ADVANCING MOLECULAR EPIDEMIOLOGY: ENHANCED METHODS AND APPLICATIONS IN PATHOGEN TRANSMISSION NETWORK ANALYSIS

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY

---

by
Steven Weaver
December 2024

Examining Committee Members:

Robert J. Kulathinal, Advisory Chair, Biology
Jody Hey, Biology
Ananias Escalante, Biology
Sudhir Kumar, Biology
Frederick A. Matsen IV, External Reader, Fred Hutchinson Cancer Research Center

ABSTRACT

The use of genomic information has gained increasing relevance in epidemiology and public health, particularly the inference of genetic networks of highly evolving pathogens like HIV. This approach offers objective evidence that allows for tracking transmission dynamics, and guiding intervention. However, the virus's high genetic diversity, combined with variations in risk factors and transmission rates among affected populations, leads to substantial differences in epidemic dynamics across different epidemiological contexts. Additionally, transmission networks inferred using a distance threshold often create many spurious edges among closely related sequences, resulting in abiological transmission cycles. Management of these networks as new sequences are generated also fuel a desire to track cluster growth over time. It is therefore advisable to develop molecular analysis tools that better capture the unique landscape of each epidemic, providing a portrayal that more closely mirrors the real epidemic scenario. This dissertation introduces novel tools aimed at optimizing genetic network analysis for epidemiological purposes, with a focus on HIV-1 but with stated potential further application in other pathogens.

Due to their greater convenience compared to phylogenetic methods, the inference of transmission networks through genetic distance methods has been widely used for near-real-time surveillance of HIV-1. Published research using the HIV-TRACE software package frequently uses the default threshold of 0.015 substitutions/site for HIV pol gene sequence. In Chapter 2, Weaver et al. introduces AUTO-TUNE, a heuristic scoring method to adaptively tune the distance threshold, to prevent giant cluster formation and maximize cluster numbers. This method was applied to various outbreaks, considering regional or temporal differences, to identify clusters with specific risk factors. For instance, we found the 0.015 threshold suits US-like epidemics, but a lower threshold of 0.005 better captures outbreaks like the CRF07 BC subtype among MSM in China. Conversely, a larger threshold is needed for regions where diverse risk factors and sparse sampling occur over longer periods. This adaptive approach aids public health officials in making more informed interventions.

Genetic distance and phylogenetic tools often yield densely connected clusters, with many spurious edges that create unrealistic cycles, overestimating connectivity and inflating node degrees. In Chapter 4, we present a scalable approach, Conditionally Orthogonal Vertices For Edge Filtering in Epidemics (COVFEFE), to prune spurious edges through a straightforward application of the PC (unroll) algorithm originally developed for Bayesian graphical models. COVFEFE removes edges likely to be indirect, preserving cluster membership. It eliminates 29%–80% of spurious edges in simulated transmission chains with low intra-host mutation rates and 4%–39% in real HIV-1 epidemic data, often simplifying dense clusters to chains. Filtered networks align better with theoretical models, affecting analysis and modeling methods that use node or edge properties. COVFEFE is part of the HIV-TRACE molecular network inference package.

The heterogeneity in risk factors and characteristics of the HIV-1 epidemic worldwide demands adapting bioinformatics tools for different contexts to meet local needs. In Chapter 4, we present HIV-TRACE GO, an extension of the HIV-TRACE software designed to enhance global HIV molecular epidemiology surveillance. This web application, developed primarily in JavaScript, enables real-time molecular surveillance with features tailored for multi-jurisdictional data management and internationalization. HIV-TRACE GO has been successfully implemented at CIENI-INER, a WHO-accredited HIV-1 sequencing reference laboratory in Mexico, demonstrating its feasibility for both local and international applications. We discuss the software architecture, key customizations, and its potential impact on public health surveillance.

I dedicate this dissertation to my parents, Rodger and April Weaver.

ACKNOWLEDGEMENTS

I would like to thank all those that were part of the UCSD CFAR Bioinformatics and Information Technologies(BIT) Core, especially Dr. Jason Young, David Mote, and Marco Martinez from 2011 - 2015 for providing the foundation and support that lead to further studies in bioinformatics. Their continuous positive feedback of my work in conjunction with exposure to how clinical trials are designed and managed, access to tours of the wet lab, and collaboration with medical doctors and statisticians both instilled in me the drive to continue learning and afforded me the opportunity to continue progressing within biotechnology.

I am thankful to Prof. Ben Murrell for continued consultation for both statistics and sequencing technologies. I would like to thank Prof. Joel Wertheim for the opportunity to work in molecular biology and epidemiology in such an impactful way. By extension, I would like to thank the Center for Disease Control's HIV Incidence and Case Surveillance Branch (HICSB) for their adoption of Secure HIV-TRACE for molecular health surveillance across the country.

Temple University, and especially the Institute for Genomics and Evolutionary Medicine has helped me grow as an individual and scientist immensely. I would like to especially thank Prof. Sudhir Kumar for explaining phylogenetic concepts and algorithms in such an elegant way, and taking the time to be available despite his demanding responsibilities as director of the institute. My gratitude also goes to program directors Dr. Rob Kulathinal Dr. Jody Hey for providing continued support throughout the duration of the program. Colleagues such as Dr. Alexander Lucaci, Dr. Craig Stanley, and Kevin Levitz also made for a welcoming environment in Philadelphia.

I am thankful to Vanessa in countless ways. The consistent feedback and contributions throughout the writing process has been invaluable.

Finally, I am thankful to Prof. Sergei L. Kosakovsky Pond for providing countless opportunties to develop professionally, technically, and academically.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1
INTRODUCTION

Advances in molecular biology have transformed the fields of epidemiology and public health by improving understanding of the pathogenesis and transmission dynamics of disease, leading to the emergence of molecular epidemiology in the 1970s and early 1980s(Tümmler, 2020). Molecular epidemiology provides valuable information for understanding the transmission, manifestation, and progression of infectious diseases. It is distinguished from other disciplines by its focus on identifying and maximizing intervention or disease prevention strategies to solve public health problems(Riley and Blanton, 2018). Molecular epidemiology is particularly relevant for infectious diseases such as HIV, where the use of genomic information allows for detailed analyses of pathogen evolution and transmission dynamics. Due to the frequent viral mutations as HIV spreads within a community network, each virus acquires a unique genetic footprint. This unique genetic sequence can be analyzed to identify potential epidemiological links between individuals carrying genetically similar viruses, forming what is known as a transmission network(German *et al.*, 2017). These networks offer objective insights into the transmission pathways, independent of participant-reported histories, which are valuable for tracking the spread of diseases and guiding public health interventions. By understanding the genetic relationships between viral sequences, researchers can identify transmission hotspots, monitor the effectiveness of interventions, and adapt strategies to emerging threats(Oster *et al.*, 2021a).

Genetic networks are composed of molecular clusters, which refers to a group of viral sequences that are closely related genetically, indicating transmission events among the individuals from whom these sequences were obtained. In these networks, sequences are represented as nodes, and the relationships between them are depicted as edges(Hevey, 2018). In network analysis, there the current key parameter that is frequently analyzed when characterizing and understanding these molecular clusters for public health purposes(Hevey, 2018; Oster *et al.*, 2018a) is:

1. **Recent and Rapidly Growing Clusters**: Identifying clusters with rapid HIV transmission for timely intervention. The national priority criteria states this as clusters detected at a $\leq$ 0.5% genetic distance threshold with $\geq$ 5 diagnoses in the past 12 months or $\geq$ 3 diagnoses in the past 12 months in low morbidity jurisdictions are prioritized.

This targeted approach enables health authorities to focus interventions where transmission is most likely. Actions such as proactive outreach, testing, and linkage to care can effectively reduce HIV spread within affected communities(Hassan *et al.*, 2020). Furthermore, timely identification of transmission clusters is highly desirable in molecular surveillance. Adding new cases to a molecular cluster indicates ongoing transmission and provides opportunities for intervention. Therefore, continuous monitoring of molecular clusters remains important once molecular surveillance is implemented. This ongoing surveillance helps health officials detect emerging outbreaks early, swiftly implement targeted interventions, and prevent further transmission within identified clusters.

The range of infectious disease epidemiological issues that can be addressed by the application of bioinformatic tools depends significantly on the reliability of the information these tools generate(Riley and Blanton, 2018). Equally important is ensuring that these tools are user-friendly and accessible to a diverse range of public health professionals, to effectively utilize these tools for surveillance, outbreak investigation, and intervention planning. This accessibility is necessary for scaling up the use of bioinformatic tools across public health systems, thereby enhancing their impact on infectious disease control and prevention efforts globally. Molecular clusters can be determined by either phylogenetic or genetic distance methods(Hassan *et al.*, 2017). This dissertation primarily explores genetic distance methods for determining molecular clusters. Unlike phylogenetic methods, genetic distance methods are favored for their simplicity, speed, and practicality. They are easy to implement and demand less computational power, facilitating near real-time surveillance for timely public health interventions(Wertheim *et al.*, 2019). A critical factor in defining molecular clusters is the genetic distance threshold, or cut-off, used to determine whether an edge should be drawn between two nodes. This cut-off defines how genetically similar two sequences must be to be considered linked. Choosing an appropriate cut-off is of utmost importance for accurately identifying meaningful clusters. If the cut-off is too strict, it may fail to capture all relevant transmission links, resulting in fragmented and incomplete clusters. Conversely, if the cut-off is too relaxed, it may create spurious connections, leading to overly large and biologically implausible clusters. To address this issue, Chapter 2 introduces AUTO-TUNE, a heuristic method that adaptively ”tunes” distance thresholds to better capture transmission clusters without generating overly large clusters. We prioritize the ratio of the sizes of the largest and the second largest cluster, and maximizing the number of clusters present in the network. The efficacy of AUTO-TUNE is demonstrated by its capability to choose thresholds that result in higher quality clustering, as evidenced by improved Degree-Weighted Homophily

(DWH) scores among related attributes observed across different datasets reflecting different epidemic scenarios. This approach proves robust to regional and temporal differences in epidemic dynamics, thereby providing more precise insights for public health officials.

Another challenge in network analysis is the tendency of genetic distance and phylogenetic tools to produce densely connected networks with many spurious edges. Such networks can create abiological cycles and overestimate connectivity, complicating the interpretation of transmission dynamics. Chapter 3 presents COVFEFE (Conditionally Orthogonal Vertices For Edge Filtering in Epidemics). This method utilizes a zero branch-length statistical test to determine whether nodes in a triangle are conditionally independent. Edges made redundant in triangles where one of the nodes are considered to have no branch length from the root are flagged for potential removal, except those that function as 'bridges' in the network to maintain cluster integrity. This iterative process continues until no further edges can be removed and can be extended to cycles of any length. With spurious edge reduction, intervention strategies may be able to incorporate more sophisticated measures that include node degree, centrality measurements, and epidemic model fitting. COVFEFE is part of the HIV-TRACE molecular network inference package. Recognizing the global heterogeneity of the HIV-1 epidemic, this dissertation also emphasizes the need for bioinformatics tools that are adaptable to different epidemiological contexts. In Chapter 4 HIV-TRACE GO, an extension of the HIV-TRACE software designed for global surveillance is introduced. Utilizing a web application developed primarily in JavaScript, HIV-TRACE GO offers real-time molecular surveillance. Translation into Spanish, customization of functionalities to manage data across multiple jurisdictions, and flexible attribute management are some of the customizations in HIV-TRACE GO. Its successful implementation at CIENI-INER, a WHO-accredited HIV-1 sequencing reference laboratory in Mexico, demonstrates its utility for both local and international public health applications. We delve into the software architecture, key customizations, and its potential impact on enhancing public health surveillance capabilities.

In summary, this dissertation contributes to the field of molecular epidemiology by developing and refining tools for genetic network analysis that address the specific challenges posed by HIV-1 (refinement of the genetic distance threshold, spurious edge filtering and expansion of bioinformatic tools for HIV surveillance), with potential applications for other pathogens. These contributions aim to improve the accuracy of transmission network inferences, enhance real-time surveillance capabilities, and provide actionable insights for public health interventions.

CHAPTER 2
AUTO-TUNE: SELECTING THE DISTANCE THRESHOLD FOR INFERRING HIV TRANSMISSION CLUSTERS

2.1 Introduction

The use of genomic data to infer and characterize transmission networks of various pathogens has grown in prominence in the past two 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 (Jombart *et al.*, 2011), and bacteria such as *M.tuberculosis* (Mai *et al.*, 2018) or *A.baumanii* (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). 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.*, 2021a, 2018a), monitoring response to an HIV outbreak (Tumpney *et al.*, 2020; Sizemore *et al.*, 2020; Tookes *et al.*, 2020), evaluating the effectiveness of interventions (Peters *et al.*, 2016; Wang *et al.*, 2015a; 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.*, 2022a). 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.

2.2 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.*, 2018a), where the genetic distance is computed using the Tamura-Nei (TN93) (Tamura and Nei, 1993a) 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.

2.2.1 Scoring Heuristic Procedure

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

1. 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.

2. 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 ith candidate, d_L^i with $W < i \leq N - W$ (W is a positive integer defined below).

   (a) The R_{12} jump is computed by looking at the normalized ratio of the mean R_{12} values computed over the leading window $d_L^{i+1} \ldots d_L^{i+W}$ and the trailing window $d_L^{i-W} \ldots d_L^{i-1}$. The width of the window, W, is defined as $\min\left(\max\left(\left[\frac{N}{100}\right], 3\right), 30\right)$. 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.

   (b) 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.

(c) 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.

3. 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.

   (a) No other thresholds have priority scores of 1.9 or higher

   (b) 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).

4. 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.

2.2.2 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 chanceGolub and Jackson (2012a). 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

$$DWH = \frac{W_{G,G} + W_{\bar{G},\bar{G}} - 2W_{G,\bar{G}}}{|G|^{-2} \sum_{i \in G} 1/d_i + |\bar{G}|^{-2} \sum_{i \in \bar{G}} 1/d_i}, \tag{2.1}$$

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.

2.2.3 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
```

5. 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.

2.2.3.1 Optional : Compute Assortativity Metrics.

6. 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.

7. 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.

2.2.4 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. 2.3 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/`).

2.2.5 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).

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 `clustuneR` method (Chato *et al.*, 2020). We procured datasets from (Wolf *et al.*, 2017) and (Vrancken *et al.*, 2017) utilizing the identical 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 `clustuneR` 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.*, 2019a). 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.

2.3 Results

2.3.1 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 (Table 2.3, 2.4). 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`) 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, ρ. 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.

2.3.1.1 Examples of How Changing Thresholds Affects Inferred Networks.

2.3.1.1.1 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 2 in that paper). A minute change to the threshold by AUTO-TUNE to 0.0194% 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 Fig 2.4 panel A). This minor change also reduced R_{12} from 21 to 2.6.

2.3.1.1.2 Cluster Size Increase. Increasing the 0.015% subs/site threshold on data from Little *et al.* (2014a) combined several small clusters (and singletons) into a single larger cluster, while preserving the overall size and properties of the network (see Fig 2.4 panel B). This change also reduced R_{12} from 2.5 to 1.5.

2.3.1.1.3 Thinning Out the Network. Reducing the 0.015% subs/site threshold on data from Rhee *et al.* (2019a) dramatically reduced the size of the largest cluster, and thinned out most clusters with five or more nodes (see Fig 2.4 panel C).

2.3.1.1.4 Materially Changing the Degree Distribution of the Network. For the sequences from Li *et al.* (2022), AUTO-TUNE suggests $D = 1.483\%$ 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 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 Fig 2.4 panel D shows, the degree distribution is dramatically affected. Many commonly used network-derived correlates (e.g. degree centrality) can be strongly affected by such changes.

2.3.1.1.5 Expanding the Network. Increasing the .015 subs/site threshold in Billings *et al.* (2019) to 2.33% more than doubles the number of nodes included (Fig 2.4 panel E)

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 Fig. 2.5, panel A). 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 Fig. 2.5, panel B).

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 2.4). 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 $D = 0.839\%$. 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.

2.3.2 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 2.5

2.3.3 Large-scale HIV-1 Database Analyses

2.3.3.1 Markedly Different Thresholds for Different Subtypes. Following the spirit of the analysis performed by Wertheim *et al.* (2014a), 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 2.1. 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 ρ, where ρ 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.

2.3.3.2 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 2.2): 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 α 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.

2.3.4 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 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.* (2018a). 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 Table 2.7 and Table 2.8. The observation of significant homophily among PWID at lower thresholds identified by AUTO-TUNE (0.0076 and 0.0019 subs/site) is consistent with epidemiological characteristics of PWID outbreaks, such as the Drug-Resistant Subtype C Outbreak in Scotland, where rapid transmission led to low genetic diversity within clusters. (Ragonnet-Cronin *et al.*, 2018b)

2.3.5 Comparison With ClustuneR

We benchmarked AUTO-TUNE versus `clustuneR` Chato *et al.* (2020), which employs the recency of sample collection or diagnosis as individual-level weights in a predictive model to 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 8×10^{-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 2.6. 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 intepretation of the filtering steps.

Both methods agree that there is a qualitative relationship of Northern Alberta $<$ Seattle $\sim$ 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 4 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 seconds 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.

2.3.6 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.

Since the (Rhee *et al.*, 2019a) 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 2.1. 50% yielded 0.018581 and 0.01871 mean and median, respectively 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 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 2.2, 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 2.2, 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.

2.4 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 `clustuneR` 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 `clustuneR` 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 (Patil *et al.*, 2022; Leung *et al.*, 2019; Di Giallonardo *et al.*, 2021).

Our evaluation 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.

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.

2.4.1 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 (figure X). 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). 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.

2.5 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.

Tables

Table 2.1: 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 2.3

Subtype	N	$E[D]$	AUTO-TUNE		Fraction	Mean	ρ
			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.* (2014a)	84527	.	1.0	N/A	15.6	3.84	1.74

Table 2.2: Distance thresholds and key network properties using four different HIV-1 genomic regions, stratified by subtype (minumum 500 sequences)

Subtype	N	AUTO-TUNE D, $subs/site$				Number of clusters				Full agreement	Krippendorff α
		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 2.3: 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 ≥ 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

Reference	N	L	$E[D]$	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, USA	1997-2008	96% B	0.02	0.01944
Chato *et al.* (2020)	808	1017	5.6%	Province	Northern Alberta, Canada	2007-2013	B	¶0.0104	0.01201
Little *et al.* (2014a)	648/646 •	1212	5.9%	City	San Diego, CA, USA	1996-2011	98.5% B	0.015	0.02495
Pérez-Losada *et al.* (2017)	1879/3411 •	1027	6.0%	City	Washington DC, USA	1987–2015	B	0.010	0.01733
Rhee *et al.* (2019a)	4553	897	6.1%	State	CA, USA	1998-2016	95.5% B	0.015	0.01139⋆
Chato *et al.* (2020)	2779/2750 •	1398	6.3%	State	Tennessee, USA	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
Chato *et al.* (2020)	1653/1840 •	1020	5.5%	City	Seattle, USA	2000-2013	B	¶0.016	0.01538

Table 2.4: 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. ρ is the fitted characteristic scale-free exponent of the corresponding degree distributions.

Reference	AUTO-TUNE	Nodes in network		Clusters in network		R_{12}		Scale parameter ρ	
	score	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.* (2019a)	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.* (2014a)	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	∞	∞
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

Table 2.5: 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 2.3

Reference	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	$\sim$ 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)

Table 2.6: 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

Table 2.7: 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

Table 2.8: 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	$0.445(-0.129, -0.012)$

Figures

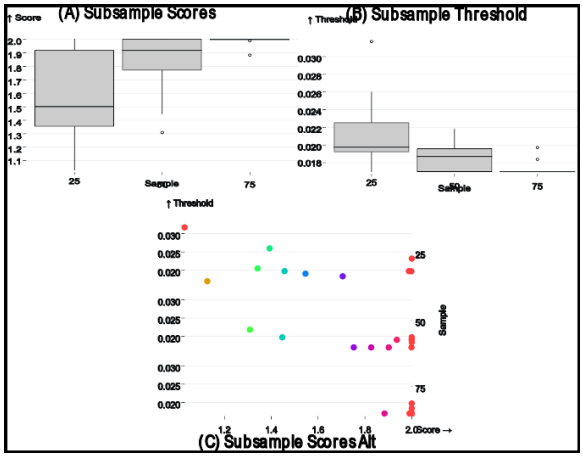

Figure 2.1: (A) Box plot representing the AUTO-TUNE scores across ten random samples at 25%, 50%, and 75% of the (Rhee *et al.*, 2019a) dataset, showing a trend of increasing confidence in score estimates with denser sampling. (B) Box plot of the selected distance thresholds across the same random samples at 25%, 50%, and 75% proportions, demonstrating improved consistency in threshold selection with increased sample size. (C) Scatterplot of the chosen thresholds (Y-axis) against their corresponding AUTO-TUNE scores (X-axis) for the three subsample proportions.

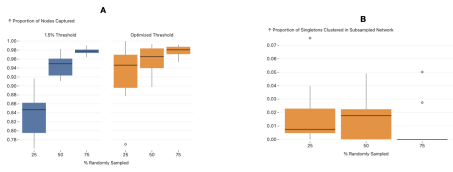

Figure 2.2: Figure A and B present the effects of subsampling on network structure using different thresholds. Figure A illustrates the proportion of nodes subsampled that remained clustered in both the original and the subsampled networks, with an observable increase in nodes captured as the threshold transitions from 1.5% to the optimized AUTO-TUNE threshold across subsampling levels (25%, 50%, and 75%). This pattern underscores the AUTO-TUNE scoring method's efficacy in preserving the network's structural integrity in subsampled datasets. Conversely, Figure B shows the proportion of nodes that were singletons in the original network and clustered in the subsampled networks. Despite varying subsampling rates and iterations, the impact of AUTO-TUNE's adaptive thresholding on clustering these singletons appears minimal, suggesting a negligible influence on the overall network structure.

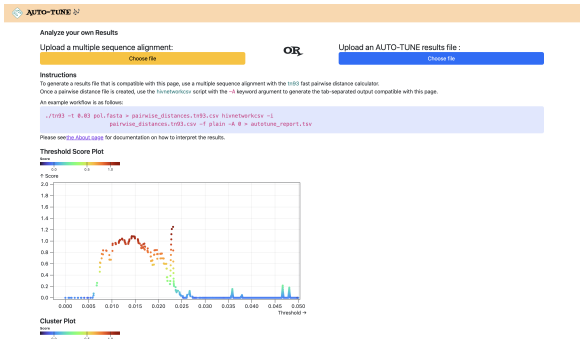

Figure 2.3: The user interface of the AUTO-TUNE web application (`http://autotune.datamonkey.org/analyze`). The platform provides a multi-faceted view of AUTO-TUNE's analysis, including a score plot that visualizes trends across different genetic distance thresholds. It also displays graphs of the number of clusters and the R1/R2 ratio—both key metrics in AUTO-TUNE's heuristic scoring system. These interactive visualizations aid researchers in making nuanced decisions for threshold selection, especially when multiple thresholds yield similar scores.

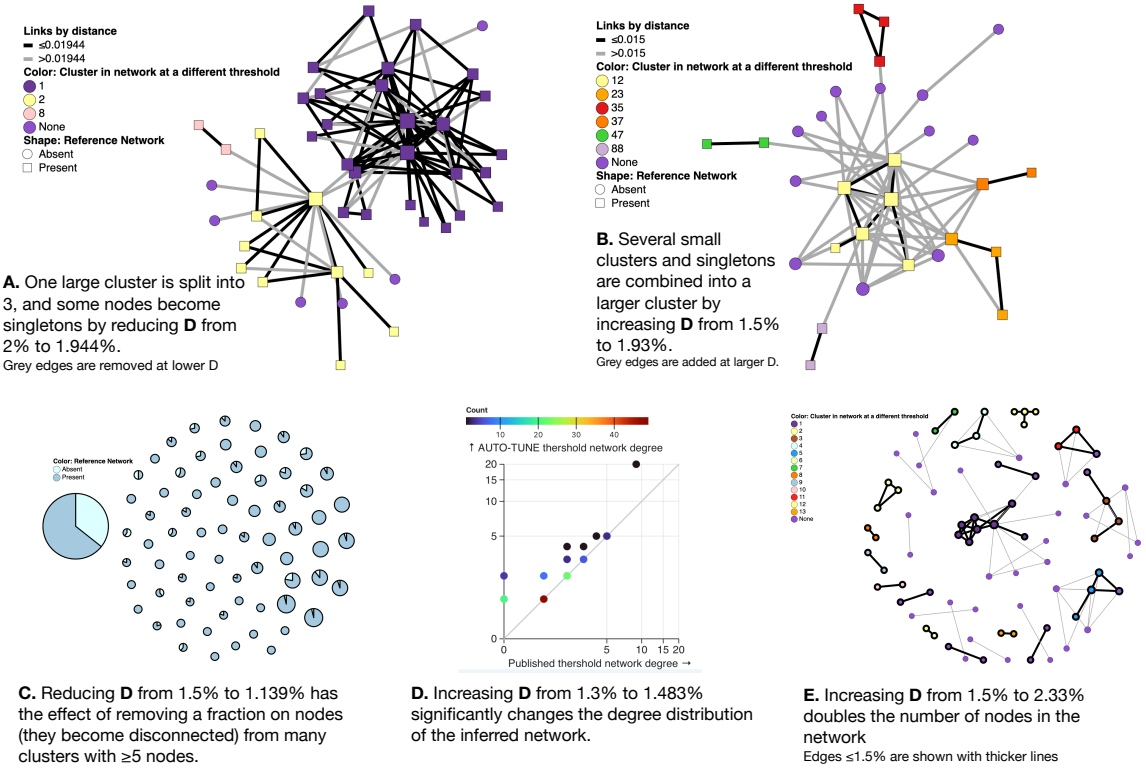

Figure 2.4: **Examples of AUTO-TUNE scores profiles.** (A). Lowering the genetic distance threshold removes some of the edges from the network (shown in grey) and disconnects a large cluster into color-coded smaller clusters; here ”None” means that the node is not connected to anything at the lower threshold. (B). Raising the genetic distance threshold adds edges to the network (shown in grey) and connectes previously separte clusters into a larger component. (C). Each circle is a cluster in the larger threshold network, and with a proportion of nodes removed when the threshold is lowered. (D). Changes to the node degree distribution (colors represent the counts of nodes with the same degree). (E). A significant enlargement of a small network at a higher threshold, with grey edges only present at the larger threshold.

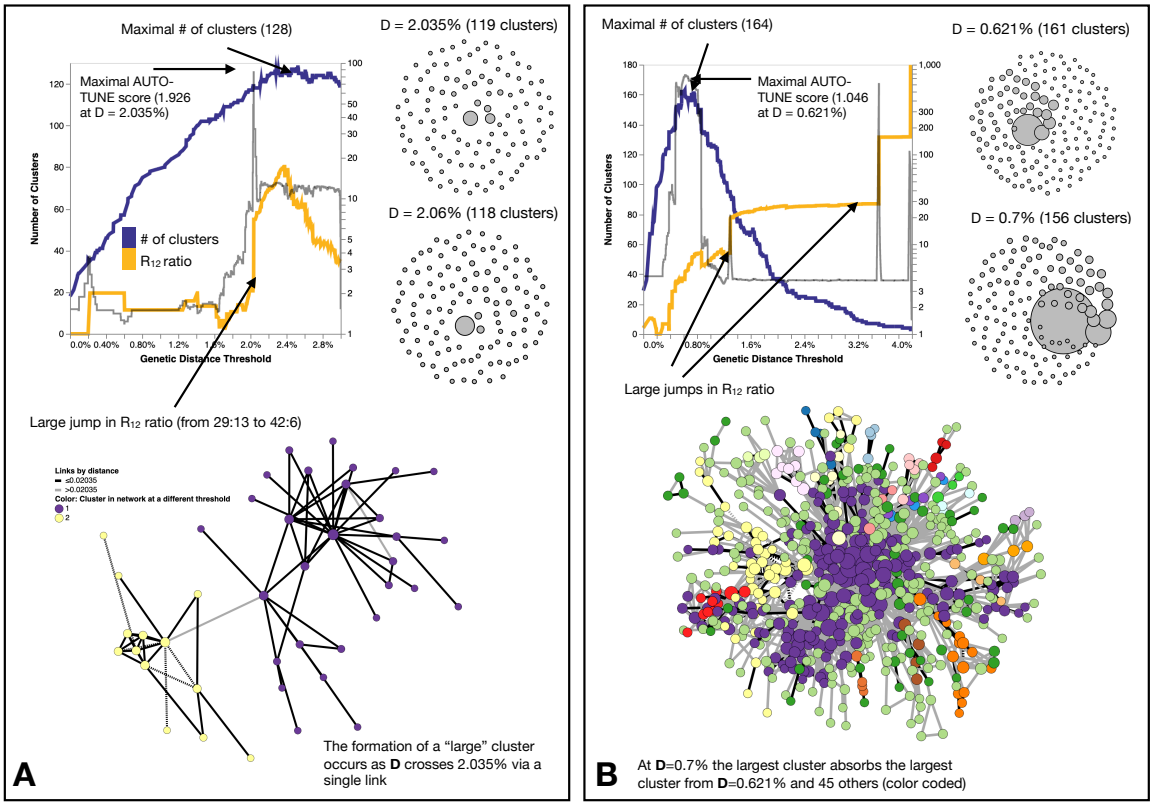

Figure 2.5: **Examples of how changing thresholds affects inferred networks.** (A). A high-scoring network Bbosa *et al.* (2020) has a distance threshold which achieves the number of clusters near the maximum, while also avoiding the formation of a large (weakly connected) cluster. (B). A low-scoring network Liu *et al.* (2020) has a misalignment between the distance for which the maximum number of clusters is found, and where the big jumps in the cluster size ratio occur. Here, AUTO-TUNE effectively optimizes the number of clusters while preventing excessive growth of the largest cluster.

CHAPTER 3
REFINEMENT OF LARGE-SCALE MOLECULAR TRANSMISSION NETWORKS THROUGH SPURIOUS EDGE FILTERING

3.1 Introduction

Genetic sequence data are routinely used to identify clusters of transmission Poon *et al.* (2016); Wertheim *et al.* (2017b); Dennis *et al.* (2021). While the definition of what constitutes a transmission cluster Liu *et al.* (2020); Erly *et al.* (2020); Novitsky *et al.* (2020); Zhao *et al.* (2021), and whether or not clusters are meaningful units of epidemic tracking and outbreak detection Tookes *et al.* (2020); Peters *et al.* (2016); Sizemore *et al.* (2020); Pan *et al.* (2018), the field of molecular epidemiology is forging ahead using this concept for surveillanceOster *et al.* (2018a); Tookes *et al.* (2020); Campbell *et al.* (2020); Tordoff *et al.* (2021), inference of epidemic dynamicsPineda-Peña *et al.* (2014); Ragonnet-Cronin *et al.* (2018c); Rouet *et al.* (2018); Novitsky *et al.* (2021), forecastingRagonnet-Cronin *et al.* (2022a); McLaughlin *et al.* (2019); Billock *et al.* (2019); Moshiri *et al.* (2019), prevention, intervention and evaluation effort targeting Poon *et al.* (2016); Little *et al.* (2014b); Samoff *et al.* (2020), and public health applicationsJenness *et al.* (2020); Wang *et al.* (2015a); Wertheim *et al.* (2016), including the formal inclusion of cluster identification in the US HIV/AIDS policyOster *et al.* (2021b); Hassan *et al.* (2020). The majority of applications are presently concerned with HIV-1, but studies of other pathogens, both viral Campigotto *et al.* (2023); Sahibzada *et al.* (2022) and bacterial Didelot *et al.* (2014); Nguinkal *et al.* (2024) have also seen the uptake of cluster-based analyses.

Two of the most common definitions for a molecular cluster are based on genetic distances, as popularized by `HIV-TRACE` (Kosakovsky Pond *et al.*, 2018a), or phylogenetic tree analysis (Balaban *et al.*, 2019; Ragonnet-Cronin *et al.*, 2013). With both approaches, clusters are defined as sets of sequences that are closely related to each other, and sufficiently distinct from the rest of the sequences in the study. With COVFEFE, sequences are linked if their genetic distance over a diagnostic locus, e.g. HIV-1 *pol* gene, is below a given threshold, d. A cluster is defined as a connected component of the network induced by this linking procedure. Phylogenetic cluster definition typically combines a requirement for sufficient bootstrap (or similar) support for the most recent ancestor branch of a cluster (separation from the rest), with an upper bound on the genetic distance between sequences in the putative cluster (close relatedness), see Hassan *et al.* (2017) for a review.

Neither procedure explicitly accounts for transitive edges which connect nodes that are in the same cluster, but need not be *directly* linked. Standard phylogenetic cluster construction implicitly assumes that every node in the cluster is linked to every other node in the cluster; the structure of the subtree defining the cluster is ignored. `HIV-TRACE` and other distance-based methods will generally not create completely connected clusters, but nonetheless, may infer many transitive edges (Figure 3.6). In the simplest case, a transmission chain $A \rightarrow B \rightarrow C$ may represented as a triangle on A, B, C if A and C are inferred to be directly linked, due to the closely relatedness of their sequences. Abundance of cycles in empirical COVFEFE transmission networks (see the Results section) implies that this error is made often.

Yet, ensuring edges that are less likely to be indicative of a direct transmission are not inferred is fundamental to the practicality of molecular epidemiology. Methods and studies that rely on node-level properties of the network, such as degree or degree distributionsBbosa *et al.* (2021); Reichmuth *et al.* (2021), will be affected by spurious edges. We aim to introduce an algorithm, COVFEFE, that identifies and removes some transitive edges between sequences. By locating simple cycles (triangles and cycles of length 4) in the network and performing a statistical test for support of each edge in each respective cycle, edges can be removed or marked as being potentially spurious. At the method's core is a phylogenetic conditional independence test which can be used to assess whether or not sequences A and B are conditionally independent, given sequence C, or two sequences C, D. Individual tests can be performed very quickly, and because tests over multiple sets of nodes can be run in parallel, the algorithm scales computationally.

3.2 Methods

The algorithm employs a zero branch-length statistical test to evaluate the necessity of edges within network triangles. Utilizing the TN93 substitution matrix(Tamura and Nei, 1993a) and Felsenstein's maximum likelihood approach (Felsenstein, 1981), it assesses whether the nodes in a triangle are conditionally independent. Edges failing the test are marked for removal unless they serve as 'bridges' in the network, ensuring no clusters are broken. The method iterates until no further edges can be removed and is extendable to cycles of arbitrary length.

3.2.1 Zero Branch-Length Statistical Test for Edge Removal

In each network triangle, a zero branch-length statistical test is executed to assess the conditional independence of nodes. The test's foundational premise posits that if a node's branch length exceeds zero within a putative transmission chain, the remaining nodes in the

triangle are not conditionally independent. The test employs the TN93 substitution matrix in conjunction with Felsenstein's maximum likelihood method for parameter optimization. We adopt a null hypothesis stating a branch length of zero and an alternative hypothesis of a non-zero branch length. The TN93 model was chosen for its balance of computational efficiency and robust results, compared to more complex alternatives like GTR.

The steps to perform the zero-length branch test is as follows.

1. Create a 3-taxon unrooted phylogenetic tree for a given triangle.
2. Calculate the maximum likelihood of the tree under the full TN93 model of evolution.
3. Iterate through each of the three sequences and calculate the maximum likelihood with the respective index branch length constrained to zero.
4. Return the hypothesis test's P-value results for each of the three sequences, using a one sided likelihood ratio test (50:50 mixture of χ_1^2 and χ_0^2). Note that in cases when more than one edge could be removed, and the degree of edge support is the same, **none** of the edges are removed in order to avoid creating spurious directionality due to tie-breaking. In the extreme case, a triangle of three identical sequences cannot be simplified to a chain simply because there is no information to do so.

3.2.1.1 Zero Branch Length Test Psuedocode. In Algorithm 1, $CChi2(x, d)$ refers to the cumulative distribution function for the χ^2 distribution with d degrees of freedom.

Data: Sequence Data A, B, C
Result: P-values for each zero branch length test
Model TN93;
Tree $T = (A, B, C)$;
$full \leftarrow$ Full Optimization of T using Model $TN93$;
foreach $s \in \{A, B, C\}$ **do**
  Constrain branch length b of s to 0;
  $constrained \leftarrow$ Constrained s Optimization of T using Model $TN93$;
  $pValues[s] \leftarrow 0.5 * (1 - CChi2(2 * (full - constrained), 1))$;
end
return $pValues$

Algorithm 1: Zero Branch Length Statistical Test

Figure 3.7 illustrates the results of the test. The zero branch-length statistical test is implemented in HyPhy (Kosakovsky Pond *et al.*, 2020) and is available under the filename TriangleSupport.bf in the 'hivclustering' Python package.

3.2.2 Procedure for Edge Pruning in Triangles

1. Compute the network and list all existing triangles. Given that the total number of triangles scales as N^3, where N represents the node count in a cluster, a preset limit (e.g., 2^{18}) is established to restrict the number of triangles examined.

2. For every triangle, identify which chains are statistically unsupportable. Edges corresponding to any non-rejectable chain are flagged for potential removal.

3. Arrange flagged edges in ascending order based on p-value support (ranging from 0.5 to a smaller decimal).

4. Eliminate the greatest possible number of statistically unsupportable edges, ensuring that no 'bridge' edges are removed, thereby maintaining the existing cluster configuration.

5. Repeat the above steps iteratively until no additional edges qualify for removal.

3.2.3 Extension to Cycles of Arbitrary Length N

1. Derive the network and catalog all cycles of length N.

2. For every identified cycle, generate $(2N-5)!!$ feasible unrooted trees.

3. Administer the zero branch-length test on each constructed tree to ascertain which edge, if any, qualifies for removal from the cycle.

4. Loop back to the first step and reiterate the process until no further edges are removable.

The zero branch-length test for edge removal in cycles of arbitrary length N is implemented in the `CycleSupport.bf` file within the `hivclustering` Python package.

3.2.4 Transmission Chain Simulation

This simulation methodology was simulated sequences based on the General Reversible Model for direct transmission chains (Rodriguez *et al.*, 1990).

3.2.4.1 Simulation Parameters.

- Chain Size: Varied from 3 to 20 nodes.

- Replications: Ten replications per size to compute average metrics such as the false positive rate.

- `internal_length`: Set to 0.005, corresponding to a 0.05% sequence divergence between nodes.
- `tip_length`: Set to 0, signifying immediate sampling.

3.2.4.2 Sequence Generation. Sequences were generated using the HyPhy script `SimulateSequence.bf`, requiring an edge list as input. This list delineates:

- Node relationships
- Distances between source and target nodes (`internal_length`)
- Within-host evolutionary dynamics prior to sampling (`tip_length`)

3.2.4.3 Simulation Workflow. The simulation and testing pipeline includes the following steps:

1. Construct transmission chains for the HyPhy simulation script.
2. Execute the HyPhy script to produce a FASTA file containing sequences.
3. Execute HIV-TRACE with an edge-filtering flag on the simulated FASTA file.
4. Run HIV-TRACE without an edge-filtering flag on the simulated FASTA file.
5. Execute HIV-TRACE with a cycle-filtering flag on the simulated FASTA file.
6. Generate an edge report, summarizing the findings.
7. Produce summary statistics in the form of plots and tables.

In Figure 3.1, we present the workflow for simulating a transmission chain of 8 with uniform distance between transmissions and uniform sampling time.

3.3 How to Perform Edge Filtering with HIV-TRACE

Edge filtering is crucial in analyzing HIV transmission networks by removing spurious transitive connections that might misrepresent the true structure of the network. Here's how you can apply edge filtering using the `hivtrace` tool and the `hivnetworkcsv` command.

3.3.1 Requirements

- HIV-TRACE installed and configured on your system.
- A FASTA file with your sequence data.

3.3.2 Using `hivtrace` for Edge Filtering

The `hivtrace` command allows you to run the entire HIV-TRACE pipeline. To apply edge filtering, use the `-f` or `--filter` option with either `remove` or `report` arguments:

- `remove`: Reports spurious transitive connections.
- `report`: Removes spurious transitive connections.

Example Command:

```
hivtrace --do-not-store-intermediate -i PATH_TO_FASTA -a
   ↪resolve -f remove -r REFERENCE_SEQUENCE -t
   ↪TRANSMISSION_THRESHOLD -m MINIMUM_CLUSTER_SIZE -g
   ↪GENETIC_DISTANCE_THRESHOLD -o OUTPUT_JSON --log
   ↪OUTPUT_LOG
```

Replace PATH_TO_FASTA, REFERENCE_SEQUENCE, TRANSMISSION_THRESHOLD, MINIMUM_CLUSTER_SIZE, GENETIC_DISTANCE_THRESHOLD, OUTPUT_JSON, and OUTPUT_LOG with your specific parameters.

3.3.3 Using `hivnetworkcsv` for Edge Filtering

If you already have pairwise distances, use the `hivnetworkcsv` command with the `--edge-filtering` option:

```
hivnetworkcsv -i INPUT_CSV -t TRANSMISSION_THRESHOLD -f
   ↪FORMAT -j -o -n {remove,report} -s SEQUENCE_FILE
```

- `INPUT_CSV`: Your input CSV file with pairwise distances.
- `TRANSMISSION_THRESHOLD`, `FORMAT`, `SEQUENCE_FILE`: Replace with your specific parameters.
- `--edge-filtering {remove,report}`: Choose remove to report spurious connections, report to view flagged edges as dashed lines in the visualization.

After performing edge filtering and generating the results JSON file using `hivtrace` or `hivnetworkcsv`, you can visualize the resulting network by visiting hivtrace.datamonkey.org. To upload and view your results, follow these steps:

1. Navigate to hivtrace.datamonkey.org.

2. Click on **Select File**.

3. Choose **Upload File** and select the results JSON file you generated.

This web-based tool provides a convenient way to visually analyze the HIV transmission network based on your data.

3.3.3.2 Code and Data Availability. The code supporting the analyses presented in this study is available in two repositories. The HIV-TRACE tool is hosted at `https://github.com/veg/hivtrace`, and the HIVClustering tool can be found at `https://github.com/veg/hivclustering`. These repositories provide the source code, documentation for installation, usage guidelines, and example datasets. Additional documentation can be found on the dedicated website, `https://covfefe.datamonkey.org`.

Additionally, the simulation pipeline used for generating synthetic data to test the edge filtering methodologies is located at `https://github.com/veg/edge-filtering`.

Datasets used in this study, along with the example outputs generated by these tools, will be made available in a dedicated repository, the details of which will be provided upon publication.

For further information or inquiries regarding the code or datasets, interested parties should refer to the documentation within the repositories or contact the corresponding author.

3.4 Results

3.4.1 Contextualizing Network Connectivity Through Simulated Transmission Chains

The complexity of HIV transmission network inference can be elucidated through the examination of viral sequence data. Inherent to the nature of such networks is the principle that as the distance between viral sequences diminishes, the number of edges in the inferred network without filtering approaches that of a fully connected network. This phenomenon underpins our simulation efforts.

To exemplify this concept, consider a transmission chain involving three individuals. In the context of HIV-TRACE, when a threshold t is applied to establish a connection between two nodes, and the sum of distances $d_1 + d_2 \leq t$, a triangle is inferred within the network, suggesting concurrent relationships despite that being abiological. This inference is illustrated in where small distances d_1 and d_2 lead to the prediction of a triangle, overestimating the connectivity within the network.

The maximum number of edges $L_{\max}$ for a network comprising N nodes is calculated as $L_{\max} = \frac{N(N-1)}{2}$. Taking a network of size $N = 8$, we deduce that $L_{\max} = 28$, representing a fully connected network. In a sequential transmission chain of 8 nodes, we inherently have $L = 7$ edges. During our simulations, which maintained a consistent rate across all chains ($d_1 = d_2 = d_n = r$), we observed the emergence of spurious edges, the distribution of which we aimed to characterize.

3.4.2 Transmission Chain Simulation

To obtain a fundamental understanding of the underlying structure of HIV transmission networks, we conducted a series of simulations that varied both intra-host (tip length) and between-host (internal node length) evolutionary distances. Specifically, we simulated a chain of length 8 for 20 repetitions at each combination of tip and internal lengths, ranging from 0.001 to 0.014 at intervals of 0.001 (4500 simulations). Networks were then constructed from the default 1.5% distance threshold.

The evolutionary distances were uniformly simulated across all combinations to ensure consistency and comparability in the resulting networks. This uniform simulation approach provided a controlled setting to discern the effects of varying evolutionary distances on the network's topology.

We evaluated the effectiveness of various intra- and inter-host evolutionary distance combinations in our HIV transmission chain simulations. Our objective was to identify configurations that achieve an optimal balance by effectively removing spurious edges while retaining true connections within the network. The performance score for each combination was calculated using a composite metric that considers effectiveness, efficiency, and the edge removal ratio. The formula for the score is as follows:

$$\text{Network Filtering Score} = w_1 \times \text{Effectiveness} + w_2 \times \text{Efficiency} + w_3 \times \text{Edge Removal Ratio}$$

Where the weights are defined as follows:

- $w_1 = \frac{1}{6}$ for Effectiveness
- $w_2 = \frac{1}{2}$ for Efficiency
- $w_3 = \frac{1}{3}$ for Edge Removal Ratio

Each component is described as:

- Effectiveness = $\frac{\text{TN}}{\text{TN+FP}}$ measures the proportion of true negatives among edges removed.
- Efficiency = $\frac{\text{TP}}{\text{TP+FN}}$ measures the proportion of true positives among edges retained.
- Edge Removal Ratio = $\frac{\text{Total Edges}-\text{Edges After Filtering}}{\text{Total Edges}}$ measures the proportion of total edges removed.

3.4.2.1 Edge Removal Metrics. Our analysis showed that performance tends to remain relatively high across various inter-host evolution rates as long as intra-host evolution remains low, highlighting the importance of taking samples as close as possible to the date of infection. Within these higher scoring inter- and intra-host combinations, the number of edges correctly identified and removed by the filter was substantial while also minimizing the number of edges incorrectly filtered, a quality that the composite score rewards most. A heatmap of the scores affiliated with each intra- and inter-host combination can be found in figure 3.2. A table of various intra- and inter-host evolutionary distance combinations in HIV transmission chain simulations can be found in Table 3.4.

3.4.3 Empirical Datasets Analysis

3.4.3.1 Northern Californian Cohort (NCC). In this cohort, HIV-TRACE was utilized to analyze a local epidemic within the context of a global HIV-1 pandemic (Rhee *et al.*, 2019b). Edge filtering was applied to remove spurious transitive connections.

Edge Filtering Results: Without edge filtering, HIV-TRACE inferred 4169 edges. With edge filtering enabled, the count reduced to 2699, a difference of 1470 edges. Additional filtering on both triangles and cycles of size 4 removed an additional 139 edges. (Figure 3.3)

The degree model fit of the NCC dataset demonstrates a clear enhancement in BIC values with successive applications of COVFEFE. The transition from no filtering to C3 and then to C3,4 edge filtering leads to progressively lower BIC scores, indicating a more precise fit of the degree model to the data. This trend implies the efficacy of incorporating COVFEFE to improve network model accuracy and reduce spurious connections. (Table 3.2)

3.4.3.2 Los Alamos National Laboratory. In this study, we utilized HIV-TRACE to analyze the global diversity of HIV-1 within the Los Alamos National Laboratory reference dataset. This dataset, generated on May 9th, 2019, includes 139,060 HIV-1 group M sequences.

Edge Filtering Results: Without edge filtering, HIV-TRACE inferred 1,794,851 edges. With C3 edge filtering enabled, the count reduced to 1,552,317, a difference of 242,534

edges. Additional filtering using C3,4 inspection removed another 77,901 edges from the network. (Figure 3.4)

The Bayesian Information Criterion (BIC) model fit values for the LANL dataset show notable improvement with the application of COVFEFE. Moving from no filtering to C3 and C3,4 edge filtering, the BIC decreases substantially, indicating better model accuracy and the effectiveness of filtering in refining the network. (Table 3.1)

3.4.3.3 Efficacy at Suboptimal Thresholds. We focused on comparing COVFEFE at both suboptimal and AUTO-TUNE optimized thresholds. These comparisons were drawn from several studies using HIV-TRACE. Using these datasets, we sought to quantitatively assess the performance of edge filtering in refining molecular epidemiological networks. Specifically, we investigated how COVFEFE performed under a commonly used, yet potentially suboptimal, threshold of 1.5% and compared these outcomes with those obtained using the AUTO-TUNE optimized thresholds. This allowed us to directly assess the robustness of COVFEFE in scenarios where precise thershold tuning might not be feasible.

To establish a baseline for comparison, we collected the datasets from selected publications, and analyzed them with both HIV-TRACE's default 1.5% threshold and thresholds determined by AUTO-TUNE. This analysis involved using all available sequence data and applying COVFEFE to evaluate its effectiveness in enhancing network inference under each threshold setting.

Our analysis highlighted that COVFEFE substantially improved the fit of degree distributions in the inferred networks, especially when used at the fixed suboptimal threshold of 1.5%. This shows that COVFEFE was potentially effective in reducing spurious edges, as shown in the marked decrease in BIC values 3.5.

After reviewing these results, three key observations were made. 1) At the unoptimized default threshold of 1.5%, the reduction in BIC after COVFEFE indicated a notable improvement in model fit. This suggests that COVFEFE help refine networks when intial conditions are not ideal. The mean BIC difference for for model fit for unoptimized thresholds was 597.5, while the mean BIC difference for optimized thresholds was 372.2. 2) Between triangle filtering C_sub3 and cycle filtering $C_sub3,4$, there is a negligible difference in BIC reductions in both networks with unoptimized and optimized thresholds, 45.0 and 59.5, respectively. 3) While AUTO-TUNE thresholds naturally provided optimal outcomes in terms of model fit, the consistent performance of COVFEFE improving BIC scores underscores its utility in network refinement.

The results affirm that COVFEFE enhances degree distribution model fit when applied to networks with both optimized and suboptimal thresholds. We find that a dual approach of AUTO-TUNE + COVFEFE results in the fitting expected models of degree distribution best. Furthermore, when an optimal threshold is difficult to ascertain, we have shown that COVFEFE still provides utility refining inferred transmission networks.

3.5 Discussion

Although edges within an HIV-TRACE network are not synonymous with actual transmission events, they do likely link individuals within the same epidemiological transmission cluster. (Wertheim *et al.*, 2013) By removing edges within cycles that lack support compared to others within the same cycle, a cluster that more closely resembles the underlying transmission events can be attained. With a sparser network that more closely resembles transmission events, researchers will be able to have greater confidence in using common network analysis metrics such as node centrality, and measures of preferential attachment to characterize their molecular transmission networks.

Currently, outbreak response plans are tailored around specific transmission clusters within a network, and whether a specific cluster is ”recent and rapid”Ragonnet-Cronin *et al.* (2018c). Recent and rapid clusters are defined by the number of cases that have a diagnosis date that fall below a pre-defined threshold. Once a transmission cluster is determined to be recent and rapid, public health officials activate a response plan that include new testing and vaccination sites, facilitating community-based organizations to form within particular populations, and other methods of intervention activities.

If the clusters themselves are more indicative of true transmission events, recent and rapid metrics can be supplemented with additional metrics such as particular populations associated with high average centrality, which could in lead to cross-jurisdictional intervention strategies tailored towards populations that would otherwise go unnoticed within localized transmission cluster metrics.

The balance between mistakenly included transmission links (false positives) and mistakenly filtered genuine transmission links (false negatives) appears to be influenced by the intra-host and inter-host evolutionary rates of the virus. Simulation results indicate that genuine transmission links are most likely to be mistakenly filtered when the rate of viral evolution within a single host (intra-host rate) exceeds the rate of evolution between hosts (inter-host rate).

The preliminary results from edge filtering have been encouraging in removing edges that are likely to not to represent underlying transmissions. With spurious edge reduction, intervention strategies will be able to implement more sophisticated techniques that include node degree, centrality measurements, and epidemic model fitting.

3.5.1 Future Direction

More robust epidemic simulations are necessary, but adapting existing epidemic networks for use molecular transmission network is nontrivial. Current models scale time between infections linearly and include multiple different "seed infectors" in the same tree, but scale them and simulate sequences from independent seeds at the same rate as those that are directly transmitted from one to other. This is at odds with observations that patients directly infected by individuals typically fall within a range below 2%, and patient pairs selected at random fall between 3% and 8.5%. (Wertheim *et al.*, 2017a)

Due to the complexity of the algorithm, $O(N^3)$, a maximum of 2^{18} nodes is set. More work needs to be done to find an appropriate maximum limit, and to implement further parallelization at this step.

Tables

Table 3.1: BIC degree model fit for LANL dataset. The table presents BIC values, shape parameter ρ, and confidence intervals of ρ for various edge filtering methods.

Type	**BIC**	ρ	ρ **CI**
No Filtering	445,963.792	1.390	[1.385, 1.396]
C3 Edge Filtering	419,908.005	1.428	[1.423, 1.434]
C3,4 Edge Filtering	410,162.331	1.443	[1.437, 1.449]

Table 3.2: BIC degree model fit for the NCC dataset showing the effect of COVFEFE.

Type	**BIC**	ρ	ρ **CI**
No Filtering	8520.592	2.659	[2.540, 2.787]
C3 Edge Filtering	6872.921	3.592	[3.391, 3.808]
C3,4 Edge Filtering	6627.654	3.713	[3.500, 3.943]

Table 3.3: Summary of PubMed Data Results

PubMed ID	Type	BIC	ρ	ρ CI
22739398	No Filtering	337.852	2.651	[2.300, 3.075]
	C3 Edge Filtering	329.420	2.682	[2.324, 3.116]
	C3,4 Edge Filtering	329.420	2.682	[2.324, 3.116]
	No Filtering @ 1.5%	474.124	2.445	[2.167, 2.773]
	C3 Edge Filtering @ 1.5%	464.840	2.468	[2.185, 2.802]
	C3,4 Edge Filtering @ 1.5%	464.840	2.468	[2.185, 2.802]
22739398	No Filtering	1206.055	2.285	[2.117, 2.471]
	C3 Edge Filtering	1106.136	2.379	[2.197, 2.582]
	C3,4 Edge Filtering	1106.136	2.379	[2.197, 2.582]
	No Filtering @ 1.5%	5607.138	1.915	[1.832, 2.005]
	C3 Edge Filtering @ 1.5%	5392.148	1.957	[1.869, 2.053]
	C3,4 Edge Filtering @ 1.5%	5384.388	1.959	[1.871, 2.054]
22739398	No Filtering	5851.277	2.588	[2.447, 2.740]
	C3 Edge Filtering	5054.319	2.488	[2.350, 2.639]
	C3,4 Edge Filtering	4921.389	2.432	[2.297, 2.579]
	No Filtering @ 1.5%	20940.347	1.557	[1.524, 1.593]
	C3 Edge Filtering @ 1.5%	20096.121	1.519	[1.487, 1.552]
	C3,4 Edge Filtering @ 1.5%	19988.813	1.512	[1.481, 1.544]
22739398	No Filtering	259.533	6.748	[4.765, 10.141]
	C3 Edge Filtering	236.598	7.450	[5.139, 11.549]
	C3,4 Edge Filtering	236.598	7.450	[5.139, 11.549]
	No Filtering @ 1.5%	220.474	7.018	[4.806, 10.999]
	C3 Edge Filtering @ 1.5%	196.887	3.511	[2.947, 4.223]
	C3,4 Edge Filtering @ 1.5%	196.887	3.511	[2.947, 4.223]
22739398	No Filtering	3551.626	3.164	[2.926, 3.426]
	C3 Edge Filtering	2753.486	3.439	[3.146, 3.768]
	C3,4 Edge Filtering	2522.581	3.722	[3.384, 4.105]
	No Filtering @ 1.5%	13685.604	2.168	[2.086, 2.253]
	C3 Edge Filtering @ 1.5%	11644.560	2.304	[2.212, 2.402]
	C3,4 Edge Filtering @ 1.5%	11637.615	2.269	[2.179, 2.364]
22739398	No Filtering	8278.747	1.628	[1.572, 1.689]
	C3 Edge Filtering	7455.014	1.597	[1.542, 1.655]
	C3,4 Edge Filtering	7429.316	1.583	[1.530, 1.641]
	No Filtering @ 1.5%	10675.717	1.210	[1.177, 1.244]
	C3 Edge Filtering @ 1.5%	9957.170	1.254	[1.220, 1.289]
	C3,4 Edge Filtering @ 1.5%	9948.451	1.254	[1.220, 1.290]

Table 3.4: Performance evaluation of various intra- and inter-host evolutionary distance combinations in HIV transmission chain simulations. Configurations are characterized by a balance between the number of edges removed, the number of spurious edges identified, and the incidence of false negatives. Combinations with too few edges to construct a chain or very low scores were excluded from this table.

Inter-host Rate	Intra-host Rate	Avg. Total Edges	Avg. TP	Avg. TN	Avg. FP	Avg. FN	Score
0.002	0.001	27	10.667	6.804	9.529	11.471	0.709
0.001	0.001	28	13.020	6.098	8.882	13.118	0.690
0.003	0.001	23.588	8.000	6.902	8.686	8.902	0.690
0.004	0.001	19.118	5.157	6.980	6.980	6.137	0.666
0.005	0.001	15.843	3.078	7.000	5.765	4.078	0.634
0.006	0.001	14.039	2.431	6.980	4.627	3.431	0.627
0.009	0.001	8.804	0.667	6.843	1.294	1.647	0.607
0.010	0.001	7.392	0.314	6.588	0.490	1.314	0.606
0.007	0.001	11.529	1.294	6.922	3.314	2.275	0.600
0.003	0.002	21.549	3.569	6.941	11.039	4.510	0.599
0.008	0.001	9.431	0.706	6.843	1.882	1.706	0.593
0.002	0.002	25.804	4.176	6.863	14.765	5.039	0.587
0.001	0.002	27.922	5.608	6.588	15.725	6.196	0.585
0.005	0.002	13.902	1.000	7.000	5.902	2.000	0.566
0.009	0.002	6.902	0.059	6.314	0.529	0.980	0.562
0.011	0.001	6.863	0.098	6.216	0.549	0.961	0.555
0.012	0.001	5.882	0.059	5.686	0.137	0.863	0.553
0.004	0.002	17.059	1.118	6.980	8.961	2.118	0.552
0.006	0.002	11.529	0.451	6.941	4.137	1.451	0.552
0.007	0.002	9.882	0.255	6.882	2.745	1.176	0.550
0.008	0.002	8.235	0.216	6.510	1.510	1.176	0.547
0.002	0.003	23.784	1.765	6.922	15.098	2.686	0.544
0.001	0.003	27.510	2.098	6.902	18.510	3.000	0.542
0.010	0.002	6.490	0.059	5.980	0.451	0.863	0.540
0.003	0.003	19.529	0.784	6.961	11.784	1.765	0.532
0.005	0.003	11.843	0.294	6.824	4.725	1.255	0.531
0.004	0.003	14.353	0.490	6.863	7.000	1.451	0.530
0.002	0.004	21.529	0.529	6.961	14.039	1.510	0.522
0.001	0.004	27	0.961	6.902	19.137	1.863	0.520
0.006	0.003	10.314	0.039	6.706	3.569	1.000	0.517
0.011	0.002	5.667	0.020	5.392	0.255	0.882	0.516
0.003	0.004	15.980	0.098	6.804	9.078	1.078	0.506

Figures

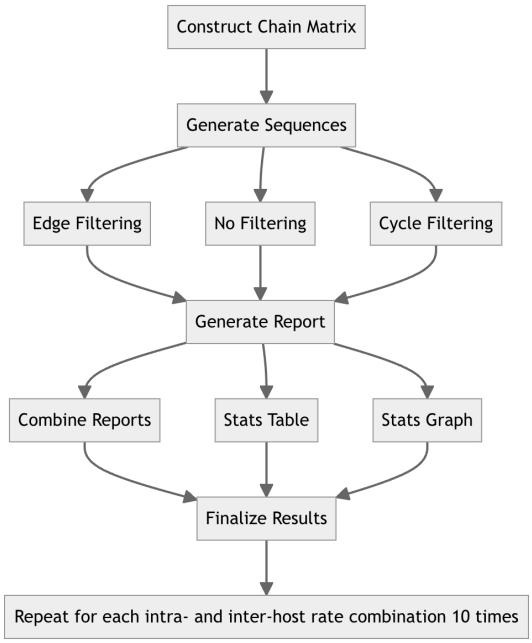

Figure 3.1: Workflow for simulating HIV transmission chains and evaluating different edge filtering methods. The workflow involves constructing transmission chains, generating sequences, and applying HIV-TRACE with various filtering options (edge, no filter, cycle). Reports are generated and combined, and summary statistics are produced. This process is repeated for each combination of intra- and inter-host evolution rates, with ten replications for each combination to compute average metrics.

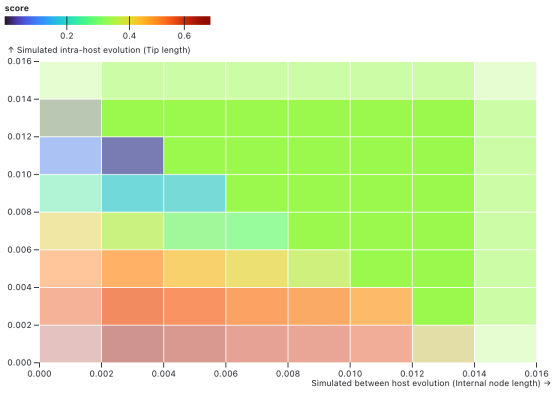

Figure 3.2: A heatmap of the network filtering scores affiliated with each intra- and inter-host combination

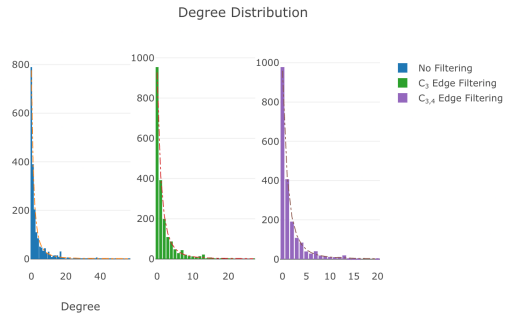

Figure 3.3: NCC degree distributions of nodes with no filtering, edge filtering, and edge filtering with cycles. Using BIC to determine model fit, the fit progressively improved with edge filtering using cycles.

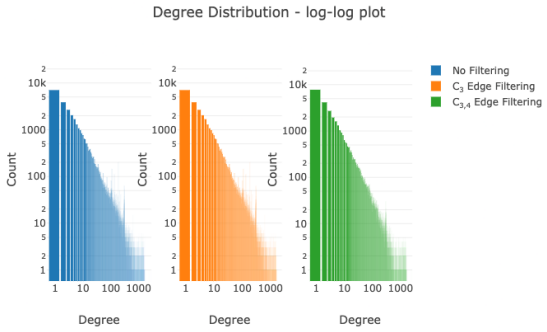

Figure 3.4: LANL degree distributions of nodes with no filtering, edge filtering, and edge filtering with cycles. Using BIC to determine model fit, the fit progressively improved with edge filtering using cycles.

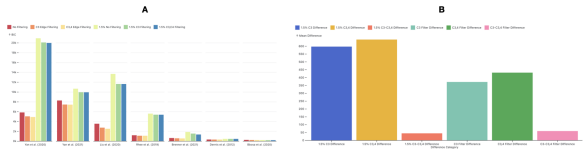

Figure 3.5: Comparative Analysis of Bayesian Information Criterion (BIC) Across Different Filtering Strategies and Their Impacts on Mean Differences. **Panel A:** This bar chart visualizes the BIC values for various filtering approaches applied to genomic data from several studies. Categories include No Filtering, C3 Edge Filtering, C3,4 Edge Filtering, and their default counterparts at a 1.5% threshold. Each bar represents the BIC value associated with a specific study and filtering type, illustrating the comparative effectiveness of these strategies in degree distribution model fitting. **Panel B:** This bar chart displays the mean differences in BIC values among various filtering categories.

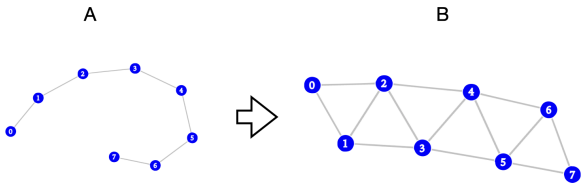

Figure 3.6: Impact of Mutation Rate on Transmission Chain Reconstruction. **Panel A**: Original reconstructed transmission chain assuming a constant mutation rate of 0.015. The chain is reconstructed by the the threshold criterion of 0.015 for linking sequences, as typically set in HIV-TRACE using the -t or –threshold flag. **Panel B**: Network configuration with an increased mutation rate set to 0.006, resulting in a denser network with the formation of multiple transitive edges or triangles. This illustrates the scenario where mutation rates are considerably below threshold values, leading to an overestimation of direct connections and potential misrepresentation of the transmission chain. The total number of links is 13 out of a maximum of 28 possible, when there should only be 7.

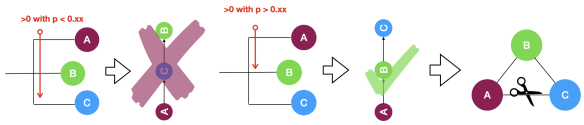

Figure 3.7: Illustration of Zero Branch-Length Statistical Test for Edge Removal in Phylogenetic Triangle Analysis. This figure depicts the sequential steps and outcomes of the zero branch-length statistical test applied to a triangle comprising nodes A, B, and C. The process begins by hypothesizing potential root positions and assessing branch lengths: (1) **Initial Hypothesis Testing:** The test initially hypothesizes C as the root, examining the significance of branch lengths connecting to nodes A and B. (2) **Rejection of C as Root:** Due to a significant branch length at C, it is rejected as the root, thus the edge between A and B cannot be removed. (3) **Identification of B as Root:** Subsequent testing with B as the hypothesized root reveals an insignificant branch length, suggesting B could act as the root node. (4) **Edge Removal:** Based on the statistical insignificance of branch lengths from B to the root, the edge between A and C is removed thanks to conditional independence, simplifying the triangle to a direct chain of A to B to C.

CHAPTER 4
HIV-TRACE GO: HIV TRANSMISSION CLUSTER ENGINE FOR GLOBAL OUTREACH

4.1 Introduction

Molecular transmission networks help in characterizing the dynamics of emerging infectious disease epidemics, including cross-national transmission, drug-resistance dynamics, and the prediction of transmission risk factors. These networks offer insight into cluster growth dynamics and the identification of cryptic transmission risk factors. Recent studies have highlighted the dual nature of antiretroviral drug resistance mutations, showing that while some decrease transmission fitness, others may establish self-sustaining drug resistance reservoirs, challenging global treatment effectiveness (Wertheim *et al.*, 2017c). Furthermore, comprehensive analyses of global HIV-1 polymerase sequences have elucidated the contemporary global transmission network of HIV-1, revealing extensive international linkage and underscoring the global diversity of HIV even in local epidemics (Wertheim *et al.*, 2014b). In the surveillance context, analyses of molecular transmission clusters have proven instrumental in predicting future cases and optimizing public health interventions in the United States, indicating the potential of molecular epidemiology to enhance prevention efforts (Wertheim *et al.*, 2018a). Furthermore, molecular epidemiology offers unique insights into the transmission dynamics among high-risk groups, such as transgender women, by identifying patterns of linkage within transmission networks that can inform targeted public health interventions, underscoring the critical role of molecular epidemiology in addressing the HIV epidemic among the most vulnerable populations (Ragonnet-Cronin *et al.*, 2019). A study in rural KwaZulu-Natal demonstrates how local and cross-community transmission contribute to sustaining high HIV incidences, emphasizing the need for incorporating regional transmission dynamics into prevention strategy designs. (Rasmussen *et al.*, 2018)

HIV-TRACE(Kosakovsky Pond *et al.*, 2018a) is a publicly available bioinformatics tool for reconstructing HIV transmission networks. In 2017, a public-health facing version of this tool was deployed in the United States, Secure HIV-TRACE, which is currently being used by over 160 professional analysts across over 50 jurisdictions conducting molecular HIV surveillance.

Over $700,000$ HIV sequences have been used to infer transmission networks as of mid-2023, informing demographic and transmission risk characteristics in lower morbidity states that have previously been understudied (Grande *et al.*, 2019), identify transmission

clusters based on risk factors (Tookes *et al.*, 2019), and inform prevention measures set forth by the U.S. Centers for Disease Control and Prevention (Oster *et al.*, 2021a).

The Centre for Research in Infectious Diseases (CIENI) of the National Institute of Respiratory Diseases (INER) in Mexico City is a WHO-accredited reference laboratory for HIV-1 sequencing that continuously generates genetic sequence data for Mexico and several Latin American countries.

Given the richness of its HIV-1 sequence database and associated socio-demographic metadata and its position as a reference HIV sequencing laboratory serving Mexico, CIENI-INER has already shown that software aiding HIV-1 molecular epidemiology is a valuable research tool with data collected outside the US well. (Dávila-Conn *et al.*, 2021)

The possibility to perform near real-time molecular surveillance in Mexico served as motivation to develop HIV-TRACE GO to serve both Mexico and the broader global community.

4.2 Design and Implementation

4.2.1 General Infrastructure

While HIV-TRACE GO was able to leverage ongoing development of HIV-TRACE and Secure HIV-TRACE (Secure HIV-TRACE; developed by Joel Wertheim, Steven Weaver, and Sergei Kosakovsky Pond in collaboration with the U.S. Centers for Disease Control and Prevention) as a basis for HIV-1 molecular epidemiology for the broader global community, and a number of features and customizations were made to meet regional and generalizable needs.

HIV-TRACE GOuses Node.js as the runtime engine for server-side scripting, and various libraries, such as d3.js and socket.io, to provide a rich client-side user experience. The result is a web application that is capable of running multiple, long running computational processes in a high-performance computing environment while maintaining a fluid user experience to all who use the site time regardless of how many computationally intensive jobs are running. The application supports running jobs on the same machine as the Node.js web application (primarily for development purposes), the Torque job scheduler, or the Slurm job scheduler. The application can also be deployed on commercial clouds, such as Amazon Web Services.

Each job run is given a persistent identifier, and researchers can select whether to append new data to their existing inferred HIV network or conduct a stand-alone analysis. Once the

job run has been completed, the researcher that initiated the analysis can be emailed with a notification of completion and a link to the results if the server is configured to do so.

User management and authentication are implemented using the `Passport` middleware. `Passport` allows for each request to be inspected and passed through an unlimited series of application-defined authentication methods. Various pre-configured user roles, such as a read-only user, and adminstrator, are available. Application data, such as users, roles, and metadata associated to each analysis are stored in `MongoDB`, a NoSQL document-oriented database. `MongoDB` stores data in JSON-like documents, which allows for tighter data integration within the JavaScript ecosystem.

Most results are generated in-browser using the `hivtrace-viz` JavaScript package, available at `https://github.com/veg/hivtrace-viz`. The stand-alone nature of the visualization package broadens its applicability across different use scenarios with `HIV-TRACE`. This flexibility enables local operation of the software, through the open-source implementation of Datamonkey, or via `HIV-TRACE GO` while maintaining consistent visualization coding. Different packages can progress developmentally at varying speeds due to the versioning of the visualization, as long as the `hivtrace-viz` dependency versions are correctly specified. The `hivtrace-viz` package leverages `d3.js` for data visualization tasks such as charting and network mapping, while `bootstrap` is used for styling and incorporating standard user interface elements, including buttons and menus.

The following is a description of the customizations that were necessary to make `HIV-TRACE GO` a viable solution for epidemiological molecular surveillance in Mexico and abroad.

4.2.2 Translation

First, `HIV-TRACE GO` and associated package dependencies, such as the JavaScript visualization `hivtrace-viz`, underwent an engineering effort known as internationalization. Internationalization was implemented using the JavaScript framework `i18next`, which provides a convenient middleware for both server-side and client-side templating libraries. Collaborating with native Spanish speakers, over ninety key phrases were identified and translated from English to Spanish. The phrases are stored in plain-text JSON and are treated as a separate package () that is a dependency of both `HIV-TRACE GO` and `hivtrace-viz`. Internationalization in this way has positioned HIV-TRACE GO for additional languages to be added to the platform by those without a software development background by simply adding an additional JSON file with translations to this repository.

4.2.3 Multi-jurisdiction Provisioning

The initial design of the Secure HIV-TRACE platform was predicated on the principle that each user would be uniquely associated with a single geographic jurisdiction, effectively restricting them from administering, uploading, or accessing data from any other site. This design was congruent with the envisaged use cases at the time across U.S public health departments.

However, CIENI-INER, serving as a key reference laboratory for HIV sequences not only for Mexico but also for neighboring countries, necessitates a more intricate data management model. The technicians at CIENI-INER oversee multiple jurisdictions, each corresponding to a Mexican state.

In response to this necessity, the system now accommodates the provisioning of multiple jurisdictions. This enhancement allows administrators to structure their data into distinct networks as required, concurrently permitting users to oversee multiple sites, thereby reflecting the complex, multi-jurisdictional nature of CIENI-INER's operations.

4.2.4 Attribute Management

Socio-demographic metadata is critical to implementing targeted prevention strategies once recent and rapidly growing clusters have been identified (Holmes *et al.*, 1995). Flexibility is required within the application in order to generalize support for institutions, like CIENI-INER, that develop and administer questionnaires that capture socio-demographic and behavioral metadata designed specifically for Mexico. HIV-TRACE GO provides administrators the autonomy to delineate, alter, and eliminate metadata attributes, circumventing the necessity for a software developer to modify the codebase and release an updated version of the website. Concurrently with the introduction of the attribute, researchers have the ability to specify the type, whether it be a "String", "Number", or "Date". This type of declaration serves dual purposes: it enables validation of the dataset upon upload and enables diverse methods for visualizing metadata within network representation (e.g., continuous, temporal, discrete). CIENI-INER collects over 40 attributes that can be annotated along with each HIV-1 sequence, ranging from various transmission risk factors to previous and current antiretroviral drug exposure.

4.3 Concluding Remarks

Implementation and customization for routine use of a powerful and user-friendly tool such as HIV-TRACE in the context of a regional reference laboratory for HIV-1 molecular

surveillance will maximize the benefits obtained from HIV-1 sequence data, providing useful information on local, national, and multi- national transmission patterns, cluster growth dynamics, drug resistance transmission dynamics, transmission risk factor characterization.

With the Condesa Clinic currently diagnosing nearly 20% of all new HIV infections registered in Mexico annually in addition to CIENI-INER's collecting sociodemographic and behavioral metadata of all recently diagnosed persons at the Condesa Clinic who consent to participate in the project, the possibility of near real-time molecular surveillance in Mexico City to inform public health policy is now stronger with the development of HIV-TRACE GO. The project has been reviewed and approved by the INER's Ethics and Research Committees (Project numbers: E02-17 and E02-20). All recently diagnosed persons at the Condesa clinic are invited to voluntarily participate in the study donating a single blood sample for HIV sequencing and answering a computer-based self-administered questionnaire to collect sociodemographic and behavioral variables.

Furthermore, localization engineering efforts have positioned HIV-TRACE for greater expansion into additional countries. For future translation efforts, only a single JSON file needs to be provided with the new language's translations to be immediately reflected on the site.

4.4 Acknowledgments

This work was supported in part by grants R01 GM093939 (NIH/NIGMS), U01 GM110749 (NIH/NIGMS), and funds from the San Diego Center for AIDS Research International Pilot Grant (P30 AI036214; Sub-award No. 112605914). HIV sequencing and matadata collection were supported by CONAHCyT (Project #303079), the Canadian Institutes of Health Research (grants PJT-148621 and PJT-159625), and the Mexican Government (Programa Presupuestal P016; Anexo 13 del Decreto del Presupuesto de Egresos de la Federación).

BIBLIOGRAPHY

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(12): e0260604.

Balaban, M., Moshiri, N., Mai, U., Jia, X., and Mirarab, S. 2019. TreeCluster: Clustering biological sequences using phylogenetic trees. *PLOS ONE*, 14(8): e0221068.

Bartlett, S. R., Wertheim, J. O., Bull, R. A., Matthews, G. V., Lamoury, F. M., Scheffler, K., Hellard, M., Maher, L., Dore, G. J., Lloyd, A. R., *et al.* 2017. A molecular transmission network of recent hepatitis c infection in people with and without hiv: Implications for targeted treatment strategies. *Journal of viral hepatitis*, 24(5): 404–411.

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 research and human retroviruses*, 36(11): 948–951.

Bbosa, N., Ssemwanga, D., Nsubuga, R. N., Kiwanuka, N., Bagaya, B. S., Kitayimbwa, J. M., Ssekagiri, A., Yebra, G., Kaleebu, P., and Leigh-Brown, A. 2021. Phylogenetic Networks and Parameters Inferred from HIV Nucleotide Sequences of High-Risk and General Population Groups in Uganda: Implications for Epidemic Control. *Viruses*, 13(6): 970.

Billings, E., Kijak, G. H., Sanders-Buell, E., Ndembi, N., O'Sullivan, A. M., Adebajo, S., Kokogho, A., Milazzo, M., Lombardi, K., Baral, S., Nowak, R., Ramadhani, H., Gramzinski, R., Robb, M. L., Michael, N. L., Charurat, M. E., Ake, J., Crowell, T. A., Tovanabutra, S., and MHRP Viral Sequencing Core and the TRUST/RV368 Study Group 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(5): 578–584.

Billock, R. M., Powers, K. A., Pasquale, D. K., Samoff, E., Mobley, V. L., Miller, W. C., Eron, J. J., and Dennis, A. M. 2019. Prediction of HIV Transmission Cluster Growth with Statewide Surveillance Data. *Journal of acquired immune deficiency syndromes (1999)*, 80(2): 152–159.

Brenner, B. G., Ibanescu, R.-I., Osman, N., Cuadra-Foy, E., Oliveira, M., Chaillon, A., Stephens, D., Hardy, I., Routy, J.-P., Thomas, R., Baril, J.-G., Leblanc, R., Tremblay, C., Roger, M., and The Montreal Primary Hiv Infection Phi Cohort Study Group, n. 2021. The Role of Phylogenetics in Unravelling Patterns of HIV Transmission towards Epidemic Control: The Quebec Experience (2002-2020). *Viruses*, 13(8): 1643.

Campbell, E. M., Jia, H., Shankar, A., Hanson, D., Luo, W., Masciotra, S., Owen, S. M., Oster, A. M., Galang, R. R., Spiller, M. W., Blosser, S. J., Chapman, E., Roseberry, J. C., Gentry, J., Pontones, P., Duwve, J., Peyrani, P., Kagan, R. M., Whitcomb, J. M., Peters, P. J., Heneine, W., Brooks, J. T., and Switzer, W. M. 2017. Detailed Transmission Network Analysis of a Large Opiate-Driven Outbreak of HIV Infection in the United States. *The Journal of Infectious Diseases*, 216(9): 1053–1062.

Campbell, E. M., Patala, A., Shankar, A., Li, J.-F., Johnson, J. A., Westheimer, E., Gay, C. L., Cohen, S. E., Switzer, W. M., and Peters, P. J. 2020. Phylodynamic Analysis Complements Partner Services by Identifying Acute and Unreported HIV Transmission. *Viruses*, 12(2): 145.

Campigotto, A., Chris, A., Orkin, J., Lau, L., Marshall, C., Bitnun, A., Buchan, S. A., MacDonald, L., Thampi, N., McCready, J., Juni, P., Parekh, R. S., and Science, M. 2023. Utility of SARS-CoV-2 Genomic Sequencing for Understanding Transmission and School Outbreaks. *The Pediatric Infectious Disease Journal*, 42(4): 324–331.

Chan, C. P., Uemura, H., Kwan, T. H., Wong, N. S., Oka, S., Chan, D. P. C., and Lee, S. S. 2020. Review on the molecular epidemiology of sexually acquired hepatitis c virus infection in the asia-pacific region. *J Int AIDS Soc*, 23(9): e25618.

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 Evolution*, 6(1): veaa011.

Chen, Y., Lan, G., Feng, Y., Ruan, Y., Shen, Z., McNeil, E. B., Tang, K., Huang, J., Shao, Y., Lin, M., and Chongsuvivatwong, V. 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(3): e0283031.

Dalai, S. C., Junqueira, D. M., Wilkinson, E., Mehra, R., Kosakovsky Pond, S. L., Levy, V., Israelski, D., de Oliveira, T., and Katzenstein, D. 2018. Combining Phylogenetic and Network Approaches to Identify HIV-1 Transmission Links in San Mateo County, California. *Frontiers in Microbiology*, 9: 2799.

Dávila-Conn, V., García-Morales, C., Matías-Florentino, M., López-Ortiz, E., Paz-Juárez, H. E., Beristain-Barreda, Á., Cárdenas-Sandoval, M., Tapia-Trejo, D., López-Sánchez, D. M., Becerril-Rodríguez, M., García-Esparza, P., Macías-González, I., Iracheta-Hernández, P., Weaver, S., Wertheim, J. O., Reyes-Terán, G., González-Rodríguez, A., and Ávila-Ríos, S. 2021. Characteristics and growth of the genetic HIV transmission network of Mexico City during 2020. *Journal of the International AIDS Society*, 24(11): e25836. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jia2.25836.

Dennis, A. M., Frost, S. D. W., Enders, K., Cressman, A. E., Volz, E., Adams, N., Miller, W. C., Cohen, M. S., Mobley, V., Samoff, E., and Eron, J. J. 2021. HIV-1 Transmission linkages among persons with incident infection to inform public health surveillance. *eClinicalMedicine*, 37.

Di Giallonardo, F., Pinto, A. N., Keen, P., Shaik, A., Carrera, A., Salem, H., Selvey, C., Nigro, S. J., Fraser, N., Price, K., Holden, J., Lee, F. J., Dwyer, D. E., Bavinton, B. R., Geoghegan, J. L., Grulich, A. E., Kelleher, A. D., and NSW HIV Prevention Partnership Project 2021. Subtype-specific differences in transmission cluster dynamics of HIV-1 B and CRF01_ae in New South Wales, Australia. *Journal of the International AIDS Society*, 24(1): e25655.

Didelot, X., Gardy, J., and Colijn, C. 2014. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Molecular Biology and Evolution*, 31(7): 1869–1879.

Ding, X., Chaillon, A., Pan, X., Zhang, J., Zhong, P., He, L., Chen, W., Fan, Q., Jiang, J., Luo, M., Xia, Y., Guo, Z., and Smith, D. M. 2022. Characterizing genetic transmission networks among newly diagnosed HIV-1 infected individuals in eastern China: 2012–2016. *PLOS ONE*, 17(6): e0269973.

Duault, H., Durand, B., and Canini, L. 2022. Methods Combining Genomic and Epidemiological Data in the Reconstruction of Transmission Trees: A Systematic Review. *Pathogens (Basel, Switzerland)*, 11(2): 252.

Erly, S. J., Herbeck, J. T., Kerani, R. P., and Reuer, J. R. 2020. Characterization of Molecular Cluster Detection and Evaluation of Cluster Investigation Criteria Using Machine Learning Methods and Statewide Surveillance Data in Washington State. *Viruses*, 12(2): 142.

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. *Journal of Acquired Immune Deficiency Syndromes (1999)*, 87(3): 912–917.

Fabeni, L., Santoro, M. M., Lorenzini, P., Rusconi, S., Gianotti, N., Costantini, A., Sarmati, L., Antinori, A., Ceccherini-Silberstein, F., d'Arminio Monforte, A., Saracino, A., Girardi, E., and Cohort, O. B. O. T. I. F. S. 2020. Evaluation of hiv transmission clusters among natives and foreigners living in italy. *Viruses*, 12(8).

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6): 368–376.

Foster, G. M., Ambrose, J. C., Hué, S., Delpech, V. C., Fearnhill, E., Abecasis, A. B., Leigh Brown, A. J., Geretti, A. M., and UK HIV Drug Resistance Database 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(1): e83337.

Friedman, N., Geiger, D., and Goldszmidt, M. 1997. Bayesian network classifiers. *Machine learning*, 29(2-3): 131–163.

Fujimoto, K., Bahl, J., Wertheim, J. O., Del Vecchio, N., Hicks, J. T., Damodaran, L., Hallmark, C. J., Lavingia, R., Mora, R., Carr, M., Yang, B., Schneider, J. A., Hwang, L.-Y., and McNeese, M. 2021. Methodological synthesis of Bayesian phylodynamics, HIV-TRACE, and GEE: HIV-1 transmission epidemiology in a racially/ethnically diverse Southern U.S. context. *Scientific Reports*, 11(1): 3325.

Ge, Z., Feng, Y., Zhang, H., Rashid, A., Zaongo, S. D., Li, K., Yu, Y., Lv, B., Sun, J., Liang, Y., Xing, H., Sönnerborg, A., Ma, P., and Shao, Y. 2021. HIV-1 CRF07_BC transmission dynamics in China: two decades of national molecular surveillance. *Emerg Microbes Infect*, 10(1): 1919–1930.

German, D., Grabowski, M. K., and Beyrer, C. 2017. Enhanced use of phylogenetic data to inform public health approaches to HIV among men who have sex with men. *Sexual Health*, 14(1): 89–96.

Gibson, K. M., Steiner, M. C., Kassaye, S., Maldarelli, F., Grossman, Z., Pérez-Losada, M., and Crandall, K. A. 2019. A 28-year history of hiv-1 drug resistance and transmission in washington, dc. *Front Microbiol*, 10: 369.

Golub, B. and Jackson, M. 2012a. Network structure and the speed of learning measuring homophily based on its consequences. *Annals of Economics and Statistics*, 2012(107-108): 33–48.

Golub, B. and Jackson, M. O. 2012b. How Homophily Affects the Speed of Learning and Best-Response Dynamics. *The Quarterly Journal of Economics*, 127(3): 1287–1338.

Goodreau, S. M. 2006. Assessing the Effects of Human Mixing Patterns on Human Immunodeficiency Virus-1 Interhost Phylogenetics Through Social Network Simulation. *Genetics*, 172(4): 2033–2045.

Gore, D. J., Schueler, K., Ramani, S., Uvin, A., Phillips, G., McNulty, M., Fujimoto, K., and Schneider, J. 2022. HIV Response Interventions that Integrate HIV Molecular Cluster and Social Network Analysis: A Systematic Review. *AIDS and behavior*, 26(6): 1750–1792.

Grabowski, M. K., Lessler, J., Redd, A. D., Kagaayi, J., Laeyendecker, O., Ndyanabo, A., Nelson, M. I., Cummings, D. A. T., Bwanika, J. B., Mueller, A. C., Reynolds, S. J., Munshaw, S., Ray, S. C., Lutalo, T., Manucci, J., Tobian, A. A. R., Chang, L. W., Beyrer, C., Jennings, J. M., Nalugoda, F., Serwadda, D., Wawer, M. J., Quinn, T. C., Gray, R. H., and Rakai Health Sciences Program 2014. The role of viral introductions in sustaining community-based hiv epidemics in rural uganda: evidence from spatial clustering, phylogenetics, and egocentric transmission models. *PLoS Med*, 11(3): e1001610.

Grande, K. M., Schumann, C. L., Ocfemia, M. C. B., Vergeront, J. M., Wertheim, J. O., and Oster, A. M. 2019. Transmission Patterns in a Low HIV-Morbidity State—Wisconsin, 2014–2017. *Morbidity and Mortality Weekly Report*, 68(6): 149. Publisher: Centers for Disease Control and Prevention.

H, Y., H, W., Y, X., L, H., Y, L., Q, L., L, C., Z, H., and S, T. 2021. Acquisition and transmission of HIV-1 among migrants and Chinese in Guangzhou, China from 2008 to 2012: Phylogenetic analysis of surveillance data. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 92.

Hassan, A., De Gruttola, V., Hu, Y. W., Sheng, Z., Poortinga, K., and Wertheim, J. O. 2020. The Relationship Between the Human Immunodeficiency Virus-1 Transmission Network and the HIV Care Continuum in Los Angeles County. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 71(9): e384–e391.

Hassan, A. S., Pybus, O. G., Sanders, E. J., Albert, J., and Esbjörnsson, J. 2017. Defining HIV-1 transmission clusters based on sequence data. *AIDS (London, England)*, 31(9): 1211–1222.

Hayes, A. F. and Krippendorff, K. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1): 77–89.

Hevey, D. 2018. Network analysis: a brief overview and tutorial. *Health Psychology and Behavioral Medicine*, 6(1): 301–328.

Holmes, E. C., Zhang, L. Q., Robertson, P., Cleland, A., Harvey, E., Simmonds, P., and Leigh Brown, A. J. 1995. The molecular epidemiology of human immunodeficiency virus type 1 in Edinburgh. *The Journal of Infectious Diseases*, 171(1): 45–53.

Inzaule, S. C., Siedner, M. J., Little, S. J., Avila-Rios, S., Ayitewala, A., Bosch, R. J., Calvez, V., Ceccherini-Silberstein, F., Charpentier, C., Descamps, D., Eshleman, S. H., Fokam, J., Frenkel, L. M., Gupta, R. K., Ioannidis, J. P. A., Kaleebu, P., Kantor, R., Kassaye, S. G., Kosakovsky Pond, S. L., Kouamou, V., Kouyos, R. D., Kuritzkes, D. R., Lessells, R., Marcelin, A.-G., Mbuagbaw, L., Minalga, B., Ndembi, N., Neher, R. A., Paredes, R.,

Pillay, D., Raizes, E. G., Rhee, S.-Y., Richman, D. D., Ruxrungtham, K., Sabeti, P. C., Schapiro, J. M., Sirivichayakul, S., Steegen, K., Sugiura, W., van Zyl, G. U., Vandamme, A.-M., Wensing, A. M. J., Wertheim, J. O., Gunthard, H. F., Jordan, M. R., and Shafer, R. W. 2023. Recommendations on data sharing in hiv drug resistance research. *PLoS Med*, 20(9): e1004293.

Jenness, S. M., Johnson, J. A., Hoover, K. W., Smith, D. K., and Delaney, K. P. 2020. Modeling an integrated HIV prevention and care continuum to achieve the Ending the HIV Epidemic goals. *AIDS (London, England)*, 34(14): 2103–2113.

Jia, Y., Zou, X., Yue, W., Liu, J., Yue, M., Liu, Y., Liu, L., Huang, P., Feng, Y., and Xia, X. 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.

Jombart, T., Eggo, R. M., Dodd, P. J., and Balloux, F. 2011. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*, 106(2): 383–390.

Junqueira, D. M., Sibisi, Z., Wilkinson, E., and de Oliveira, T. 2019. Factors influencing HIV-1 phylogenetic clustering. *Current opinion in HIV and AIDS*, 14(3): 161–172.

Kosakovsky Pond, S. L., Posada, D., Stawiski, E., Chappey, C., Poon, A. F., Hughes, G., Fearnhill, E., Gravenor, M. B., Leigh Brown, A. J., and Frost, S. D. 2009. An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in hiv-1. *PLoS computational biology*, 5(11): e1000581.

Kosakovsky Pond, S. L., Weaver, S., Leigh Brown, A. J., and Wertheim, J. O. 2018a. HIV-TRACE (TRAnsmission Cluster Engine): a tool for large scale molecular epidemiology of HIV-1 and other rapidly evolving pathogens. *Molecular biology and evolution*, 35(7): 1812–1819.

Kosakovsky Pond, S. L., Weaver, S., Leigh Brown, A. J., and Wertheim, J. O. 2018b. HIV-TRACE (TRAnsmission Cluster Engine): a Tool for Large Scale Molecular Epidemiology of HIV-1 and Other Rapidly Evolving Pathogens. *Molecular Biology and Evolution*, 35(7): 1812–1819.

Kosakovsky Pond, S. L., Poon, A. F., Velazquez, R., Weaver, S., Hepler, N. L., Murrell, B., Shank, S. D., Magalis, B. R., Bouvier, D., Nekrutenko, A., and others 2020. HyPhy 2.5—a customizable platform for evolutionary hypothesis testing using phylogenies. *Molecular biology and evolution*, 37(1): 295–299.

Köster, J. and Rahmann, S. 2012. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19): 2520–2522.

Leal, É., Arrais, C. R., Barreiros, M., Farias Rodrigues, J. K., Silva Sousa, N. P., Duarte Costa, D., Rodrigo Pereira Santos, F. D., Dantas Silva, A., Silva Viana, A. I. E., Barros, A. K., and Lima, K. 2020. Characterization of hiv-1 genetic diversity and antiretroviral resistance in the state of maranhão, northeast brazil. *PLoS One*, 15(3): e0230878.

Leigh Brown, A. J., Lycett, S. J., Weinert, L., Hughes, G. J., Fearnhill, E., and Dunn, D. T. 2011. Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *Journal of Infectious Diseases*, 204(9): 1463–1469.

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. *Current HIV research*, 17(1): 53–64.

Li, D., Chen, H., Li, H., Ma, Y., Dong, L., Dai, J., Jin, X., Yang, M., Zeng, Z., Sun, P., Song, Z., and Chen, M. 2022. Hiv-1 pretreatment drug resistance and genetic transmission network in the southwest border region of china. *BMC Infect Dis*, 22(1): 741.

Liesenborgs, J., Hendrickx, D. M., Kuylen, E., Niyukuri, D., Hens, N., and Delva, W. 2018. SimpactCyan 1.0: An Open-source Simulator for Individual-Based Models in HIV Epidemiology with R and Python Interfaces. *bioRxiv*, page 440834.

Little, S. J., Kosakovsky Pond, S. L., Anderson, C. M., Young, J. A., Wertheim, J. O., Mehta, S. R., May, S., and Smith, D. M. 2014a. Using hiv networks to inform real time prevention interventions. *PLoS One*, 9(6): e98443.

Little, S. J., Pond, S. L. K., Anderson, C. M., Young, J. A., Wertheim, J. O., Mehta, S. R., May, S., and Smith, D. M. 2014b. Using HIV networks to inform real time prevention interventions. *PloS one*, 9(6): e98443.

Liu, M., Han, X., Zhao, B., An, M., He, W., Wang, Z., Qiu, Y., Ding, H., and Shang, H. 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. *Frontiers in Microbiology*, 11: 604993.

Mai, T. Q., Martinez, E., Menon, R., Van Anh, N. T., Hien, N. T., Marais, B. J., and Sintchenko, V. 2018. Mycobacterium tuberculosis Drug Resistance and Transmission among Human Immunodeficiency Virus–Infected Patients in Ho Chi Minh City, Vietnam. *The American Journal of Tropical Medicine and Hygiene*, 99(6): 1397–1406.

McLaughlin, A., Sereda, P., Oliveira, N., Barrios, R., Brumme, C. J., Brumme, Z. L., Montaner, J. S. G., and Joy, J. B. 2019. Detection of HIV transmission hotspots in British

Columbia, Canada: A novel framework for the prioritization and allocation of treatment and prevention resources. *EBioMedicine*, 48: 405–413.

Moshiri, N., Ragonnet-Cronin, M., Wertheim, J. O., and Mirarab, S. 2019. FAVITES: simultaneous simulation of transmission networks, phylogenetic trees and sequences. *Bioinformatics*, 35(11): 1852–1861.

Murphy, D. G., Dion, R., Simard, M., Vachon, M. L., Martel-Laferrière, V., Serhir, B., and Longtin, J. 2019a. Molecular surveillance of hepatitis C virus genotypes identifies the emergence of a genotype 4d lineage among men in Quebec, 2001-2017. *Canada Communicable Disease Report = Releve Des Maladies Transmissibles Au Canada*, 45(9): 230–237.

Murphy, D. G., Dion, R., Simard, M., Vachon, M. L., Martel-Laferrière, V., Serhir, B., and Longtin, J. 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(9): 230–237.

Ng, K. T., Ng, L. J., Oong, X. Y., Chook, J. B., Chan, K. G., Takebe, Y., Kamarulzaman, A., and Tee, K. K. 2022. Application of a vp4/vp2-inferred transmission clusters in estimating the impact of interventions on rhinovirus transmission. *Virol J*, 19(1): 36.

Nguinkal, J. A., Zoclanclounon, Y. A. B., Molina, A., Roba, A., Nyakio, N. M., Lokamar, P. N., Nzoyikorera, N., Ihorimbere, T., Nyandwi, J., Aguer, M. A., Maror, J. A., Lokore, M. L., Francis, M. F., Mapunda, L. A., Beyanga, M., Muyigi, T., Pimundu, G., Nabadda, S. N., Kabalisa, E., Umuringa, J. d., Tare, I. M., Lagu, H. I., Achol, E., May, J., Affara, M., and Gehre, F. 2024. Assessment of the pathogen genomics landscape highlights disparities and challenges for effective AMR Surveillance and outbreak response in the East African community. *BMC public health*, 24(1): 1500.

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

Novitsky, V., Steingrimsson, J. A., Howison, M., Gillani, F. S., Li, Y., Manne, A., Fulton, J., Spence, M., Parillo, Z., Marak, T., Chan, P. A., Bertrand, T., Bandy, U., Alexander-Scott, N., Dunn, C. W., Hogan, J., and Kantor, R. 2020. Empirical comparison of analytical approaches for identifying molecular HIV-1 clusters. *Scientific Reports*, 10(1): 18547.

Novitsky, V., Steingrimsson, J., Howison, M., Dunn, C., Gillani, F. S., Manne, A., Li, Y., Spence, M., Parillo, Z., Fulton, J., Marak, T., Chan, P., Bertrand, T., Bandy, U., Alexander-Scott, N., Hogan, J., and Kantor, R. 2021. Longitudinal Typing of Molecular

HIV Clusters in a Statewide Epidemic. *AIDS (London, England)*, 35(11): 1711–1722.

Oster, A. M., France, A. M., Panneer, N., Ocfemia, M. C. B., Campbell, E., Dasgupta, S., Switzer, W. M., Wertheim, J. O., and Hernandez, A. L. 2018a. Identifying clusters of recent and rapid HIV transmission through analysis of molecular surveillance data. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 79(5): 543–550.

Oster, A. M., France, A. M., Panneer, N., Bañez Ocfemia, M. C., Campbell, E., Dasgupta, S., Switzer, W. M., Wertheim, J. O., and Hernandez, A. L. 2018b. Identifying Clusters of Recent and Rapid HIV Transmission Through Analysis of Molecular Surveillance Data. *Journal of Acquired Immune Deficiency Syndromes (1999)*, 79(5): 543–550.

Oster, A. M., Lyss, S. B., McClung, R. P., Watson, M., Panneer, N., Hernandez, A. L., Buchacz, K., Robilotto, S. E., Curran, K. G., Hassan, R., Ocfemia, M. C. B., Linley, L., Perez, S. M., Phillip, S. A., and France, A. M. 2021a. HIV Cluster and Outbreak Detection and Response: The Science and Experience. *American Journal of Preventive Medicine*, 61(5 Suppl 1): S130–S142.

Oster, A. M., Panneer, N., Lyss, S. B., McClung, R. P., Watson, M., Saduvala, N., Ocfemia, M. C. B., Linley, L., Switzer, W. M., Wertheim, J. O., Campbell, E., Hernandez, A. L., and France, A. M. 2021b. Increasing Capacity to Detect Clusters of Rapid HIV Transmission in Varied Populations-United States. *Viruses*, 13(4): 577.

Pan, X., Jiang, J., Ma, Q., Zhang, J., Yang, J., Chen, W., Ding, X., Fan, Q., Guo, Z., Xia, Y., Xia, S., and Wu, Z. 2018. Outbreak of HIV Infection Linked to Nosocomial Transmission, China, 2016–2017. *Emerging Infectious Diseases*, 24(12): 2141–2149.

Paraschiv, S., Banica, L., Nicolae, I., Niculescu, I., Abagiu, A., Jipa, R., Pineda-Peña, A.-C., Pingarilho, M., Neaga, E., Theys, K., Libin, P., Otelea, D., and Abecasis, A. 2017. Epidemic dispersion of HIV and HCV in a population of co-infected Romania injecting drug users. *PLoS One*, 12(10): e0185866.

Paraskevis, D., Nikolopoulos, G. K., Magiorkinis, G., Hodges-Mameletzis, I., and Hatzakis, A. 2016. The application of HIV molecular epidemiology to public health. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 46: 159–168.

Patil, A., Patil, S., Rao, A., Gadhe, S., Kurle, S., and Panda, S. 2022. Exploring the Evolutionary History and Phylodynamics of Human Immunodeficiency Virus Type 1 Outbreak From Unnao, India Using Phylogenetic Approach. *Frontiers in Microbiology*, 13: 848250.

Penn, O., Stern, A., Rubinstein, N. D., Dutheil, J., Bacharach, E., Galtier, N., and Pupko, T. 2008. Evolutionary modeling of rate shifts reveals specificity determinants in hiv-1 subtypes. *PLoS Comput Biol*, 4(11): e1000214.

Pérez-Losada, M., Castel, A. D., Lewis, B., Kharfen, M., Cartwright, C. P., Huang, B., Maxwell, T., Greenberg, A. E., Crandall, K. A., and DC Cohort Executive Committee 2017. Characterization of HIV diversity, phylodynamics and drug resistance in Washington, DC. *PLoS One*, 12(9): e0185644.

Peters, P. J., Pontones, P., Hoover, K. W., Patel, M. R., Galang, R. R., Shields, J., Blosser, S. J., Spiller, M. W., Combs, B., Switzer, W. M., Conrad, C., Gentry, J., Khudyakov, Y., Waterhouse, D., Owen, S. M., Chapman, E., Roseberry, J. C., McCants, V., Weidle, P. J., Broz, D., Samandari, T., Mermin, J., Walthall, J., Brooks, J. T., Duwve, J. M., and Indiana HIV Outbreak Investigation Team 2016. HIV Infection Linked to Injection Use of Oxymorphone in Indiana, 2014-2015. *The New England Journal of Medicine*, 375(3): 229–239.

Pineda-Peña, A.-C., Schrooten, Y., Vinken, L., Ferreira, F., Li, G., Trovão, N. S., Khouri, R., Derdelinckx, I., De Munter, P., Kücherer, C., Kostrikis, L. G., Nielsen, C., Littsola, K., Wensing, A., Stanojevic, M., Paredes, R., Balotta, C., Albert, J., Boucher, C., Gomez-Lopez, A., Van Wijngaerden, E., Van Ranst, M., Vercauteren, J., Vandamme, A.-M., and Van Laethem, K. 2014. Trends and Predictors of Transmitted Drug Resistance (TDR) and Clusters with TDR in a Local Belgian HIV-1 Epidemic. *PLoS ONE*, 9(7): e101738.

Poon, A. F. Y. 2016. Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks. *Virus Evolution*, 2(2).

Poon, A. F. Y., Gustafson, R., Daly, P., Zerr, L., Demlow, S. E., Wong, J., Woods, C. K., Hogg, R. S., Krajden, M., Moore, D., Kendall, P., Montaner, J. S. G., and Harrigan, P. R. 2016. Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. *The lancet. HIV*, 3(5): e231–238.

Potterat, J. J., Phillips-Plummer, L., Muth, S. Q., Rothenberg, R. B., Woodhouse, D. E., Maldonado-Long, T. S., Zimmerman, H. P., and Muth, J. B. 2002. Risk network structure in the early epidemic phase of HIV transmission in Colorado Springs. *Sexually Transmitted Infections*, 78(suppl 1): i159–i163.

Ragonnet-Cronin, M., Hodcroft, E., Hué, S., Fearnhill, E., Delpech, V., Brown, A. J. L., and Lycett, S. 2013. Automated analysis of phylogenetic clusters. *BMC Bioinformatics*, 14(1): 317.

Ragonnet-Cronin, M., Hué, S., Hodcroft, E. B., Tostevin, A., Dunn, D., Fawcett, T., Pozniak, A., Brown, A. E., Delpech, V., Brown, A. J. L., and UK HIV Drug Resistance Database 2018a. Non-disclosed men who have sex with men in UK HIV transmission networks: phylogenetic analysis of surveillance data. *Lancet HIV*, 5(6): e309–e316.

Ragonnet-Cronin, M., Jackson, C., Bradley-Stewart, A., Aitken, C., McAuley, A., Palmateer, N., Gunson, R., Goldberg, D., Milosevic, C., and Leigh Brown, A. J. 2018b. Recent and rapid transmission of hiv among people who inject drugs in scotland revealed through phylogenetic analysis. *The Journal of infectious diseases*, 217(12): 1875–1882.

Ragonnet-Cronin, M., Jackson, C., Bradley-Stewart, A., Aitken, C., McAuley, A., Palmateer, N., Gunson, R., Goldberg, D., Milosevic, C., and Leigh Brown, A. J. 2018c. Recent and Rapid Transmission of HIV Among People Who Inject Drugs in Scotland Revealed Through Phylogenetic Analysis. *The Journal of Infectious Diseases*, 217(12): 1875–1882.

Ragonnet-Cronin, M., Jackson, C., Bradley-Stewart, A., Aitken, C., McAuley, A., Palmateer, N., Gunson, R., Goldberg, D., Milosevic, C., and Leigh Brown, A. J. 2018d. Recent and rapid transmission of HIV among people who inject drugs in Scotland revealed through phylogenetic analysis. *The Journal of infectious diseases*, 217(12): 1875–1882.

Ragonnet-Cronin, M., Hu, Y. W., Morris, S. R., Sheng, Z., Poortinga, K., and Wertheim, J. O. 2019. HIV transmission networks among transgender women in Los Angeles County, CA, USA: a phylogenetic analysis of surveillance data. *The Lancet HIV*, 6(3): e164–e172.

Ragonnet-Cronin, M., Benbow, N., Hayford, C., Poortinga, K., Ma, F., Forgione, L. A., Sheng, Z., Hu, Y. W., Torian, L. V., and Wertheim, J. O. 2021. Sorting by race/ethnicity across hiv genetic transmission networks in three major metropolitan areas in the united states. *AIDS research and human retroviruses*, 37(10): 784–792.

Ragonnet-Cronin, M., Hayford, C., D'Aquila, R., Ma, F., Ward, C., Benbow, N., and Wertheim, J. O. 2022a. Forecasting HIV-1 Genetic Cluster Growth in Illinois,United States. *Journal of Acquired Immune Deficiency Syndromes (1999)*, 89(1): 49–55.

Ragonnet-Cronin, M., Hayford, C., D'Aquila, R., Ma, F., Ward, C., Benbow, N., and Wertheim, J. O. 2022b. Forecasting HIV-1 Genetic Cluster Growth in Illinois,United States. *Journal of Acquired Immune Deficiency Syndromes (1999)*, 89(1): 49–55.

Rasmussen, D. A., Wilkinson, E., Vandormael, A., Tanser, F., Pillay, D., Stadler, T., and de Oliveira, T. 2018. Tracking external introductions of hiv using phylodynamics reveals a major source of infections in rural kwazulu-natal, south africa. *Virus Evolution*, 4(2): vey037.

Reichmuth, M. L., Chaudron, S. E., Bachmann, N., Nguyen, H., Böni, J., Metzner, K. J., Perreau, M., Klimkait, T., Yerly, S., Hirsch, H. H., Hauser, C., Ramette, A., Vernazza, P., Cavassini, M., Bernasconi, E., Günthard, H. F., Kusejko, K., Kouyos, R. D., and Swiss HIV Cohort Study 2021. Using longitudinally sampled viral nucleotide sequences to characterize the drivers of HIV-1 transmission. *HIV medicine*, 22(5): 346–359.

Rhee, S.-Y., Magalis, B. R., Hurley, L., Silverberg, M. J., Marcus, J. L., Slome, S., Kosakovsky Pond, S. L., and Shafer, R. W. 2019a. National and International Dimensions of Human Immunodeficiency Virus-1 Sequence Clusters in a Northern California Clinical Cohort. *Open Forum Infectious Diseases*, 6(4): ofz135.

Rhee, S.-Y., Magalis, B. R., Hurley, L., Silverberg, M. J., Marcus, J. L., Slome, S., Kosakovsky Pond, S. L., and Shafer, R. W. 2019b. National and International Dimensions of Human Immunodeficiency Virus-1 Sequence Clusters in a Northern California Clinical Cohort. *Open Forum Infectious Diseases*, 6(4).

Riley, L. W. and Blanton, R. E. 2018. Advances in Molecular Epidemiology of Infectious Diseases: Definitions, Approaches, and Scope of the Field. *Microbiology Spectrum*, 6(6).

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 Computational Biology*, 9(6): e1003105.

Rodriguez, F., Oliver, J. L., Marin, A., and Medina, J. R. 1990. The general stochastic model of nucleotide substitution. *Journal of theoretical biology*, 142(4): 485–501.

Rose, R., Cross, S., Lamers, S. L., Astemborski, J., Kirk, G. D., Mehta, S. H., Sievers, M., Martens, C., Bruno, D., Redd, A. D., and Laeyendecker, O. 2020. Persistence of HIV transmission clusters among people who inject drugs. *AIDS (London, England)*, 34(14): 2037–2044.

Rouet, F., Nouhin, J., Zheng, D.-P., Roche, B., Black, A., Prak, S., Leoz, M., Gaudy-Graffin, C., Ferradini, L., Mom, C., Mam, S., Gautier, C., Lesage, G., Ken, S., Phon, K., Kerleguer, A., Yang, C., Killam, W., Fujita, M., Mean, C., Fontenille, D., Barin, F., Plantier, J.-C., Bedford, T., Ramos, A., and Saphonn, V. 2018. Massive Iatrogenic Outbreak of Human Immunodeficiency Virus Type 1 in Rural Cambodia, 2014–2015. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 66(11): 1733–1741.

Sahibzada, K. I., Ganova-Raeva, L., Dimitrova, Z., Ramachandran, S., Lin, Y., Longmire, G., Arthur, L., Xia, G.-L., Khudyakov, Y., Khan, I., and Sadaf, S. 2022. Hepatitis C virus transmission cluster among injection drug users in Pakistan. *PloS One*, 17(7): e0270910.

Samoff, E., Mobley, V., Hudgins, M., Cope, A. B., Adams, N. D., Caputo, C. R., Dennis, A. M., Billock, R. M., Crowley, C. A., Clymore, J. M., and Foust, E. 2020. HIV Outbreak Control With Effective Access to Care and Harm Reduction in North Carolina, 2017-2018. *American Journal of Public Health*, 110(3): 394–400.

Sivay, M. V., Hudelson, S. E., Wang, J., Agyei, Y., Hamilton, E. L., Selin, A., Dennis, A., Kahn, K., Gomez-Olive, F. X., MacPhail, C., Hughes, J. P., Pettifor, A., Eshleman, S. H., and Grabowski, M. K. 2018. HIV-1 diversity among young women in rural South Africa: HPTN 068. *PloS One*, 13(7): e0198999.

Sizemore, L., Fill, M.-M., Mathieson, S. A., Black, J., Brantley, M., Cooper, K., Garrett, J., Switzer, W. M., Peters, P. J., and Wester, C. 2020. Using an Established Outbreak Response Plan and Molecular Epidemiology Methods in an HIV Transmission Cluster Investigation, Tennessee, January-June 2017. *Public Health Reports (Washington, D.C.: 1974)*, 135(3): 329–333.

Spirtes, P. and Glymour, C. 1991. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1): 62–72.

Stecher, M., Chaillon, A., Eberle, J., Behrens, G. M. N., Eis-Hübinger, A.-M., Lehmann, C., Jablonka, A., Bogner, J., Fätkenheuer, G., Spinner, C. D., Wasmuth, J.-C., Kaiser, R., Mehta, S. R., Vehreschild, J. J., and Hoenigl, M. 2018. Molecular epidemiology of the hiv epidemic in three german metropolitan regions - cologne/bonn, munich and hannover, 1999-2016. *Sci Rep*, 8(1): 6799.

Tamura, K. and Nei, M. 1993a. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10(3): 512–526.

Tamura, K. and Nei, M. 1993b. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10(3): 512–526.

Temereanca, A., Oprea, C., Wertheim, J. O., Ianache, I., Ceausu, E., Cernescu, C., Mehta, S. R., and Ruta, S. 2017. Hiv transmission clusters among injecting drug users in romania. *Rom Biotechnol Lett*, 22(1): 12307–12315.

Thoma, R., Seneghini, M., Seiffert, S. N., Vuichard Gysin, D., Scanferla, G., Haller, S., Flury, D., Boggian, K., Kleger, G.-R., Filipovic, M., Nolte, O., Schlegel, M., and Kohler, P. 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. *Antimicrobial Resistance and Infection Control*, 11: 12.

Tookes, H., Bartholomew, T. S., Geary, S., Matthias, J., Poschman, K., Blackmore, C., Philip, C., Suarez, E., Forrest, D. W., Rodriguez, A. E., and others 2019. Rapid Identification and Investigation of an HIV Risk Network Among People Who Inject Drugs–Miami, FL, 2018. *AIDS and Behavior*, pages 1–11. Publisher: Springer.

Tookes, H., Bartholomew, T. S., Geary, S., Matthias, J., Poschman, K., Blackmore, C., Philip, C., Suarez, E., Forrest, D. W., Rodriguez, A. E., Kolber, M. A., Knaul, F., Colucci, L., and Spencer, E. 2020. Rapid Identification and Investigation of an HIV Risk Network Among People Who Inject Drugs -Miami, FL, 2018. *AIDS and behavior*, 24(1): 246–256.

Tordoff, D. M., Buskin, S., Lechtenberg, R., Golden, M. R., Kerani, R. P., and Herbeck, J. T. 2021. Combining traditional and molecular epidemiology methods to quantify local HIV transmission among foreign-born residents. *AIDS (London, England)*, 35(4): 655–664.

Tumpney, M., John, B., Panneer, N., McClung, R. P., Campbell, E. M., Roosevelt, K., DeMaria, A., Buchacz, K., Switzer, W. M., Lyss, S., and Cranston, K. 2020. Human Immunodeficiency Virus (HIV) Outbreak Investigation Among Persons Who Inject Drugs in Massachusetts Enhanced by HIV Sequence Data. *The Journal of Infectious Diseases*, 222(Suppl 5): S259–S267.

Tümmler, B. 2020. Molecular epidemiology in current times. *Environmental Microbiology*, 22(12): 4909–4918.

Volz, E. M. and Frost, S. D. 2014. Sampling through time and phylodynamic inference with coalescent and birth–death models. *Journal of The Royal Society Interface*, 11(101): 20140945.

von Rotz, M., Kuehl, R., Durovic, A., Zingg, S., Apitz, A., Wegner, F., Seth-Smith, H. M. B., Roloff, T., Leuzinger, K., Hirsch, H. H., Kuster, S., Battegay, M., Mariani, L., Schaeren, S., Bassetti, S., Banderet-Uglioni, F., Egli, A., and Tschudin-Sutter, S. 2023. A systematic outbreak investigation of SARS-CoV-2 transmission clusters in a tertiary academic care center. *Antimicrobial Resistance and Infection Control*, 12(1): 38.

Vrancken, B., Adachi, D., Benedet, M., Singh, A., Read, R., Shafran, S., Taylor, G. D., Simmonds, K., Sikora, C., Lemey, P., Charlton, C. L., and Tang, J. W. 2017. 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: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 52: 100–105.

Wang, X., Wu, Y., Mao, L., Xia, W., Zhang, W., Dai, L., Mehta, S. R., Wertheim, J. O., Dong, X., Zhang, T., Wu, H., and Smith, D. M. 2015a. Targeting HIV Prevention Based on Molecular Epidemiology Among Deeply Sampled Subnetworks of Men Who Have Sex With Men. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 61(9): 1462–1468.

Wang, X., Wu, Y., Mao, L., Xia, W., Zhang, W., Dai, L., Mehta, S. R., Wertheim, J. O., Dong, X., Zhang, T., Wu, H., and Smith, D. M. 2015b. Targeting HIV Prevention Based on Molecular Epidemiology Among Deeply Sampled Subnetworks of Men Who Have Sex With Men. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 61(9): 1462–1468.

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. *Molecular Biology and Evolution*, 35(3): 773–777.

Wertheim, J. O., Richman, D. D., Mehta, S. R., Kosakovsky Pond, S. L., Smith, D. M., Leigh Brown, A. J., and Hepler, N. L. 2013. The Global Transmission Network of HIV-1. *The Journal of Infectious Diseases*, 209(2): 304–313.

Wertheim, J. O., Leigh Brown, A. J., Hepler, N. L., Mehta, S. R., Richman, D. D., Smith, D. M., and Kosakovsky Pond, S. L. 2014a. The global transmission network of HIV-1. *J Infect Dis*, 209(2): 304–13.

Wertheim, J. O., Leigh Brown, A. J., Hepler, N. L., Mehta, S. R., Richman, D. D., Smith, D. M., and Kosakovsky Pond, S. L. 2014b. The global transmission network of hiv-1. *The Journal of infectious diseases*, 209(2): 304–313.

Wertheim, J. O., Oster, A. M., Hernandez, A. L., Saduvala, N., Bañez Ocfemia, M. C., and Hall, H. I. 2016. The International Dimension of the U.S. HIV Transmission Network and Onward Transmission of HIV Recently Imported into the United States. *AIDS Research and Human Retroviruses*, 32(10-11): 1046–1053.

Wertheim, J. O., Pond, S. L. K., Forgione, L. A., Mehta, S. R., Murrell, B., Shah, S., Smith, D. M., Scheffler, K., and Torian, L. V. 2017a. Social and genetic networks of HIV-1 transmission in New York City. *PLoS pathogens*, 13(1): e1006000.

Wertheim, J. O., Oster, A. M., Johnson, J. A., Switzer, W. M., Saduvala, N., Hernandez, A. L., Hall, H. I., and Heneine, W. 2017b. Transmission fitness of drug-resistant HIV revealed in a surveillance system transmission network. *Virus Evolution*, 3(1): vex008.

Wertheim, J. O., Oster, A. M., Johnson, J. A., Switzer, W. M., Saduvala, N., Hernandez, A. L., Hall, H. I., and Heneine, W. 2017c. Transmission fitness of drug-resistant hiv revealed in a surveillance system transmission network. *Virus evolution*, 3(1): vex008.

Wertheim, J. O., Murrell, B., Mehta, S. R., Forgione, L. A., Kosakovsky Pond, S. L., Smith, D. M., and Torian, L. V. 2018a. Growth of hiv-1 molecular transmission clusters in new york city. *The Journal of infectious diseases*, 218(12): 1943–1953.

Wertheim, J. O., Murrell, B., Mehta, S. R., Forgione, L. A., Kosakovsky Pond, S. L., Smith, D. M., and Torian, L. V. 2018b. Growth of HIV-1 molecular transmission clusters in New York city. *The Journal of infectious diseases*, 218(12): 1943–1953.

Wertheim, J. O., Chato, C., and Poon, A. F. Y. 2019. Comparative analysis of HIV sequences in real time for public health. *Current opinion in HIV and AIDS*, 14(3): 213–220.

Wolf, E., Herbeck, J. T., Van Rompaey, S., Kitahata, M., Thomas, K., Pepper, G., and Frenkel, L. 2017. Short Communication: Phylogenetic Evidence of HIV-1 Transmission Between Adult and Adolescent Men Who Have Sex with Men. *AIDS research and human retroviruses*, 33(4): 318–322.

Xiang, F., Frost, S. D. W., and Dearlove, B. L. 2017. Biased phylodynamic inferences from analysing clusters of viral sequences. *Virus Evolution*, 3(2).

Yan, H., He, W., Huang, L., Wu, H., Liang, Y., Li, Q., Shui, J., Wang, C., Dzakah, E. E., Han, Z., and Tang, S. 2020. The Central Role of Nondisclosed Men Who Have Sex With Men in Human Immunodeficiency Virus-1 Transmission Networks in Guangzhou, China. *Open Forum Infectious Diseases*, 7(5): ofaa154.

Yan, H., Wu, H., Xia, Y., Huang, L., Liang, Y., Li, Q., Chen, L., Han, Z., and Tang, S. 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.

Ye, J., Sun, Y., Li, J., Lu, X., Zheng, M., Liu, L., Yu, F., He, S., Xu, C., Ren, X., Wang, J., Chen, J., Ruan, Y., Feng, Y., Shao, Y., Xing, H., and Lu, H. 2023. Distribution pattern, molecular transmission networks, and phylodynamic of hepatitis c virus in china. *PLoS One*, 18(12): e0296053.

Yebra, G., Ragonnet-Cronin, M., Ssemwanga, D., Parry, C. M., Logue, C. H., Cane, P. A., Kaleebu, P., and Brown, A. J. L. 2015. Analysis of the history and spread of HIV-1 in uganda using phylodynamics. *J Gen Virol*, 96(Pt 7): 1890–8.