AUTO-TUNE: FINDING AN OPTIMAL DISTANCE THRESHOLD FOR INFERRING HIV TRANSMISSION CLUSTERS

Steven Weaver [1*], **Vanessa Davila Conn** [2], **Hannah Verdonk** [1], **Joel Wertheim** [3], **and Sergei L. Kosakovsky Pond** [1]

[1] *Center for Viral Evolution, Temple University, Philadelphia, PA, USA*
[2] *Center for Research in Infectious Diseases, National Institute of Respiratory Diseases, Mexico City, Mexico*
[3] *Department of Medicine, University of California, San Diego, CA*

Correspondence*:
Steven Weaver
sweaver@temple.edu

ABSTRACT

Choosing an appropriate distance threshold is an important part of inferring a transmission network to determine the relative growth of clusters within a localized epidemic. This distance threshold determines how close two consensus sequences must be in order for a link to be created between them in the network. Using a distance threshold that is too high can result in a network with many unnecessary links, making it difficult to interpret and analyze. On the other hand, using a distance threshold that is too low can result in a network with too few links, which may not capture key insights into rapidly growing clusters among patients with shared attributes that could benefit from public health intervention measures.

Here, we present a heuristic scoring approach for tuning a distance threshold by associating each tested threshold against the maximal number of clusters created across all thresholds and the difference between the ratio (R_{12}) of the largest cluster in the network to the second largest cluster at each iteration. The number of clusters is normalized between $[0, 1]$ then gated via a Gompertz function transform. Meanwhile, the distribution of all R_{12} ratios are converted to Z scores, and normalized relative to the largest positive Z score across all candidate distances. The priority score is the sum of aforementioned two components.

Published research using the HIV-TRACE software package frequently use the default threshold of 1.5% for HIV pol gene sequences. We apply our scoring heuristic to outbreaks with different characteristics, such as regional or temporal variability, and demonstrate the utility of using the scoring mechanism's suggested distance threshold to identify clusters exhibiting risk factors that would have otherwise been more difficult to identify. For example, while we found that a 1.5% distance threshold is typical for US-like epidemics, recent outbreaks like the CRF07_BC subtype among men who have sex with men (MSM) in China has been found to have a lower optimal threshold of 0.5% to better capture the transition from injected drug use (IDU) to MSM as the primary risk factor. Alternatively, in communities surrounding Lake Victoria, where there has been sustained transmission for several years, we found that a larger distance threshold is suitable to

capture a more risk factor diverse populace with sparse sampling over a longer period of time. Such identification may allow for more informed intervention action by respective public health officials.

Keywords: molecular epidemiology, HIV, network, transmission cluster, surveillance

1 INTRODUCTION

Choosing an appropriate distance threshold is an important part of using a transmission network to track the spread of a contagious disease. This distance threshold determines how close two individuals must be in order for a link to be created between them in the network.

Using a distance threshold that is too small can result in a network with many unnecessary links, making it difficult to interpret and analyze. On the other hand, using a distance threshold that is too large can result in a network with too few links, making it difficult to accurately track the spread of the disease.

To ensure that the transmission network is useful and informative, it is important to carefully consider the appropriate distance threshold. This may vary depending on the specific disease and the context in which it is spreading. For example, a highly contagious respiratory illness may require a smaller distance threshold than a less contagious illness that is primarily spread through direct contact.

In general, the goal is to strike a balance between having enough links to accurately track the spread of the disease, while not having so many links that the network becomes difficult to interpret. This can be achieved through careful analysis and consideration of the specific disease and context.

Overall, choosing an appropriate distance threshold is an important step in using a transmission network to track the spread of a contagious disease. It can help ensure that the network is useful and informative, and can ultimately aid in efforts to control and prevent the spread of the disease.

2 METHODS

2.1 Scoring Heuristic Procedure

Network threshold selection procedure proceeds as follows:

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 largest cluster to the second largest cluster, and C – the number of clusters in the network.
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 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 $d_L^{i+1} \dots d_L^{i+W}$ and the trailing window $d_L^{i-W} \dots d_L^{i-1}$. The width of the window, W, is defined as $(([\frac{N}{100}], 3), 30)$. The distribution of ratios is converted to Z scores, and normalized relative to the largest positive Z score across all candidate distances, yielding the jump component of the score.

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, while a threshold that yields 85% - a score of 0.376.

c. The priority score for d_L^i is the sum of the two components defined in (a) and (b).

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 the plot like the one on the analyze page is recommended to ensure that the threshold is sensible.

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 (i.e., the number of connections they have to other nodes in the network). It is used to quantify the extent to which nodes with similar attributes tend to be connected to each other more frequently than would be expected by chance. 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 degree.

In mathematical terms, it is defined as:

$$DWH = \frac{W_M + W_C - 2W_X}{\frac{d_{in}}{nodes_{in}^2} + \frac{d_{out}}{nodes_{out}^2}} \tag{1}$$

Where

- W_M : Weight of in-group connections
- W_C : Weight of out-group connections
- W_X : Weight of cross-group connections
- d_{in} : In-group degree
- d_{out} : Out-group degree
- $nodes_{in}$: number of in-group nodes
- $nodes_{out}$: number of out-group nodes

DWH ranges from -1 to 1. A DWH value of 0 indicates that there is no more homophily than expected with chance, while a value of 1 indicates that there is perfect homophily (e.g. Birds always link to birds, and only birds). A value of -1 is achieved for perfectly disassortative networks (e.g. Bird never linking with another bird).

DWH is used in social network analysis and in the study of how different attributes are related to the formation of connections between individuals. It is used as a way to measure the similarity of attributes between individuals in a network. Additionally, randomization is performed by shuffling attribute labels among nodes, then performing DWH computation. This is useful in creating a null distribution of DWH scores under random mixing. A panmictic range is reported by shuffling attributes multiple times and reporting the minimum and maximum score.

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 wth the tn93 tool and a supplied multiple sequence alignment. This generated pairwise distances are supplied to the hivnetworkcsv script while providing the -A keyword argument. A brief outline of the software's implementation are as follows

1. Calculate pairwise distances: The user first calculates the pairwise distances using the tn93 fast pairwise distance calculator, providing the necessary threshold value and the input FASTA file. The command for this step is

   ```
   tn93 -t 0.030 pol.fasta > pairwise_distances.15.tn93.csv
   ```

   Please note that the threshold should include the maximal range one is intending to test.
2. Compute distance threshold scores: The hivnetworkcsv script is then executed with the required input file, format, and autotune option to generate a tab-separated output file, as shown below

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

3. Visualize the report: Users can upload the generated autotune_report.tsv file to `http://autotune.datamonkey.org/analyze` for visualization and further analysis of the data. This web-based site extends the Datamonkey platform Weaver et al. (2018) to provide an interactive environment to explore scores and other metrics across the range of tested outputs.
4. Run HIV-TRACE: Once AUTO-TUNEd threshold(s) are settled upon after review, the user runs the HIV-TRACE command with the appropriate input FASTA file, distance threshold, and other required arguments. The output is saved as a JSON file. An example command is

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

2.3.1 Optional : Compute Assortatviity Metrics

5. 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.
6. 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.

   AUTO-TUNE is readily accessible on GitHub as part of the hivclustering repository (https://github.com/veg/hivclustering). It is integrated into the command-line interface of the software as the -A or –auto-profile argument. hivclustering is a key component of the HIV-TRACE suite of tools, a resource for the inference, analysis and visualization of HIV transmission networks.

The Degree Weighted Homophily (DWH) calculation tool, an integral component of the assortativity step, is developed using TypeScript, a statically typed superset of JavaScript that ensures robustness and scalability. In an effort to promote accessibility and ease of integration, the DWH tool is packaged and distributed through the Node Package Manager (NPM), enabling researchers and developers to conveniently incorporate this advanced analytical tool into their own projects and workflows. DWH can be used in-browser or as a command line tool, allowing researchers and developers to employ the tool in an interactive command-line interface or integrate it into larger software applications, thus catering to a diverse array of technical needs and preferences. Instructions for usage and installation is found on Github (https://github.com/veg/dwh).

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

2.4 Visualization

Visualizations of AUTO-TUNE results are accessible at http://autotune.datamonkey.org/analyze. It is a dynamic and interactive web-based platform that offers visualization and analysis of results generated by AUTO-TUNE. The website provides a comprehensive view of the data by generating various plots across candidate distance thresholds. These include a score plot, allowing users to identify trends and anomalies across the full range of thresholds. Additionally, it generates a graph showing the number of clusters across candidate thresholds, one of the components that contribute to the score. The site also includes an R1/R2 plot that displays the ratio of the largest cluster to the second largest cluster across candidate thresholds, which is the other metric that contributes to the scoring heuristic.

An assortativity tool is available at http://autotune.datamonkey.org/assortativity, and is an advanced 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. A notable feature is the computation of the panmictic 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. Lastly, the site also provides a plot of the fraction of pairwise connections, normalized by degree, for each value pertinent to the selected field. This visual depiction facilitates an intuitive comprehension of the distribution and interconnections within the dataset.

The site aims to offer a user-friendly interface for data visualization, playing an important role in interpreting and understanding AUTO-TUNE's output data. The visualization code is available on Github (https://github.com/stevenweaver/autotune-app/).

2.5 Comparisons with previously published analyses

In conducting our comparisons with the established clustuneR method, we procured our 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 prescribed in Section 2.3. This enabled us to determine an optimal threshold for each dataset using our proposed method, 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).

To perform comparisons, we computed the average degree-weighted homophily score over a set of three-year sliding windows. Specifically, the homophily among nodes was calculated for a collection of date ranges as follows:

$$\bar{H} = \frac{1}{N} \sum_{i=1}^{N} H(w_i) \tag{2}$$

where $\bar{H}$ represents the average degree-weighted homophily score, N is the total number of sliding windows, $H(w_i)$ is the homophily score for the i-th window, and the windows w_i correspond to the date ranges, e.g., ’2012-2015’, ’2013-2016’, ’2014-2017’, etc. This methodology allowed us to compare the ”best thresholds” derived from our proposed AUTO-TUNE method against those defined as optimal in Chato et al. (2020).

Second, we set out to compare the thresholds obtained in original investigations 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. Thus, we 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.

In order to evaluate the influence of sampling density on the genetic distance threshold as determined by AUTO-TUNE, we implemented a strategy of random subsampling from the original dataset sourced from Rhee et al. (2019). This study was selected due to its satisfactory AUTO-TUNE score when utilized in its entirety, as well as its inherent design as a Geographically-Stratified set of 716 Pol Subtype/CRF (GSPS) reference sequence dataset. The dataset, which comprises 6034 samples gathered between 1959 and 2016, was subjected to random subsampling ten times at proportions of 25%, 50%, and 75% of the original sample size. For each subsample, the optimal threshold and associated scores were determined via AUTO-TUNE.

3 RESULTS

3.1 Comparison with clustuneR

We compared results to clustuneR, 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 determined optimal by clustuneR were found by finding the minimum GAIC (generalized AIC) 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 model that includes individual-level attributes among known cases in the cluster. Using the minimum GAIC metric, it was found that 0.016 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 the same datasets analyzed by clustuneR Chato et al. (2020), AUTO-TUNE found the thresholds with the highest scores to be 0.01431 for Middle Tennessee, 0.01354 for Seattle, and 0.01099 for Northern Alberta. Table 1

Performance of the inferred optimal thresholds were performed using an average degree-weighted homophily (DWH) score across 3-year collection date windows starting from the oldest collection year for each respective dataset, as that is the metadata that was consistently available across all three datasets and was the attribute of focus used by clustuneR. DWH in this case measures the affinity for nodes within the network to link with other nodes in the same collection date window. For example, samples collected between 2012-2015 linking more often with other samples within the same time window would result in a higher DWH score. It was found that, across all three datasets, using the threshold with the highest score reported by AUTO-TUNE resulted in a higher average DWH score across all three datasets.

When reviewing scores across all candidate thresholds with AUTO-TUNE . none of the three datasets reached one confident score over any other. The most confident score was achieved by the Seattle dataset at 1.53325, then Tennessee with a high score of 1.25807, and lastly Canada at 1.01678. When finding peaks, a second peak in Seattle denotes that 0.01166 may also be an optimal threshold to consider, as its score, 1.52976, is only 0.003 less than the highest score. The optimal score for Tennessee and Canada are a bit more dubious, as there are multiple peaks within close scores of each other. Indeed, after applying a 0.75 minimum score threshold after visual inspection for peak ranges, standard deviation among scores were 0.0475, 0.0089, and 0.0084 for Seattle, northern Alberta, and Tennessee, respectively. This implies there may be multiple thresholds that would be considered reasonable, and downstream homophily metrics with attributes may aid in coming to a decision.

3.2 Comparison with Prior Publications Citing HIV-TRACE

Next, we curated publications citing HIV-TRACE that also had publicly accessible data associated with the study Rhee et al. (2019); Brenner et al. (2021); H et al. (2021); Liu et al. (2020); Bbosa et al. (2020); Yan et al. (2020); Dalai et al. (2018); Sivay et al. (2018). We found that a variety of different ways were used to determine distance thresholds, from precedent set by the CDC for detecting recent and rapid clusters Yan et al. (2020), because they've been used before in other studies Sivay et al. (2018), to visual inspection of number of clusters and nodes across candidate distance thresholds Liu et al. (2020). As a result, it was found that since these thresholds are determined largely qualitatively, the distance thresholds used tend to be round numbers and only somewhat tuned to their respective geographic region of research. Table 2 When utilizing AUTO-TUNE for the same datasets, more granularity in selecting thresholds can be performed in a straightforward manner. While it may seem appealing to assume that being within a 0.5% threshold of what AUTO-TUNE considers optimal is good enough, further inspection into results reveals that scores can vary widely within a very short distance. For example, when reviewing the distribution of scores for the dataset used by Dalai et al. (2018), the score is exactly 2 at 0.01848, but at thresholds tested at just 0.00002 difference, scores drop preciptously to 1.638 and 0.826 for candidates 0.01846 and 0.0185, respectively. Another example of a seemingly close threshold yet perhaps not optimal is found with Bbosa et al. (2020). While no score across candidate thresholds was found to be above 1.9, a high score was found at distance 1.707, with 1.2415. Contrast this with the threshold used, 1.5%, with a score of 0.0124.

3.3 Evaluating Performance with DWH

To evaluate the performance of an AUTO-TUNEd optimized threshold using degree-weighted homophily, we first evaluated a CRF07_BC network with data from China. 8178 HIV-1 CRF07_BC pol sequences were used from China to construct longitudinal transmission networks, each pol sequence were annotated with risk factor detailing whether the patient was heterosexual (Hetero), Person With Injected Drug Use (PWID), or Men who have Sex with Men (MSM), among other attributes. Using AUTO-TUNE, no distance threshold receives a score above 1.9, but using the default 1.5% threshold is clearly suboptimal . Using a 1.5% threshold, the network captures 5923 nodes, of which 559 are PWID, 3371 MSM, 1993 Hetero, and received an AUTO-TUNE score of 0.029. When evaluating DWH among risk factors, MSM, Hetero, and PWID had scores of 0.211, 0.133, and 0.168, respectively.

When using AUTO-TUNE, two separate ranges appear. The highest score is obtained with distance threshold 0.76% at 1.1369. The second highest score is 1.0303 at distance threshold 0.0019. Networks at both thresholds were also evaluated with DWH based on risk factor. The network at 0.76% captured 3537 nodes, of which 236 are PWID, 2271 MSM, 1030 Hetero. When evaluating DWH among risk factors, MSM and Hetero both had slightly increased scores of 0.237 and 0.185, respectively. PWID DWH dramatically increased to 0.401. The network at 0.19% captures 1654 nodes, of which 151 are PWID, 1075 MSM, 428 Hetero. When evaluating DWH among risk factors, MSM, Hetero, and PWID had slightly increased scores over 0.76% of 0.292, 0.25, and 0.445, respectively.

We next evaluated Rhee et al. (2019) with DWH. The dataset received a clear optimal threshold of 0.01699 with an AUTO-TUNE score of 1.9998. At the default threshold of 1.5%, the AUTO-TUNE score was 0.9782. At threshold 1.5%, the network captured 1351 nodes, compared with 1592 nodes out of 6034 captured at threshold 1.699%. When evaluating DWH with country, substantial change were only found for China and Thailand, improving from no better than random linkage at threshold 1.5% with DWH -0.166 and -0.051, to 0.116 and 0.132 at threshold 1.699%, respectively. Notably, no country scored worse with the optimal AUTO-TUNE threshold, despite it being larger than the default 1.5%.

3.4 Subsampling's Effect on Optimal Thresholds and AUTO-TUNE Scores

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

Sampling at 25% yielded a mean top threshold of 0.021509, median at 0.019765, and standard deviation of 0.004388 . 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 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.

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

4 DISCUSSION

AUTO-TUNE addresses the challenge of manually setting an appropriate genetic distance threshold in HIV transmission network analysis by implementing a heuristic scoring system that measures two properties of networks generated by candidate genetic distance thresholds. In the application of AUTO-TUNE to various datasets, the results demonstrated its efficacy across different epidemic contexts, risk groups, and network sizes. AUTO-TUNE consistently selected thresholds that were comparable, if not better, to those manually chosen in prior studies using the same data, illustrating the value of a more systematic, automated, and data-adaptive 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 the clustuneR methodology. 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.

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. **?** 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.

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

4.1 When a Score is Below 1.9

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, 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.

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 tests positive for HIV, 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.

Consequently, AUTO-TUNE's performance is consistently controlled, irrespective of the fluctuations seen in data collection protocols after a positive 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.

CONFLICT OF INTEREST STATEMENT

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

AUTHOR CONTRIBUTIONS

The Author Contributions section is mandatory for all articles, including articles by sole authors. If an appropriate statement is not provided on submission, a standard one will be inserted during the production process. The Author Contributions statement must describe the contributions of individual authors referred to by their initials and, in doing so, all authors agree to be accountable for the content of the work. Please see here for full authorship criteria.

FUNDING

Details of all funding sources should be provided, including grant numbers if applicable. Please ensure to add all necessary funding information, as after publication this is no longer possible.

ACKNOWLEDGMENTS

This is a short text to acknowledge the contributions of specific colleagues, institutions, or agencies that aided the efforts of the authors.

SUPPLEMENTAL DATA

Supplementary Material should be uploaded separately on submission, if there are Supplementary Figures, please include the caption in the same file as the figure. LaTeX Supplementary Material templates can be found in the Frontiers LaTeX folder.

DATA AVAILABILITY STATEMENT

Data are available at GenBank accession numbers JX160108-JX161480,JX498971-JX498972,JX498976-JX498990,JX498992-JX499018,KU190031-KU190839,KY34691-KY37792,KY883695-KY883762,KY888784-KY888875,KY921717-KY921757,MG434786-MG435347,MG435358-MG436769,MH352627-MH355541,MK25 MK25548,MN424584-MN427369,MT336755-MT336776,MT368043-MT369927.

REFERENCES

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, 948–951. doi:10.1089/AID.2020.0033

Brenner, B. G., Ibanescu, R.-I., Osman, N., Cuadra-Foy, E., Oliveira, M., Chaillon, A., et al. (2021). The Role of Phylogenetics in Unravelling Patterns of HIV Transmission towards Epidemic Control: The Quebec Experience (2002-2020). *Viruses* 13, 1643. doi:10.3390/v13081643

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, veaa011. doi:10.1093/ve/veaa011

Dalai, S. C., Junqueira, D. M., Wilkinson, E., Mehra, R., Kosakovsky Pond, S. L., Levy, V., et al. (2018). Combining Phylogenetic and Network Approaches to Identify HIV-1 Transmission Links in San Mateo County, California. *Frontiers in Microbiology* 9, 2799. doi:10.3389/fmicb.2018.02799

Ding, X., Chaillon, A., Pan, X., Zhang, J., Zhong, P., He, L., et al. (2022). Characterizing genetic transmission networks among newly diagnosed HIV-1 infected individuals in eastern China: 2012–2016. *PLOS ONE* 17, e0269973. doi:10.1371/journal.pone.0269973. Publisher: Public Library of Science

H, Y., H, W., Y, X., L, H., Y, L., Q, L., et al. (2021). Acquisition and transmission of HIV-1 among migrants and Chinese in Guangzhou, China from 2008 to 2012: Phylogenetic analysis of surveillance data. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* 92. doi:10.1016/j.meegid.2021.104870. Publisher: Infect Genet Evol

Holmes, E. C., Zhang, L. Q., Robertson, P., Cleland, A., Harvey, E., Simmonds, P., et al. (1995). The molecular epidemiology of human immunodeficiency virus type 1 in Edinburgh. *The Journal of Infectious Diseases* 171, 45–53. doi:10.1093/infdis/171.1.45

Kosakovsky Pond, S. L., Weaver, S., Leigh Brown, A. J., and Wertheim, J. O. (2018). 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, 1812–1819. doi:10.1093/molbev/msy016

Liu, M., Han, X., Zhao, B., An, M., He, W., Wang, Z., et al. (2020). Dynamics of HIV-1 Molecular Networks Reveal Effective Control of Large Transmission Clusters in an Area Affected by an Epidemic of Multiple HIV Subtypes. *Frontiers in Microbiology* 11, 604993. doi:10.3389/fmicb.2020.604993

Rhee, S.-Y., Magalis, B. R., Hurley, L., Silverberg, M. J., Marcus, J. L., Slome, S., et al. (2019). National and International Dimensions of Human Immunodeficiency Virus-1 Sequence Clusters in a Northern California Clinical Cohort. *Open Forum Infectious Diseases* 6, ofz135. doi:10.1093/ofid/ofz135

Sivay, M. V., Hudelson, S. E., Wang, J., Agyei, Y., Hamilton, E. L., Selin, A., et al. (2018). HIV-1 diversity among young women in rural South Africa: HPTN 068. *PloS One* 13, e0198999. doi:10.1371/journal.pone.0198999

Vrancken, B., Adachi, D., Benedet, M., Singh, A., Read, R., Shafran, S., et al. (2017). The multi-faceted dynamics of HIV-1 transmission in Northern Alberta: A combined analysis of virus genetic and public health data. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 52, 100–105. doi:10.1016/j.meegid.2017.04.005

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, 773–777. doi:10.1093/molbev/msx335

Wolf, E., Herbeck, J. T., Van Rompaey, S., Kitahata, M., Thomas, K., Pepper, G., et al. (2017). Short Communication: Phylogenetic Evidence of HIV-1 Transmission Between Adult and Adolescent Men Who Have Sex with Men. *AIDS research and human retroviruses* 33, 318–322. doi:10.1089/AID.2016.0061

Yan, H., He, W., Huang, L., Wu, H., Liang, Y., Li, Q., et al. (2020). The Central Role of Nondisclosed Men Who Have Sex With Men in Human Immunodeficiency Virus-1 Transmission Networks in Guangzhou, China. *Open Forum Infectious Diseases* 7, ofaa154. doi:10.1093/ofid/ofaa154

TABLES

Table 1. clustuneR Comparison

Dataset	clustuneR		AUTO-TUNE		
	Threshold	Avg. Homophily	Threshold	Avg. Homophily	Max Score
Middle Tennessee	0.0160	0.0079	0.01431	0.0147	1.25807
Seattle	0.0160	0.0259	0.01354	0.0348	1.53325
Northern Alberta	0.0104	-0.0536	0.01099	-0.0448	1.01678

Table 2. Threshold Comparison with Prior Publications Citing HIV-TRACE

PMID	Country	Collection Date	Threshold Used	AUTO-TUNE
29975689	South Africa	2011-2015	2.5%	2.584%
30574123	USA	1997-2008	2%	1.848%
32500089	China	2008-2015	0.5%	0.675%
32693608	Uganda	2009-2016	1.5%	1.707%
33281803	China	2000-2016	0.5%/0.7%	0.676%
33901684	China	2008-2012	1.5%	1.215%
34452506	Canada	1996-2017	1.5%/2.5%	0.547%
31041344	USA	1997-2017	1.5%	0.927%

6 FIGURE CAPTIONS

Table 3. CRF07_BC DWH and Panmictic Range at Different Thresholds

Record	Threshold 1.5%		Threshold 0.76%		Threshold 0.19%	
	DWH	Panmictic Range	DWH	Panmictic Range	DWH	Panmictic Range
MSM	0.211	[-0.105, -0.205]	0.237	[-0.120,-0.240]	0.292	[-0.146, -0.280]
Hetero	0.133	[-0.092, -0.190]	0.185	[-0.100,-0.211]	0.25	[-0.093, -0.256]
PWID	0.168	[-0.001, -0.089]	0.401	[-0.005,-0.081]	0.445	[-0.012, -0.129]

Figure 1. AUTO-TUNE scoring across candidate thresholds ranging from 0% to 2.5% genetic distance. The plots represent the datasets from Seattle, Middle Tennessee, and Northern Alberta as described by clustuneR. The y-axis represents the AUTO-TUNE score, with higher scores suggesting more optimal thresholds. The x-axis represents the candidate thresholds. None of the three datasets exhibited an extreme peak of over 1.9, implying multiple thresholds could serve well and that a more complex decision-making process that includes downstream metrics such as DWH or careful inspection may be necessary.

Figure 2. AUTO-TUNE scoring across candidate thresholds from 0% to 2.5% genetic distance for eight datasets from various studies that have previously employed HIV-TRACE with qualitatively defined thresholds. Each plot represents one dataset, and the y-axis shows the AUTO-TUNE score. Higher scores indicate more optimal thresholds for clustering. The x-axis represents the range of candidate thresholds. Each plot is labeled by the respective studies' PubMed ID.

Figure 3. Hola 3

Figure 4. Hola 4

Figure 5. Hola 5