

# UC San Diego

## UC San Diego Previously Published Works

### Title

AUTO-TUNE: selecting the distance threshold for inferring HIV transmission clusters

### Permalink

<https://escholarship.org/uc/item/1nh876bz>

### Authors

Weaver, Steven

Conn, Vanessa M Dávila

Ji, Daniel

et al.

### Publication Date

2024

### DOI

10.3389/fbinf.2024.1400003

### Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed



## OPEN ACCESS

## EDITED BY

Helen Piontkivska,  
Kent State University, United States

## REVIEWED BY

Robin Paul,  
St. Jude Children's Research Hospital,  
United States  
Jean Lutanyo Mbisa,  
UK Health Security Agency (UKHSA),  
United Kingdom

## \*CORRESPONDENCE

Sergei L. Kosakovsky Pond,  
✉ spond@temple.edu

RECEIVED 12 March 2024

ACCEPTED 17 May 2024

PUBLISHED 10 July 2024

## CITATION

Weaver S, Dávila Conn VM, Ji D, Verdonk H,  
Ávila-Ríos S, Leigh Brown AJ, Wertheim JO and  
Kosakovsky Pond SL (2024), AUTO-TUNE:  
selecting the distance threshold for inferring  
HIV transmission clusters.  
*Front. Bioinform.* 4:1400003.  
doi: 10.3389/fbinf.2024.1400003

## COPYRIGHT

© 2024 Weaver, Dávila Conn, Ji, Verdonk, Ávila-Ríos, Leigh Brown, Wertheim and Kosakovsky Pond. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# AUTO-TUNE: selecting the distance threshold for inferring HIV transmission clusters

Steven Weaver<sup>1</sup>, Vanessa M. Dávila Conn<sup>2</sup>, Daniel Ji<sup>3</sup>,  
Hannah Verdonk<sup>1</sup>, Santiago Ávila-Ríos<sup>4</sup>, Andrew J. Leigh Brown<sup>3</sup>,  
Joel O. Wertheim<sup>3</sup> and Sergei L. Kosakovsky Pond<sup>1\*</sup>

<sup>1</sup>Center for Viral Evolution, Temple University, Philadelphia, PA, United States, <sup>2</sup>Center for Research in Infectious Diseases, National Institute of Respiratory Diseases, Mexico City, Mexico, <sup>3</sup>Department of Medicine, University of California San Diego, La Jolla, CA, United States, <sup>4</sup>National Institute of Respiratory Diseases, Mexico City, Mexico

Molecular surveillance of viral pathogens and inference of transmission networks from genomic data play an increasingly important role in public health efforts, especially for HIV-1. For many methods, the genetic distance threshold used to connect sequences in the transmission network is a key parameter informing the properties of inferred networks. Using a distance threshold that is too high can result in a network with many spurious links, making it difficult to interpret. Conversely, a distance threshold that is too low can result in a network with too few links, which may not capture key insights into clusters of public health concern. Published research using the HIV-TRACE software package frequently uses the default threshold of 0.015 substitutions/site for HIV pol gene sequences, but in many cases, investigators heuristically select other threshold parameters to better capture the underlying dynamics of the epidemic they are studying. Here, we present a general heuristic scoring approach for tuning a distance threshold adaptively, which seeks to prevent the formation of giant clusters. We prioritize the ratio of the sizes of the largest and the second largest cluster, maximizing the number of clusters present in the network. We apply our scoring heuristic to outbreaks with different characteristics, such as regional or temporal variability, and demonstrate the utility of using the scoring mechanism's suggested distance threshold to identify clusters exhibiting risk factors that would have otherwise been more difficult to identify. For example, while we found that a 0.015 substitutions/site distance threshold is typical for US-like epidemics, recent outbreaks like the CRF07\_BC subtype among men who have sex with men (MSM) in China have been found to have a lower optimal threshold of 0.005 to better capture the transition from injected drug use (IDU) to MSM as the primary risk factor. Alternatively, in communities surrounding Lake Victoria in Uganda, where there has been sustained heterosexual transmission for many years, we found that a larger distance threshold is necessary to capture a more risk factor-diverse population with sparse sampling over a longer period of time. Such identification may allow for more informed intervention action by respective public health officials.

## KEYWORDS

molecular epidemiology, HIV, network, transmission cluster, surveillance

## 1 Introduction

The use of genomic data to infer and characterize transmission networks of various pathogens has grown in prominence in the past 2 decades, with applications to a growing list of pathogens, including viruses such as HIV (Paraskevis et al., 2016), hepatitis C virus (HCV) (Murphy et al., 2019a), or influenza A virus (IAV) (Jombart et al., 2011), and bacteria such as *M. tuberculosis* (Mai et al., 2018) or *A. baumannii* (Thoma et al., 2022). Notably, genomic surveillance had a prominent role during the COVID-19 pandemic, including the use of sequencing for the study of transmission clusters (von Rotz et al., 2023; Campigotto et al., 2023).

Many competing approaches for inferring transmission clusters, transmission parameters, and source attribution have been described (e.g., for an HIV-1 centric review, see Grabowski et al. (2018)). These approaches can be roughly categorized into distance-based (infer clusters from pairwise genetic distances, e.g., (Kosakovsky Pond et al., 2018)), phylogeny-based (infer phylogenetic trees from the data, then process the resulting tree, e.g., Ragonnet-Cronin et al. (2013)), or phylodynamic (transmission model is directly incorporated into tree inference, e.g., Volz et al. (2017)). These methodological categories differ considerably in model and computational complexity, as well as in interpretability of results. Comparisons of different methods have been undertaken, showing broad compatibility of results, but also highlighting application-specific differences between them (Novitsky et al., 2020). Our goal here is not to develop a conceptually new method, but rather to propose a systematic approach to selecting the key parameter (distance threshold) for the popular HIV-TRACE (Kosakovsky Pond et al., 2018) class distance-based methods for identifying transmission clusters.

Choosing an appropriate genetic distance threshold is an important part of using a molecular transmission network to track the spread of rapidly evolving pathogens (Liu et al., 2020; Rose et al., 2020). This distance threshold defines the degree of genetic closeness between pathogen sequences, isolated from two individuals, required to suggest them as potential transmission partners in the network. Using a distance threshold that is too large can result in a network with many spurious, making it difficult to interpret and analyze. On the other hand, using a distance threshold that is too small can result in a network with too few links, underestimating connections between individuals and making it difficult to accurately track the spread of the disease (Gore et al., 2022).

To enhance the utility of inferred transmission networks, it is important to carefully consider the appropriate distance threshold,  $d$ . This threshold may vary depending on the specific disease and the context in which it is spreading. For example, a highly contagious acute respiratory illness (e.g., SARS-CoV-2) may require a smaller  $d$  than a less contagious chronic illness that is primarily spread through direct contact (e.g., HIV-1). Viruses are more amenable to molecular studies compared to bacteria due to their high genetic divergence and compact genomes. Given the relatively high evolutionary rate of RNA viruses detectable genetic fingerprints can be prioritized for epidemiological studies over short time periods (Paraskevis et al., 2016).

For chronic infections such as HIV, the most appropriate genetic distance threshold should be determined according to the

characteristics of the epidemic such as the speed of transmission, and the evolutionary rate of the genomic region analyzed (Liu et al., 2020). Sampling density and possible delays between infection and diagnosis should be considered, since samples close to the time of seroconversion are more likely to cluster than samples from well after infection. Lower thresholds will capture the most closely related sequences, while higher thresholds will capture long-term epidemics and chronically infected individuals (Junqueira et al., 2019).

Cluster analysis, i.e., identification and analysis of connected network components, in public health has been used for early identification of increased transmission (Oster et al., 2021; 2018), monitoring response to an HIV outbreak (Sizemore et al., 2020; Tookes et al., 2020; Tumpney et al., 2020), evaluating the effectiveness of interventions (Wang et al., 2015; Peters et al., 2016; Liu et al., 2020) or predicting clusters that are most likely to grow in the near future (Erly et al., 2021; Ragonnet-Cronin et al., 2022). This balance can be achieved through careful analysis and consideration of the specific disease and context.

This study introduces AUTO-TUNE, a method that offers a systematic approach to select genetic distance thresholds for molecular HIV transmission network analysis, based purely on the structure of the collected data. By autonomously optimizing clustering metrics derived from pairwise genetic distances, AUTO-TUNE has the potential to improve the accuracy and reliability of network inference, irrespective of data attributes. The AUTO-TUNE methodology's independence from supplementary data makes it less sensitive to variations in data collection protocols and enhances its adaptability to various contexts, including potentially other viral diseases.

## 2 Methods

Assume that there are  $S$  aligned genomic sequences (full or partial, e.g., the HIV-1 *pol* gene) for a pathogen of interest, each representing the "consensus" circulating viral diversity at the time of sampling in a single infected individual. We shall infer a putative transmission network comprising  $S$  nodes, and  $E$  links (edges), where an edge is drawn between a pair of sequences if the genetic distance between them is at or below a threshold  $d$ . In such a network, there will be  $0 \leq C < S$  connected components with more than one node (clusters), which are the primary object of inference. This network inference strategy is used by HIV-TRACE (Kosakovsky Pond et al., 2018), where the genetic distance is computed using the Tamura-Nei (TN93) (Tamura and Nei, 1993) model, with a variety of options controlling how to deal with ambiguous nucleotide bases; for HIV-1 such bases are informative since they often represent variants co-circulating in the infected individual at the time of sampling at substantial frequencies (Kosakovsky Pond et al., 2009).

We begin by describing an approach to assign a score to each of the choices of  $d$  in a plausible/informative range of distances. Note that while such a range is continuous, it is sufficient to only consider distance cutoffs that are in the array of pairwise distances between the sequences, as those are the cut-points where one or more additional edges will be added to the network as  $d$  is increased.

## 2.1 Scoring heuristic procedure

The network threshold selection procedure proceeds as follows (we provide an example in the Results section as well).

- For each candidate threshold  $d_L$ , in increasing order, ranging from the smallest genetic distance in the dataset, up to either the largest distance or a predetermined maximal threshold, we compute two network statistics:  $R_{12}$ , the ratio of the size of the largest cluster to the size of the second largest cluster, and  $C$ , the number of clusters in the network at this threshold. A cluster is defined as a connected component in the network with at least two nodes.
- A priority score is assigned to each  $d_L$ . This score measures two properties of the threshold: Does  $R_{12}$  jump at  $d_L$ ? How far is the number of clusters  $C$  at  $d_L$  from the maximal number of clusters computed over all threshold values? Let there be  $N$  overall  $d_L$  candidate values, and assume we are examining the  $i$ th candidate,  $d_L^i$  with  $W < i \leq N - W$  ( $W$  is a positive integer defined below).
  - The  $R_{12}$  jump is computed by looking at the normalized ratio of the mean  $R_{12}$  values computed over the leading window  $d_L^{i+1} \dots d_L^{i+W}$  and the trailing window  $d_L^{i-W} \dots d_L^{i-1}$ . The width of the window,  $W$ , is defined as  $\min(\max(\lfloor \frac{N}{100} \rfloor, 3), 30)$ . The distribution of ratios is converted to  $Z$  scores, and normalized relative to the largest positive  $Z$  score across all candidate distances, yielding the jump component of the score.
  - The number of clusters,  $C_i$  at threshold  $d_L^i$  is first normalized to  $[0,1]$  through  $\frac{C_{max} - C_i}{C_{max} - C_{min}}$  and next gated via a Gompertz function transform  $1 - e^{-e^{-25x+3}}$ . This function provides an *ad hoc* means for penalizing having too few clusters relative to the maximum over all ranges. For example, a threshold that yields 95% of the maximal number of clusters receives a score of 0.996, a threshold that yields 85% - a score of 0.376, and a threshold that yields 60% - a score of 0.0009.
  - The priority score for  $d_L^i$  is the sum of the two components defined in a) and b), and ranges from 0 to 2.
- The threshold with the highest priority score will be selected as the suggested automatic distance threshold, if the score is high enough (1.9 or more), and either of the two conditions hold.
  - No other thresholds have priority scores of 1.9 or higher
  - If other thresholds have priority scores of 1.9 or higher, then the range of thresholds represented by these options is small (no more than  $\log N$  times the mean step between successive  $d_L^i$ ).
- If no single threshold can be selected in step 3, then the one with the highest priority score is suggested, and an inspection of a plot of scores is recommended to ensure that the threshold is sensible.

The corresponding flowchart can be found in Figure 1. The  $R_{12}$  jump component of the score is motivated by the giant component formation result from network theory: when the degree distribution satisfies particular conditions, most nodes in the network will belong to a single component, or cluster (Molloy and Reed, 1995). This situation leads to epidemiologically uninformative networks, and should be

avoided. The default configuration considers all clusters (i.e., with two or more nodes), but users can specify that only clusters with  $K$  or more members ( $K \geq 2$ ) should be included in  $R_{12}$  and  $C_i$  calculations.

## 2.2 Assortativity

Degree-weighted homophily (DWH) is a measure of similarity between nodes in a network based on their attributes (such as demographic characteristics or behaviors) and their degree centrality (i.e., the number of connections they have to other nodes in the network). It is used to quantify the extent to which nodes with similar attributes tend to be connected to each other more frequently than would be expected by chance (Golub and Jackson, 2012). DWH is calculated as the ratio of the observed number of connections between nodes with similar attributes to the expected number of connections between such nodes, based on their network degree.

For any two subsets  $A$  and  $B$  of nodes in a network without singletons (each node has a positive degree), define the weight between  $A$  and  $B$  as

$$W_{A,B} = \frac{1}{|A||B|} \sum_{i \in A, j \in B, (i,j) \text{ are connected}} \frac{1}{d_i d_j},$$

where  $d_i$  is the degree of node  $i$ , and  $|X|$  is the cardinality (size) of subset  $X$ .

Then for any proper (not empty and not the complete network) subset of the network,  $G$ , e.g., a group of nodes sharing an attribute, e.g., transmission risk factor, define

$$DWH = \frac{W_{G,G} + W_{\bar{G},\bar{G}} - 2W_{G,\bar{G}}}{|G|^{-2} \sum_{i \in G} 1/d_i + |\bar{G}|^{-2} \sum_{i \in \bar{G}} 1/d_i}, \quad (1)$$

with

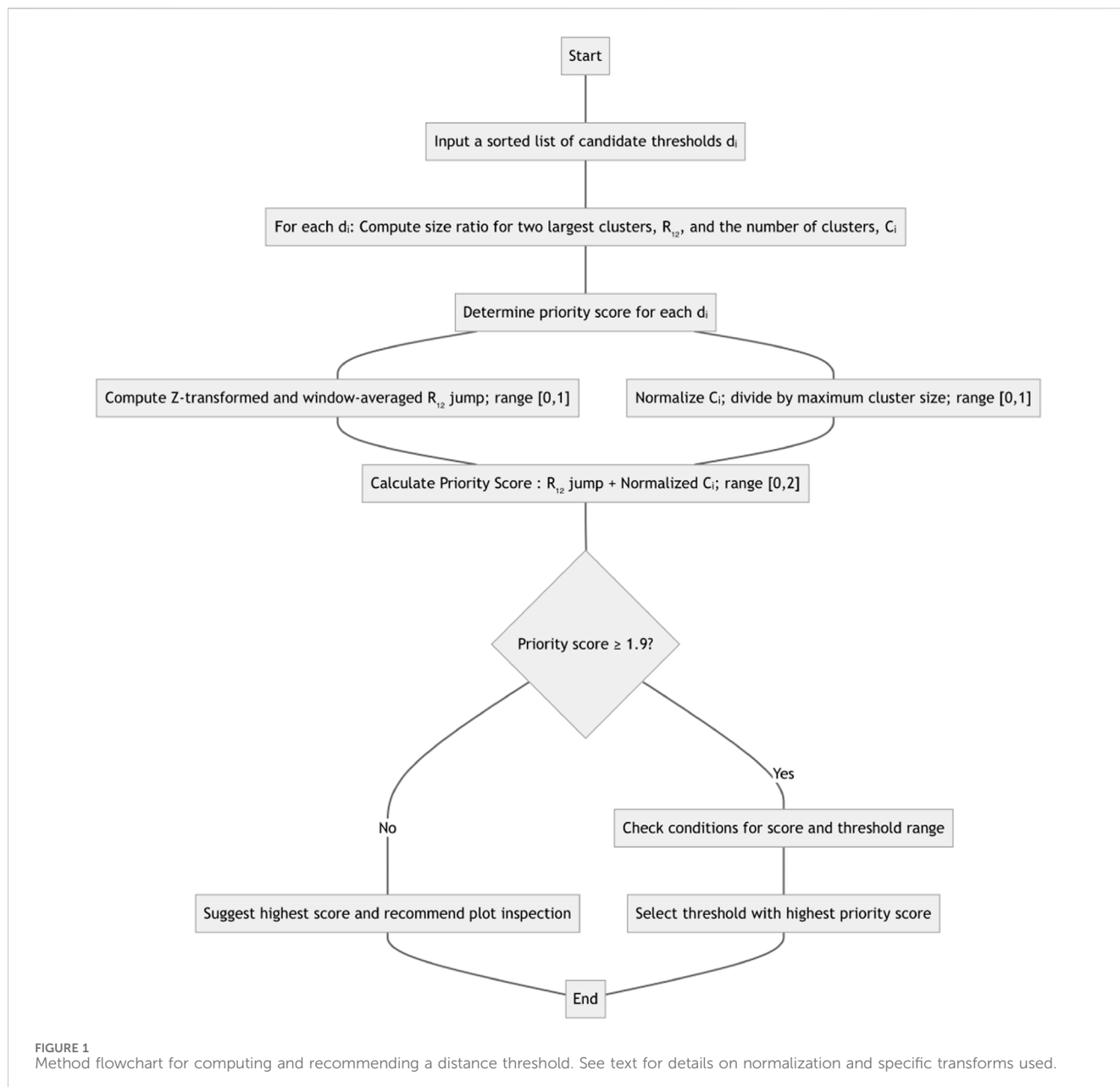
- $\bar{G}$ : the complement of  $G$  (all nodes not in  $G$ )
- $d_i$ : the degree of node  $i$

DWH ranges from  $-1$  to  $1$ . A DWH value of  $0$  indicates that there is no more homophily than expected by chance (conditioned on network structure), while a value of  $1$  indicates that there is perfect homophily ( $G$  consists of connected components disconnected from the rest of the network). A value of  $-1$  is achieved for perfectly disassortative networks (the only links are between  $G$  and  $\bar{G}$ ).

Homophily metrics have been used in social network analysis and in the study of how different attributes are related to the formation of connections between individuals (Ragonnet-Cronin et al., 2021). To assess whether or not DWH is significantly different from  $0$  (and from random expectation), we generate the null distribution of DWH obtained by randomly reshuffling node attributes used to define group  $G$  and recomputing DWH for each such replicate.

## 2.3 Implementation

The software implementation involves a step-by-step process that utilizes the HIV-TRACE suite of packages. It starts with calculating pairwise distances with the `tn93` tool and a supplied



multiple sequence alignment. Thus generated pairwise distances are supplied to the `hivnetworkcsv` script while providing the `-A` keyword argument. A brief outline of the software's implementation is as follows.

1. Calculate pairwise distances: The user first calculates the pairwise distances using the `tn93` fast pairwise distance calculator, providing the maximum threshold value to consider (0.03 in this case, which may be revised upwards for sufficiently divergent sequences, as this provides an upper bound of thresholds to consider) and the input FASTA file. The command for this step is

```
tn93 -t 0.030 -a resolve -g 0.05 pol.fasta > pairwise_distances.15.tn93.csv
```

Please note that the threshold should include the maximal range one is intending to test.

2. Compute priority scores for each candidate threshold: The `hivnetworkcsv` script is then executed with the required input file, format, and autotune option to generate a tab-separated output file, as shown below

```
hivnetworkcsv -i pairwise_distances.15.tn93.csv -f plain -A 0 > autotune_report.tsv
```

3. Visualize the report: Users can upload the generated `autotune_report.tsv` file to

<http://autotune.datamonkey.org/analyze> for visualization and further analysis of the data. This web-based site extends the

Datamonkey platform (Weaver et al., 2018) to provide an interactive environment to explore scores and other metrics across the range of tested outputs.

4. Run HIV-TRACE: Once AUTO-TUNED threshold(s) are settled upon after review, the user runs the HIV-TRACE command with the appropriate input FASTA file, distance threshold, and other required arguments. The output is saved as a JSON file. An example command is

```
hivtrace -i ./INPUT.FASTA -a resolve -r HXB2_prrt -t <autotune.threshold> -m 500 -g .05 >
↳hivtrace.results.json
```

### 2.3.1 Optional: compute assortativity metrics

1. Annotate results: The `hivnetworkannotate` script is used to annotate the results obtained from the HIV-TRACE step with attributes. The script takes the JSON results file, node attributes file, schema file, and a resolve flag as input.

```
hivnetworkannotate -n hivtrace.results.json -a node.attributes.json -g schema.json -r
```

For more information, users can refer to the `hivnetworkannotate` documentation.

2. Analyze the results with DWH: After the results file has been annotated, the user can proceed to the assortativity page,

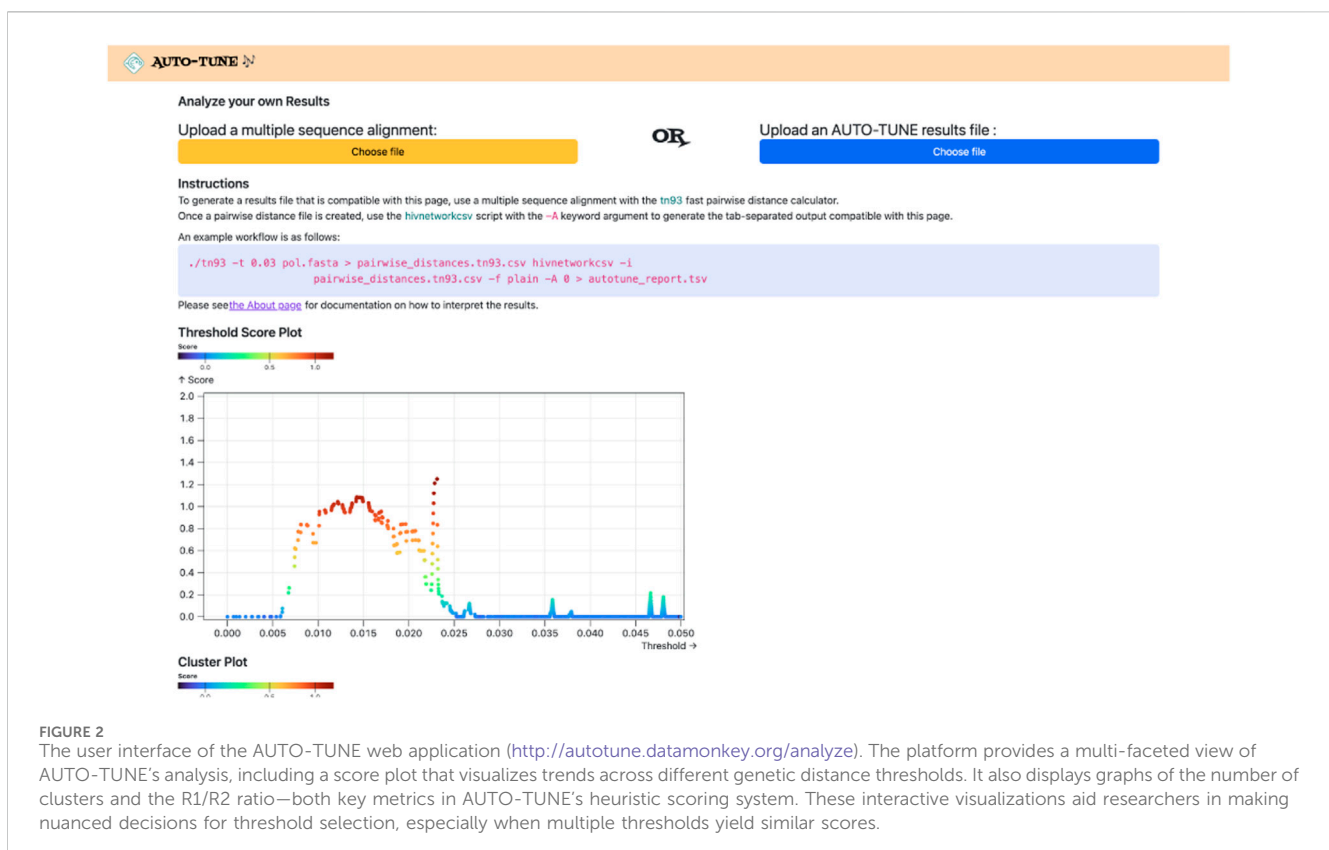
<http://autotune.datamonkey.org/assortativity>, for further analysis of the output.

The described workflow offers a systematic approach to analyze potential distance thresholds for one's data with AUTO-TUNE, from calculating pairwise distances to visualizing and annotating results.

## 2.4 Visualization

Visualizations of AUTO-TUNE results are accessible at <http://autotune.datamonkey.org/analyze>. These include the priority score plot, and the two contributing statistics: cluster count relative to the maximum and the ratio of two largest cluster sizes (Figure 2). An assortativity tool is available at <http://autotune.datamonkey.org/assortativity>, and is an analytical tool engineered to facilitate the calculation of Degree-weighted homophily (DWH) values. It utilizes the DWH NPM package to generate a tabular representation of DWH values corresponding to each value for a selected attribute annotation, providing an exhaustive examination of the interrelationships for the field. The tool also computes the panmictic (null) range, which involves a label permutation test to generate the null distribution of DWH values. This feature establishes a comparative baseline that aids in determining the significance of homophily *versus* what would be expected by chance.

The visualization code is available on Github (<https://github.com/stevenweaver/autotune-app/>).



## 2.5 Comparisons with previously published analyses

First, we set out to compare the thresholds used in numerous published studies with those obtained by AUTO-TUNE. To select the data sets for this analysis, we conducted a scientific literature search to identify studies focused on HIV networks for public health purposes. We then filtered the studies that utilized HIV-TRACE to infer genetic networks and had publicly available sequences. Due to privacy concerns, HIV-1 sequences are frequently not released in the public domain (Inzaule et al., 2023). Some of the best-sampled datasets are national-level cohorts, such as the UK HIV Drug Resistance Database (Dunn and Pillay, 2007), the Swiss HIV cohort (Scherrer et al., 2022), or the Dutch ATHENA cohort (Boender et al., 2018). However, because sequences from these cohorts are not in the public domain and are typically subject to strong usage restrictions, we elected not to use such data, for reasons of reproducibility, practicality, and data transparency.

We also attempted to include studies from different countries and regions, enabling us to assess the performance of our method across various epidemic contexts, risk groups, and network sizes in real-data sets that used variable clustering thresholds.

Second, we compared AUTO-TUNE with the most direct published alternative: the `clustner` method (Chato et al., 2020). We procured datasets from Wolf et al. (2017) and Vrancken et al. (2017) utilizing the approach delineated in Chato et al. (2020). These datasets, namely, Middle Tennessee, Seattle, and Alberta were processed using the workflow described in Section 2.3. This enabled us to determine an optimal threshold for each dataset using AUTO-TUNE. We further executed the command as detailed in step 4 of Section 2.3, deploying thresholds previously established as optimal by Chato et al. (2020). Note that `clustner` requires and uses temporal information (dates sequences were collected), whereas AUTO-TUNE does not.

Lastly, we evaluated the effect of sampling density on the genetic distance threshold as determined by AUTO-TUNE, we implemented a strategy of random subsampling from the original dataset sourced from Rhee et al. (2019). This study was selected due to its satisfactory AUTO-TUNE score when utilized in its entirety, as well as its inherent design as a Geographically-Stratified set of 716 *pol* Subtype/CRF (GSPS) reference sequence dataset. The dataset, which comprises 6034 samples gathered between 1989 and 2016, was subjected to random subsampling ten times at proportions of 25%, 50%, and 75% of the original sample size. For each subsample, the optimal threshold and associated scores were determined via AUTO-TUNE.

## 3 Results

### 3.1 Comparisons with published HIV-1 molecular epidemiology studies

We selected several publications citing HIV-TRACE for our analysis, primarily because these studies not only referenced the tool but also made some or all of their sequence data publicly available (Tables 1, 2). These studies adopted several different approaches for selecting genetic distance thresholds, including using US CDC guidelines (Yan et al., 2020), picking thresholds based on prior studies (Sivay et al., 2018), and visually inspecting the numbers of

clusters and nodes in the networks across candidate distance thresholds (Liu et al., 2020). These thresholds, often qualitatively determined, tended to be round numbers, and were usually determined using *ad hoc* or subjective procedures. Some studies stratified their analyses by viral subtype (major clade), while others did not (or this was not applicable).

A direct comparison with published networks is not feasible because only the underlying sequence data (and often only some of the sequences) are made available, not the networks themselves. To facilitate comparison here, we used distance thresholds and all available sequences from primary publications to infer transmission networks anew (the scripts for doing so and the corresponding settings are available in [github.com/veg/auto-tune](https://github.com/veg/auto-tune)) and compare them with the networks obtained using the highest scoring AUTO-TUNE threshold.

With a few exceptions (e.g., Dalai et al. (2018); Sivay et al. (2018)), both the distance thresholds and the inferred networks were quite different, in terms of the numbers of connected nodes, clusters, degree distributions, and even hyper-parameters, such as the characteristic exponent of the scale free degree distribution,  $\rho$ . This is true even for the studies where the published threshold was tuned (typically to maximize the number of clusters). AUTO-TUNE thresholds were larger than the published values in 13/21 datasets, and smaller in 8/21 datasets.

#### 3.1.1 Examples of how changing thresholds affects inferred networks

##### 3.1.1.1 Cluster size reduction

The 0.02 subs/site (substitutions/site) threshold used by Dalai et al. (2018), yielded one large cluster composed of two loosely connected components (one PWID/HSX, one MSM, see Figure 3 in that paper). A minute change to the threshold by AUTO-TUNE to 0.0194 subs/site splits one large cluster into three (some nodes also became disconnected), separating the two major risk groups; this is because the “bridging” connections were between these two thresholds (see Figure 4A). This minor change also reduced  $R_{12}$  from 21 to 2.6.

##### 3.1.1.2 Cluster size increase

Increasing the threshold from 0.015 to 0.02495 subs/site on data from Little et al. (2014) combined several small clusters (and singletons) into a single larger cluster, while preserving the overall size and properties of the network (see Figure 4B). This change also reduced  $R_{12}$  from 2.5 to 1.5.

##### 3.1.1.3 Thinning out the network

Reducing the threshold from 0.015 to 0.01139 subs/site on data from Rhee et al. (2019) dramatically reduced the size of the largest cluster, and thinned out most clusters with five or more nodes (see Figure 4C). This is different from the Dalai et al. (2018) case above, because the entire network is affected, rather than a single or a few clusters.

##### 3.1.1.4 Materially changing the degree distribution of the network

For the sequences from Li et al. (2022), AUTO-TUNE suggests  $D = 0.01483$  subs/site with robust (1.76) confidence, whereas the original  $D = 0.013$  subs/site was selected based on maximizing the number of clusters (and likely rounding to the nearest decimal). While the total number of the clusters only increases by 1, the

**TABLE 1** Comparison of AUTO-TUNE and published thresholds from prior studies using partial HIV-1 polymerase gene sequences. *N*: the number of sequences; *L*: length of the multiple sequence alignment, bp; *E* [*D*] mean pairwise TN93 distance; (the studies are sorted on this column, in ascending order) ¶: the original study performed threshold tuning (varied methods); †: distance thresholds were specific to subtypes; ★: the corresponding AUTO-TUNE score is  $\geq 1.9$ ; ●: only a subset of the complete dataset was made available (privacy, data use restrictions, incomplete GenBank submissions), the number of sequences analyzed here is shown after the/symbol; N.R.: not reported.

References	<i>N</i>	<i>L</i>	<i>E</i> [ <i>D</i> ] (%)	Scope	Location/Country	Timespan	Common subtypes	Distance threshold, sub/site	
								Published	AUTO-TUNE
Zai et al. (2020)	209	1056	1.5	Country	China	2007–2015	CRF55/01B	¶ 0.002	0.00255
Liu et al. (2020)	2087/ 1907 ●	1053	5.3	City	Shenyang, China	2008–2016	CRF01, CRF07, B	¶ 0.005/0.007 †	0.00621
Dalai et al. (2018)	317	1044	5.5	City	San Mateo, CA, United States of America	1997–2008	96% B	0.02	0.01944
Chato et al. (2020)	1653/ 1840 ●	1020	5.5	City	Seattle, United States of America	2000–2013	B	¶ 0.016	0.01538
Chato et al. (2020)	808	1017	5.6	Province	Northern Alberta, Canada	2007–2013	B	¶ 0.0104	0.01201
Little et al. (2014)	648/ 646 ●	1212	5.9	City	San Diego, CA, United States of America	1996–2011	98.5% B	0.015	0.02495
Pérez-Losada et al. (2017)	1879/ 3411 ●	1027	6.0	City	Washington DC, United States of America	1987–2015	B	0.010	0.01733
Rhee et al. (2019)	4553	897	6.1	State	CA, United States of America	1998–2016	95.5% B	0.015	0.01139 ★
Chato et al. (2020)	2779/ 2750 ●	1398	6.3	State	Tennessee, United States of America	2001–2015	B	¶ 0.016	0.01872
Temereanca et al. (2017)	37	1302	6.7	City	Bucharest, Romania	2010–2013	F1, G, B	0.015	0.00194 ★
Brenner et al. (2021)	10945/ 448 ●	738	6.7	Province	Quebec, Canada	2002–2020	B	0.015/0.025	0.02741
Sivay et al. (2018)	201	1302	6.9	Province	Mpumalanga, South Africa	2011–2015	C	0.025	0.02506
Li et al. (2022)	295	1206	7.8	Prefecture	Pu'er, China	2021	CRF08, CRF01, CRF07	¶ 0.013	0.01483 ★
Yu et al. (2022)	316	1074	7.9	Province	Guangxi, China	2012–2018	CRF01, CRF08, CRF07	¶ 0.013	0.01178
Yan et al. (2021)	1695/ 1569 ●	1569	8.4	City	Guangzhou, China	2008–2012	CRF01, CRF07, CRF55,G	0.015	0.00839
Billings et al. (2019)	150	1597	8.5	City	Lagos, Nigeria	2013–2016	CRF02, URF	0.015	0.0233
Chen et al. (2023)	1975/ 209 ●	1050	8.7	Province	Guangxi, China	2016–2018	CRF01, CRF07, CRF08	¶ 0.0075	0.01295
Fabeni et al. (2020)	726/ 3499 ●	1029	9.2	Country	Italy	1998–2018	B	0.010	0.0037
Leal et al. (2020)	630/ 633 ●	990	9.2	State	Maranhão, Brazil	2008–2017	B	0.020	0.04033
Bbosa et al. (2020)	2018	1257	9.3	Country	Uganda	2009–2016	N.R.	0.015	0.02035 ★
Stecher et al. (2018)	2774	1028	12.1	Multi-City	Germany	1999–2016	B	0.015	0.03056

number of nodes connected in the network grows from 95 to 119, and the scale free exponent of the distribution is dramatically affected. The latter is informed by the degree distribution of the network, and Figure 4D shows, the degree distribution is dramatically affected. The degree distribution of network, which

tabulates the number edges connected to each node, is a fundamental feature of network analysis. For each integer  $0 \leq K \leq K_{\max}$ , the degree distribution function counts how many nodes have exactly  $K$  edges connected to them.  $K_{\max}$  is simply the highest such number for a given network. Many commonly used



**TABLE 2** Network properties at the published and AUTO-TUNE thresholds. In cases when the original paper used more than one threshold, we selected the largest for comparison. The datasets are ordered by the AUTO-TUNE priority score from highest to lowest.  $\rho$  is the fitted characteristic scale-free exponent of the corresponding degree distributions.

References	AUTO-TUNE score	Nodes in network		Clusters in network		$R_{12}$		Scale parameter $\rho$	
		Published	AUTO-TUNE	Published	AUTO-TUNE	Published	AUTO-TUNE	Published	AUTO-TUNE
Li et al. (2022)	2.00	1364	1224	277	277	1.7	2.4	2.8	2.6
Chato et al. (2020) TN	2.00	394	445	108	109	1.0	1.7	2.7	2.9
Rhee et al. (2019)	1.95	2044	1636	524	488	13.2	1.5	2.6	2.7
Bbosa et al. (2020)	1.93	222	296	102	119	2.2	1.6	3.2	2.6
Dalai et al. (2018)	1.89	60	54	9	11	22	2.6	2.0	2.2
Temereanca et al. (2017)	1.79	30	16	5	3	3	1.5	N/A	2.8
Yu et al. (2022)	1.76	55	51	19	19	2.75	1.75	10.4	34.0
Sivay et al. (2018)	1.42	51	51	19	19	1.5	1.5	3.2	3.0
Zai et al. (2020)	1.40	96	98	26	27	1.5	1.5	24.1	17.7
Little et al. (2014)	1.31	301	394	98	87	2.5	6.1	3.6	3.1
Brenner et al. (2021)	1.22	363	379	71	70	5.6	5.5	2.7	2.8
Stecher et al. (2018)	1.20	97	558	36	155	2.2	4.9	3.2	3.3
Chato et al. (2020) Seattle	1.16	505	484	148	149	2.5	1.7	2.7	2.6
Billings et al. (2019)	1.16	38	78	13	23	2	2.3	2.6	11.5
Yan et al. (2021)	1.14	1084	753	124	116	2.0	1.8	1.2	2.0
Chen et al. (2023)	1.11	20	47	8	16	1.3	2.0	$\infty$	$\infty$
Leal et al. (2020)	1.11	50	270	25	57	1	1.6	53.6	3.1
Pérez-Losada et al. (2017)	1.06	172	431	76	134	5.1	1.4	5.2	2.9
Liu et al. (2020)	1.05	885	797	156	161	6.0	4.5	3.1	3.0
Chato et al. (2020) Alberta	1.03	394	445	108	109	1.0	1.7	2.7	2.9
Fabeni et al. (2020)	1.00	626	221	197	83	2.1	3.2	2.1	3.2

network-derived correlates (e.g., degree centrality) can be strongly affected by such changes.

### 3.1.1.5 Expanding the network

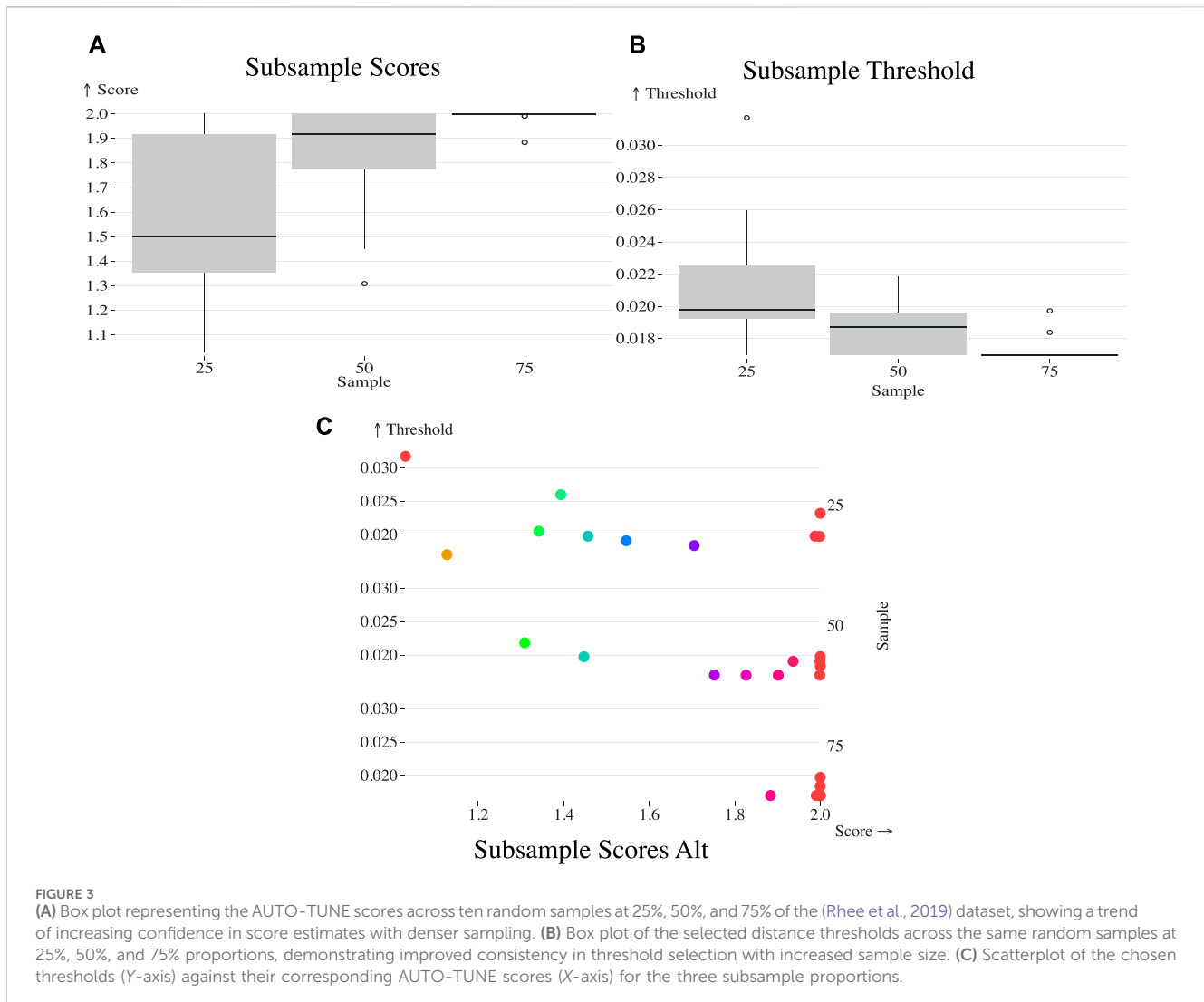
Increasing the .015 subs/site threshold in Billings et al. (2019) to 0.0233 subs/site more than doubles the number of nodes included (Figure 4E). This is distinct from the Rhee et al. (2019) case above, because, once again, most of the network is affected, rather than a few key clusters.

## 3.2 Interpretation

Networks with high AUTO-TUNE scores are exemplified by the alignment (in the distance space) of the points where the number of clusters is maximized and where the network

transitions to having an “unusually” large cluster (see Figure 5A). In cases of low scores, AUTO-TUNE effectively falls back to maximizing the number of clusters as a function of the distance thresholds, which is a common strategy found in empirical studies (see Figure 5B).

As expected, AUTO-TUNE inferred smaller thresholds for younger (e.g., studies based in China) epidemics. While AUTO-TUNE will always return a score, in the majority of cases there is no clear “winner” threshold, with priority scores exceeding 1.5 in only 6/18 cases (Table 2). One interpretation for such lack of clarity is that the underlying network has several different (e.g., spatial, temporal, or subtype-specific) thresholds which cannot be well-represented by any single value. For instance, when analyzing the data from Yan et al. (2021), AUTO-TUNE returned a low score of 1.14 for  $D = 0.00839$  subs/site. However, when we split the data into major constituent subtypes and ran AUTO-TUNE on each one



separately, starkly discrepant thresholds were found for different subtypes:  $D = 0.0102$  subs/site (score = 1.59) for CRF01,  $D = 0.00193$  subs/site (score = 2) for CRF05,  $D = 0.02615$  subs/site (score = 1.65) for B, and  $D = 0.0111$  subs/site (score = 1.04) for CRF07. Although many networks from the literature tend to be dominated by sequences from the same subtype, in more heterogeneous settings it seems prudent to partition the data by subtype (corresponding to major phylogenetic clades), and perform network analyses within subtypes.

### 3.3 Minimum cluster size setting

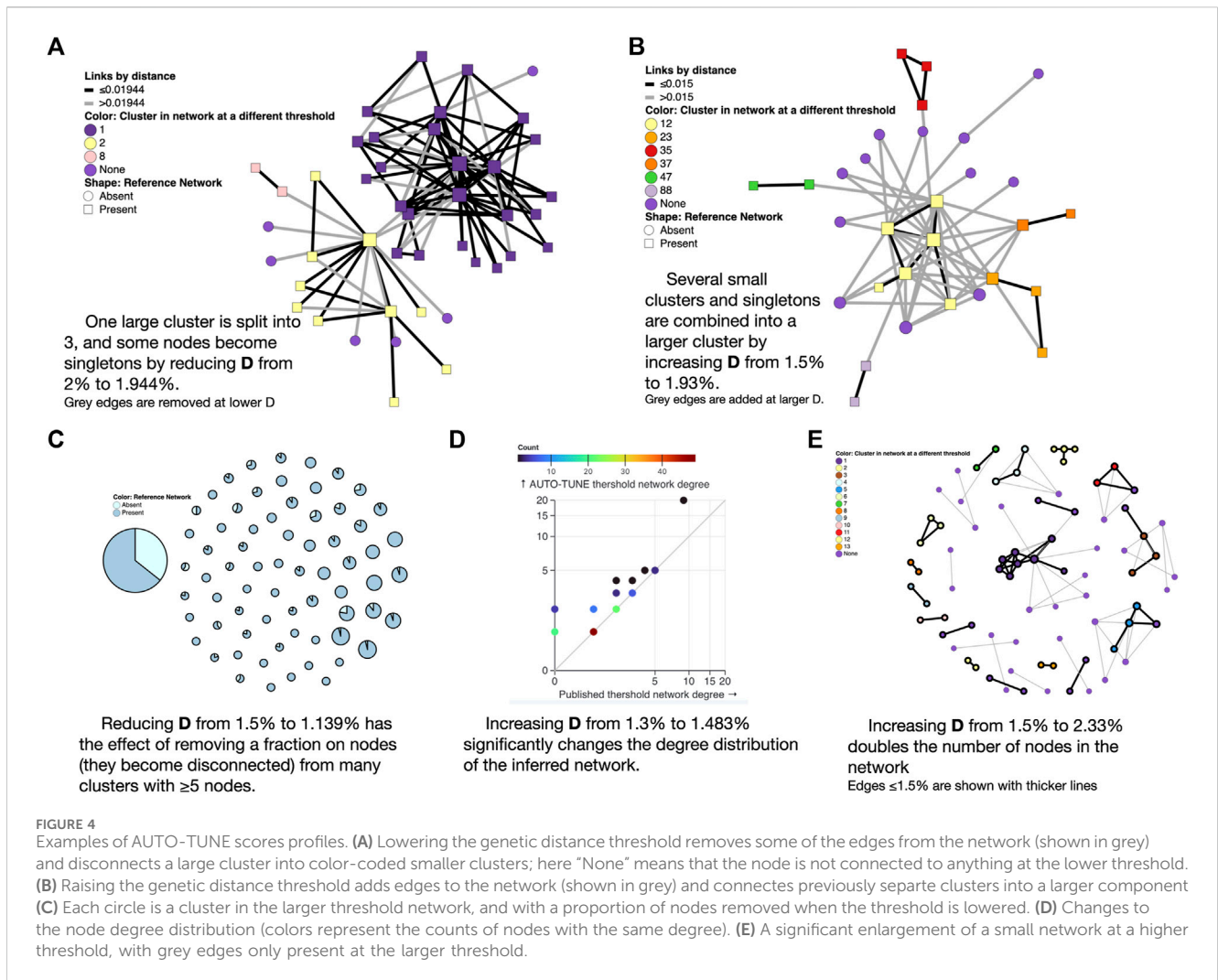
We explored how selecting the minimum number of connected nodes needed to define a cluster affected the selected threshold and the score for the same collection of 21 empirical datasets (Table 3). For the majority of the datasets, requiring three or more connected nodes to define a cluster had a minor effect on the selected distance, with the maximal priority score achieved for the standard minimum of two nodes. The few exceptions where higher distances yield larger

scores come from very small networks, where there are only a few clusters, e.g., Chen et al. (2023); Dalai et al. (2018).

### 3.4 Comparisons with published non-HIV molecular epidemiology studies

While HIV-1 epidemiology is the predominant niche for distance-based molecular transmission analyses, other rapidly evolving viruses, especially HCV, have also been analyzed with these approaches (Bartlett et al., 2017). Unlike HIV-1, there is considerably less work on how to choose an appropriate distance threshold, further complicated by the use of different genes to build networks (see Chan et al. (2020) for a comprehensive summary). Two commonly seen methods exist: use some measure of intra-host variation (obtained by deep sequencing) as a lower bound for the threshold, or tune  $D$  to obtain some desired network property, e.g., the maximal number of clusters. Like with HIV-1, we searched the literature for relevant studies, and selected several with publicly available sequence data.

Most of the datasets are much smaller and less systematically sampled than those for HIV-1, and often combine highly divergent



subtypes in the same collection, making a joint analysis challenging. As with HIV-1, AUTO-TUNE returns a wide range of scores and *D* thresholds. For example, effectively maximizing the number of clusters on rhinovirus sequences from Ng et al. (2022) yields a *D* estimate very similar to that obtained by the authors from intra-host variability (information not available to AUTO-TUNE). Table 4.

### 3.5 Large-scale HIV-1 database analyses

#### 3.5.1 Markedly different thresholds for different subtypes

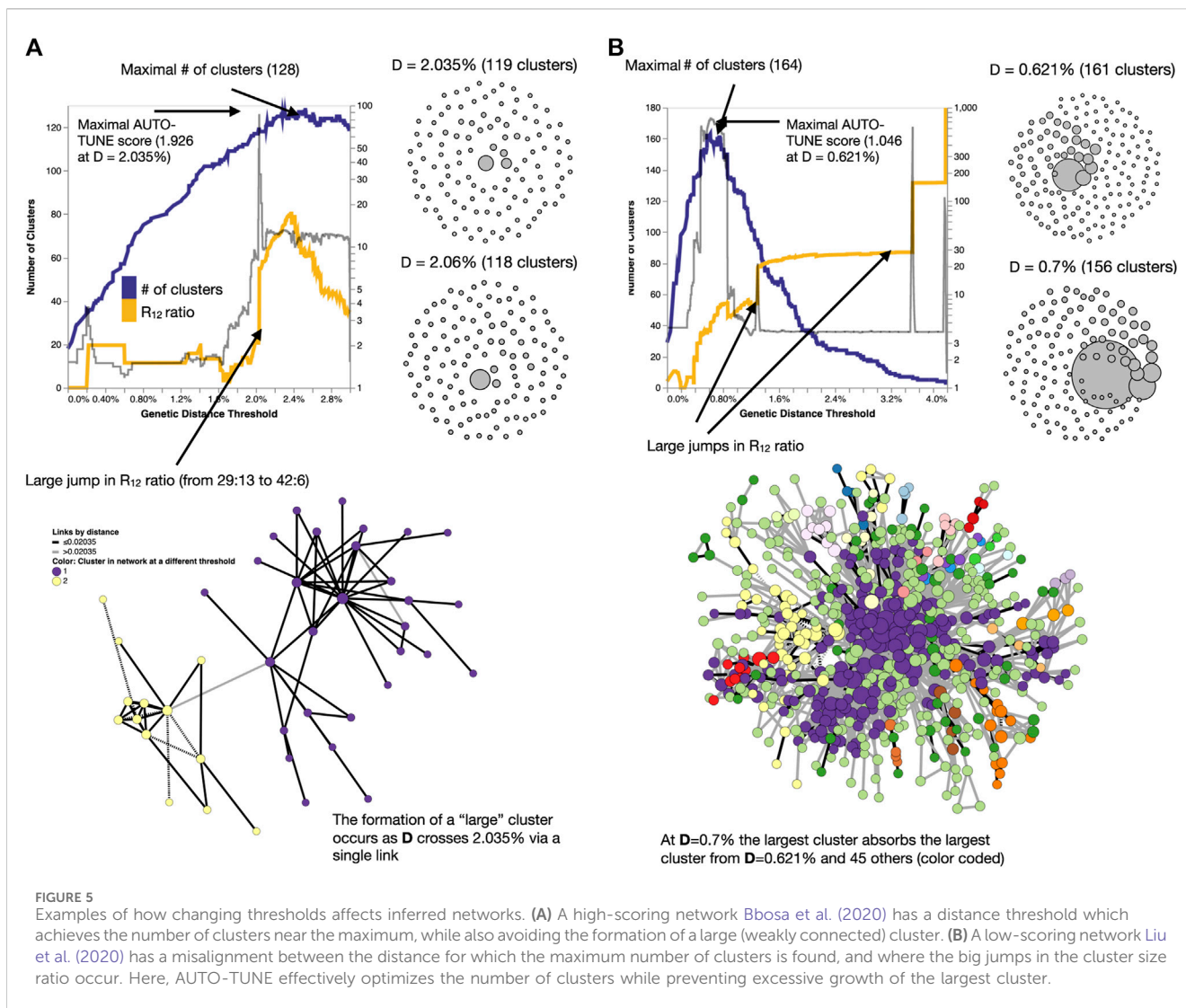
Following the spirit of the analysis performed by Wertheim et al. (2014), we downloaded partial *pol* sequences (between HXB-2 coordinates 2253 and 3200, one sequence per patient) from the Los Alamos HIV-1 Database, split them by annotated subtype and applied AUTO-TUNE to individual subtypes with 1000 or more sequences.

Some (but not all) HIV-1 subtypes often act as strong correlates of regional and temporal distributions of sequences, and are expected to represent epidemics with different sampling rates and transmission dynamics. These differences are reflected in a wide range of mean pairwise distances and inferred AUTO-TUNE

thresholds shown in Table 5. For example, the relatively young subtype A6, which is the most common subtype in the countries of the former Soviet Union (Abidi et al., 2021), has a low mean pairwise distance (0.046) and a low AUTO-TUNE threshold (0.0056). In contrast A1D recombinant sequences have high distance and threshold values (0.089 and 0.0323, respectively), because sequences of this “subtype” represent broadly circulating strains with complex backgrounds, and extensive histories of recombination (Foster et al., 2014; Yebra et al., 2015).

There was extensive variability among subtypes in all high-level network statistics, including the mean degree, fractions of nodes that were in the network, and the characteristic exponent  $\rho$ , where  $\rho$  is inferred from by fitting the degree distribution to various network formation models, and with  $\text{Prob}(\text{degree} = k) \sim 1/k^\rho$  for large  $k$ .

For A1, B, C, and CRF08 networks there’s very strong support for a single AUTO-TUNE threshold (score >1.9), while for many other subtypes there is extreme ambiguity in which threshold to choose (score <1.1). We suggest that networks where AUTO-TUNE fails to find a single threshold may comprise heterogeneous data which require multiple thresholds to resolve.



### 3.5.2 Congruence of networks inferred from different genes

Very few published studies of HIV-1 transmission networks use genes other than *pol*, and nearly all of the extrinsically motivated thresholds have been derived for this gene, the utility of other genes and the appropriate  $D$  values for them are unclear. Because of different rates of evolution in HIV-1 genes and, possibly, subtypes (Penn et al., 2008), one would expect  $D$  to be different for different subtypes and genes. As a simple exercise, we downloaded full-length HIV-1 genomes from the LANL database, stratified them by subtype, and conducted AUTO-TUNE inference using four genomic segments: protease + reverse transcriptase, integrase, matrix (*gag*), and the less variable *gp41* segment of the envelope gene.

Only three subtypes had  $\geq 500$  full-length sequences in the LANL HIV database (Table 6): B, C, and CRF01. As expected, the inferred thresholds differed by gene and subtype, with lower thresholds inferred for more slowly evolving segments (PR + RT and INT), and similar numbers of clusters found in the resulting subtype-level networks. For all three subtypes, the level of agreement between the four networks on whether or not nodes were clustered

or not (present/absent from the network), measured by Krippendorff's  $\alpha$  (Hayes and Krippendorff, 2007), were substantially higher than expected by chance ( $\alpha = 0$ ). All four networks also had between a quarter and a half of all the clusters in perfect agreement.

### 3.6 Evaluating inferred networks using homophily

Non-random mixing and attribute-based homophily are intrinsic characteristics of human contact networks and can be expected to be present in transmission networks, particularly in the context of HIV transmission. People frequently engage in relationships with those who share similar attributes or behaviors, such as risk factors (e.g., PWID, MSM). Recent evidence suggests that race/ethnicity is also a strong predictor of homophily in HIV networks (Ragonnet-Cronin et al., 2021). The effect of these nonrandom mixing patterns on the genetic diversity of HIV-1 has not only been extensively explored via modeling and simulations (Goodreau, 2006), but the structure of sexual contact

**TABLE 3** AUTO-TUNE distance thresholds and scores as a function of the minimal cluster size. Rows are sorted by the score at cluster size  $\geq 2$ . The maximum score (or scores) for each row are highlighted in boldface.

References	AUTO-TUNE minimum cluster size							
	Size $\geq 2$		Size $\geq 3$		Size $\geq 4$		Size $\geq 5$	
	Threshold, %	Score	Threshold, %	Score	Threshold, %	Score	Threshold, %	Score
Li et al. (2022)	1.15	<b>2.0</b>	1.15	1.97	1.8	1.4	1.15	<b>2.0</b>
Chato et al. (2020) TN	1.87	<b>2.0</b>	1.87	<b>2.0</b>	1.87	<b>2.0</b>	1.87	<b>2.0</b>
Rhee et al. (2019)	1.14	<b>1.94</b>	1.14	1.17	2.05	1.12	1.14	1.11
Bbosa et al. (2020)	2.04	<b>1.94</b>	2.49	1.01	2.04	1.00	2.31	1.03
Dalai et al. (2018)	1.94	1.89	1.50	1.02	1.95	1.00	1.95	<b>2.0</b>
Temereanca et al. (2017)	0.19	<b>1.78</b>	0.27	2.0	2.79	1.00	2.79	1.06
Yu et al. (2022)	1.18	<b>1.76</b>	1.22	1.00	2.67	1.18	2.67	1.37
Sivay et al. (2018)	2.51	<b>1.41</b>	3.68	1.07	3.67	1.07	3.67	1.07
Zai et al. (2020)	0.26	<b>1.41</b>	0.26	1.41	0.26	1.19	0.27	1.18
Little et al. (2014)	2.35	<b>1.58</b>	2.35	1.20	1.76	1.04	2.10	1.03
Brenner et al. (2021)	2.74	1.22	2.74	1.22	2.01	<b>1.26</b>	2.01	1.26
Stecher et al. (2018)	3.06	<b>1.20</b>	3.46	1.14	3.49	1.12	3.88	1.09
Chato et al. (2020) Seattle	1.53	<b>1.16</b>	1.76	1.11	1.54	1.16	1.54	1.09
Billings et al. (2019)	2.33	1.16	2.97	1.12	2.33	1.17	3.00	<b>1.31</b>
Yan et al. (2021)	1.22	<b>1.09</b>	1.29	1.01	1.48	1.00	1.62	1.01
Chen et al. (2023)	1.30	1.11	2.7	1.31	1.89	1.08	2.73	<b>1.83</b>
Leal et al. (2020)	4.03	1.11	4.03	1.12	4.24	1.19	4.05	<b>1.21</b>
Pérez-Losada et al. (2017)	1.73	<b>1.06</b>	2.04	1.04	2.04	1.03	2.04	1.04
Liu et al. (2020)	0.62	<b>1.13</b>	0.62	<b>1.13</b>	0.62	1.13	1.29	1.00
Chato et al. (2020) Alberta	1.20	<b>1.03</b>	3.87	1.00	3.87	1	1.15	1.01
Fabeni et al. (2020)	1.77	1.00	3.05	1.00	3.97	1.15	3.97	<b>1.45</b>

networks has been found to directly influence pathogen phylogenies (Robinson et al., 2013). Phylogenetic analysis of HIV type 1 sequences has revealed distinct grouping patterns based on risk behaviors (Holmes et al., 1995). The expectation of homophily is so strong, that its disruption, e.g., the presence of a self-reported heterosexual risk group individual in a cluster exclusively composed of MSMs has been used as a marker of undisclosed/incomplete risk factor reporting (Ragonnet-Cronin et al., 2018). Therefore, when subject-level attributes are available, homophily is an expected and desired feature of the network.

To assess the performance of an AUTO-TUNED optimized threshold using degree-weighted homophily, we first evaluated a CRF07\_BC network with national survey data from China (Ge et al., 2021). Each of the 8178 pol sequences was annotated with a transmission risk factor: heterosexual contact (Hetero), people who use injection drugs (PWID), or men who have sex with men (MSM), among other attributes.

When we analyze the dataset with AUTO-TUNE, local maxima of AUTO-TUNE scores were achieved with 0.0076 sub/site and 0.0019 sub/site thresholds, at scores 1.137 and 1.030, respectively.

Notably, the DWH scores for PWID exhibited a significant surge at these thresholds, indicating a robust pattern of increased PWID homophily even when relatively low scoring. The close proximity of AUTO-TUNE scores and the consistent rise in PWID homophily at 0.0076 and 0.0019 thresholds suggest comparable performance at these thresholds compared to the default 0.015 threshold, suggesting that these thresholds might be more effective in representing homophilic relationships in this network. At each threshold—0.015, 0.0076, and 0.0019—all DWH scores for the risk groups (MSM, Hetero, and PWID) lie outside their respective panmictic ranges. This consistently indicates non-random mixing and attribute-based homophily across the network. Detailed results are in Tables 7, 8.

### 3.7 Comparison with clustner

We benchmarked AUTO-TUNE versus clustner (Chato et al., 2020), which employs the recency of sample collection or diagnosis as individual-level weights in a predictive model to

TABLE 4 Comparison of AUTO-TUNE and published thresholds from prior studies using sequences from viruses other than HIV-1. "N/A": no distance-based clustering analyses were done. Other notation is the same as in Table 1.

References	Virus	Gene	N	L	E [D] (%)	Scope	Location/Country	Timespan	Distance threshold, subs/site	
									Published	AUTO-TUNE (score)
Jia et al. (2023)	HCV	NS5B	503	315	34.9	Province	Yunnan, China	2008–2018	N/A	1.933 (1.92)
			97		8.0	Genotype 1b only			♣ 2.3	1.944 (2.0)
			53		7.4	Genotype 2a only			♣ 3.3	3.3 (1.3)
			110		5.4	Genotype 3a only			♣ 2.0	3.6 (1.0)
			189		5.5	Genotype 3b only			♣ 1.7	1.6 (1.0)
Murphy et al. (2019b)	HCV	NS5B	119	340–850	5.6	Province	Quebec, Canada	2001–2017	N/A	0.0251 (1.05)
Paraschiv et al. (2017)	HCV	NS5B	117	~ 300	24.6	Country	Romania	2011–2014	N/A	1.394 (1.11)
Ye et al. (2023)	HCV	NS5B	1603/399 •	701	27.6	Country	China	1999–2017	♣ 0.010	0.359 (1)
		C/E2	865/396 •	837	37.3				♣ 0.0325	1.572 (1.98)
Ng et al. (2022)	Rhinovirus	VP2/VP4	977	388	43.2	City	Kuala Lumpur, Malaysia	2012–2014	♣ 0.005	0.523 (1)

estimate the growth of HIV clusters. The thresholds deemed optimal by `clustuneR` were found by a grid-search for the minimum GAIC (generalized Akaike Information Criterion) across candidate distances between 0 and 0.04 in steps of  $8 \times 10^{-4}$ . GAIC is the difference between a null model that is only influenced by cluster size, and a weighted model that includes individual-level attributes among known cases in the cluster. Using the minimum GAIC metric, it was found that 0.016 ( $\pm 0.5 \times 10^{-4}$ ) was the optimal threshold for Tennessee and Seattle, and 0.0104 for Northern Alberta.

In contrast, AUTO-TUNE does not incorporate any attribute data in its scoring heuristic. Instead, it relies on clustering metrics constructed purely from pairwise distances between sequences. Using nearly same datasets analyzed by `clustuneR` (Chato et al., 2020), AUTO-TUNE found the thresholds with the highest scores to be 0.01872 for Middle Tennessee, 0.01538 for Seattle, and 0.01201 for Northern Alberta Table 9. We use the adjective “nearly” because we were not able to exactly match the number of sequences analyzed in Chato et al. (2020) by obtaining the referenced GenBank accession number and our best-effort interpretation of the filtering steps.

Both methods agree that there is a qualitative relationship of Northern Alberta < Seattle ~Tennessee for distance thresholds. AUTO-TUNE thresholds, while not optimal in the GAIC sense all yield improvements over the null model, hence they are qualitatively similar to `clustuneR` (Figure 2 in Chato et al. (2020)). AUTO-TUNE is notably faster in computation than `clustuneR` due to the fact that AUTO-TUNE only clusters based on pairwise distances rather than inferring a maximum-likelihood phylogeny. For example, the entire pipeline for the Seattle dataset took less than 16 s on an Apple M1 Max. Alternatively, the tree

inference step alone with `clustuneR` takes several hours to complete.

Because the methods optimize very different objectives and `clustuneR` makes use of additional data, broad agreement between the inferred thresholds is encouraging.

### 3.8 The effect of subsampling on optimal thresholds and AUTO-TUNE scores

To address the challenges of applying network inference algorithms to incompletely sampled datasets, this study includes a focused evaluation of AUTO-TUNE’s performance across varying data densities. Given logistical limitations, obtaining a fully sampled HIV transmission network is often infeasible. Therefore, we label a dataset as ‘full’ to serve as a closest approximation of a fully sampled network. Using the selected dataset as a benchmark, we assess AUTO-TUNE’s adaptability and robustness when applied to sparser datasets, a prevalent issue in real-world settings. In this analysis there is no expectation that any specific fraction of underlying infections was sampled, but simply that the complete dataset acts as the upper bound for inference. This was done in Dasgupta et al. (2019), for example,.

Since the Rhee et al. (2019) dataset exhibited a clear optimal peak, we used the dataset for analysis, and randomly sampled 10 times from the entire dataset at 25%, 50%, and 75% each. The original full dataset confidently determined 0.01699 (AUTO-TUNE score 1.9998).

Sampling at 25% yielded a mean top threshold of 0.021509, median at 0.019765, and standard deviation of 0.004388 (Figure 3). 50% yielded 0.018581 and 0.01871 mean and median, respectively

**TABLE 5** An application of AUTO-TUNE to subtype stratified HIV-1 pol sequences from the LANL database. Fraction clustered is the proportion of all sequences that are connected to at least one other sequence. Subtypes are sorted by the inferred threshold, lowest first. Other notation is the same as in Table 1.

Subtype	N	E [D]	AUTO-TUNE		Fraction	Mean	$\rho$
			Threshold, %	Score	Clustered, %	Degree	
CRF55	2237	2.6	0.187	1.20	29.6	1.41	2.41
CRF07	11682	3.3	0.26	1.00	26.9	3.42	1.87
CRF63	1649	3.6	0.505	1.01	22.1	4.85	1.8
01B	2237	7.8	0.518	1.08	22.2	0.97	5.05
A6	11991	4.6	0.558	1.09	18.6	5.55	1.6
CRF08	2538	3.9	0.82	1.95	25.6	1.95	2.12
CRF01	25689	5.1	0.875	1.73	47.0	1.94	5.54
B	106261	6.4	1.084	2.00	46.4	4.77	1.95
D	3561	6.6	1.133	1.12	20.8	3.65	0.79
C	30714	6.7	1.438	2.00	19.3	1.26	2.22
A1	7154	7.0	1.89	2.00	17.0	5	1.7
CRF02	7821	6.3	1.97	1.01	34.3	10.44	1.53
BF1	4825	8.1	2.046	1.03	25.1	2.27	1.95
G	2162	7.3	2.407	1.03	49.0	9.1	1.66
F1	3986	7.6	2.941	1.34	50.4	15.03	1.40
A1D	1284	8.9	3.23	1.70	27.5	1	4.3
BC	2724	8.0	3.54	1.00	81.4	71.2	1.32
Wertheim et al. (2014)	84527		1.0	N/A	15.6	3.84	1.74

**TABLE 6** Distance thresholds and key network properties using four different HIV-1 genomic regions, stratified by subtype (minimum 500 sequences).

Subtype	N	AUTO-TUNE $D$ , subs/site				Number of clusters				Full agreement	Krippendorff $\alpha$
		pr + rt	Integrase	gag	gp41	pr + rt	Integrase	gag	gp41	Clusters	
B	1843	0.02081	2.0005	3.137	5.095	115	128	119	144	64	0.723
C	877	0.03266	0.02	4.754	5.325	44	35	47	46	21	0.588
CRF01/AE	624	0.01635	0.818	2.285	2.037	40	30	40	41	12	0.610

**TABLE 7** CRF07\_BC nodes count at different thresholds.

Threshold subs/site	AUTO-TUNE score	Nodes	PWID	MSM	Hetero
0.015	0.029	5923	559	3371	1993
0.0076	1.1369	3537	236	2271	1030
0.0019	1.0303	1654	151	1075	428

with a standard deviation of 0.001629. Finally, 75% calculated mean is approximately 0.017403, with a median of approximately 0.01699. The standard deviation was 0.000924.

As the dataset becomes sparser due to subsampling, the algorithm tends to select higher distance thresholds. This

phenomenon can be understood by considering the effect of reduced sampling density on the network topology. Sparse datasets naturally result in less interconnected clusters. To capture a comparable level of network connectivity as in denser datasets, higher distance thresholds are necessary. This is evidenced

TABLE 8 Panmictic ranges for CRF07\_BC DWH at different thresholds.

Threshold subs/site	Risk group	DWH (panmictic range)
0.015	MSM	0.211 (−0.213, −0.085)
	Hetero	0.133 (−0.190, −0.087)
	PWID	0.168 (−0.091, 0.002)
0.0076	MSM	0.237 (−0.240, −0.120)
	Hetero	0.185 (−0.211, −0.100)
	PWID	0.401 (−0.081, −0.005)
0.0019	MSM	0.292 (−0.280, −0.146)
	Hetero	0.250 (−0.256, −0.093)
	PWID	0.445 (−0.129, −0.012)

TABLE 9 ClustuneR Comparison.

Dataset	AUTO-TUNE			clustuneR	
	Threshold subs/site	Avg. Homophily	Max score	Threshold	Avg. Homophily
Seattle	0.01354	0.0348	1.53325	0.0160	0.0259
Tennessee	0.01431	0.0147	1.25807	0.0160	0.0079
Canada	0.01099	−0.0448	1.01678	0.0104	−0.0536

by the observed mean thresholds: 0.021509 at 25%, 0.018581 at 50%, and 0.017403 at 75%. The standard deviations also narrow as the sampling density increases, corroborating the increased precision of the threshold selection in denser datasets.

As the proportion increased from 25% to 50% and 75%, observable shifts were also noted in the mean, median, and standard deviation of the AUTO-TUNE scores. At 25%, the mean and median scores were 1.5585 and 1.5014 respectively, with a standard deviation of 0.3568. At 50%, both mean and median scores significantly increased to 1.8171 and 1.9191 respectively, and the standard deviation dropped to 0.2482. Upon reaching an AUTO-TUNE of 75%, the mean and median scores rose further to 1.9870 and 1.9997 respectively, while the standard deviation shrank substantially to 0.0364, indicating higher consistency in scores.

Next to determine how well subsampled datasets aligned with the full dataset, we used two primary outcomes to gauge this concordance: the proportion of nodes that remained clustered after subsampling and the proportion of singletons from the original network that clustered in the subsampled networks.

We observed a consistent increase in the proportion of nodes that remained clustered from the 0.015 sub/site threshold to the AUTO-TUNE threshold for each respective subsampling proportion, with 25% subsampling being the most profound difference rising from a roughly 80%–86% interquartile range (IQR) for 0.015 threshold to a 90%–96% IQR for AUTO-TUNE, which indicates that the AUTO-TUNE thresholds retain a higher degree of stability in the network's structure across sampling density (Please see Figure 6, Panel A).

Since the thresholds inferred by AUTO-TUNE for the subsampled networks were larger than the “fully” sampled

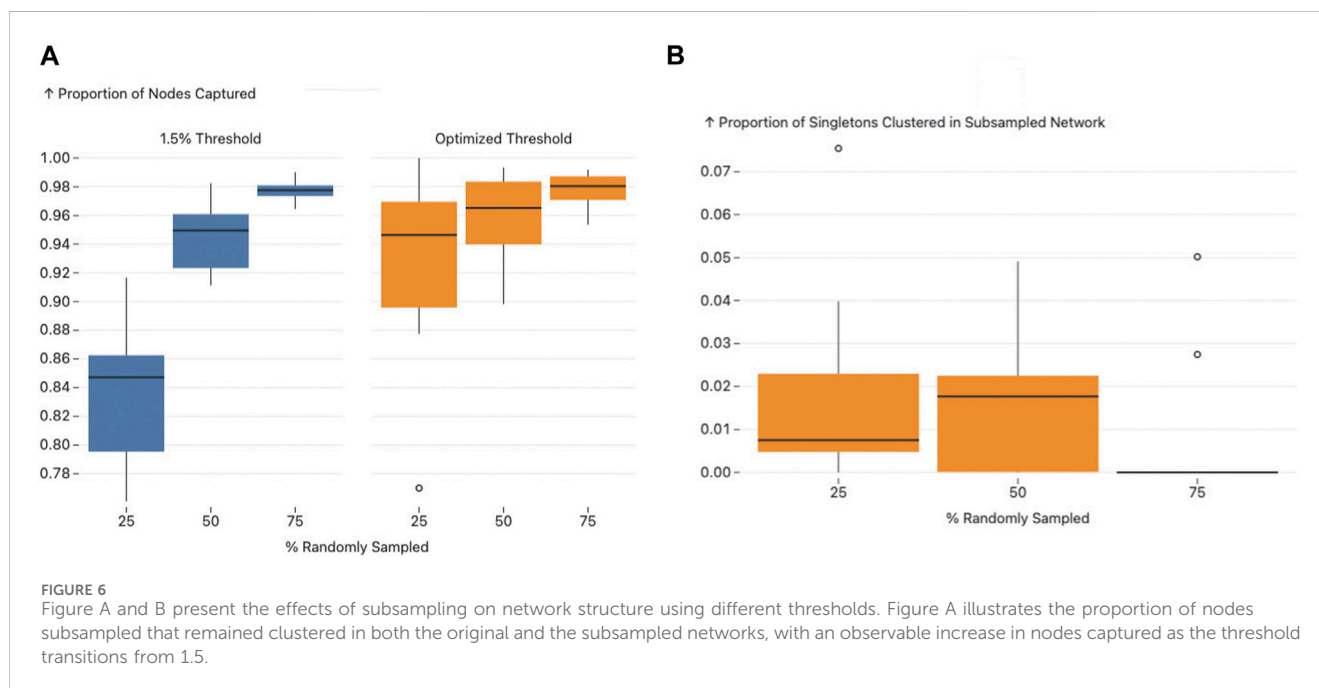
network, we also measured the impact of thresholding on the network's nodes that were originally singletons. Across all variations in subsampling rates, the proportion of sampled singletons that clustered all maintained low IQRs (See Figure 6, Panel B). This implies that while AUTO-TUNE is effective in maintaining the core structure of the network, it does not significantly alter the clustering of nodes that were singletons in the full dataset.

As the sample proportion increased, an upward trend was noted in average AUTO-TUNE scores. Additionally, the standard deviation reduced significantly when increasing sample proportion. This implies that as sampling becomes denser, AUTO-TUNE will become more confident in determining the optimal threshold for a particular dataset.

## 4 Discussion

AUTO-TUNE addresses the challenge of selecting an appropriate genetic distance threshold to construct HIV transmission networks by implementing a heuristic scoring system. This system is predicated on two key features of networks generated by candidate genetic distance thresholds: a high number of clusters and the absence of a giant component. Few small clusters indicate an excessively low threshold, while a giant cluster comprising numerous sequences signals an overly high threshold. The efficacy of AUTO-TUNE is evidenced by its ability to select thresholds that yield higher quality clustering, as demonstrated by improved Degree-Weighted Homophily (DWH) scores across various datasets, epidemic contexts, and risk groups.





Furthermore, AUTO-TUNE thresholds not only matched but often outperformed those manually selected in prior studies, thus underlining the benefits of a more systematic, automated, and data-responsive approach.

For example, the results of our study suggest that AUTO-TUNE, which relies solely on clustering metrics from pairwise distances, could be an effective alternative to other distance-based methods, such as `clusterR` while less time-consuming and possessing a gentle learning curve, which makes it easy to use by personnel not specialized in bioinformatics and computer science. Furthermore, the simplicity of the method without compromising results represents an advantage over phylogenetic methods where, in addition to the calculation of genetic distances, it must also determine a support/distance threshold where a rationale for the selection of these thresholds is rarely provided (Junqueira et al., 2019).

AUTO-TUNE generated thresholds for all three examined datasets (Middle Tennessee, Seattle, and Northern Alberta) that outperformed `clusterR` using DWH on 3-year collection date windows across all three datasets. This indicates that even without incorporating attribute data, AUTO-TUNE's scoring heuristic could provide reliable thresholds for HIV clusters. However, for the determination of the optimal genetic distance threshold, time-related and context-specific factors might need to be considered if there is no significant score for any one candidate threshold, especially if there are multiple peaks. For example, during HIV outbreaks in injection drug users (that usually occur over several months), it may be more appropriate to use the shorter genetic distance threshold (Peters et al., 2016; Campbell et al., 2017) between multiple high-scoring thresholds. On the contrary, larger and more extended epidemics over time exhibit a tendency toward larger genetic distance thresholds in order to capture transmission than younger epidemics and less densely sampled epidemic investigations (Leung et al., 2019; Di Giallonardo et al., 2021; Patil et al., 2022).

Our review of publications citing HIV-TRACE revealed the largely qualitative determination of distance thresholds. This

approach may result in less accurate or suboptimal thresholds due to a lack of systematic analysis. In contrast, AUTO-TUNE offers a more systematic and granular approach to threshold selection, with our findings demonstrating that even minor adjustments to the distance can drastically change the score. Therefore, using AUTO-TUNE could potentially improve the quality of HIV clustering and transmission network studies.

The Degree-Weighted Homophily (DWH) evaluation showed that AUTO-TUNE could improve network quality based on specific attributes, such as risk factor, which is an important part of HIV studies and informing prevention measures (Potterat et al., 2002; Fujimoto et al., 2021). For example, the use of AUTO-TUNE resulted in an increased DWH among the MSM, Hetero, and PWID groups when analyzing a CRF07\_BC network. Additionally, the results from the Rhee et al. dataset also demonstrated AUTO-TUNE's ability to improve DWH geographically, enhancing the network's ability to accurately reflect transmission dynamics. However, in contexts with overlapping risk factors, the interpretation of these improvements requires caution. The complexities of risk group interactions mean that applying AUTO-TUNE's thresholds should be tailored to the specific epidemiological setting to ensure accurate modeling of HIV transmission networks. More broadly, the ultimate impact of AUTO-TUNE on network quality and interpretation will be case specific, strongly dependant on how clusters are ultimately used in the study, and what types of data in addition to sequences alone are available.

Our analysis of AUTO-TUNE's performance on subsamples of a dataset revealed its sensitivity to sample size. The results indicated a correlation between increased sample size and higher average AUTO-TUNE scores, as well as lower score variability. This suggests that denser sampling could enhance AUTO-TUNE's ability to determine the optimal threshold for a dataset. Further studies might be needed to establish the minimum sample size required for reliable threshold determination.

## 4.1 When a score is below 1.9

In some cases, multiple scores at different thresholds could suggest the presence of inherently different scales in the network. For instance, if a network combines both global and local transmission patterns, AUTO-TUNE may produce more than one high score, reflecting these different scales. This was observed in a study on HIV-1 CRF07\_BC transmission networks in China, where two distinct clusters, 07BC\_N and 07BC\_O, showed different transmission routes and geographic concentrations (Ding et al., 2022). Such network complexities could mean that different thresholds might offer more accurate insights into subpopulations or transmission dynamics.

The use of AUTO-TUNE, while offering a method for automated threshold selection, may not always provide a single, decisive score that unambiguously determines the optimal threshold. In certain situations, such as datasets with lower sampling densities or those reflecting heterogenous dynamics within an epidemic, several candidate thresholds may yield similar AUTO-TUNE scores, making it difficult to single out one as the clear-cut 'optimal' threshold. In these scenarios, the process of threshold selection becomes more nuanced and requires a deeper analysis. The plot of AUTO-TUNE scores across candidate thresholds can serve as a valuable tool in these cases. For instance, researchers could identify a range of thresholds that all produce similar scores, suggesting that the specific choice of threshold within this range may not significantly impact the resulting network. Moreover, combining AUTO-TUNE with the DWH measure can enhance the interpretation of such plots. By considering how assortativity changes across the range of candidates, researchers can make more informed decisions about the appropriate choice. If there is a certain threshold at which the DWH measure noticeably changes for an attribute of interest, this could suggest a meaningful shift in the network structure that would be worth considering when selecting a threshold. The symbiotic approach of combining AUTO-TUNE scores, DWH measure, and visual analysis of score plots provides a more nuanced method for threshold selection when no clear optimal threshold emerges from the AUTO-TUNE scores alone.

The AUTO-TUNE methodology has several limitations. First, even though it provides the advantage of operating without the need for metadata, the size and the subgenomic region analyzed may affect the accuracy of transmission inference (Junqueira et al., 2019). Second, our analysis of AUTO-TUNE's performance on subsamples of a dataset revealed its sensitivity to sample size, as the performance of the method can be affected by sampling density, improving the reliability of the test as the sampling density increases. However, our results were consistent with previous studies, which have suggested an optimal sampling density of 50–70% for HIV-1 cluster analysis (Novitsky et al., 2014), and the significant drop-off in power to detect clusters using a fixed distance threshold as the sampling fraction was reduced (Dasgupta et al., 2019). Third, even when it provides an insight of the optimal threshold to analyze a network, the supplied information might still need validation by experts, especially when no clear threshold is identified. In this case, it has been recommended to combine genetic data with clinical and sociodemographic information for a better characterization of the network structure. Finally, the performance of the method needs to be assessed in pathogens different from HIV, leading to opportunities for future research.

## 5 Conclusion

AUTO-TUNE operates solely utilizing genetic sequence data to ascertain a decisive threshold. It employs a scoring heuristic, which is based on the number of clusters produced by a pairwise distance threshold and the ratio of the largest cluster to the second largest across a range of possible thresholds using sliding windows.

A key advantage of this approach is its autonomy from supplementary data. When a patient receives an HIV diagnosis, data collection protocols can greatly vary, and additional data are not always available or consistent. However, by leveraging only genetic sequence data, AUTO-TUNE eliminates the need for such information in some cases, and at minimum serves as a preliminary assessment of candidate thresholds.

Consequently, AUTO-TUNE's performance is consistently controlled, irrespective of the fluctuations seen in data collection protocols after an HIV diagnosis. This level of adaptability demonstrates its suitability for integration into various contexts related to HIV, and possibly other viral cluster detection and response protocols. This versatility underscores the strong methodological foundation of AUTO-TUNE and its potential utility.

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://github.com/veg/autotune-paper>.

## Author contributions

SW: Writing–review and editing, Writing–original draft, Visualization, Supervision, Software, Project administration, Methodology, Formal Analysis, Data curation, Conceptualization. VD: Writing–review and editing, Writing–original draft, Data curation. DJ: Writing–review and editing, Software. HV: Writing–review and editing, Writing–original draft, Visualization. SÁ-R: Writing–review and editing. AL: Conceptualization, Writing–review and editing. JW: Writing–review and editing, Conceptualization. SK: Writing–review and editing, Writing–original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Funding acquisition, Data curation, Conceptualization.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article. SK and SW were supported in part by grant funding from the NIH, grants AI134384, AI140970, GM144468, GM110749, GM151683. JOW was supported in part by AI135992.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

- Abidi, S. H., Aibekova, L., Davlidova, S., Amangeldiyeva, A., Foley, B., and Ali, S. (2021). Origin and evolution of HIV-1 subtype A6. *PLoS One* 16, e0260604. doi:10.1371/journal.pone.0260604
- Bartlett, S. R., Wertheim, J. O., Bull, R. A., Matthews, G. V., Lamoury, F. M., Scheffler, K., et al. (2017). A molecular transmission network of recent hepatitis c infection in people with and without hiv: implications for targeted treatment strategies. *J. viral Hepat.* 24, 404–411. doi:10.1111/jvh.12652
- Bbosa, N., Ssemwanga, D., and Kaleebu, P. (2020). Short communication: choosing the right program for the identification of HIV-1 transmission networks from nucleotide sequences sampled from different populations. *AIDS Res. Hum. retroviruses* 36, 948–951. doi:10.1089/AID.2020.0033
- Billings, E., Kijak, G. H., Sanders-Buell, E., Ndemi, N., O'Sullivan, A. M., Adebajo, S., et al. (2019). New subtype b containing hiv-1 circulating recombinant of sub-saharan africa origin in nigerian men who have sex with men. *J. Acquir Immune Defic. Syndr.* 81, 578–584. doi:10.1097/QAI.0000000000002076
- Boender, T. S., Smit, C., Sighem, A. v., Bezemer, D., Ester, C. J., Zaheri, S., et al. (2018). AIDS therapy evaluation in The Netherlands (ATHENA) national observational HIV cohort: cohort profile. *BMJ Open* 8, e022516. doi:10.1136/bmjopen-2018-022516
- Brenner, B. G., Ibanescu, R.-I., Osman, N., Cuadra-Foy, E., Oliveira, M., Chaillon, A., et al. (2021). The role of phylogenetics in unravelling patterns of HIV transmission towards epidemic control: the quebec experience (2002–2020). *Viruses* 13, 1643. doi:10.3390/v13081643
- Campbell, E. M., Jia, H., Shankar, A., Hanson, D., Luo, W., Masciotra, S., et al. (2017). Detailed transmission network analysis of a large opiate-driven outbreak of HIV infection in the United States. *J. Infect. Dis.* 216, 1053–1062. doi:10.1093/infdis/jix307
- Campigotto, A., Chris, A., Orkin, J., Lau, L., Marshall, C., Bitnun, A., et al. (2023). Utility of SARS-CoV-2 genomic sequencing for understanding transmission and school outbreaks. *Pediatr. Infect. Dis. J.* 42, 324–331. doi:10.1097/INF.0000000000003834
- Chan, C. P., Uemura, H., Kwan, T. H., Wong, N. S., Oka, S., Chan, D. P. C., et al. (2020). Review on the molecular epidemiology of sexually acquired hepatitis c virus infection in the asia-pacific region. *J. Int. AIDS Soc.* 23, e25618. doi:10.1002/jia2.25618
- Chato, C., Kalish, M. L., and Poon, A. F. Y. (2020). Public health in genetic spaces: a statistical framework to optimize cluster-based outbreak detection. *Virus Evol.* 6, veaa011. doi:10.1093/ve/veaa011
- Chen, Y., Lan, G., Feng, Y., Ruan, Y., Shen, Z., McNeil, E. B., et al. (2023). Inferring potential non-disclosed men who have sex with men among self-reported heterosexual men with hiv in southwest China: a genetic network study. *PLoS One* 18, e0283031. doi:10.1371/journal.pone.0283031
- Dalai, S. C., Junqueira, D. M., Wilkinson, E., Mehra, R., Kosakovsky Pond, S. L., Levy, V., et al. (2018). Combining phylogenetic and network approaches to identify HIV-1 transmission links in san mateo county, California. *Front. Microbiol.* 9, 2799. doi:10.3389/fmicb.2018.02799
- Dasgupta, S., France, A. M., Brandt, M.-G., Reuer, J., Zhang, T., Panneer, N., et al. (2019). Estimating effects of HIV sequencing data completeness on transmission network patterns and detection of growing HIV transmission clusters. *AIDS Res. Hum. Retroviruses* 35, 368–375. doi:10.1089/AID.2018.0181
- Di Giallonardo, F., Pinto, A. N., Keen, P., Shaik, A., Carrera, A., Salem, H., et al. (2021). Subtype-specific differences in transmission cluster dynamics of HIV-1 B and CRF01\_ae in New South Wales, Australia. *J. Int. AIDS Soc.* 24, e25655. doi:10.1002/jia2.25655
- Ding, X., Chaillon, A., Pan, X., Zhang, J., Zhong, P., He, L., et al. (2022). Characterizing genetic transmission networks among newly diagnosed HIV-1 infected individuals in eastern China: 2012–2016. *PLoS ONE* 17, e0269973. Publisher: Public Library of Science. doi:10.1371/journal.pone.0269973
- Dunn, D., and Pillay, D. (2007). UK HIV drug resistance database: background and recent outputs. *J. HIV Ther.* 12, 97–98.
- Erly, S. J., Naismith, K., Kerani, R., Buskin, S. E., and Reuer, J. R. (2021). Predictive value of time-space clusters for HIV transmission in Washington state, 2017–2019. *J. Acquir. Immune Defic. Syndromes* 87, 912–917. doi:10.1097/QAI.0000000000002675
- Fabeni, L., Santoro, M. M., Lorenzini, P., Rusconi, S., Gianotti, N., Costantini, A., et al. (2020). Evaluation of hiv transmission clusters among natives and foreigners living in Italy. *Viruses* 12, 791. doi:10.3390/v12080791
- Foster, G. M., Ambrose, J. C., Hué, S., Delpech, V. C., Fearnhill, E., Abecasis, A. B., et al. (2014). Novel HIV-1 recombinants spreading across multiple risk groups in the United Kingdom: the identification and phylogeography of circulating recombinant form (crf) 50\_a1d. *PLoS One* 9, e83337. doi:10.1371/journal.pone.0083337
- Fujimoto, K., Bahl, J., Wertheim, J. O., Del Vecchio, N., Hicks, J. T., Damodaran, L., et al. (2021). Methodological synthesis of Bayesian phylodynamics, HIV-TRACE, and GEE: HIV-1 transmission epidemiology in a racially/ethnically diverse Southern U.S. context. *Sci. Rep.* 11, 3325. doi:10.1038/s41598-021-82673-8
- Ge, Z., Feng, Y., Zhang, H., Rashid, A., Zaongo, S. D., Li, K., et al. (2021). HIV-1 CRF07\_BC transmission dynamics in China: two decades of national molecular surveillance. *Emerg. Microbes Infect.* 10, 1919–1930. doi:10.1080/22221751.2021.1978822
- Golub, B., and Jackson, M. (2012). Network structure and the speed of learning measuring homophily based on its consequences. *Ann. Econ. Statistics*, 33–48. doi:10.2307/23646571
- Goodreau, S. M. (2006). Assessing the effects of human mixing patterns on human immunodeficiency virus-1 interhost phylogenetics through social network simulation. *Genetics* 172, 2033–2045. doi:10.1534/genetics.103.024612
- Gore, D. J., Schueler, K., Ramani, S., Uvin, A., Phillips, G., McNulty, M., et al. (2022). HIV response interventions that integrate HIV molecular cluster and social network analysis: a systematic review. *AIDS Behav.* 26, 1750–1792. doi:10.1007/s10461-021-03525-0
- Grabowski, M. K., Herbeck, J. T., and Poon, A. F. Y. (2018). Genetic cluster analysis for hiv prevention. *Curr. HIV/AIDS Rep.* 15, 182–189. doi:10.1007/s11904-018-0384-1
- Hayes, A. F., and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Commun. Methods Meas.* 1, 77–89. doi:10.1080/19312450709336664
- Holmes, E. C., Zhang, L. Q., Robertson, P., Cleland, A., Harvey, E., Simmonds, P., et al. (1995). The molecular epidemiology of human immunodeficiency virus type 1 in Edinburgh. *J. Infect. Dis.* 171, 45–53. doi:10.1093/infdis/171.1.45
- Inzaule, S. C., Siedner, M. J., Little, S. J., Avila-Rios, S., Ayitewala, A., Bosch, R. J., et al. (2023). Recommendations on data sharing in hiv drug resistance research. *PLoS Med.* 20, e1004293. doi:10.1371/journal.pmed.1004293
- Jia, Y., Zou, X., Yue, W., Liu, J., Yue, M., Liu, Y., et al. (2023). The distribution of hepatitis C viral genotypes shifted among chronic hepatitis c patients in yunnan, China, between 2008–2018. *Front. Cell Infect. Microbiol.* 13, 1092936. doi:10.3389/fcimb.2023.1092936
- Jombart, T., Eggo, R. M., Dodd, P. J., and Balloux, F. (2011). Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity* 106, 383–390. doi:10.1038/hdy.2010.78
- Junqueira, D. M., Sibisi, Z., Wilkinson, E., and de Oliveira, T. (2019). Factors influencing HIV-1 phylogenetic clustering. *Curr. Opin. HIV AIDS* 14, 161–172. doi:10.1097/COH.0000000000000540
- Kosakovsky Pond, S. L., Posada, D., Stawiski, E., Chappay, C., Poon, A. F., Hughes, G., et al. (2009). An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in hiv-1. *PLoS Comput. Biol.* 5, e1000581. doi:10.1371/journal.pcbi.1000581
- Kosakovsky Pond, S. L., Weaver, S., Leigh Brown, A. J., and Wertheim, J. O. (2018). HIV-TRACE (TRANmission cluster engine): a tool for large scale molecular epidemiology of HIV-1 and other rapidly evolving pathogens. *Mol. Biol. Evol.* 35, 1812–1819. doi:10.1093/molbev/msy016
- Leal, E., Arrais, C. R., Barreiros, M., Farias Rodrigues, J. K., Silva Sousa, N. P., Duarte Costa, D., et al. (2020). Characterization of HIV-1 genetic diversity and antiretroviral resistance in the state of maranhão, northeast Brazil. *PLoS One* 15, e0230878. doi:10.1371/journal.pone.0230878
- Leung, K. S.-S., To, S. W.-C., Chen, J. H.-K., Siu, G. K.-H., Chan, K. C.-W., and Yam, W.-C. (2019). Molecular characterization of HIV-1 minority subtypes in Hong Kong: a recent epidemic of CRF07\_bc among the men who have sex with men population. *Curr. HIV Res.* 17, 53–64. doi:10.2174/1570162X17666190530081355
- Li, D., Chen, H., Li, H., Ma, Y., Dong, L., Dai, J., et al. (2022). HIV-1 pretreatment drug resistance and genetic transmission network in the southwest border region of China. *BMC Infect. Dis.* 22, 741. doi:10.1186/s12879-022-07734-3
- Little, S. J., Kosakovsky Pond, S. L., Anderson, C. M., Young, J. A., Wertheim, J. O., Mehta, S. R., et al. (2014). Using hiv networks to inform real time prevention interventions. *PLoS One* 9, e98443. doi:10.1371/journal.pone.0098443
- Liu, M., Han, X., Zhao, B., An, M., He, W., Wang, Z., et al. (2020). Dynamics of HIV-1 molecular networks reveal effective control of large transmission clusters in an area affected by an epidemic of multiple HIV subtypes. *Front. Microbiol.* 11, 604993. doi:10.3389/fmicb.2020.604993
- Mai, T. Q., Martinez, E., Menon, R., Van Anh, N. T., Hien, N. T., Marais, B. J., et al. (2018). *Mycobacterium tuberculosis* drug resistance and transmission among human immunodeficiency virus-infected patients in Ho chi minh city, vietnam. *Am. J. Trop. Med. Hyg.* 99, 1397–1406. doi:10.4269/ajtmh.18-0185

- Molloy, M., and Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random Struct. Algorithms* 6, 161–180. doi:10.1002/rsa.3240060204
- Murphy, D. G., Dion, R., Simard, M., Vachon, M. L., Martel-Laferrrière, V., Serhir, B., et al. (2019a). Molecular surveillance of hepatitis C virus genotypes identifies the emergence of a genotype 4d lineage among men in Quebec, 2001–2017. *Can. Commun. Dis. Rep. = Relevé Des. Mal. Transm. Au Can.* 45, 230–237. doi:10.14745/ccdr.v45i09a02
- Murphy, D. G., Dion, R., Simard, M., Vachon, M. L., Martel-Laferrrière, V., Serhir, B., et al. (2019b). Molecular surveillance of hepatitis c virus genotypes identifies the emergence of a genotype 4d lineage among men in quebec, 2001–2017. *Can. Commun. Dis. Rep.* 45, 230–237. doi:10.14745/ccdr.v45i09a02
- Ng, K. T., Ng, L. J., Oong, X. Y., Chook, J. B., Chan, K. G., Takebe, Y., et al. (2022). Application of a vp4/vp2-inferred transmission clusters in estimating the impact of interventions on rhinovirus transmission. *Viol. J.* 19, 36. doi:10.1186/s12985-022-01762-w
- Novitsky, V., Moyo, S., Lei, Q., DeGruttola, V., and Essex, M. (2014). Impact of sampling density on the extent of HIV clustering. *AIDS Res. Hum. retroviruses* 30, 1226–1235. doi:10.1089/aid.2014.0173
- Novitsky, V., Steingrimsson, J. A., Howison, M., Gillani, F. S., Li, Y., Manne, A., et al. (2020). Empirical comparison of analytical approaches for identifying molecular HIV-1 clusters. *Sci. Rep.* 10, 18547. doi:10.1038/s41598-020-75560-1
- Oster, A. M., France, A. M., Panneer, N., Bañez Ocfemia, M. C., Campbell, E., Dasgupta, S., et al. (2018). Identifying clusters of recent and rapid HIV transmission through analysis of molecular surveillance data. *J. Acquir. Immune Defic. Syndromes* 79, 543–550. doi:10.1097/QAI.0000000000001856
- Oster, A. M., Lyss, S. B., McClung, R. P., Watson, M., Panneer, N., Hernandez, A. L., et al. (2021). HIV cluster and outbreak detection and response: the science and experience. *Am. J. Prev. Med.* 61, S130–S142. doi:10.1016/j.amepre.2021.05.029
- Paraschiv, S., Banica, L., Nicolae, I., Niculescu, I., Abagiu, A., Jipa, R., et al. (2017). Epidemic dispersion of HIV and HCV in a population of co-infected Romanian injecting drug users. *PLoS One* 12, e0185866. doi:10.1371/journal.pone.0185866
- Paraskevis, D., Nikolopoulos, G. K., Magiorkinis, G., Hodges-Mameletzis, I., and Hatzakis, A. (2016). The application of HIV molecular epidemiology to public health. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* 46, 159–168. doi:10.1016/j.meegid.2016.06.021
- Patil, A., Patil, S., Rao, A., Gadhe, S., Kurlle, S., and Panda, S. (2022). Exploring the evolutionary history and phylodynamics of human immunodeficiency virus type 1 outbreak from unnao, India using phylogenetic approach. *Front. Microbiol.* 13, 848250. doi:10.3389/fmicb.2022.848250
- Penn, O., Stern, A., Rubinstein, N. D., Dutheil, J., Bacharach, E., Galtier, N., et al. (2008). Evolutionary modeling of rate shifts reveals specificity determinants in hiv-1 subtypes. *PLoS Comput. Biol.* 4, e1000214. doi:10.1371/journal.pcbi.1000214
- Pérez-Losada, M., Castel, A. D., Lewis, B., Kharfen, M., Cartwright, C. P., Huang, B., et al. (2017). Characterization of HIV diversity, phylodynamics and drug resistance in Washington, DC. *PLoS One* 12, e0185644. doi:10.1371/journal.pone.0185644
- Peters, P. J., Pontones, P., Hoover, K. W., Patel, M. R., Galang, R. R., Shields, J., et al. (2016). HIV infection linked to injection use of oxymorphone in Indiana, 2014–2015. *N. Engl. J. Med.* 375, 229–239. doi:10.1056/NEJMoa1515195
- Potterat, J. J., Phillips-Plummer, L., Muth, S. Q., Rothenberg, R. B., Woodhouse, D. E., Maldonado-Long, T. S., et al. (2002). Risk network structure in the early epidemic phase of HIV transmission in Colorado Springs. *Sex. Transm. Infect.* 78, i159–i163. doi:10.1136/sti.78.suppl\_1.i159
- Ragonnet-Cronin, M., Benbow, N., Hayford, C., Poortinga, K., Ma, F., Forgiione, L. A., et al. (2021). Sorting by race/ethnicity across hiv genetic transmission networks in three major metropolitan areas in the United States. *AIDS Res. Hum. retroviruses* 37, 784–792. doi:10.1089/aid.2020.0145
- Ragonnet-Cronin, M., Hayford, C., D'Aquila, R., Ma, F., Ward, C., Benbow, N., et al. (2022). Forecasting HIV-1 genetic cluster growth in Illinois, United States. *J. Acquir. Immune Defic. Syndromes* 89, 49–55. doi:10.1097/QAI.0000000000002821
- Ragonnet-Cronin, M., Hodcroft, E., Hué, S., Fearnhill, E., Delpech, V., Brown, A. J. L., et al. (2013). Automated analysis of phylogenetic clusters. *BMC Bioinforma.* 14, 317. doi:10.1186/1471-2105-14-317
- Ragonnet-Cronin, M., Hué, S., Hodcroft, E. B., Tostevin, A., Dunn, D., Fawcett, T., et al. (2018). Non-disclosed men who have sex with men in UK HIV transmission networks: phylogenetic analysis of surveillance data. *Lancet HIV* 5, e309–e316. doi:10.1016/S2352-3018(18)30062-6
- Rhee, S.-Y., Magalis, B. R., Hurley, L., Silverberg, M. J., Marcus, J. L., Slome, S., et al. (2019). National and international dimensions of human immunodeficiency virus-1 sequence clusters in a northern California clinical cohort. *Open Forum Infect. Dis.* 6, ofz135. doi:10.1093/ofid/ofz135
- Robinson, K., Fyson, N., Cohen, T., Fraser, C., and Colijn, C. (2013). How the dynamics and structure of sexual contact networks shape pathogen phylogenies. *PLOS Comput. Biol.* 9, e1003105. Publisher: Public Library of Science. doi:10.1371/journal.pcbi.1003105
- Rose, R., Cross, S., Lamers, S. L., Astemborski, J., Kirk, G. D., Mehta, S. H., et al. (2020). Persistence of HIV transmission clusters among people who inject drugs. *AIDS Lond. Engl.* 34, 2037–2044. doi:10.1097/QAD.0000000000002662
- Scherrer, A. U., Traytel, A., Braun, D. L., Calmy, A., Battegay, M., Cavassini, M., et al. (2022). Cohort profile update: the Swiss HIV cohort study (SHCS). *Int. J. Epidemiol.* 51, 33–34j. doi:10.1093/ije/dyab141
- Sivay, M. V., Hudelson, S. E., Wang, J., Ageyi, Y., Hamilton, E. L., Selin, A., et al. (2018). HIV-1 diversity among young women in rural South Africa: HPTN 068. *PLoS One* 13, e0198999. doi:10.1371/journal.pone.0198999
- Sizemore, L., Fill, M.-M., Mathieson, S. A., Black, J., Brantley, M., Cooper, K., et al. (2020). Using an established outbreak response plan and molecular epidemiology methods in an HIV transmission cluster investigation, Tennessee, January–June 2017. *Public Health Rep. Wash. D.C.* 1974 135, 329–333. doi:10.1177/0033354920915445
- Stecher, M., Chaillon, A., Eberle, J., Behrens, G. M. N., Eis-Hübinger, A.-M., Lehmann, C., et al. (2018). Molecular epidemiology of the hiv epidemic in three German metropolitan regions - Cologne/Bonn, Munich and Hannover, 1999–2016. *Sci. Rep.* 8, 6799. doi:10.1038/s41598-018-25004-8
- Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10, 512–526. doi:10.1093/oxfordjournals.molbev.a040023
- Temereanca, A., Oprea, C., Wertheim, J. O., Ianache, I., Ceausu, E., Cernescu, C., et al. (2017). HIV transmission clusters among injecting drug users in Romania. *Rom. Biotechnol. Lett.* 22, 12307–12315.
- Thoma, R., Seneghini, M., Seiffert, S. N., Vuichard Gysin, D., Scanferla, G., Haller, S., et al. (2022). The challenge of preventing and containing outbreaks of multidrug-resistant organisms and *Candida auris* during the coronavirus disease 2019 pandemic: report of a carbapenem-resistant *Acinetobacter baumannii* outbreak and a systematic review of the literature. *Antimicrob. Resist. Infect. Control* 11, 2. doi:10.1186/s13756-022-01052-8
- Tookes, H., Bartholomew, T. S., Geary, S., Matthias, J., Poschman, K., Blackmore, C., et al. (2020). Rapid identification and investigation of an HIV risk network among people who inject drugs -Miami, FL, 2018. *AIDS Behav.* 24, 246–256. doi:10.1007/s10461-019-02680-9
- Tumpney, M., John, B., Panneer, N., McClung, R. P., Campbell, E. M., Roosevelt, K., et al. (2020). Human immunodeficiency virus (HIV) outbreak investigation among persons who inject drugs in Massachusetts enhanced by HIV sequence data. *J. Infect. Dis.* 222, S259–S267. doi:10.1093/infdis/jiaa053
- Volz, E. M., Ndemi, N., Nowak, R., Kijak, G. H., Idoko, J., Dakum, P., et al. (2017). Phylodynamic analysis to inform prevention efforts in mixed hiv epidemics. *Virus Evol.* 3, vex014. doi:10.1093/ve/vex014
- von Rotz, M., Kuehl, R., Durovic, A., Zingg, S., Apitz, A., Wegner, F., et al. (2023). A systematic outbreak investigation of SARS-CoV-2 transmission clusters in a tertiary academic care center. *Antimicrob. Resist. Infect. Control* 12, 38. doi:10.1186/s13756-023-01242-y
- Vrancken, B., Adachi, D., Benedet, M., Singh, A., Read, R., Shafran, S., et al. (2017). The multi-faceted dynamics of HIV-1 transmission in Northern Alberta: a combined analysis of virus genetic and public health data. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* 52, 100–105. doi:10.1016/j.meegid.2017.04.005
- Wang, X., Wu, Y., Mao, L., Xia, W., Zhang, W., Dai, L., et al. (2015). Targeting HIV prevention based on molecular epidemiology among deeply sampled subnetworks of men who have sex with men. *Clin. Infect. Dis. Official Publ. Infect. Dis. Soc. Am.* 61, 1462–1468. doi:10.1093/cid/civ526
- Weaver, S., Shank, S. D., Spielman, S. J., Li, M., Muse, S. V., and Kosakovsky Pond, S. L. (2018). Datamonkey 2.0: a modern web application for characterizing selective and other evolutionary processes. *Mol. Biol. Evol.* 35, 773–777. doi:10.1093/molbev/msx335
- Wertheim, J. O., Leigh Brown, A. J., Hepler, N. L., Mehta, S. R., Richman, D. D., Smith, D. M., et al. (2014). The global transmission network of HIV-1. *J. Infect. Dis.* 209, 304–313. doi:10.1093/infdis/jit524
- Wolf, E., Herbeck, J. T., Van Rompaey, S., Kitahata, M., Thomas, K., Pepper, G., et al. (2017). Short communication: phylogenetic evidence of HIV-1 transmission between adult and adolescent men who have sex with men. *AIDS Res. Hum. retroviruses* 33, 318–322. doi:10.1089/AID.2016.0061
- Yan, H., He, W., Huang, L., Wu, H., Liang, Y., Li, Q., et al. (2020). The central role of nondisclosed men who have sex with men in human immunodeficiency virus-1 transmission networks in guangzhou, China. *Open Forum Infect. Dis.* 7, ofaa154. doi:10.1093/ofid/ofaa154
- Yan, H., Wu, H., Xia, Y., Huang, L., Liang, Y., Li, Q., et al. (2021). Acquisition and transmission of hiv-1 among migrants and Chinese in guangzhou, China from 2008 to 2012: phylogenetic analysis of surveillance data. *Infect. Genet. Evol.* 92, 104870. doi:10.1016/j.meegid.2021.104870
- Ye, J., Sun, Y., Li, J., Lu, X., Zheng, M., Liu, L., et al. (2023). Distribution pattern, molecular transmission networks, and photodynamic of hepatitis c virus in China. *PLoS One* 18, e0296053. doi:10.1371/journal.pone.0296053
- Yebrá, G., Ragonnet-Cronin, M., Ssemwanga, D., Parry, C. M., Logue, C. H., Cane, P. A., et al. (2015). Analysis of the history and spread of HIV-1 in Uganda using phylodynamics. *J. Gen. Virol.* 96, 1890–1898. doi:10.1099/vir.0.000107
- Yu, D., Liang, B., Yang, Y., Liu, J., Liang, H., Zhang, F., et al. (2022). Prevalence of drug resistance and genetic transmission networks among human immunodeficiency virus/acquired immunodeficiency syndrome patients with antiretroviral therapy failure in guangxi, China. *AIDS Res. Hum. Retroviruses* 38, 822–830. doi:10.1089/AID.2021.0181
- Zai, J., Liu, H., Lu, Z., Chaillon, A., Smith, D., Li, Y., et al. (2020). Tracing the transmission dynamics of hiv-1 crf55\_01b. *Sci. Rep.* 10, 5098. doi:10.1038/s41598-020-61870-x