# An outcome-based statistical framework to select and optimize molecular clustering methods for infectious diseases.

Connor Chato[1] & Art FY Poon[1,2,3]

1. Pathology & Laboratory Medicine, Schulich School of Medicine & Dentistry, Western University, London, ON, Canada
2. Microbiology and Immunology, Schulich SchoSchulich School of Medicine & Dentistry, Western University, London, ON, Canada
3. Department of Applied Mathematics, Western University, London, ON, Canada

A cluster of genetically similar infections found in different patients imply an elevated rate of transmission in a subpopulation. For diseases such as HIV, where time-space clustering methods are inaccurate due to the time-lag between transmission and diagnosis, genetic clusters can indicate outbreaks. However, these often rely on an arbitrary similarity cutoff to draw connections, which dramatically changes the clusters created (**Figure 1)**. We validated cluster growth models on real data, while modulating thresholds for clustering and time-scales for predictors. Optimal parameters are then those providing the best improvement of fit provided by additional variables.

## Data Set

808 patient-matched HIV-1 B *pol* sequences Collected between 2007-2013 in Northern Alberta, Canada.

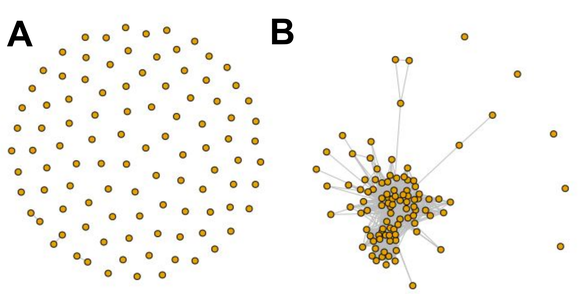


**Figure 1.** Clusters defined as components of a network. Vertices represent a random subset (n=100) of equences from the data set. Edges represent connections, where TN93 distances between sequences fall under different thresholds of 0.005 (A) and 0.040 (B) expected substitutions per site. .

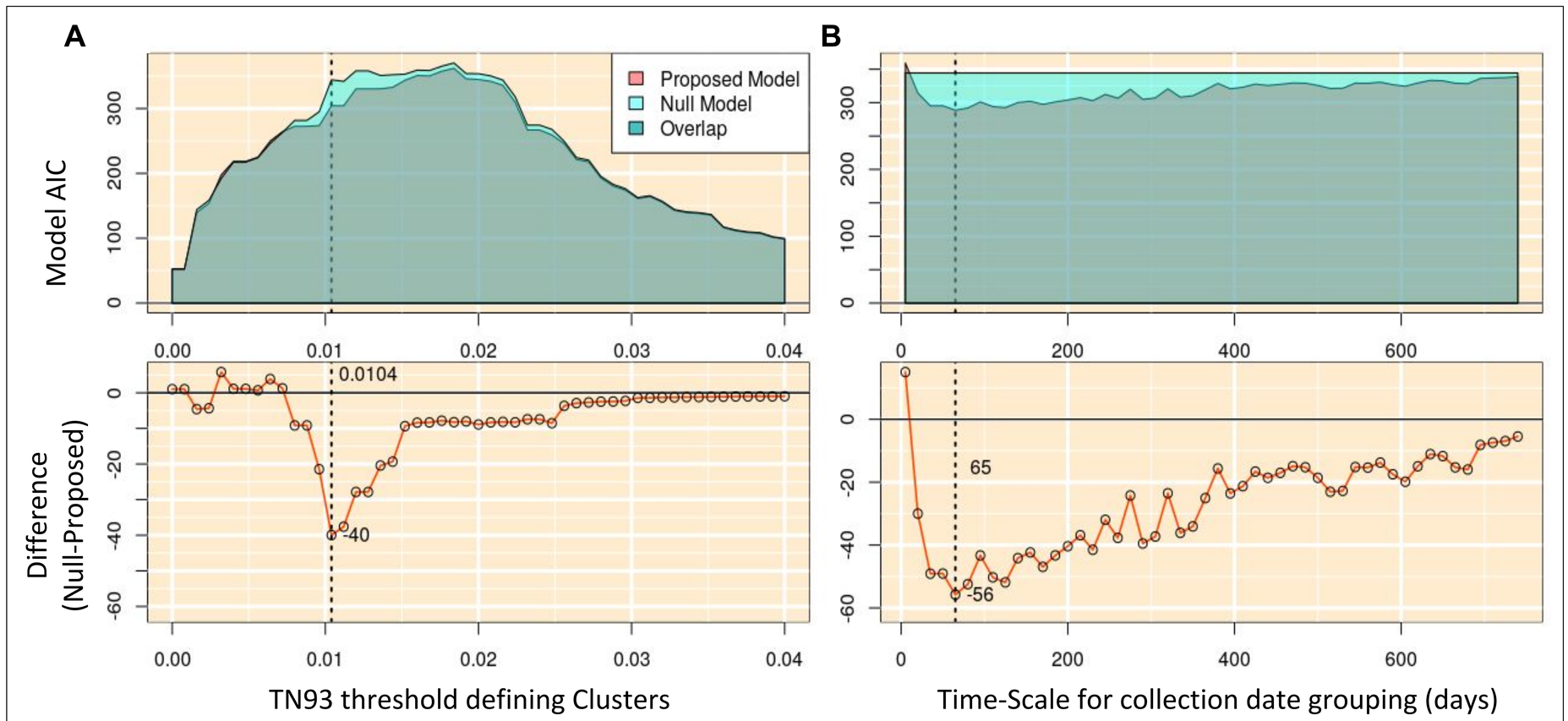


**Figure 2.** Akaike Information Criterion (AIC) for two Poisson-linked, predictive models. Higher AIC, indicates worse fit, and light blue regions indicate a higher AIC from the null model compared to the Proposed model. Fits were obtained using different TN93 cutoff thresholds to define clusters (**A**) and different time scales (ie. Weeks, Months, Years) for sequence collection time (**B**). While varying time scales, cluster threshold was static at the optimal 0.0104 obtained in (**A**) for consistency.

## Poisson-Linked Cluster Growth Models

- Predicting the number of "new" (collected in 2013) sequences that connect to a cluster of "old" (collected before 2013) sequences

***Null Model***

*Growth is predicted by cluster size alone.*

***Proposed Model***

*Growth is predicted by weighted cluster size. Recently collected sequences are weighted higher. This is effected by time-scale.*

## Conclusions

- AIC difference quantifies the benefits of new information (ex. sequence collection date) for a predictive model (**Figure 2**).
- Using a TN93 Cutoff threshold of 0.0104 for cluster definition and grouping sequence collection dates into 65 day time-blocks to train our proposed model, maximizes AIC difference. We view these as optimal parameters. for the Northern Alberta study area

# *Further Details*

→ Conflict of interest disclosure: I have no conflict of interest

→ For correspondence: ***ConnorChato@gmail.com***

→ More detailed discussion of this method covered in associated publication.

Chato, C. J., & Poon, A. F. (2019). An Application of the Modifiable Areal Unit Problem: Optimizing Cluster Method Parameters to Produce Predictive Data for HIV Outbreaks.

→ Associated publication for Northern Alberta HIV-1 Data set (Genbank Popset # 1033910942)

Vrancken, B., Adachi, D., Benedet, M., Singh, A., Read, R., Shafran, S., ... & Charlton, C. L. (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, 52, 100-105.

***Tools Used in Figure Creation and analysis***


- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

- Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. InterJournal, complex systems, 1695(5), 1-9.

- Tamura, K., & Nei, M. (1993). 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.

- Nakaya, T. (2000). An information statistical approach to the modifiable areal unit problem in incidence rate maps. Environment and Planning A, 32(1), 91-109.