MAPPING GROWTH OF LARGE TRANSMISSION NETWORKS USING DIFFERENT CLUSTERING ALGORITHMS

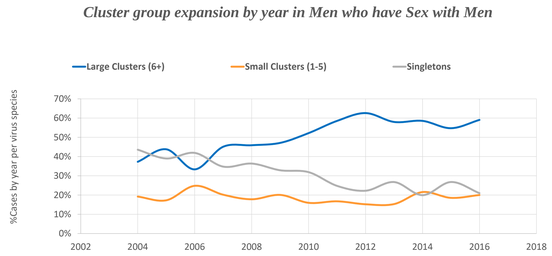
Nathan Osman[1,3], Antoine Chaillon[2], Ernesto Cuadra-Foy[1,3], Ruxandra-Ilinca Ibanescu[3], Isabelle Hardy[4], Irene Vrbik[5], Bonnie Spira[3], Luc Villandré[1], David Stephens[1], Michel Roger[4], Bluma G. Brenner[1,3]

[1]McGill University, Montreal, QC, Canada, [2]University of California San Diego, La Jolla, CA, USA, [3]Lady Davis Institute for Medical Research, Montreal, QC, Canada, [4]Centre Hospitalier de l'Université de Montréal, Montreal, QC, Canada, [5]University of British Columbia, Vancouver, BC, Canada

Introduction

Concept of phylogenetic surveillance: Transmission bottlenecks leads to single monophyletic founder viral strain establishing most HIV-1 infections. Phylogenetics is a molecular strategy that explores viral evolution, diversity, transmissions and characterizes epidemics based on the genetic interrelatedness of founder HIV-1 viral sequences. In the context of HIV-1, this analysis takes advantage of the high error rate and rapid pace of HIV replication, which leads to predictable evolution over long periods of time, or a "molecular clock". Numerous studies have demonstrated that HIV transmission networks can be inferred using standard HIV *pol* sequence data from drug resistance testing across the conserved protease, reverse transcriptase, and integrase regions targeted by antiretroviral therapy.

A.



B.

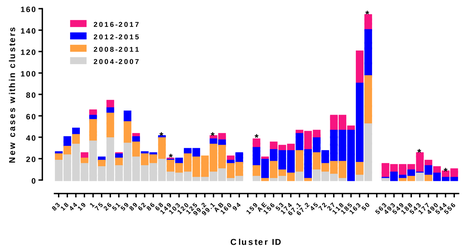


Fig. 1: Monitoring of the growth of clusters allows to identify transmission networks

A. The epidemic in MSM is fueled by a shift towards large cluster outbreaks (20-149 members/cluster)

B. Time trends in the genesis, expansion and persistence of 30 large cluster micro-epidemics over the 2004-2007, 2008-2011, and 2012-2015 quadrennial periods. Nine new "active" clusters (13-19 members) have recently emerged (2016 to August 2018).

Materials and methods

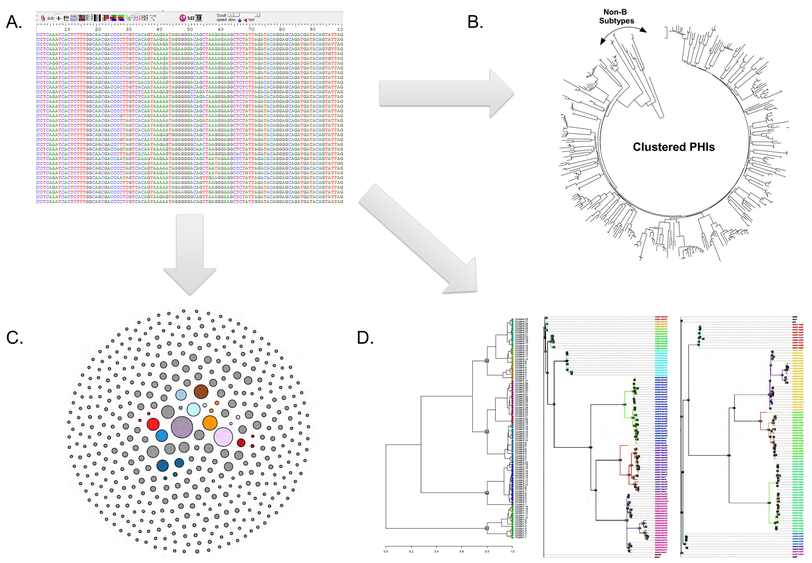


Fig. 2: Overview of the different methodologies compared

A. HIV-1 sequences of the *pol* gene (PR and RT) were obtained from the Quebec HIV genotyping program database.[3]

B. Unique non-nominative patient identifiers were assigned based on putative cluster group association, ascertained by Maximum likelihood (ML) methodology (high bootstrap support >97% and short genetic distance <0.015).[2]

C. The San Diego-based HIV-TRACE (Transmission Cluster Engine) software[4] (non-phylogenetic) was used to infer clusters based on genetic distance under pre-set thresholds.

D. The Montreal-based Gap Procedure[5] (non-phylogenetic), MrBayes[6] (Bayesian phylogenetic), and DM-PhyClus[7] (Bayesian) methodologies were used and compared to build chain estimates.

Results

There is no consensus on how to best define clusters and link putative transmission chains. Viral sequences within a cluster show more similarity to each other than remaining sequences outside of the group. The partitioning schemes, algorithms, cut-offs and thresholds to determine viral interrelatedness are highly subjective.

1) HIV-TRACE software assemble pairs of sequences across pre-set thresholds.

2) Neighbor-joining/ Maximum Likelihood (ML) methods convert pairwise distances into a tree (phylogeny) as descendants from a common ancestor. Maximum Likelihood methods jointly evaluate all genetic sequences instead of two sequences at a time, providing genetic-distance estimates and bootstrap support. The analysis is far slower to compute but the improved accuracy may be more informative.

3) Bayesian methods employ time-stamped sequences to establish transmission chains with posterior probability support and branch length distributions,

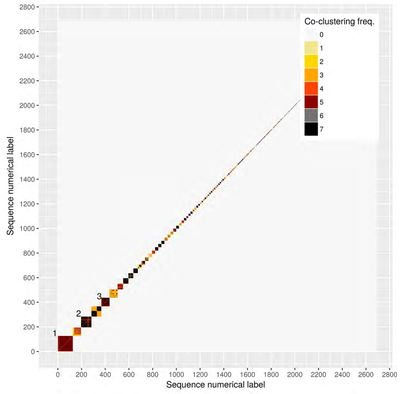
4) DM-PhyClus is a Bayesian phylogenetic algorithm that estimates transmission chains by identifying disjoint sets of sequences supported by subtrees with distinctive branch length distributions. It therefore avoids patristic distance and clade support.

5) The Gap Procedure is a pure distance-based partitioning algorithm that avoids reliance on ad hoc cutpoints by separating sets of sequences into distinctive components without requiring phylogenetic estimation.

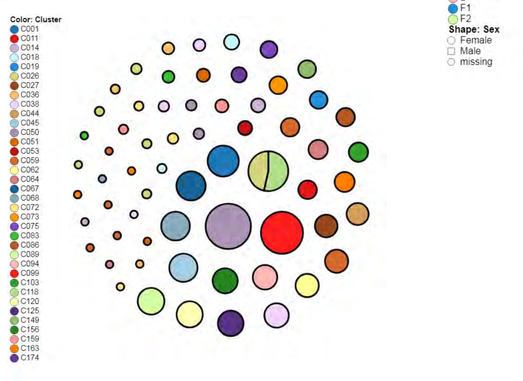
Our analyses revealed considerable overlap in defining thirty 20+ transmission chains.

DM-PhysClus was able to partition cluster 99 into two waves, where all members in the first wave harbored K103N drug resistance (>99% quasispecies) while members in the second waves were wild-type viruses.

A.



B.





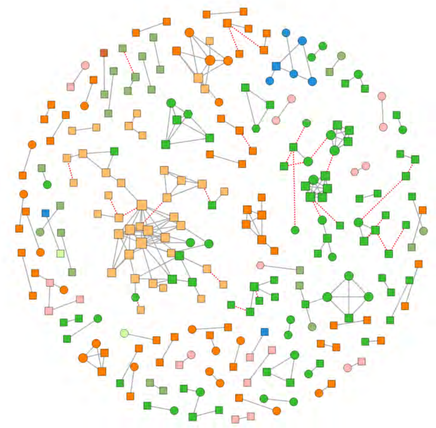


Fig. 3: Co-clustering of sequences into the same transmission chains by all the different methods used

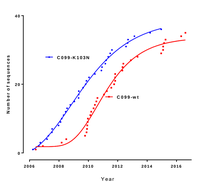
A. Heat map indicating overlap between estimates produced by conventional ML (RAxML), conventional Bayes (MrBayes), Gap Procedure and DM-PhyClus.

B. Clusters obtained with HIV-TRACE overlapping with the reference clusters identified by Brenner et al.[2]

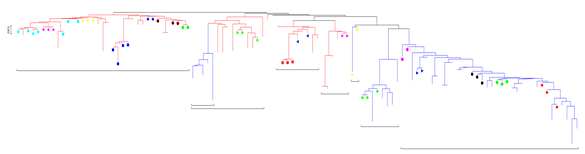
Fig. 5: Co-clustering of non-B subtype sequences into the same transmission chains by HIV-TRACE

Non-B subtype sequences were not properly segregated into distinct clusters regardless of the different parameters tested in the software

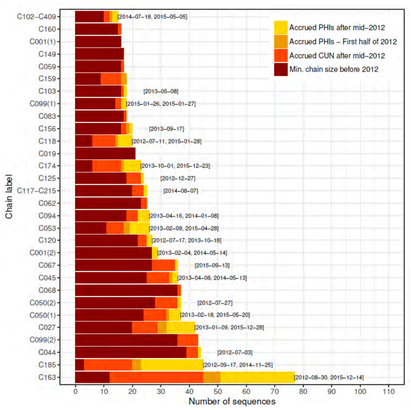
A.



B.



C.



D.

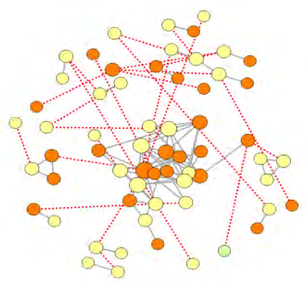


Fig. 4: Segregation of cluster C099 by DM-PhyClus and not by HIV-TRACE

A. Growth of C099 in two distinct waves: K103N (blue, n=34) and WT (red, n=34)

B. Reference ML tree from Brenner et al. separating K103N and WT species

C. Co-clustering of sequences by DM-PhyClus with the reference clusters, splitting C099 in two sub clusters

D. Links between K103N (orange) and WT (yellow) variants grouped in the same clusters by HIV-TRACE

Conclusion

In this study, we compared the sensitivity and accuracy of different phylogenetic and non-phylogenetic-based methodologies in estimating transmission linkage and mapping epidemic growth in close to real-time.

Maximum Likelihood-bootstrap methods build very informative trees but take a lot of time and computing power. HIV-TRACE, the Gap Procedure, and DM-PhyClus methodologies are quick and more suitable for scalibity. Our recent study suggests that HIV-TRACE may not have enough specificity and sensitivity to reliably establish linkages between individual nodes (infections) within some clusters. All estimates stress that longer older chains are responsible for a sizeable proportion of sampled incident infections among MSMs. Curbing the HIV epidemic will require strategies aimed specifically at preventing such growth.

References and acknowledgements

1. WHO HIV/AIDS (http://www.who.int/gho/hiv/en/)
2. Brenner et al., *AIDS* 2017; 31(5):707-717. doi: 10.1097/QAD.0000000000001383
3. Brenner et al., *J Infect Dis*. 2007;195(7):951-959. doi: 10.1086/512088
4. Kosakovsky Pond et al., *Mol Biol Evol*. 2018 Jul 1;35(7):1812-1819. doi: 10.1093/molbev/msy016.
5. Vrbik et al., *BMC Bioinformatics*. 2015; 16: 355. doi: 10.1186/s12859-015-0791-x
6. Ronquist et al., *Syst Biol*. 2012;61(3):539{542. doi:10.1093/sysbio/sys029
7. Villandré et al., *BMC Bioinformatics*. 2018 Sep 14;19(1):324. doi: 10.1186/s12859-018-2347-3

This work is supported by the Canadian Institutes of Health Research and Réseau SIDA du Fonds de Recherche du Québec – Santé.

I do not have any conflict of interest.

Contact: nathan.osman@mail.mcgill.ca
+1-514-340-8222 ext. 25283 or 23454