weight of known cases in the i-th cluster, where each weight is calculated from a linear combination of individual attributes including the sample collection date, *e.g.*, Δt.

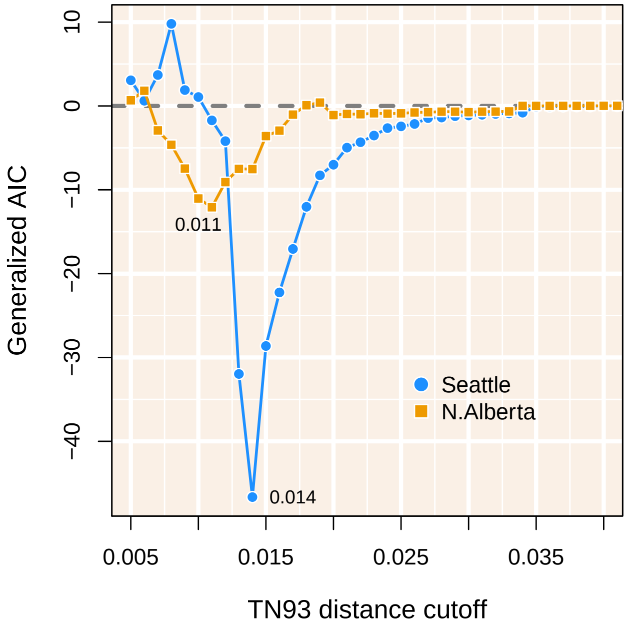

Figure 4: The GAIC relative to cuttoff d. This is calculated as the difference in AIC between $\bar{R}(c)$ fit to $R(c)$ minus $\bar{R}_0(c)$ fit to $R(c)$. The statistics from G_{na} are shown in Yellow and those taken from G_{st} are shown in blue. The highlighted minimums, where $\bar{R}(c)$ most out preforms $\bar{R}_0(c)$ in cluster by cluster growth prediction are shown by the dashed line. These minimums are found at cutoffs $d = 0.011$ and $d = 0.014$.

Figure 4 summarizes the distributions of GAIC for varying distance cutoffs for the Seattle and Northern Alberta data sets. We observed that GAIC tended to be near zero for relaxed cutoffs ($w_{\max} \geq 0.03$), which indicated that the ability of the weighted model, which incorporated individual-level information, to predict new cases was indistinguishable from the naive model. At these high cutoffs, the majority of known cases tended to become grouped into a single large cluster, thereby homogenizing any individual-level variation that could be used by the weighted model to predict the distribution of new cases. The GAIC values tended to decline monotonically with decreasing cutoffs, which split the known cases into progressively smaller clusters, until they reached the minimum values at $w_{\max} = 0.014$ for Seattle and $w_{\max} = 0.011$ for Northern Alberta. These minima identified the optimal distance cutoffs for the respective data sets where the weighted model was significantly better at predicting new cases. For cutoffs below these optimal values, we observed that the GAIC increased to positive values indicating that the weighted model was inferior to the naive model. Hence, lowering the cutoff $w_{\max}$ to these levels resulted in the dissolution of clusters into large numbers of singletons where our ability to predict new cases was overwhelmed by individual-level variation in Δt. In other words, the distribution of new cases among the small clusters defined by a low cutoff becomes more stochastic.

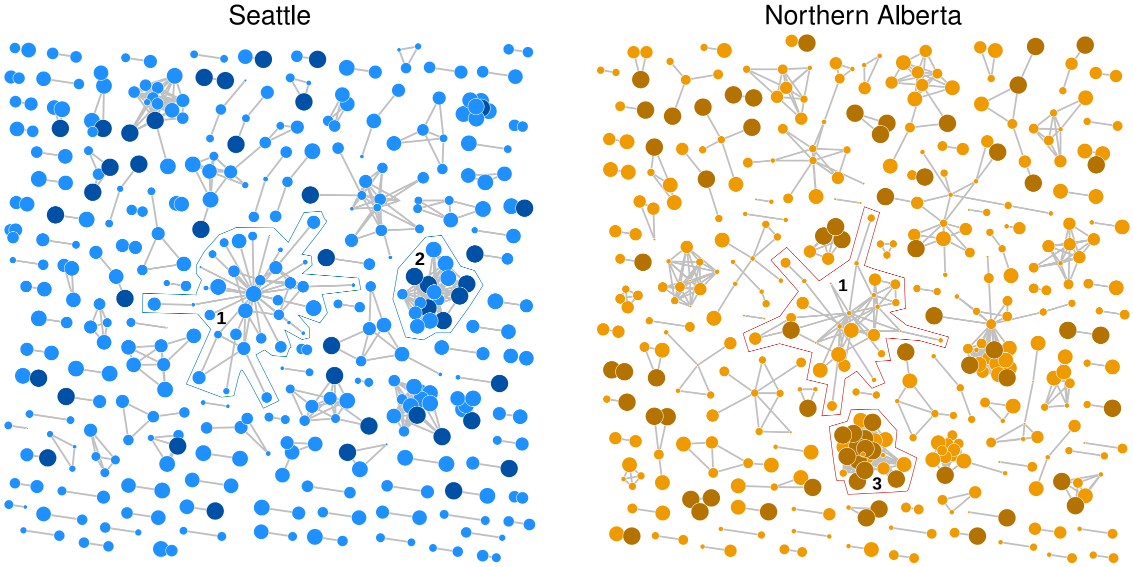

Figure 5: Visualizations of the graphs of the Seattle and Northern Alberta data obtained by the respective GAIC-optimized cutoffs. Vertices that corresponded to new cases were coloured in a darker shade, and the width of each vertex was scaled to the sample collection year with more recent cases drawn at a larger size. The vertices with zero degrees (singletons) were not drawn for clarity; the numbers of singletons were 1191 (Seattle) and 384 (N. Alberta), of which 68 and 42 were new cases respectively. Specific clusters discussed in the text were manually outlined and labeled with their rank according to size. The graph layouts were rendered using the implementation of the Kamada-Kawai [36] algorithm in Graphviz [37].

The graphs for these optimal cutoffs are summarized in Figure 5. In the Seattle graph, the largest cluster (1) comprising 35 known cases was not adjacent to any of the new cases. The sample collection years associated with this cluster range from 2000 to 2011 with a mean of 2006.0. In contrast, the second largest cluster (2) comprised 16 cases of which 6 are new, and the sampling years of known cases range from 2007 to 2011 with a mean of 2009.6. Similarly, the largest cluster in the Northern Alberta graph (1) comprised 27 cases of which only 1 was new, with sample collection years ranging from 2007 to 2013 with a mean of 2009.3. In the same graph the third largest cluster (3) of 10 known cases, of which 5 were collected in 2012, gained 12 new cases in 2013. These simple examples illustrate the effect of optimizing the clustering threshold on the covariation of new case counts and the recency of known cases among clusters. On the other hand, only about half of new cases were adjacent to clusters of known cases in either graph (67.3% for Seattle, 41.6% for Northern Alberta) given the optimized cutoffs.

Discussion

Our results demonstrate that an apparently small difference in pairwise genetic distances — for instance, between 0.5% and 1.5% — can make the difference between accurate forecasting of new cases among clusters and becoming misled by stochastic noise. Specifically, both cutoffs cited above are routinely used as customary settings in pairwise genetic distance clustering studies of the same HIV-1 subtype in the same country [16, 17, 29]. To investigate the sensitivity of clustering thresholds, we have examined two published data sets of anonymized HIV-1 subtype B *pol* sequences that were collected in different regions of North America within similar time frames. Both cutoffs are located in the left tails of the respective empirical distributions of pairwise distances, with no clear demarcation that might motivate the selection *a priori* of one cutoff over another (Figure 1). However, our information-based criterion reveals a stark difference between these cutoffs when we evaluate the ability of genetic clusters to ‘forecast’ the occurrence of new cases (Figure 4). This discordance is a result of a tradeoff between the coverage and predictive value of spatial information that is encapsulated by the modifiable areal unit problem (MAUP). As we relax the clustering threshold in our example, instances of cluster growth become more frequent such that aggregate effects (*viz.*, case recency) can be distinguished against the background of stochastic effects. At the same time, there is declining variation in growth rates among clusters, making it more difficult to detect associations between growth and the variation in covariates among clusters.

Unlike most instances of the MAUP that arise in spatial epidemiology, our outcome variable (the number of new cases per genetic cluster) is directly dependent on the same parameters that reshape the partition of the spatial distribution of covarates into units. This dual dependency results in an asymptotically increasing model likelihood with increasing distance thresholds, plateauing at the point where all known cases were assigned to the same giant cluster, such that any new cases are effectively guaranteed to be adjacent to this cluster. We addressed this unique problem by formulating a null model R_0, where the predicted growth of a cluster was directly proportional to its relative size in the number of known cases. Hence, R_0 provided a useful baseline that controlled for the proportionate effect of the largest cluster with increasing cutoffs, thereby enabling us to focus on the predictive value of variation in covariates among clusters.

We selected the simplest clustering method (pairwise TN93 distance clustering [15]) to demonstrate our new framework for evaluating clusters, based on Nakaya's generalized Akaike information criterion (GAIC) [27]. Despite the simplicity of pairwise clustering, it has been widely adopted in health jurisdictions around the world, including the U.S. Centers for Disease Control and Prevention [35], in part due to the growing popularity of the HIV-TRACE software package [34]. However, we emphasize that our framework can be used to evaluate any clustering method on the merits of its ability to forecast new cases. Put another way, any clustering criterion that changes the degree of connectivity can be optimized through this method. For example, if we require some minimum bootstrap support value to define clusters as subtrees extracted from the total phylogenetic tree [38], then this bootstrap support threshold can represent a second dimension to locate the minimal GAIC in combination with a distance threshold. Additionally, our framework enables us to evaluate clustering methods that do not use any genetic information. Thus we propose that an informative assessment on the potential value of genetic clustering for public health would be to compare the GAIC of the genetic clustering method against the GAIC obtained from the prioritization of groups by experts in public health, *e.g.*, medical health officers. However, the specialized and confidential information comprisign the latter case is unlikely to be found in the public domain.

Similarly, we can evaluate any linear combination of predictor variables, such as viral load or risk group, because our framework is based on a Poisson regression model. For the purpose of demonstrating the framework, we used only sample collection dates to derive a predictor variable (case recency). These dates in units of years are most frequently available as sample metadata in association with published genetic sequences. We had a strong *a priori* expectation for an association between new case adjacency and known case recency that we subsequently confirmed from these data (Figure 2). On the other hand, we are also aware that samples may be collected well after the start of a new infection, due to the long asymptomatic period of HIV-1 infection and social barriers to HIV testing [39]. Although dates of HIV diagnosis or estimated dates of infection, *e.g.*, the midpoint between the last HIV seronegative and first seropositive visit dates, will inevitably be closer to the actual date of infection, sample collection dates are substantially more readily available in the public domain. Furthermore, we recognize that more precise dates of sample collection would likely confer greater prediction accuracy. The granularity of time in the context of genetic cluster analysis represents another extension of MAUP, known as the modifiable temporal unit problem [40]. While reducing the length of time intervals may produce more timely predictions, *e.g.*, new cases in the next three months instead of the next year, the accuracy of prediction will erode with progressively shorter intervals.

Another caveat is that the expected probability of a specific edge between known and new cases is very small. Consequently, our framework requires a substantial number of new cases to parameterize models of the variation in edge densities among clusters and, ultimately, to discriminate between the null and weighted models. The results that we obtained with the smaller of the two data sets (N. Alberta) implies that averaging about 100 sampled cases per year over a 6 year period is adequate for the relatively simple models evaluated here, although this was almost certainly influenced by the relatively higher proportion of pairwise distances below 0.03 (Figure 1). Note that the number of cases sampled in a given year does not correspond to the annual incidence. In summary, we were able to infer the optimal cutoffs as distinct minima in the GAICs between the weighted and baseline models, despite using only the most rudimentary clustering method and a single covariate. This implies that employing more complex clustering methods (*e.g.*, [38, 41, 42]) and more extensive individual- and group-level covariates (*e.g.*, [7, 18, 43–45]) can identify increasingly lower minima, *i.e.*, more effective predictive models of cluster growth, provided adequate data. A model selection procedure for optimizing the combination of covariates in the context of predicting cluster growth was recently described by Billock and colleagues [6], for instance, for pairwise TN93 clusters at a prespecified cutoff of 1.5%.

Our method is not specific to HIV-1, although the proliferation of clustering methods in HIV molecular epidemiology — driven by the abundance of genetic sequence data and the relatively rapid rate of evolution and low transmission rate of the virus — does make this approach particularly applicable to HIV-1. Similar pairwise distance clustering methods, for instance, have been used for *Mycobacterium tuberculosis* [46] and hepatitis C virus [47] to infer epidemiological characteristics from molecular sequence variation. In these cases, it may be necessary to rescale time-frame for cluster growth or the step size/range of clustering thresholds to locate the minimum GAIC. Genetic clustering is used increasingly for near real-time monitoring of clinical populations for the purpose of guiding public health activities [5, 6, 10, 35]. Because of this, it has become very important which tools we use for phylodynamic analysis and how we use them. Inadequately calibrated clustering methods may result in outbreak false-positives, diverting limited public health resources away from subpopulations where the immediate need for prevention and treatment services was greatest. Improving the predictive potential of clustering techniques will be an important part of optimizing the public health response to HIV, as increased normalization and access to HIV testing [48], proper data anonymization and network security practices [49], and faster and more cost-effective resistance genotyping continue to lower barriers to detecting outbreaks in shorter time frames.

Acknowledgements

This study was supported by a Project Grant from the Canadian Institutes of Health Research (PJT-156178 to AFYP). CC was supported by a Dr. Frederick Winnett Luney Graduate Research Award from the Department of Pathology at Western University, Canada. An early version of this work was presented at the 26th International HIV Dynamics & Evolution meeting in Cascais, Portugal. We thank John R. J. Thompson for a helpful discussion on applications of spatial statistics to model forest fire risk, which motivated us to look beyond autoregressive models; and Dr. Tomoki Nakaya for encouraging feedback at an early stage of this study.

References

[1] Robertson C, Nelson TA, MacNab YC, Lawson AB. Review of methods for space–time disease surveillance. Spatial and spatio-temporal epidemiology. 2010;1(2-3):105–116.

[2] Kulldorff M, Heffernan R, Hartman J, Assunçao R, Mostashari F. A space–time permutation scan statistic for disease outbreak detection. PLoS medicine. 2005;2(3):e59.

[3] Huang SS, Yokoe DS, Stelling J, Placzek H, Kulldorff M, Kleinman K, et al. Automated detection of infectious disease outbreaks in hospitals: a retrospective cohort study. PLoS medicine. 2010;7(2):e1000238.

[4] Cummings MJ, Tokarz R, Bakamutumaho B, Kayiwa J, Byaruhanga T, Owor N, et al. Precision surveillance for viral respiratory pathogens: virome capture sequencing for the detection and genomic characterization of severe acute respiratory infection in Uganda. Clinical Infectious Diseases. 2018;68(7):1118–1125.

[5] Rose R, Lamers SL, Dollar JJ, Grabowski MK, Hodcroft EB, Ragonnet-Cronin M, et al. Identifying transmission clusters with Cluster Picker and HIV-TRACE. AIDS research and human retroviruses. 2017;33(3):211–218.

[6] Billock RM, Powers KA, Pasquale DK, Samoff E, Mobley VL, Miller WC, et al. Prediction of HIV Transmission Cluster Growth With Statewide Surveillance Data. Jaids Journal of Acquired Immune Deficiency Syndromes. 2019;80(2):152–159.

[7] Ragonnet-Cronin M, Jackson C, Bradley-Stewart A, Aitken C, McAuley A, Palmateer N, et al. Recent and Rapid Transmission of HIV Among People Who Inject Drugs in Scotland Revealed Through Phylogenetic Analysis. The Journal of infectious diseases. 2018;217(12):1875–1882.

[8] Poon AF. Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks. Virus evolution. 2016;2(2):vew031.

[9] Levi J, Raymond A, Pozniak A, Vernazza P, Kohler P, Hill A. Can the UNAIDS 90-90-90 target be achieved? A systematic analysis of national HIV treatment cascades. BMJ global health. 2016;1(2):e000010.

[10] Poon AF, Gustafson R, Daly P, Zerr L, Demlow SE, Wong J, et al. Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. The lancet HIV. 2016;3(5):e231–e238.

[11] Gonsalves GS, Crawford FW. Dynamics of the HIV outbreak and response in Scott County, IN, USA, 2011–15: a modelling study. The Lancet HIV. 2018;5(10):e569–e577.

[12] Volz EM, Le Vu S, Ratmann O, Tostevin A, Dunn D, Orkin C, et al. Molecular epidemiology of HIV-1 subtype B reveals heterogeneous transmission risk: Implications for intervention and control. The Journal of infectious diseases. 2018;217(10):1522–1529.

[13] Hassan AS, Pybus OG, Sanders EJ, Albert J, Esbjörnsson J. Defining HIV-1 transmission clusters based on sequence data. AIDS (London, England). 2017;31(9):1211.

[14] Little SJ, Pond SLK, Anderson CM, Young JA, Wertheim JO, Mehta SR, et al. Using HIV networks to inform real time prevention interventions. PloS one. 2014;9(6):e98443.

[15] Aldous JL, Pond SK, Poon A, Jain S, Qin H, Kahn JS, et al. Characterizing HIV transmission networks across the United States. Clinical Infectious Diseases. 2012;55(8):1135–1143.

[16] Oster AM, Wertheim JO, Hernandez AL, Ocfemia MCB, Saduvala N, Hall HI. Using molecular HIV surveillance data to understand transmission between subpopulations in the United States. Journal of acquired immune deficiency syndromes (1999). 2015;70(4):444.

[17] National Center fo HIV/AIDS, Viral Hepatitis, STD, and TB Prevention. Detecting and responding to HIV transmission clusters: a guide for health departments; 2018. https://www.cdc.gov/hiv/pdf/funding/announcements/ps18-1802/CDC-HIV-PS18-1802-AttachmentE-Detecting-Investigating-and-Responding-to-HIV-Transmission-Clusters.pdf.

[18] Poon AF, Joy JB, Woods CK, Shurgold S, Colley G, Brumme CJ, et al. The impact of clinical, demographic and risk factors on rates of HIV transmission: a population-based phylogenetic analysis in British Columbia, Canada. The Journal of infectious diseases. 2014;211(6):926–935.

[19] Wertheim JO, Pond SLK, Forgione LA, Mehta SR, Murrell B, Shah S, et al. Social and genetic networks of HIV-1 transmission in New York City. PLoS pathogens. 2017;13(1):e1006000.

[20] Novitsky V, Moyo S, Lei Q, DeGruttola V, Essex M. Impact of sampling density on the extent of HIV clustering. AIDS research and human retroviruses. 2014;30(12):1226–1235.

[21] De Oliveira T, Kharsany AB, Gräf T, Cawood C, Khanyile D, Grobler A, et al. Transmission networks and risk of HIV infection in KwaZulu-Natal, South Africa: a community-wide phylogenetic study. The lancet HIV. 2017;4(1):e41–e50.

[22] Santoro MM, Perno CF. HIV-1 genetic variability and clinical implications. ISRN microbiology. 2013;2013.

[23] Hemelaar J. Implications of HIV diversity for the HIV-1 pandemic. Journal of Infection. 2013;66(5):391–400.

[24] Dasgupta S, France AM, Brandt MG, Reuer J, Zhang T, Panneer N, et al. Estimating Effects of HIV Sequencing Data Completeness on Transmission Network Patterns and Detection of Growing HIV Transmission Clusters. AIDS research and human retroviruses. 2019;35(4):368–375.

[25] Openshaw S, Taylor P. Statistical applications in the spatial sciences, chapter A million or so correlation coefficients: three experiments on the modifiable areal unit problem. Wrigley N Publishers, London, Pion. 1979;p. 127–144.

[26] Swift A, Liu L, Uber J. Reducing MAUP bias of correlation statistics between water quality and GI illness. Computers, Environment and Urban Systems. 2008;32(2):134–148.

[27] Nakaya T. An information statistical approach to the modifiable areal unit problem in incidence rate maps. Environment and Planning A. 2000;32(1):91–109.

[28] Vrancken B, Adachi D, Benedet M, Singh A, Read R, Shafran S, et al. The multi-faceted dynamics of HIV-1 transmission in Northern Alberta: A combined analysis of virus genetic and public health data. Infection, Genetics and Evolution. 2017;52:100–105.

[29] Wolf E, Herbeck JT, Van Rompaey S, Kitahata M, Thomas K, Pepper G, et al. Phylogenetic evidence of HIV-1 transmission between adult and adolescent men who have sex with men. AIDS research and human retroviruses. 2017;33(4):318–322.

[30] Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422–1423.

[31] Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular biology and evolution. 1993;10(3):512–526.

[32] Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. 2006;Complex Systems:1695. Available from: http://igraph.org.

[33] Akaike H. Information theory and an extension of the maximum likelihood principle. In: Selected papers of hirotugu akaike. Springer; 1998. p. 199–213.

[34] Pond SLK, Weaver S, Leigh Brown AJ, Wertheim JO. HIV-TRACE (TRAnsmission Cluster Engine): a tool for large scale molecular epidemiology of HIV-1 and other rapidly evolving pathogens. Molecular biology and evolution. 2018;35(7):1812–1819.

[35] Oster AM, France AM, Panneer N, Ocfemia MCB, Campbell E, Dasgupta S, et al. Identifying clusters of recent and rapid HIV transmission through analysis of molecular surveillance data. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2018;79(5):543–550.

[36] Kamada T, Kawai S. An algorithm for drawing general undirected graphs. Information processing letters. 1989;31(1):7–15.

[37] Ellson J, Gansner E, Koutsofios L, North SC, Woodhull G. Graphviz—open source graph drawing tools. In: International Symposium on Graph Drawing. Springer; 2001. p. 483–484.

[38] Ragonnet-Cronin M, Hodcroft E, Hué S, Fearnhill E, Delpech V, Brown AJL, et al. Automated analysis of phylogenetic clusters. BMC bioinformatics. 2013;14(1):317.

[39] Mahajan AP, Sayles JN, Patel VA, Remien RH, Ortiz D, Szekeres G, et al. Stigma in the HIV/AIDS epidemic: a review of the literature and recommendations for the way forward. AIDS (London, England). 2008;22(Suppl 2):S67.

[40] Cheng T, Adepeju M. Modifiable temporal unit problem (MTUP) and its effect on space-time cluster detection. PLoS One. 2014;9(6):e100465.

[41] McCloskey RM, Poon AF. A model-based clustering method to detect infectious disease transmission outbreaks from sequence variation. PLoS computational biology. 2017;13(11):e1005868.

[42] Barido-Sottani J, Vaughan TG, Stadler T. Detection of HIV transmission clusters from phylogenetic trees using a multi-state birth–death model. Journal of The Royal Society Interface. 2018;15(146):20180512.

[43] Wertheim JO, Murrell B, Mehta SR, Forgione LA, Kosakovsky Pond SL, Smith DM, et al. Growth of HIV-1 Molecular Transmission Clusters in New York City. The Journal of infectious diseases. 2018;218(12):1943–1953.

[44] McVea DA, Liang RH, Joy JB, Harrigan R, Poon AF. A framework for predicting phylogenetic clusters at high-risk for growth [CROI Abstract 848]. Top Antivir Med. 2017;25(Suppl 1):360s–361s.

[45] Dennis AM, Volz E, Frost AMSD, Hossain M, Poon AF, Rebeiro PF, et al. HIV-1 transmission clustering and phylodynamics highlight the important role of young men who have sex with men. AIDS research and human retroviruses. 2018;34(10):879–888.

[46] Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. The Lancet infectious diseases. 2013;13(2):137–146.

[47] Jacka B, Applegate T, Krajden M, Olmstead A, Harrigan PR, Marshall BD, et al. Phylogenetic clustering of hepatitis C virus among people who inject drugs in Vancouver, Canada. Hepatology. 2014;60(5):1571–1580.

[48] Hull MW, Wu Z, Montaner JS. Optimizing the engagement of care cascade: a critical step to maximize the impact of HIV treatment as prevention. Current Opinion in HIV and AIDS. 2012;7(6):579–586.

[49] Mehta SR, Schairer C, Little S. Ethical issues in HIV phylogenetics and molecular epidemiology. Current Opinion in HIV and AIDS. 2019;14(3):221–226.

1 ## Supplementary Figures

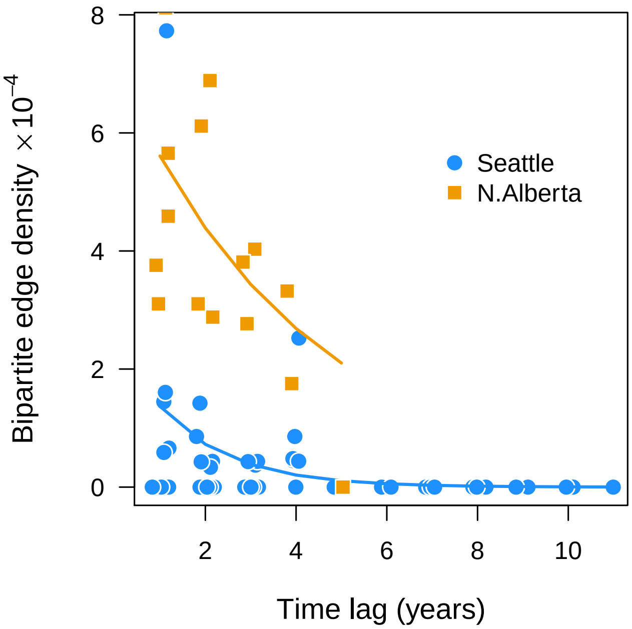

Figure S1: Decline of bipartite edge density with increasing time lag for $w_{\max} = 0.005$. Unlike Figure 2, the y-axis (bipartite edge density) is not log-transformed because of a substantial number of zero densities in the Seattle data set at this cutoff. The estimated effects of lag (Δt) on the log-odds of bipartite edges were -0.63 (95% C.I. $-0.89, -0.42$) and -0.25 $(-0.46, -0.05)$ for Seattle and Northern Alberta, respectively. The respective coefficients of determination (R^2) were 0.37 and 0.34. In sum, lowering the distance cutoff $w_{\max}$ reduced the observed edge densities but consistently maintained a negative association between the log-odds and Δt.