7-31-2020 12:00 PM

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

Connor Chato, *The University of Western Ontario*

Supervisor: Art Poon, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Pathology and Laboratory Medicine

Abstract

A molecular cluster is a set of highly similar genetic sequences from pathogens. If each of these sample pathogens are infecting a different host, it can imply rapid spread between hosts. In practice, these clusters are often qualified by a genetic similarity threshold (ie. less than 0.015 expected substitutions per site between sequences indicates ”high” similarity). This thesis demonstrates an information-based approach to threshold selection based on the performance of models predicting cluster growth. Optimal thresholds maximize the loss of Akaike's information criterion (which measures inaccuracy and complexity) associated with predictive variables. Three sets of North American HIV-1 sequences and two different popular clustering methods were used to demonstrate this framework, using recency of sequence collection and patient diagnosis as predictive variables for future clustering. This addresses the issue of arbitrary, unspecified threshold selection for molecular clustering, showing different optimal thresholds depending on the source data and clustering algorithm.

Keywords: Bioinformatics, infectious disease, HIV and AIDs

Lay Summary

If a fast-evolving virus has little time to mutate in one host before being transmitted to the next, the result is that many hosts share genetically similar viruses. This can be evidence of an outbreak, and such evidence is vital for public health authorities, especially as similar viral sequences are collected from new patients (indicating a growth of the outbreak). Such methods have been particularly well-used for Human Immunodeficiency Virus (HIV), the causative agent for AIDs. However, it is difficult to establish how genetically similar these viruses need to be before a group of cases is labelled an outbreak and arbitrary thresholds of similarity are often used for this task. A poor choice for this threshold can lead to overestimation or the underestimation of the outbreak. Furthermore, this may make the predictive models which estimate how the outbreak will grow ineffective. This work shows a statistical method which chooses such a threshold based on how accurately it will predict the growth of outbreaks. Three different data sets of HIV genetic sequences are used as an example, each of which were collected from North America. We used two different examples of these sequence-based outbreak detection methods and found that the ideal threshold of similarity for predicting outbreak growth differs between location and method.

Acknowledgements

First and foremost, I would like to thank Dr. Art Poon for the exceptional level of support and guidance he has committed to this work. I hope to follow his example of bold, care-driven research. I'd also like to thank the other members of the Poon Lab, as well as the students, teachers and administrators of the Western Pathology and Laboratory medicine program - all have been excellent friends and colleagues to learn from/with. To my advisory committee, Dr. Jessica Prodger and Dr. Michael Silverman, I also extend my thanks for your vital feedback throughout this work, which has kept me grounded in reality instead of lost in my code. I would also like to acknowledge Josie the dog and Boo the cat, who have both been at my side during the majority of this writing process. Although they have offered me little in the way of advice or feedback, I am thankful for their general support. Finally, I would like thank my wonderful parents, who have supported my lengthy voyage through academia. Thank you dad for sharing your love of science with me and thank you mom for showing me how important it is to be a teacher and a learner.

Contents

List of Figures

Chapter 1

Molecular Clustering in Epidemiology

1.1 Introduction

A common goal of epidemiological analysis is to group a set of patients into ”clusters” based on a particular feature. Each cluster is a partition of the data set with reduced variation in that feature, implying important similarities between patients. These can also be thought of on a pairwise level, with connections drawn between pairs that have similar characteristics. Observing large clusters can reveal areas in need of public health resource priority and identify a common or novel source of infection. This has seen use historically, starting with an essay by John Snow [Sno55] linking an 1854 outbreak of cholera to a specific water pump in Soho, London. When compared to the more regular distribution of cholera throughout the city, the unusually high prevalence of 616 recent cases in the Broad street pump service area qualified as a cluster, sparking investigation by the London board of health and resulting in lasting changes to London's wastewater system. To build such clusters, a *proximity measure* is required to interpret similarity between cases and a *clustering criterion* is required to assign patients to a cluster [HBV01]. For example, a set of clusters established based on temporal and spacial proximity would indicate several cases of a particular disease or injury occurring in a relatively small area, over a short period of time. However, the location and time of infection are often difficult to obtain. Alternatively a set of genetically related pathogens in multiple patients can constitute a cluster, implying that an infectious agent spread through the population fast enough to accumulate few mutations. This proximity can be determined by phylogenetic techniques - the same genetic comparisons which allow for the construction of evolutionary trees.

The criteria for clustering becomes more of a theoretical problem. Similar to other groupings, molecular clustering is a binary characteristic (ie. a set of sequences either is or is not a cluster) based on a continuous measurement (genetic similarity). In practice, this often re-

sults in the use of a threshold in order to define clustering [BPP+19, WHVR+17, DVF+18], the selection of which has an effect on the degree to which observations aggregate into large clusters. For molecular clusters, the literature currently does not discuss the effect that threshold selection may have on the outcomes of clustering studies, despite the continuous use of standardized thresholds. In the following thesis, I will demonstrate how a change in threshold effects the performance of a predictive model which estimates the growth of known clusters over time. These demonstrations will use real sequence data published in previous studies, as well as my own implementations of popular clustering methods. The metric of performance will provide some basis by which an optimal threshold can be selected, defining ideal thresholds as those which result in the greatest gain in accuracy associated with predictive variables for a model predicting cluster growth. The differences in optimum threshold between locations and between clustering method are important, as these illustrate that the best results require tailoring of this parameter.

1.2 Requirements for molecular clustering

The creation of molecular clusters has several practical and technological requirements. First, the sequences used as a point of comparison must accumulate mutations relatively quickly, as some evolutionary divergence must occur within the time scale of a local epidemic (ie. months). Ideally, the differences in sequence data would become apparent between transmissions - viewing ongoing epidemics in real-time and guiding a public health response toward large clusters. This makes RNA viruses ideal candidates, as the RNA genome has been noted for its extremely fast mutation rate [HSH+82]. In particular, viruses within the *Retroviridae* family demonstrate an error prone replication cycle due to the low fidelity of reverse transcriptase [SCZ+03], a protein which synthesizes DNA sequences from viral RNA during replication. This requirement for fast mutation is unlikely to be met for species with a longer generation time such as parasites, but it can be met for some of the more slowly evolving viruses or bacteria by studying genetic differences on the scale of the whole genome [FFRF20, WIH+13]. Another requirement for molecular clustering is a large amount of available data. Fast, next generation sequence technology is the current method by which this data is obtained, where pathogen samples are taken from hosts and their whole genomes are assembled from fragments of genetic material [WBL+12, DBC+17]. The novelty of this sequencing power and the computational demands from even short-sequence comparisons, have limited large molecular clustering studies to occur only within the last few decades[Tom92]. However as computing

power increases and the collection of pathogen sequence data becomes a more routine part of diagnosis, we see a trend towards their use as a standard tool [GL18], as well as a trend towards large data-bases of sequence data, [SNH+12, FKL+18, SM17].

Molecular clustering is particularly useful for diseases that normally fail to create informative clusters with space or time criteria. For instance, when diagnostic date is likely to vary significantly from the actual date of infection due to a long asymptomatic period, time-based data is not necessarily informative for epidemiology. In addition, geographic location can be insufficient when trying to explain the pattern of transmissions for diseases with a lower transmission rate, as shared spaces alone may not be sufficient for transmission. Some studies need to rely on an overlay of complicated social networks to confirm feasible transmission patterns in this case [WPF+17]. The sexually transmitted Human Immunodeficiency Virus (HIV), which requires intimate contact for transmission, has a relatively low per-act transmission rate (<2 per 100 exposures [PBB+14]) and manifests with a long and variable asymptomatic period, making it an excellent candidate for this clustering approach. HIV also boasts an immensely high rate of mutation overall [CGG+15], which is partially based on its use of reverse transcription and RNA genome. This allows researchers to see measurable differences in the viral-genome between pair of patients months after a transmission between them. HIV is also a remarkably well studied species, with the full genome sequenced in 1985 [RHP+85], a standardized reference genome for comparison between studies, and a detailed understanding of gene function [WDG+09]. Fast, open source software [WBL+12] is available to screen for the presence of drug resistance in HIV, with the polycistronic *pol* gene, acting as a regular target [Kan06]. This gene can express mutations which confer resistance to highly active antiretroviral therapy (HAART) [DBK+10]. Because of this, large data sets of *pol* sequences are available and regularly obtained Genbank hosts numerous sets of published HIV sequence data sets, which represent cohort populations of over 1000 individuals each [BKML+11], in addition to the dedicated HIV sequence database hosted by the Los Alamos National Laboratory in the United states [FKL+18]. Not only does HIV meet the above criteria for molecular clustering candidates, but it also has no current effective vaccine or cure [MS16], making prevention a vital part of fighting the disease.

Although many molecular clustering studies focus on HIV for the reasons discussed above, there are other candidate pathogens for these techniques. Other RNA viruses which infect humans, such as *Flaviviridae* and *Coronaviridae* are major topics in the field of infectious disease, have widespread prevalence and meet the criterion of relatively fast evolution, when compared to other bacteria, or parasite-based diseases. The hepatitus C virus (HCV) (a member

of *Flaviviridae*) has been well studied through molecular clustering techniques [MDS+19b, MDS+19a, SDDA+12, JAK+14], paying special attention to injection drug use as a mode of transmission. Like HIV, HCV requires intimate contact, often manifests no symptoms, and currently has no available vaccine. Zika virus (another member of *Flaviviridae*), has shown the potential for molecular clustering studies, using hundreds of whole genome sequences to characterize the spread of the virus to new locations [LLH+16, GLK+17, ZMM+15]. Because the pattern insect-borne diseases like Zika can be deeply complex, conventional contact-tracing methods are not as useful as phylogenetic studies. A whole-genome based analysis has also been used to study the spread of coronaviruses, such as Middle East respiratory syndrome (MERS) in Saudi Arabia [CWK+13] and severe acute respiratory syndrome (SARS) in China [C+04]. In addition, the global spread of the SARS-like novel 2019 Coronivirus (SARS-CoV-2) [FFRF20] continues to be a major topic of study, as the amount of patient-matched sequence data expands [SM17]. Even non-viral species, such as *Mycobacterium tuberculosis* and *Vibrio cholerae* bacteria have been studied through genetic clusters despite their slower mutation rate [WIH+13, CSH+11].

1.3 Common molecular clustering methods

For the purposes of this work, it is important to discuss some molecular clustering methods in further detail. Both methods in the following description use aligned sequence data, where an algorithm has matched sequences by position. This means introducing gaps or blank characters to account for regions which may not exist in all sequences due to insertion mutations [Lar14]. Because bioinformatics is a fast moving field, it is important to at least note the alternative methods which may define new standards. For instance, it is becoming more common for researchers to meet the high computational demands of bayesian tree-building approaches, allowing accurate, time-scaled phylogenetic trees [YR97, RY96]. Also, parametric methods define the criterion for clustering based on a model, which avoids some of the difficulties involved in manual parameter selection [MP17, HPMSR19]. However, these novel methods are not as commonly used and have not dominated the literature to the same extent that the following approaches have.

1.3.1 Graph-Based Clustering

The ”network” - or more formally ”graph” structure visualizes data as a set of vertices (points on a plane) connected by edges (lines which connect the points). This is used to represent epi-

demiological relationships, with vertices representing individuals infected with the pathogen and edges representing some epidemiological relationship, such as direct or indirect transmission. These edges can be given a numerical "genetic distance", a simple genetic proximity measure which indicates the estimated number of point mutations (substitutions of one nucleotide base to another) that have occurred between sequences. Genetic distance is calculated in a similar manner to the Hamming distance [Ham50], where the number of mismatched characters are counted between two aligned sequences. Like the hamming distance, this is often reported relative to the length of the sequence, sometimes as a percentage of the sequences that differ. The modern standard of genetic distance measurement, the TN93 distance [TN93], takes into account that not all mismatches between sequences are equally likely to occur. For a pair of pathogen samples from two different patients, a relatively small genetic distance implies a genetic relatedness and a higher likelihood of close epidemiological relationship between the hosts: be that a short chain of transmissions, a common source of infection, or a direct transmission relationship.

Given a set of n sequences we can obtain $C(n, 2)$ genetic distance measurements between all possible pairs of sequences, where C is the *choose function*; the binomial coefficient of n and 2. The set of sequences and their associated distance measurements can be implemented as a complex, undirected graph with n vertices, each representing a sequence from an infected individual as described previously. The edges are then weighted with pairwise genetic distances and clusters are typically defined by imposing a maximum distance threshold for edges, highlighting only those which are considered highly similar [KPWLBW18, LLR+20, OFP+18, BPP+19, RLD+17, DVF+18]. The graph (Figure 1.1) is shown as an example from a subset of published HIV-1B *pol* sequences [WHVR+17]. All edges within this graph represent a pairwise TN93 distance under 0.02, and any individual vertices with no connections to any other have been removed.

This allows the use of a clustering definition described early, where clusters are simply a series of connected observations - in this case, a series of vertices (patients) connected by edges (infected with highly similar pathogens). Although this definition is used for the clustering of HIV, thresholds are not the only way to define clusters using a network. Other terms that could apply as clustering criterion could reference the order of vertices (ie. how many edges exist per vertex) or the average pairwise distance between sequences in a subset. Fast, open-source software is available to calculate tn93 distances [kPWV18] (https://github.com/veg/tn93) and to define pairwise graph-based clusters [KPWLBW18], making this method especially popular and accessible.

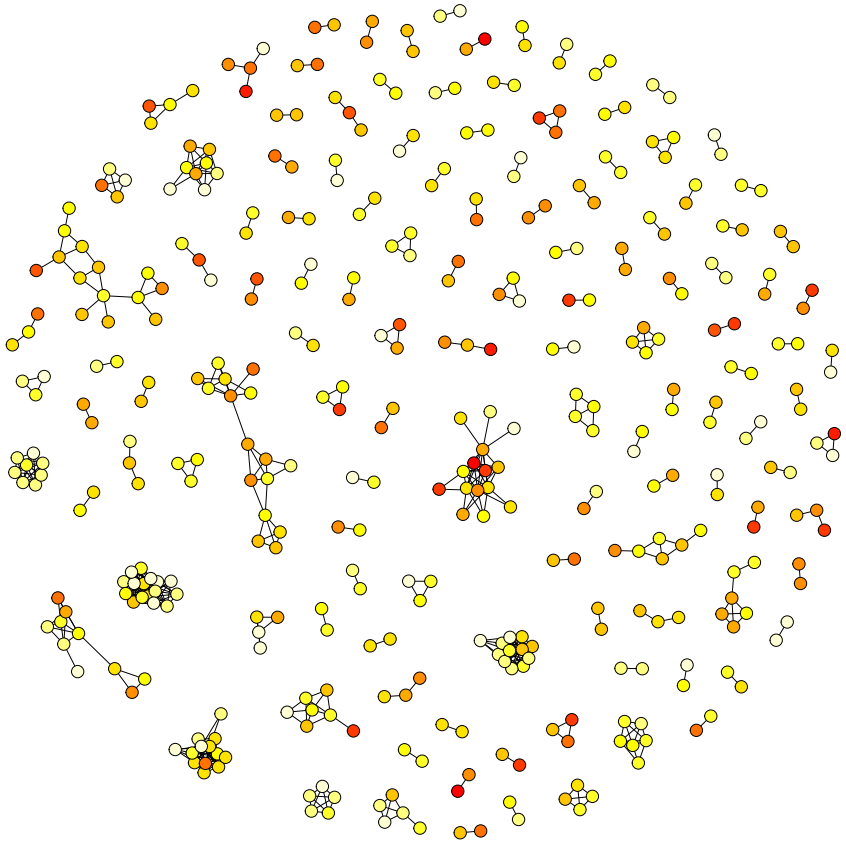

Figure 1.1: An example graph, representing pairwise graph-based clusters from 1200 HIV-1B *pol* sequences [WHVR+17]. The vertices are coloured based on how recently the corresponding sequences were collected, with darker red representing the most recently collected sequences.

1.3.2 Tree-Based Clustering

A bifurcating tree structure reflects the way we often conceptualize evolution; as individual taxa expanding from a single common ancestor in a series of branching events. For phylogenetic trees, the terminal ”tips” of the tree represent the sequences used to build it and the ”internal nodes” represent common ancestors; a diagram is shown below (Figure 1.2).

The horizontal length of branches represent relative amounts of divergence, meaning that two similar tips are likely to have shorter branches leading to the internal node which represents their common ancestor. Vertical lengths only exist for clarity and do not represent divergence. The total branch length traversed along the tree to get from one tip to another is termed the ”patristic distance” and often acts as a proximity measure, sharing some similari-

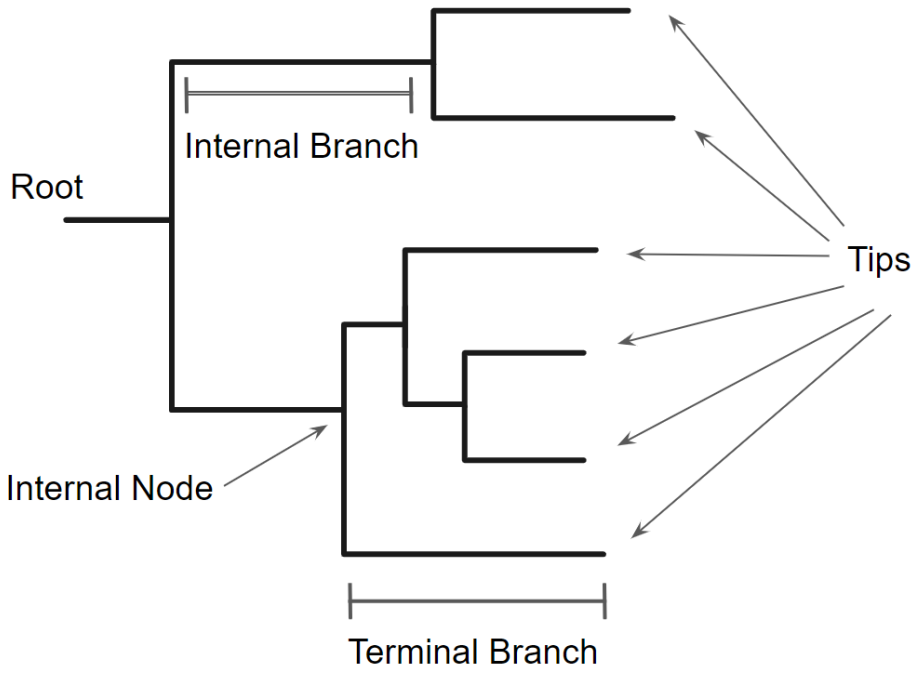

Figure 1.2: An example tree, labelled with terminology

ties with the pairwise genetic distance measurement described in section 1.3.1. However, in maximum likelihood trees, these branch lengths are not measured empirically from pairwise sequence comparisons, but instead estimated by a model-fitting process. Maximum likelihood trees, represent a proposed evolutionary timeline, and the likelihood of this proposition can be quantified through the use of an evolutionary model, taking in the relative length of branches and the order of branching events as parameters [LPD00]. For instance a tree that presents two very different sequences with a relatively short patristic distance and an immediate common ancestor is not a very likely representation of the actual genetic relationship of these sequences. Maximum likelihood algorithms [NSVHM15, Sta06] use heuristics to accomplish this, taking in aligned sequences and outputting the tree with the most likely branching order and branch lengths. The certainty of the node placements in a tree can be quantified through the process of ”bootstrapping”, which is traditionally a repeated rebuilding of the tree with different segments of the sequences sampled with replacement. The variation in the tree’s overall topology with each sample indicates the ancestral relationships that are more likely to vary. Bootstrap values for internal nodes in the final tree are then reported as the percentage of trees where that node was held at that particular place with respect to the descendant tips.

Tree-based clusters are often defined as closely related monophyletic clades: specifying a

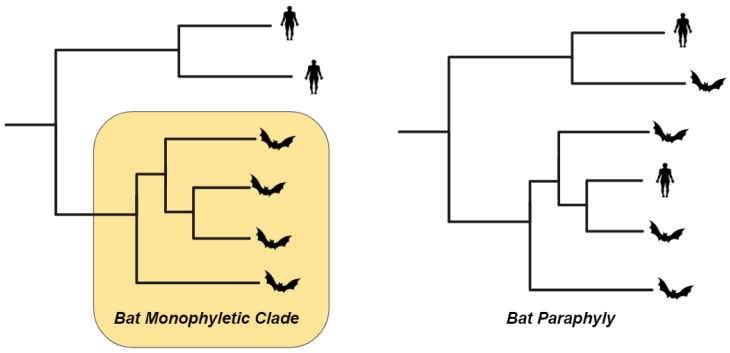

Figure 1.3: An example monophyly (left) compared to a paraphyly (right) within a phylogenetic tree.

set of tips which converge to a relatively close internal node with high bootstrap certainty. A monophyletic clade is exemplified in Figure 1.3, where it is contrasted with paraphyly. Similar to the genetic distance thresholds discussed in the previous method outline, the clustering criterion used to qualify close relatedness is often a constraint to a maximum patristic distance within a monophyletic clade, such that no two sequences in a cluster can have a patristic distance beyond that length [RCHH+13, VLVR+18, WHVR+17]. In addition, a minimum bootstrap certainty is also sometimes required, specifying that a set of sequences in a cluster must converge to the same ancestral node with certainty [DOKG+17, RCHH+13]. The tree in Figure 1.4 is shown as an example built from 20 published HIV-1 B *pol* sequences [WHVR+17] using IqTree software [NSVHM15]. Tips sharing a highly confident ancestor (bootstrap ≥ 75) and no patristic distances greater than 0.04 between them are highlighted in blue to represent clustering.

For a small sample of highly related sequences, all branch lengths are relatively short and bootstrap confidence is relatively low. A different expected degree of divergence would ultimately change the average patristic distance between sequences placed in the tree, as well as the certainty of their placement and branching order. The open source "Cluster-Picker" [RCHH+13] software package has been created to determine clusters using the clustering criteria mentioned above and is well used in the literature on HIV clustering [RLD+17, WHVR+17, DOKG+17]

1.4 The Goals of Molecular Clustering

In practice, there are two common goals of molecular clustering. The first is source attribution, which seeks to determine the potential vector by which a disease entered a new population.

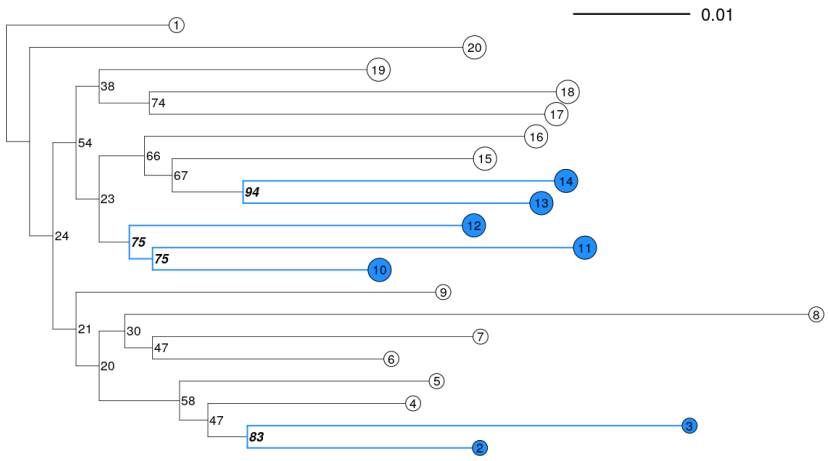

Figure 1.4: An example tree, built from 20 sequences using maximum likelihood methods (IqTree, default parameters). The clusters are highlighted in blue based on relatively confident relationships (Bootstrap $\geq$ 75) and relatively short terminal branch lengths between cases ($\leq$ 0.04). Relative tip size corresponds to terminal branch length, a scale bar is given in the top left to reference branch lengths, and branch lengths under 0.002 have been resolved to 0.002 for the purposes of clarity.

In this case, connections between individual sequences are taken to represent pairs who are relatively close to each other in the chain of transmissions defining the epidemic. For maximum likelihood trees, the direction can also be revealed through paraphyly [VF13], when sequences from multiple sources share a clade. Looking at the right panel of Figure 1.3 as an example, the human sequence is nested within the lower subtree of bat sequences. This would point to a disease which transferred from bats to humans, as it appears that the pathogens found in humans descended from a subset of those found in bats.

The second use of molecular clustering is ”outbreak detection”, which aims to identify heterogeneity in the rate of transmission. Large clusters, with very close proximity measures can often be labelled as outbreaks, as they indicate a pathogen which has been transmitted through multiple hosts in a short amount of time. Locating an ongoing outbreak can provide an opportunity for intervention, distributing any available treatment, vaccination, prophylaxis, or known preventative measures to the population associated with clusters. It also may identify an area with a higher number of unknown infections, guiding testing and diagnostic efforts. For HIV, Pre-Exposure Prophylaxis (PREP), is a preventative drug which protects individuals at high risk from attaining an infection [TKP$^+$12]. In addition, HIV treatment with HAART

[HCCP04, DBK+10] reduces the chance that an HIV positive individual will transmit the disease, so confirming a connection to care is of priority importance for individuals in clusters. For other transmissible diseases, this may indicate priority populations for available vaccines, ensuring that individuals at the highest risk of obtaining a new infection are protected. Training a predictive model to identify the indicators of future clustering is also useful in this context, allowing the prevention of outbreaks in real-time. The following sub-sections will discuss each method with further depth using example case studies.

1.4.1 Studies in Source Attribution

In cases of zoonotic infection (ie. diseases which are obtained from animals), source attribution can determine key animal vectors which act as points of entry into populations. For insect-borne diseases such as malaria, west Nile and Zika virus, this is the primary mode of transmission. For instance, the 2019 novel Coronavirus pandemic caused by the SARS-CoV-2 virus was initially attributed to bat species in eastern China through a clustering study by Lu, et al (2020) [LZL+20]. The SARS-CoV-2 genomes taken from 9 individuals who had been in contact with the Huanan seafood market in Wuhan China, were sorted within the *sarbecovirus* sub-genus, a subset of the *Betacoronavirus* genus. When a maximum likelihood tree was constructed, bat-borne SARS-CoV-2 sequences formed a paraphyly with the human SARS CoV-2 sequences, suggesting that the disease had transferred from bats to humans (Similarly to the configuration previously described in Figure 1.3). However, it is unclear how directly this disease is transferred and intermediate hosts could have existed. Further work has suggested that coronavirus sequences taken from pangolin species *Manis javanica* [ZWZ20] fall within this same cluster, suggesting that all 3 animals are likely able to host the same virus.

The origin of a major 2010 cholera outbreak discusses transmission directly between human hosts. Chin et al [CSH+11] obtained genomes from 5 isolates of cholera obtained from patients within this outbreak. These sequences were found to be more closely to cholera strains from Bangladesh than from the more local sequence data obtained in Peru, when comparing the number of substitutions between sequences. The results ultimately implied that the disease was most likely introduced by international aide forces stationed there in response to a major earthquake occurring in the same year. For both situations, illuminating the cause of infection provides extremely valuable information for public health agencies, as future testing efforts can be guided towards source populations to prevent a growing infected population and control the ways that an infection could enter a new population.

In the context of direct human-to-human transmission, it becomes important to consider

the controversies associated with criminalized diseases such as HIV, where epidemiological association can have legal consequences. Studies with the goal of identifying patient to patient transmission do exist within Canada and have seen recent use [MPL+20], however, there are practical challenges and ethical conflicts in the identification of direct transmission relationships. As implicit in any source attribution method with incomplete sampling, there is the potential for an unsampled individual to act as a bridge between an alleged donor and receiver pair - or for two infections to appear as a donor-reciever pair because they have been infected by the same unsampled source [RHR+19, Poo16, NML+14]. This creates doubt in the accuracy of transmission pair classification, as a connection may not exclusively represent direct transmission [RHR+19]. Furthermore, the potential for patient information to be requested by subpoena in a status non-disclosure case, has been shown to be counter productive to the goals of public health, decreasing the likelihood that infected individuals seek treatment or disclose their infection [SKS+07, PMO+15, Myk15]. Sequence data exists in ample amounts for HIV and large-scale source attribution studies are done with regularity in order to discuss the transmission dynamics of HIV on a larger, population level [DVF+18, LLR+20, HD03, San14]. These seek to identify if there is a particular subpopulation with a higher risk of onward transmissions. However, to remain sensitive to the problems of direct patient to patient transmission analysis, HIV sequence data from patients is anonymized as a standard practice and rarely published with any associated meta-data that could be traced back to a specific patient.

1.4.2 Studies in HIV Outbreak Detection

In the context of this study, an outbreak will describe a sub-population with an unusually high disease incidence. Identification of any outbreak often demands large data sets, with enough variation between observations to contextualize the significance of any incidence change. There is also an imperative to discover ongoing outbreaks, as the clustering of more recently diagnosed cases may indicate an opportunity for intervention. The standard surveillance of drug resistance mutations has been a major factor in allowing outbreak detection to be done on a large scale [SNH+12, FKL+18, SM17].

My work has been particularly focused on outbreak detection as it pertains to HIV, because compared to other uses of molecular clustering, this particular field has seen a remarkably well developed history and discourse, spanning three decades before the initiation of this study. The initial studies of HIV transmission using genetic sequences were criminal cases, with one of the most well known stories in North America being that of the ”Florida Dentist” [SW92], a criminal case in the United States where a dentist was charged with knowingly transmitting HIV to

multiple patients. Pathogen sequence data from patients and the defendant were admitted as evidence. The first HIV molecular clustering studies with public health goals occurred slightly afterwards [HZR+95, LEF+96], showcasing the simple idea that phylogenetic trees can mimic known transmission patterns of HIV. As it was discovered that the HIV-1 pol gene could confer drug resistance and the regular sequencing of patient's viral genome became common as a part of diagnosis, the available data sparked larger studies with implications for a whole infected population [Kan06, YVR+01]. The specific methods (HIV-Trace [KPWLBW18] and ClusterPicker [RCHH+13]) described earlier in this chapter were developed in the last decade, and the following case studies describe some examples of their use. Most recently, these methods are used to prioritize which clusters are the most likely to attain new cases, estimating where new cases are most likely to appear.

Once an outbreak is identified, the clusters which comprise it are often used to detect specific risk factors. For instance, a recent study Ragonnet-Cronin, et al (2018) [RCJBS+18] identified a relatively large pairwise graph-based cluster of 104 HIV-1 sequences from over 2000 patient-matched *pol* sequences Glasgow, Scotland using a TN93 distance threshold of 0.01 expected substitutions per site. The sequences identified in the cluster shared known drug resistance mutations E138A and V179E to the *pol* gene, which potentially conferred a higher transmission likelihood from individuals who were receiving treatment in the form of non-nucleoside reverse-transcriptase inhibitor (NNRTIs). In addition, of the individuals identified in this major cluster, 102 reported injection drug use, which, as a form of blood to blood contact, was also suggested to contribute to the unusually high rate of spread among members. The identification of this cluster and the analysis of potential causes allows for the support of specific public health tools based on the demographic. In this case, harm reduction programs such as needle exchanges and safe injection sites could be suggested as a tailored response. The genetics of the disease itself is also an important factor, understanding the particular subtype and likely drug resistance mutation could guide decisions in effective treatment regimes.

For large data sets which are updated over time, it is possible to correlate cluster growth with certain characteristics and identify key predictors of an outbreak. In this case, known clusters can be updated to include new cases based on proximity measures between known clusters and new cases. This essentially means training a predictive model to identify the frequency of a connection between a known patient and a new patient. Figure 1.5 shows an example of this, where a circled cluster represents a theoretical target for priority because of the likelihood to attain two new cases.

A study by Wertheim et al (2018) [WMM+18] implements this predictive method, observ-

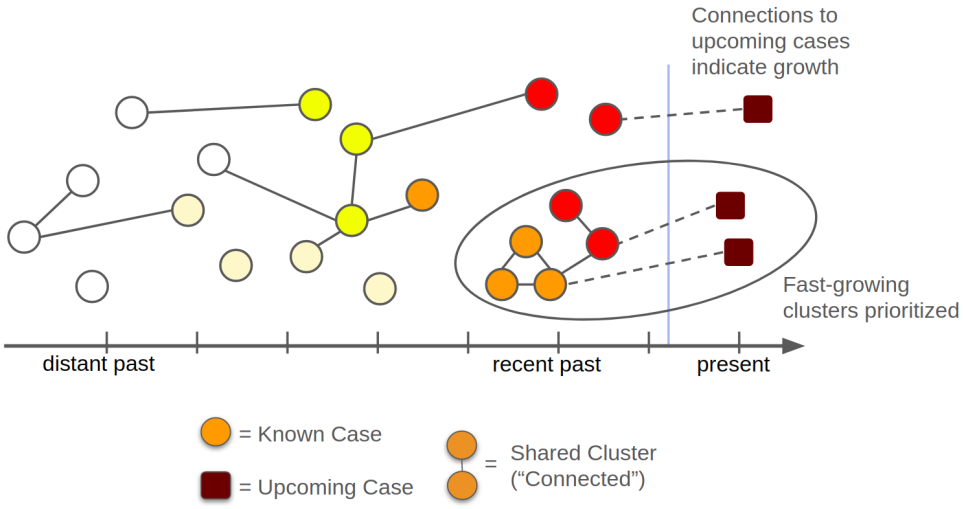

Figure 1.5: An illustration of how cluster growth may be interpreted over time. Older cases clustering with newer cases as an indication of onward transmission. The prediction of the connections which attach known cases to upcoming cases is prioritized, as it implicates a cluster with a high likelihood of significant growth. Such a cluster is circled in this figure. Darker red colour indicates a higher likelihood of onward transmission

ing the growth of graph-based clusters built from 65,736 HIV-1 B *pol* sequences in New York City using a TN93 distance threshold of 0.015 expected substitutions per site. The goal of the study was to train a predictive model to prioritize 500 known cases most likely to connect to cases which were newly incorporated into the data set. New cases are incorporated on an annual basis over several years. This effectively defines of binomial regression model of the following form

$$log\left(\frac{P}{1-P}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 ... + \beta_i x_i \tag{1.1}$$

where P represents the probability that a cluster will grow (ie. some case in that cluster will connect to a new case) as a function of some series of predictor variables x, with coefficients β and intercept α estimated by regression. It was shown that prioritizing whole clusters based on their size and recent growth yielded more accurate predictions than prioritizing individuals based on things like risk factor or past transmission history. This demonstrates a key part of what makes clusters appealing for predictive modelling, that the large partitions smooth out the stochasticity associated with individual cases. In a more recent study by Billock et al (2019) [BPP+19], clusters were built from 8,202 HIV-1 B *pol* sequences corresponding to individuals

diagnosed with HIV in North Carolina, USA from 2015-2017 using a similar methodology. Clusters with an average diagnostic date that falls within a year of growth observation period, clusters with any proportion of cases displaying high virus populations in the blood, or clusters where over 50% of the members had no named contacts were all at a higher risk of displaying growth. Potentially confounding variables were taken into account, in this calculation. Interestingly, both of these studies treated cluster growth as a binary outcome, instead of investigating the magnitude of cluster growth (number of new cases) as a Poisson distributed outcome. A paper by Le Vu et al, [LVRD+18] describes this an important future goal to allow for more informed prioritization of clusters given the larger (in some cases international [WLBH+14]) scale of HIV outbreak detection studies In this alternate case, the equation would follow

$$E[y] = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 ... + \beta_i x_i) \tag{1.2}$$

where $E[y]$ represents the number of cases which join a particular cluster as a function of some series of predictor variables x, with coefficients β and intercept α estimated by regression.

Despite observing populations which likely show different sample coverage, time range and study area size, both studies described above use the same 0.015 threshold criterion to build clusters. This threshold was initially based on the distribution of expected genetic distance between any two *pol* sequences in the United States given a national level [Kan06, APP+12], however, it has become a widely used standard at the municipal and state-wide level in North America. This same threshold has also seen use in different continents [LLR+20, VLVR+18, RCOAM+10, DOKG+17], despite the potential for differences in population densities, modes of transmission and sampling efforts. In addition, these thresholds have been treated as equivalents across multiple clustering methods; both the tree-based clustering methods and graph-based clustering methods described in this chapter. What a connection indicates changes in response to the threshold chosen, and by extension, so does the interpretation of a molecular cluster [LVRD+18, RLD+17] In further chapters, I will discuss the benefit of tailoring these often standardized threshold criteria to the area of study and the method used.

Chapter 2

Applications of the modifiable areal unit problem

While the previous chapter described how these molecular clustering methods are used in practice, this chapter will discuss the statistical problem created by the selection of a threshold. Although this trade-off is not formalized as a problem in the molecular clustering literature for HIV, the degree to which observations should be aggregated is a well discussed topic in other fields. In particular, there is a well defined problem described as the modifiable areal unit problem, which addresses the potential for different spatial partitions of the same data to change outcomes. If the problem of threshold selection is analogous to the Modifiable areal unit problem, there may be applicable solutions that can be borrowed for this cause, however, because genetic data often exists in a dimensionless space, the known solutions to the modifiable areal unit problem must be adapted for use on genetic data. This chapter also aims to discuss how the data sets are expected respond to a change in threshold under the two popular clustering methods described previously (graph-based clustering methods such as HIV-Trace [KPWLBW18] and Tree-Based clustering methods such as Cluster Picker [RCHH$^+$13]). These response are important to analyze, as they define the potential costs or benefits associated with threshold changes. These responses may also act as the key points of difference between two clustering methods, potentially resulting in different optimal thresholds from one clustering method to another..

2.1 Variance-bias trade offs and the modifiable areal unit problem

As described in the previous chapter, the partitioning of a data set into discrete clusters can offer important information in an epidemiological context. However, for the clustering methods described in the previous chapter, the selection of a threshold may introduce a trade off between bias and variance. Higher, more relaxed clustering thresholds allow more connections to exist, which leads to a greater proportion of larger clusters, but too much connectivity may begin to make those connections less meaningful. This also means that a set of clusters with high variance (ie. many clusters with significant differences between them) also tend to result in high bias (ie. smaller clusters which do not representative of consistent trends in the whole population). Although some molecular clustering studies evaluate the effects of threshold selection by using multiple thresholds to define clusters [RLD+17, VLVR+18, OFP+18, VAB+17], the relationship between the number of clusters and the outcomes of molecular cluster analysis is not well characterized, especially as they apply to predictive models of cluster growth. This is a well discussed issue in the field of machine learning, where the relationship between dependant and independent variables is not initially specified with a formula [NMB+18, LSG11]. In this case, the variance-bias trade-off must be addressed in order to extrapolate beyond the initial data set. If a predictive model aims to predict outcomes for a set of partitions, but the entire data set falls into a single partition then only one prediction and one outcome would be generated. This is known as ”undertraining”, which results in limited information for the effect of predictors on outcomes. As an epidemiological example, this would be similar to taking the average age for a whole population as a single predictor and counting the total number of deaths by heart disease as a single outcome. If each individual in the population were treated as an observation with its own predictor (age) and its own outcome (death by heart disease), then a predictive model which only aims to minimize error would become biased, treating exceptional cases as part of the rule. A classic example of overtraining and undertraining a predictive model is visualized in Figure 2.1, using an R script and simulated data. This represents the consequences of favouring each of the two extremes in the variance-bias trade off discussed above, where neither a single global mean (blue) or the exact values associated with each observation (red) represent a pattern which is useful for future analysis. For this example, the true nature of this relationship is linear, but in an unsupervised machine-learning context, the nature of the relationships between predictors and outcomes are usually assumed to be complex and beyond definition by a simple, known model.

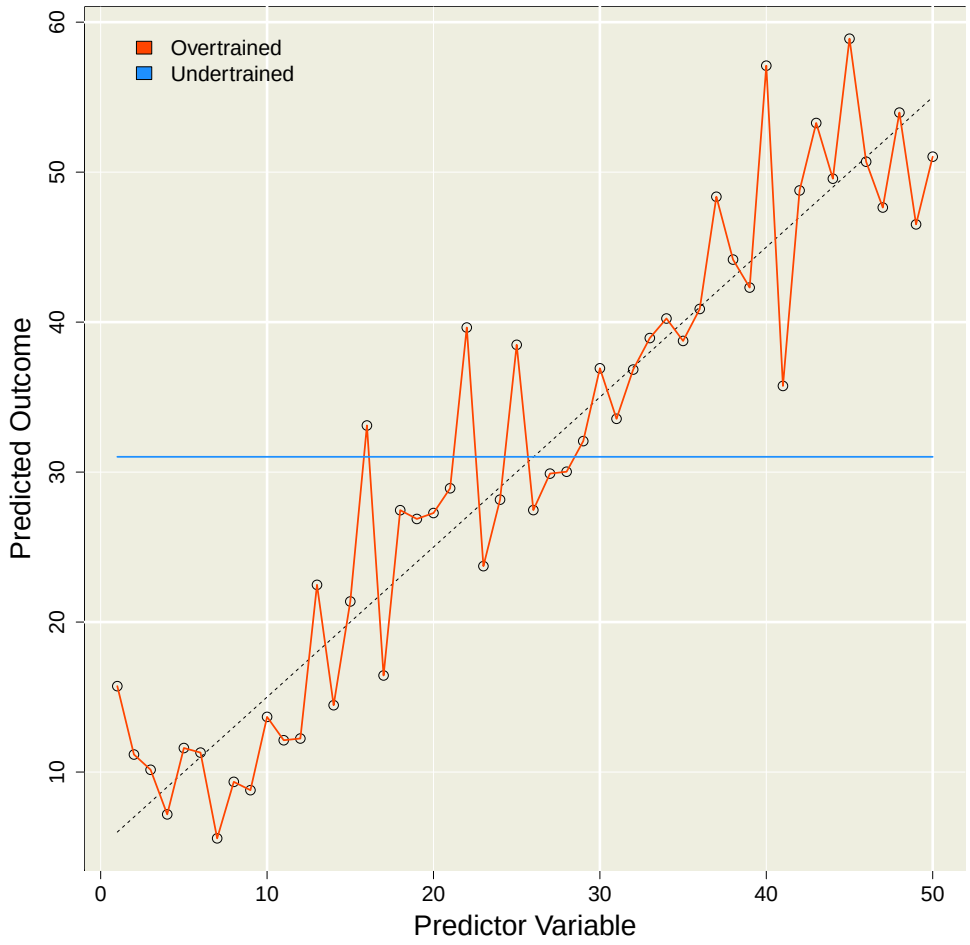

Figure 2.1: A visual example of overtraining and undertraining a predictive model. The predicted values of single mean of the complete data set (blue), as well as a line which goes through the values of each point individually (red) each contrast the actual relationship (black) between predictor and outcome.

Although variance-bias trade offs appear in multiple fields [Nak00, FW91, ZLZ16], the partitioning of a data set into k clusters is comparable to a specific type of variance-bias trade-off problem: termed the modifiable areal unit problem (MAUP) by Openshaw and Taylor (1979) [Ope77]. The MAUP describes the trade off inherently involved in the partitioning of an area for a geographic study. Each resulting partition or "spatial unit" contains a set of observations, similar to the clusters which are defined based on spatial or genetic proximity measures. There are two aspects to the MAUP. The first is the "scaling problem", which is implicit when selecting a scale for hierarchical data. For instance, spatial units could be defined as nations, provinces or households for a census study [Nak00]. A parameter which affects scale can then

define the number of clusters (k), with a set of n observations offering n different possible values k. The second aspect is the "zoning problem", referring to the inherent problem of drawing borders within a data set. This challenge is well discussed in political gerrymandering research, where voting outcomes may be biased to over-represent a particular group when borders are drawn to create voting districts [Won09]. This second aspect complicates the relatively simple variance-bias trade off discussed above, as it allows the membership of a given partition to vary without changing the number of partitions. The number of permutations this allows follows the Stirling partition number series [RD69], a particularly fast growing number series which scales exponentially with larger data sets, posing computational challenges for any data set larger than 100 observations.

2.2 The modifiable areal unit problem for molecular clustering methods

Although the MAUP has been discussed almost exclusively in the field of geography and environmental statistics [FW91, NB17, JW96], a similar problem occurs when choosing a scale in the definition of molecular clusters. It is first important to clarify that clusters are a special type of partition, which are specifically qualified by relatively high connectivity or relatively low variation in a given proximity measure. The MAUP does not necessarily require its partitions to be clusters, although the geographic context of the MAUP implies that partitions are assigned on a spatial basis, with all observations in a given partition sharing a fairly constrained set of possible locations. Although the parameters which control scale may very depending on the clustering method, a smaller scale would be expected to identify numerous small clusters and a larger scale would capture a much smaller number of large clusters. Representing individual sequences as their own cluster is rarely done when clustering HIV sequence data from different patients, however it does represent the theoretical lowest scale for the modifiable areal unit problem and can be thought of as the highest possible number of partitions for a data set. Fortunately, at a particular scale, the membership of clusters that are formed based on proximity measures is fixed. This is because the connections between individual sequences cannot be changed unless without changing the empirical measurements that define them (ie. genetic sequence comparisons). The more complex zoning aspect of the MAUP is then not a part of this study, leaving only the scaling aspect to consider. This hierarchical organization of sequence data can be shown by creating a tree using the unweighted pair group method with arithmetic mean (UPGMA) [D'h05], which iterates through all possible ways that a set of sequences

could be separated into partitions by iterating through a series of collapsing events. A visual example of this process and how it is interpreted as a tree is illustrated below in three steps, where partitions are collapsed into an internal node based on the order of largest to smallest mean pairwise distances (Figure 2.2) .

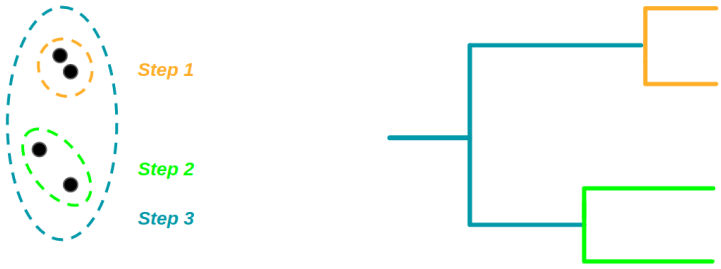

Figure 2.2: This example of the UPGMA acts on four points in a two-dimensional plane. In this case, the distance between points is analogous to pairwise genetic distance. The series of collapsing events is also interpreted as a colour coded tree (right), with branches scaled to illustrate the relative magnitudes of pairwise distances.

For n sequences, a parameter based on this process would have $n-1$ possible values, each representing a stage in the algorithm and an associated set of partitions. Assigning a value x to this parameter would then mean stopping at step x of the algorithm, defining how many partitions the sequences are sorted into. For the remainder of this text, this parameter and any others which determines the level of aggregation for the data set will be referred to as a "scaling parameter". A paper by Bull et, al. (1993) [BHC+93] discusses the use of different partitioning schemes on large sets of sequence data and considering the advantages of running separate analysis on different partitions of the data. This approach views the partition decision as an optimization problem, where a specific scale is most optimal for their classification analysis. Too large a scale misses the relevant differences between partitions, but too small a scale erroneously assumes multiple individuals to be separate. The goals of molecular clustering are similar, as epidemiologists essentially try to classify if pathogen sequences are similar enough to represent a connection. In order to characterize their molecular cluster analysis, the scaling parameters must be identified for the tree-based and graph-based clustering methods developed for HIV.

2.2.1 Scaling parameters for TN93 graph-based clustering

In the graph-based methods such as HIV-Trace [KPWLBW18], the TN93 distance threshold imposed upon the edges between vertices (Section 1.3.1) acts as a scaling parameter. As this threshold is increased, fewer edges are filtered out of the graph, leading to a higher overall connectivity and a smaller number of clusters. Unlike the finite scale selection for UPGMA trees (ie. stopping at a specific step in the algorithm), the TN93 threshold is a value imposed on a continuous measurement. This means that unlike the UPGMA, changes in the threshold can effect multiple partitions at once. In addition, this is a "single-linkage" basis for clustering, meaning that individuals may join a cluster based on a single connection to one of that cluster's members. Because of this single linkage characteristic, large clusters are likely to attain new members simply by virtue of their size [WMM+18], as they contain many potential sequences to connect to. This also allows long bridges to exist between clusters, as illustrated in Figure 2.3, allowing relatively dissimilar sequences to exist in the same cluster [ST06].

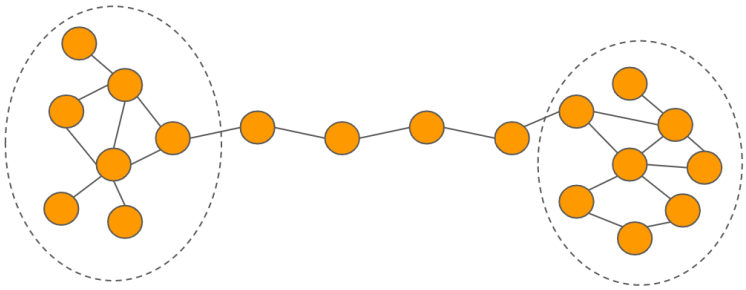

Figure 2.3: An example of bridge formation, in this case allowing the circled sets of sequences to exist within the same cluster.

To Illustrate the effect of pairwise distance threshold as a scaling parameter, a series of networks have been constructed from a subset of *pol* sequences collected in Seattle, USA (Figure 2.4) using three different TN93 distance thresholds to define clusters. The rightmost panel represents a fairly extreme threshold and a course scale, with few partitions being imposed upon the data set. Because 0.05 is the expected pairwise distance between any two HIV-1 subtype B *pol* sequences found in the United States [APP+12], the threshold of 0.05 chosen for the third panel fails to specify a truly unusual degree of similarity, as the connections represent common proximity measures. In the context of a variance-bias trade off, this could be effectively classified as low variance and undertraining. By contrast, the strict threshold in the first panel fails to capture many connections at all, resulting in limited information to act upon. This reduces the number of connections between cases, resulting in a large number of small clusters,

often comprised of a single individual. Although this could be seen as the same overtraining problem as demonstrated earlier, the more significant effect is often the lack of connections between cases. If the research goal requires some connections to be observed, than this provides less outcome data. The low number of individuals with connections drawn between them results in a biased set of clusters, as such a small portion of the overall sample may be effected by the connections that occur by random chance. Intuitively, a threshold of 0 represents the lower bound for this clustering method, representing all points as their own partition of size one, with the exception of completely identical sequences. Again, methods typically would not call one-sequence single partitions a cluster [WLBH+14], but this exclusion would be driven by epidemiological interest and doesn't consider their theoretical importance to an analysis of scale. The upper bound requires more computation, as this would be defined by the highest distance in the minimum spanning tree [KVS72], the set of edges which sums to the lowest total distance while still connecting all individual points in the graph. An important characteristic here is that a graph where many connections are excluded, still may contain the minimum spanning tree, resulting in all sequences placed into a single cluster by a relatively low scaling parameter.

If the connections between cases are treated as a binomial outcome associated with some predictor, such as patient age difference [DOKG+17] or the difference in time between the date of diagnosis [RCJBS+18], then the threshold selection process changes the number of growth events associated with this model. This corresponds to many of the connections which would indicate priority in Figure 1.5 The limited connectivity shown by the low threshold specified in the left panel of Figure 2.4 would not just result in the exclusion of potentially relevant connections between known sequences, but it would also limit the number of connections which define growth. This has been shown to change the effect size of predictors for clustering [VLVR+18, OFP+18, RLD+17] and poses a problem more unique to the goal of predictive clustering, where low scaling parameters have additional negative effects. For studies which are interested in growth, low clustering thresholds may then underestimate the scale of onward transmission.

2.2.2 Scaling parameters for maximum likelihood tree-based clustering

In the tree-based methods such as Cluster-Picker [RCHH+13], the Maximum likelihood trees provide some initial structure through a proposed branching order. This means that monophyletic clades alone can be treated as a partitions, especially those with a high bootstrap support for their common ancestor as discussed in Section 1.3.2. However, molecular clusters are

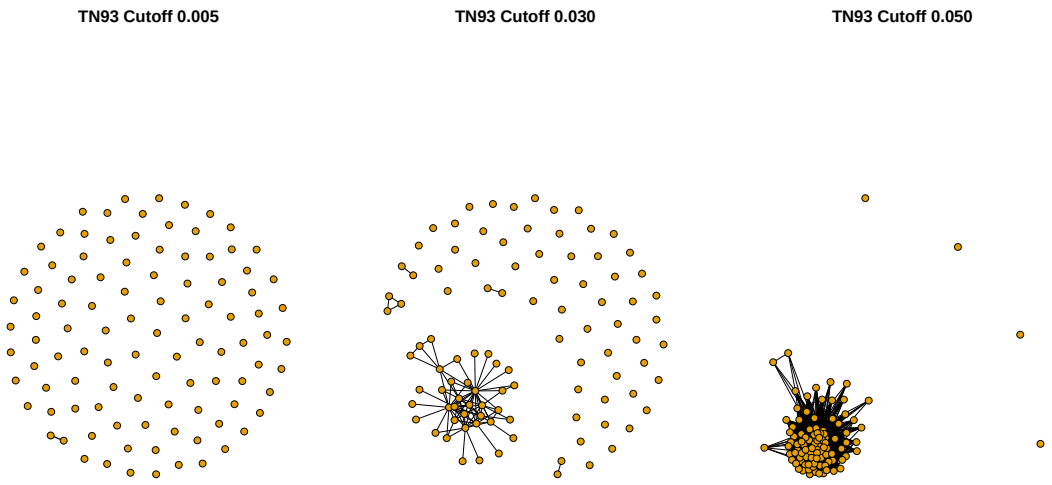

Figure 2.4: An set of graphs built from 153 HIV-1 subtype B *pol* sequences taken from Seattle USA in 2010 [WHVR+17]. Edges represent the pairwise TN93 distances between cases, with each graph showing only the remaining edges beneath a given cutoff threshold. The edges are scaled for visual clarity, and the placement of points on the plane does not represent genetic distance.

typically further qualified by imposing specific requirements on the branch-lengths of a sub-tree. These requirements act as the scaling parameters, with each resulting in a different scale. The influence of these scaling parameters are not necessarily the same as the pairwise graph-based method however. For instance, because these maximum branch length criterion are not based on a single linkage from one sequence to another, tree-based clusters are less tolerant to divergence between members compared to those built from the graph-based method [RLD+17] - connecting a sequence to a given cluster requires that that sequence be close to all members of a given cluster. This manages to avoid the problem of bridging nodes outlined in Figure 2.3, as all sequences in one cluster must be relatively similar to all the sequences in another cluster before the two are merged. A criterion for bootstrap certainty is not strictly hierarchical as a highly confident sub tree may be composed of several sub trees with significantly less confidence. In fact, this would be expected for connections within highly related clusters, as the context of a highly related tree makes it difficult to establish which sequence relationships are closest. These are still used as a criterion for clustering in practice [DOKG+17, RCLH+16],

however they may not necessarily qualify as a scaling parameter.

To demonstrate the effect that branch length threshold selection has on maximum likelihood subtree clusters, a series of subtrees are highlighted in Figure 3.3 as part of a maximum likelihood tree. These are based on a subset of *pol* sequences collected in Seattle, USA using three different thresholds for maximum branch length. The range of meaningful values for this threshold are bounded more simply by the minimum distance between cases and the maximum distance between cases. These represent the points at which all cases are represented by their own individual partition, and all cases are sorted into the same partition (respectively). Unlike the graph-based clusters built on based on single-linkage connections, the use of a maximum branch-length threshold of 0.05 (the expected divergence [APP+12]) may provide a more reasonable indicator of unusual similarity for large clusters, as this threshold selects sub-trees where all cases are at or below the expected rate of divergence. However, this may also indicate a large number of pairs which are connected by chance, thus misleading studies that attempt to discribe clustering connections as instances of transmission [WHVR+17, DOKG+17]. Again, the left and right panels represent the same extremes as discussed by the MAUP, with their associated disadvantages: sparse connections with potentially uninformative clusters and large clusters with less meaningful connections.

If new sequences were to be added to the data set, they provide new information that may indicate the presented tree no-longer holds the most likely branching order or branch lengths. In order to incorporate new sequences into the tree, it is then often necessary to re-construct the entire tree in order to insure that the most likely evolutionary history is presented. To avoid this - either in the interest of time or because the expansion of a fixed tree is being simulated - there are multiple methods that calculate the branch length and most likely placement of new sequences without changing any characteristics of the tree [MKA10, ICCSS14]. This grafting of tips onto a fixed tree has been used to simulate cluster growth [LSM19] in an evolutionary biology context, although has not yet been used in the context of HIV clustering research. If the growth of clusters is of interest (similarly to Figure 1.5), then it is important consider how this outcome is effected by the scaling parameter, as newly grafted tips may join known clusters. In a similar fashion to graph-based clustering, strict maximum branch lengths are likely to prevent these instances of a new tip joining a known cluster.

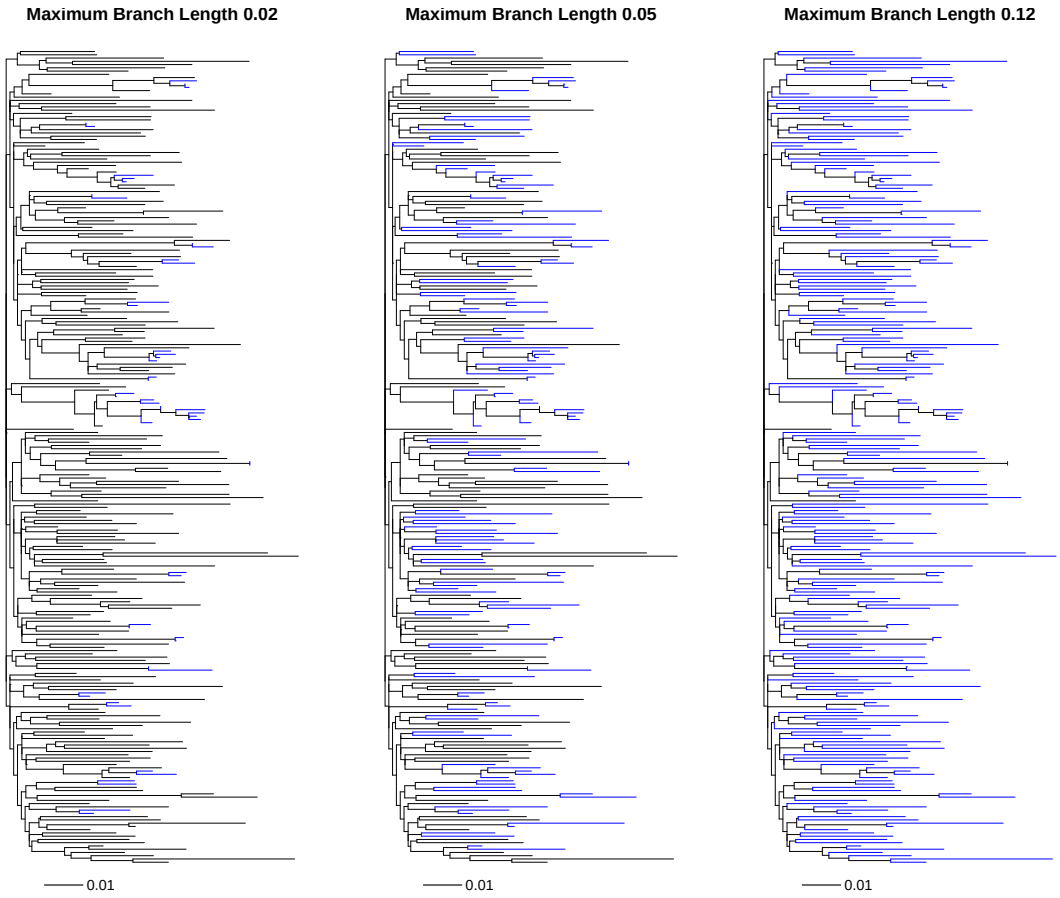

Figure 2.5: A 250-tip subtree shown within a large maximum likelihood tree constructed from 1503 HIV-1 subtype B *pol* sequences. These were collected in Seattle USA between 2000 and 2011 [WHVR+17]. iqTree software with default settings was used to construct the overall tree [NSVHM15]. Branches highlighted in blue represent complete monophyletic clades where all pairwise branch lengths fall below the maximum branch length referenced above the tree.

2.3 Optimal scaling parameters

The selection of scale effectively becomes an optimization problem, and therefore, an external measurement of performance is needed to assess the quality of a given partition set. While some molecular clustering studies have discussed the effect of scaling parameters using measurements of complexity [HYW+18], these are fairly infrequent and not commonly used for the study of HIV. Instead, HIV studies will often receiver operator curves for the detection of known transmission pairs [RHR+19, MP17], which is often interpreted as a measure of how accurately clustering connections represent transmission [MPL+20]. The predictive growth studies referenced in chapter 1, also talk about the fit of a predictive model for future clustering

connections given a particular threshold [WMM+18, DVF+18, BPP+19]. However, there is no generally accepted way to assess the information content of a set of clusters which takes into account the threshold used for clustering. For the purposes of predictive growth modeling in real-time clustering studies, this becomes important, as extreme clustering thresholds may be useless to public health, while yielding impressive measurements of fit [HYW+18]. This section will discuss which statistical tools can be used to judge the information content of clusters with respect to the MAUP and move towards a metric to judge predictive clustering models while taking into account scaling parameters.

Tomoki Nakaya (2000) [Nak00] proposes several information-based solutions to the MAUP, using an estimation of mortality rate at different administrative scales (ie. districts, cities, wards, ect.) as an example. One of these solutions treats mortality as a countable outcome predicted by a Poisson model, and uses the absolute gain in model accuracy as an indicator of appropriate scale-parameters. In this framework, two predictive models are used to predict mortality for an area. A "full" model predicts mortality in the smallest possible administrative regions with each representing its own partition. The paper's term for these are "Basic Spatial Units" (BSUs). The number of deaths y for some BSU i with death rate α_i and population size B_i is given by $y_i = \alpha_i B_i + \varepsilon_i$ where ε_i represents some error from the expected number of deaths. The log-likelihood for the full model is represented by the following equation.

$$l_f = \sum_i (y_i \ln(\alpha_i B_i) - \alpha_i B_i) \tag{2.1}$$

The "restricted" model, makes mortality-rate predictions within aggregated partitions. Each partition then contains multiple BSUs, but no BSU is contained in multiple different partitions. In this case, the death rates for all BSUs within a partition j is simplified to be the mean death-rate for BSUs within the partition. The estimate of mortality for each BSU i within partition j is then $y_i = \alpha_j^A B_j + \varepsilon_i$ where α_j^A represents the mean death rate across all BSUs in partition j. The log-likelihood for a restricted model with some partition scheme A is represented by the following equation.

$$l_A = \sum_i \left(y_i \ln(\alpha_j^A B_i) - \alpha_i^A B_i\right) \tag{2.2}$$

Given these equations as well and the likelihood measurements, it is possible to measure Akaike's information criterion (AIC) [Aka73], an absolute measure of inaccuracy and model complexity. The AIC for some model A is calculated as shown below, given that p represents the number of parameters for the model and l represents the log likelihood of that model in the

context of outcome data. Therefore, both poor model fit and high model complexity contribute to the AIC.

$$AIC_A = 2p - 2l \tag{2.3}$$

Given the AIC measurements of the two models, a difference in fit can be calculated as $AIC_{restricted} - AIC_{full}$, such that negative values would imply that the restricted model outperforms the full model.

Because each administrative scale creates a new set of partitions, and each partition adds to the number of parameters estimating the death rate, p in the AIC calculation is equivalent to the overall number of partitions at a given scale. Therefore, the reduction in overall likelihood is counter-balanced by the increases in model complexity, with the full model acting as a baseline, showing both the highest likelihood and the highest penalty for partitions. Any benefits from fewer partitions for some restricted model are quantified by a negative value in the AIC difference. A positive value in this difference would correlate to a partitioning scheme which sacrificed enough model information (decreasing l) to outweigh the benefits of a simpler set of estimates (low p).

When considering how this AIC-loss metric applies to predictive molecular cluster models, it is also necessary to consider that cluster growth is based on connections between individuals, and would therefore also be effected by the scaling parameter [VLVR+18]. There is basis for this in ecological research, where this effect has been characterized for binary classification outcomes, using 11 different metrics to measure performance in response to the threshold which determines a positive classification [FM08]. However, this goal of binary classification does not apply to quantitative, Poisson-linked growth outcomes for clusters over time. Despite common criticism as to the accuracy of genetic clustering techniques for HIV [VKW+12, Poo16, RLD+17, NML+14], there exists no in-depth analysis which characterizes the way in which Poisson-linked growth outcomes for each cluster respond to clustering threshold. Furthermore, there is no available framework which informs the selection of an appropriate threshold based on the characteristics of a data set.

This thesis work aims proposes such a framework by implementing an altered version of Nakaya's approach to the modifiable areal unit problem. This framework was developed and implemented using both tree-based clustering and graph-based clustering methods then demonstrated on multiple sets of real HIV sequence data in North America. I propose that the theoretical optimum threshold value would then provide the greatest amount of information for the training of a quantitative predictive cluster growth model that estimates which clusters are

likely to attain the most new cases. Furthermore, I aim to show that this information can be evaluated with respect to thresholds by calculating the loss of AIC relative to a baseline ”null” model.

It is my hypothesis that the threshold values used for the identification of HIV clusters can be optimized through this framework and that this optimization is a necessary step as the optimal threshold values vary between research locations. The flowchart used to outline this framework is summarized below in Figure 2.6 and discussed in more detail within the following chapters

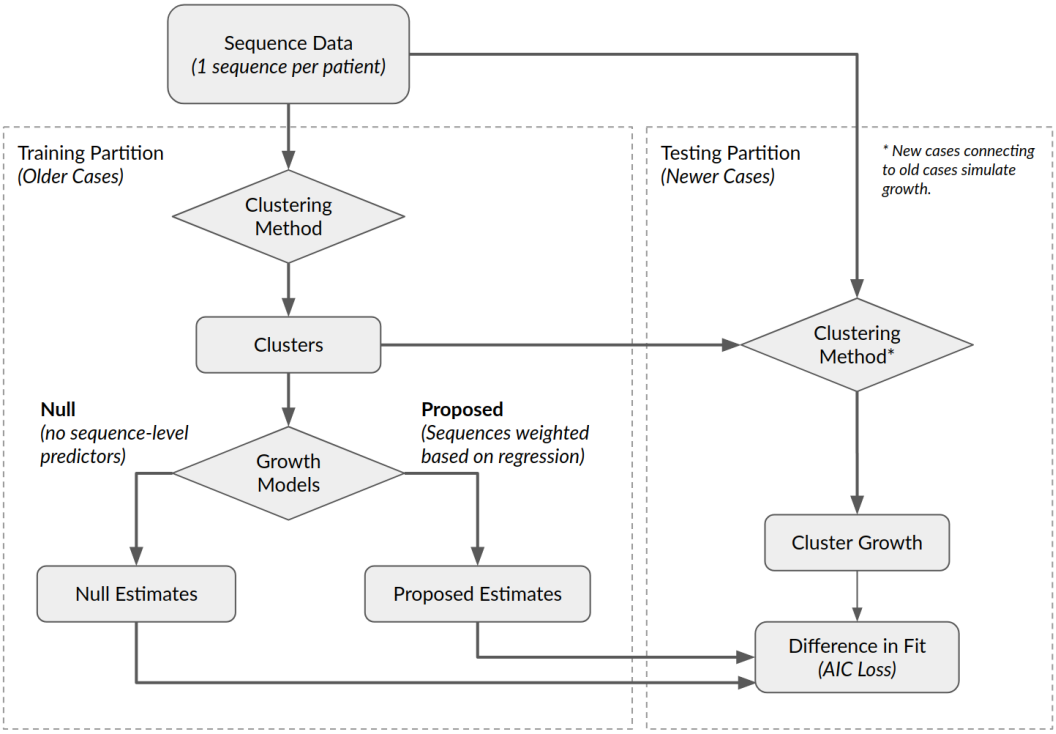

Figure 2.6: The initial flowchart which outline the stages of my optimization framework

Chapter 3

Methods

3.1 Methods overview

This work develops and demonstrates a framework to optimize the selection of a scaling parameter for popular threshold-based methods used on HIV such as HIV-Trace and Cluster Picker, this means choosing the most appropriate threshold (ie. TN93 distance threshold for pairwise edges or maximum within-subtree patristic distance). Here, the optimal threshold produces clusters with the greatest information content and this information content will be measured by AIC loss associated with predictive variables. The outcomes which obtain AIC, are effectively simulations of real time cluster growth, performed by adding ”new” sequences to the data set, and observing which known clusters those sequences join. A predictive model observes the set of connections in known clusters to estimate to associate the likelihood of cluster connections with some kind of predictor variable. The threshold acts as a scaling parameter by altering the number of connections for any given clustering method, changing the number of clusters, the number of connections which could train a predictive model, and the number of growth outcomes, Adjusting Nakaya's [Nak00] information-based approach to the modifiable areal unit problem, performance is measured by a comparison between two models which predict which predict which clusters new cases will join: a ”null model” with no predictor variables; and a ”proposed model”, with one or more predictor variables associated with connection, such as risk factor, age difference between patients or patient location. In effect, this replaces the ”full model” in Nakaya's approach for a ”null model”, which responds to the same scaling parameter as the proposed model, but assumes all sequences are equally likely to connect to each other and cluster growth is predicted only by size. The AIC loss calculated between the proposed model and a null model captures how much accuracy is gained when using a model that allows

predictor variables to predict connections between individual. Put another way, this measures how useful the predictor variables are. Optimal thresholds should result in clusters that make the predictor variables appear most useful (ie. the highest AIC loss). As an example, time lag is used as a predictor variable, with the proposed model assuming that connections are more likely between sequences with a similar associated time (either sequence collection date or patient diagnostic date). It follows that clusters with a large number of recent cases would then be most likely to grow. The framework is implemented mostly in the R programming language [R C13], with some supporting scripts written in Python [VRDJ95]. The results in the following chapter are obtained from a demonstration on three separate sets of HIV-1 subtype B data collected in North America. Further interests explored in this project include factors within the data which may influence the optimum threshold, the robustness of the optimum parameter to random sub-sampling, the stability of the optimum threshold over time, and the difference in performance between multiple indicators of time point (ie. date of sequence collection versus date of diagnosis).

3.2 HIV data sets and data processing

3.2.1 Sequence data

Three different anonomized data sets of aligned HIV-1 *pol* sequences were obtained for the purposes of this study. These were population data sets, reported such that individuals are represented by only one sequence in the alignment. Two of these were publicly accessible through the Genbank Database [BKML+11] - $n = 1648$ sequences collected in Seattle, USA [WHVR+17] and $n = 1020$, sequences collected in Northern Alberta [VAB+17]. After gaps were inserted in the alignment process, the lengths of these sequences were 1095bp and 1077bp for Seattle and Alberta respectively. A Biopython module was used to query Genbank for the sequence collection date associated with each accession number [CAC+09]. This collection date was used as the associated timepoint for sequences, separating new versus old cases. The third data set was collected by the Vanderbilt comprehensive care clinic in Nashville Tennessee and surrounding area. Patient meta data including the year of sequence collection and year of patient diagnosis were available as part of this set [DVF+18]. The Tennessee data contained a total of 2,779 sequences, each 1500 bp in length after alignment. Each data set was filtered to remove any sequences which were annotated as a subtype other than B, as well as any sequences with ambiguous bases at 5% of the positions, which would generally indicated a problem during sequencing. This filtration step removed 211 subtype C sequences and 1

ambiguous sequence from the Alberta data set. 163 sequences with high ambiguity were also from the Tennessee data set. In addition, the time range of the Seattle data set was adjusted, removing the sequences collected in 2013 due to an unusually low sample size of 35 sequences for that year. This likely indicates that sampling was not carried out through the entire length of the year at the time these sequences were published. The sequences within the Tennessee data set were also truncated using the Aliview [Lar14] tool due to a poorly aligned terminal region, reducing the overall sequence length from 1500 bp to 1305 bp. Within the filtered Tennessee data set, only 2,077 of the remaining 2,616 sequences contained diagnostic years. Further information regarding each data set after filtration is summarized in the following table for reference.

Location	**Seq Length**	**Sample Size**	**Date Range**	**Time Information**
Seattle, USA	1095 bp	Collection Year	1,613	2000-2012
Northern Alberta, Canada	1077 bp	Collection Date	808	2007-2013
Middle Tennessee, USA	1305 bp	Collection Year	2,616	2001-2015
-Subset-		Diagnostic Year	2,527	1977-2013

Each sequence in the data set had some associated timepoint; either sequence collection year or the year that the host patient was diagnosed with HIV. Before clustering, the distributions of time information were collected for all data sets, and summarized in Figure 3.1. These time points were used to define the time lag between sequences and establish subset of each data set to be defined as ”new sequences”. Subsets of ”new sequences” contained only the sequences diagnosed at the newest time point and would later be used to validate the predictive growth models. All data sets contained at least the year of collection for each sequence, with a relatively even distribution of sampling effort (number of sequences collected per year). Collection rates averaged 124, 115, and 174 sequences per year for the Seattle, Alberta, and Tennessee data sets respectively, although all data sets saw fewer sequences collected in the first few years of sampling. For all data sets, there were over 100 sequences associated with the newest time point with 110, 110, and 153, for Seattle, Alberta and Tennessee as well as 129 for the diagnostic Tennessee subset. This ensured that there was a sufficient number of cases to be used for validation.

For the Alberta data set, sequence collection information was also given at the resolution of complete dates, with September averaging the lowest number of sequences collected per month (8.5) and December averaging the highest (14.3). The collection rate does not necessarily correlate to the incidence of HIV during these years, as no estimates were available for time since infection or the proportion of the epidemic that was sampled. For the sequences collected

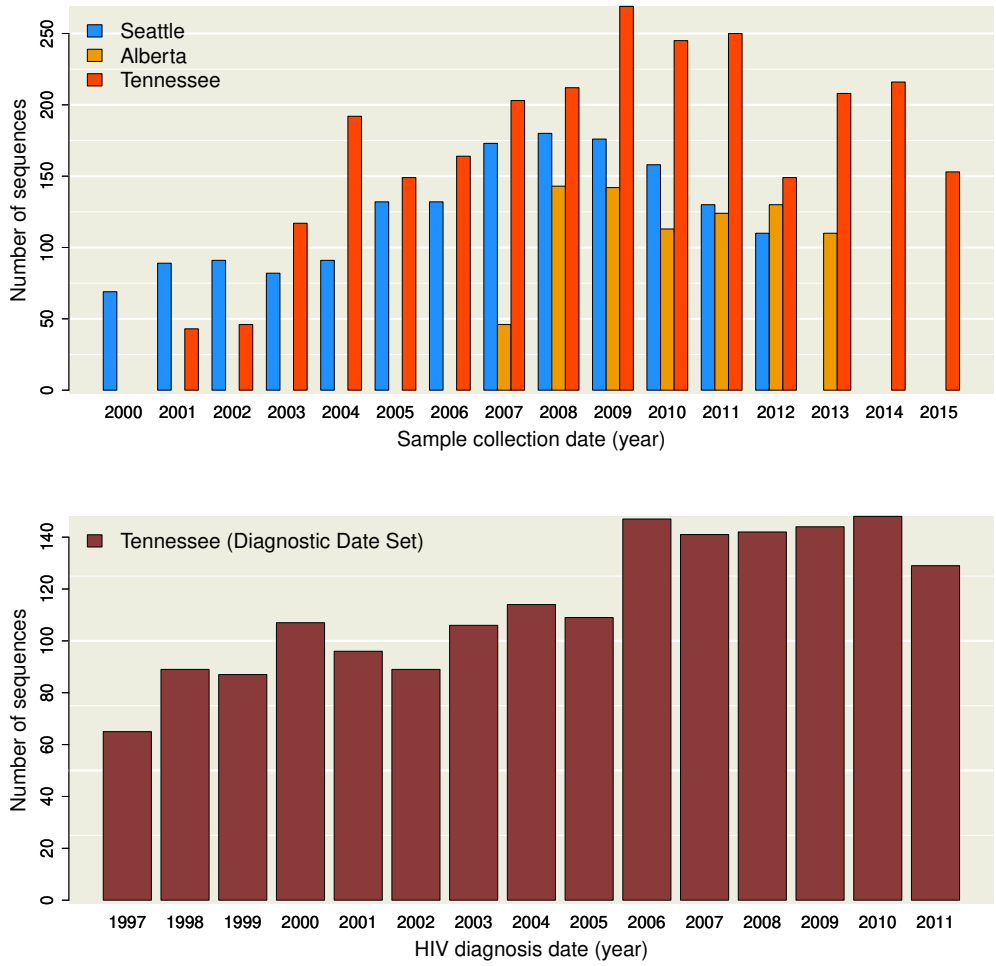

Figure 3.1: **(top)** Distribution of sequence collection years for the Seattle (blue), Alberta (orange) and Tennessee (red) data sets. Absent bars indicate that no sampling was carried out in the respective years, and does not reflect an absence of cases. **(bottom)** Distribution of sample diagnostic years for the cases in the Tennessee data set. For clarity, this excludes the sparse tail to the left of this distribution, which would contain cases diagnosed between 1977 and 1997.

in Tennessee with an associated diagnostic date, the early dates of diagnosis were particularly sporadic, with a total of 429 sequences corresponding to patients diagnosed between 1977 and 1997, compared to 1,648 diagnosed from 1997-2013. For this reason, these years are excluded from the bottom half of the presented timepoint distribution figure (Figure 3.1 (bottom)) for clarity. These early timepoints are still included for the purposes of training a predictive model. Given the extremely early dates of diagnosis for some of these sequences, it is likely that some of these cases, particularly those tagged before 1980 were diagnosed retrospectively, as HIV was not nationally reported in the United States until the 1980s [GSF+81]. Across the

Tennessee diagnostic subset, sequences were collected, on average, 5.39 years after the date of diagnosis. However, because data surveillance programs did not see widespread use in the United States until after the year 2000 [OWH+15, CGO+14, WZZ+10], the subset of cases diagnosed after 2000 (n = 1365) may be more informative, with a mean time lag of 2.20 years between HIV diagnosis and sequence collection.

3.2.2 TN93 Distances and Tree building

TN93 distances were calculated between all possible pairs of using the open-source TN93 calculation tool associated with HIV-Trace [kPWV18] (https://github.com/veg/tn93) which is implemented in the C++ programming language [ES90]. Any unknown or ambiguous bases in the sequence were resolved to whatever base would minimize the distance between sequences in the overall alignment. Maximum likelihood trees were built using the open source IqTree software [NSVHM15] using their ultrafast bootstrap approximation with 1000 bootstraps [MNvH13]. This contrasts the more traditional method of obtaining bootstrap values described in the first chapter by using a statistical model to approximate the certainty of placements, instead of fully rebuilding any parts of the tree. A general time reversible model of evolution as described by [LPSS84] with free rate variation among sites to determine likelihood [Yan95] and optimized base frequencies. These trees did not include the sequences in the most recent time point, which were withheld to represent new cases in the measurement of cluster growth. In order to measure growth on a fixed tree, I used the open source pplacer software (https://github.com/matsen/pplacer) version 1.1 alpha19 [MKA10]. This tool calculates the branch length and placement of new tips onto a fixed tree, effectively allowing new tips to be added without requiring the recalculation of the tree. Further, pplacer computes a posterior probability of that placement, allowing for a metric similar to bootstrap support when a new node is created through placement. Within pplacer, the guppy command with the *sing* subcommand was used to produce a new tree for each new tip, placing each new tip at its most likely location.

3.3 Implementation of cluster methods

This section aims to clarify the logic which is used by the R code to implement the popular clustering methods described previously and formally describe how a growth connection was defined and how the predictive model was trained. For both methods, the associated code creates clusters, trains a Poisson-linked predictive model based on connected sequences and

validates that model by adding new sequences to the data set and simulating growth. It also responds to a given scaling parameter (the TN93 distance threshold for graph-based clustering and maximum internal patristic for maximum-likelihood subtree based clustering). Minimum bootstrap criterion is considered as a secondary parameter for clustering, as this is not considered a requirement for two cases being connected on an individual basis or a requirement for a new case joining a new cluster.

3.4 Graph-based clusters

3.4.1 Defining Clusters

The first clustering method implemented in this work mirrors the graph-based approach to clustering taken by methods such as HIV Trace [KPWLBW18]. After all the pairwise TN93 distances are calculated from a set of aligned sequences, they are taken as an edge list for the creation of a complex, undirected graph. We will let G represent the complete graph built from vertices $V(G)$ and edges $E(G)$. Initially, this graph contains all possible edges between vertices G constitutes a training set, reserving a subset of data from the newest timepoint for validation. Each edge $e \in E(G)$ has an associated TN93 distance $d(e)$ and each vertex $v \in V(G)$ has an associated time-point $t(v)$, representing the point at which the sequence was collected or, if available, the point at which the associated host was diagnosed with HIV. Although timepoint data is not an inherent part of the graph-based clustering methods, it is the most regularly available piece of information that could be used to train a predictive model [BKML+11]. Because $E(G)$ initially represents the complete set of edges, it can be assumed that all vertex pairs $\{v_i, v_j\} \subseteq V(G)$ have some edge $e(v_i, v_j)$ that connects them directly. $E_{\text{filt}}(\max_d)$ represents the set of edges constrained by some maximum distance $\max_d$, implying that for all edges in $E_{\text{filt}}(\max_d)$, $d(e) \leq \max_d$. Each cluster C is a "component" of G; a sub-graph containing edges $E(C)$ and vertices $V(C)$ where for all vertex pairs $\{v_i, v_j\} \subseteq V(C)$ there is some set of edges $\{E_{\text{path}}(v_i, v_j) \,|\, E_{\text{path}} \subseteq E_{\text{filt}}(\max_d)\}$ that connects them indirectly. Finally, single vertices with no incident edges are considered clusters of size 1. 3.2 shows a list of these features, and clarifies some important terms that distinguish certain types of edges explained in the following subsections.

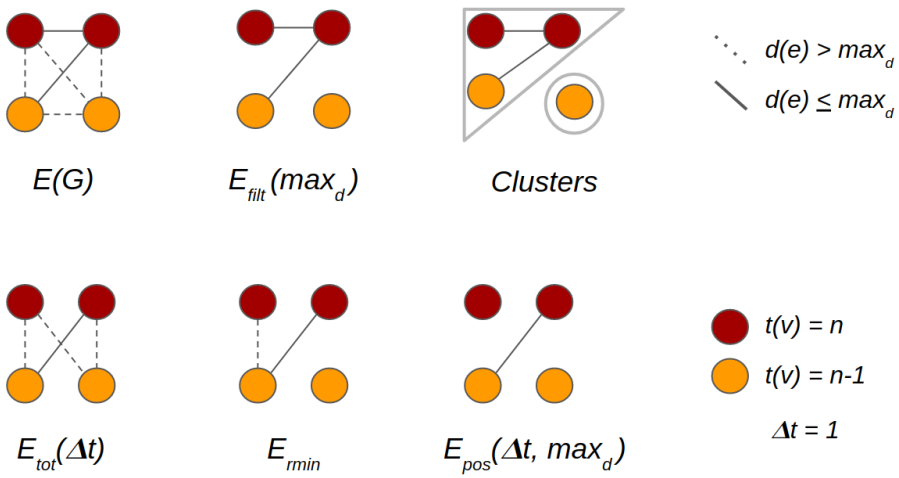

Figure 3.2: An example graph with four vertices spread across two different time points. Each illustration clarifies the remaining graph after each a given filter is placed upon it, corresponding to the different subgraphs referenced in the following subsection. The top right illustration simply clarifies the definition of clusters as a component of a graph.

3.4.2 Predictive model training

For each v in $V(G)$, the "minimum retrospective edge" $e_{\text{rmin}}(v)$ can be obtained for a given vertex v. For all "retrospective" edges connected to v and another case at an earlier time point than v, $e_{\text{rmin}}(v)$ has the smallest distance. The set of all of these edges for G is termed E_{rmin} and represents important connections for the purposes of predicting cluster growth. Each edge $e(v_i, v_j) \in E(G)$ has a time lag $\Delta t(e) = t(v_i) - t(v_j)$ which can be used as a predictor for a given edge's membership in $E_{\text{rmin}} \cap E_{\text{filt}}$. $|E_{\text{tot}}(\Delta t)|$ to refer to the total number of edges $e(v_i, v_j) \in E_{\text{filt}}(\max_d)$ such that $t(v_i) - t(v_j) = \Delta t$. We can then use $|E_{\text{pos}}(\Delta t, \max_d)|$ to refer to the size of intersection $E_{\text{rmin}} \cap E_{\text{tot}}(\Delta t)$. Given a specific $\max_d$ value to limit edges in the graph, the following logistic regression quantifies how frequently edges with a given time lag appear as minimum retrospective edges in $E_{\text{filt}}(\max_d)$.

$$\log\left(\frac{|E_{\text{pos}}(\Delta t, \max_d)|}{|E_{\text{tot}}(\Delta t)| - |E_{\text{pos}}(\Delta t, \max_d)|}\right) = \alpha + \beta\Delta t \tag{3.1}$$

If this time lag between vertices has a negative correlation with the likelihood that those vertices will be connected by a minimum retrospective edge, then it follows that vertices with a time-point closer to the newest time point $t_{\max}$ are more likely to connect to new vertices. As will be detailed in the next subsection, this is because new vertices join whichever cluster is connected to them via their minimum retrospective edge. If the minimum retrospective

edge of a new vertex is filtered out via $\max_d$, then that case doesn't join any cluster. We can then weight each vertex based on time point using the coefficients obtained through regression $w(v) = \alpha + \beta(t_{\max} - t(v))$. The growth prediction for a given cluster $\hat{R}_{\text{proposed}}(C)$ will then be based on the sum of $w(v)$ for each vertex in $V(C)$.

$$\hat{R}_{\text{proposed}}(C) = \exp\left(\sum_{v \in C} w(v)\right) \tag{3.2}$$

A baseline "null" model compares to the overall effect of weighted vertices. This would assume $w(v) = 1$ for all $v \in V(T)$ and can be calculated as

$$\hat{R}_{\text{null}}(C) = \exp\left(|V(C)|\right) \tag{3.3}$$

3.4.3 Validation through growth

In order to obtain an actual growth measurement $R(C)$, we measure the growth of clusters through the addition of new vertices. A "new" vertex $v' \notin V(G)$ has $t(v') > \max_t$ where $\max_t$ represents the max time-point value from the set $\{t(v) \mid v \in V(G)\}$. The "growth" of G is a process simulating an update of cluster over time where individual new vertices $V(G)$ through the following actions. For some new vertex v', we let e' represent the minimum retrospective edge of v', with some distance $d(e')$. If $d(e') < \max_d$, then this edge will connect v' to some vertex in $V(G)$. After this growth process, individual clusters may have obtained new vertices, but because only one minimum retrospective edge exists per new sequence, we can assume that no new sequence in V' was added to multiple clusters. This is done so that no clusters are "merged" together, step does not change the partitions used in the training step. "Growth" value $R(C)$ for individual clusters can be defined as the number of new vertices attained, calculated as the number of new sequences that join any given cluster.

3.5 Tree-based clusters

3.5.1 Defining Clusters

This mirrors the maximum likelihood subtree approach taken by Cluster Picker [RCHH+13]. We let T represent a midpoint-rooted, phylogenetic tree with internal nodes $N(T)$, tips $V(T)$, branches $E(V)$ and some root node $r \in N(T)$. Each branch $e \in E(T)$ has an associated branch-length $d(e)$, each tip $v \in V(G)$ has an associated time point $t(v)$ and each internal node $n \in N(G)$

has an associated bootstrap support value $b(n)$. It is important to note that the root node in a midpoint-rooted tree cannot normally have its own bootstrap support value, as the root is defined here as the ancestor to all tips. However, we assign $b(r) = 0$ in order to ensure that even the lowest possible minimum bootstrap threshold would allow all tips to be sorted into the same cluster. Each branch $e \in E(T)$ connects either two internal nodes, or one internal node and one tip. For any two tips $\{v_i, v_j\} \subseteq V(T)$, there exists exactly one set of edges $E_{\text{path}}(v_i, v_j) \subseteq E(T)$ that connects them with the minimum number of branches. The patristic distance is the total branch-length traversed by any particular $E_{\text{path}}(v_i, v_j)$ and can be calculated by the following equation

$$d(E_{\text{path}}(v_i, v_j)) = \sum_{e \in E_{\text{path}}} d(e) \tag{3.4}$$

A "subtree" is a monophyletic clade T_{n_i}, defined as the subset of nodes and branches converging to a given internal node $n_i \in N(T)$ with branches $E(T_{n_i}) \subseteq E(T)$, internal nodes $N(T_{n_i}) \subseteq N(T)$, and tips $V(T_{n_i}) \subseteq V(T)$. This is represented in figure 3.3, as are the branch paths between tips.

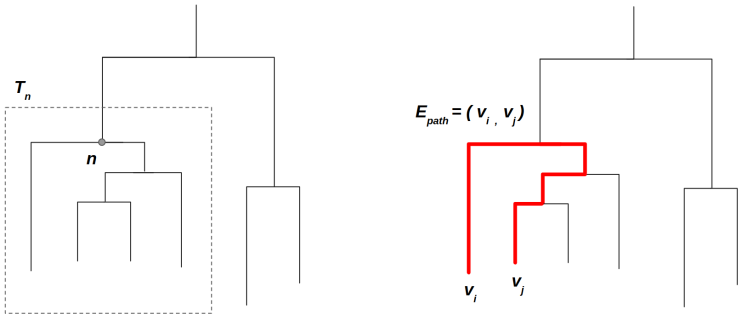

Figure 3.3: Some clarification on subtrees and branch paths between tips. The patristic distance is the total vertical distance traversed throughout the branch path

Any sub-tree T_n has an element $\bar{t}(V(T_{n_i}))$, representing the mean $t(v)$ for all tips in $V(T_{n_i})$ as well as an element $max_d(T_n)$, representing the largest patristic distance in the tree. $N_{\text{filt}}(\max_d, \min_b) \subseteq N(T)$ represent the set of nodes constrained by two parameters $\max_d$ and $\min_b$, implying that $b(n) \geq min_b$ and $\max_d(T_n) \leq \max_d$ for any $n \in N_{\text{filt}}(\max_d, \min_b)$. A cluster C is defined as a given sub-tree T_n where $n \in N_{\text{filt}}(\max_d, \min_b)$ and n is not a member of any larger cluster. Individual tips which are not a member of any cluster can be considered clusters of size 1. We take $\max_d$ and $\min_b$ as inputs for the following calculations.

3.5.2 Predictive model training

For T, we may obtain the subset of internal nodes that connect to tips $N_{\min}$. These represent the "direct ancestors" of at least one tip. Each direct ancestor node $n \in N_{\min}$ has a time lag $\Delta t(n)$ which is based on the t values of tips which descend from that n. $\Delta t(n)$ can be calculated in two different ways depending on whether n connects to two tips or one tip and one internal node. These two cases are shown below in 3.4

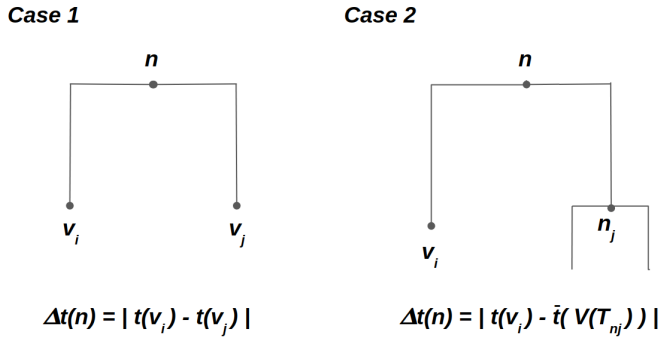

Figure 3.4: The two cases that encapsulate how a node n in $N_{\min}$ will exist in the tree. The two cases of how time difference are calculated are also shown - either between the time point of each tip, or the time point of one tip and the mean time point of all tips in a subtree

- For two neighbouring tips $\{v_i, v_j\} \subseteq V(T)$ such that there exists branches $\{e_i(v_i, n_i), e_j(v_j, n_i)\} \subseteq E(T)$ and $n_i \in N_{min}$, the time lag $\Delta t(n_i)$ can be calculated as $|t(v_i) - t(v_n)|$
- For a tip $v_i \in V(T)$ with a neighboring internal node $n_j \in N(T)$ such that there exists branches $\{e_i(v_i, n_i), e_j(v_j, n_i)\} \subseteq E(T)$ and $n_i \in N_{min}$, the time lag $\Delta t(n_i)$ can be calculated using the mean time values of tips in T_{n_j}: $|t(v_i) - \bar{t}(T_{n_j})|$

Given $\max_d$, we may limit the number of nodes in $N_{\min}$ if they meet either one of two criteria.

- If $n \in N_{\min}$ is connected to two different tips $\{v_i, v_j\}$ and $d(E_t extrmpath(v_i, v_j)) > \max_d$
- If $n \in N_{\min}$ is connected to an internal node n_i and a tip v_i and there exists some tip in $v_j \in T_{n_i}$ such that $d(E_{\text{path}}(v_i, v_j)) > \max_d$

Nodes in $N_{\min}$ then represent an instance where at least one tip meets the clustering criterion - either clustering with a neighbouring tip, or joining a neighbouring subtree. We will use

$|N_{\text{tot}}(\Delta t)|$ to refer to the total number of tips which could theoretically form a node $n \in N_{\text{min}}$ such that $\Delta t(n) = \Delta t$. This will often not include all tips in $V(T)$ as many tips with more moderate $t(v)$ values cannot experience the largest time difference possible in the set. We can then use $|N_{\text{pos}}(\Delta t, \max_d)$ to refer to the total number of nodes $n \in N_{\text{min}} \cap N(C)$ for any cluster C in the tree such that $\Delta t(n) = \Delta t$. The following logistic regression quantifies how frequently tips join the tree such that their direct ancestor spans a given time lag.

$$\log\left(\frac{|N_{\text{pos}}(\Delta t, \max_d)|}{|N_{\text{tot}}(\Delta t)| - |N_{\text{pos}}(\Delta t, \max_d)|}\right) = \alpha + \beta\Delta t \tag{3.5}$$

In order to predict cluster growth, we can then weight each tip based on time point using using the coefficients obtained through the regression above. $w(v) = \alpha + \beta(t_{max} - t(v))$ The growth prediction for a given cluster $\hat{R}_{\text{proposed}}(C)$ will then be based on the sum of $w(v)$ for each tip in $V(C)$.

$$\hat{R}_{\text{proposed}}(C) = \exp\left(\sum_{v \in C} w(v)\right) \tag{3.6}$$

A baseline "null" model compares to the overall effect of weighted tips. A null point of comparison with no weight, would assume $w(v) = 1$ for all $v \in V(T)$. This is calculated by the following equation

$$\hat{R}_{\text{null}}(C) = \exp\left(|V(C)|\right) \tag{3.7}$$

3.5.3 Validation through growth

In order to obtain an actual growth measurement $R(C)$, we measure the growth of clusters through the addition of "new" vertices. The "growth" of T is a process where an individual tip v' joins the tree . This creates a new internal node n', effectively splitting some edge $e(n_i, v_i) \in E(T)$ into $e(n_i, n')$ and $e(n', v_i)$ such that $d(e(n_i, n')) + d(e(n', v_i)) = d(e(v_i, n_i))$. It is important to note that newly added tips do not split other newly created branches. Each added tip also creates a new branch $e(v', n') \in E(T)$ with branch length $d(e')$. For clarity, the edge splitting process is visualized in 3.5.

Given $\max_d$, if the distance from v' to any tip in the cluster it joins is greater than $\max_d$ (ie. large enough to "break" the cluster), then we remove the associated tip v' from consideration. Otherwise, the tip potentially joins a given cluster. After the growth process, individual clusters may have obtained new tips. We can define the "growth" value $R(C)$ for individual clusters as

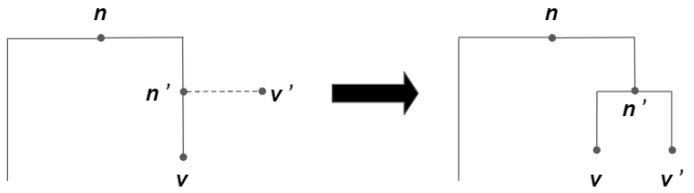

Figure 3.5: The edge splitting process, by which new cases are appended to a tree

the number of new cases attained, calculated as the number of new tips that joined a given cluster.

3.5.4 AIC Calculation

In order to quantify the information content of a particular set of clusters, we use the *AIC* measurements associated with our growth models $\hat{R}_{\text{null}}$ and $\hat{R}_{\text{proposed}}$ given observed growth. In order to calculate, AIC, first a likelihood measurement must be obtained for the growth of a given cluster. The likelihood for a proposed model given a set of clusters built at some threshold $\max_d$ is given by the following equation

$$l_{\max_d} = \sum_i \left(R(C_i) \ln(\hat{R}_{\text{proposed}}(C_i)) - \hat{R}_{\text{proposed}}(C_i) \right) \tag{3.8}$$

The likelihood of a null model is then given by a similar equation, but substituting $\hat{R}_{\text{proposed}}$ for the proposed $\hat{R}_{\text{null}}$

$$l_T = \sum_i \left(R(C_i) \ln(\hat{R}_{\text{null}}(C_i)) - \hat{R}_{\text{null}}(C_i) \right) \tag{3.9}$$

The AIC is then calculated for each. Because only the proposed model takes in predictors, the AIC effectively only penalizes the $\hat{R}_{\text{proposed}}$. The difference $AIC_{proposed} - AIC_{null}$ then represents the loss of AIC that accompanies the use of predictor variables in the predictive model $\hat{R}_{\text{proposed}}$. This is inspired by the solution to the MAUP proposed by Nakaya et al, [Nak00]. However instead of comparing a set of clusters to a set of completely individual predictor variables, two different models are compared on the same set of clusters. This is done because the same parameter which partitions the set of sequences into clusters ($\max_d$) can also control the cluster growth outcomes. Therefore a null model with no clustering would have either no outcomes, or outcomes which are inconsistent with the rules by which the data is partitioned.

Instead, this AIC loss represents the gain in predictive model accuracy associated with a new predictive variable (time point) in response to a specific maximum distance.

3.6 Framework testing

A default "run" of this framework reports the above AIC difference calculation responding to a series of 50 scaling parameter values. These values are used to define $\max_d$ in the previous equations, defining either the largest TN93 distance which could represent an edge in a graph, or the largest patristic distance allowed in a subtree for the purposes of clustering. The sequence collection year is used as the default time point for the purposes of assigning $t(v)$ and the default training set contains all sequences excluding only those collected in the most recent year (which is withheld for the validation step). For the diagnostic subset of the Tennessee data, the diagnostic year of the patient is used. An initial version of the code which executes such a run has been made public under the name *Mountain Plot* (https://github.com/PoonLab/MountainPlot) with an associated publication [CKP20] For the graph-based clustering method, TN93 distance thresholds of 0 to 0.040 (0.0008 increments) were used as $\max_d$ for the retention of edges. For the maximum-likelihood tree based method, maximum internal patristic distance thresholds of 0 to 0.15 (0.003 increments) were used as $\max_d$ for the classification of some subtrees as clusters. These were chosen based on the upper and lower bounds of clustering, with the highest values in each set of $\max_d$ values representing the point at which all sequences are sorted into the same cluster. In order to assess the effect of bootstrap values as an additional requirement for clustering, the confidence requirement for common ancestors ($\min_b$) was held static at 0.90 during these changes to $\max_d$. For runs which do not provide large or consistent AIC loss, a random model is available as a control, where sequences are weighted randomly using a random sample from a distribution with a mean of 1 and a standard deviation of 0.25. This random model then replaces the proposed model, to interpret whether or not the advantages of the proposed model are meaningful compared to an irrelevant predictor variable.

3.6.1 Robustness testing and time information analysis

The robustness of this framework is assessed through two different tests. The first assesses the effect of incomplete sampling, a known influence for clustering methods [NML+14]. For the Tennessee data set, 3 different rounds of 30 random samples are taken without replacement. Each round of sampling took a different proportion of the sequences (0.4, 0.6 and 0.8) to investigate the effect of sampling density on maximum AIC loss. The framework is then run

on each sample using the graph-based clustering method. For practical reasons, this test was limited to the significantly faster graph-based clustering method, as a feasible way to quickly re-construct many maximum likelihood trees is not yet part of this framework. In order to insure a large enough sample, only the Tennessee data set was used, to prevent the stochastic associated with small samples less than 800 sequences each. This test was run on both the complete data set and the diagnostic subset. The second test assesses the way that outcomes of the framework may change over time, simulating continued use of the same data set while new cases are sequentially added. A separate set of samples from the middle Tennessee data sets were selected using a set of sliding maximum collection dates ranging from 2011 to 2015 for the full data sets. For the diagnostic subset, the maximum diagnostic dates ranged from 2007 to 2011. Finally, when a run directly compares the complete Tennessee data set, to the diagnostic subset the more complete data set is filtered. This is done in order to ensure this comparison is not confounded by differences in the size of the validation set. For instance, for the Tennessee data set the number of sequences with a collection date of 2015 (the newest collection year) was reduced to 129 in order to equal the number of cases with a diagnostic date of 2013 (the newest diagnostic date).

Chapter 4

Results

4.1 Genetic variation in populations

4.1.1 Pairwise TN93 distances

The pairwise TN93 distances between all sequences were calculated using open source software affiliated with the publication of HIV-TRACE [kPWV18]. The means of the resulting distributions were 0.545, 0.563, and 0.576, for sequences collected in Seattle, Alberta and Tennessee. The portion of this distribution containing distances at or below 0.05 was used in future analysis and is summarized in Figure 4.1. Although these distributions contain a large number of observations and a relatively even skew, normality was not assumed. The Seattle and Alberta data sets appeared normal through a Shapiro test [SF72] on a random sample of 5000 sequence ($p<0.001$), but the Tennessee data set did appear normal to the same extent ($n = 5000$, $p>0.1$). Given this outcome, normality was not assumed for these observations. Because these distances were not normally distributed, pairwise ranked-sum Wilcoxon tests [Geh65] were then used to determine differences between data sets as opposed to a statistical test which requires a normal distribution for data.

Using a pairwise series of these tests, each data set was determined to be differ significantly ($p<0.001$) from all other distributions. The subset of Tennessee data annotated with diagnostic dates, also differed significantly compared to the complete set (Wilcoxon test, $p<0.001$) with a lower mean TN93 distance of 0.0555. Because of this, the individuals in this data set with associated diagnostic dates cannot be assumed to be a truly representative sample of the whole data set with regards to expected divergence between sequences. The highest distance in the range of thresholds used in the analysis of graph-based methods is 0.04 expected substitutions per site. Therefore, distances above 0.05 are excluded from figure 4.1 for clarity as they do not

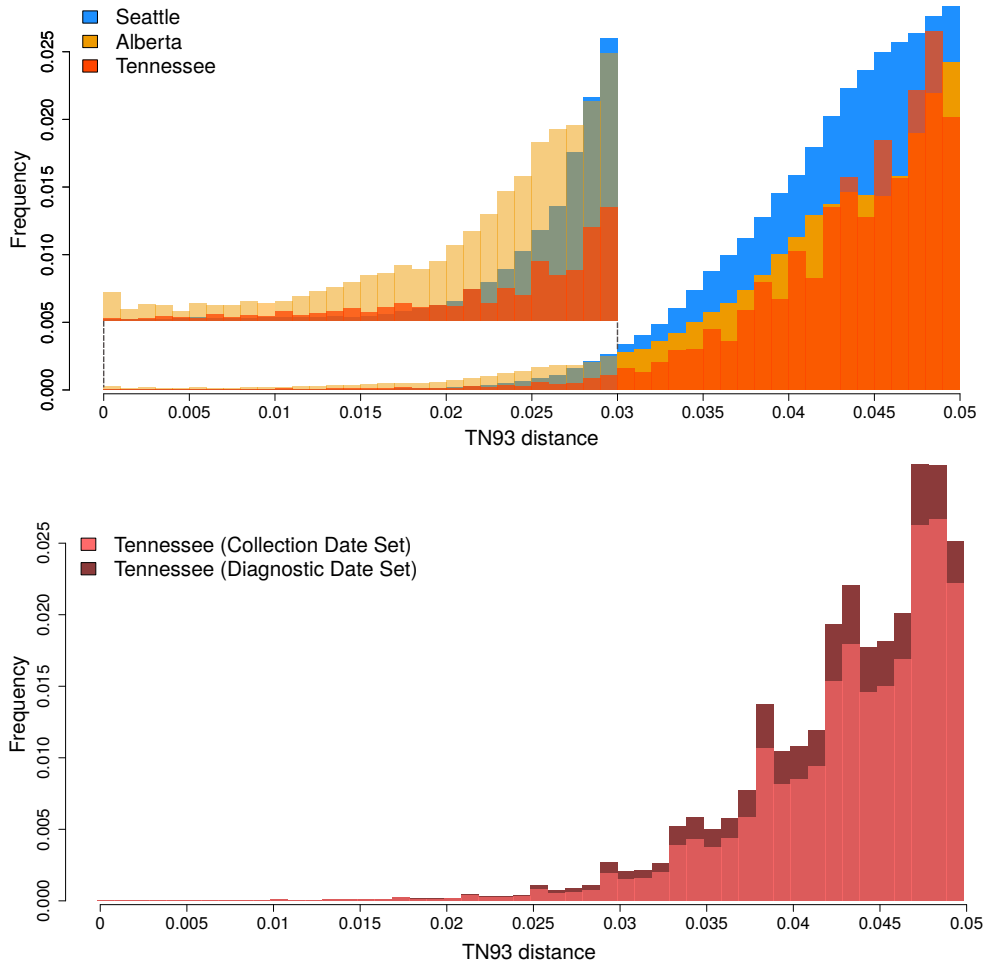

Figure 4.1: **(top)** Histogram, representing the distribution of pairwise TN93 distances for the Seattle (blue), Alberta (orange) and Tennessee (red) data sets of HIV-1 subtype B *pol* sequences. An expanded section of the bar plots in the range (0, 0.03) is provided as a figure inset to clarify differences among the distributions. **(bottom)** Distribution of pairwise TN93 distances for the full data set of HIV-1 subtype B *pol* sequences collected in Middle Tennessee (pink), compared to the subset of sequences with associated diagnostic dates (dark red). The height of each bin has been re-scaled to reflect the total number of pairwise comparisons, for which the majority (above 0.05) were excluded from analysis.

appear in any of the graphs used to define clusters and do not represent any connections that would contribute to cluster growth. These distances were still obtained to analyze the overall average distances calculated above, as well as the normality of distributions and the differences between distributions. Within the Alberta data set, this check of the overall pairwise distance distribution identified an outlier sequence (genbank ID KU190160), with unusually high TN93 distances to all other sequences, ranging from 0.52 to 0.61 expected substitutions per cite.

Upon visual inspection through aliview software [Lar14], this did not appear to be based upon unusual ambiguity or a frame-shift error that could be easily corrected through the addition of gaps and did not effect the rest of the alignment in a significant way. In addition, the use of the COMET subtyping tool [SLT+14], confirmed that this was, in fact, a subtype B *pol* sequence, as opposed to a mislabelled sequence of a different subtype. This individual can have a misleading effect on the overall distribution of mean pairwise distances between sequences, especially given that the Alberta data set contains the smallest number of sequences. For instance excluding this sequence reduces the mean pairwise distance in the data set from 0.0563 to 0.0548. However, given that any sequence above 0.04 is excluded at even the most relaxed thresholds used in the analysis, this sequence is unlikely to have any effect on the following results as it is excluded from any and all clusters. As shown in the highlighted section of the bottom component of Figure 4.1 (top), the Alberta data set has a heavier left-tail compared to the two American data sets, containing a higher number of sequence pairings with a TN93 distance below 0.03. Within the Alberta data set, the 0.1% quantile represented was marked by a distance of 0.005 compared to 0.020 and 0.015 for Seattle and Tennessee respectively. Although this is unlikely to result in any major differences in the overall distribution, it is likely to represent large differences in a clustering analysis, which focuses more exclusively on highly similar pairs of sequences.

4.1.2 Patristic distances in maximum-likelihood trees

Iqtree [NSVHM15] was used to construct maximum likelihood trees from the three data sets using a general time reversible model of evolution [LPSS84] with free rate variation among sites to determine likelihood [Yan95] as well as optimized base frequencies and 1000 iterations of the ultrafast bootstrap algorithm [MNvH13]. Given the large size of these trees, each data set is difficult to visualize in its entirety, however, the figures highlighting specific subtrees within the tree (Figures 4.11,4.12, 4.13, and 4.14), show these clusters in the context of a complete tree. The pairwise patristic distances were significantly larger than the pairwise TN93 distances calculated directly between sequences (Wilcoxon Test, $p<0.001$), with means closer to 0.075. Despite this, the overall branch lengths of each tree suggested no specific tree encountered significant problems during construction, with average branch lengths of 0.0122, 0.0097, and 0.0105 for Seattle, Alberta and Tennessee trees respectively. Long branch lengths overall would indicate that the trees settled on a model where all sequences were unrelated. The average terminal branch lengths (from tips to their common ancestors) were all above these averages, with mean values of 0.020, 0.013, and 0.016. This is often taken as indication that the

virus evolves rapidly within hosts, but less rapidly on a population level and that a less divergent sample of virus is transmitted [LF12]. The poorly aligned sequence from Alberta (KU190160) was effectively represented as an outlier with an extremely long terminal branch length (1.3) in the final tree; this is a visible feature in 4.12. The outlier sequence only rejoins the other descendants at the root of the tree, meaning that it would participate in no clusters, unless the complete tree was labeled a cluster. For each of these trees, terminal branch lengths were significantly greater (Wilcoxon test, $p<0.001$) for the tips which were added on to the fixed tree using pplacer [MKA10], with mean branch lengths of 0.024, 0.017 and 0.027 (Seattle, Alberta and Tennessee) from new tips to newly created nodes. A separate tree was constructed for the subset of the Tennessee data set with diagnostic dates using the same parameters for IqTree 4.14. This differed significantly from the tree constructed from the complete data set, holding an average terminal branch length of 0.017, an average branch length to new tips of 0.024 and an average overall branch length of 0.011 (Wilcoxon Test, $p<0.001$). This is consistent with the differing TN93 branch lengths between the diagnostic subset and the complete set of TN93 distances.

All pairwise patristic distances were calculated for each tree using the *dist.nodes* function within the *ape* R package. The patristic distances between pairs of tips are summarized below in Figure 4.2

The Seattle, Alberta and Tennessee data sets, hold respective mean pairwise patristic distances of 0.079, 0.085, and 0.088 with the largest distances in each set being 0.186, 2.71, and 0.174. The diagnostic subset held a mean patristic distance of 0.082 and a maximum patristic distance of 0.156. All three pairwise comparisons between these distributions resulted in significant differences by a Wilcoxon rank sum test ($p<0.001$). Disregarding, the outlier sequence, the Alberta set holds a mean patristic distance more similar to that of Seattle, (0.078) and a maximum patristic distance of 0.147, which is actually included in the range of maximum patristic distances used for analysis (0 to 0.15). The relative order of these distributions does not contradict the distributions of TN93 distances, with the Tennessee data set seeing the highest divergence overall, and the diagnostic subset representing a more similar set of cases. Further, the relatively left-weighted tail of the Alberta distribution is visible in the distribution of pairwise patristic distances 4.2, with a 0.1% quantile of 0.008, compared to 0.030 and 0.021 for the Seattle and Middle Tennessee distributions.

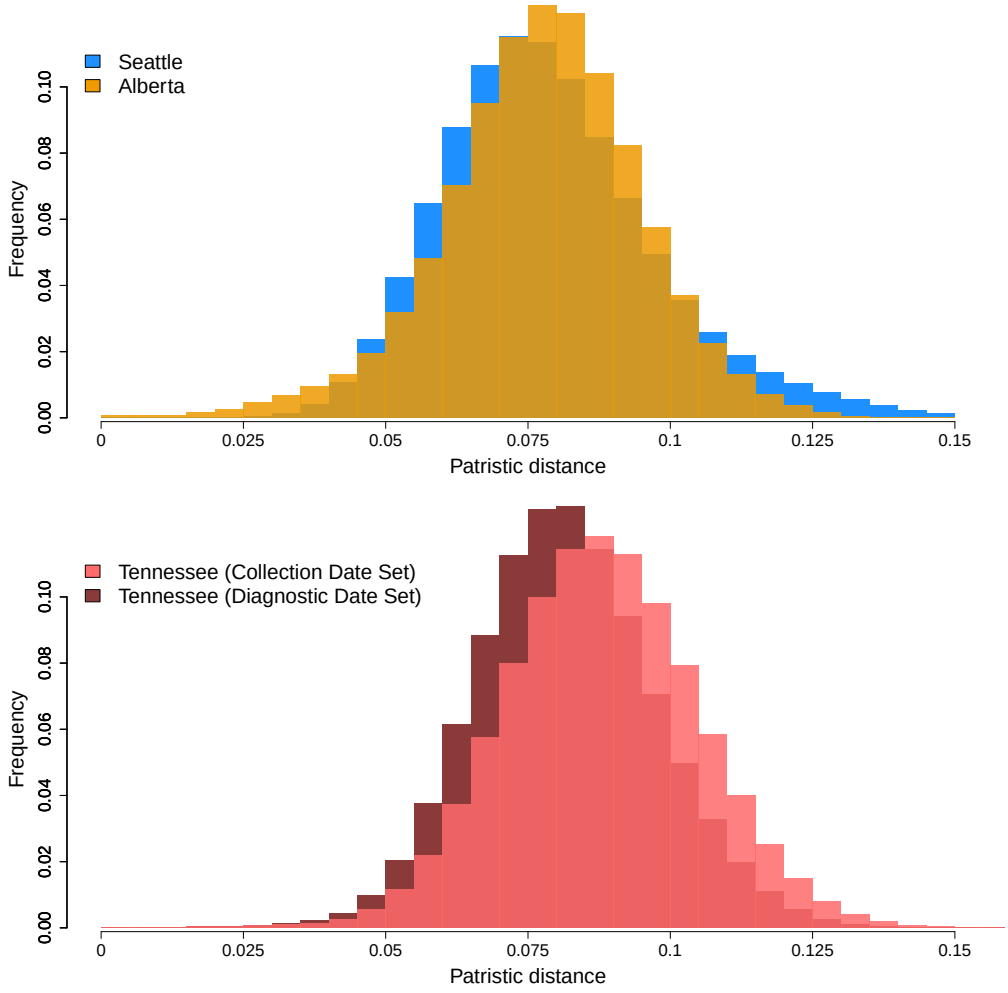

Figure 4.2: **(top)** Histograms representing the distribution of Patristic distances between tips in a maximum likelihood tree made from HIV-1 subtype B *pol* sequences from the Seattle (blue) and Alberta (orange) data sets. **(bottom)** Distribution of Patristic distances between tips in a maximum likelihood tree made from the full set of HIV-1 subtype B *pol* sequences collected in Tennessee (pink) compared to the subset of those sequences with associated diagnostic dates (dark red).

4.2 Time lag affects cluster growth

For both methods, the probability that a pair of sequences are connected is modeled as a function of the time-lag Δt between them. This also captures any change probability that sequences with a more recent time point connect to any member of the subset of new sequences. Any increased probability can inform relatively increased weights of recent cases in clusters, as they increase the likelihood of a connection to a new case. The connections of interest were calculated separately from clusters, in order to specify that a connection which defines growth based

on the closest connection between a new sequence and older sequence, not just a connection beneath a given threshold. For graph-based methods, this stipulation prevents the ambiguous case of one sequence joining multiple clusters simultaneously. For tree-based methods,the inherent tree structure ensures that no one sequence joins two clusters at the same time; nevertheless, the particular node a sequence is closest to still provides information, referencing that tip's location within a cluster. When obtaining an AIC loss measurement between two different models that predict cluster growth, the number of these close connections would change in response to a threshold, thus changing the effectiveness of the proposed model. A greater number of these is more informative as a proposed model based on time is not informative if connections rarely occur regardless of time.

4.2.1 Growth defined by graph-based connections

The connections which represent growth events for graph-based methods are specifically minimum retrospective edges, meaning that the edge of interest from a vertex with some time point t must be the shortest TN93 distance compared to all other edges to vertices with a time point less than t. To review the graph-based model of cluster growth introduced in Chapter 3, the following equation models the number of minimum retrospective edges at a particular time lag $|E_{\text{pos}}(\Delta t)|$, where $|E_{\text{tot}}|$ is the total number of minimum retrospective edges that could occur at that time lag.

$$\log\left(\frac{|E_{\text{pos}}(\Delta t)|}{|E_{\text{tot}}(\Delta t)| - |E_{\text{pos}}(\Delta t)|}\right) = \alpha + \beta\Delta t \tag{4.1}$$

For each data set, the effect of time lag on minimum retrospective edge frequency was viewed with the complete set of pairwise TN93 distances in Figure 4.3 A and B. This figure includes those measured from new sequences which would normally be censored for model training. The effect size (ie. the α in the above equation) for the Seattle, Alberta and Tennessee data sets was then measured as -0.416, -0.402 and -0.235 when using collection date to measure time lag and -0.467 when using diagnostic date in the Tennessee data set. This implies that the log odds of a minimum retrospective edge connection were lower with increasing time lag between cases. Despite these negative trends, time lag between sequences showed no clear effects on overall genetic distances for any of the data sets, which is accounted for by the previous observation that the vast majority of edges have TN93 distances at or above the expected pairwise distance between any two sequences. Sequences with closer collection dates do become more commonly linked when a threshold is imposed upon the graph however.

When excluding distances above 0.015, the proportion of remaining edges between sequences collected or diagnosed in the same year (ie.Δt=0) is 0.192 in Seattle, 0.223 in Alberta, 0.127 in Tennessee, and 0.138 for the diagnostic subset of Tennessee. This compares to proportions of 0.090, 0.178, 0.082, and 0.052 in the complete graph Also important, for the Seattle and Tennessee data sets, the mean time lag for minimum retrospective edges was significantly smaller (Wilcoxon test, sample of 5000, p<0.05) than the mean time differences for all edges in the graph. No significant effect was identified for the Alberta data set, however, likely in part due to a limited range of possible time differences (1-5 years). Unexpectedly, in the diagnostic Tennessee subset, cases diagnosed in 1992, maintained a high degree of connectivity after filtering out edges above a TN93 distance of 0.015. Per sequence, an average of 3.39 of these high-similarity edges connect to individuals diagnosed in a year other than 1997 This compares to a much lower average of 1.27±0.81 across all other years, potentially over-representing the frequency of growth for cases from this year.

4.2.2 Growth as defined connections in maximum likelihood tree

For tree-based methods, an instance of growth is represented by the placement of an individual tip onto the tree. This means that "direct ancestor" nodes for a given tip are counted in the place of minimum retrospective edges, as they represent the most immediate internal node associated with the sequence - the closest location with respect to the rest of the tree. These nodes have an associated time lag Δt, which is calculated between the tip and the other descendant of its direct ancestor. If that other descendant is a subtree as opposed to a single tip, the average time of all tips in the subtree is taken for this calculation. To review the tree-based model of cluster growth introduced in Chapter 3, the following equation models the number of direct ancestor nodes with a particular time lag $|N_{\text{pos}}(\Delta t)|$, where $|N_{\text{tot}}|$ is the total number of direct ancestor nodes that could occur at that time lag.

$$\log\left(\frac{|N_{\text{pos}}(\Delta t)|}{|N_{\text{tot}}(\Delta t)| - |N_{\text{pos}}(\Delta t)|}\right) = \alpha + \beta\Delta t \tag{4.2}$$

For each data set, the effect of time lag on direct ancestor node frequency was viewed for each tree as shown in Figure 4.3 B and C.

Because these trees were built before the addition of new tips, the newest sequences were excluded from this data. The trees built from Seattle, Alberta and Tennessee sequences had a total of 1,014, 478, and 1,696 possible direct ancestor nodes respectively, with the diagnostic subset holding 1,342. This number varies depending on how often multiple tips share the same

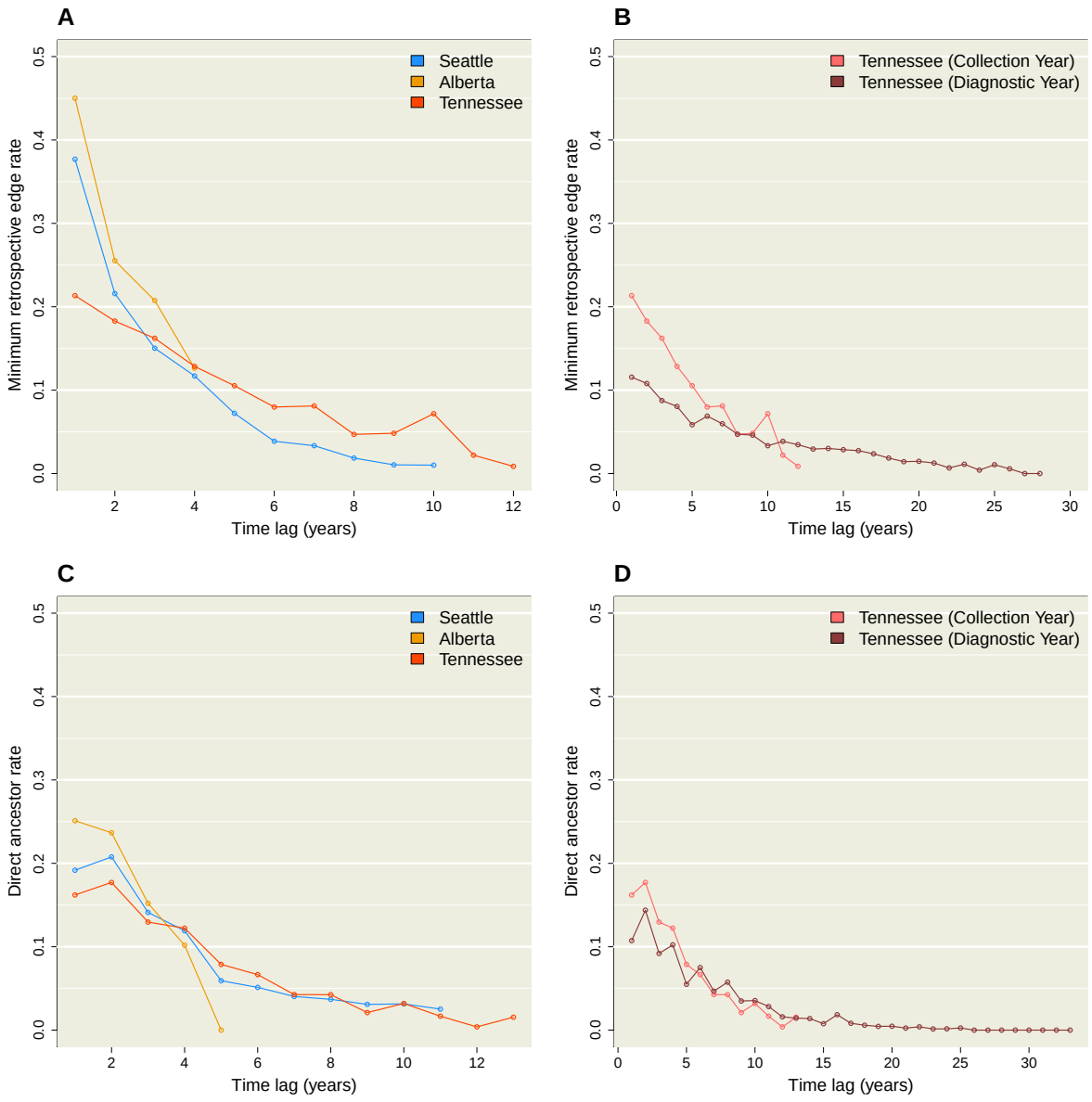

Figure 4.3: **(A)** Minimum retrospective edge frequency with respect to time lag for the Seattle (blue), Alberta (orange) and Tennessee (red) data sets. This is calculated as the number of minimum retrospective edges with a given time lag, over the number of possible minimum retrospective edges with that time lag. **(B)** Minimum retrospective edge frequency with respect to time lag for the diagnostic subset of the Tennessee data (red) compared to the full set using collection dates (pink). **(C)** Direct ancestor node frequency with respect to time lag for the Seattle (blue), Alberta (orange) and Tennessee (red) data sets. This is calculated as the number of minimum retrospective edges with a given time lag, over the number of possible minimum retrospective edges with that time lag. **(D)** Direct ancestor node frequency with respect to time lag for the diagnostic subset of the Tennessee data (red) compared to the full set using collection dates (pink).

direct ancestor. The effect size (ie. the α in the above equation) for the Seattle, Alberta and Tennessee data sets was then measured as -0.227, -0.315 and -0.266 when using collection date to measure time difference and -0.189 when using diagnostic date in the Tennessee data set. This implies that the log odds of a direct ancestor node were lower with increasing time lag between the two descendants. These odds decayed with similar consistency to the graph-based outcomes, with the Alberta data set appearing visibly steeper than the others due to its relatively short time frame and lower number of cases. Also consistent with the graph-based method was the lack of a relationship between the time lag for a pairs of tips and the patristic distances. However, a filtering step once again illustrates the relatively high number of closely related sequences collected in the same year. When a filter is used to only consider only pairs of sequences with a patristic distance below 0.015 the proportion of these pairs with a time lag of 0 increases from 0.090 to 0.243 for Seattle, from 0.179 to 0.262 for Alberta, from 0.081 to 0.148 for the full set of Tennessee data and from 0.052 to 0.194 for the diagnostic subset of the Tennessee data. The time lag associated with immediate ancestor nodes was significantly smaller than the time lag between all tips in the tree (Wilcoxon test, sample of 5000, $p<0.001$)) for all trees, including that which was made from sequences collected in Alberta.

4.3 Effect of cluster threshold

4.3.1 Cluster frequency

Various TN93 thresholds were imposed upon the training partitions of each data set to form clusters using a graph based clustering method. These thresholds act as scaling parameters and their effects are summarized in Figure 4.4.

The TN93 threshold for edges effect the number of clusters created for graph-based clustering methods (Figure 4.4 A). At the lower bound of the TN93 cutoff threshold, almost every individual sequence is considered its own cluster, resulting in a total number of clusters close to the total number of sequences in each data set. The exceptional clusters that do contain multiple sequences are bound together by some TN93 distances of 0, which are particularly frequent ($n = 33$) in the Alberta data set. By comparison, the highest threshold used (a TN93 distance threshold of 0.04) places most cases into a single large cluster. For all data sets, including the diagnostic subset, this resulted in over 90 percent of all sequences placed into a single cluster. Although, this is a higher threshold than what is normally used in this context [APP+12], it does not represent the absolute upper bound for any data set, as no data set reached a point where all sequences were placed into a single cluster.

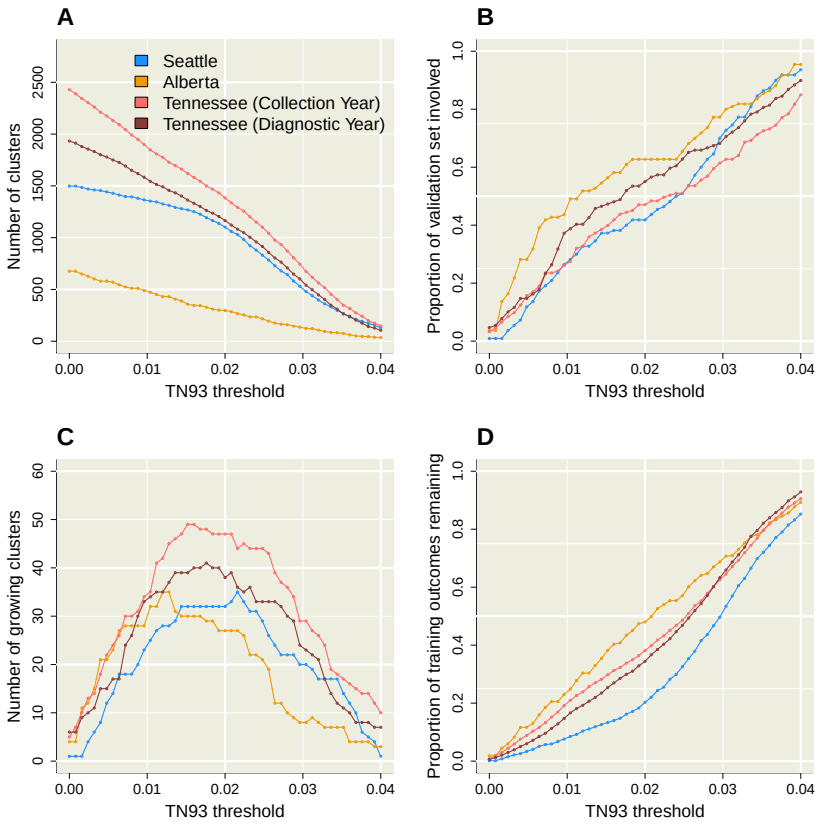

Figure 4.4: Several characteristics of graph based clusters which respond to a change in TN93 threshold. The number of individual clusters (including clusters of size 1) **A**, the proportion of new cases from the data set involved in growth **B**, the number of clusters which experience some growth **C** and the proportion of the training set which is counted as positive (ie. the proportion of minimum retrospective edges below the threshold) **D**.

The TN93 threshold also limits the number of connections which contribute cluster growth (new sequences joining known clusters) and events which act as the basis for training predictive models (minimum retrospective edges and direct ancestors). The proportion of new cases that joined clusters, as well as the proportion of minimum retrospective edges included after filtering the edges, increased steadily in response to the TN93 distance threshold (Figure 4.4 B and D) for all data sets. No data set contained 100 % of either of these edges at the most relaxed threshold, meaning that for all data sets, some new sequences did not participate in cluster growth, and some minimum retrospective edges were never included in the training set. Conversely, at the most strict distance threshold (ie. 0 expected substitutions per site), all data sets experienced some growth outcomes and had several minimum retrospective edges present for model training. However, these vary in how they are distributed across clusters. The number of clusters which are growing reaches an intermediate peak for all data sets, with Alberta's

occurring earliest at a threshold of 0.0120 and Seattle's occurring latest (0.0216), albeit after a long period with no change (Figure 4.4 C). This is another indicator of information content, as this would ultimately correspond to the maximum variance in cluster size and growth.

The same adjustment of clustering threshold was repeated for tree based clustering methods using maximum internal patristic distances for a subtree. Initially, this was done without any restrictions on bootstrap certainty, in order to compare to the effects of the maximum TN93 distance threshold more directly. The trends shown in Figure 4.4 are ultimately repeated in response to the change in maximum patristic distance threshold which specifies subtrees as clusters (Figure 4.5).

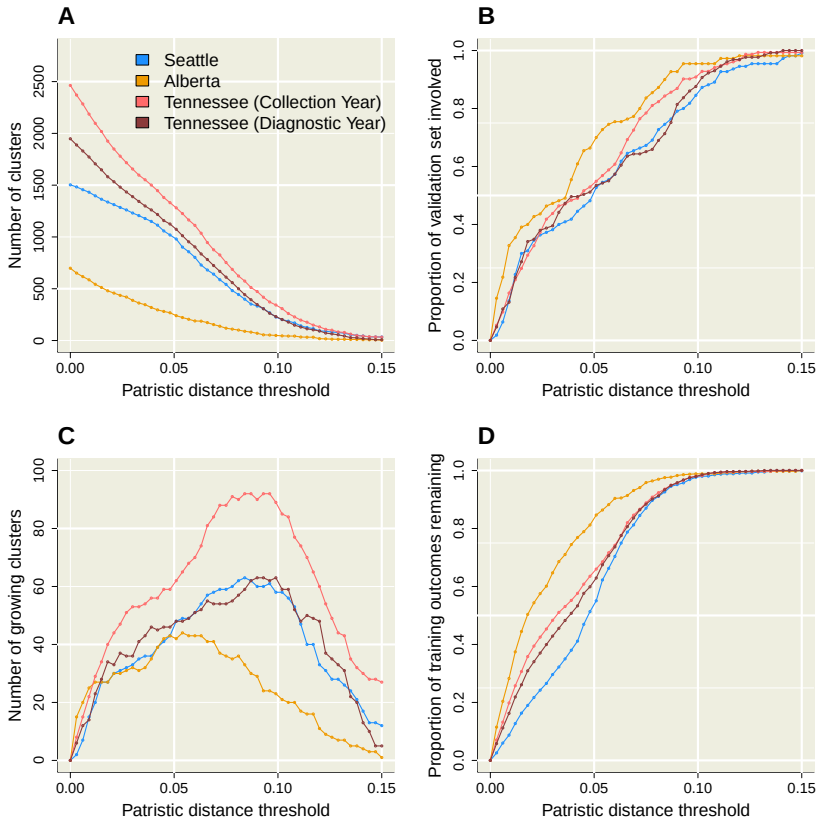

Figure 4.5: Several characteristics of clustering which respond to a change in maximum patristic distance threshold. The number of individual clusters (including singletons) **A**, the proportion of new cases from the data set involved in growth **B**, the number of clusters which experience some growth **C** and the proportion of the training set which is counted as positive (ie. the proportion of direct ancestors below the threshold) **D**

The peak for the number of growing clusters again occurs earliest for the Alberta Data set (at a maximum patristic distance threshold of 0.0144) compared to the other three (0.0224, for Seattle and Tennessee, 0.0240 for the diagnostic subset of the Tennessee data). As is true of

the most relaxed threshold for the graph-based clustering methods, the highest threshold used (maximum patristic distance of 0.15) does not place any data set into a single cluster, as the range of thresholds used does not contain the maximum patristic distances for any of these. Because the maximum likelihood tree construction settings did not allow for true "polytomies" (pairs of tips with no branch length between them), the lowest threshold of 0 placed all sequences in their own cluster. Further differing from the graph-based trends in clustering, the proportion of new sequences connecting to a given cluster (Figure 4.5 B) and the proportion of potential outcomes for the logistic model predicting growth (Figure 4.5 D), increase quickly with more relaxed cutoffs, eventually reaching a point where all new sequences are connected to a cluster and all potential direct ancestor nodes are included in the predictive model training. Furthermore, the majority of this information is present before the greatest number of growing clusters has occurred (Figure 4.5 C), implying that a large number of new cases are dispersed across an appropriately large number of clusters.

The model for tree-based cluster growth does not incorporate bootstrap certainty for either training or growth measurement, so the use of a bootstrap threshold only had an effect on the number of clusters created. With a threshold of 0.90 for bootstrap certainty limiting clusters, the number of cases that were considered singletons (individual sequences in their own cluster of size 1) increased for all data sets, particularly at the largest maximum distance thresholds, where an additional 421, 400, and 278 singletons were created for Seattle, Tennessee and the diagnostic Tennessee subset respectively. Due to the outlier sequence from the North Alberta data set, the highest confidence requirement and most relaxed cutoff threshold effectively divide all sequences in the data set from this single outlier, creating no additional singletons. However, without this outlier, the next largest patristic distance in the Alberta set (0.148) would be included within the clustering threshold allowing for the existence of a single cluster.

4.3.2 Obtaining AIC loss and optimizing threshold

For each clustering method, AIC measurements were obtained for two cluster growth models. The first is a null model (introduced in the previous chapter as $\hat{R}_{null}$) which assumes that all individuals in clusters are equally likely to connect to new cases. The second is a proposed model (introduced in the previous chapter as $\hat{R}_{proposed}$) which assigns higher weights to sequences collected or diagnosed more recently. The relative weighting for newer cases is based off of the effect sizes established in the log-linked training models described in the previous section 4.2 - which measures the same connections that indicate growth as a function of time lag between sequences. For each data set, the loss of AIC was calculated in a reasonable time

frame across 51 different thresholds, with each run of the program finishing in under 10 minutes using modest computational resources - a single 1.6 GHz processor with 8 GBs RAM. This involved the creation of clusters, training of the predictive growth model and growth measurement. The current implementation only requires the creation of a maximum likelihood tree and the calculation of TN93 distances once. Importantly, this time complexity is effected by the number of pairwise edges for graph-based methods, but not for tree-based methods, meaning that the most relaxed thresholds held the largest computational demands in this case. For tree-based methods, the largest computational demands were based on the highest number of clusters, meaning that computation time is shortest for the most relaxed clustering threshold.

The resulting AIC loss calculated by the difference $AIC_{proposed} - AIC_{null}$ is shown in the following figures (4.6 and 4.9) and constitutes the primary outcome of this framework. For the graph-based clustering method, the loss in AIC associated with the proposed model reaches a central minimum for all data sets, corresponding to TN93 thresholds of 0.0152, 0.0104, and 0.0160 respectively (Figure 4.6 (left)).

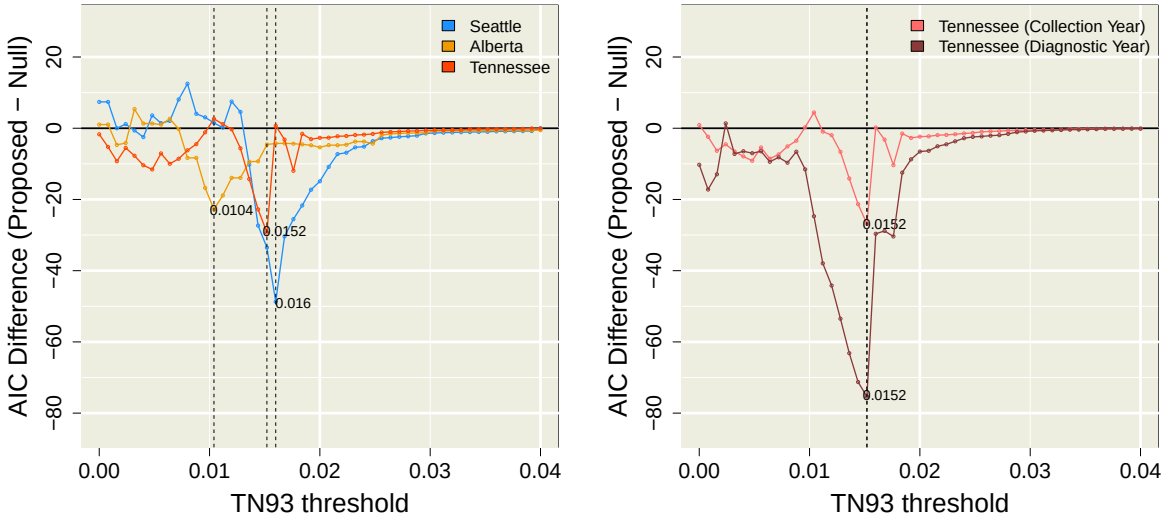

Figure 4.6: The AIC loss for a graph-based predictive growth model in response to the TN93 thresholds used to define clustering. Loss is calculated between a proposed model, which weights clusters more heavily based on the recency of members, and a null model which weights all cases equally The greatest loss in AIC is highlighted. (left) shows the model performance responding to threshold for each location dated with collection dates, while (right) shows this response for the Tennessee diagnostic subset of the compared to the full data set (with the set of new sequences filtered to to only include 129 sequences)

This is interpreted as the point where the additional information provided by case weights

contributes the most. The relative depth (quantity of AIC loss) and location (threshold which produces the largest AIC loss) of these minima is of particular interest, as it illustrates the different optimal scaling parameters for each data set for the purposes of information. The relatively strict threshold leading to the largest AIC loss in Alberta matches loosely corresponds the earlier observations in Figure 4.4 C, where the largest number of growing clusters occurred earlier for this data set. Another important difference is between diagnostic dates and collection dates in the time-based predictive model. For the diagnostic subset of the Tennessee data, the overall profile of AIC loss is consistent with the full set, however the loss is amplified, owing to the relative difference in the effect of time when using recent patient diagnosis compared to recent sequence collection (Figure 4.6 (left)). For all data sets, these profiles of AIC loss change asymmetrically, with strict TN93 thresholds producing stochastic changes in AIC loss to the left of the minimum. Some of these even corresponded to positive AIC differences, suggesting that the use of a weighted model acted as a misleading predictor of cluster growth and the rare connections between cases were ultimately driven by chance. For example, a threshold of 0.068 produces a peak AIC gain for the Seattle data set, corresponding to a small set of edges that made it appear as though closer sequence collection dates implied less frequent connections between cases. To the right of these optimal values, the AIC loss approaches 0 more steadily, as the large number of growth cases and the more complete training sets offer more stable predictive models, but suffer from a lower number of clusters and an overall lack of variation in potential cluster membership.

The following figures 4.7 and 4.8 show each set of clusters at their optimum threshold, using open source graphviz software [EGK+01] implementing the Kamada and Kawai algorithm for visualization [KK89]. These optimum cutoff thresholds, reveal key distinctions between large clusters and fast growing clusters showing tangible examples that explain the differences in performance between a model which only considers cluster size and the proposed model, which acknowledges the effect of collection date recency. For example, at the optimum cutoff threshold of 0.016, the Seattle data set shows a large cluster of 28 individuals which only grows by 2 (labeled as *Se1*), while the largest growth of 6, is seen by a smaller but more recent cluster of 10 individuals (*Se2*). This fastest growing cluster has a mean collection year of 2010.5, compared to the larger, yet older cluster, with a mean collection year of 2007.4. Similar situations are visible in all cluster sets at these thresholds, with the largest cluster failing to attain the largest number of new cases. The labels *NA1* and *Tn1* in 4.7 also indicate relatively large clusters which don't grow as much as clusters with a more recent average collection date. In the diagnostic subset of the Tennessee data, these differences are most dramatic 4.8,

where the fastest growing (*Tn_Diag2*) cluster of 58 sequences obtains 6 new sequences with a mean diagnostic date of 2007.1, while the largest cluster of 73 sequences (*Tn_Diag1*) obtains no new sequences, containing a mean collection of 1999.7. The extremely high connectivity of these early sequences could correspond to the unusually high connectivity of sequences from the early 90's that was previously identified for sequences with a pairwise distance under 0.015 expected substitutions per site. These older clusters possibly indicate past outbreaks - transmission chains that are unlikely to connect to new cases.

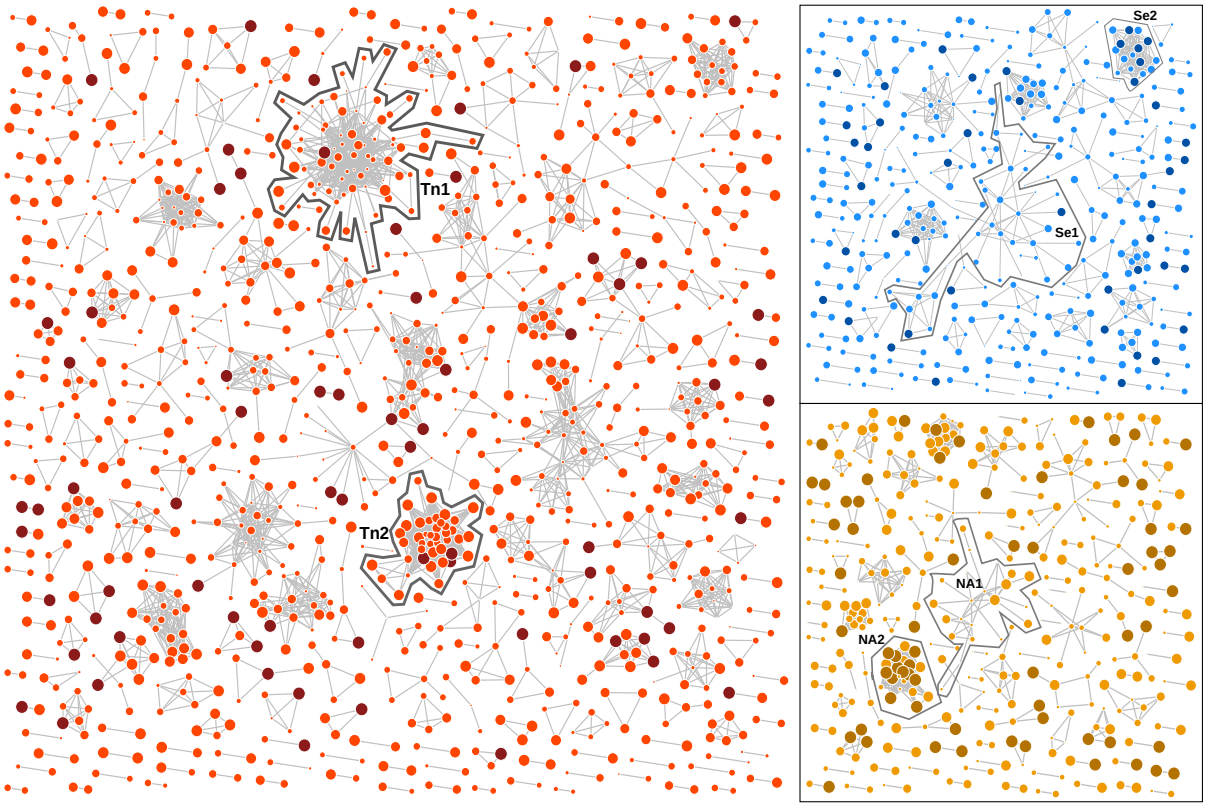

Figure 4.7: Graphs created from each data set at the optimal TN93 threshold parameters. 0.016 for Seattle (blue), 0.0104 for Alberta (orange), and 0.0152 for Tennessee (red). Relative sizes of dots indicate how recently sequences were collected. Darker dots indicate new cases. The largest cluster is labelled with an identifier and a 1 (ex. id1) and the cluster which obtains the most new sequences is labeled with an identifier and a 2 (ex. id2) for each data set. Clusters of size 1 are excluded for clarity.

The corresponding profiles for tree-based clustering methods are more complex - in part owing to a much larger step size (0.003 vs. 0.0008) and a much wider range of scaling parameters being explored (0 to 0.15 vs. 0 to 0.04). It then appears less specific which maximum distance threshold is optimal for these methods. These profiles are detailed in the following figure 4.9, with the largest AIC loss highlighted.

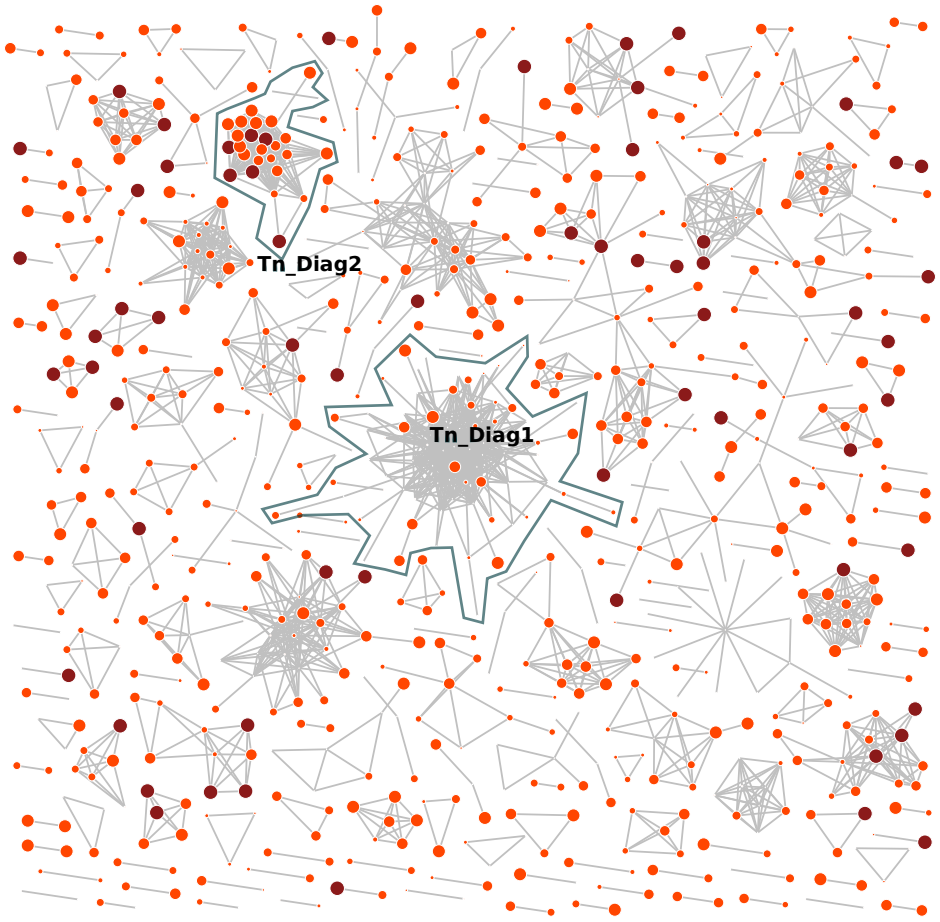

Figure 4.8: A graph created from the subset of the Tennessee data set with diagnostic dates at the threshold for TN93 distance (0.0152). Relative sizes of dots indicate how recently the patient associated with the sequence was diagnosed. Darker dots indicate new cases. The largest cluster is labeled with an identifier and a 1 (ex. id1) and the cluster which obtains the most new sequences is labelled with an identifier and a 2 (ex. id2). Clusters of size 1 are excluded for clarity.

Because these profiles have a less clear minimum value, they are compared to a control, where a random model is used to weight individual sequences (Figure 4.10). This is ultimately done to ensure that the differences between the null and proposed model for tree based methods are not simply due to random chance.

The random model leads to regular fluctuations between positive and negative AIC difference values with more thresholds providing a situation where a random model is outperformed by the null model. By comparison, the proposed model is more structured with larger negative components, however, there is no clear asymptotic relationship with 0, as all sequences join the same cluster. For the Seattle data set, the AIC loss falls to its largest negative value at a

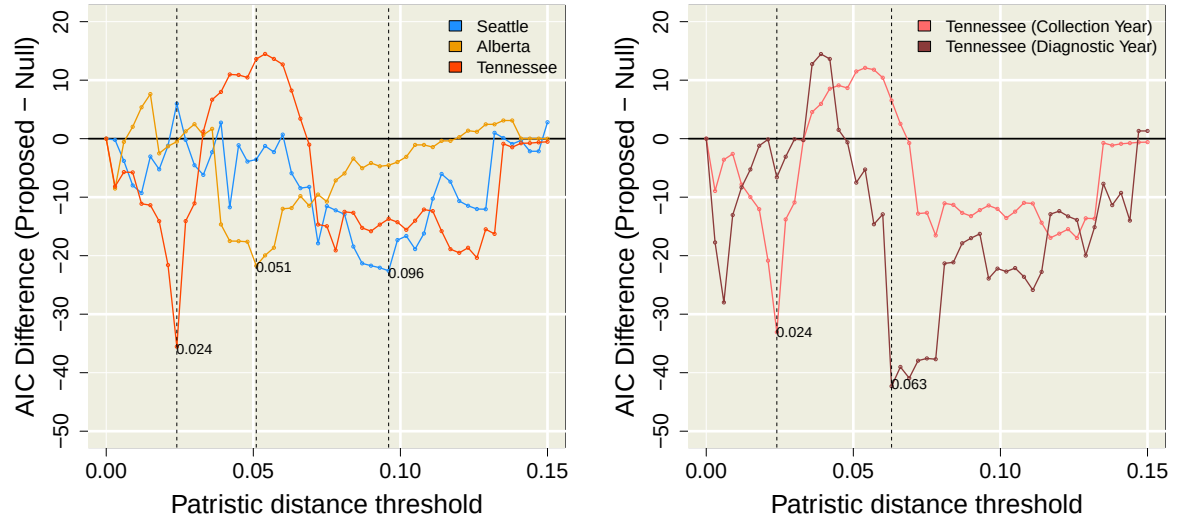

Figure 4.9: The AIC loss for a tree-based predictive growth model in response to the Maximum patristic distances thresholds used to define clustering. Loss is calculated between a proposed model, which weights clusters more heavily based on the recency of members, and a null model which weights all cases equally The greatest loss in AIC is highlighted. (left) shows the model performance responding to threshold for each location dated with collection dates, while (right) shows this response for the Tennessee diagnostic subset of the compared to the full data set (with the set of new sequences filtered to to only include 129 sequences).

relatively high maximum patristic distance threshold of 0.096, marking the center of a wide range of stable AIC loss values from 0.072 to 0.129. The Alberta data set sees this optimal threshold at a lower maximum distance of 0.051 and sees a similar magnitude of loss to Seattle (-22 vs. -23) before rising gradually to a value of 0. The Tennessee data set was noteworthy for a much earlier optimum than the other two data sets - inconsistent with the relaxed optimal parameters seen in the graph-based clustering methods. After a significant portion of positive AIC differences (ie. poor threshold choices) from 0.033 to 0.066, a second area of more consistent negative values occurred, with a minimum loss of -19 at a maximum patristic distance of 0.12. For the subset of Tennessee data with diagnostic dates, a similar sharp, local minimum value occurs at 0.006 before a brief region of positive differences and much more prominent minimum occurring at a maximum distance of 0.063 figure 4.9 (right). Interestingly, these two characteristics in the profile of the Diagnostic subset occur at earlier scaling parameters when compared to the full data set and appear to span a shorter span of threshold values. This loosely corresponds to the relatively lower and narrower distribution of pairwise patristic distances seen in (Figure 4.2 (bottom)). The AIC values for both the null model and predictive

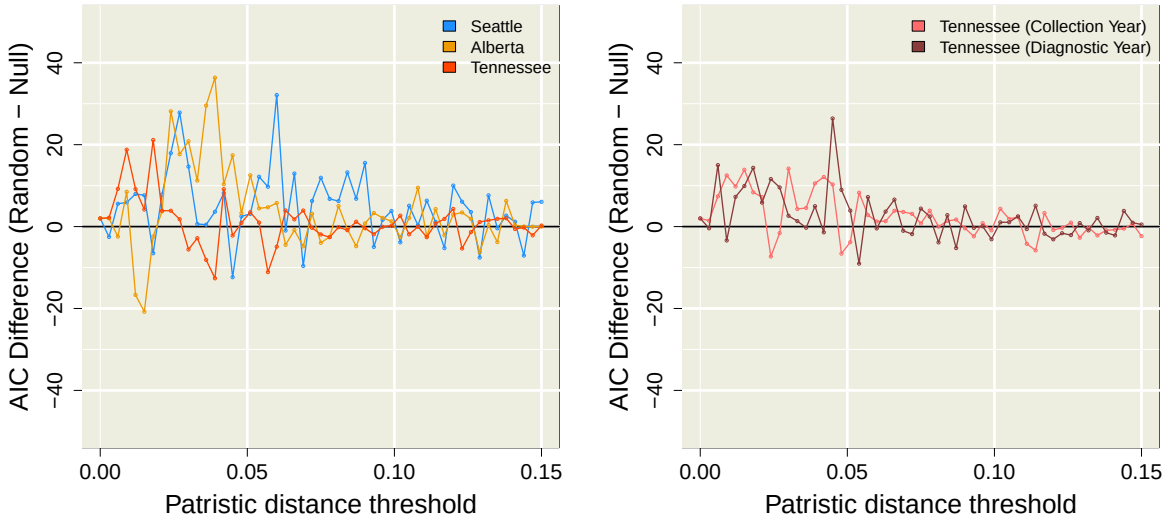

Figure 4.10: The AIC loss for a tree-based predictive growth model in response to the Maximum patristic distances thresholds used to define clustering. Loss is calculated between a random model, which weights individual cases randomly with a mean of 1±0.25, and a minimum of 0 and a null model which weights all cases equally The greatest loss in AIC is highlighted. (left) shows the model performance responding to threshold for each location dated with collection dates, while (right) shows this response for the Tennessee diagnostic subset of the compared to the full data set (with the set of new sequences filtered to to only include 129 sequences)

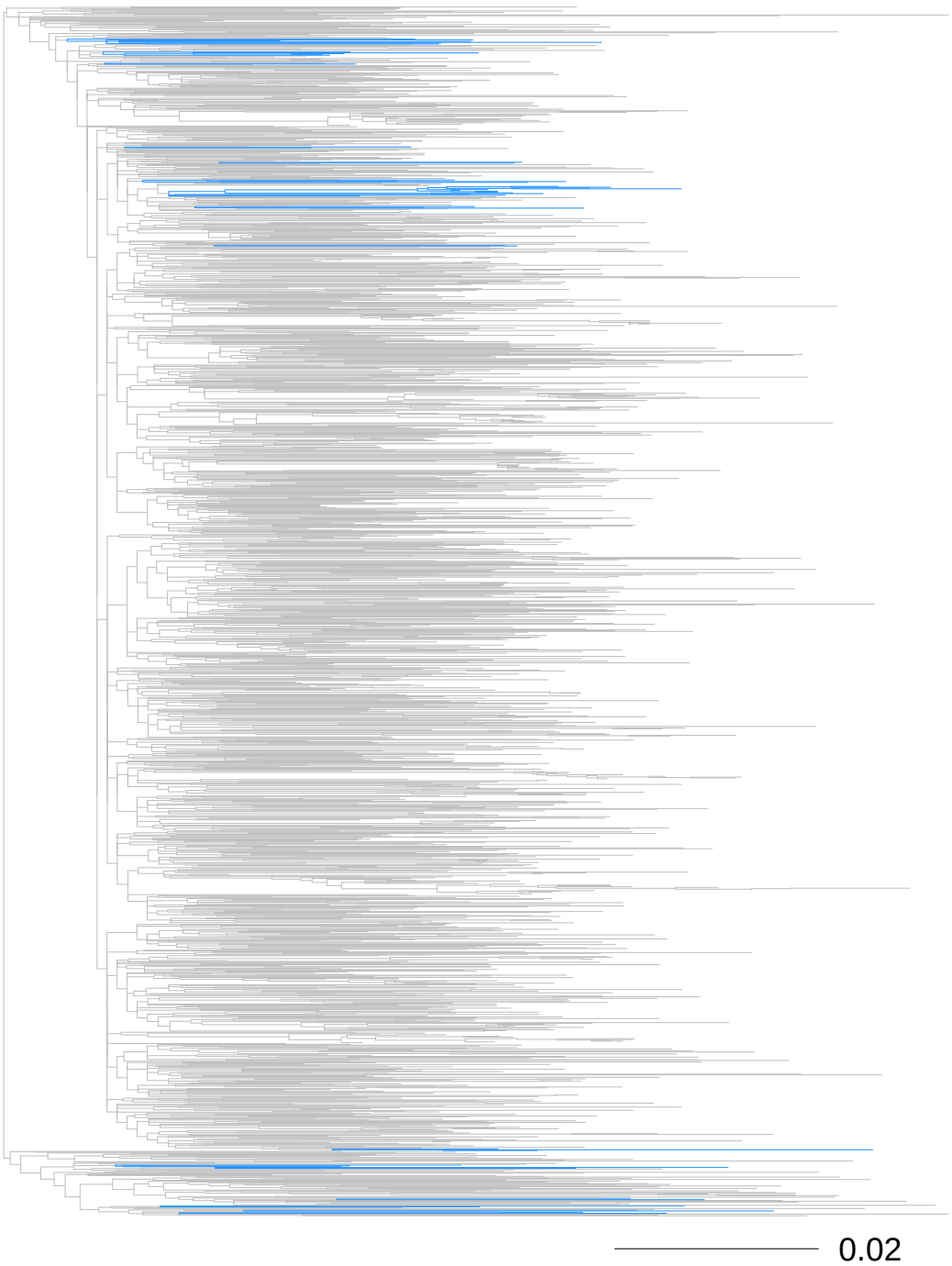

Figure 4.11: The complete maximum likelihood tree constructed from HIV-1 B *pol* sequences collected from patients in Seattle, USA. Specific subtrees within it are highlighted to show the extent of important cluster formation using the optimized maximum patristic distance threshold (0.096). Blue highlighted regions indicate the 20 clusters in the data set which obtain more than one new case.

Figure 4.12: The complete maximum likelihood tree constructed from HIV-1 B *pol* sequences collected from patients in Northern Alberta, Canada. Specific subtrees within it are highlighted to show the extent of important cluster formation, using the optimized maximum patristic distance threshold (0.054). Orange highlighted regions indicate the 14 clusters in the data set which obtain more than one new case. Due to highly divergent sequences, branch lengths are limited at 0.06.

Figure 4.13: The complete maximum likelihood tree constructed from HIV-1 B *pol* sequences collected from patients in Nashville and Surrounding Area, USA. Specific subtrees within it are highlighted to show the extent of important cluster formation, using the optimized maximum patristic distance threshold (0.024). Red highlighted regions indicate the 9 clusters in the data set which obtain more than one new case.

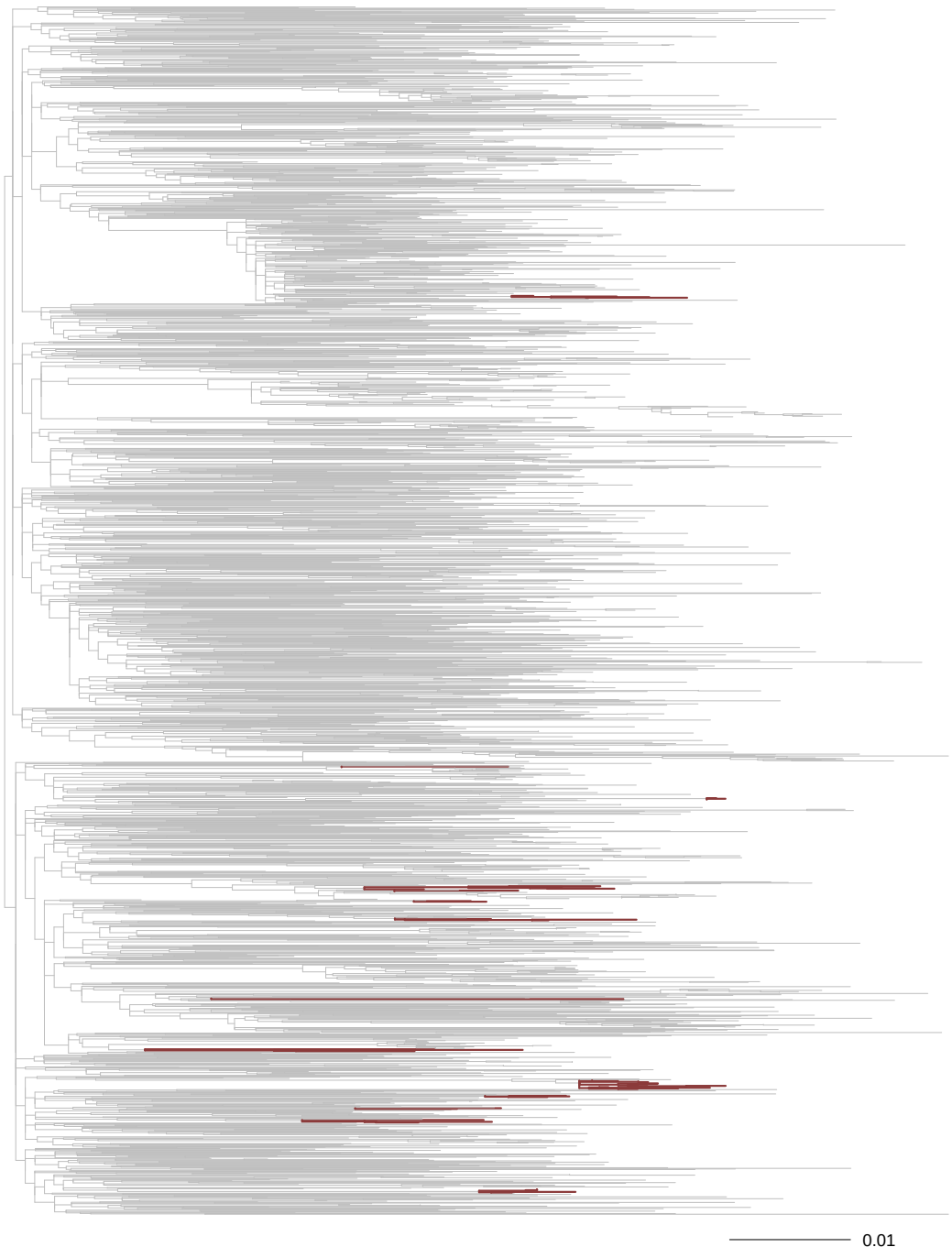

Figure 4.14: The complete maximum likelihood tree constructed from HIV-1 B *pol* sequences collected from patients in Nashville and Surrounding Area, USA. Specific subtrees within it are highlighted to show the extent of important cluster formation, using the optimized maximum patristic distance threshold (0.063). Red highlighted regions indicate the 16 clusters in the data set which obtain more than one new case.

model are higher for these tree based methods, owing to the fact that these simulations are dealing with a much higher number of large clusters, and much more dispersed growth 4.5.

At an optimum value of 0.096, the Seattle data set has a similar situation to its optimal set of graph based clusters, where the maximum growth is seen by a recent cluster of 12 individuals with an average sequence collection date of 2009.2, despite the existence of 35 clusters with a larger size. Although all cases are arguably somewhat recent within the Alberta data set due to the limited time range, the cluster which obtains the highest number of cases (11) under the optimum threshold of 0.051 is again smaller than the largest cluster over all (10 vs. 25) while ultimately being comprised of more recent cases (mean collection year of 2010.5 vs. 2009.5). For the Tennessee data set, the initial optimum seen at a maximum patristic distance reveals a highly recent cluster: 4 individuals with mean collection date of 2012.2. This cluster obtains 3 new cases, and the 3 largest clusters (sizes 16, 11 and 10) have less recent collection dates and obtain no cases. At a maximum patristic distance of 0.053, a different set of clusters is obtained for this data set, which indicates a much lower importance for the sequence collection date as much less recent clusters begin to obtain new cases by virtue of their size - this corresponds to brief range of positive values seen in the profile of AIC loss values. The prioritized cluster corresponding to multiple new cases at optimal thresholds are highlighted within the context of the complete tree in figures 4.11, 4.12, 4.13 and 4.14. At optimal thresholds, the membership of these clusters ultimately differed between the tree based and graph-based methods, especially in the case of the Seattle data set, where this optimal threshold differs so drastically. For the Seattle data set, only 16 percent of sequence pairs which shared a cluster under one of the clustering methods at optimal thresholds, shared a cluster in both methods at their respective optimal thresholds. The equivalent measurements for the other two data sets are 34 percent for Alberta and 28 percent for Tennessee. These differences are ultimately due to a much larger proportion of cases joining clusters under the tree-based method, compared to the graph-based method at their respective optimal thresholds.

4.3.3 Robustness and further optimization

The TN93 distances are a fast and independent measurements which allowed for the assessment of how robust these optimal parameters are to subsampling and use over time. This involved the repeated recalculation of AIC loss for graph-based predictive clustering models using multiple random resamples of the full data sets without replacement. Because of its large size, and consistent number of sequences sampled per year the Tennessee data was used for these robustness tests. The full data set in Figure 4.15 was compared to the diagnostic subset

in Figure 4.16 after filtering a number of cases from the new year such that each had an equally large set of new sequences ($n = 129$). This random sub-sampling (without replacement) shows the occasional movement of the optimum distance threshold, with an interquartile range of 0.0144 to 0.0168 for the full Tennessee data set and 0.0152 to 0.0168 for the diagnostic subset. A smoothed spline function was used to obtain a trend based on the AIC loss values for each subsample at each threshold. This function obtained its highest negative value at a threshold of 0.0152 for the complete data set and 0.0160 for the diagnostic subset. Interestingly, the use of this smoothed function obtains a different optimal value for the diagnostic subset compared to then single complete run (0.0152). This process also illustrates the relative vulnerability to stochastic changes associated with the earliest section of the plot, with a visibly wider set of AIC difference values for thresholds below 0.012 visible in each data set. Finally, the different sampling proportions reveal an important characteristic about the magnitude of AIC loss for both diagnostic data and collection date data, as the overall amplitude of these runs decreases with smaller sampling proportion, due to a loss of information used to train and validate the predictive models.

An additional series of runs were performed on subsets of the Tennessee data set with subsamples and specific time ranges in mind. This series of runs progressively right-censored the range of case collection dates from a maximum year of 2015 to a maximum year of 2011 and the range of diagnostic dates from a maximum year of 2011 to a maximum year of 2007 (Figure 4.17). The TN93 Thresholds which obtain the largest AIC loss vary based on this time range, from 0.0136 to 0.0176 for the diagnostic subset and from 0.0112 to 0.052 for the complete data set. These values are well outside of the interquartile range established earlier in figures 4.15 and 4.16. No particular trend can be claimed to be associated with a shortening time frame, although based on the effects of smaller sample size, the depth of the AIC loss would be expected to decrease, as was clear in the smaller random sample sizes for each data set. Interestingly, the complete set of Tennessee data saw the smallest maximum depth of AIC loss when the data set was most complete (Figure 4.17 (red)), implying that the inclusion of cases with the most recent collection date decreases the difference in performance between the proposed and null models.

Finally, bootstrap support was used to explore the potential of further optimization for the tree based methods, although as discussed in chapter two, bootstrap support does not act as a scaling parameter under the strictest definitions. A bootstrap support threshold of 90 percent had the effect of flattening the AIC loss profiles explored in Figure 4.9, limiting the return to an AIC loss of 0 for the Seattle and Tennessee data sets (Figure 4.18 (right)). This corresponds

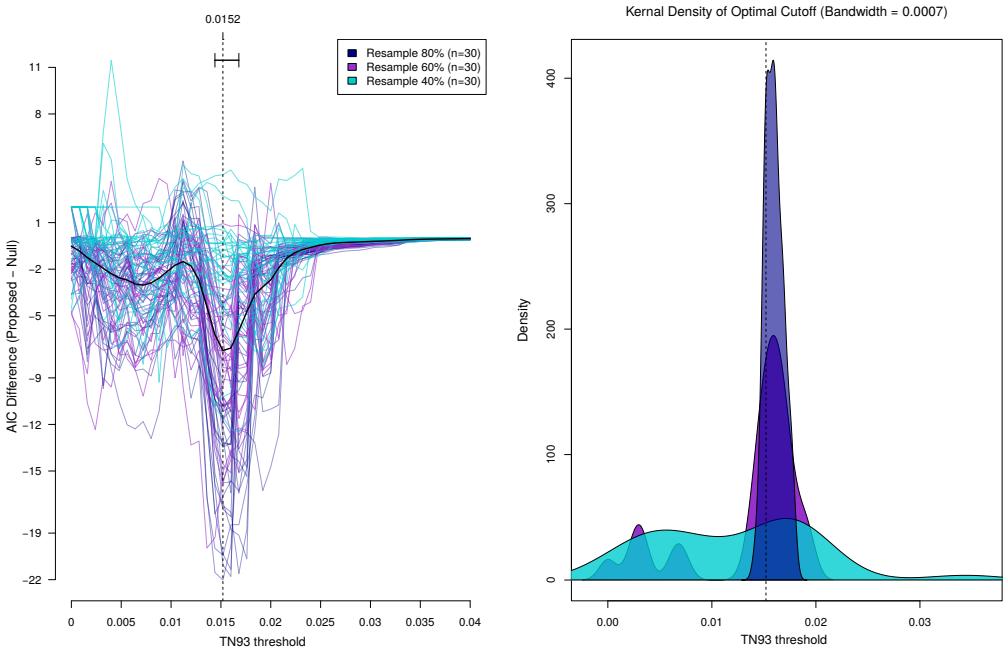

Figure 4.15: **left** The AIC loss for a graph-based predictive growth model in response to the TN93 thresholds used to define clustering. Loss is calculated between a proposed model, which weights clusters more heavily based on the recency of members, and a null model which weights all cases equally. 30 random draws of 3 different sample sizes were taken from the full Tennessee data set and run. A smoothed spline function (black) calculates the general trend and the minimum value of this function is highlighted. The interquartile range for the threshold which obtains the largest AIC loss is also highlighted. **right** The kernal density function for the location of the highest AIC loss

to many large clusters becoming unable to collapse further due to their dependence on an uncertain parent node. This effectively offers the ability to keep sequences separated into a higher number of clusters, while still maintaining a high proportion of the set of new sequences involved in growth. This also allows for a high proportion of direct ancestor nodes to be included in the data set. Given this bootstrap requirement, the minimum number of clusters for Seattle, Alberta and Tennessee are 702, 110, 728 respectively. For Seattle this appears to allow for a larger AIC loss than either the highlighted local minimum at 0.054 or the previously identified minimum of 0.096 can obtain, with the final three threshold values obtaining AIC loss values of -38.

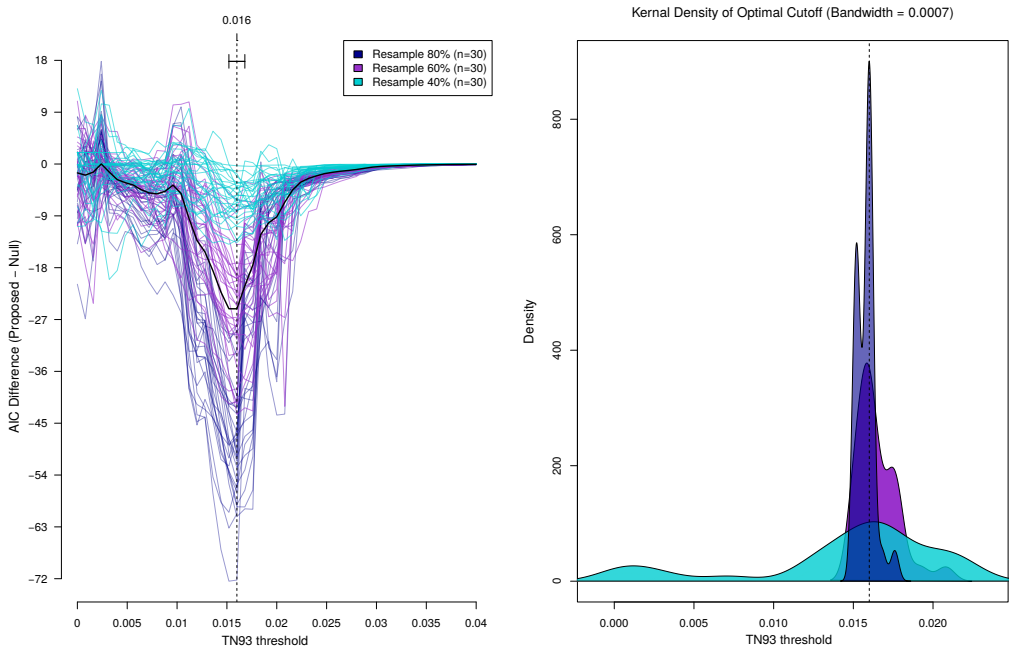

Figure 4.16: **left** The AIC loss for a graph-based predictive growth model in response to the TN93 thresholds used to define clustering. Loss is calculated between a proposed model, which weights clusters more heavily based on the recency of members, and a null model which weights all cases equally. 30 random draws of 3 different sample sizes were taken from the subset of the Tennessee data set with diagnostic dates and run A smoothed spline function (black) calculates the general trend and the minimum value of this function is highlighted. The inter-quartile range for the threshold which obtains the largest AIC loss is also highlighted. **right** The kernal density function for the location of the highest AIC loss

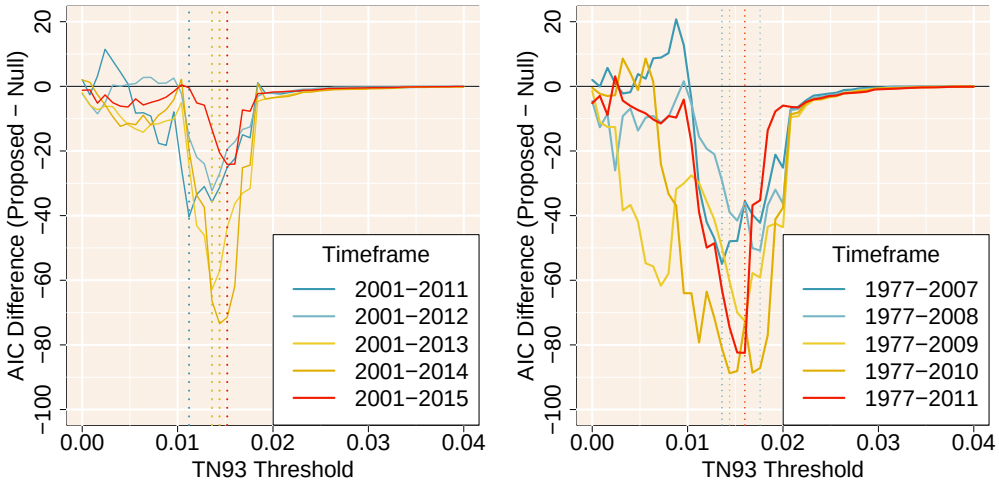

Figure 4.17: The AIC loss for a graph-based predictive growth model in response to the TN93 thresholds used to define clustering. Loss is calculated between a proposed model, which weights clusters more heavily based on the recency of members, and a null model which weights all cases equally. **right** 5 different subsets of the Tennessee data set with diagnostic dates were taken, each representing date ranges with a progressively later final year. **right** 5 different subsets of the Tennessee data set with diagnostic dates, each represents date ranges with a progressively later final year.

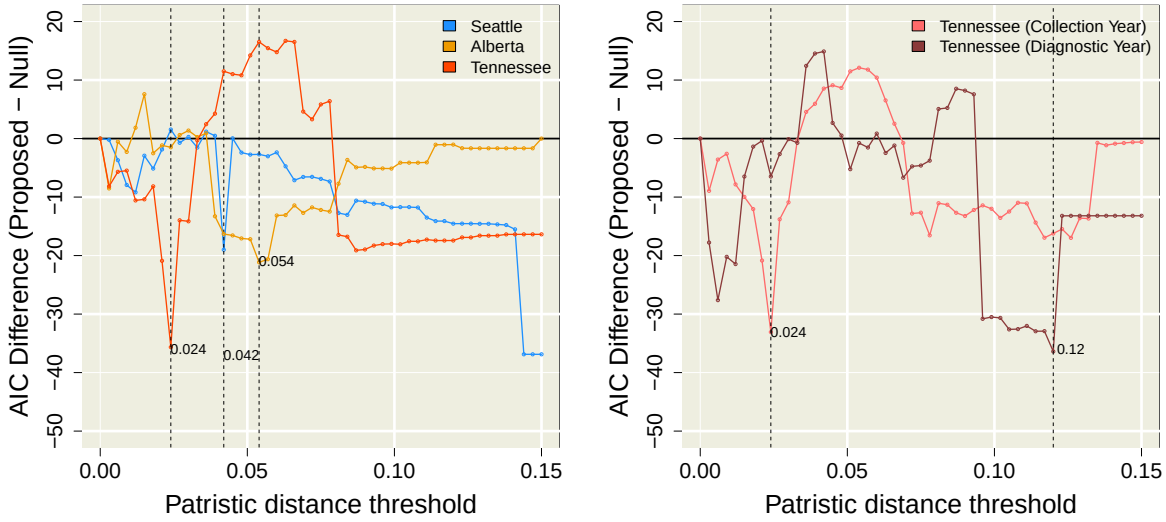

Figure 4.18: The AIC loss for a tree-based predictive growth model in response to the maximum patristic distances thresholds used to define clustering. Trees and clusters are further restricted by a minimum bootstrap requirement of 90 percent certainty. Loss is calculated between a proposed model, which weights clusters more heavily based on the recency of members, and a null model which weights all cases equally The greatest loss in AIC is highlighted. (left) shows the model performance responding to threshold for each location dated with collection dates, while (right) shows this response for the Tennessee diagnostic subset of the compared to the full data set (with the set of new sequences filtered to to only include 129 sequences)

Chapter 5

Discussion

In this chapter, I will summarize the key results of my thesis work, mainly focusing on the implications of differing AIC loss results between different clustering methods and different data sets. The actual threshold which obtains optimum clustering values, as well as the magnitude of AIC loss associated with that threshold are both of particular interest. The metric chosen to measure performance is also specific to the prediction of cluster growth, and the merits of that goal are compared to other traditional uses of molecular clusters, specific to HIV. Finally, this section will discuss the future work necessary to improve upon this framework and make it easier to implement.

5.1 Direct comparisons

The data sets I have analyzed in the previous chapters are associated with three previously published studies, each of which uses molecular clusters of HIV sequences to make suggestions about public health priority. This offers an interesting point of comparison, as the clusters identified in those studies will inevitably differ from those identified here due to the use of different threshold parameters. In addition, the goal of my work differs, as the use of predictive growth models on clusters is not always common in the literature. These studies all aim to treat clustering connections between patients as an indication of direct transmission, while my work defines clusters as an indication that a sub population containing the associated sample of cases may be experiencing an elevated rate of transmission. Ultimately, it would not be appropriate to compare the ”accuracy” of these outcomes, unless they too aimed to predict onward transmission with the associated data. Unfortunately, the study which incorporates the Northern Alberta data set uses a fundamentally different tree-building method based on Bayesian

statistics [YR97, VAB+17]. This is difficult to compare directly, as the branch lengths in their tree are scaled to estimate time, not just the similarity of sequence data [BLD+12]. Both other studies (using the sequence data from Seattle and Tennessee) can be more easily compared to my results, although the methods used to construct maximum likelihood trees differed. These utilized FastTree software [PDA10] which uses heuristics to approximate the maximum likelihood tree instead of converging to it with more certainty. What this tree represents does not differ fundamentally from those that I have built with IqTree [NSVHM15], it simply uses a different algorithm which prioritizes speed and does not explore the possible values for branch length and branching order to the same extent as other more commonly used tree-building methods. This results in a limited maximum precision value for these branch lengths unless additional measures are taken. Another important difference is that neither of these studies use the same quantitative predictive model of clustering that I have demonstrated in chapters 3 and 4 to make statements at the level of populations. Instead, they both focus on more individual connections, utilizing available meta data such as age, injection drug use, race and sexual behavior (particularly whether or not the individual associated with a sequence self-identifies as a man who has sex with other men) to determine whether or not these characteristics are associated with sequence clustering. This means that the definition of clusters is restricted to a size greater than 1, as sequences with no connections and sequences with at least 1 connection are being compared as the positives and negatives in a logistic regression, similar to studies mentioned previously in chapter 1 [DOKG+17, VLVR+18]. Also, because of the retrospective nature, the associations with known clusters does not necessarily indicate a high-likelihood of onward transmission. Even if a population level analysis was done, these clusters would only be indicating what may have driven past outbreaks [LVRD+18]. This was supported by my results, as the largest clusters identified in the previous results were not always associated with onward transmission. Finally, it's important to note that these studies both use slightly larger data sets than what was available for my project, showing sequences from 1,953 individuals in Seattle and 2,915 from Tennessee. This compared to the 1,648 and 2,779 individuals represented in my work.

Keeping those caveats in mind, I will clarify how these previous studies could have obtained additional information through the use of a predictive model at the population level in combination with threshold optimization. In addition, I will reiterate the critiques stated in the literature about how this study goal may be less beneficial to prevention efforts. I will begin with the Seattle data set, which comes from a study by Wolf, et al [WHVR+17]. This study aimed to use strict thresholds to retrospectively show connections that indicate a high likelihood

of direct transmission between individuals. As discussed in chapter 1, there are a number of problems associated with this goal, for instance, the similarity of two individual sequences may indicate infection by the same source, not transmission between hosts [VF13]. The tree-based Cluster Picker [RCHH+13] method was used, with a maximum patristic distance of 0.015 and a bootstrap requirement of 0.95 to define a set of 42 clusters with a size greater than 1, representing a total of 168 individuals (8.6% of the data set). By comparison, the optimal maximum patristic distance I identified for tree-based clustering was 0.096, which identifies 277 clusters of size two or greater, even when constrained by a confidence requirement of 0.95. These clusters represent the sequence data of 918 individuals - more than half (61%) of the total training set in my results. This suggests that their thresholds limit the range of cluster sizes, resulting in mostly the clusters of size 2 (”dyads”) that were discussed as potential transmission pairs. Similarly to another case study by De Olivera , et al [DOKG+17], the primary outcome of this study was the unusually high rate of adolescents (age 13-24) connecting with non-adolescents (mean age 34 years), suggesting that interventions should be aimed at age-discrepant pairs due to the apparent regularity with which they were observed. The highly strict clustering threshold used by Wolf identifies connections that are very unlikely to appear by chance compared to those identified in my work, however this still does not reliably imply direct transmission For instance, their use of multiple sequences from the same patient does not support these pairs consistently. The primary phylogeny was constructed with the first sequence collected from each individual, however a second phylogeny was constructed with the same method using all sequences. This allowed multiple sequences from the same individual to exist in the tree, with the expectation that individual sequences that formed clusters with the first sequence sampled from a given host would form clusters with all sequences sampled from that host. The paper states that only 24 of the 42 clusters identified in the first tree were also present in the second, indicating that clustering between sequences can be dependent on within-host evolution of the virus [LF12, LRP06]. This implies that theoretically, real transmissions between patients may not appear consistently depending on whether or not sequences were taken closer to the time of infection, clarifying the importance of fast sampling time for this form of molecular cluster analysis.

For the original study using the Tennessee data set [DVF+18], Dennis, et al interpreted clusters using a similar assumption about connections (ie. that they representing transmission events), although their primary results were not as dependent on this assumption. Interestingly, the clusters in this study were not defined as subtrees with a maximum branch-length requirement similar to the Cluster-Picker method. Dennis, et al used single-linkage graph-based

methods to define clusters, with a maximum distance of 0.015 identifying a connection between sequences. However, the pairwise distances were calculated using the patristic distances from FastTree, instead of the TN93 distance calculator used as a component of HIV-TRACE [KPWLBW18]. In this case, a much greater proportion of the study population was linked to clustering compared to the Seattle study, with 1113 individuals in 292 clusters. This is associated with the single linkage requirement for graph-based clustering methods. At the optimal threshold obtained in my work for graph-based methods (TN93 distance threshold of 0.016), a total of 1205 individuals were sorted into 259 clusters of size 2 or greater using a graph based method. This resulted in slightly larger clusters overall, with the largest cluster containing 86 individuals at the optimized threshold compared to 39 individuals at the threshold chosen in Dennis' study. Going beyond cluster size however, Dennis focused on identifying clusters with at least one sequence collected between the years of 2011 and 2015, as well as any associated meta-data such as the over-representation or under-representation of a particular risk factor in these clusters. While this is more indicative of onward transmission likelihood, the window for what was defined as recent was relatively large compared to the indications of recency used in my study (5 years), with roughly a third of all clustered sequences (32%) considered recent. It is also based on collection date, which may vary significantly from when infections took place, especially given that this study had such a wide range of diagnostic dates. Given that the primary results of this study were not as interested in transmission pairs as Wolf, the relatively high number of larger clusters is helpful for this study as it captures a larger percentage of the individuals enrolled in the study. However, the model in question classifies priority clusters based on whether or not they contain recent cases instead of quantifying how many recent individuals they contain, making it difficult to compare the benefits of information content. The question of whether or not a cluster will attain any recent cases is very different from how many recent cases will a cluster obtain, and in large epidemics with large variations in cluster growth, the latter helps establish a better understanding of priority [LVRD+18].

5.2 Scaling parameter response

For both clustering methods, the threshold for proximity measures (TN93 distance and patristic distance) acts as a scaling parameter for the modifiable areal unit problem [FW91]. As these thresholds were relaxed, each data set was divided into fewer clusters, with both tree-based and graph-based clustering methods showing similar relationships between cluster number and threshold. The upper and lower bounds of scale are not reached for all runs of the framework,

however, the range of thresholds used in this study expands beyond the range of thresholds used in the field [VLVR+18, RLD+17, WPF+17, OFP+18], showing the behaviour of clustering techniques beyond what is seen in practice. At extremely high clustering thresholds, new sequences were not distributed across a large number of clusters leading to uninformative estimates of growth (panels A and C for Figures 4.4 and 4.5). At extremely low thresholds, most if not all individuals constitute their own cluster. Because this threshold value also effects growth outcomes (as shown in panel B and D for Figures 4.4 and 4.5), the accuracy of growth predictions became less relavent than the complete lack of connections to new cases, as clustering connections in general were either extremely rare or not at all present. This demanded a modified version of Nakaya's solution to the modifiable areal unit problem [Nak00] in order to identify central optimal thresholds for the purposes of predictive model performance. Using this original solution, the information content for a set of clusters would be measured by comparing clustered data to a data set with no clustering and therefore, no outcomes. This would lead to a more trivial result, as the information content would increase unidirectionally as the threshold relaxes. The inevitable result is the selection of an "optimal" extreme threshold that shows all possible connections between sequences, resulting in all new cases joining a single, large cluster. Instead, the results of this study show a comparison between two models that both partition the data set into clusters, with the difference being whether or not predictive variables are used to determine the log likelihood of individuals in clusters connecting to new sequences. This means that the scaling parameter which obtains the largest AIC loss represents the point where the use of predictive variables provides the most additional information.

Although the distribution of cluster size is expected to be exponential for both clustering methods [Poo16], the previous comparisons to published work demonstrate that the number of clusters created under a given threshold is a key difference. As the scaling parameter is increased, clusters defined by the graph-based method aggregate together at a much more rapid rate compared to the tree-based methods due to the requirement that only one edge is needed to collapse two clusters into a single cluster [ST06]. This is an important distinction, as a relatively low number of edges in the pairwise graph need to exist before all cases are members of the same cluster. For tree-based clustering, two groups of sequences only come together if all edges between the sequences in each cluster are below a given threshold. Despite this difference, the standard thresholds used for the graph-based clustering methods (a maximum distance of 0.015) are often translated to tree-based methods as a similar default [APP+12, VLVR+18, WMM+18, BPP+19], meaning that tree-based methods are inherently less prone to large clusters, an observation which is also noted by Rose, et all in their use of both HIV-

TRACE [KPWLBW18] and Cluster Picker [RCHH+13] on the same data set [RLD+17]. My results also demonstrate how an additional requirement for certainty also limits the existence of large clusters for tree-based methods. For the purposes of this predictive framework, it was important to treat this separately from the distance parameter, as highly uncertain connections between new cases and old cases are still counted as growth. The placement of a new tip onto a fixed tree through the use of a tool such as pplacer [MKA10], reports the most likely location given a model of evolution. If this likelihood is low, it may be because the tip is not closely related to any other sequence, in which case it is unlikely to be included in any known clusters due to the distance requirement. The alternative is that it is highly related to multiple sequences, implying that this tip should join a known cluster, but clarity is needed as to which. It follows that certainty requirements do nothing to constrain the number of connections used to train or test the predictive model, they only limit the size of clusters. This explains the lack of a central optimum for the runs which were limited by a bootstrap certainty requirement shown in Figure 4.18, as the most permissive distance requirements for these runs would eventually lead to the largest possible clusters allowed by a given bootstrap requirement. Although this was associated with a relatively high amount of stable AIC loss, this state does not necessarily act as an optimum for all data sets. The Seattle data set sees its greatest AIC loss at this point, however, the North Alberta and Tennessee data sets still see optimal parameters well before this state, with their final ”stable state” closer to an even performance between a proposed and null model.

5.2.1 Location of maximum AIC loss

The threshold which produces the maximum AIC loss when comparing between a null model and a proposed model of cluster growth is arguably the most important result of this framework, as it effectively represents the point where our choice of predictor variables is most effective given the data and clustering method. Because of the complex nature of factors which lead to predictive model performance, it is difficult to determine exactly which characteristics lead to a particular optimum value, as there are many which can contribute to predictive model accuracy. However, perhaps the most identifiable factor shown in my study results is expected rate of variation between sequences, which has been well studied with respect to HIV clustering. For instance, a slow sampling rate can allow for a high amount of divergence to occur in the infected population between the time of infection and the time of sampling [LRP06]. A higher threshold may be more appropriate for the detection of clusters in this situation, as the expected diversity between individuals may be higher. Previous work has also discussed the opposite situation,

where samples with a recent infection were over-represented in clusters simply due to the fact that they had been sampled recently [VKW+12]. The proportion of the population which is represented in the sample ("sampling density") has also been described as an important factor in clustering [NML+14, DFGC17]. Incomplete sampling, where only a small portion of the infected population is captured in the study reduces the frequency of clustering possibly demanding a threshold which is more relaxed in order to obtain meaningful clusters. The scale of the study is inherently associated with these factors; for example, observing HIV on a global scale [WLBH+14] would come with the expectation of a much lower sampling density, compared to a more feasible statewide or province-wide surveillance program [NMC+14]. This was observed in this study's results, where the more rural study setting of Northern Alberta [VAB+17] produced different optimal scaling parameters compared to the urban centers of Seattle and Nashville (for the Tennessee data set), in response to the higher number of similar sequences between hosts shown in Figure 4.6 and 4.9. In addition, different rates of diversity are expected depending on the region of the genome being studied, as some locations are able to accumulate a greater number of mutations [NML+15]. Although the *pol* gene is highly available due to its association with drug resistance mutations [Kan06] and displays a high rate of variation [HCCP04], the *env* gene responsible for encoding the spike proteins which surround the viral capsule has even more extensive diversity [LJM+95, MSCB96]. This gene is used less frequently in clustering, however in the context of an incredibly rapid outbreak which is captured early, the additional diversity may be useful for distinguishing between closely related individuals [BKK+01, MWT+90]. Alternatively, a lower threshold could be used to distinguish unusually similar pairs within the context of the outbreak, but identical sequences put a theoretical lower bound on threshold selection for a given gene of interest.

Beyond the effects that study design may have on the expected variation between sequences, the selection of a new threshold may also be necessary for the study of different subtypes of the virus [KGY+01, GIA+04]. Due to the potential for HIV to recombine, extensive variation in sequence data and recombinant subtypes can occur in a location where HIV has been endemic and circulating for an extended period of time. Ongoing HIV epidemics in Africa [BJRP+06, VKR+03, TBS+99] may be challenging to investigate for this reason, especially using a threshold developed based on the expected variation between two individual's infections in North America. For an entirely new disease of interest, the process of selecting an appropriate threshold becomes an important initial step before clustering methods are applied. This is particularly relevant in the context of the ongoing SARS CoV2 pandemic, as sequence databases quickly grow in size [SM17] and molecular clusters are used to observe outbreaks

[RB20]. The much slower rate of evolution for SARS CoV2 often demands the use of whole genome sequences to see this variation [FFRF20]. Although this study uses HIV as an example due to the availability of data, this framework is applicable to any pathogen with significant variation between hosts, and could be applied to other diseases which have already been studied through molecular clustering techniques such as Zika [FFI+14, GLK+17], SARS [C+04], HCV [JAK+14, SDDA+12, MDS+19a] or malaria [HYW+18] in order to improve the performance of predictive models which are used to identify areas of public health priority.

A less discussed issue is the stability of this optimum value, as well as the consequences of choosing a threshold other than the optimum value. The sudden spike in AIC loss seen for the Seattle and Tennessee data sets using the tree-based clustering method represent an interesting decision for the purposes of long-term study design. This low clustering threshold results in a set of clusters which provide a high amount of usable information, with the use of sequence time-point acting as a relatively good predictor of cluster growth. This would offer impressive results for a purely retrospective study on the association of clusters with a given predictor. However the potential for this optimum value to move as the data set is updated with new cases is also demonstrated with the graph-based clustering methods shown in Figure 4.17. This kind of stochastic behaviour was particularly associated with lower threshold values, with a low threshold value subject to change drastically in its ability to effectively represent epidemics The same low threshold values were also most likely to be associated with positive values in AIC difference, indicating sections where the use of predictor variables were counterproductive. Consistent AIC loss associated with a particular threshold parameter over time is then a useful characteristic for an ongoing study. If such a consistent optimum is not found, it is then important to constantly update the parameters used to identify clusters and track growth using a small window for updates (ie. measuring growth on a monthly or quarterly basis), in order to ensure the criterion for clustering does not become outdated by the time it is used to measure growth.

5.2.2 Depth of maximum AIC loss

The depth of AIC loss can be driven by some combination of two factors. The first is the poor performance (High AIC) of a null model, which would indicate that cluster size alone is a poor predictor of cluster growth. This is particularly true within the field of HIV, as previous studies have affirmed the need for some predictive variable beyond population size to accurately predict cluster growth [LVRD+18]. In part, this can be accounted for by the low, per-act transmission rate [PBB+14] and a number of active treatment and testing programs [oHU+16] which keep

unrestrained exponential growth of an epidemic rare. The more important and highly variable aspect of the AIC loss measurement however, is the proposed model, which in the case of this study, estimates larger growth for clusters with a large number of recent cases. This is a relatively simple model and it is therefore important to classify this work as a proof of concept. The results are only intended to show that a central optimum threshold parameter exists for the purposes of predicting future cluster growth, and that the selection of this threshold has non-trivial effects on the data set. Recent collection date is only one possible indicator of a likely connection to new sequences, and this framework scales well with additional parameters due to it's use of AIC as a measurement of model fit. The penalization of excessively complicated models counters the threat of the same over-fitting situation described in chapter two.

The AIC for this predictive model, even at it's most optimal parameters still indicates a relatively small effect, with a consistent but narrow margin between the performance of the null model and the proposed model. This is shown in the supplementary figure 5.1, which offers a slightly more detailed view of the AIC loss calculation.

This is especially important in the case of the tree-based clustering model (Figure 4.9) , where a random model was used for visual comparison (Figure 4.10) to insure the the patterns of AIC loss were not just based on random effects. The smaller AIC loss values are an indication that the proposed model did not outperform the null model to the same extent in these methods compared to the graph-based methods. It is tempting to conclude that graph-based clustering methods were more effective in this task, however, these results are also dependant on how cluster growth was defined, the predictor variable used (time lag) and how the proposed model was trained. Certain characteristics inherent to the tree structure could explain this: for example, the way that time lag is sometimes calculated between an individual tip and a neighboring node due to the bounds of the framework. The same Bayesian methods used in the Alberta study [VAB+17] which scale tree branches to indicate time could be employed for future implementations of tree-based cluster growth models in this framework. However, even with the profiles shown in this study, it is worth noting that the tree-based optima appeared less "sharp" than those found in the graph-based methods, indicating that these methods may be overall less specific in their effective threshold values.

Although small, the magnitude of AIC loss at optimal values was sufficient to quantify actual improvements in model performance [SIK86, HYW+18]. Furthermore, when using the same data set and the same clustering method, an increased AIC loss is associated with a more representative predictor variable. For instance, because diagnostic date is closer to the time of infection, it is expected to be better representative of time point in an epidemiological

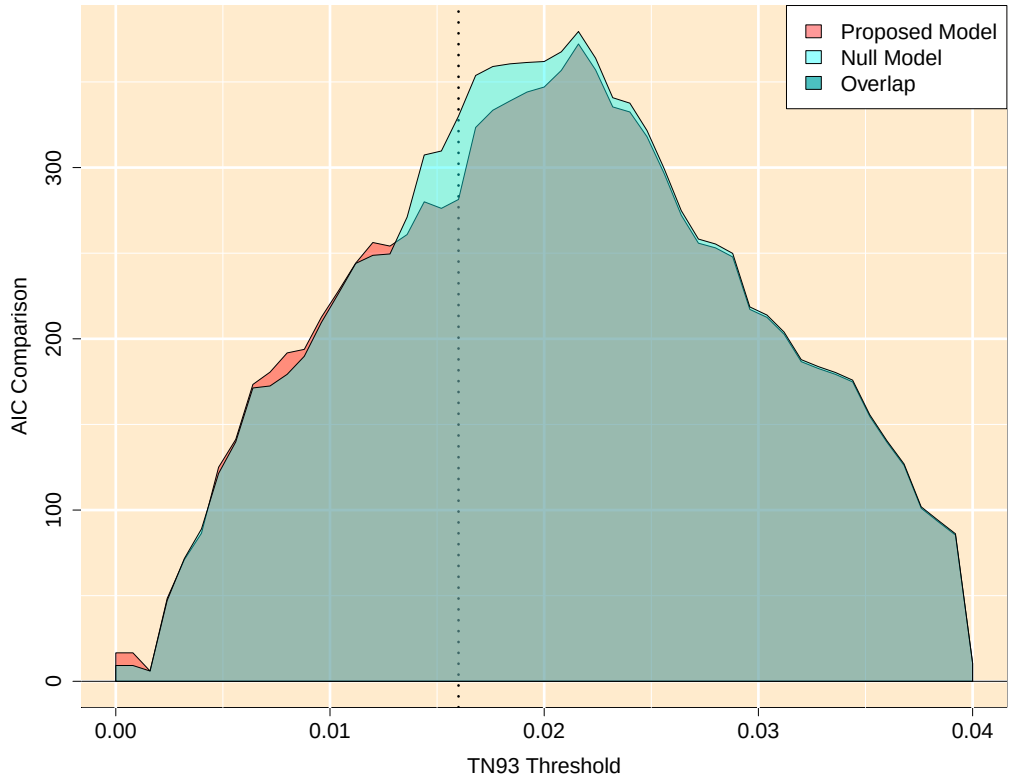

Figure 5.1: Akaike's Information Criterion (AIC) shown for both the null model and proposed model performing in a cross validation test using various TN93 distance thresholds to partition the Seattle data set. Red represents a higher AIC for the proposed model, light blue represents a higher AIC for the null model. The optimal point established in AIC loss calculation is highlighted with a dotted line.

context. This is visible in the plots where models based on the diagnostic date are compared to models based on the collection date such as the rightmost panels in Figures 4.6 and 4.9. In addition, comparing between Figures 4.15 and 4.16 can show that this difference is robust to re-sampling and Figure 4.17 shows that this difference is somewhat robust over time. Given that the full Tennessee data set is the largest, it was expected that the ample amount of data for predictive model training would be result in a relatively effective proposed model and some of the greatest magnitudes of AIC loss in the study. Instead, the relatively small AIC loss at the optimal parameter appears somewhat counter-intuitive. This correlates to the limited effect size of the time-based predictive model, implying that collection dates are a particularly poor indicator of actual infection time for Tennessee. Given the left panel of Figure 4.17, where some of the lowest optimal AIC loss values were seen when the cases with the newest collection dates were included, this appears to be particularly true for the most recent years of

sample collection. The availability of diagnostic dates helps clarify this, as the performance of the proposed model increases dramatically in the comparisons mentioned above. Within this same data set, the unusually high connectivity for cases diagnosed in the early 90s is an important, albeit incidental outcome, as it may point to a past outbreak in that time-frame. A large number of transmissions may have occurred in a short time-frame, but the associated patient sequences were collected over a longer period of time. This is acknowledged as a weakness of time-based clustering models for HIV, especially those which rely on sample collection date, as the time lag between infection and collection can vary so largely [DFGC17, HYW+18]. If my goal was to show these results to the public health authorities associated with each site, other predictive models of cluster growth which take in a larger amount of patient meta data would be necessary for a more certain identification of clusters with a high risk of onward transmission. Variable selection algorithms such as LASSO [Mei07] determine the predictors with the greatest effect on cluster growth and have already been implemented in some predictive studies [WMM+18]. This produces effective predictive models which would the become tailored to the study location and more robust to change. In combination with an optimal range of scaling parameters accounting for any changes that might occur due to sub-sampling, this could offer more consistent predictions of clusters at a high risk of onward transmission.

5.3 Applications and novel components of the presented framework

5.3.1 Optimization based on predictive model outcomes

Predicting the growth of clusters in near real time is a relatively recent innovation [BPP+19, WMM+18, LPA+14, RCLH+16], with the more common alternative use for population-based clusters being the retrospective identification of characteristics associated with clustering [VAB+17, WHVR+17, DVF+18, RCJBS+18, DOKG+17]. These regularly assess recent clusters as a point of high priority [RCJBS+18, DVF+18], which inspired the use of time point to weight cases in my example proposed model. Although purely retrospective clustering studies can be useful in determining areas of low genetic diversity within a given sample, the potential to intervene and prevent an outbreak is dependent on the ability to determine the drivers of future transmission [LVRD+18]. Figure 4.17 shows that the effectiveness for a given predictor variable is shown to fluctuate and over the course of several years, studies have identified differences in the key

drivers of a given epidemic [RCLH+16]. The outcomes used to validate this framework are intended to ensure that clusters are prioritized based on their likelihood to associate with future transmission events. This also does not require that connections between cases represent actual direct transmissions, avoiding some of the previously mentioned inaccuracies with the assumption of direct transmission [Poo16, VF13]. In addition, observing heterogeneity on the level of populations, does not make any specific individual liable in a source attribution case, avoiding the specific problems associated with a hesitation to seek treatment or diagnosis given the criminalization of HIV [SKS+07].

The need for some selection process for scaling has been stated in the literature [Poo16], and the information-based metric provided by this framework allows for a less threshold-dependant comparison between clustering methods, as the optimization step ensures that clustering method performance is being judge at it's most optimal parameters. For example, the use of any distance threshold above 0.05 for graph-based methods would likely result in a singular large cluster containing most, if not all cases. As shown by the AIC loss profiles of graph based methods (Figure 4.6), this provides no real difference in performance between a proposed model and a null model, as both are likely to predict large growth for the single large cluster, which will obtain new cases indiscriminately. For a tree-based clustering method on the same data, high clustering thresholds provide the most difference between a null model and proposed model. This suggests that the information associated with the predictor variables is most significant with a more relaxed clustering threshold. Comparing both clustering methods at the same threshold could produce differences in performance purely based on the choice of scaling parameter. New alternatives to the clustering methods used in the field are quickly being developed [MP17, HPMSR19], often without depending on the manual selection of a scaling parameter. For these newer methods which automatically select a scaling parameter, the threshold optimization step is built into the definition of clusters. Therefore, in order to fairly compare the performance of such a parametric method to a threshold-based method, the parameters of the threshold based method should be optimized. The entirety of the AIC loss profile also allows for comparison in a larger context, providing information such as robustness to sub-sampling (4.15 and 4.16), change of optimal paramaters over time 4.17 and required precision for optimal parameters (ie. the breadth of acceptable scaling parameters). This is not strictly based on predictive model accuracy, but does speak to the re-usability and reliability of a particular method, an important component of performance.

5.3.2 Acknowledging the differences in optimal scaling parameters

The main objective for this study was to discuss the effect of threshold selection on molecular clusters and move towards location-specific threshold parameters as opposed to standardized parameters for the study of population-level HIV transmission dynamics. This addresses the occasional poor performance of standardized clustering thresholds due to the effect of study-dependent characteristics, a well discussed issue in the literature [RLD+17, Poo16, LVRD+18, VLVR+18]. As discussed in Chapter 1, the initial standard TN93 threshold choice of 0.015 is based on the expected distribution of pairwise TN93 distance between any two HIV *pol* sequences in the United states [APP+12], representing the 5% quantile for this distribution. In the context of that study, pairwise distances under that length indicate a level of similarity unlikely to arise by chance. Even disregarding the change in diversity expected with a different viral subtype or gene of interest [BKK+01, LJM+95], Figures 4.1 and 4.2 show that diversity for the subtype B *pol* gene within North America differs visibly from site to site. The associated profile of AIC loss in the previous figures (4.6 and 4.9) show that the differences between study sites lead to distinct differences in optimal threshold for the same subtype. The fact that the optimal TN93 threshold for the Alberta sequence data exists well outside of the IQR shown for the random subsampling tests on the Tennessee data 4.16 indicates a high likelihood that these differences are not simply random. For graph-based methods, using the Northern Alberta optimum threshold of 0.0104 for either of the other two data set results in a predictive model which does worse than a null model with no predictors. In tree-based methods, these differences are also very important, as the use of the optimal maximum patristic distance for Alberta leads to some of the worst possible performance for the Tennessee data set. For public health, this implies that priority clusters identified using a conventional threshold may have a poor basis for their high-priority label, as the predictive model that indicates their high likelihood of onward transmission may be less accurate than model which uses only cluster size. It is important to address that even once priority clusters have been identified, effective intervention cannot be assumed. However, some success has been seen in specific HIV cluster-based responses [SPP+17, GAM+15, IOGT+13], decreasing the prevalence of the disease based on well-prioritized public health intervention. In addition, retrospective analysis of the 2011 Scott County outbreak in the US suggest that the identification of clusters could have effectively prevented an enormous rise in HIV prevalence [GC18]. Due to the widespread nature of the virus, consistent effectiveness of HIV surveillance tools is extremely important [WLBH+14]. The global HIV programme UNAIDS [oHU+16] set the ambitious ”90-90-90” goal for the current year (2020). This had three components: 90 percent of HIV positive individuals globally

should be aware of their status, 90 percent of diagnosed patients should be accessing treatment, and 90 percent of the individuals on treatment should have the amount of virus circulating in their blood suppressed to undetectable levels. These goals are not currently met on a global scale: 79% of the HIV-positive individuals know their status, 60% of diagnosed individuals are on antiretroviral therapy and only 53% of the individuals on anti-retroviral therapy have undetectable viral loads [oHUU+19]. Although the original goal has been criticized in the literature for being unrealistic within the given time-frame [BNN17], it's worth noting that many specific areas achieved this goal quickly [GSL+17, XLB+16, GCK+17], indicating that the lack of widespread success may be due in part to consistency of surveillance effectiveness, as populations with poor access to treatment are unidentified and outbreaks are not met with a fast intervention. With the optimization of these common HIV clustering tools, there is a potential for some of these disparities to be mitigated, allowing for a better overall detection of priority clusters in a wider range of contexts. These prevention efforts offer some hope in reducing HIV prevalence, despite the lack of a vaccine or cure, inspiring further work to refine the clustering techniques used to guide prevention and detection efforts.

5.4 Conclusions

1. Molecular clustering methods which require the manual selection of a parameter must first consider the context of the study site, taking into account the expected variation between sequences. A failure to do so could result in a set of clusters which fail to capture epidemiological dynamics in an informative way, either providing such infrequent connectivity between sequences that few conclusions can be made or such frequent connectivity between sequences that those connections become less meaningful for the purposes of prioritization.

2. An intermediate optimal threshold can be identified for both the tree-based clustering methods similar to Cluster Picker [RCHH+13] and the graph-based methods similar to HIV-TRACE [KPWLBW18] using an information-based metric from the application of predictive clustering models. This estimates the growth of known clusters by viewing the connections within them. In order to avoid the selection of extreme thresholds (which may offer perfect, yet uninformative fit to such a model), a difference in AIC (Akaike's Information Criterion) can be used as an information-based metric, showing the gain in accuracy offered by the use of predictor variables. The optimal threshold is then that which results in the greatest AIC loss calculated between a proposed and null model.

3. The optimal threshold varies depending on the research location, and may vary within a location over time, or in response to incomplete sampling. The magnitude associated with this optimum (ie. largest AIC loss), is difficult to compare between locations, but may become larger with a change in response to a more accurate proposed model (ex. using recent diagnosis to predict onward transmission instead of recent sequence collection).

4. The optimal value can allow for a less context dependent comparison of predictive models and molecular clustering methods, allowing for comparison in the light of threshold parameters that are known to provide informative clusters.

5.5 Future directions

There are several potential future works associated with this project. Most relevant to the results presented here, is the need for an efficient way to obtain the same information regarding robustness of optimal values for tree-based methods. The advantage of single-linkage graph based methods is speed, allowing many iterations of the graph-building process to happen quite quickly. Unfortunately, the same cannot be said for maximum likelihood tree-based methods, with the tree-building algorithms used here taking hours to complete, even with impressive computing resources. Alternatives such as FastTree exist [PDA10], however, methods which more accurately represent the evolutionary divergence between sequences would be preferable. Even disregarding the robustness of such methods to time or sub-sampling, there are other fundamentally different tree-building methods to consider. For instance, Bayesian tree building approaches are becoming a new standard for molecular clustering studies [YR97, VAB+17, BHK+14], providing more detailed information about the probability of the observed tree. While the maximum likelihood methods described previously try to determine the best parameters of an evolutionary model by maximizing the likelihood of the data, Bayesian methods consider the probability of the parameters given the data (the ”posterior probability”) as well as considering the probability of the parameters themselves (the ”prior” probability). These tools are also computationally intensive, but accommodating such tree-building methods in future implementations of this framework may be more representative of the new standards used in the field. The differences seen in the AIC loss profiles for tree and graph based methods could also indicate a relatively ineffective predictive model for tree based methods. A limitation of this study was that my implementation of graph-based methods had the benefit of significantly more peer review [CKP20] prior to this report, as well as a clear methodology outlined in the literature [WMM+18, BPP+19]. The use of pplacer [MKA10] to simulate the growth of known clusters using a maximum likelihood tree is much more novel, and further experimentation with the definition of growth and the calculation of time-lag between sequences would be valuable, potentially revealing a context where these methods perform significantly better. This would allow for more confident and fair comparison between the outcomes obtained by each clustering method.

Code which implements this framework to optimize TN93 distance thresholds for graph-based clustering methods has been released under the general public license v3.0 under the name ”MountainPlot” (https://github.com/PoonLab/MountainPlot), with an associated publication [CKP20] in an effort to translate this work into a usable, open-source tool. However, there are currently many limitations to the current release as a piece of software. Ideally, fur-

ther releases could allow for different proposed models, tree based methods, robustness testing (seen in Figures 4.17, 4.15, 4.16), and comparison to a random model's AIC loss profile. In addition, a more user-friendly interfacing could do more to help a broad audience use this work, given that the end users of this framework would be more likely to be public health agencies and not bioinformaticians. There is also an additional application of the modifiable areal unit problem termed the modifiable temporal unit problem [CA14], which discusses the agglomeration of time points into different resolutions (ie. viewing the data at the resolution of years, months, weeks or days). This can be done in parallel with the other threshold-based optimization, obtaining an optimal precision of time information. I compiled some preliminary results using the Northern Alberta data set and the Graph-based clustering methods and presented them in virtual poster form at the Canadian Association for HIV Research conference in May 2020 [Cha]. This involved running an altered version of the framework discussed in previous chapters, using numerous different time-frame granularities while keeping the threshold used to define clustering constant. Establishing a method of accomplishing this task for tree based methods and implementing this into MountainPlot is another dimension of optimization which could improve predictive model performance for models which use some degree of time-point information.

Bibliography

[Aka73] Htrotugu Akaike. Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 60(2):255–265, 1973.

[APP+12] Jeannette L Aldous, Sergei Kosakovsky Pond, Art Poon, Sonia Jain, Huifang Qin, James S Kahn, Mari Kitahata, Benigno Rodriguez, Ann M Dennis, Stephen L Boswell, et al. Characterizing hiv transmission networks across the united states. *Clinical Infectious Diseases*, 55(8):1135–1143, 2012.

[BHC+93] James J Bull, John P Huelsenbeck, Clifford W Cunningham, David L Swofford, and Peter J Waddell. Partitioning and combining data in phylogenetic analysis. *Systematic Biology*, 42(3):384–397, 1993.

[BHK+14] Remco Bouckaert, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A Suchard, Andrew Rambaut, and Alexei J Drummond. Beast 2: a software platform for bayesian evolutionary analysis. *PLoS Comput Biol*, 10(4):e1003537, 2014.

[BJRP+06] Salima Bouzeghoub, Valérie Jauvin, Patricia Recordon-Pinson, Isabelle Garrigue, Achour Amrane, El-Hadj Belabbes, and Hervé J Fleury. High diversity of hiv type 1 in algeria. *AIDS Research & Human Retroviruses*, 22(4):367–372, 2006.

[BKK+01] Aleksei Bobkov, Elena Kazennova, Tatyana Khanina, Marina Bobkova, Ludmila Selimova, Aleksei Kravchenko, Vadim Pokrovsky, and Jonathan Weber. An hiv type 1 subtype a strain of low genetic diversity continues to spread among injecting drug users in russia: study of the new local outbreaks in moscow and irkutsk. *AIDS research and human retroviruses*, 17(3):257–261, 2001.

[BKML+11] Dennis A Benson, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. Genbank. *Nucleic acids research*, 39(Database issue):D32, 2011.

[BLD+12] Guy Baele, Wai Lok Sibon Li, Alexei J Drummond, Marc A Suchard, and Philippe Lemey. Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. *Molecular biology and evolution*, 30(2):239–243, 2012.

[BNN17] Luchuo Engelbert Bain, Clovis Nkoke, and Jean Jacques N Noubiap. Unaids 90–90–90 targets to end the aids epidemic by 2020 are not realistic: comment on can the unaids 90–90–90 target be achieved? a systematic analysis of national hiv treatment cascades. *BMJ global health*, 2(2):e000227, 2017.

[BPP+19] Rachael M Billock, Kimberly A Powers, Dana K Pasquale, Erika Samoff, Victoria L Mobley, William C Miller, Joseph J Eron, and Ann M Dennis. Prediction of hiv transmission cluster growth with statewide surveillance data. *Jaids Journal of Acquired Immune Deficiency Syndromes*, 80(2):152–159, 2019.

[C+04] Chinese SARS Molecular Epidemiology Consortium et al. Molecular evolution of the sars coronavirus during the course of the sars epidemic in china. *Science*, 303(5664):1666–1669, 2004.

[CA14] Tao Cheng and Monsuru Adepeju. Modifiable temporal unit problem (MTUP) and its effect on space-time cluster detection. *PLoS One*, 9(6):e100465, 2014.

[CAC+09] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[CGG+15] José M Cuevas, Ron Geller, Raquel Garijo, José López-Aldeguer, and Rafael Sanjuán. Extremely high mutation rate of hiv-1 in vivo. *PLoS biology*, 13(9), 2015.

[CGO+14] Stacy M Cohen, Kristen Mahle Gray, M Cheryl Bañez Ocfemia, Anna Satcher Johnson, and H Irene Hall. The status of the national hiv surveillance system, united states, 2013. *Public health reports*, 129(4):335–341, 2014.

[Cha] C. Chato. An outcome-based statistical framework to select and optimize molecular clustering methods for infectious diseases.

[CKP20] Connor Chato, Marcia L Kalish, and Art FY Poon. Public health in genetic spaces: a statistical framework to optimize cluster-based outbreak detection. *Virus Evolution*, 6(1):veaa011, 2020.

[CSH+11] Chen-Shan Chin, Jon Sorenson, Jason B Harris, William P Robins, Richelle C Charles, Roger R Jean-Charles, James Bullard, Dale R Webster, Andrew Kasarskis, Paul Peluso, et al. The origin of the haitian cholera outbreak strain. *New England Journal of Medicine*, 364(1):33–42, 2011.

[CWK+13] Matthew Cotten, Simon J Watson, Paul Kellam, Abdullah A Al-Rabeeah, Hatem Q Makhdoom, Abdullah Assiri, Jaffar A Al-Tawfiq, Rafat F Alhakeem, Hossam Madani, Fahad A AlRabiah, et al. Transmission and evolution of the middle east respiratory syndrome coronavirus in saudi arabia: a descriptive genomic study. *The Lancet*, 382(9909):1993–2002, 2013.

[DBC+17] Ruud H Deurenberg, Erik Bathoorn, Monika A Chlebowicz, Natacha Couto, Mithila Ferdous, Silvia García-Cobos, Anna MD Kooistra-Smid, Erwin C Raangs, Sigrid Rosema, Alida CM Veloo, et al. Application of next generation sequencing in clinical microbiology and infection prevention. *Journal of biotechnology*, 243:16–24, 2017.

[DBK+10] Deborah Donnell, Jared M Baeten, James Kiarie, Katherine K Thomas, Wendy Stevens, Craig R Cohen, James McIntyre, Jairam R Lingappa, Connie Celum, Partners in Prevention HSV/HIV Transmission Study Team, et al. Heterosexual hiv-1 transmission after initiation of antiretroviral therapy: a prospective cohort analysis. *The Lancet*, 375(9731):2092–2098, 2010.

[DFGC17] Xavier Didelot, Christophe Fraser, Jennifer Gardy, and Caroline Colijn. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular biology and evolution*, 34(4):997–1007, 2017.

[D'h05] Patrik D'haeseleer. How does gene expression clustering work? *Nature biotechnology*, 23(12):1499–1501, 2005.

[DOKG+17] Tulio De Oliveira, Ayesha BM Kharsany, Tiago Gräf, Cherie Cawood, David Khanyile, Anneke Grobler, Adrian Puren, Savathree Madurai, Cheryl Baxter,

Quarraisha Abdool Karim, et al. Transmission networks and risk of hiv infection in kwazulu-natal, south africa: a community-wide phylogenetic study. *The lancet HIV*, 4(1):e41–e50, 2017.

[DVF+18] Ann M Dennis, Erik Volz, AS Md Simon DW Frost, Mukarram Hossain, Art FY Poon, Peter F Rebeiro, Sten H Vermund, Timothy R Sterling, and Marcia L Kalish. Hiv-1 transmission clustering and phylodynamics highlight the important role of young men who have sex with men. *AIDS research and human retroviruses*, 34(10):879–888, 2018.

[EGK+01] John Ellson, Emden Gansner, Lefteris Koutsofios, Stephen C North, and Gordon Woodhull. Graphvizopen source graph drawing tools. In *International Symposium on Graph Drawing*, pages 483–484. Springer, 2001.

[ES90] Margaret A Ellis and Bjarne Stroustrup. *The annotated C++ reference manual*. Addison-Wesley, 1990.

[FFI+14] Oumar Faye, Caio CM Freire, Atila Iamarino, Ousmane Faye, Juliana Velasco C de Oliveira, Mawlouth Diallo, Paolo MA Zanotto, and Amadou Alpha Sall. Molecular evolution of zika virus during its emergence in the 20th century. *PLoS neglected tropical diseases*, 8(1), 2014.

[FFRF20] Peter Forster, Lucy Forster, Colin Renfrew, and Michael Forster. Phylogenetic network analysis of sars-cov-2 genomes. *Proceedings of the National Academy of Sciences*, 117(17):9241–9243, 2020.

[FKL+18] Brian Thomas Foley, Bette Tina Marie Korber, Thomas Kenneth Leitner, Cristian Apetrei, Beatrice Hahn, Ilene Mizrachi, James Mullins, Andrew Rambaut, and Steven Wolinsky. Hiv sequence compendium 2018. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2018.

[FM08] Elizabeth A Freeman and Gretchen G Moisen. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling*, 217(1-2):48–58, 2008.

[FW91] A Stewart Fotheringham and David WS Wong. The modifiable areal unit problem in multivariate statistical analysis. *Environment and planning A*, 23(7):1025–1044, 1991.

[GAM+15] Raquel González, Orvalho J Augusto, Khátia Munguambe, Charlotte Pierrat, Elpidia N Pedro, Charfudin Sacoor, Elisa De Lazzari, John J Aponte, Eusébio Macete, Pedro L Alonso, et al. Hiv incidence and spatial clustering in a rural area of southern mozambique. *PloS one*, 10(7), 2015.

[GC18] Gregg S Gonsalves and Forrest W Crawford. Dynamics of the hiv outbreak and response in scott county, in, usa, 2011–15: a modelling study. *The Lancet HIV*, 5(10):e569–e577, 2018.

[GCK+17] Anneke Grobler, Cherie Cawood, David Khanyile, Adrian Puren, and Ayesha BM Kharsany. Progress of unaids 90-90-90 targets in a district in kwazulu-natal, south africa, with high hiv burden, in the hipss study: a household-based complex multilevel community survey. *The Lancet HIV*, 4(11):e505–e513, 2017.

[Geh65] Edmund A Gehan. A generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52(1-2):203–224, 1965.

[GIA+04] Zehava Grossman, Valery Istomin, Diana Averbuch, Margalit Lorber, Klaris Risenberg, Itzchak Levi, Michal Chowers, Michael Burke, Nimrod Bar Yaacov, Jonathan M Schapiro, et al. Genetic variation at nnrti resistance-associated positions in patients infected with hiv-1 subtype c. *Aids*, 18(6):909–915, 2004.

[GL18] Jennifer L Gardy and Nicholas J Loman. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nature Reviews Genetics*, 19(1):9, 2018.

[GLK+17] Nathan D Grubaugh, Jason T Ladner, Moritz UG Kraemer, Gytis Dudas, Amanda L Tan, Karthik Gangavarapu, Michael R Wiley, Stephen White, Julien Thézé, Diogo M Magnani, et al. Genomic epidemiology reveals multiple introductions of zika virus into the united states. *Nature*, 546(7658):401–405, 2017.

[GSF+81] Michael S Gottlieb, Howard M Schanker, Peng Thim Fan, Andrew Saxon, Joel D Weisman, Irving Pozalski, et al. Pneumocystis pneumonialos angeles. *Mmwr*, 30(21):250–2, 1981.

[GSL+17] M Gisslen, V Svedhem, L Lindborg, L Flamholc, H Norrgren, S Wendahl, M Axelsson, and A Sönnerborg. Sweden, the first country to achieve the joint united nations programme on hiv/aids (unaids)/world health organization (who) 90-90-90 continuum of hiv care targets. *HIV medicine*, 18(4):305–307, 2017.

[Ham50] Richard W Hamming. Error detecting and error correcting codes. *The Bell system technical journal*, 29(2):147–160, 1950.

[HBV01] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of intelligent information systems*, 17(2-3):107–145, 2001.

[HCCP04] Stéphane Hué, Jonathan P Clewley, Patricia A Cane, and Deenan Pillay. Hiv-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *Aids*, 18(5):719–728, 2004.

[HD03] Françoise F Hamers and Angela M Downs. Hiv in central and eastern europe. *The Lancet*, 361(9362):1035–1044, 2003.

[HPMSR19] Alvin X Han, Edyth Parker, Sebastian Maurer-Stroh, and Colin A Russell. Inferring putative transmission clusters with phydelity. *Virus Evolution*, 5(2):vez039, 2019.

[HSH+82] John Holland, Katherine Spindler, Frank Horodyski, Elizabeth Grabau, Stuart Nichol, and Scott VandePol. Rapid evolution of rna genomes. *Science*, 215(4540):1577–1585, 1982.

[HYW+18] Peter Haddawy, Myat Su Yin, Tanawan Wisanrakkit, Rootrada Limsupavanich, Promporn Promrat, Saranath Lawpoolsri, and Patiwat Sa-angchai. Complexity-based spatial hierarchical clustering for malaria prediction. *Journal of Healthcare Informatics Research*, 2(4):423–447, 2018.

[HZR+95] Edward C Holmes, Lin Qi Zhang, Pamela Robertson, Alexander Cleland, Elizabeth Harvey, Peter Simmonds, and Andrew J Leigh Brown. The molecular epidemiology of human immunodeficiency virus type 1 in edinburgh. *Journal of Infectious Diseases*, 171(1):45–53, 1995.

[ICCSS14] Fernando Izquierdo-Carrasco, John Cazes, Stephen A Smith, and Alexandros Stamatakis. Pumper: phylogenies updated perpetually. *Bioinformatics*, 30(10):1476–1477, 2014.

[IOGT+13] Collins C Iwuji, Joanna Orne-Gliemann, Frank Tanser, Sylvie Boyer, Richard J Lessells, France Lert, John Imrie, Till Bärnighausen, Claire Rekacewicz, Brigitte Bazin, et al. Evaluation of the impact of immediate versus who recommendations-guided antiretroviral therapy initiation on hiv incidence: the anrs 12249 tasp (treatment as prevention) trial in hlabisa subdistrict, kwazulu-natal, south africa: study protocol for a cluster randomised controlled trial. *Trials*, 14(1):230, 2013.

[JAK+14] Brendan Jacka, Tanya Applegate, Mel Krajden, Andrea Olmstead, P Richard Harrigan, Brandon DL Marshall, Kora DeBeck, M-J Milloy, Francois Lamoury, Oliver G Pybus, et al. Phylogenetic clustering of hepatitis c virus among people who inject drugs in vancouver, canada. *Hepatology*, 60(5):1571–1580, 2014.

[JW96] Dennis E Jelinski and Jianguo Wu. The modifiable areal unit problem and implications for landscape ecology. *Landscape ecology*, 11(3):129–140, 1996.

[Kan06] Rami Kantor. Impact of hiv-1 pol diversity on drug resistance and its clinical implications. *Current opinion in infectious diseases*, 19(6):594–606, 2006.

[KGY+01] Bette Korber, Brian Gaschen, Karina Yusim, Rama Thakallapally, Can Kesmir, and Vincent Detours. Evolutionary and immunological implications of contemporary hiv-1 variation. *British medical bulletin*, 58(1):19–42, 2001.

[KK89] Tomihisa Kamada and Satoru Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1):7–15, 1989.

[KPWLBW18] Sergei L Kosakovsky Pond, Steven Weaver, Andrew J Leigh Brown, and Joel O Wertheim. 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, 2018.

[kPWV18] Sergei L kosakovsky Pond, Steven Weaver, and Ryan Velazquez. Tn93 fast distance calculator. `https://github.com/veg/tn93`, 2018.

[KVS72] A Kershenbaum and R Van Slyke. Computing minimum spanning trees efficiently. In *Proceedings of the ACM annual conference-Volume 1*, pages 518–527, 1972.

[Lar14] Anders Larsson. Aliview: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, 30(22):3276–3278, 2014.

[LEF+96] Thomas Leitner, David Escanilla, Christer Franzen, Mathias Uhlen, and Jan Albert. Accurate reconstruction of a known hiv-1 transmission history by phylogenetic tree analysis. *Proceedings of the National Academy of Sciences*, 93(20):10864–10869, 1996.

[LF12] Katrina A Lythgoe and Christophe Fraser. New insights into the evolutionary rate of hiv-1 at the within-host and epidemiological levels. *Proceedings of the Royal Society B: Biological Sciences*, 279(1741):3367–3375, 2012.

[LJM+95] Joost Louwagie, Wouter Janssens, John Mascola, Leo Heyndrickx, Patricia Hegerich, Guido Van Der Groen, Francine E McCutchan, and Donald S Burke. Genetic diversity of the envelope glycoprotein from human immunodeficiency virus type 1 isolates of african origin. *Journal of Virology*, 69(1):263–271, 1995.

[LLH+16] Robert S Lanciotti, Amy J Lambert, Mark Holodniy, Sonia Saavedra, and Leticia del Carmen Castillo Signor. Phylogeny of zika virus in western hemisphere, 2015. *Emerging infectious diseases*, 22(5):933, 2016.

[LLR+20] Yuruo Li, Hongjie Liu, Habib O Ramadhani, Nicaise Ndembi, Trevor A Crowell, Gustavo Kijak, Merlin L Robb, Julie A Ake, Afoke Kokogho, Rebecca G Nowak, et al. Genetic clustering analysis for hiv infection among msm in nigeria: implications for intervention. *Aids*, 34(2):227–236, 2020.

[LPA+14] Susan J Little, Sergei L Kosakovsky Pond, Christy M Anderson, Jason A Young, Joel O Wertheim, Sanjay R Mehta, Susanne May, and Davey M Smith. Using hiv networks to inform real time prevention interventions. *PloS one*, 9(6), 2014.

[LPD00] Shuying Li, Dennis K Pearl, and Hani Doss. Phylogenetic tree construction using markov chain monte carlo. *Journal of the American statistical Association*, 95(450):493–508, 2000.

[LPSS84] Cecilia Lanave, Giuliano Preparata, Cecilia Sacone, and Gabriella Serio. A new method for calculating evolutionary substitution rates. *Journal of molecular evolution*, 20(1):86–93, 1984.

[LRP06] Philippe Lemey, Andrew Rambaut, and Oliver G Pybus. Hiv evolutionary dynamics within and among hosts. *AIDs Rev*, 8(3):125–140, 2006.

[LSG11] Annan Li, Shiguang Shan, and Wen Gao. Coupled bias–variance tradeoff for cross-pose face recognition. *IEEE Transactions on Image Processing*, 21(1):305–315, 2011.

[LSM19] Jessica A Lee, Sergey Stolyar, and Christopher J Marx. An aerobic link between lignin degradation and c1 metabolism: growth on methoxylated aromatic compounds by members of the genus methylobacterium. *bioRxiv*, page 712836, 2019.

[LVRD+18] Stéphane Le Vu, Oliver Ratmann, Valerie Delpech, Alison E Brown, O Noel Gill, Anna Tostevin, Christophe Fraser, and Erik M Volz. Comparison of cluster-based and source-attribution methods for estimating transmission risk using large hiv sequence databases. *Epidemics*, 23:1–10, 2018.

[LZL+20] Roujian Lu, Xiang Zhao, Juan Li, Peihua Niu, Bo Yang, Honglong Wu, Wenling Wang, Hao Song, Baoying Huang, Na Zhu, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet*, 395(10224):565–574, 2020.

[MDS+19a] DG Murphy, R Dion, M Simard, ML Vachon, V Martel-Laferrière, B Serhir, and J Longtin. Molecular phylogenetic used to identify hcv transmission. *CCDR*, 45:9, 2019.

[MDS+19b] DG Murphy, R Dion, M Simard, ML Vachon, V Martel-Laferrière, B Serhir, and J Longtin. Outbreaks: 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*, 45(9):230, 2019.

[Mei07] Nicolai Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, 2007.

[MKA10] Frederick A Matsen, Robin B Kodner, and E Virginia Armbrust. pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics*, 11(1):538, 2010.

[MNvH13] Bui Quang Minh, Minh Anh Thi Nguyen, and Arndt von Haeseler. Ultrafast approximation for phylogenetic bootstrap. *Molecular biology and evolution*, 30(5):1188–1195, 2013.

[MP17] Rosemary M McCloskey and Art FY Poon. A model-based clustering method to detect infectious disease transmission outbreaks from sequence variation. *PLoS computational biology*, 13(11):e1005868, 2017.

[MPL+20] Lauren Mak, Deshan Perera, Raynell Lang, Pathum Kossinna, Jingni He, M John Gill, Quan Long, and Guido van Marle. Evaluation of a phylogenetic pipeline to examine transmission networks in a canadian hiv cohort. *Microorganisms*, 8(2):196, 2020.

[MS16] Alyssa R Martin and Robert F Siliciano. Progress toward hiv eradication: case reports, current efforts, and the challenges associated with cure. *Annual review of medicine*, 67:215–228, 2016.

[MSCB96] Francine E McCutchan, Mika O Salminen, Jean K Carr, and Donald S Burke. Hiv-1 genetic diversity. *AIDS (London, England)*, 10:S13–20, 1996.

[MWT+90] Terry McNearney, Peter Westervelt, Benjamin J Thielan, David B Trowbridge, Juan Garcia, Robert Whittier, and Lee Ratner. Limited sequence heterogeneity among biologically distinct human immunodeficiency virus type 1 isolates from individuals involved in a clustered infectious outbreak. *Proceedings of the National Academy of Sciences*, 87(5):1917–1921, 1990.

[Myk15] Eric Mykhalovskiy. The public health implications of hiv criminalization: past, current, and future research directions, 2015.

[Nak00] Tomoki Nakaya. An information statistical approach to the modifiable areal unit problem in incidence rate maps. *Environment and Planning A*, 32(1):91–109, 2000.

[NB17] Jonathan K Nelson and Cynthia A Brewer. Evaluating data stability in aggregation structures across spatial scales: revisiting the modifiable areal unit

problem. *Cartography and Geographic Information Science*, 44(1):35–50, 2017.

[NMB+18] Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks. *arXiv preprint arXiv:1810.08591*, 2018.

[NMC+14] Bohdan Nosyk, Julio SG Montaner, Guillaume Colley, Viviane D Lima, Keith Chan, Katherine Heath, Benita Yip, Hasina Samji, Mark Gilbert, Rolando Barrios, et al. The cascade of hiv care in british columbia, canada, 1996–2011: a population-based retrospective cohort study. *The Lancet infectious diseases*, 14(1):40–49, 2014.

[NML+14] Vlad Novitsky, Sikhulile Moyo, Quanhong Lei, Victor DeGruttola, and Myron Essex. Impact of sampling density on the extent of hiv clustering. *AIDS research and human retroviruses*, 30(12):1226–1235, 2014.

[NML+15] Vlad Novitsky, Sikhulile Moyo, Quanhong Lei, Victor DeGruttola, and M Essex. Importance of viral sequence length and number of variable and informative sites in analysis of hiv clustering. *AIDS research and human retroviruses*, 31(5):531–542, 2015.

[NSVHM15] Lam-Tung Nguyen, Heiko A Schmidt, Arndt Von Haeseler, and Bui Quang Minh. Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, 32(1):268–274, 2015.

[OFP+18] Alexandra M Oster, Anne Marie France, Nivedha Panneer, M Cheryl Bañez Ocfemia, Ellsworth Campbell, Sharoda Dasgupta, William M Switzer, Joel O Wertheim, and Angela L Hernandez. Identifying clusters of recent and rapid hiv transmission through analysis of molecular surveillance data. *Journal of acquired immune deficiency syndromes (1999)*, 79(5):543, 2018.

[oHU+16] Joint United Nations Programme on HIV/AIDS (UNAIDS) et al. Global aids update 2016. 2016. *Joint United Nations Programme on HIV/AIDS (UNAIDS): Geneva, Switzerland*, 2016.

[oHUU+19] Joint United Nations Programme on HIV/AIDS (UNAIDS), Data UNAIDS, et al. Geneva, switzerland; 2018. *North American, Western and Central Europe: AIDS epidemic update regional summary*, pages 1–16, 2019.

[Ope77] Stan Openshaw. Optimal zoning systems for spatial interaction models. *Environment and Planning A*, 9(2):169–184, 1977.

[OWH+15] Alexandra M Oster, Joel O Wertheim, Angela L Hernandez, M Cheryl Bañez Ocfemia, Neeraja Saduvala, and H Irene Hall. Using molecular hiv surveillance data to understand transmission between subpopulations in the united states. *Journal of acquired immune deficiency syndromes (1999)*, 70(4):444, 2015.

[PBB+14] Pragna Patel, Craig B Borkowf, John T Brooks, Arielle Lasry, Amy Lansky, and Jonathan Mermin. Estimating per-act hiv transmission risk: a systematic review. *AIDS (London, England)*, 28(10), 2014.

[PDA10] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3), 2010.

[PMO+15] Sophie E Patterson, M-J Milloy, Gina Ogilvie, Saara Greene, Valerie Nicholson, Micheal Vonn, Robert Hogg, and Angela Kaida. The impact of criminalization of hiv non-disclosure on the healthcare engagement of women living with hiv in canada: a comprehensive review of the evidence. *Journal of the International AIDS Society*, 18(1):20572, 2015.

[Poo16] Art FY Poon. Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks. *Virus evolution*, 2(2):vew031, 2016.

[R C13] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[RB20] Hussin A Rothan and Siddappa N Byrareddy. The epidemiology and pathogenesis of coronavirus disease (covid-19) outbreak. *Journal of autoimmunity*, page 102433, 2020.

[RCHH+13] Manon Ragonnet-Cronin, Emma Hodcroft, Stéphane Hué, Esther Fearnhill, Valerie Delpech, Andrew J Leigh Brown, and Samantha Lycett. Automated analysis of phylogenetic clusters. *BMC bioinformatics*, 14(1):317, 2013.

[RCJBS+18] Manon Ragonnet-Cronin, Celia Jackson, Amanda Bradley-Stewart, Celia Aitken, Andrew McAuley, Norah Palmateer, Rory Gunson, David Goldberg, Catriona Milosevic, and Andrew J Leigh Brown. 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, 2018.

[RCLH+16] Manon Ragonnet-Cronin, Samantha J Lycett, Emma B Hodcroft, Stéphane Hué, Esther Fearnhill, Alison E Brown, Valerie Delpech, David Dunn, and Andrew J Leigh Brown. Transmission of non-b hiv subtypes in the united kingdom is increasingly driven by large non-heterosexual transmission clusters. *The Journal of infectious diseases*, 213(9):1410–1418, 2016.

[RCOAM+10] Manon Ragonnet-Cronin, Marianna Ofner-Agostini, Harriet Merks, Richard Pilon, Michael Rekart, Chris P Archibald, Paul A Sandstrom, and James I Brooks. Longitudinal phylogenetic surveillance identifies distinct patterns of cluster dynamics. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 55(1):102–108, 2010.

[RD69] Basil Cameron Rennie and Annette Jane Dobson. On stirling numbers of the second kind. *Journal of Combinatorial Theory*, 7(2):116–121, 1969.

[RHP+85] Lee Ratner, William Haseltine, Roberto Patarca, Kenneth J Livak, Bruno Starcich, Steven F Josephs, Ellen R Doran, J Antoni Rafalski, Erik A Whitehorn, Kirk Baumeister, et al. Complete nucleotide sequence of the aids virus, htlv-iii. *Nature*, 313(6000):277–284, 1985.

[RHR+19] Rebecca Rose, Matthew Hall, Andrew D Redd, Susanna Lamers, Andrew E Barbier, Stephen F Porcella, Sarah E Hudelson, Estelle Piwowar-Manning, Marybeth McCauley, Theresa Gamble, et al. Phylogenetic methods inconsistently predict the direction of hiv transmission among heterosexual pairs in the hptn 052 cohort. *The Journal of infectious diseases*, 220(9):1406–1413, 2019.

[RLD+17] Rebecca Rose, Susanna L Lamers, James J Dollar, Mary K Grabowski, Emma B Hodcroft, Manon Ragonnet-Cronin, Joel O Wertheim, Andrew D Redd, Danielle German, and Oliver Laeyendecker. Identifying transmission clusters with cluster picker and hiv-trace. *AIDS research and human retroviruses*, 33(3):211–218, 2017.

[RY96] Bruce Rannala and Ziheng Yang. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of molecular evolution*, 43(3):304–311, 1996.

[San14] Chris Sanders. Discussing the limits of confidentiality: The impact of criminalizing hiv nondisclosure on public health nurses counseling practices. *Public Health Ethics*, 7(3):253–260, 2014.

[SCZ+03] Evguenia S Svarovskaia, Sara R Cheslock, Wen-Hui Zhang, Wei-Shau Hu, and Vinay K Pathak. Retroviral mutation rates and reverse transcriptase fidelity. *Front Biosci*, 8(4):d117–134, 2003.

[SDDA+12] Rachel Sacks-Davis, Galina Daraganova, Campbell Aitken, Peter Higgs, Lilly Tracy, Scott Bowden, Rebecca Jenkinson, David Rolls, Philippa Pattison, Garry Robins, et al. Hepatitis c virus phylogenetic clustering is associated with the social-injecting network in a cohort of people who inject drugs. *PloS one*, 7(10), 2012.

[SF72] Samuel S Shapiro and RS Francia. An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67(337):215–216, 1972.

[SIK86] Yosiyuki Sakamoto, Makio Ishiguro, and Genshiro Kitagawa. Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, 81, 1986.

[SKS+07] Leickness C Simbayi, Seth C Kalichman, Anna Strebel, Allanise Cloete, Nomvo Henda, and Ayanda Mqeketo. Disclosure of hiv status to sex partners and sexual risk behaviours among hiv-positive men and women, cape town, south africa. *Sexually transmitted infections*, 83(1):29–34, 2007.

[SLT+14] Daniel Struck, Glenn Lawyer, Anne-Marie Ternes, Jean-Claude Schmit, and Danielle Perez Bercoff. Comet: adaptive context-based modeling for ultrafast hiv-1 subtype identification. *Nucleic acids research*, 42(18):e144–e144, 2014.

[SM17] Yuelong Shu and John McCauley. Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, 22(13), 2017.

[SNH+12] R Burke Squires, Jyothi Noronha, Victoria Hunt, Adolfo García-Sastre, Catherine Macken, Nicole Baumgarth, David Suarez, Brett E Pickett, Yun

Zhang, Christopher N Larsen, et al. Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. *Influenza and other respiratory viruses*, 6(6):404–416, 2012.

[Sno55] John Snow. *On the mode of communication of cholera*. John Churchill, 1855.

[SPP+17] Vana Sypsa, Mina Psichogiou, Dimitrios Paraskevis, Georgios Nikolopoulos, Chrissa Tsiara, Dimitra Paraskeva, Katerina Micha, Meni Malliori, Anastasia Pharris, Lucas Wiessing, et al. Rapid decline in hiv incidence among persons who inject drugs during a fast-track combination prevention program after an hiv outbreak in athens. *The Journal of infectious diseases*, 215(10):1496–1505, 2017.

[ST06] Jerry Scripps and Pang-Ning Tan. Clustering in the presence of bridge-nodes. In *Proceedings of the 2006 SIAM International Conference on Data Mining*, pages 270–281. SIAM, 2006.

[Sta06] Alexandros Stamatakis. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

[SW92] Temple F Smith and Michael S Waterman. The continuing case of the florida dentist. *Science*, 256(5060):1155–1157, 1992.

[TBS+99] Karine Triques, Anke Bourgeois, Sentob Saragosti, Nicole Vidal, Eitel Mpoudi-Ngole, Nzila Nzilambi, Christian Apetrei, Michel Ekwalanga, Eric Delaporte, and Martine Peeters. High diversity of hiv-1 subtype f strains in central africa. *Virology*, 259(1):99–109, 1999.

[TKP+12] Michael C Thigpen, Poloko M Kebaabetswe, Lynn A Paxton, Dawn K Smith, Charles E Rose, Tebogo M Segolodi, Faith L Henderson, Sonal R Pathak, Fatma A Soud, Kata L Chillag, et al. Antiretroviral preexposure prophylaxis for heterosexual hiv transmission in botswana. *New England Journal of Medicine*, 367(5):423–434, 2012.

[TN93] Koichiro Tamura and Masatoshi Nei. 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, 1993.

[Tom92] Lucy S Tompkins. The use of molecular methods in infectious diseases. *New England Journal of Medicine*, 327(18):1290–1297, 1992.

[VAB+17] Bram Vrancken, D Adachi, M Benedet, A Singh, R Read, S Shafran, GD Taylor, K Simmonds, C Sikora, Philippe Lemey, et al. 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, 2017.

[VF13] Erik M Volz and SD Frost. Inferring the source of transmission with phylogenetic data. *PLoS computational biology*, 9(12):e1003397–e1003397, 2013.

[VKR+03] Nicole Vidal, Donato Koyalta, Vincent Richard, Catherine Lechiche, Thomas Ndinaromtan, Abakar Djimasngar, Eric Delaporte, and Martine Peeters. High genetic diversity of hiv-1 strains in chad, west central africa. *Journal of acquired immune deficiency syndromes (1999)*, 33(2):239–246, 2003.

[VKW+12] Erik M Volz, James S Koopman, Melissa J Ward, Andrew Leigh Brown, and Simon DW Frost. Simple epidemiological dynamics explain phylogenetic clustering of hiv from patients with recent infection. *PLoS computational biology*, 8(6), 2012.

[VLVR+18] Erik M Volz, Stephane Le Vu, Oliver Ratmann, Anna Tostevin, David Dunn, Chloe Orkin, Siobhan OShea, Valerie Delpech, Alison Brown, Noel Gill, et al. Molecular epidemiology of hiv-1 subtype b reveals heterogeneous transmission risk: implications for intervention and control. *The Journal of infectious diseases*, 217(10):1522–1529, 2018.

[VRDJ95] Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[WBL+12] Conan K Woods, Chanson J Brumme, Tommy F Liu, Celia KS Chui, Anna L Chu, Brian Wynhoven, Tom A Hall, Christina Trevino, Robert W Shafer, and P Richard Harrigan. Automating hiv drug resistance genotyping with recall, a freely accessible sequence analysis tool. *Journal of clinical microbiology*, 50(6):1936–1942, 2012.

[WCP19] Joel O Wertheim, Connor Chato, and Art FY Poon. Comparative analysis of hiv sequences in real time for public health. *Current Opinion in HIV and AIDS*, 14(3):213–220, 2019.

[WDG+09] Joseph M Watts, Kristen K Dang, Robert J Gorelick, Christopher W Leonard, Julian W Bess Jr, Ronald Swanstrom, Christina L Burch, and Kevin M Weeks. Architecture and secondary structure of an entire hiv-1 rna genome. *Nature*, 460(7256):711–716, 2009.

[WHVR+17] Elizabeth Wolf, Joshua T Herbeck, Stephen Van Rompaey, Mari Kitahata, Katherine Thomas, Gregory Pepper, and Lisa Frenkel. 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, 2017.

[WIH+13] Timothy M Walker, Camilla LC Ip, Ruth H Harrell, Jason T Evans, Georgia Kapatai, Martin J Dedicoat, David W Eyre, Daniel J Wilson, Peter M Hawkey, Derrick W Crook, et al. Whole-genome sequencing to delineate mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet infectious diseases*, 13(2):137–146, 2013.

[WLBH+14] Joel O Wertheim, Andrew J Leigh Brown, N Lance Hepler, Sanjay R Mehta, Douglas D Richman, Davey M Smith, and Sergei L Kosakovsky Pond. The global transmission network of hiv-1. *The Journal of infectious diseases*, 209(2):304–313, 2014.

[WMM+18] Joel O Wertheim, Ben Murrell, Sanjay R Mehta, Lisa A Forgione, Sergei L Kosakovsky Pond, Davey M Smith, and Lucia V Torian. Growth of hiv-1 molecular transmission clusters in new york city. *The Journal of infectious diseases*, 218(12):1943–1953, 2018.

[Won09] David Wong. The modifiable areal unit problem (maup). *The SAGE handbook of spatial analysis*, 105(23):2, 2009.

[WPF+17] Joel O Wertheim, Sergei L Kosakovsky Pond, Lisa A Forgione, Sanjay R Mehta, Ben Murrell, Sharmila Shah, Davey M Smith, Konrad Scheffler, and Lucia V Torian. Social and genetic networks of hiv-1 transmission in new york city. *PLoS pathogens*, 13(1), 2017.

[WZZ+10] William H Wheeler, Rebecca A Ziebell, Helena Zabina, Danuta Pieniazek, Joseph Prejean, Ulana R Bodnar, Kristen C Mahle, Walid Heneine, Jeffrey A Johnson, H Irene Hall, et al. Prevalence of transmitted drug resistance asso-

ciated mutations and hiv-1 subtypes in new hiv-1 diagnoses, us–2006. *Aids*, 24(8):1203–1212, 2010.

[XLB+16] Qiang Xia, Rachael Lazar, Marie A Bernard, Paul McNamee, Demetre C Daskalakis, Lucia V Torian, and Sarah L Braunstein. New york city achieves the unaids 90-90-90 targets for hiv-infected whites but not latinos/hispanics and blacks. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 73(3):e59–e62, 2016.

[Yan95] Ziheng Yang. A space-time process model for the evolution of dna sequences. *Genetics*, 139(2):993–1005, 1995.

[YR97] Ziheng Yang and Bruce Rannala. Bayesian phylogenetic inference using dna sequences: a markov chain monte carlo method. *Molecular biology and evolution*, 14(7):717–724, 1997.

[YVR+01] S Yerly, S Vora, P Rizzardi, J-P Chave, PL Vernazza, Markus Flepp, A Telenti, M Battegay, A-L Veuthey, J-P Bru, et al. Acute hiv infection: impact on the spread of hiv and transmission of drug resistance. *Aids*, 15(17):2287–2292, 2001.

[ZLZ16] Dan Zhu, Yue Li, and Chao Zhang. Automatic time picking for microseismic data based on a fuzzy c-means clustering algorithm. *IEEE Geoscience and Remote Sensing Letters*, 13(12):1900–1904, 2016.

[ZMM+15] Camila Zanluca, Vanessa Campos Andrade de Melo, Ana Luiza Pamplona Mosimann, Glauco Igor Viana dos Santos, Claudia Nunes Duarte dos Santos, and Kleber Luz. First report of autochthonous transmission of zika virus in brazil. *Memórias do Instituto Oswaldo Cruz*, 110(4):569–572, 2015.

[ZWZ20] Tao Zhang, Qunfu Wu, and Zhigang Zhang. Probable pangolin origin of sars-cov-2 associated with the covid-19 outbreak. *Current Biology*, 2020.

Curriculum Vitae

Name:	Connor Chato
Post-Secondary Education and Degrees:	University of Western Ontario London Ontario 2012 - 2017 B.Sc
Honours and Awards:	Dr. Frederick Winnett Luney Graduate Research Award 2019 Dr. P.C. Raju & Jyoti Shah Resident Research Prize 2020
Related Work Experience:	Teaching Assistant The University of Western Ontario 2018 - 2020

Publications:

- Connor Chato, Marcia L Kalish, and Art FY Poon. Public health in genetic spaces: a statistical framework to optimize cluster-based outbreak detection. *Virus Evolution*, 6(1):veaa011, 2020
- Joel O Wertheim, Connor Chato, and Art FY Poon. Comparative analysis of hiv sequences in real time for public health. *Current Opinion in HIV and AIDS*, 14(3):213–220, 2019