0929

CONCORDANCE OF METHODS IN IDENTIFYING MOLECULAR HIV CLUSTERS

Vlad Novitsky[1], Jon Steingrimsson[1], Mark Howison[2], Fizza Gillani[1], Yuanning Li[3], Akarsh Manne[1], Matthew Spence[4], Zoanne Parillo[4], Theodore Marak[4], Philip A. Chan[1,4], Thomas Bertrand[4], Utpala Bandy[4], Nicole Alexander-Scott[4], Casey Dunn[3], Joseph Hogan[1], Rami Kantor[1]

[1]Brown University, Providence, RI, USA; [2]Research Improving People's Life, Providence, RI, USA; [3]Yale University, New Haven, CT, USA; [4]Rhode Island Department of Health, Providence, RI, USA

BACKGROUND

Public health interventions guided by clustering of HIV molecular sequences may be impacted by choices of analytical approaches. In this study, we compare concordance of commonly-used analytical approaches, which vary in methods and models, in identifying molecular HIV-1 clusters. Determining concordance *within-* and *between-* analytical approaches can lead to better understanding of their strengths and limitations, and ultimately may be used to improve public health efforts to prevent and disrupt HIV transmissions.

METHODS

We identified commonly-used clustering analytical approaches and applied them to 1,886 HIV-1 Rhode Island sequences from 2004–2018. We compared number and proportion of clustered sequences *within* approaches across topological supports and distance thresholds; analyzed proportionate difference of clustered sequences *between* methods; and estimated methods' concordance by comparing proportions of clustered sequence-pairs, identical clusters, and non-clustered sequences. We used strict (topological support ≥0.95; distance 0.015 substitutions/site) and relaxed (topological support 0.80–0.95; distance 0.030–0.045 substitutions/site) thresholds to reflect different public health and scientific scenarios.

RESULTS

Clustering differed by method, and *within* method, by threshold (Figure 1); 22–54% of sequences were in 146–234 clusters by different methods/thresholds (Table). Clustering depended more on distance thresholds than topological support thresholds. Proportions of clustered sequences *between* methods were in worse agreement, and variation was more pronounced, at stringent topological support and relaxed distances (Figure 2). Differences in proportions of clustered sequences between model-based methods were 21% for strict and 31% for relaxed thresholds. Multi-level clustering concordance analyses demonstrated some differences in proportion of sequence-pairs clustered (Figure 3) and of identical clusters (Figure 4) across analytical approaches, with RAxML-based analytical approaches having the highest (91%) mean summary concordance when strict thresholds are applied, and three (RAxML-, FastTree regular bootstrap- and IQ-Tree regular bootstrap-based) analytical approaches having the highest (86%) mean summary concordance when relaxed thresholds are applied.

Different analytical approaches can yield diverse HIV-1 clustering outcomes, and may need to be differentially used in diverse public health scenarios.

Recognizing the variability and limitations of commonly-used methods in identification of clusters is important for guiding clustering-triggered interventions in order to disrupt new transmissions and end the HIV epidemic.

Table: HIV-1 clusters identified by seven commonly-used analytical approaches according to strict and relaxed sets of topological support and distance thresholds

Methods	Strict thresholds: Topological support	Strict thresholds: Mean TN93 p/w distances	Strict thresholds: # of clusters (# of sequences in clusters; %)	Relaxed thresholds: Topological support	Relaxed thresholds: Mean TN93 p/w distances	Relaxed thresholds: # of clusters (# of sequences in clusters; %)
RAxML	**0.95**; rapid btstr; 1000 replicates	**0.015**	**167** (500; 27%)	**0.80**; rapid btstr; 1000 replicates	**0·045**	**220** (847; 45%)
FastTree aLRT	**0.95**; aLRT	**0.015**	**163** (500; 27%)	**0.90**; aLRT	**0·030**	**212** (856; 45%)
FastTree bootstrap	**0.95**; regular btstr; 1000 replicates	**0.015**	**146** (451; 24%)	**0.80**; regular btstr; 1000 replicates	**0·045**	**201** (772; 41%)
PhyML aLRT	**0.95**; aLRT	**0.015**	**162** (496; 26%)	**0.90**; aLRT	**0·045**	**234** (1019; 54%)
MEGA	MCL; **0.95**; regular btstr; 1000 replicates	**0.015**	**156** (411; 22%)	MCL; **0.80**; regular btstr; 1000 replicates	**0·045**	**223** (712; 38%)
IQ-Tree ufast	**1.0**; ufast btstr; 1000 replicates	**0.015**	**187** (573; 30%)	**0.95**; ufast btstr; 1000 replicates	**0·030**	**231** (913; 48%)
IQ-Tree regular	**0.95**; regular btstr; 100 replicates	**0·015**	**146** (439; 23%)	**0.80**; regular btstr; 100 replicates	**0·045**	**198** (758; 40%)

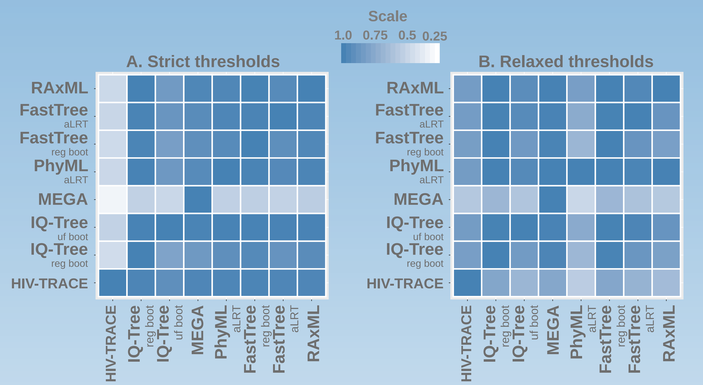


Figure 3: Concordance of HIV-1 clustering: Proportion of sequence pairs clustered by method-pairs.

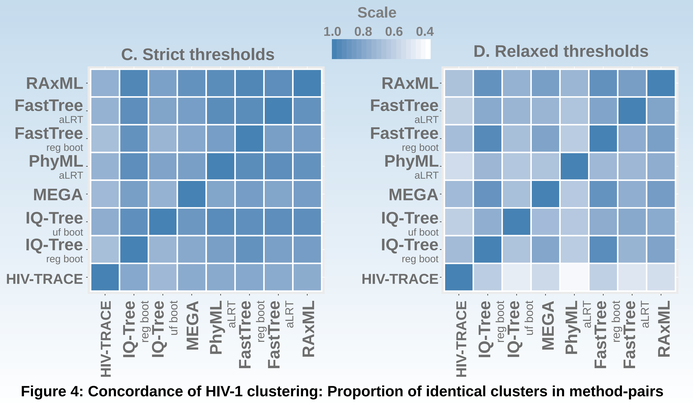


Figure 4: Concordance of HIV-1 clustering: Proportion of identical clusters in method-pairs

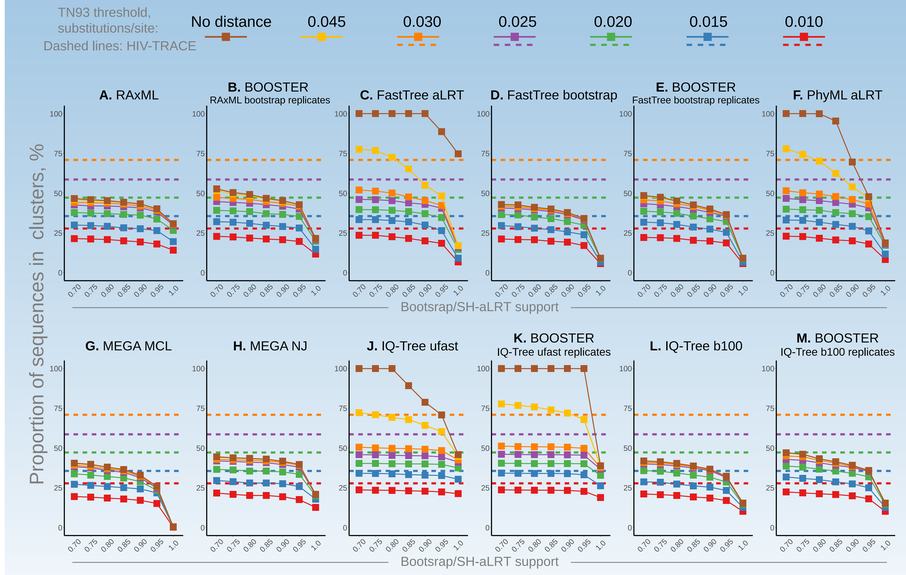


Figure 1: Comparison of proportion of HIV-1 sequences in clusters within commonly-used analytical approaches. Graphs A-M each represent the 12 model-based methods/variations examined. Solid lines in each graph represent the range of proportions of clustered sequences (Y axis) according to topological support (X axis) and distance (colored squares (legend at the top of the Figure), matching the line colors) thresholds. Color-matching dashed lines in each graph represent the range of proportions of clustered sequences identified by HIV-TRACE according to five different thresholds.

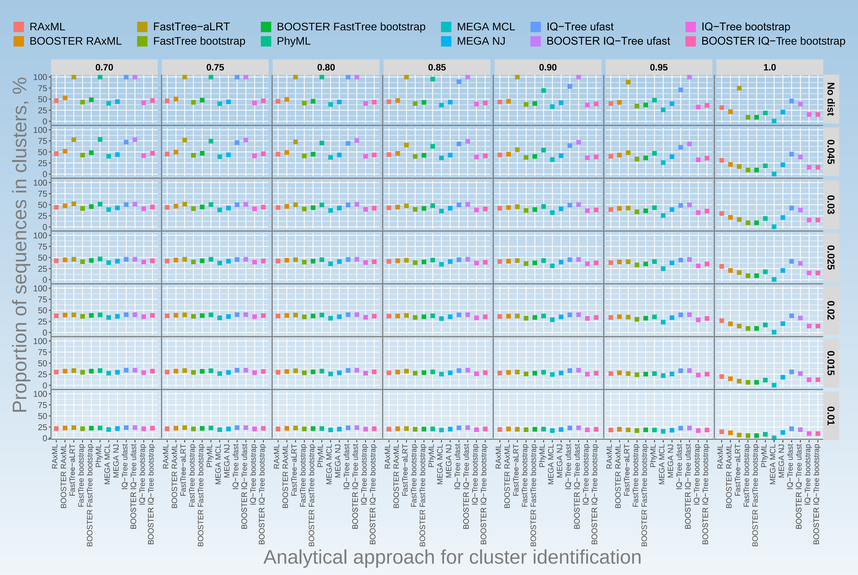


Figure 2: Comparison of proportion of HIV-1 sequences in clusters between commonly-used analytical approaches. Each of the 49 panels demonstrates proportions of HIV sequences in clusters (Y axis) identified by the 12 selected methods (X axis; also represented by colors and outlined in the legend above the panels), representing a distinct combination of topological support (outlined in the gray line above the panels) and distance thresholds (outlined in the gray line to the right of the panels).