An investigation of disease transmission clusters with Bayesian phylogenetic clustering methods

Luc Villandré - Department of Epidemiology, Biostatistics, and Occupational Health - McGill University, Montreal

November 30, 2017

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of PhD Biostatistics

Acknowledgements

Completing this thesis would not have been possible without the advice, guidance, and financial support offered by Prof. Aurélie Labbe, my supervisor, and Prof. David Stephens, my co-supervisor. Thanks to their valuable input over the years, I learned not only to do research, but also to write manuscripts I knew were solid, both from a substantive and stylistic standpoint. I would also like to acknowledge Dr. Bluma Brenner, as well as Dr. Michel Roger, for granting me access to the Québec HIV genotyping database, whose analysis forms the core of my PhD work. Conversations with Dr. Brenner and Ilinca Ibanescu gave me a much needed boost in my understanding of HIV-1 sequencing and the practical reality of HIV-1 epidemics in Québec.

Manuscript 1 was written in close collaboration with Prof. Tanja Stadler, who I wish to thank sincerely for supporting my ThinkSwiss application, and welcoming me in her team. I learned a lot thanks to her extensive knowledge of phylogenetics. Her kindness, as well as that of the students at the D-BSSE, made my stay in Basel eye-opening, instructive, and fun.

I would also like to express heartfelt gratitude to David J. Vasilevsky, who was kind enough to let me take advantage of his extensive skill in programming to write the software used in manuscript 1. What he taught me about Linux and object-oriented programming in the last decade or so also had a profound influence on the coding work done for manuscripts 2 and 3. His professionalism, as well as his focus on code clarity and brevity, still inspire me to this day.

I am also deeply grateful to Eric van de Sande for patiently helping me with the re-design of the C++ modules in the implementation of the algorithm that forms the core methodological contribution of this thesis. The structure he proposed paved the way for

vast improvements in computational efficiency, that made the analyses in manuscript 3 possible. I should also thank Calvin Ference, who was kind enough to provide useful tips on parallelization and let me use some of his code, and Naoto Hieda, who on many occasions helped me greatly by answering my many questions on C++.

I must also mention the contribution of the McGill Department of Epidemiology, Biostatistics, and Occupational Health, that helped fund my travels, and provided an environment in which I could thrive as a researcher. I should also acknowledge the kindness of the support staff, who always went out of their way to answer my questions and help me navigate the university bureaucracy. I am also very grateful to Prof. Jinko Graham for inviting me to present my early research results at the Department of Statistics and Actuarial Science of Simon Fraser University, and generously contributing financially to make the trip possible.

The Fonds de Recherche du Québec - Santé also made my PhD possible through their *Bourse de formation doctorale* program. A lot of the computational work was done on machines administered by the Calcul Québec and Westgrid branches of Compute Canada, whose technicians kindly assisted me with the preparation of my code for running on large clusters.

Of course, I would also like to thank the many friends I have made over the years in the Department of Epidemiology, Biostatistics, and Occupational Health. They were there to cheer me up when I thought I would not make it through my PhD, and that support helped me find the courage to persevere despite the many obstacles I encountered.

Finally, I would not have reached this point if it were not for the patience and generosity of my close family: it is mainly thanks to them that I found the courage to do a PhD in the first place. In the last number of years, they provided the emotional support I needed to keep going and so, this thesis exists in great part thanks to their contribution to my life.

Abstract

A number of studies have investigated transmission clusters in the Human Immunodeficiency Virus (HIV) epidemic among Men who have Sex with Men (MSM) in the province of Québec, Canada, stressing the contribution of clusters to incidence. Studies of that type usually rely on a sample of HIV-1 genetic sequences, whose ancestry is inferred with a phylogenetic model, yielding a tree used to partition the sample. Understanding of clusters found through phylogenetic analyses is still limited, which is reflected in the many ad hoc criteria used in their estimation. This manuscript-based thesis aims to improve understanding of phylogenetic clusters, propose improvements to transmission cluster inference methods, and provide updated estimates of HIV-1 transmission clusters in Québec, by means of a thorough comparison between results from conventional approaches and the new method.

The first manuscript in the thesis addresses the issue of phylogenetic cluster interpretation. Through simulations of epidemics on several categories of contact networks, we explore the association between phylogenetic clusters, found under a variety of distance-based clustering methods, and communities, distinctive groups of densely-connected individuals in the network. We find limited overlap between clusters and communities, suggesting that a network interpretation of phylogenetic clusters may not be warranted.

The second manuscript presents a new phylogenetic clustering algorithm, *DM-PhyClus*, that readily weaves into cluster inference a clear definition of transmission clusters, resulting in a straightforward interpretation for the inferred clusters. Unlike conventional phylogenetic clustering approaches, the method does not rely on arbitrary genetic distance or clade-confidence cutpoints applied a posteriori to the estimated phylogeny. Simulations reveal that *DM-PhyClus* can outperform a number of conventional clustering methods in terms of mean cluster recovery. We apply *DM-PhyClus* to a sample of real HIV-1 sequences obtained from

the Québec HIV genotyping program database, revealing a set of clusters whose estimates are in line with the conclusions of a previous curated analysis.

The third manuscript includes a detailed clustering analysis of HIV-1 cases among MSMs based on DNA sequences collected for the Québec HIV genotyping program. We first cluster the data with two conventional approaches, maximum likelihood phylogenetic inference coupled with bootstrap estimation of confidence in clades, and pure Bayesian phylogenetic estimation, under a variety of clustering criteria and cutpoints. We then partition the sample with the help of *DM-PhyClus* and the Gap Procedure, both approaches aiming to avoid arbitrary selection of cutpoints. The analyses based on conventional methods reveal largely overlapping sets of clusters, while DM-PhyClus and the Gap Procedure propose moderately different partitions. An examination of more recently-diagnosed cases that are known to have been infected at most six months prior to diagnostic shows considerable expansion of large clusters, and hint at the emergence of a few new transmission clusters. The analyses stress the continued importance of clustering in maintaining the HIV epidemic among MSMs, and suggest that the frequency of early transmission events might explain why improvements in antiretroviral therapy have not lead to the end of the epidemic.

Résumé

Plusieurs études se sont penchées sur les grappes de transmission au sein de l'épidemie de VIH-1 parmi les hommes ayant des relations sexuelles avec d'autres hommes (HARSAH) au Québec, au Canada, mettant en lumière la contribution de ces grappes à l'incidence. Les études de ce type se fient habituellement à un échantillon de séquences génétiques de VIH-1, dont l'histoire ancestrale est inférée à l'aide d'un modèle phylogénétique, produisant un arbre utilisé pour partitionner l'échantillon. La compréhension des grappes trouvées par l'intermédiaire d'une analyse phylogénétique est toutefois limitée, ce qui se traduit par une multitude de critères arbitraires pour leur estimation. La thèse, comportant trois articles, a comme but d'améliorer la compréhension des grappes phylogénétiques, de proposer des améliorations aux méthodes actuelles d'inférence des grappes de transmission, et de fournir une mise à jour des estimés des grappes de transmission du VIH-1 au Québec, au moyen d'une comparaison rigoureuse entre les résultats d'approches conventionnelles et ceux de la nouvelle méthode.

Le premier article de la thèse traite de l'interprétation des grappes phylogénétiques. À l'aide de simulations d'épidémies sur plusieurs catégories de réseaux de contacts, nous explorons l'association entre les grappes phylogénétiques, obtenues par l'application de différentes méthodes de regroupement basées sur la distance, et les communautés, des groupes distinctifs d'individus dans le réseau. Nous remarquons une correspondance limitée entre les grappes et les communautés, et concluons qu'une interprétation des grappes phylogénétiques en termes de la structure du réseau de contacts pourrait être difficile à justifier.

Le deuxième article présente un nouvel algorithme de regroupement phylogénétique, *DM-PhyClus*, qui mêle à l'inférence des grappes une définition claire des grappes de transmission, donnant ainsi une interprétation sans ambiguïté aux grappes inférées. Contrairement aux approches conventionnelles de regroupement, *DM-PhyClus* ne nécessite pas l'application à la

phylogénie inférée de critères arbitraires de distance génétique ou de confiance en les clades obtenues. Les simulations révèlent que *DM-PhyClus* peut battre les méthodes conventionnelles sur le plan du taux moyen de détection des grappes. Nous appliquons la méthode à un échantillon de séquences véritables de VIH-1 tirées de la base de données du programme québécois de génotypage du VIH, ce qui révèle un ensemble de grappes très similaires à celles proposées par une étude précédente dont les estimés ont été partiellement validés.

Le troisième article inclut une analyse détaillée de regroupement des cas de VIH-1 parmi des HARSAH basée sur les séquences d'ADN collectées dans le cadre du programme québécois de génotypage du VIH. Tout d'abord, nous regroupons les données à l'aide de deux méthodes conventionnelles: l'inférence phylogénétique par maximum de vraisemblance, couplée avec l'estimation de la confiance en les clades par le *bootstrap*, et l'estimation phylogénétique bayésienne pure. Nous partitionnons par la suite l'échantillon à l'aide de *DM-PhyClus* et du *Gap Procedure*, deux approches cherchant à éviter la sélection arbitraire de critères de regroupement. Les analyses basées sur les méthodes conventionnelles produisent des estimés de grappes très similaires, tandis que DM-PhyClus et le *Gap Procedure* proposent des partitions modérément distinctives. Un coup d'oeil aux cas diagnostiqués récemment, et dont la date d'infection se situe au maximum six mois auparavant, met en lumière une expansion considérable des plus grandes grappes de transmission et l'émergence potentielle de quelques nouvelles grappes. Les résultats de l'étude soulignent le rôle persistant des grappes de transmission dans la survie de l'épidémie de VIH parmi les HARSAH. De plus, ils suggèrent que les événements de transmission hâtifs expliqueraient pourquoi les améliorations aux traitements antirétroviraux n'ont pas mené à une résorption de l'épidémie.

Contents

List of Figures

List of Tables

Acronyms

AIC: Akaike Information Criterion

AIDS: Acquired Immunodeficiency Syndrome

ARI: Adjusted Rand Index

ART: Antiretroviral Therapy

BA: Barabasi-Albert

BIC: Bayesian Information Criterion

CP: ClusterPicker

CRF: Circulating Recombinant Form

DNA: Deoxyribonucleic Acid

DP: Dirichlet Process

EM: Expectation-Maximization

ER: Erdos-Renyi

GTR: General Time Reversible

HAART: Highly Active Antiretroviral Therapy

HIV: Human Immunodeficiency Virus

HKY85: Hasegawa-Kishino-Yano 1985

IDU: Injection Drug User

IS: Importance Sampling

JC69: Jukes-Cantor 1969

JTT: Jones, Taylor, and Thornton

K80: Kimura 1980

MAP: Maximum Posterior probability
MCMC: Markov Chain Monte Carlo
MH: Metropolis-Hastings
ML: Maximum Likelihood
MRCA: Most Recent Common Ancestor
MSM: Men who have Sex with Men

NJ: Neighbour-Joining
NNI: Nearest-Neighbour Interchange

PAM: Percent Accepted Mutation
PHI: Primary HIV Infection
PR: Protease
PYP: Pitman-Yor process

RT: Reverse Transcriptase

SD: Standard Deviation
SE: Standard Error
SHCS: Swiss HIV Cohort Study
SIR: Susceptible-Infected-Recovered
SIS: Sequential Importance Sampling
SMC: Sequential Monte Carlo
SPR: Subtree Pruning and Regrafting
STI: Sexually-Transmitted Infection

TasP: Treatment as Prevention
TN93: Tamura-Nei 1983

URF: Unique Recombinant Form

WAG: Whelan and Goldman

WHO: World Health Organization

WPGMA: Weighted Pair-Group Method of Analysis

WS: Watts-Strogatz

Contribution of authors

This thesis includes an overview of notions pertaining to HIV-1 and related epidemics, contact network modelling, phylogenetic inference, and phylogenetic clustering. Its original contribution to knowledge in the field is threefold.

(1) It clarifies the link between phylogenetic clusters and the contact network serving as a substrate for epidemics,
(2) It proposes a method that overcomes one of the main weaknesses in conventional methods for clustering HIV-1 sequences, namely their reliance on poorly-understood cutpoints, whose selection is often arbitrary and loosely explained,
(3) It provides an up-to-date portrait of clustering in the HIV-1 epidemic among men who have sex with men in the province of Québec, Canada.

Luc Villandré (LV) wrote all chapters in the thesis, which David A. Stephens (DAS) and Aurélie Labbe (AL) thoroughly reviewed. LV wrote the software used to produce results for Chapters 3, 4, and 5. DAS and AL supervised the thesis work from beginning to end, proposing and fleshing out several analyses in Chapters 3, 4, and 5, pointing out shortcomings in the results and presentation throughout. DAS identified the core methodological issues with conventional phylogenetic clustering methods that LV addressed in the thesis. LV, AL, and DAS jointly formulated the algorithm used to solve the identified problems, presented in Chapter 4. Bluma G. Brenner and Michel Roger provided the HIV-1 sequencing data, and reviewed Chapters 4 and 5.

LV and Tanja Stadler (TS) jointly formulated the idea for the analyses in Chapter 3. TS supervised the work for that chapter, reviewed it, and helped LV gain access to the Swiss HIV Cohort Study (SHCS) data. Roger Kouyos and Huldrych Günthard provided the SHCS data used in the analyses for Chapter 3, which they also reviewed.

CHAPTER 1

Introduction

1. The HIV-1 pandemic

Since the World Health Organization (WHO) initiated surveillance of Human Immunodeficiency Virus (HIV) in 1983, conditions related to Acquired Immunodeficiency Syndrome (AIDS) have claimed the lives of more than 35 million people. In 2015, an estimated 36.7 million people were living with HIV/AIDS worldwide, most of them in Sub-Saharan Africa [67]. Thanks to major improvements in Antiretroviral Therapy (ART), the prognostic of people diagnosed with HIV/AIDS has improved tremendously, increasingly turning HIV-1 into a manageable chronic condition. Nevertheless, in 2015, life expectancy of HIV-positive individuals with access to care was still 8 years under that of uninfected people [83]. ART's side-effects can also be debilitating in some cases, leading to imperfect adherence [8, 115].

2. HIV-1 among men who have sex with men (MSM) in Canada

In Canada, an estimated $75,500$ people were living with HIV-AIDS in 2014, among which $39,630$ belong to the Men who have Sex with Men (MSM) risk category. HIV-1 prevalence in MSMs in major cities ranged from an estimated 11% in Ottawa to over 23% in Toronto, making HIV-1 a major public health concern in large urban areas [3].

3. HIV-1 genotyping programs

Imperfect adherence to ART may lead to the emergence of drug resistance, and the need to detect drug-resistant subspecies and monitor their transmission has lead to the onset of HIV-1 genotyping efforts. The Québec HIV genotyping program and Swiss HIV Cohort Study (SHCS), for example, have created large HIV sequence databases, making it possible to obtain a very accurate molecular description of HIV epidemics in their regions of coverage.

4. Transmission clusters

Phylogenetics is the study of the ancestral relationships between genetic sequences, the ancestry being modelled with a tree structure known as a *phylogeny*. Phylogenetic analyses of HIV-1 sequencing data from different HIV-positive MSMs have revealed the existence of distinctive sets of genetically-close sequences, so-called *transmission clusters*, which may result from transmission cascades known as *quick transmission chains* [19]. The presence of clustering in an HIV-1 epidemic has implications for prevention strategies, insomuch as it might reasonably stress the major contribution of Primary HIV Infection (PHI) and recent stage infection to incidence [18]. Successive clustering analyses of data collected for the Québec HIV genotyping program also indicate that large transmission clusters may be the driving force behind the epidemic, with 51% of newly-infected MSMs in Montreal belonging to a large transmission cluster in 2011, compared to 25% in 2005 [18]. Understanding the transmission dynamics of HIV among MSMs therefore requires identifying the factors behind cluster expansion.

Partitioning a sample of genetic sequences using an inferred phylogeny requires a cluster definition, usually combining an arbitrary within-cluster genetic distance criterion with a minimum confidence requirement for the existence of the cluster. The lack of any convention regarding transmission cluster estimation from phylogenies has lead to researchers using a large array of criteria, whose implications are unclear [94]. What transmission clusters can teach us about HIV transmission networks is also an open question. More importantly, phylogenetic estimation does not aim first and foremost to produce cluster estimates, but rather, an estimate of the sample's ancestry, which is then used as a tool in the partitioning. Such *post hoc* cluster estimation procedures make cluster point estimates much harder to interpret.

5. Objectives

This manuscript-based thesis has three main objectives,

(1) Clarify the meaning of transmission clusters in terms of contact network structure,

(2) Propose a new phylogenetic clustering method that results in straightforward inference for transmission clusters,
(3) Apply the new clustering algorithm to the MSM sequencing data in the Québec HIV genotyping program database to update and improve current cluster estimates, obtained using conventional approaches.

The work is split into three manuscripts, each of which addresses one of the stated objectives.

6. Manuscript overview

6.1. Manuscript 1: "Assessment of Overlap of Phylogenetic Transmission Clusters and Communities in Simple Sexual Contact Networks: Applications to HIV-1". Sexual transmission of HIV among MSMs in the province of Québec occurs on an intricate sexual contact network, whose properties may affect epidemic dynamics. Mapping transmission clusters obtained from viral phylogenies onto the sexual contact network graph may reveal links between network structure, more specifically so-called *communities*, and transmission clusters. The first manuscript aims to verify the latter statement in a simulation setting, and consequently provide a more intuitive understanding of clusters.

In a nutshell, we simulate epidemics on different random sexual contact networks under the assumption of a simple diagnostic scheme. For each epidemic, we track viral transmission and infection diagnostic, producing a phylogeny that we then use to cluster the resulting sample of diagnosed individuals. We finally measure overlap of phylogenetic clusters with communities.

The study revealed limited correspondence between phylogenetic clusters and communities, mainly attributable to the way phylogenetic clusters are conventionally defined. Therefore, we concluded that the intuition that phylogenetic clusters map precisely onto network communities was often misleading, but that the phylogeny could still be helpful in detecting community structure, provided an alternative cluster definition be proposed.

6.2. Manuscript 2: "DM-PhyClus: A Bayesian phylogenetic algorithm for infectious disease transmission cluster inference". Transmission cluster inference from phylogenetic estimates usually relies on an *ad hoc* within-cluster *patristic distance* requirement, where patristic distance between any two sequences is computed by summing branch lengths along the shortest path between the corresponding tips in the tree. Such distance criteria are often paired with a confidence threshold for inferred clades, resulting in non-straightforward inference. Clusters obtained by applying those criteria are thought to result from quick transmission chains, albeit accidentally. The manuscript aims to propose a new Bayesian phylogenetic clustering method, *DM-PhyClus*, based on a new cluster definition. A cluster is now a set of sequences supported by a phylogeny with short mean branch lengths. We show through simulations that DM-PhyClus is on average more successful at recovering transmission clusters than more conventional approaches, under a variety of prior assumptions. We also apply the algorithm to a sample of 526 sequences from patients belonging to the MSM risk category in the Québec HIV genotyping program sequence database, confirming the results of a previous analysis largely based on heuristics.

6.3. Manuscript 3: "Characterizing HIV-1 transmission clusters among men who have sex with men in Quebec, Canada". Previous studies have revealed the existence of several large transmission clusters in the HIV-1 epidemic among MSMs in the province of Québec [19, 20]. Manuscript 3 presents an up-to-date clustering analysis of MSM sequences in the Québec HIV genotyping program database. We compare cluster estimates obtained following the application of the conventional maximum likelihood and Bayesian methods to those produced by DM-PhyClus and the Gap Procedure [130], that both aim to simplify phylogenetic clustering by avoiding arbitrary cutpoint selection. All methods reveal fairly similar cluster estimates, with DM-PhyClus, and the Gap Procedure to a lesser extent, proposing the most distinctive partitions. Clustering patterns in sequences obtained from cases in the Primary HIV Infection (PHI) stage, corresponding to the six months following seroconversion, highlight the sizeable contribution of large clusters to recent incidence and the emergence of several new transmission clusters. We conclude by emphasising the link between early transmission events and clustering, the understanding of which is crucial to inform public health strategies to curb transmission.

CHAPTER 2

Background

As an introduction to the core ideas covered in this thesis, we present a summary overview of several topics pertaining to Human Immunodeficiency Virus (HIV)-1 and phylogenetics. We start with a brief genetic and epidemiological description of HIV-1, followed by a reminder of fundamental concepts in network theory and of random network models. We then move on to formally introduce phylogenies from a genetic, probabilistic, and statistical standpoint. We describe a number of algorithms used for phylogenetic reconstruction and inference, and finally, we tackle the subject of clustering, first presenting model-based and distance-based clustering, then illustrating how those approaches are applied to phylogenetic data.

1. The Human Immunodeficiency Virus

1.1. The pandemic. Thirty-two years after the World Health Organization (WHO) started monitoring the Acquired Immunodeficiency Syndrome (AIDS) pandemic, HIV prevalence is still a major public health burden in many regions of the world. In 2015, most people living with HIV were found in sub-Saharan Africa, followed by southeast Asia [4]. In Canada, more than $75,500$ people are HIV-positive, most of them belonging either to the men who have sex with men or injection-drug user risk category, and in 2014, incidence estimates ranged between $2,570$ and $3,200$ [3]. Improvements in Antiretroviral Therapy (ART) have drastically improved the long-term prognosis of HIV, although an 8-year gap in curtate life expectancy may still remain [83]. Further, side-effects of ART are known to be debilitating, leading to imperfect compliance.

1.2. Classification and landmarks of HIV-1. HIV is a lentivirus of the retrovirus family, known for its large degree of genetic diversity. It is classified into groups, types, and subtypes. HIV type 1, abbreviated *HIV-1*, the viral type responsible for the pandemic,

belongs to group M, for "Major", and originates from several zoonotic transmissions from *Pan Troglodytes Troglodytes* [57].

HIV-1 is further broken down into genetic subtypes, denoted by letters ranging from A to J, and recombinants. Common recombinants are denoted CRF-xy, for "Circulating Recombinant Forms" mixing subtypes x and y. For example, a recombinant between subtypes A and E, denoted CRF-AE, represents a majority of infections in southeast Asia [57]. Rare recombinants are instead denoted URF-xy, where the U stands for "Unique". Although subtype C represents roughly half of infections worldwide, in Canada, subtype B is predominant. However, due to new introductions of the virus from other regions of the world, genetic diversity has been increasing [21].

The HIV genome comprises 9 genes. The *pol* gene codes for enzymes essential for RNA transcription and replication, which are targeted by ART. More specifically, ART works by disrupting the action or production of the enzymes Protease (PR), Reverse Transcriptase (RT), and integrase, coded by the *PR/RT* and *int* regions of the *pol* gene. The *env* gene codes for the viral membrane and lets the virus target and bind to lymphocytes. Finally, the *gag* (group-specific antigen) gene codes for several structural core proteins. The other genes, *rev*, *tat*, *vpr*, *vif*, *nef*, and *vpu* are regulatory. They control the virus's metabolism and life cycle, and help increase infectivity [1].

1.3. Genotyping HIV-1. Due to the virus's high mutation rate, in the absence of perfect compliance, selective pressure created by ART may lead to the emergence of drug resistance. The need to monitor the transmission of drug-resistant strains has lead to the creation of large HIV-1 genotyping programs [19, 117]. Sequencing efforts for HIV-1 systematically include the *pol* gene, more specifically the *PR/RT* region, since its action is directly affected by ART. A list of sites associated with drug resistance is available, and updated periodically [134].

2. Contact networks

2.1. Networks and epidemics. Classical epidemic models work under the *random mixing* assumption. In a random mixing population, each individual has a small and equal chance of coming into contact with any other individual, and knowing the actual contact structure is unnecessary. The original Susceptible-Infected-Recovered (SIR) model, for example, relies on the random mixing assumption and consisted of differential equations representing the expected temporal variation in the number of subjects in each category [69].

For epidemics such as HIV-1 however, the random mixing hypothesis fails to hold: subjects have a fixed number of contacts, usually not sampled uniformly among individuals in the population. In such a case, employing network models is necessary: they work by constraining transmissions to the "neighbourhood" of infected subjects. Network structure matters for public health authorities, as it determines to what extent a vaccination campaign can help curb epidemic spread. Under certain network types, random vaccination will be systematically ineffective in preventing epidemic outbreaks [69].

2.2. Graph representation. Transmission of HIV-1 through sexual contact or needle-sharing occurs on an intricate contact network, which we represent with a *static undirected* graph, whose *vertices*, or nodes, correspond to individuals and *edges*, to a transmission route between them, cf. Figure 1. We denote the graph structure with an *adjacency matrix*, with element (i, j) taking value 1 if node i connects to node j, and 0 otherwise. Unless *loops* are allowed, that is, edges can connect a node to itself, diagonal elements in the matrix take value 0. In a static network, an edge is permanent: in other words, it represents a contact that at one point in time made transmission of the virus possible. In *undirected* graphs, edges are not assigned a direction, and it follows that transmission between two connected individuals, called *neighbours*, can originate from any of them, and the adjacency matrix is symmetrical. Directional edges could help constrain the number of potential paths followed by the virus, but would be hard to infer in the case of HIV-1, where sensitive information would need to be collected [109].

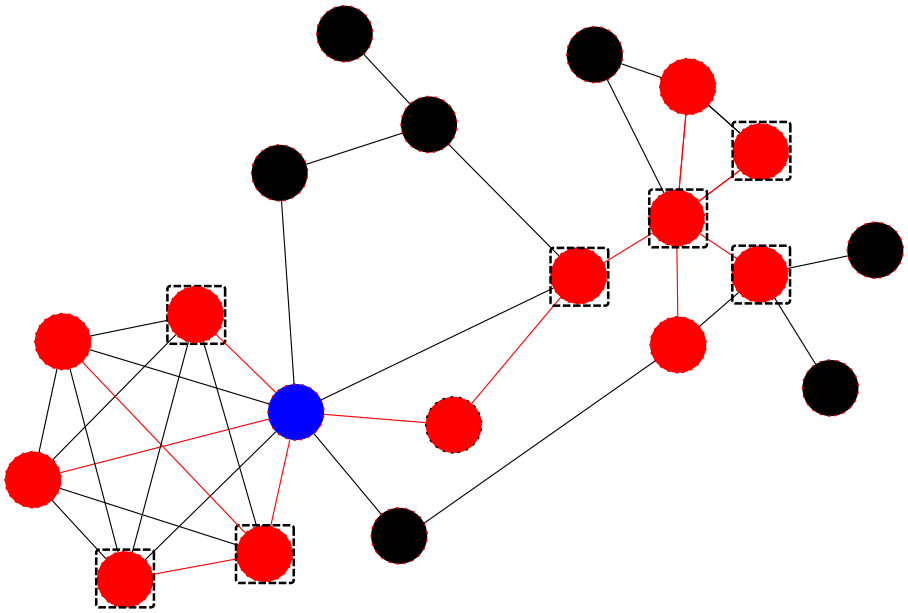

FIGURE 1. **A graph representing a contact network between 20 individuals**. The vertices in red represent infected individuals and the edges in red, the connections used by the virus to infect new subjects. A vertex in a box represents a diagnosed infection. The individual represented by the vertex in blue contracted the virus first and triggered the epidemic. We call the collection of edges and vertices in red, plus the vertex in blue, the *transmission network*. The collection of all edges and vertices is called the *contact network*.

Embedded within this graph is the transmission network, representing solely HIV-positive individuals and the edges along which HIV transmission occurred. The transmission network is a snapshot of the epidemic at a given point in time. Because the graph is undirected, we cannot deduce from it a deterministic history of the epidemic.

2.3. Network characteristics. Contact networks are characterised, among other things, by their size, *degree distribution*, *clustering*, and *modularity*.

2.3.1. *Degree distribution.* We define a node's *degree* as the number of edges connecting to it. For example, in Figure 1, the blue node has degree 9. In other words, it has 9 first-degree neighbours.

2.3.2. *Clustering coefficient.* A network's clustering indicates its level of interconnectedness. The *clustering coefficient* usually consists in a ratio of the number of triangles within

the network to the number of connected triples [69]. If we denote the adjacency matrix $\boldsymbol{A}$, we have that the clustering coefficient takes value

$$C = \frac{\mathrm{Tr}(\boldsymbol{A}^3)}{\sum_{i,j} A_{i,j}^2 - \mathrm{Tr}(\boldsymbol{A}^2)},$$

where Tr denotes the trace of the matrix. For example, the subnetwork represented by the hexagon on the left side of Figure 1 has clustering coefficient 1.0, since all possible sets of three edges are interconnected. The entire network, on the other hand, has a clustering coefficient equal to 0.61.

2.3.3. *Modularity.* We define *communities* as distinctive, non-overlapping sets of densely-interconnected nodes in a network graph. We can rate the quality of any partition of vertices into communities with the so-called *modularity score*, which is proportional to the difference between the number of edges falling within sets in the identified partition and the expected number in a same-size network with randomly-placed edges [87]. Let element (i, j) of matrix $\boldsymbol{e}$ denote the proportion of edges connecting components i and j in the partition. We have that the trace of $\boldsymbol{e}$ is equal to the proportion of edges within the identified components. The modularity score is then expressed

$$\text{Mod.} = \mathrm{Tr}(\boldsymbol{e}) - \sum_{i,j} e_{i,j}^2,$$

where, as expected, the second term corresponds to the proportion of edges that would fall within the identified components if the edges were placed at random.

The best community estimates are found by identifying the partition that maximizes the modularity score. However, even that partition may still contain components that are not obviously distinctive: there is no absolute criterion for calling any set of nodes a "community". Partitions are merely more or less modular. Since the number of ways in which vertices can be split into non-overlapping sets grows quickly with network size, community detection usually relies on heuristic algorithms.

The walktrap algorithm [91], for example, is based on the intuition that if a set of vertices forms a community, then any short random walk launched from within it is unlikely to leave it. Accordingly, the walktrap algorithm works by performing a large number of

Term	Definition
Degree	The number of edges connecting to a node.
x-th degree neighbour	(With respect to an arbitrary node) All nodes reachable by crossing exactly x different edges.
Clustering coefficient	Measure of the network's interconnectedness, equal to the ratio of connected trios of nodes to the total number of trios in the network.
Modularity	The extent to which the network comprises distinctive components.
Community	A subset of a network comprising densely-interconnected nodes with comparatively few connections with nodes outside the set.
Small-world property	Can be said of networks with the mean length of shortest path between any two nodes growing logarithmically with network size.
Connection probability	In network generation, probability that any two nodes become first-degree neighbours.
Lattice	Network with no disconnected component whose vertices all have the same number of first-degree neighbours.
Rewiring probability	In Watts-Strogatz network, probability that the edge between any two neighbours is disconnected at one end to form a bridge between more distant regions in the lattice.
Preferential attachment	Phenomenon guiding network formation, where the probability that an individual already in the network attracts a new (disconnected) contact increases with that individual's current number of connections (degree).
Scale-free network	Network whose degree distribution follows a power law.

TABLE 1. **Glossary of terms used in network analysis.**

random walks on a graph, usually involving between two and five steps, from a starting point selected at random. Using these short random walks, a distance matrix is derived, which can then be employed in a standard hierarchical clustering algorithm, such as those described in section 10.2. The cutpoint for the dendrogram is selected in such a way as to optimise the modularity score. Other popular community detection algorithms include label propagation, Girvan-Newman, spinglass, and Louvain [91, 96].

2.4. Network models. Real contact networks rarely have a straightforward structure, and estimating them can be challenging [69]. Consequently, to study the effects of network characteristics on epidemics, we often rely on simpler network models, whose properties are well known. Such models include those of Erdos-Renyi [40], Watts-Strogatz [132], and Barabasi-Albert [12]. Table 1 presents a list of terms used in describing contact networks such as those introduced in the upcoming sections.

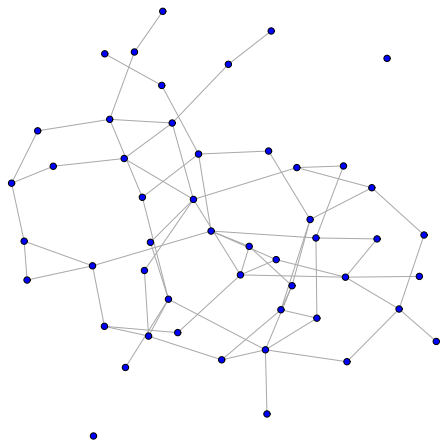

FIGURE 2. **Graph representing an Erdos-Renyi network with a** 5% **connection probability**.

2.4.1. *Erdos-Renyi networks.* Erdos-Renyi (ER) graphs are generated by randomly connecting any two vertices based on a draw from a Bernoulli distribution with an arbitrary success rate, cf. Figure 2. In large networks, ER graphs have an approximately Poisson-distributed degree distribution, with mean equal to (connection probability)*(network size - 1). Such networks tend to have low clustering, and long *mean shortest path lengths*, that is, the average number of edges along the shortest path between any two vertices.

2.4.2. *Watts-Strogatz networks.* Watts-Strogatz (WS) networks, also known as "small-world" networks, are formed by randomly rewiring edges in a *lattice*, cf. Figure 3. A lattice is a graph consisting solely of nodes with an equal degree. The lower part of Figure 3, for example, is an intact lattice with degree 4. By "rewiring", we mean that an edge is disconnected at one end and reconnected to a randomly-selected vertex. The rewiring creates

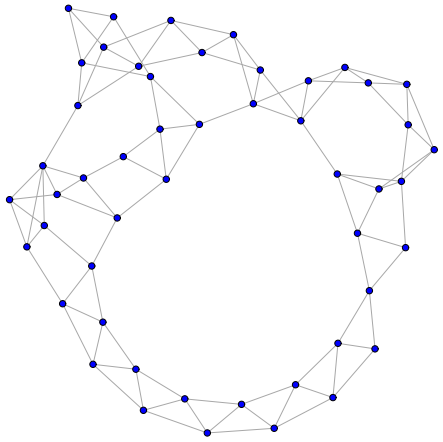

FIGURE 3. **Graph representing a Watts-Strogatz network with a 5% rewiring probability**.

bridges between different regions of the lattice, granting it the "small-world property", thus dramatically reducing the mean shortest path lengths. Further, such graphs are characterised by a relatively high clustering coefficient and very low variance in the degree distribution.

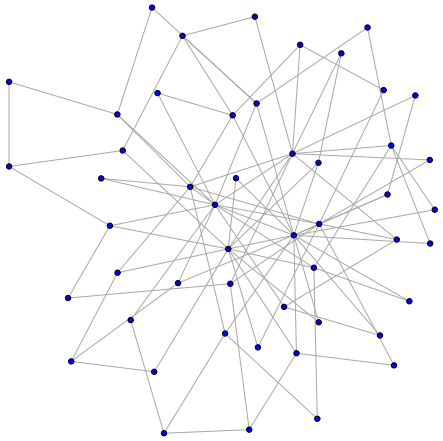

FIGURE 4. **Graph representing a Barabasi-Albert network**. Several vertices in the center have a distinctively high degree.

2.4.3. *Barabasi-Albert networks.* Barabasi-Albert (BA), or "scale-free", networks result from *preferential attachment*, cf. Figure 4. When preferential attachment guides network formation, new subjects are more likely to form connections to already well-connected subjects. In the graph, we therefore observe a large number of vertices with a small degree, and a few so-called "hubs", nodes with a disproportionately high degree. For example, the degree of central vertices in Figure 4 is much larger than that of peripheral nodes. It follows that such networks produce a heavy right-tail, termed *power-law*, degree distribution and low clustering [12]. In other words,

$$P(D = d) \propto d^{-\psi},$$

where D is node degree and ψ is a positive constant.

2.4.4. *Network models and real contact networks.* All three network models, while unrealistic, have desirable features for the modelling of epidemics. Although ER networks tend to have overly long mean shortest path lengths, they are a valuable alternative to the classical SIR model, with a lower growth rate in the number of infected subjects [69]. They might offer a reasonable simplification to describe the transmission of airborne diseases such as influenza or measles.

WS networks allow for bursts in the number of infected subjects, as crossing a bridge may lead to a drastic increase in the number of susceptible subjects. Their high clustering is also noteworthy, because of its effect on transmission dynamics. Human social networks, for instance, are believed to exhibit the small-world property [124]. The very small variance in their degree distribution remains problematic though.

Finally, preferential attachment may be an important phenomenon guiding the formation of sexual contact networks [80]. It follows that BA networks are well-suited for modelling the spread of HIV-1, even though their very low clustering coefficients tend to be a weakness.

More realistic network models have been proposed to represent empirical findings on network structure. For example, the model of [123] is characterised by high clustering, allows for *assortative mixing*, that is, a tendency of subjects to form connections with similar

individuals, and has a power-law degree distribution. Accordingly, it has small mean shortest path lengths that grow slowly with network size.

3. Features of a phylogeny

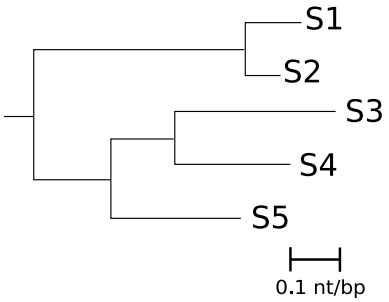

FIGURE 5. **Phylogeny for a sample of five viral DNA sequences obtained from five different subjects, labelled S1 to S5**. External nodes, at the right end of the tree, are called *tips* or *leaves*. The internal node at the left end of the phylogeny is called the *root*. Distance is measured in expected number of nucleotide substitutions per base pair, denoted "nt/bp". For instance, an expected 2 nucleotide substitutions per 10 base pairs separate the viral sequences from individuals S1 and S2. Lineages merge once they find a common ancestor.

Because of shared ancestry, sequence data sampled in different organisms are not marginally independent and identically distributed (*iid*). Phylogenetic approaches usually represent the ancestral relationship between such data with a bifurcating tree structure known as a *phylogeny*.

Figure 5 is an example of a phylogeny for five DNA sequences. In that tree, branch lengths, expressed in expected nucleotide substitutions per base pair (nt/bp), measure *genetic distance* between DNA sequences. Genetic distance is defined broadly as the expected number of substitutions per site between two sequences. By making a molecular clock assumption, branch lengths can be converted to a more intuitive time unit such as month or day. The *topology* consists of hierarchically-nested sets of tips called *clades*. A clade is defined as a set that comprises all and only leaves from a given ancestral node. In Figure 5 for

example, $\{S3, S4, S5\}$ and $\{S3, S4\}$ form clades, but not $\{S1, S5\}$ or $\{S3, S5\}$. The combination of the topology with branch lengths provides a complete unambiguous representation of the phylogeny.

4. Genetic distance estimation

We can naively estimate distance between two aligned sequences by counting the number of differences across loci and dividing by sequence length. For example, under that model, nucleotide sequences AATA and ACTA are separated by 0.25 nt/bp. This measure corresponds to the so-called *Hamming* or *raw* distance [53]. Because of potential multiple substitutions at the same site, Hamming distances tend to underestimate the true distance [136]. We can refine genetic distance estimation by assuming that substitutions occur according to a continuous-time Markov chain, with rate matrix Q and limiting probabilities $\boldsymbol{\pi}$.

The simplest model of DNA evolution is called Jukes-Cantor 1969 (JC69). It assumes that all transitions occur with equal rate λ. In other words, we have transition rate matrix Q,

$$Q = \begin{bmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{bmatrix}$$

with rows and columns corresponding to states T, C, A, and G, respectively. Under this model, limiting probabilities are uniformly equal to $1/4$ and a method of moments estimate of distance can be obtained,

$$\hat{d} = -\frac{3}{4} \log\left(1 - \frac{4}{3}\hat{p}\right),$$

with $\hat{p}$ corresponding to the Hamming distance estimate.

Since *transitions* are known to occur at higher rates than *transversions*, a common extension to JC69, called Kimura 1980 (K80) [71], involves letting *transition* and *transversion* rates differ in the Q matrix. *Transitions* are mutations from one purine to another purine or from one pyrimidine to another pyrimidine, that is, A $\leftrightarrow$ G or T $\leftrightarrow$ C, while *transversions*

are mutations from a purine to a pyrimidine, or vice versa, that is, A $\leftrightarrow$ C, T $\leftrightarrow$ G, A $\leftrightarrow$ T, or C $\leftrightarrow$ G. The K80 rate matrix then corresponds to,

$$Q = \begin{pmatrix} -(\alpha+2\beta) & \alpha & \beta & \beta \\ \alpha & -(\alpha+2\beta) & \beta & \beta \\ \beta & \beta & -(\alpha+2\beta) & \alpha \\ \beta & \beta & \alpha & -(\alpha+2\beta) \end{pmatrix},$$

with α and β being the transition and transversion rates, respectively.

Let S and V correspond to the proportion of sites with transitional and transversional differences, respectively. The method of moments distance estimate is now equal to,

$$\hat{d} = -0.5\log(1-2S-V) - 0.25\log(1-2V),$$

and limiting probabilities are still uniformly 1/4.

Multiple other extensions have been proposed: Hasegawa-Kishino-Yano 1985 (HKY85), Tamura-Nei 1983 (TN93), and the General Time Reversible (GTR) [54, 119, 102] model. The GTR model is the most flexible: its substitution rate matrix has nine parameters, corresponding to all distinct substitution rates (six in total) and limiting probabilities (three, since they are constrained to sum to 1.0). Note that method of moments estimates may not be readily available for more complicated models. In that case, maximum likelihood estimates can be used instead.

Pairwise estimation of genetic distance can however result in conflicting estimates when samples of three sequences or more are involved, hence the need for phylogenetic models, that help overcome this problem by permitting joint estimation of all pairwise genetic distances.

5. Phylogenetic likelihood

Felsenstein's *tree-pruning*, or *tree-peeling*, algorithm is the standard method for computing phylogenetic likelihoods [136, 42]. Its computational complexity is linear in the number of sequences and loci. We illustrate it in Figure 6. The notation was suggested by [136].

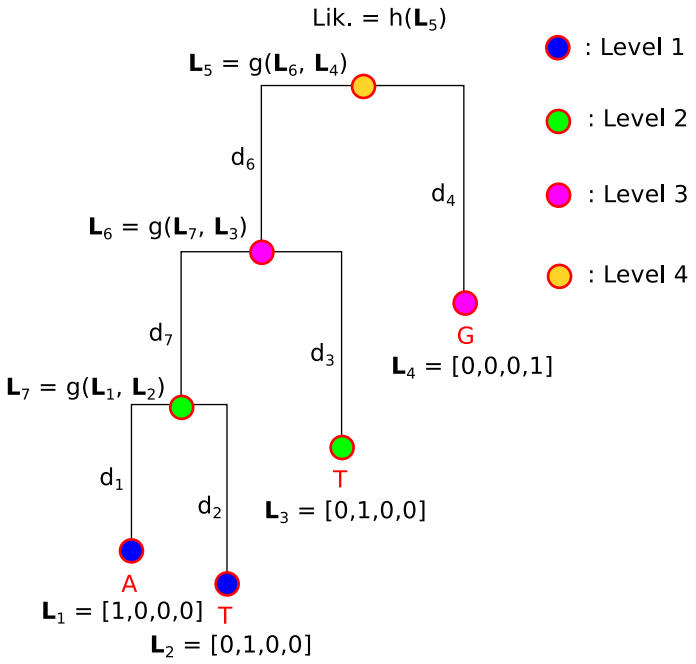

FIGURE 6. **An illustration of Felsenstein's tree-pruning algorithm for one locus**. The algorithm starts by assigning unit vectors $\boldsymbol{L}_1$ and $\boldsymbol{L}_2$ to nodes 1 and 2, supported by branches of lengths d_1 and d_2, respectively, found in level 1 of the phylogeny. The value for (non-unit) vector $\boldsymbol{L}_7$ is obtained by applying Eq. 1. The substitution rate matrix and limiting probabilities are assumed known. We then move up to level 2. Since node 3 is a tip, we assign to it a unit vector corresponding to its state. We just computed the value for $\boldsymbol{L}_7$. Now that we have $\boldsymbol{L}_7$ and $\boldsymbol{L}_3$, we can compute $\boldsymbol{L}_6$. We use the same method to obtain $\boldsymbol{L}_5$. Finally, by applying Eq. 2 to $\boldsymbol{L}_5$, we obtain the required likelihood.

Let us consider an alignment of n sequences, each denoted $\boldsymbol{y}_i$, with n_l loci, supported by a bifurcating phylogeny whose nodes and *levels* are numbered 1 to $n(n+1)/2$, and 1 to n_{lev}, respectively. A level is a set of nodes separated by a given number of splits from the root node. The level farthest from the root is labelled 1. We arbitrarily number the root node $n+1$. For example, the phylogeny in Figure 6 has four tips ($n = 4$), and three internal nodes, for a total of seven nodes. Nodes are split between four levels ($n_{lev} = 4$), and the root node is labelled 5. We compute the log-likelihood by following Algo. 1.

Algorithm 1: Felsenstein's tree pruning algorithm.

for $l = 1, \ldots, n_l$ **do**

  for $a = 1, \ldots, n$ **do**

    Assign to tip a a length-m unit vector $\boldsymbol{L}_i^{(l)}$ ($m = 4$ for DNA alignments), with the position of the 1 determined by state $y_{a,l}$;

  end

  for $b = 2, \ldots, n_{lev}$ **do**

    List nodes at level b whose children, at level $b-1$, have been assigned vectors $\boldsymbol{L}_i^{(l)}$;

    Iterate across nodes listed in the previous step. For each of them, compute,

$$L_i^{(l)}(x_i) = \sum_{x_j} p_{x_i,x_j}(d_j) L_j^{(l)}(x_j) \sum_{x_k} p_{x_i,x_k}(d_k) L_k^{(l)}(x_k), \tag{1}$$

    where i is the node label, j and k are the labels for the children of node i, x_i is a state label that indexes $\boldsymbol{L}_i^{(l)}$, $p_{x_i,x_j}(d_j)$ is the probability that a continuous-time Markov chain transitions from state x_i to x_j in time d_j, and d_j being the length of the branch supporting node j. That transition probability depends on the assumed genetic distance model, e.g. JC69 or K80 ;

  end

  Obtain the likelihood contribution of locus l by computing,

$$h_l(\boldsymbol{\theta}) = \sum_{x_i} \pi_{x_i} L_{n+1}^{(l)}(x_i), \tag{2}$$

  where $\boldsymbol{\theta}$ is the vector of phylogenetic parameters, $\boldsymbol{\pi}$ are the limiting probabilities, and $n+1$ is the label of the root node

end

Result: Compute log-likelihood by calculating,

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n_l} \log[h_i(\boldsymbol{\theta})]. \tag{3}$$

In Figure 6, we only consider one locus, which is why we omit the (l) superscript. The nucleotides at that sole locus are A, T, T, and G, corresponding to tips 1 to 4, respectively. The first inner loop in the algorithm assigns tips 1 to 4 vectors $\boldsymbol{L}_1 = [1,0,0,0]$, $\boldsymbol{L}_2 = [0,1,0,0]$, $\boldsymbol{L}_3 = [0,1,0,0]$, and $\boldsymbol{L}_4 = [0,0,0,1]$. We then move to the second inner loop. We start at level 1, formed by the blue tips, numbered 1 and 2. The parent node, at level 2, is numbered 7. We know $\boldsymbol{L}_1$, $\boldsymbol{L}_2$, d_1, and d_2 and so, we obtain $\boldsymbol{L}_7$ by using Eq. 1. We obtain the required transition probabilities $p_{.,.}(.)$ by computing the exponential of the, assumed known, transition rate matrix scaled by the branch length, e.g. for tip 1, we have $p_{A,i}(d_1) = \exp(Qd_1)|_{A,i}$. We have that x_i indexes vector $\boldsymbol{L_i}$ and can take one of four values, A, T, C, or G.

Node 7 is the only internal node at level 2. We therefore move up to level 3, where we find node 6, the parent of nodes 7 and 3. Since both $\boldsymbol{L}_7$ and $\boldsymbol{L}_3$ are known, we obtain $\boldsymbol{L}_6$ exactly as we did for $\boldsymbol{L}_7$. Since node 7 is the only internal node at level 3, we move up to level 4, where root node 5 is located. We compute $\boldsymbol{L}_5$, which concludes the second inner loop. We then use Eq. 2 to obtain the log-likelihood contribution of the locus. Note that the limiting probabilities are also assumed known. More specifically, we have,

$$h_1(\boldsymbol{\theta}) = \sum_{x_i=\{A,T,C,G\}} \pi_{x_i} L_5^{(1)}(x_i).$$

Since the simple alignment used in the example only has one locus, that concludes the outer loop, and we have that $l(\boldsymbol{\theta}) = \log[h_1(\boldsymbol{\theta})]$. If the alignment had loci 1 to n_l, we would instead move to locus $l = 2$, then $l = 3$, ..., then $l = n_l$ and for each of them, repeat the previous steps until we get $\log[h_l(\boldsymbol{\theta})]$. We would then use Eq. 3 to obtain the required log-likelihood.

5.1. Mutation rate variation. In the example above, we assume that the cumulative mutation rate at each locus is the same. That is however unrealistic: in real genotyping data, certain sites evolve much faster than others [136]. We can incorporate among-sites mutation rate variation into the likelihood by treating it as a random effect. The assumption that each site has its own mutation rate is cumbersome from a computational standpoint and so, it is common to assume instead that each site belongs to one of a few rate variation categories.

In that case, Eq. 1 becomes,

$$L_{i,r}^{(l)}(x_i) = \sum_{x_j} p_{x_i,x_j}(\xi_r d_j) L_{j,r}^{(l)}(x_j) \sum_{x_k} p_{x_i,x_k}(\xi_r d_k) L_{k,r}^{(l)}(x_k), \tag{4}$$

where $r = 1, \ldots, n_r$ indexes rate variation categories and ξ_r acts as a frailty parameter.

The likelihood contribution of locus l is now,

$$h_l(\boldsymbol{\theta}) = \sum_{r=1}^{n_r} P(R = r) \sum_{x_i} \pi_i L_{n+1,r}^{(l)}(x_i), \tag{5}$$

where R is a random variable giving the value of the rate variation parameter, and $l(\boldsymbol{\theta})$ is computed as before, with Eq. 3. In practice, rate variation is often assumed to follow a discrete gamma distribution with three or four components, and with equal shape and rate parameters, in order for the distribution to have mean 1. The non-zero values of the distribution correspond to ξ_r, $r = 1, \ldots, n_r$, in Eq. 4, and area under each corresponding segment is equal to one divided by n_r, and it follows that $P(R = r) = 1/n_r$ in Eq. 5. Extending Eq. 5 to the case where rate variation is assumed to follow a standard gamma distribution is straightforward. In practice however, as was noted before, this assumption is seldom made, as it involves a much heavier computational burden.

Several phylogenetic analyses also involve a fixed proportion of invariant sites. This assumption, although very common in practice, may not however be essential, as gamma-distributed rate variation with a relatively small shape parameter already allows for sites with very low mutation rates [136].

6. Rooted and unrooted trees

All models mentioned in Section 4 are *time-reversible*, that is, they all meet the detailed balance condition,

$$\pi_i p_{i,j}(d) = \pi_j p_{j,i}(d).$$

Time-reversibility precludes identification of the ancestor of the sample, the root of the tree, and accordingly, phylogenetic likelihood is invariant to root placement. It follows that, for example, moving root node 5 between nodes 6 and 7 in Figure 6 does not affect likelihood.

Under time-reversibility, without further assumptions, phylogenetic models lead to unrooted trees, like the one in Figure 7 a).

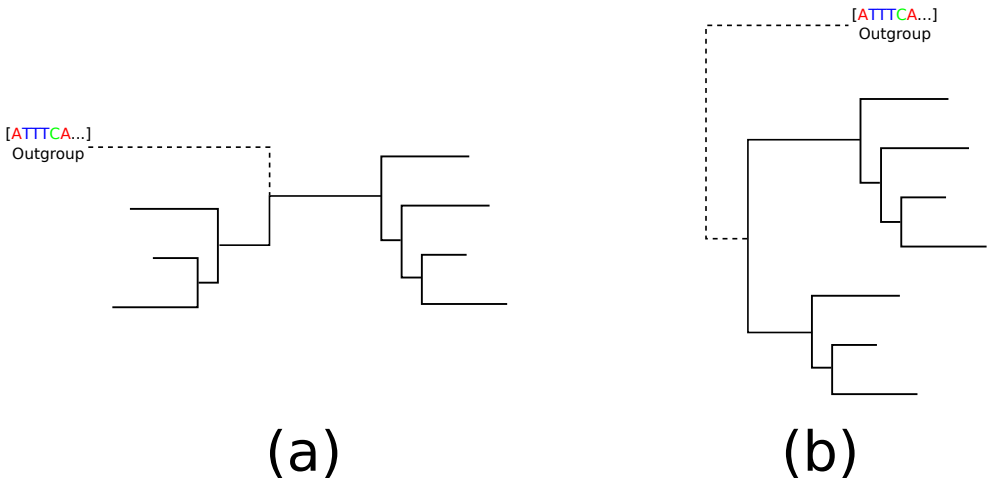

FIGURE 7. **Rooting a tree with an outgroup**. The tree in (a) is unrooted. At one of its tips, we find an outgroup sequence, that we use to place the root of the phylogeny for the sampled sequences, as in b).

Rooting the tree consists in choosing a common ancestor for sampled sequences. In order for the phylogeny to better reflect evolution in a given population, it is customary to select a sequence external to the sample as a root, called an *outgroup*. We provide an illustration in Figure 7. For example, phylogenies for HIV-1 subtype B sequence data are often rooted with a subtype C sequence.

7. Phylogenetic inference

Phylogenetic inference commonly involves estimation of the following parameters [136, 104],

(1) The topology,
(2) The branch lengths,
(3) The Markov chain substitution rate matrix and limiting probabilities,
(4) The among-sites rate variation parameters.

7.1. Tree construction. Phylogenetic inference requires at least a *tree-construction algorithm*, but many applications also involve a *tree-searching algorithm* and a method to estimate confidence in elements of the topology [60].

Tree construction produces an initial estimate of the sample phylogeny. The basic Neighbour-Joining (NJ) algorithm, an agglomerative distance-based clustering method, is a very popular tree construction method [106, 86]. It starts with a *star tree*, a tree whose tips are all connected directly to a central node, and reconstructs the phylogeny by recursively joining branches, selecting at each step the change that minimises cumulative branch length or equivalently, total genetic distance along the entire tree [136]. Algo. 2 lists the steps involved [114]. [48] proposed improvements to the algorithm: the so-called *BIONJ* algorithm is like basic NJ, except that mergers are selected as to minimise the variance of the updated distance matrix. Neighbour-joining undoubtedly owes its popularity to its simplicity and speed: the algorithm can produce a phylogeny for even large samples in a few seconds [120].

The Weighted Pair-Group Method of Analysis (WPGMA) is also commonly used for tree construction. It is another name for simple hierarchical clustering with McQuitty's method used to compute inter-cluster distances [55]. In agglomerative hierarchical clustering, we use the matrix of pairwise distances to successively aggregate elements, or sets of elements, until all elements form a single cluster, forming an *ultrametric* tree structure known as *dendrogram*. At first, each element is in a distinct cluster. At each step, the algorithm aggregates the two closest clusters, and updates the distance matrix accordingly, depending on the clustering method. Nodes in the dendrogram are placed as to reflect the distance between the merged elements. In McQuitty's method, the distance between the merged cluster M, obtained after merging clusters K and L, and any other cluster N, denoted D_{MN} is simply $(D_{KN}+D_{LN})/2$.

7.2. Tree searching. Tree construction algorithms often fail to produce an optimal estimate of the sample phylogeny, that is, a phylogeny that maximises measures such as the posterior probability or the likelihood of the phylogenetic parameters. Taking the tree construction algorithm output as a starting value, tree-searching algorithms recursively explore the phylogenetic space in order to improve the estimate. Algorithms that score all trees in

Algorithm 2: The Neighbour-Joining (NJ) algorithm.

Data: Start with D_0, a $n \times n$ matrix, n being the sample size, containing pairwise distance estimates between any two sampled sequences,

for $a = 1, \ldots, n-1$ **do**

Calculate the $\mathcal{E}$ matrix, with,

$$\mathcal{E}(i,j) = (n-2)D_{a-1}(i,j) - \sum_{k=1}^{n} D_{a-1}(i,k) - \sum_{k=1}^{n} D_{a-1}(j,k),$$

with $D_{a-1}(i,j)$ being element (i,j) in matrix D_{a-1} ;

Select for merger nodes k and l such that $\mathcal{E}(k,l)$ is the minimum value in the lower-triangular section of $\mathcal{E}$. Let the new node be labelled $n+a$;

Compute incremented distance matrix D_a with

$D_a(1:(n+a-1), 1:(n+a-1)) = D_{a-1}$ and

$$\begin{aligned} D_a(k, n+a) = D_a(n+a, k) &= 0.5 D_{a-1}(k,l) + \\ &+ \frac{1}{2(n-2)} \left[\sum_{m=1}^{n} D_{a-1}(k,m) - \sum_{m=1}^{n} D_{a-1}(l,m) \right], \\ D_a(l, n+a) = D_a(n+a, l) &= 0.5 D_{a-1}(k,l) - \\ &- \frac{1}{2(n-2)} \left[\sum_{m=1}^{n} D_{a-1}(k,m) - \sum_{m=1}^{n} D_{a-1}(l,m) \right], \\ D_a(m, n+a) = D_a(n+a, m) &= 0.5 [D_{a-1}(k,m) + D_{a-1}(l,m) - \\ &- D_{a-1}(k,l)], \quad m \neq k, l. \end{aligned}$$

end

the tree space, deemed *exhaustive*, quickly become impractical as the number of sequences increases and so, *heuristic algorithms* are commonly preferred. At each iteration, they propose a new phylogeny based on the input phylogeny, a "move in the tree space", and score it [60].

Maximum likelihood-based tree-searching algorithms are usually strict hill-climbers: they accept a proposed move if and only if it improves likelihood. Markov Chain Monte Carlo (MCMC) algorithms, on the other hand, accept the move only if a uniformly-distributed

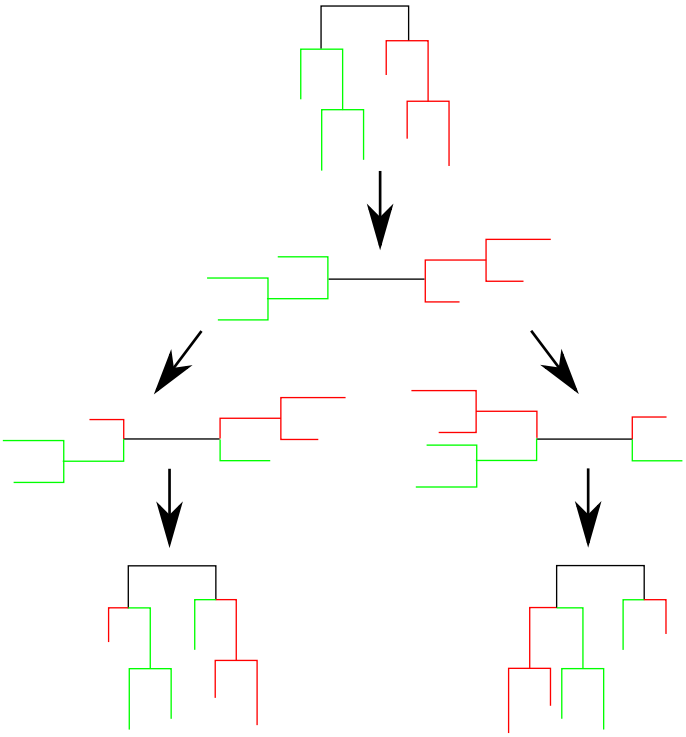

FIGURE 8. **An illustration of nearest-neighbour interchange (NNI).** We start with a rooted phylogeny. After unrooting it (line 2), the algorithm proposes transitions (line 3) by interchanging clades. Each tree with four or more tips has two distinct nearest neighbours. Finally, the algorithm restores the root (line 4).

random number falls below the Metropolis-Hastings acceptance ratio [56]. If the move is rejected, the proposed phylogeny is discarded and the input phylogeny is kept. The process is repeated until a stopping rule is activated.

7.2.1. *Tree proposal mechanisms.* Popular tree proposal algorithms include Nearest-Neighbour Interchange (NNI) and Subtree Pruning and Regrafting (SPR). We illustrate NNI in Figure 8. In SPR, a subtree is first selected at random to be pruned, and then, it is reattached to a randomly selected branch of the amputated tree. NNI and SPR propose moves in the topological space. Changes in branch lengths can then be proposed jointly, in order to limit the total length of the tree for instance [98], or independently for each branch.

Another method, used by [84] and [74], involves representing a phylogeny with a peak-and-valley graph giving distances between neighbouring leaves and their most recent common

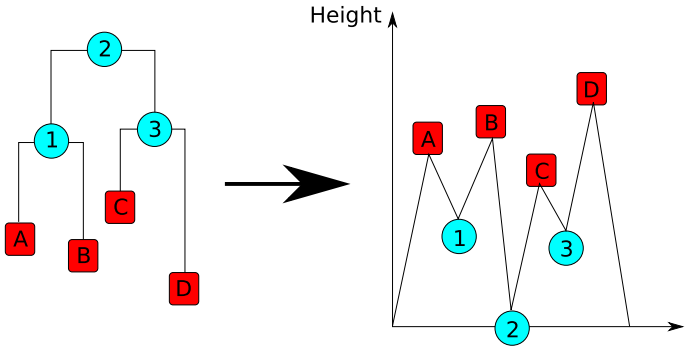

FIGURE 9. **Valley-and-peak graph for a 4-leaf phylogeny**. There is a one-to-one correspondence between the phylogeny on the left and the graph on the right: the ordering of the leaves (red squares, peaks) and internal nodes (teal circles, valleys) is identical, and the vertical distance between the circles and the squares is maintained. The root of the tree is placed at 0 on the vertical axis, and it follows that the height coordinate gives the distance between each node and the root.

ancestor. We provide an example in Figure 9. There is a one-to-one correspondence between such a graph and the phylogeny. The proposal mechanism is in two steps. First, a permutation of the tip labels is obtained by swapping tips around each node (valley) based on a fair coin toss. Then, the algorithm perturbs the length of the lines in the graph by sampling updated values from a uniform distribution centred at the original length whose support is determined by a tuning parameter.

7.3. Bayesian tree-searching. We now focus on on Bayesian phylogenetic inference.

7.3.1. *The basic algorithm.* Algorithms most commonly used for Bayesian phylogenetic inference, such as those implemented in BEAST and MrBayes [33, 105], rely on MCMC, and differ mainly in their transition kernels, the user-defined function used to propose moves in the parameter space. We summarise the basic MCMC algorithm applied to phylogenetic inference in Algo. 3.

If the chain is irreducible and aperiodic, then it can be shown that the sample produced is a draw, albeit correlated, from the posterior of parameter vector $\boldsymbol{\theta}$ [122]. We can determine reasonable values for N, the total number of iterations, and B, the size of the burn-in, by

Algorithm 3: The basic MCMC algorithm.

Data: Starting values for the phylogenetic parameters $\boldsymbol{\theta}$

Compute likelihood $L(\boldsymbol{\theta})$ and prior probability $p(\boldsymbol{\theta})$;

for *iter* $= 1, \ldots, N$ **do**

  for $i = 1, \ldots, m$ **do**

    Use transition kernel $q(\theta_i, \theta')$ to propose an update to the parameter, i.e. a move from θ_i to θ' ;

    Compute the likelihood and prior at θ';

    Evaluate the Metropolis-Hastings acceptance ratio,

$$AR = \min\left[1, \frac{q(\theta', \theta_i)L(\boldsymbol{\theta'})p(\boldsymbol{\theta'})}{q(\theta_i, \theta')L(\boldsymbol{\theta})p(\boldsymbol{\theta})}\right] \tag{6}$$

    Draw a random number from $U(0, 1)$;

    if *the number falls under AR* **then**

      Accept the move and update the parameter value

    end

    else

      Reject the move and do nothing

    end

    Record the current parameter values.

  end

end

Discard the first B iterations of the chain, called the *burn-in*

looking at a posterior probability graph from the chain. If non-negligible autocorrelation is observed, larger values of N are recommended. In real phylogenetic analyses of large datasets, the total number of iterations is often larger than three million, and more than 25% of iterations are discarded as a burn-in. The posterior probability graph normally shows a steep increase in posterior probabilities at the start of the chain, followed by a stabilisation. B should be chosen as to keep only values sampled after stabilisation. Another heuristic diagnostic for the chain consists in graphing individual parameter values across iterations. The discreteness of the topology, and the inherent correlation in the limiting probability

parameters in the substitution rate matrix may limit their applicability though. [30] discuss other convergence diagnostics.

Specifying a transition kernel that can efficiently explore the parameter space, while simultaneously ensuring a reasonably high acceptance ratio, is key. NNI, for instance, tends to produce high acceptance ratios, but since the proposed moves are small, a larger number of iterations is required to properly map the posterior probability space. Also, if the posterior probability surface has several peaks and valleys, it may get trapped in a local posterior probability maximum region [131].

7.3.2. *The algorithm of Larget and Simon 1999.* The algorithm of [74] is in two steps, both of them based on Metropolis-Hastings (MH). First, it attempts a conventional update of all phylogenetic parameters conditional on the topology. Then, conditioning on other parameter values, it proposes a new topology by performing multiple successive MH scans starting from the current topology. That scheme leads to a proposal that belongs to a higher posterior probability region and consequently, increases the acceptance ratio. In each scan, a tree is proposed following the peak-and-valley mechanism described in section 7.2. In order to control acceptance ratios, tree proposals can be either global or local, that is, new trees may be obtained by permuting elements across the whole initial tree or only in the neighbourhood of a randomly-selected internal branch [131].

7.3.3. *Metropolis-coupled Markov Chain Monte Carlo ((MC)-cubed).* [49] proposed Metropolis-coupled MCMC, also called $(MC)^3$, which takes advantage of parallel computing to improve mixing and cross regions of low posterior probability. The main idea resembles that of simulated annealing. Multiple chains are run in parallel following the general MCMC scheme: the regular one is called the *cold chain*, while the others, meant to help improve mixing, are called *heated chains*. Chains are "heated" by applying exponents between 0 and 1 to the posterior probability expression, thus flattening it, which in turn makes moves across posterior probability troughs more likely. The algorithm attempts to swap the states of two randomly-selected chains periodically, with probability of acceptance corresponding to a Metropolis-Hastings ratio. In the end, inference is based on the cold chain only. [105] implemented this algorithm in a phylogenetic setting, and improvements were proposed by [45] and [7].

7.3.4. *Stochastic approximation Monte Carlo (SAMC).* Stochastic approximation Monte Carlo (SAMC) is also based on MH [9, 27, 79, 131]. It deals with ruggedness in the posterior probability space by dynamically modulating, thus biasing, the jump acceptance ratio in such a way that the chain is increasingly likely to leave regions where it spends much time. It is based on an arbitrary partition of the *energy* landscape, the energy function corresponding to the negative logarithm of the posterior probability density. The user specifies a vector of weights that determines the limiting proportion of time the chain spends in each subregion of the partition.

That scheme forces the chain into regions of lower posterior probability, which lets it reach other regions of high posterior probability. Once the sample from the posterior has been generated, a correction factor is included in the estimator of quantities of interest to account for the bias in sampling induced by the modulations in the jump ratio.

7.3.5. *Sequential Monte Carlo (SMC).* Sequential Monte Carlo (SMC) phylogenetic algorithms have also been devised, mainly to grapple with the heavy computational load of conventional MCMC methods applied to very large sequence datasets [16]. SMC methods are based on Sequential Importance Sampling (SIS), a reparameterisation of Importance Sampling (IS).

IS lets us infer quantities of interest, functions of the mean for example, defined with respect to a certain distribution by independently simulating values from a proposal distribution, which is convenient when simulating values from the original distribution is difficult. The bias resulting from the altered sampling scheme is subsequently fixed by weighting sampled values.

In SIS, we generate particles sequentially and update those weights along the way. Whereas in IS, we sample independent values from the proposal distribution, in SIS, we instead sample independent chains of values by drawing from conditional distributions. Each intermediate value is considered a particle and forms a *generation*. Particles found in a given generation across the different chains form a *particle population*.

One problem shared by IS and SIS is that of *particle degeneracy*, in which a few particles contribute overwhelmingly to expectation estimates by having weights much higher than

that of other particles in a population. A particle resampling step can be added to alleviate the issue, laying the foundation of SMC.

7.4. Priors in Bayesian phylogenetic inference. Attributing a uniform prior to the tree topology is common. Such a prior should reflect the assumption that all hierarchical partitionings of the sample are equally probable *a priori*. However, such a prior implies different prior probabilities for clades, depending on both their sizes and the total number of leaves in the tree [131, 89]. Clades containing few or a large number of sequences have higher prior probabilities than middle-size clades. It may not be possible to define a prior that would give equal probability to all clades, irrespective of size [131, 113]. However, if the data are informative, the likelihood should dominate the prior and mitigate the issue [131, 17].

Priors on branch lengths are known to be highly informative [131, 138]. For convenience, they are usually assumed to be independent and identically distributed exponential or uniform, which leads to unrealistically long trees in the posterior. Those priors may also create local peaks in the posterior and cause poor mixing in MCMC algorithms [131, 98, 139]. Assuming instead compound Dirichlet priors may help alleviate the problem [98], at the cost of a heavier computational load.

8. Evaluating confidence in inferred clades

Once an optimal tree is found, we assess confidence in its clades. In non-Bayesian phylogenetic inference, non-parametric bootstrapping is by far the most common approach [39]. In a nutshell, each bootstrap iteration involves sampling with replacement site indices, constructing a new alignment by stitching together configurations at the selected sites in the original alignment, and performing tree construction and tree searching for that new alignment. Each iteration yields a tree, whose clades are then listed. After performing a large number of bootstrap iterations, we merge all the lists and compute clade frequencies. We call these frequencies *bootstrap support for clades*.

Bayesian phylogenetic inference, on the other hand, merges tree searching and confidence assessment. Topologies sampled using the MCMC procedure let us obtain directly posterior probability support for the different clades.

9. Summarising a sample of trees

Once a sample of phylogenies has been obtained, with MCMC for example, it is customary to propose a "best" phylogenetic estimate. In this context, the Maximum Posterior probability (MAP) estimate is a natural choice [136]. We obtain a measure of posterior probability support for its clades by listing clades found in each of the sampled trees. The posterior probability support for any clade found in the MAP tree then corresponds to the proportion of times it is observed in the sample. The MAP estimate may not be ideal however, given the usual flatness of the phylogenetic posterior probability surface. In other words, many different phylogenies may have more or less equal posterior probability scores and so, the MAP tree may not be much preferable to other configurations.

Another solution consists in generating the so-called *majority-rule consensus tree* [74], a tree whose bifurcations only support clades found in at least half the sampled trees. It can be shown that such a tree always exists [136]. Note that although it is useful for summarising a large number of topologies, its lack of branch length measures may be problematic in certain analyses.

10. Cluster inference

Phylogenies provide estimated distances between any two sequences in a sample and so, they are ideal for the clustering of genotyping data. There are however several model-based and distance-based alternatives for clustering such data.

10.1. Model-based clustering.

10.1.1. *The core model.* Conventionally, in model-based clustering, data are assumed iid, with likelihood given by a finite mixture,

$$L(\boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{j=1}^{K} \pi_j g(y_i \mid \theta_j), \tag{7}$$

where π_j and θ_j, $j = 1, \ldots, K$, are, respectively, the mixture weight and cluster-specific parameter for cluster j.

10.1.2. *Maximum likelihood estimation.* We can use the Expectation-Maximization (EM) algorithm to obtain maximum likelihood estimates for cluster assignment probabilities [32].

We first expand the likelihood by adding the latent cluster assignment indices s_i, $i = 1, \ldots, n$. Moving to the log-scale, we get the so-called *complete-data log-likelihood*,

$$l(\boldsymbol{\theta}; \boldsymbol{s}) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{K} I(s_i = j) g(y_i \mid \theta_j) \right], \tag{8}$$

where $I(s_i = j)$ is an indicator function taking value 1 when the cluster assignment index for sequence i takes value j, and 0 otherwise. Note that the expectation of the complete-data likelihood with respect to $\boldsymbol{s}$ yields Eq. 7.

We iteratively refine the starting value for $\boldsymbol{\theta}$ with Algo. 4, yielding estimates $(\theta^{(1)}, \ldots, \theta^{N^{(EM)}})$. Once the algorithm has converged, we use Eq. 10 to compute cluster assignment probabilities for each data point. By assigning each observation to the cluster that maximises that probability, we obtain the required point estimate for $\boldsymbol{s}$.

The number of clusters, K, must be specified *a priori*. A common approach for choosing K consists in fitting mixture models with varying number of components and comparing them using the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC). However, the dependence between genetic sequences, resulting from shared ancestry, limits the applicability of mixture models for clustering sequence data.

10.1.3. *Bayesian inference of mixture parameters with a known number of components.* When K is known, the standard MCMC algorithm, described in section 7.3, can be used to infer posteriors for $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$. *Conjugacy* is achieved when the prior and posterior belong to the same parametric family, and can provide considerable computational benefits.

Algorithm 4: The Expectation-Maximisation (EM) algorithm

Data: Starting value for $\boldsymbol{\theta}$, denoted $\boldsymbol{\theta}^{(0)}$

for *iteration* $l = 1, \ldots, N^{(EM)}$ **do**

(Expectation step) Compute,

$$\begin{aligned} E_{\boldsymbol{s}|\boldsymbol{y},\boldsymbol{\theta}^{(l)}}[l(\boldsymbol{\theta};\boldsymbol{s})] &= \sum_{i=1}^{n} E_{\boldsymbol{s}|\boldsymbol{y},\boldsymbol{\theta}^{(l)}} \left\{ \log \left[\sum_{j=1}^{K} I(s_i = j) g(y_i \mid \theta_j) \right] \right\} \\ &= \sum_{i=1}^{n} E_{\boldsymbol{s}|\boldsymbol{y},\boldsymbol{\theta}^{(l)}} \{ \log[g(y_i \mid \theta_{s_i})] \} \\ &= \sum_{i=1}^{n} \sum_{s_i=1}^{K} \log[g(y_i \mid \theta_{s_i})] P(s_i \mid y_i, \boldsymbol{\theta}^{(l)}), \end{aligned} \tag{9}$$

with,

$$\begin{aligned} P(s_i \mid y_i, \boldsymbol{\theta}^{(l)}) &= \frac{g(y_i \mid \theta_{s_i}^{(l)}) p(\theta_{s_i}^{(l)})}{P(y_i \mid \boldsymbol{\theta}^{(l)})} \\ &= \frac{\pi_{s_i} g(y_i \mid \theta_{s_i}^{(l)})}{\sum_{j=1}^{K} \pi_j g(y_i \mid \theta_j^{(l)})}. \end{aligned} \tag{10}$$

(Maximisation step) Find $\boldsymbol{\theta}$ that maximises Eq. 9 and set $\boldsymbol{\theta}^{(l+1)}$ equal to this value.

end

The Dirichlet distribution, a multivariate expansion of the beta distribution, is a a popular prior for mixture weights. Its density is given by,

$$f(\pi_1, \ldots, \pi_K; \boldsymbol{\alpha}, K) = \frac{1}{\boldsymbol{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} \pi_i^{\alpha_i - 1}, \quad 0 \leq \pi_i \leq 1, \sum_{i=1}^{K} \pi_i = 1,$$

with,

$$\boldsymbol{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)},$$

and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, $\alpha_i > 0$.

The mixture distribution given by Eq. 7 can also spring from the assumption of a multinomial distribution for the cluster assignment indices. To see that, we generate each observation y_i in two steps. First, a cluster assignment index j is sampled from a multinomial distribution with parameters $\pi_1, \ldots, \pi_K$ and then, a value is sampled from distribution $g(y_i \mid$

θ_j), θ_j being the centroid of cluster j. It can then be shown that the density of y_i is indeed given by Eq. 7.

We can exploit the conjugacy between the multinomial and Dirichlet distributions to simplify computations. Indeed, by giving the multinomial sampling probabilities $\boldsymbol{\pi}$ a Dirichlet prior, we find that the posterior for the mixture weights follows a Dirichlet distribution with parameters $(\boldsymbol{n} + \boldsymbol{\alpha})$, where $\boldsymbol{n}$ is the number of occurrences of each cluster in the data, that is, $n_i = \sum_{j=1}^{n} I(s_j = i)$. If we select the posterior as a transition kernel for the MCMC algorithm, the jump acceptance probability reduces to 1, and we are left with *Gibbs sampling*. In the case where Eq. 7 corresponds to a mixture of normal distributions, priors for $\boldsymbol{\theta}$ can also be selected to ensure conjugacy [51].

To obtain cluster assignment probabilities, we note that,

$$P(s_i \mid \theta_{s_i}) \propto g(y_i \mid \theta_{s_i}) p(s_i), \quad i = 1, \ldots, n,$$

where $P(.)$ and $p(.)$ are the posterior and prior distributions, respectively. It follows that the posterior of cluster assignment index s_i is multinomial with sampling probabilities proportional to $\pi_{s_i} g(y_i \mid \theta_{s_i})$.

10.1.4. *Bayesian inference of mixture parameters with an unknown number of components.* If K is unknown, we can make it a parameter and give it, for example, a Poisson or discrete uniform prior. An increase in K implies however an increase in the dimensionality of the parameter space, precluding use of the standard MCMC procedure we described in section 7.3. The 'reversible jump' sampler suggested in [100, 52] lets us overcome that issue.

The approach involves a conventional update mechanism for parameters that do not affect the dimension of the parameter space - the cluster centroids, for example. The algorithm moves in the K-space by first, either splitting an existing cluster or merging two existing clusters, and then, by spawning a new empty cluster or deleting an existing empty cluster. Each split or spawn move increases the dimension of the parameter space by an amount equal to the number of cluster-specific parameters. Equivalently, merge or deletion moves reduce the dimension of the parameter space by an equal amount.

When a split move is proposed, we obtain the cluster-specific parameters for the resulting two clusters by multiplying the old cluster-specific parameters by independent random numbers, denoted $\boldsymbol{u}$. Computing the value of the transition kernel for the move, $q((\boldsymbol{s}_t, \boldsymbol{u}, K, \boldsymbol{\theta}_t), (\boldsymbol{s}_{t+1}, K+1, \boldsymbol{\theta}_{t+1}))$, is straightforward. Computing that value for the reverse move, $q((\boldsymbol{s}_{t+1}, K+1, \boldsymbol{\theta}_{t+1}), (\boldsymbol{s}_t, \boldsymbol{u}, K, \boldsymbol{\theta}_t))$ is not trivial.

Let us denote by f the function mapping $(\boldsymbol{\theta}_t, K, \boldsymbol{u})$ to $(\boldsymbol{\theta}_{t+1}, K)$, that is, $f : (\boldsymbol{\theta}_t, K, \boldsymbol{u}) \rightarrow (\boldsymbol{\theta}_{t+1}, K+1)$. If f is 1-to-1, then the inverse function f^{-1} exists, and we can obtain the probability density for f^{-1} by applying standard results pertaining to the transformation of random variables. Merge and split moves are reciprocal and so, we handle cluster mergers the same way as splits. In other words, we compute the transition kernel ratio for the reciprocal split move like before and simply swap the numerator and denominator.

A cluster, albeit non-empty, is born when a split move occurs, and it follows that we can obtain the value of the transition kernel ratio for an empty cluster birth move by adapting the technique used to handle splits. Since cluster death is reciprocal to cluster birth, we can use the strategy outlined before to process such moves.

10.1.5. *Dirichlet process clustering.* To avoid a priori specification of the number of clusters, we may assume that the data were sampled from a mixture distribution with an infinite number of components. The Dirichlet Process (DP) defines a *random distribution*, whose support is made up of infinite mixture distributions [14, 10]. The DP can serve as a prior for cluster assignment indices when the number of clusters is unknown. Thanks to the Polya urn or stick-breaking algorithms, sampling from the DP is straightforward [66]. Such a prior is exchangeable and has one tuning parameter, the *concentration parameter*, that controls the expected number of clusters [59]. A higher concentration parameter value increases the expected number of clusters, which also grows logarithmically with sample size. The cluster size distribution is characterised by the 'rich-get-richer' property: when sequentially simulating from the DP, large clusters tend to attract new sequences, which translates as exponential decay in the tail of the distribution.

The Pitman-Yor process (PYP) generalises the DP by adding a so-called *discount parameter*, that improves flexibility in terms of the cluster size distribution [90]. The PYP reduces

to the DP when the discount parameter is brought to 0, and its tail is characterised by a power-law decay. Sampling from the PYP is still straightforward, as it involves an extension of the stick-breaking algorithm.

10.2. Distance-based clustering. Distance-based clustering algorithms take as input a matrix of pairwise distances and, in many cases, a number of tuning parameters. Defining an appropriate distance measure is key in ensuring the quality of the final partition.

Conventional distance-based approaches include simple hierarchical clustering, K-means, and K-medoid. We briefly described hierarchical clustering in section 7.1. Clusters returned by hierarchical clustering approaches are a partition of the sample into non-overlapping clades. We can obtain such a partition by snipping the dendrogram at a fixed height from the bottom, or by selecting multiple cutpoints in order to optimise a cluster quality measure such as, for instance, the Dunn index [35] or the Calinski-Harabasz index [23].

In K-means, we select a number of clusters *a priori* and try to find,

$$\arg\min_{\boldsymbol{s}} \sum_{i=1}^{n} E(y_i, \mu_{s_i})^2,$$

with μ_{s_i} being the centroid of cluster s_i, and $E(x, y)$ denoting the Euclidean distance between points x and y.

Optimisation is performed iteratively, by sequential updates of s_i and μ_{s_i}. Convergence of the method is assured, but stabilisation may occur around a local minimum.

The K-medoid method extends K-means by allowing an arbitrary distance measure and categorical data. It also consists of two steps. First, for each cluster, it selects as centroid the element that minimises the sum of distances to co-clustering elements. Then, it iterates across all elements, potentially assigning each of them to a new cluster, in order to minimise the sum of all distances between elements and centroids. The two steps are repeated until stabilisation.

However, both K-means and K-medoid may be ill-suited for clustering epidemic data. The Euclidean distance is not a natural measure of distance between genetic sequences, thus

disqualifying K-means. Also, just like for the mixture approach described in section 10.1, picking the number of clusters a priori might be difficult.

More recently, [130] proposed a distance-based clustering approach focusing on sequence data, called the *Gap Procedure*. It does not require any user-specified threshold and is very fast, since it avoids phylogenetic estimation entirely.

10.3. Phylogenetic clustering. Tree-searching algorithms, like those described in section 7.2, can produce an optimal phylogeny, or set of phylogenies, that serves as the base for cluster inference. Just like in simple hierarchical clustering, *phylogenetic clusters* are conventionally disjoint clades. They are typically found from an estimated phylogeny by applying an *ad hoc* rule, usually involving a maximum within-cluster distance requirement and a confidence requirement. A *depth-first* tree-traversal algorithm can be used for that purpose [36], cf. Algo 5.

Algorithm 5: A depth-first tree-traversal algorithm for clustering sequence data

Data: A phylogeny

Create an empty list of nodes to inspect;

Add to the list the root node;

repeat

  if *the first node in the list is a tip* **then**

    Skip the next check

  end

  if *the clade formed by tips descended from the first node meets the clustering requirements* **then**

    Record the clade (it is a cluster)

  end

  else

    Append the children nodes to the end of the list

  end

  Remove the first node from the list of nodes to inspect

until *the list of nodes to inspect is empty*;

10.3.1. *Confidence requirements in non-Bayesian clustering.* A large confidence estimate for a clade is usually required to conclude in clustering. For instance, under strict evolutionary assumptions, a 70% bootstrap support corresponds roughly to a 95% probability that the inferred clade be found in the true phylogeny [60, 58].

Adopting an exceedingly high bootstrap frequency requirement results in the exclusion of many valid clusters, while increasing marginally the probability that the selected clusters are found in the true phylogeny. Simulations shown in [41] reveal a modest gain ($< 0.02\%$) in the probability of an inferred cluster appearing in the true phylogeny when the bootstrap requirement is increased from 0.9 to 0.99. Caution must be exercised before drawing conclusions from bootstrap confidence levels: the sole consideration of a fixed bootstrap threshold, irrespective of the genetic distance model, is insufficient to draw reliable inferences of co-clustering [63]. Also, several papers propose different interpretations of bootstrap support for clades [43, 58, 44, 39, 110, 116, 82]. Bootstrap estimation of confidence in groupings is further known to be affected by the number of taxonomic units considered [62].

10.3.2. *Confidence requirements in Bayesian clustering.* Bayesian and non-Bayesian clustering can usually be tuned to uncover largely-overlapping sets of clusters, but the approaches do not consistently agree. Bayesian support for inferred clades is systematically higher than that derived from non-parametric bootstrapping [41]. For higher support values, the increase in the probability that a cluster be correctly inferred resulting from a fixed-size increase in posterior probability is greater than that resulting from a same-size increase in bootstrap frequency. Adopting a higher posterior probability requirement is therefore justified.

Some limitations of bootstrap-based cluster inference extend to Bayesian inference. In both cases, misspecification of the genetic distance model results in an increase in the number of misidentified clades [41]. Even under model misspecification, a fixed posterior probability requirement produces a considerably larger number of clusters than an equal non-parametric bootstrap frequency requirement. The proportion of misidentified clades will be much lower with the latter requirement, at the cost of missing a considerable number of clusters. Unlike bootstrap frequency, posterior probability support for a clade has the advantage of having a straightforward interpretation. It may however produce conflicting sets of clusters when

different sets of loci are used, all of them supported by an empirical posterior probability estimate of 1.0 [103].

Finally, as emphasised in section 7.4, unrealistic priors may have a detrimental effect on cluster recovery. Indeed, if total tree length is overestimated, a fixed distance requirement will result in spuriously small clusters. Since a uniform prior on tree topologies involves different prior probabilities for clades of different sizes, the support for any two inferred clades may not be readily comparable, which makes choosing a uniform minimum posterior probability requirement difficult.

CHAPTER 3

Manuscript 1: "Assessment of Overlap of Phylogenetic Transmission Clusters and Communities in Simple Sexual Contact Networks: Applications to HIV-1"

1. Preamble

The association between phylogenies for HIV-1 and sexual contact networks is of interest to public health specialists and accordingly, several studies have formally tackled the subject [133, 77, 101, 85, 31], leading to conflicting results. The research presented in manuscript 1 aims first and foremost to improve substantive understanding of phylogenetic clusters, by highlighting and quantifying their potential association with communities, community structure being a defining characteristic of sexual contact networks [123].

Experts have observed episodic spikes in HIV-1 incidence, that lead to clustering in phylogenetic samples [19], but the reason behind their occurrence is unclear. We hypothesise they might be caused by the pathogen entering non-infected communities, suddenly boosting the number of subjects exposed to the virus, and thus, contributing to an increased incidence. That conjecture motivated the study.

Because of the error associated with phylogenetic inference and the lack of relevant and reliable estimates of sexual contact networks, we investigate overlap between phylogenetic clusters and network communities in a simulation setting, that lets us track the epidemic. The study reveals that this correspondence is often limited, mainly due to the requirement that clusters should correspond to clades in the phylogeny, thus casting doubt over the intuitive interpretation proposed before. Nevertheless, the study emphasises features of a

phylogeny that may be used to recover communities, stressing once more that the clade assumption is fundamentally flawed for that purpose.

Assessment of Overlap of Phylogenetic Transmission Clusters and Communities in Simple Sexual Contact Networks: Applications to HIV-1

Luc Villandre[1] David A. Stephens[2¶] Aurelie Labbe[1,3¶] Huldrych F. Günthard[4,5&] Roger Kouyos[4,5&] Tanja Stadler[6,7*] The Swiss HIV Cohort Study^

1 Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montréal, Québec, Canada
2 Department of Mathematics and Statistics, McGill University, Montréal, Québec, Canada
3 Department of Psychiatry, Douglas Mental Health University Institute, Montréal, Québec, Canada
4 Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, Zurich, Kanton Zurich, Switzerland
5 Institute of Medical Virology, University of Zurich, Zurich, Switzerland
6 Department of Biosystems Science and Engineering, ETH Zürich, Basel, Basel-Landschaft, Switzerland
7 Swiss Institute of Bioinformatics, Lausanne, Switzerland

¶These authors contributed equally to this work.

& These authors also contributed equally to this work.

*** E-mail: tanja.stadler@bsse.ethz.ch (TS)**

^Membership of the Swiss HIV Cohort Study is provided in the Acknowledgements.

Abstract

Background. Transmission patterns of Sexually-Transmitted Infections (STIs) could relate to the structure of the underlying sexual contact network, whose features are therefore of interest to clinicians. Conventionally, we represent sexual contacts in a population with a graph, that can reveal the existence of communities. Phylogenetic methods help infer the history of an epidemic and incidentally, may help detecting communities. In particular, phylogenetic analyses of Human Immunodeficiency Virus (HIV)-1 epidemics among Men who have Sex with Men (MSM) have revealed the existence of large transmission clusters, possibly resulting from within-community transmissions. Past studies have explored the association between contact networks and phylogenies, including transmission clusters, producing conflicting conclusions about whether network features significantly affect observed transmission history. As far as we know however, none of them thoroughly investigated the role of communities, defined with respect to the network graph, in the observation of clusters. **Methods.** The present study investigates, through simulations, community detection from phylogenies. We simulate a large number of epidemics over both unweighted and weighted, undirected random interconnected-islands networks, with islands corresponding to communities. We use weighting to modulate distance between islands. We translate each epidemic into a phylogeny, that lets us partition our samples of infected subjects into transmission clusters, based on several common definitions from the literature. We measure similarity between subjects' island membership indices and transmission cluster membership indices with the adjusted Rand index. **Results and conclusion.** Analyses reveal modest mean correspondence between communities in graphs and phylogenetic transmission clusters. We conclude that common methods often have limited success in detecting contact network communities from phylogenies. The rarely-fulfilled requirement that network communities correspond to clades in the phylogeny is their main drawback. Understanding the link between transmission clusters and communities in sexual contact networks could help inform policymaking to curb HIV incidence in MSMs.

2. Introduction

2.1. Background and objectives. Basic epidemiologic models rest on the *random mixing* assumption [6, 77]. In the presence of random mixing, each individual in a population has a small and equal probability of coming into contact with any other individual, which can lead to very quick epidemic spread. For Sexually-Transmitted Infections (STIs) however, the random mixing hypothesis fails to hold: STIs spread within sexual contact networks, that limit their propagation. In particular, the random mixing assumption seems ill-suited for modelling HIV-1 epidemics [77].

We conventionally represent a sexual contact network with a graph whose *nodes*, also called *vertices*, correspond to individuals, and edges, to a sexual association between them. Connected nodes are called *neighbours*. The number of neighbours for a given node is called its *degree*, and the *degree distribution* is one of the defining features of contact network graphs. The sexual contact network therefore maps potential transmission paths for sexually-transmitted pathogens. A graph may also be characterised by *community structure*, that is, it may contain distinctive, non-overlapping sets of nodes within which we observe a high connection density [87]. We call these sets *communities*.

Communities in sexual contact networks could very well leave a footprint in the observed epidemiologic history. On one hand, we expect quick (early) transmission within a recently-infected community. Indeed, right after a virus infects a first node in a community, the number of edges leading to uninfected nodes is high, which decreases the mean time until the next infection event. In other words, when the virus enters a new community, the number of *exposed* subjects, that is, uninfected neighbours of infected subjects, tends to rise quickly, and the incidence is expected to spike as a result. On the other hand, we expect slow (late) transmission between different communities [76]. After all, in a contact network sense, communities have to be distinctive, which implies that their member nodes tend to be considerably more frequently connected with one another than with non-member nodes.

Our study explores the association between communities and epidemic spread, the transmission history being represented with a bifurcating tree known as a *phylogeny*. *Phylogenetic*

methods make use of sequence data to reconstruct the ancestral history of a set of organisms, such as HIV-1, and it is still unclear to what extent community structure can be recovered from the phylogeny.

2.2. A formal definition of community structure. There are two requirements for a set of nodes to be called a community: a large enough proportion of nodes in the set must be connected to one another, and the set must be sparsely connected to external nodes [50]. Formally, a proposed split of nodes into communities is evaluated using a *modularity* score [87]. Different node partitions are proposed and scored, the objective being to identify the partition that maximises this score: a higher maximum modularity implies a more apparent community structure.

Computing the modularity of all possible node partitions quickly becomes intractable as network size increases and so, communities in graphs are usually found heuristically. For example, the *walktrap* and *label propagation* algorithms are two basic community-detection algorithms [91, 96]. Several more sophisticated approaches can also be used [50]. In large, complicated graphs, different algorithms may yield vastly different results, whose validity can be assessed based on substantive criteria such as a priori knowledge of the relatedness of subjects in the network, if this information is available, or through the modularity score.

In *undirected interconnected-islands models* however, characterised by disjoint subnetworks connected by bridges (see Figure 10), it is straightforward to make all algorithms return the optimal partition: in these models, each island corresponds to a community. Although simplistic, such networks are not entirely unrealistic. Islands could be thought of as subnetworks in separate countries, cities, or neighbourhoods. They can also result from a high degree of *assortative mixing*. Assortative mixing refers to the tendency of subjects to form connections with individuals sharing similar characteristics, for example, age, socio-economic status, or profession. A high degree of assortative mixing tends to produce easily-distinguishable regions in the contact network graph, made up of nodes representing similar subjects. More specifically, sociodemographic characteristics such as age could contribute to producing islands within a localised sexual contact network.

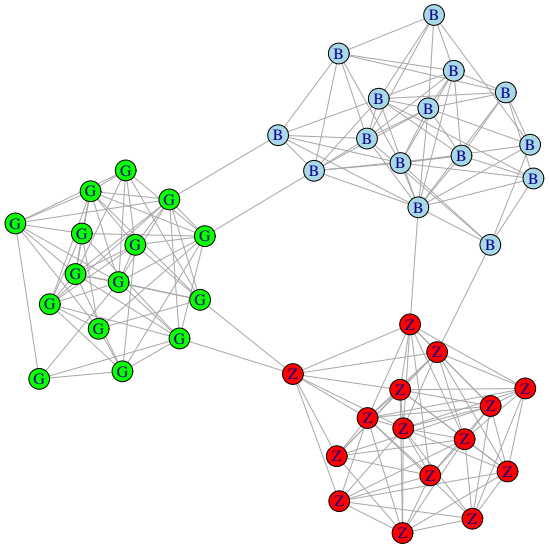

FIGURE 10. **A simple undirected interconnected-islands network, representing subjects living in three different islands, corresponding to cities**. The graph was randomly-generated. It has three islands of of size 15, and the connection probability of any two vertices within an island is 0.6. There are two edges, called "bridges", between any two islands. The label within each vertex indicates in which city the subject lives.

2.3. The link between contact networks and phylogenies. Several studies have looked at the link between contact networks and phylogenies, producing seemingly mixed conclusions. [133] investigated how *clustering* in a contact network graph, defined as the propensity of pairs of connected nodes to share a common neighbour, affects epidemics, in terms of their phylogenies. They found that changes in network clustering result in very little variation in trees. [77], on the other hand, found a strong association between the shape of phylogenies resulting from epidemics on four types of static contact networks. [101] extended the investigation of [77] by, among other things, looking at a family of dynamic networks. Unlike the latter, they found only a modest direct effect of network configuration on phylogenies. Another study presents an attempt to directly reconstruct a sexual contact network,

underlying HIV transmission, based on epidemiological and genetic information [140]. The approach is based on a filtering scheme: starting with a fully-connected graph, the method removes edges by comparing individuals based on socio-demographic and medical covariates, and then, adds directions to the remaining edges by using estimates of seroconversion dates.

2.4. The relevance of sexual contact networks: HIV-1 epidemics among men who have sex with men. HIV-1 remains at a high prevalence among Men who have Sex with Men (MSM) in developed countries. In the United States, in 2010, MSMs represented 63% of the estimated total number of new HIV infections, and the prevalence of HIV in this subpopulation was approximately 18% [24]. Despite an increase in the proportion of individuals treated with *Highly Active Antiretroviral Therapy (HAART)*, the incidence rate of HIV in MSMs in the United States has increased by an estimated 12% between 2008 and 2010.

2.4.1. *Quick transmission chains.* Studies have revealed the existence of *quick transmission chains* in HIV epidemics among MSMs [19, 20], that is, distinct groups of infected subjects with genetically-similar viruses. Such groups can only be formed through series of infection events close in time. Indeed, the fast evolution of HIV-1 ensures that correspondence in the genetic makeup of viruses in different subjects indicates not only epidemiologic relatedness, but also a recent viral common ancestor. In other words, sets of genetically similar infections must result from chains of quickly-occurring transmissions. We stress that a quick transmission chain does not necessarily involve a first individual transmitting the virus to a second individual, who then infects a third one, and so on. We can also have a single infected subject transmitting the virus to an arbitrary number of individuals in a short amount of time. The "chain" is defined with respect to transmission events, irrespective of the transmission path.

Understanding the reasons for the existence of quick transmission chains is crucial, as it could inform public health interventions in MSMs. For instance, the WHO now advocates Treatment as Prevention (TasP) [135]. However, the potential of TasP to curb the HIV-1 epidemic in this subpopulation depends in great part on the timing of onward transmission. If it tends to occur mostly within the first month after infection, corresponding to the *acute*

infection stage [2], TasP is unlikely to succeed in significantly reducing incidence rates, since HIV-1 is rarely diagnosed so early [72]. Indeed, the detection of numerous quick transmission chains may suggest that early transmissions are mainly responsible for the rising HIV-1 incidence rates in MSMs [20]. In this context, TasP could still potentially prevent a number of infections, but would be unsuccessful in containing the epidemic.

2.4.2. *Transmission clusters.* Quick transmission chains for HIV correspond to *transmission clusters*, loosely defined as sets of HIV-positive individuals whose viruses share a "close" common genetic ancestor [21]. How close this common ancestor ought to be is typically decided in an ad hoc way. Sets of co-clustered subjects have viruses that are not only close in genetic terms, but also distinct from any viruses from non-co-clustered subjects.

Numerous studies have used phylogenetics for transmission cluster inference, and have reported the existence of large transmission clusters in HIV epidemics among MSMs [20, 75, 78, 73]. In MSMs, transmission clusters may explain as much as 75% of incidence, with one infection leading up to an estimated 10 to 13 onward transmissions [21, 76].

2.4.3. *Sexual contact networks in MSMs.* The existence of quick transmission chains should inform public health approaches and so, gaining insight into the factors contributing to time between transmissions is important. One such factor could be the structure of the sexual contact network: it might play a major role in the production of quick transmission chains.

Sexual contact networks in MSMs are characterised by high clustering, assortative mixing, and the presence of several nodes with a distinctively larger degree [123]. Clustering, in a contact network graph, refers to the frequency at which any pair of nodes share a common neighbour. The existence of several distinctive high-degree nodes may result from *preferential attachment*, which implies that each subject, when forming a new connection, tends to prefer individuals with an already large number of connections. Preferential attachment leads to graphs with several "hubs", nodes with an exceptionally high number of neighbours. The resulting degree distribution, called a *power-law distribution*, therefore has a heavy right tail. In other words, we have

$$P(D = d) \propto d^{-\psi},$$

where d is node degree and ψ is a positive constant. Communities may result from a combination of assortative mixing and preferential attachment.

2.4.4. *Understanding communities for prevention of HIV-1 in MSMs.* The existence of numerous transmission clusters, because of their correspondence with quick transmission chains, may explain the difficulties in containing HIV-1 incidence in MSMs [21], but its link with community structure in MSM sexual contact networks is poorly understood. Assessing the contribution of community structure to HIV-1 incidence in MSMs may be helpful to design more effective intervention strategies. With this goal in mind, our study looks specifically at how well transmission clusters map onto communities in a graph. In other words, findings in the present study will help understand the extent to which HIV-positive MSMs in the same transmission cluster tend to belong to the same community, defined with respect to a graph, within a known network structure with easily-distinguishable communities.

3. Methods

Since we do not know of any extensively mapped sexual contact network for MSMs and infection tracing is difficult in this population [15], we have no choice but to rely on simulations.

3.1. Simulating the sexual contact networks. In each simulation, we simulate a sexual contact network, and then, an epidemic spreading onto it. We simulate epidemics on three classes of randomly-generated networks, all within the framework of undirected interconnected-islands models: two deliberately simplistic, called "type" A and "type B", and the other, called "type C", tailored in such a way that it displays prominent features of real sexual contact networks. Weights in our networks modulate distance between any two connected subjects. In unweighted networks, any two connected subjects are at arbitrary distance 1. In weighted networks, we attribute weights below 1 to edges serving as bridges. Distance between any two subjects is equal to one over the sum of weight values for edges on the shortest path between them. It follows that the two subjects delimiting a bridge are more distant when we attribute it a lower weight. In each weighted network, we give all bridges the same weight.

3.1.1. *Network consisting of many islands of equal size (Type A).* In the first network structure, we find 13 islands of 20 subjects each, with each island being a *fully-connected graph*, and one bridge linking any two islands. In a fully-connected graph, all vertices are connected to each other. We let bridge weights take values 0.25, 0.5, 0.75, or 1. In the latter case, the network is unweighted.

3.1.2. *Network consisting of one large island connected to many small islands (Type B).* The second network structure consists of one central island of size 60, representing a foreign sexual contact network, connected by single bridges to 25 islands of size 20. Each small island is a fully-connected graph. We implement weighting exactly as before. The set of small islands represents disjoint sexual contact subnetworks in a population of interest.

3.1.3. *The more realistic network (Type C).* The more realistic networks are made of 100 islands each. To ensure that all islands are accessible, we first link islands in a chain. We then create additional bridges by connecting any two vertices belonging to different islands with probability 0.00075. As in networks of types A and B, we consider networks in which bridge weights take values 0.25, 0.5, 0.75, or 1.

We introduce preferential attachment by making each island a *Barabasi-Albert* graph [12]. In our simulations, each subnetwork is generated by first creating three interconnected subjects. New subjects are then added one by one. When introduced into the network, a subject randomly forms three connections with existing subjects, who become first-degree neighbours. The probability that a subject is selected increases linearly with his degree. We call this *linear preferential attachment*, and this process gives subjects the expected power-law degree distribution.

Further, islands in these networks are of variable size. We sample island sizes independently from an empirical distribution obtained by inspection of the maximum likelihood phylogeny for HIV-1 subtype B sequences used in [11]. All viral sequences come from participants in the Swiss HIV Cohort Study (SHCS). We use this phylogeny to partition the 5395 sampled HIV-positive subjects into transmission clusters. We consider co-clustered sets of subjects whose viral sequences are separated by a tree distance of at most 0.1 nt/bp, forming a *clade* of size five or more, with this clade having bootstrap support greater than 70%. A

set of tips forms a clade if and only if there is a node in the tree with its descending tips corresponding exactly to this set of tips. Based on these criteria, we obtain a list of clusters, for which we derive a cluster size distribution. This distribution may have gaps and so, we apply a loess smoother to it [29].

Because [11] are working with empirical HIV sequence data, many HIV-positive subjects who would co-cluster with individuals in the sample may be missing. This incomplete sampling may result from undiagnosed subjects or diagnosed subjects that could or would not participate in the SHCS [73]. In order to account for this, we assume that each known infection is connected to either 0, 1, or 2 unknown infections, with an average of 1 unknown infection.

3.2. The link between islands and communities. In our analyses, we assume that each island is a community. We ensure that two conventional community-detection algorithms validate this assumption. That is why we measure, with the Adjusted Rand Index (ARI) [129], correspondence between subjects' island and community membership indices, with the latter indices produced by the walktrap and label-propagation community-detection algorithms [91, 96]. The ARI is equal to the ratio of correctly co-clustered and separated elements to the total number of pairs of elements, with a correction term for chance. It is bounded above by 1, which in our case indicates perfect overlap between islands and communities. It gives insight similar to the adjusted mutual information score.

3.3. Simulating the epidemics. Subjects can be in one of three states: "susceptible", "infected", or "removed". All subjects start in the "susceptible" state. After introduction of the virus in the network, susceptible subjects may contract the virus from an infected neighbour, at which point they enter the "infected" state. Infections are eventually diagnosed, and subjects move to the "removed" state. In the context of HIV, removal corresponds to diagnosis and uptake of HAART, which prevents transmission by drastically reducing the viral load. Two parameters are involved: an infection rate that scales linearly with the number of susceptible neighbours of an infected individual, and a removal rate, that determines when infected individuals become diagnosed.

We stop simulating new transmissions once a predetermined number of subjects are in the "removed" state or, trivially, once the epidemic becomes *extinct*, that is, all infected subjects have moved on to the "removed" state.

We represent each simulated epidemic with a transmission tree connecting the infected individuals. We terminate a lineage, that is, we obtain a tip, upon an individual leaving the "infected" state and entering the "removed" state. In the analyses, we consider only the subtree connecting all tips corresponding to removed individuals. This phylogeny lets us partition those subjects into transmission clusters. This scheme reflects real data collection, in which an infection can only be recorded after diagnosis.

In type A sexual contact networks, epidemics result from one random introduction of the virus. In type B networks, we randomly introduce the virus once in the large island, but we assume that only subjects in the small islands can be observed. It follows that removed subjects in the large island are not counted for the purpose of determining when to stop the simulations, as they belong to a foreign epidemic. Thus, the tips in the phylogenies representing the simulated epidemics only correspond to removed individuals in the small islands. Given the large size of the type C networks, we increase the number of introductions to two. This increase also aims to produce a more plausible epidemic, as multiple introductions are common in practice [73].

Transmission time along any edge follows an exponential distribution with rate directly proportional to the associated weight, or equivalently, inversely proportional to the distance between the subjects delimiting the edge. In type A and type B networks, we stop simulating new infections once 60 subjects have entered the "removed" state. In type C sexual contact networks, because of their large size, we raise this requirement to 200 subjects. Those numbers have been chosen to represent epidemics with moderate prevalence. If an epidemic fails to reach this threshold, we discard it along with the sexual contact network on which it was simulated. We then generate a new network and a new epidemic. We simulate epidemics until 300 manage to reach the threshold.

The minimal size requirement ensures that all sampled epidemics are of the same size and span more than one community, which makes comparisons across and within scenarios more

straightforward. Indeed, since an increase in the distance between islands reduces mean time to extinction, without the size requirement, scenarios with more distant islands would involve systematically smaller epidemics, with more of them becoming extinct before a single bridge is crossed. In that extreme case, the ARI for community recovery would be trivially 0, which could be misleading. Indeed, the ARI is meaningful only when the data are partitioned in a non-trivial way. Further, averaging ARIs obtained on epidemics of different sizes would be questionable, since epidemic size affects the difficulty of the clustering problem.

In order to reduce the probability of epidemics going extinct too early, we use a shifted exponential distribution to generate time until removal. In other words, this time corresponds to the sum of a fixed time and a value simulated from an exponential distribution. It follows that the distribution of time until removal is bounded below by a value greater than zero. The shift gives all newly-infected subjects a minimum amount of time to potentially transmit the virus to their susceptible neighbours. A subject enters the "removed" state at the moment of diagnosis and so, this delay is understood as reflecting the time it takes for an infection to become diagnosable. We select the shift parameter by generating epidemics across network types and weighting scenarios while ensuring that at most 10% of them go extinct before the epidemic size requirement is reached.

Time in the simulation of epidemics corresponds to genetic distance. In other words, it is understood as the expected number of mutations divided by the number of loci in the DNA alignment. Under a strict molecular clock assumption, genetic distance translates directly to calendar time. We represent each simulated epidemic by a rooted phylogeny, with which, if needed, we can obtain simulated samples of DNA sequences. To get such a sample, first, we simulate a *root sequence*, that is, the genetic makeup of the virus that first entered the sexual contact network, assuming equal frequency for all four bases. Evolution along the phylogeny follows a continuous time Markov process, whose rate matrix we inferred from a subsample of HIV-1 subtype B sequences collected in Québec, Canada. By letting the root sequence evolve along the phylogeny, we obtain the required sample.

3.4. Finding the clusters and assessing correspondence to islands. There is no widely-accepted method for partitioning samples of sequencing data into transmission clusters, and different methods tend to yield contradicting results. To ensure that our analyses are not overly affected by the arbitrary selection of one method, we obtain transmission clusters in four different ways, all commonly employed in HIV transmission cluster inference. We define a transmission cluster as,

(1) A clade whose elements are separated by a fixed tree distance of at most x, where tree distance is the sum of branch lengths between any two tips [19],
(2) A clade whose elements are separated by a fixed distance of at most x, where distance is the standardised number of different nucleotides between any two sequences [97],

(3) A clade whose elements are separated by a median pairwise tree distance below x, an arbitrary percentile of the tree's between-tip tree distance distribution [94], where tree distance once again corresponds to the sum of branch lengths,
(4) A clade in a dendrogram, that is, an ultrametric tree obtained from the matrix of between-tip tree distances, cut at height x, where the dendrogram is obtained using one of three methods: average linkage, complete linkage, or Weighted Pair-Group Method of Analysis (WPGMA),

where we let x vary in order to inspect its effect on community recovery.

Methods in definition 4 are standards in agglomerative hierarchical clustering, which involves recursively combining the two clusters that are closest one to another [55]. The three selected methods differ in how they define inter-cluster distance. In the average linkage method, distance between two clusters is an average of genetic distance within all non-co-clustered pairs of elements from these clusters. In the complete linkage method, this distance is instead the maximum distance within those pairs. In WPGMA, distance between two clusters is the distance between two elements, one in each cluster, each deemed the most representative of its cluster.

Under definitions 1, 2, and 3, the *clade-based methods*, we explore the tree from root to tips to verify if nested clades meet the clustering requirements. More specifically, we employ a *depth-first* algorithm,

(1) Start exploration at the root node,
(2) Check if the clade whose Most Recent Common Ancestor (MRCA) is the current node meets the cluster definition. If so, stop. If not, move down to the two immediate children nodes and for each of them, repeat the current step,
(3) Stop exploration along any given path once a tip is reached.

In real studies, transmission cluster inference additionally relies on a confidence requirement for clades in the phylogeny. In our analyses however, we use the true phylogeny and so, all clades are known with confidence 1. We assume that each island is a community and we measure overlap between transmission clusters and islands with the Adjusted Rand Index (ARI) [129].

3.5. Ethics. Our study is a pure simulation study and so, did not require approval by an ethics committee. We used trees inferred in a study by Avila et al. [11] to obtain a component of our simulation algorithm. The latter study was conducted within the framework of the Swiss HIV Cohort Study (SHCS). The SHCS has been approved by the following ethics committees: Ethikkommission beider Basel (EKBB), Kantonale Ethikkommission Bern (KEK), Comité départemental d'éthique des spécialités médicales et de médecine communautaire et de premier recours des Hôpitaux Universitaires de Genève, Commission cantonale d'éthique de la recherche sur l'être humain du canton de Vaud, Comitato etico cantonale - Cantone Ticino, Ethikkommission des Kantons St.Gallen, Kantonale Ethik-Komission Zürich (KEK). The data collection was anonymous and written informed consent was obtained from all participants. The detailed ethics approval for the SHCS can be found at `http://www.shcs.ch/206-ethic-committee-approval-and-informed-consent`.

3.6. Software. We perform simulations and cluster inference in R v3.1.2, using functions contained in the *igraph*, *ape*, and *phangorn* packages [108, 88]. We validate results

obtained for cluster definitions 2 and 3 by comparing them to those from ClusterPicker v1.3 and PhyloPart v2.0, respectively. Code is available upon request.

4. Results

4.1. Networks and phylogenies. In order to highlight the association between phylogenetic and network representations of epidemics, we plot two toy examples, involving simulated epidemics on unweighted and weighted networks of type A.

We show a simulated network on the right in Figure 11. Each island is identified by a colour: purple, orange, and green. The virus was introduced in the green island and then travelled to the purple island. A subject in the green island was also responsible for transmitting the virus to the orange island. On the left in Figure 11, we have the epidemic's phylogenetic representation. Branch colour indicates on which island sequences are evolving and transmissions are taking place. Tip label colours indicate to which transmission cluster, black or grey, a given sequence belongs and are matched with node colours in the network graph. We obtained those clusters by applying the WPGMA method (Def. 4) and selecting the cutpoint that maximised correspondence between islands and clusters. The network graph emphasises the limited correspondence between the transmission clusters we obtained and the islands. Indeed, subjects in the grey cluster were found on all three islands and shared the green island with the 10 subjects in the black cluster.

The plotted phylogeny emphasises that any clustering method assuming that a cluster is a clade cannot achieve high island recovery, as infected individuals from the same island may not form a clade. Indeed, in our example, infected individuals in the orange and the purple islands formed a clade in the phylogeny, while infected individuals in the green island did not. In general, the correspondence between island membership and a clade in the phylogeny is broken once an individual in the island transmits to another island.

A visual inspection of the phylogeny can however reveal the existence of community structure, as emphasised by Figure 12, which shows an epidemic simulated on a weighted network of type A. We attributed a weight of 0.25 to bridges, thus multiplying by 4 the mean

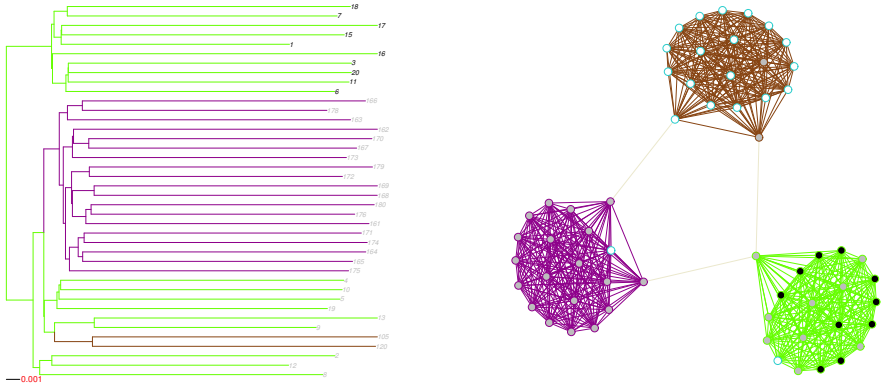

FIGURE 11. **Phylogeny and associated unweighted network graph for a simulated epidemic.** Branch colour indicates on which island transmission and evolution takes place, while tip label colour indicates cluster membership based on snipping the WPGMA dendrogram at height 0.007. Only islands with at least one infected vertex are displayed. Vertex and edge colours are matched with tip label and branch colours, respectively. A white vertex with a teal frame is infected, but undiagnosed. A white vertex with frame colour matching that of the island is uninfected.

time required for the virus to cross to another island. Both islands ended up saturated. We can easily identify the branch in the phylogeny supporting the subtree representing the sub-epidemic in the pink island. This pattern is typical: when a sufficiently large number of infection events take place after a bridge is crossed, we generally observe a subphylogeny supported by a long branch, consequence of the limited number of between-island, compared to within-island, connections.

Each island was a fully-connected graph. In epidemics on fully-connected graphs, the incidence rate is proportional to the number of edges connecting infected subjects to their susceptible neighbours. Per lineage however, the transmission rate decreases with the number of infected individuals. In a phylogeny, this translates to short branches right after a crossing event, and then an increase in branch lengths.

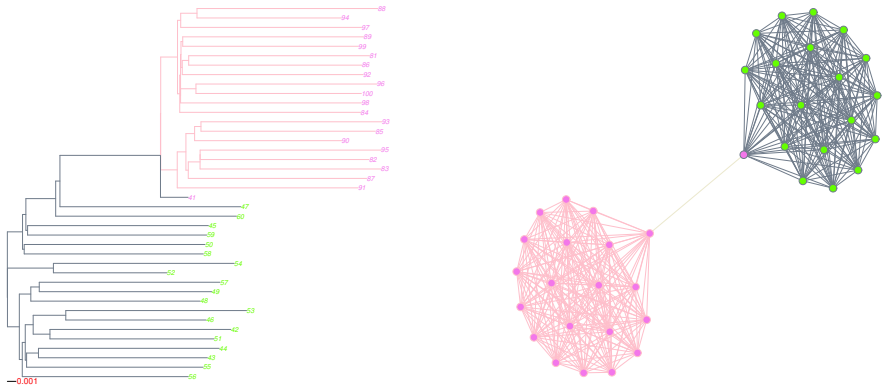

FIGURE 12. **Phylogeny and associated weighted network graph for a simulated epidemic.** Branch colour indicates on which island transmission and evolution takes place, while tip label colour indicates cluster membership based on snipping the WPGMA dendrogram at height 0.0071. Only islands with at least one infected vertex are displayed. Vertex and edge colours are matched with tip label and branch colours, respectively. A white vertex with a teal frame is infected, but undiagnosed. A white vertex with frame colour matching that of the island is uninfected.

4.2. Epidemics in type A networks. In type A sexual contact networks, correspondence between islands and communities returned by the label propagation and walktrap algorithms was close to perfect, with ARIs slightly under 1. Type A networks had a clustering coefficient of 0.94, a mean degree of 19.6, and a mean shortest path length of 2.76. Epidemics on these networks had an estimated reproduction number of 3.7.

ARIs varied widely across simulations. For instance, at cutpoint 0.02, on networks with bridge weights set at 0.25, clusters obtained from the complete linkage method produced ARIs between 0, meaning that we concluded in the absence of clusters, to 0.88, yielding a variance of 0.02. Figure 13 gives mean island recovery rates across distance requirements under the different weighting schemes for the cluster definitions stated previously. Overall, we observed low to moderate correspondence between transmission clusters and communities, with mean recovery rates lower when epidemics spread on unweighted networks. Around

the optimal cutpoint, clusters obtained under Def. 4 tended to agree more with the island structure, as evidenced by Figure 13 D, with complete linkage usually producing better correspondence.

Figure 13 A and 13 B further suggest that complete linkage worked better under a wider range of cutpoints. However, past this optimal range, Def. 2 produced greater overlap with the islands. Figure 13 C and 13 D show that even under the optimal percentile threshold, Def. 3 produced clusters that overlapped modestly with the islands, although it did slightly outperform Def. 1 and Def. 2 under optimal setting.

Figure 13 A and 13 B emphasise how weighting affects cluster recovery. In clusters resulting from the use of hierarchical clustering with complete linkage, from above 0.58 when bridge weights were set at 0.25, the ARI decreased to approximately 0.43 in the unweighted network case. Weighting also improved optimal recovery rates for Def. 1, 2, and 3, although the variation was much smaller.

4.3. Epidemics in type B networks. The mean ARIs for correspondence between the small islands in type B sexual contact networks and communities returned by the walktrap and label propagation algorithms were also very close to 1. Type B networks had a mean clustering coefficient of 0.98, a mean degree of 17.7, and a mean shortest path length of 5.9. Epidemics on these networks also had an estimated reproduction number of 3.7.

As before, we observed large variations in ARIs across simulations. For example, on networks with bridge weights taking value 0.25, clusters obtained using the complete linkage method, at optimal cutpoint 0.04, had ARIs ranging from 0.02 to 1, producing a variance of approximately 0.02. Figure 14 shows that all cluster definitions could be tuned to recover the island structure fairly accurately. In the weighted case (Fig 14 B), optimal island recovery reached over 0.90. Once again, hierarchical clustering (Def. 4) tended to perform better, but the difference in optimal recovery rates was rather small. Like before, Def. 2 clearly outperformed other definitions at higher cutpoints. The improved performance of clade-based definitions was not a surprise. After all, in type B networks, all epidemic outbreaks in the smaller islands form clades in the phylogeny.

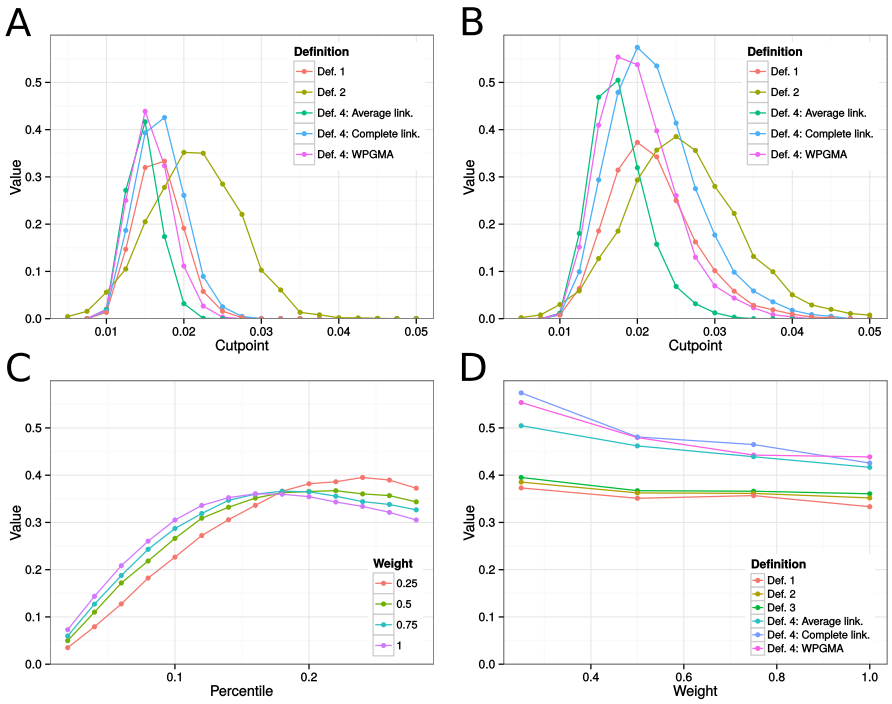

FIGURE 13. **Estimates of island recovery for epidemics on type A sexual contact networks, measured with the adjusted Rand Index (ARI).** Figure A indicates the mean ARI across cutpoints for the different cluster definitions applied to epidemics simulated on unweighted networks. In Figure B, we show corresponding results for weighted networks with between-island transmission rate equal to 25% the within-island transmission rate. In Figure C, each curve gives mean island recovery rates across between-tip distance percentile requirements in Def. 3 for networks with between-island transmission rates 25%, 50%, 75%, or 100% the within-island transmission rate. Figure D gives the optimal ARIs across between-island transmission rates. Each curve represents the maximum achieved mean island recovery under the different cluster definitions. For example, the values at bridge weight 0.25 are the suprema in Figure B, as well as the supremum of the red curve in Figure C. Def. 1-4 correspond to the methods for cluster detection described in the Methods section.

4.4. Epidemics in type C networks. The more realistic networks were of mean size 985.70, with mean degree 5.70, mean clustering coefficient 0.37, and mean shortest path

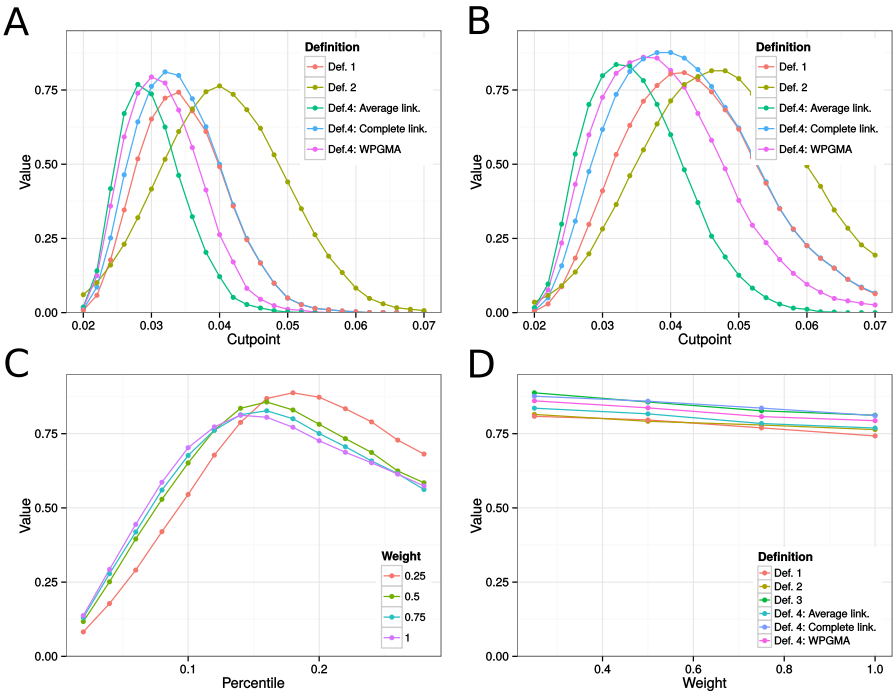

FIGURE 14. **Estimates of island recovery for epidemics on type B sexual contact networks, measured with the adjusted Rand Index (ARI).** Figure A indicates the mean ARI across cutpoints for the different cluster definitions applied to epidemics simulated on unweighted networks. In Figure B, we show corresponding results for weighted networks with between-island transmission rate equal to 25% the within-island transmission rate. In Figure C, each curve gives mean island recovery rates across between-tip distance percentile requirements in Def. 3 for networks with between-island transmission rates 25%, 50%, 75%, or 100% the within-island transmission rate. Figure D gives the optimal ARIs across between-island transmission rates. Each curve represents the maximum achieved mean island recovery under the different cluster definitions. For example, the values at bridge weight 0.25 are the suprema in Figure B, as well as the supremum of the red curve in Figure C. Def. 1-4 correspond to the methods for cluster detection described in the Methods section.

length 5.64. Islands had sizes ranging from 5 to 16, with mean 9.84. Approximately 9.63%

of islands were of size 5. The island size distribution frequency decreased slowly, but systematically, with increasing size until it reached 2.90% for islands of size 16 (Figure 16 in Supplementary Material). Within-island degree distributions had the heavy right tail resulting from preferential attachment (Figure 17 in Supplementary Material). Communities returned by the label propagation algorithm overlapped strongly with the islands, with an ARI close to 0.99. With an ARI of 0.80, the communities proposed by the walktrap algorithm still matched the islands rather closely. Epidemics on type C networks had an estimated reproduction number of 2.35.

Because of the larger size of the epidemics, variability in ARIs obtained across simulations was lower than before. For example, in networks with bridge weights set to 0.25, at optimal cutpoint 0.025, clusters inferred using the complete linkage method lead to ARIs ranging from 0.39 to 0.70, with a variance of 0.003. Figure 15 presents mean island recovery rates across distance requirements under the different weighting schemes and cluster definitions. Despite considerable differences between networks of type A and type C, the results were similar to those outlined before. Once again, hierarchical clustering algorithms (Def. 4) suggested transmission clusters that overlapped more with islands, and among those algorithms, the complete linkage method performed better, as it worked better under a wider range of cutpoints and produced slightly higher suprema.

The other three definitions offered a better performance than in networks of type A. We suggest that this is due to the increase in the number of introductions and the larger number of infected islands, leading to infected individuals from the same island forming clades in the phylogeny more often.

5. Discussion

Disagreements persist about the interpretation of transmission clusters inferred from phylogenies [94]. The present work illustrates how transmission clusters obtained by applying common methods may overlap only partially with islands in the underlying network, especially when epidemics result from a small number of introductions. Because hierarchical clustering methods do not require clusters to be clades in the phylogeny, they tended to

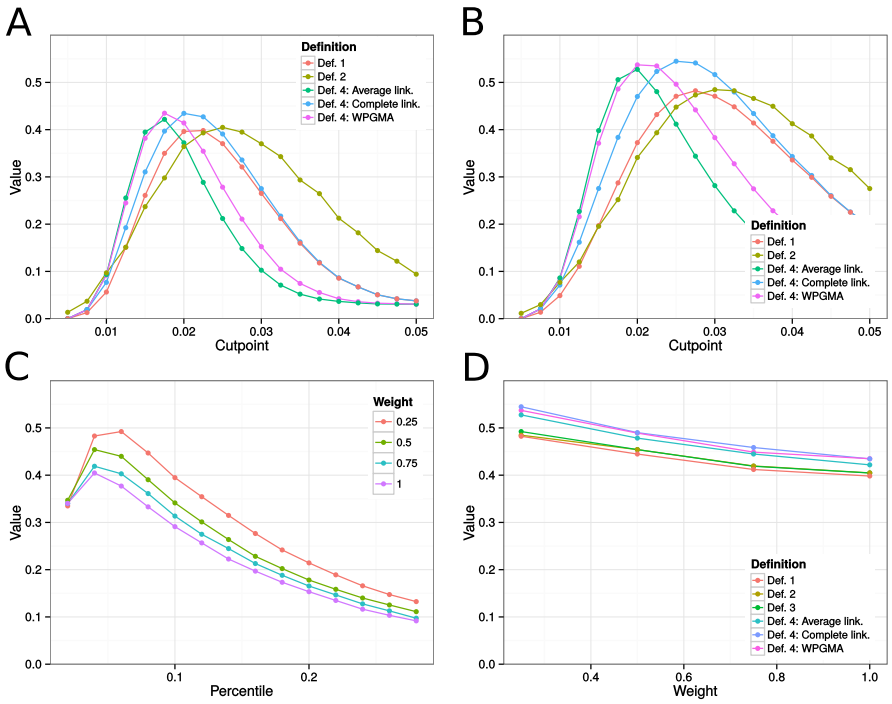

FIGURE 15. **Estimates of island recovery for epidemics on type C sexual contact networks, measured with the adjusted Rand Index (ARI).** Figure A indicates the mean ARI across cutpoints for the different cluster definitions applied to epidemics simulated on unweighted networks. In Figure B, we show corresponding results for weighted networks with between-island transmission rate equal to 25% the within-island transmission rate. In Figure C, each curve gives mean island recovery rates across between-tip distance percentile requirements in Def. 3 for networks with between-island transmission rates 25%, 50%, 75%, or 100% the within-island transmission rate. Figure D gives the optimal ARIs across between-island transmission rates. Each curve represents the maximum achieved mean island recovery under a different cluster definition. For example, the values at bridge weight 0.25 are the suprema in Figure B, as well as the supremum of the red curve in Figure C. Def. 1-4 correspond to the methods for cluster detection described in the Methods section.

produce clusters that corresponded more closely to islands. The three clade-based methods performed almost as well as hierarchical clustering when we simulated under scenarios where

each transmission chain within an island formed a clade (type B networks). Since the simulated islands form clear-cut communities, we conclude that community structure cannot be inferred reliably using the existing phylogenetic clustering tools.

The present study does have limitations. Real genotyping data from a mapped transmission network was unavailable and so, we had to rely on simulations. It follows that we cannot establish without a doubt that our results would generalise beyond the scenarios we selected. Nevertheless, the similarities between the results we obtained across network types make us confident that our conclusions are reasonably robust to the characteristics of the network structure. We recognise that the networks we simulated are major simplifications of empirical networks, but they do exhibit some important properties of real sexual contact networks, namely moderate to high clustering, community structure, and low mean shortest path lengths.

The island network model has the important advantage of comprising unambiguous communities. If transmission clusters do not approximate communities in such networks, it is unlikely they will correspond to communities in other network models.

In summary, current widely-used phylogenetic clustering methods (Def. 2 and 3, which are extensions of Def. 1) often failed to recover sexual contact network communities reliably. A major drawback was their assumption that clusters form clades in the phylogeny, while many clusters were not clades, like in the epidemics simulated in type A and type C networks. Conventional hierarchical clustering methods outperformed the more sophisticated methods in the case where clusters were not clades, and produced similar performance in networks where clusters always formed clades, like in networks of type B. However, particularly for epidemics in type A and type C networks, hierarchical clustering was still far from optimal. Therefore, we call for new clustering methods improving on existing approaches, by dropping for instance the clade assumption.

Understanding the link between HIV transmission clusters and community structure in sexual contact networks could help inform policymaking to curb HIV incidence in MSMs. This work stresses the need for new clustering algorithms that focus on community recovery,

which remains limited and incidental under current methods. We speculate that covariate information, such as sociodemographic characteristics, could be helpful in community detection, and could best be integrated within a novel parametric approach that, unlike conventional methods, would not rely on the clade assumption and ad hoc requirements.

6. Acknowledgements

L. V. would like to thank the ThinkSwiss program, as well as David J. Vasilevsky for his invaluable help in designing and writing the program whose outputs form the core of this work. L.V. would also like to thank Dorita Avila for providing phylogenies used in configuring the network simulations and the people working in T.S.'s team for their support.

This study has been conducted within the framework of the Swiss HIV Cohort Study (SHCS project number 776). The members of the SHCS are: Aubert V, Battegay M, Bernasconi E, Böni J, Bucher HC, Burton-Jeangros C, Calmy A, Cavassini M, Dollenmaier G, Egger M, Elzi L, Fehr J, Fellay J, Furrer H (Chairman of the Clinical and Laboratory Committee), Fux CA, Gorgievski M, Günthard H (President of the SHCS), Haerry D (deputy of "Positive Council"), Hasse B, Hirsch HH, Hoffmann M, Hösli I, Kahlert C, Kaiser L, Keiser O, Klimkait T, Kouyos R, Kovari H, Ledergerber B, Martinetti G, Martinez de Tejada B, Metzner K, Müller N, Nadal D, Nicca D, Pantaleo G, Rauch A (Chairman of the Scientific Board), Regenass S, Rickenbach M (Head of Data Centre), Rudin C (Chairman of the Mother & Child Substudy), Schöni-Affolter F, Schmid P, Schüpbach J, Speck R, Tarr P, Telenti A, Trkola A, Vernazza P, Weber R, Yerly S.

Supplementary Material

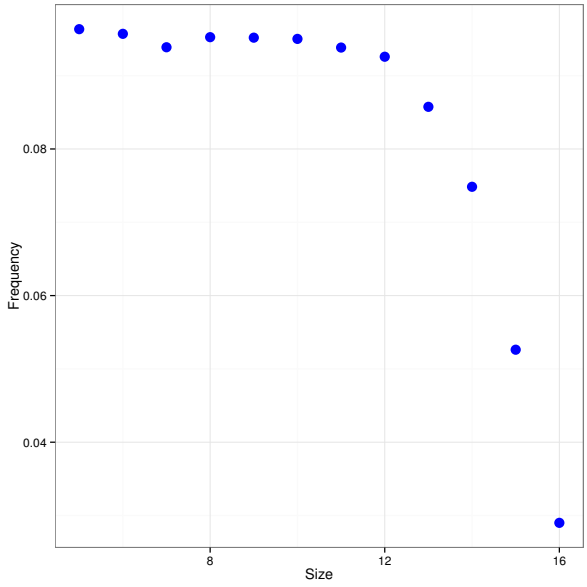

FIGURE 16. **Island size distribution in type C networks.**

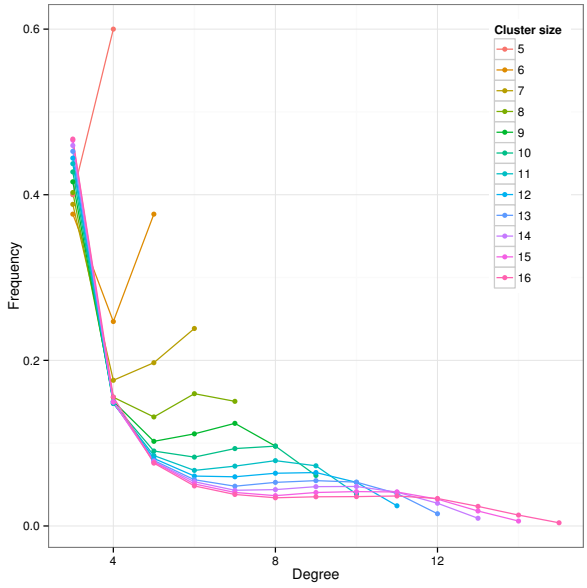

FIGURE 17. **Within-island degree distributions, stratified on island size, in type C networks.**

7. Bridge between manuscript 1 and manuscript 2

Manuscript 1's aim, in the context of the thesis, was mainly to clarify the meaning of phylogenetic clusters. The variety of standards used for clustering HIV-1 sequence data is problematic, as it reflects a lack of agreement in the literature as to what phylogenetic clusters correspond to. Clinicians view them as estimates of transmission clusters [19], but the implications of that interpretation remain unclear. Manuscript 1 sought to address this issue.

Although the work presented in manuscript 1 did not end up producing a very clear network interpretation of phylogenetic clusters, it did highlight the need for phylogenetic clustering methods based on an intuitive cluster definition that agrees with conventional epidemiologic insight. Manuscript 2 describes such a method. Rather than making clusters the results of snipping a phylogenetic estimate at arbitrary locations, it weaves them into the inference process itself. By explicitly linking clusters to phylogenetic branch length patterns, it produces unambiguous cluster estimates that can be readily assumed to result from quick transmission chains.

CHAPTER 4

Manuscript 2: "DM-PhyClus: A Bayesian phylogenetic algorithm for infectious disease transmission cluster inference"

1. Preamble

HIV-1 transmission cluster inference conventionally relies on a phylogenetic estimate, with a measure of confidence in clades. The lack of a clear definition for transmission clusters has lead to studies using a wide array of criteria for selecting clusters from a phylogeny, with varying implications [26].

Transmission clusters are thought to result from quick transmission chains, that is, series of concurrent transmission events close in time. Manuscript 2 introduces a new Bayesian phylogenetic clustering algorithm, called *DM-PhyClus*, that estimates clusters now defined as sets of sequences produced by quick transmission chains.

The algorithm splits the sample phylogeny into several components: each cluster has its own *within-cluster phylogeny*, whose root is linked to that of other clusters by the *between-cluster phylogeny*. The between-cluster phylogeny, in other words, is a tree whose tips correspond to the Most Recent Common Ancestors (MRCAs) of each cluster. Since we assume transmission clusters result from quick transmission chains, we assign branch lengths in the within-cluster phylogenies priors with a small mean. We let cluster membership follow the Dirichlet-multinomial distribution, with a Poisson weight accounting for the number of clusters observed in the sample. Explicit modelling of the partitioning of the sample into clusters lets us obtain uncertainty estimates for cluster configurations, making cluster inference more rigorous.

We estimate the posterior distribution with MCMC, with cluster point estimates obtained from either the maximum posterior probability state, called the *MAP estimate*, or

by requiring a minimum co-clustering frequency, computed from all cluster assignment index vectors sampled from the posterior, for any two sequences to be assigned to the same cluster, producing the *linkage* estimate. We let the chain move within a subspace of cluster membership indices defined by either the maximum likelihood topological estimate produced by RAxML [111], or a maximum posterior probability tree estimate obtained by exploring neighbouring configurations. We come up with starting values for cluster membership indices by comparing partitions resulting from the application of a range of genetic distance cutpoints. We pick the partition found to maximise the Dunn index [35]. After testing cluster recovery in a simulation setting, we cluster a sample of 526 real HIV-1 subtype B sequences, collected for the Quebec HIV genotyping program [19].

Simulations revealed cluster recovery rates higher than those obtained with the conventional maximum likelihood phylogenetic clustering approach. The real data analysis, on the other hand, uncovered a set of clusters that largely overlaps with the one computed in a previous study [20].

The method does have some limitations. The results can be sensitive to prior specification for the concentration parameter, whose interpretation is not straightforward. The simulations do however provide some helpful insight into configuring that parameter. Further, the use of a fixed topology, necessary to ensure good mixing in the chain, results in an underestimation of the uncertainty around returned cluster configurations. Defining a transition kernel that would solve the mixing problem encountered when the chain also performs transitions in the topological space would be a worthwhile improvement.

DM-PhyClus: A Bayesian phylogenetic algorithm for HIV-1 transmission cluster inference

Luc Villandré[1] Aurelie Labbe[2¶] Bluma Brenner[3¶] Michel Roger[4,5¶] David A. Stephens[6¶]

1 Department of Epidemiology, Biostatistics, and Occupational Health, McGill University

2 Department of Decision Science, HEC Montréal

3 McGill AIDS Centre, Lady Davis Institute, Jewish General Hospital

4 Centre de Recherche du Centre Hospitalier de l'Université de Montréal (CRCHUM)

5 Département de microbiologie, infectiologie et immunologie, Université de Montréal

6 Department of Mathematics and Statistics, McGill University

Abstract

Background. Conventional phylogenetic clustering approaches rely on arbitrary cutpoints applied a posteriori to phylogenetic estimates. Although in practice, Bayesian and bootstrap-based clustering tend to lead to similar estimates, they often produce conflicting measures of confidence in clusters. The current study proposes a new Bayesian phylogenetic clustering algorithm, which we refer to as *DM-PhyClus*, that identifies sets of sequences resulting from quick transmission chains, thus yielding easily-interpretable clusters, without using any ad hoc distance or confidence requirement. **Results.** Simulations reveal that DM-PhyClus can outperform conventional clustering methods, as well as the Gap procedure, a pure distance-based algorithm, in terms of mean cluster recovery. We apply DM-PhyClus to a sample of real HIV-1 sequences, producing a set of clusters whose inference is in line with the conclusions of a previous thorough analysis. **Conclusions.** DM-PhyClus, by eliminating the need for cutpoints and producing sensible inference for cluster configurations,

can facilitate transmission cluster detection. Future efforts to reduce incidence of infectious diseases, like HIV-1, will need reliable estimates of transmission clusters. It follows that algorithms like DM-PhyClus could serve to better inform public health strategies.

2. Introduction

The collection and, often public, availability of viral genotyping data has made phylogenetics, the field concerned with the inference from genetic data of the ancestral history of organisms, a popular tool for modelling epidemics [46, 64]. Phylogenetic models represent the ancestral relationships between sequences of nucleotides or amino acids with a hierarchical tree structure known as a *phylogeny*. Phylogenetics can help guide public health efforts to curb incidence of HIV-1 and tuberculosis [21, 18, 125], by revealing the existence of *transmission clusters*, epidemiologically-linked individuals infected by a genetically-similar pathogen. Transmission clusters are known to affect incidence and may hinder the implementation of effective intervention strategies [20].

2.1. Transmission cluster inference. Observed clustering in viral sequencing data, thought to result from series of fast onward transmission events called *quick transmission chains*, is a convenient proxy for transmission clusters [19]. To estimate transmission clusters from an inferred phylogeny, a collection of ad hoc rules are conventionally applied. One normally looks for a partition of the sample into *clades*. A clade is a set of sequences corresponding to all tips descended from a given ancestral node in the tree. Usually, a clade corresponds to a cluster only when it is known with high confidence, and when its sequences are similar. Unsurprisingly, disagreements over clustering rules are common, and what the resulting partitions mean in an epidemiological sense is still unclear [26, 128].

2.2. Study objective. In the present study, we aim to propose a new Bayesian phylogenetic clustering algorithm, called *DM-PhyClus*, that eliminates the need for arbitrary distance and confidence criteria. DM-PhyClus looks directly for sets of sequences resulting from quick transmission chains, thus also improving interpretability of clusters.

2.3. Phylogenetic inference and clustering. Bayesian phylogenetic inference is commonly used in the clustering of sequencing data, mainly because it readily provides an intuitive confidence measure for inferred clades [137, 105]. Popular software implementations include BEAST and MrBayes [33, 105], which both rely on variations of the Markov Chain Monte Carlo (MCMC) approach. Convergence issues have prompted the development of several other approaches, based, for example, on Sequential Monte Carlo [16] and Stochastic Approximation Monte Carlo [28].

Software like MEGA and PAUP* [121, 118] have made Maximum Likelihood (ML) phylogenetic reconstruction a popular alternative. RAxML [111] and FastTree [93] are more recent options, designed specifically to handle large datasets. They both rely on heuristic tree-searching strategies to considerably speed up likelihood optimisation. Generally, methods for maximum likelihood phylogenetic reconstruction do not yield measures of confidence for clades, which are necessary to apply conventional clustering rules. To solve that problem, they are combined with a bootstrap scheme. However, the interpretation of bootstrap support for clades remains controversial [41, 116, 82].

Bayesian and ML phylogenetic approaches involve generating a large collection of trees. The Maximum Posterior probability (MAP) or ML estimates are natural choices for the tree that best describes the ancestry of the data. However, especially in large samples, the score for those estimates may not be much higher than that for many other trees. Therefore, summarising a collection of phylogenies by building a so-called *consensus tree* [74, 22, 61] is common. Unlike conventional point estimates, consensus trees provide measures of uncertainty for elements in the tree *topology*, an unambiguous representation of the hierarchical nesting of clades in the phylogeny.

After computing a sensible phylogenetic estimate, one can then proceed to estimate clusters. [19] define a cluster as a clade known with high confidence, and with *patristic distances* bounded above by a reasonably low value, where the patristic distance between any two sequences is calculated by summing branch lengths along the path linking the corresponding tips in the tree. The method itself however does not specify how confidence and distance requirements should be selected. In their ML-bootstrap analysis for example, [19] used a confidence threshold of 98% and a patristic distance requirement of 0.015 nt/bp.

[94] designed PhyloPart, a method that also defines clusters as clades known with high confidence. The genetic distance requirement is now formulated in terms of the median patristic distance in a clade. To conclude in clustering, we must have median patristic distance in a clade below a value equal to a reasonably low percentile of patristic distances in the entire tree. In their analyses, [94] used the 1st, 10th, 15th, and 30th percentiles. The choice of a percentile threshold is arbitrary: in their study, it was selected to maximise agreement with a number of "confirmed clusters". The paper does not however mention how those clusters were validated.

Alternatively, [97] proposed ClusterPicker, that also finds clusters by identifying clades inferred with reasonably high confidence. The distance requirement in ClusterPicker does not involve patristic distances, but rather simple pairwise estimates of genetic distance, computed for example with the JC69, K80, HKY85, or raw (Hamming) model [68, 71, 54]. The method is convenient, as it can be applied readily to consensus trees, which do not naturally have branch lengths. Once again, the tuning of the clustering requirements is left entirely to the investigator.

Clustering criteria are often arbitrary, and tend to be poorly justified. In Bayesian phylogenetic clustering, posterior probability requirements of 1 are the most common [136, 41], although studies may opt for a lower value [5]. In the ML-bootstrap framework, clade support requirements as low as 70% [25, 65, 81], or above 90% [75, 97, 19] are common. A lot of variability is also observed in genetic distance requirements. For instance, [75] use the HKY85 + γ model [54] to assess pairwise distances between sequences and impose a maximum distance of 0.045 nt/bp within any cluster. [94] instead find that a median patristic distance requirement of 0.07 nt/bp maximises correspondence with known clusters.

The variety of standards encountered in the literature may reflect a lack of agreement as to what clusters correspond to [26]. More recently, [130] proposed the *Gap Procedure*, a distance-based clustering approach that avoids phylogenetic reconstruction and cutpoint selection altogether by defining clusters based on a measure of distinctiveness. Although it is very fast, it does not provide any means to evaluate uncertainty around its point estimates. Like the Gap Procedure, the method presented in this paper aims to avoid cutpoint selection

by giving clusters a straightforward definition. However, it should also offer an intuitive measure of confidence in cluster estimates. We designed it specifically for clustering HIV-1 sequencing data, which will be the main substantive focus in the remainder of the paper.

3. Methods

DM-PhyClus is a MCMC-based algorithm [56] that innovates by relying on a definition of transmission clusters that better reflects clinical understanding, and by avoiding ad hoc distance and confidence requirements. DM-PhyClus makes use of a likelihood formulation that distinguishes between *between-cluster* and *within-cluster* components of the phylogeny, cf. Figure 18. The between-cluster phylogeny represents the ancestral relationships between each cluster's Most Recent Common Ancestor (MRCA), and the within-cluster phylogenies, the ancestral history of each cluster.

Under DM-PhyClus, clusters have a clear definition: they are sets of sequences whose ancestral history is characterised by a specific distribution for branch lengths. In order for clusters to reflect quick transmission chains, we attribute branch lengths in the within-cluster phylogenies a prior with a reasonably low mean, in comparison to that for branches in the between-cluster phylogeny.

3.1. Likelihood. We compute the tree likelihood recursively with Felsenstein's tree-pruning algorithm [42]. Let $(y_1, \ldots, y_n)$ denote the sequence data, and $y_{i,s}$, the state at the s'th site, $s = 1, \ldots, S$, of sequence i. If sequences are made up of nucleotides, $y_{i,s}$ can take one of 4 values, each represented by a unit vector of length 4. For example, nucleotides A and T are represented by vectors $(1, 0, 0, 0)$ and $(0, 1, 0, 0)$, respectively.

At each site, evolution along branches of the tree, whose topology is denoted τ, follows a continuous time Markov chain with rate matrix Q. We denote branch lengths in the between-cluster and within-cluster phylogenies $\boldsymbol{l}^{(b)}$ and $\boldsymbol{l}^{(w)}$, respectively, with $\boldsymbol{l} = (\boldsymbol{l}^{(b)}, \boldsymbol{l}^{(w)})$. Further, we assume that among-loci variation in evolution rates follows a discrete gamma

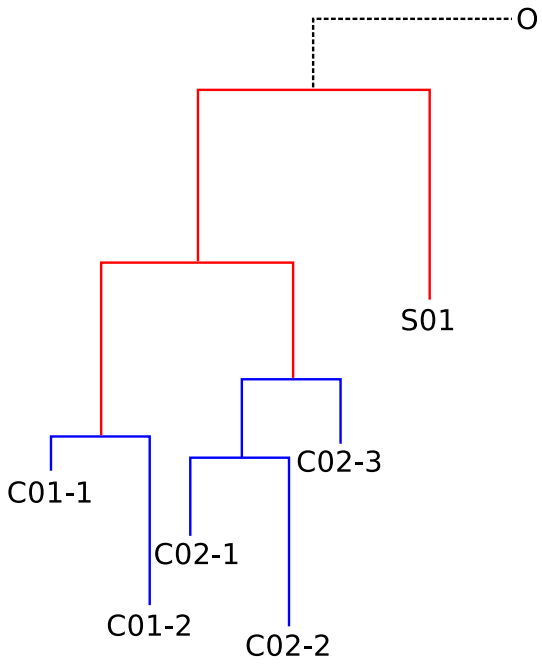

FIGURE 18. **A phylogeny split into between- and within- cluster components**. Sequences C01-1 and C01-2 belong to cluster 1, while C02-1, C02-2, and C02-3 belong to cluster 2. Sequence S01 is a singleton, that is, a cluster of size 1, and O is an outgroup, used to root the sample phylogeny. The red sub-phylogeny is called the *between-cluster* phylogeny, while the blue sub-phylogenies are called the *within-cluster* phylogenies.

distribution with n_r categories and parameter r. Evolution occurs independently at different loci and so, the likelihood takes value,

$$\zeta(\tau, \boldsymbol{l}, n_r, r, Q) = \prod_{s=1}^{S} \zeta_s(\tau, \boldsymbol{l}, n_r, r, Q), \tag{11}$$

where $\zeta_s(\tau, \boldsymbol{l}, n_r, r, Q)$ represents the likelihood contribution of site s.

Let j and k index the two children of an arbitrary internal node i in topology τ, and $x_.$ be a numerical code for the state at node ., e.g. $A = 1$, $C = 2$, $T = 3$, $G = 4$. Omitting superscripts for branches in the between-cluster and within-cluster phylogenies, we have,

$$L_{x_i}^{(s,i,m)} = \sum_{x_j} p_{x_i,x_j}(\xi_m l_j) L_{x_j}^{(s,j,m)} \sum_{x_k} p_{x_i,x_k}(\xi_m l_k) L_{x_k}^{(s,k,m)}, \tag{12}$$

where $p_{x_i,x_.}(\xi_m l_.)$ represents the transition probability from state x_i to $x_.$ along a branch of length $l_.$, with coefficient ξ_m being a scaling factor resulting from the conditioning on rate

variation category m. We note that x_i indexes the $\boldsymbol{L}^{(s,i,m)}$ vector, and it follows that the vector has as many elements as there are states in the data, e.g. 4 for nucleotide data. From the Markov assumption, it follows that,

$$p_{x_i,x_\cdot}(\xi_m l_\cdot) = \exp(Q\xi_m l_\cdot).$$

When index i is for a tip, we have that $\boldsymbol{L}^{(s,i,m)} = y_{i,s}$. We must compute $\boldsymbol{L}^{(s,i,m)}$ for each combination of locus s, node i, and rate variation category m.

We start by computing $\boldsymbol{L}^{(s,i,m)}$ for all nodes i whose children j and k are both tips. Then, we list all pairs of nodes j and k for which both $\boldsymbol{L}^{(s,j,m)}$ and $\boldsymbol{L}^{(s,k,m)}$ are known, and compute $\boldsymbol{L}^{(s,i,m)}$ for each of them.

Let the root of the tree have index ϑ. We have that the likelihood contribution of site s takes value,

$$\zeta_s(\tau, \boldsymbol{l}, n_r, r, Q) = \frac{1}{n_r} \sum_{m=1}^{n_r} \sum_{x_\vartheta} L_{x_\vartheta}^{(s,\vartheta,m)} p_{x_\vartheta},$$

where $\boldsymbol{p}$ represents the limiting probabilities of the Markov chain.

In real DNA sequences, sequencing may reveal that two or more nucleotides can be found at certain loci, producing an *ambiguity*. In Felsenstein's tree-pruning algorithm, ambiguities are expressed as a sum of the unit vectors for the potential states. For example, if A and T are observed at site m in sequence i, we get that $y_{i,m} = [1, 1, 0, 0]$.

3.2. Priors. We assign branch lengths in the between-cluster phylogeny a log-normal prior with parameters μ and σ. We picked that distribution because of its potentially heavy right tail, which allows for a small number of distinctively long branches. We tune priors for those parameters based on a desired mean and coefficient of variation. To lighten the computational load, we assign that mean a uniform prior over a finite number of discrete values, and the coefficient of variation is fixed. We assign branch lengths in within-cluster phylogenies an exponential prior with rate δ, whose prior is, like before, discrete uniform over a finite range of sensible values.

We assign cluster membership indices $(c_1, \ldots, c_n)$ a multinomial prior with probability parameters $(p_1, \ldots, p_{\max(\boldsymbol{c})})$, weighted by values from a Poisson distribution, with rate parameter λ, evaluated at $\max(\boldsymbol{c})$ and an indicator function giving probability 0 to configurations not meeting the clade assumption,

$$P(c_1, \ldots, c_n \mid \tau, \lambda, \boldsymbol{\pi}) \propto \binom{n}{n_1 \ldots n_{\max(\boldsymbol{c})}} \pi_1^{n_1} \cdots \pi_{\max(\boldsymbol{c})}^{n_{\max(\boldsymbol{c})}} \frac{\exp(-\lambda)\lambda^{\max(\boldsymbol{c})}}{\max(\boldsymbol{c})!} \times \\ \times I[\text{Partition allowed by } \tau], \tag{13}$$

with $n_k = \sum_{i=1}^{n} I[c_i = k]$ and $I[.]$ being an indicator function.

The probability parameters $(\pi_1, \ldots, \pi_{\max(\boldsymbol{c})})$ have a symmetric Dirichlet hyperprior with concentration parameter α, to which we assign a gamma hyperprior with shape and scale parameters η and β. We summarise parameters in Figure 19.

3.3. Posterior probability derivation. We are interested primarily in the posterior distribution of cluster membership indices $\boldsymbol{c}$ and so, we marginalise out probability parameters $\boldsymbol{\pi}$, as well as all branch lengths. Marginalising out $\boldsymbol{\pi}$ from Equation 13, we obtain,

$$P(c_1, \ldots, c_n \mid \tau, \boldsymbol{\alpha}, \lambda) \propto \frac{B(\boldsymbol{n} + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})} \binom{n}{n_1 \ldots n_{\max(\boldsymbol{c})}} \frac{\lambda^{\max(\boldsymbol{c})} \exp(-\lambda)}{\max(\boldsymbol{c})!} \times \\ \times I[\text{Partition allowed by } \tau], \tag{14}$$

with,

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{\max(\boldsymbol{c})} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{\max(\boldsymbol{c})} \alpha_i\right)}.$$

We use Monte Carlo integration to marginalise out branch lengths from the likelihood. When the number of Monte Carlo replications N_{MC} is large enough, the probability of a transition from state x_i to x_j over any given branch is approximately,

$$P(x_j \mid x_i, \boldsymbol{c}) = \int_{\mathcal{D}(l|\boldsymbol{c})} [\exp(Ql)]_{(x_i,x_j)} p(l \mid \boldsymbol{c}) dl \approx \frac{1}{N_{MC}} \sum_{k=1}^{N_{MC}} [\exp(Ql_k)]_{(x_i,x_j)}, \tag{15}$$

where $\mathcal{D}(l \mid \boldsymbol{c})$ is the domain of $l \mid \boldsymbol{c}$, $p(l \mid \boldsymbol{c})$ is the prior distribution of l conditional on $\boldsymbol{c}$, and l_k is drawn from that distribution. $[\exp(Ql)]$ denotes the transition probability matrix along a branch of length l, and $[\exp(Ql)]_{(x_i,x_j)}$ represents element (x_i, x_j) of that matrix. The

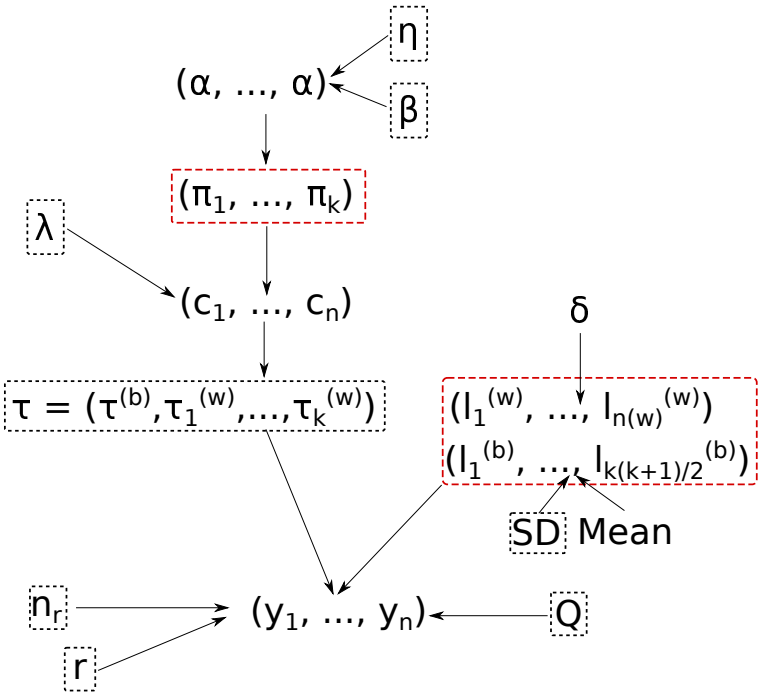

FIGURE 19. **Graphical representation of the relationships between parameters and the data**. Parameters in a black box are fixed. Parameters in a red box are marginalised out. The vector $(y_1, \ldots, y_n)$ is the sample, and "SD" stands for standard deviation. We denote the within-cluster phylogenies $(\tau_1^{(w)}, \ldots, \tau_k^{(w)})$, k being the number of clusters, and the between-cluster phylogeny, $\tau^{(b)}$. Within-cluster phylogenies are degenerate when they support a cluster of size 1, while the between-cluster phylogeny is degenerate when the sample comprises only one cluster. The log-normal prior distribution for the between-cluster branch lengths is reparameterised in such a way that it has mean and standard deviation parameters, like in the normal distribution.

conditioning on $\boldsymbol{c}$ appears as a result of the marginalisation, because of the different priors for branch lengths in the within-cluster phylogenies and the between-cluster phylogeny.

The posterior distribution of the cluster membership indices is denoted,

$$P(c_1, \ldots, c_n \mid y_1, \ldots, y_n, \tau, \boldsymbol{\alpha}, \lambda) \propto \zeta(\tau, n_r, r, Q \mid c_1, \ldots, c_n) P(c_1, \ldots, c_n \mid \tau, \boldsymbol{\alpha}, \lambda),$$

where $P(c_1, \ldots, c_n \mid \tau, \boldsymbol{\alpha}, \lambda)$ is given by Equation 14 and $\zeta(\tau, n_r, r, Q \mid c_1, \ldots, c_n)$ is obtained by replacing $p_{x_i, x_\cdot}(\xi_m l_\cdot)$ in Equation 12 by the approximation derived in Equation 15, but with simulated branch lengths l_k being multiplied by ξ_m. There is a one-to-one correspondence

between $(c_1, \ldots, c_n)$ and the breakdown of τ into within-cluster phylogenies and between-cluster phylogeny, and the conditioning on $(c_1, \ldots, c_n)$ in the marginal likelihood appears as a result.

3.4. Transition kernels and Metropolis-Hastings (MH) ratios. DM-PhyClus first searches for a sensible phylogenetic estimate, that acts to restrict the space of potential cluster membership indices, and then, conditional on that phylogeny, performs successive Metropolis-Hastings (MH) updates of the concentration parameter and the cluster membership indices.

We sample tentative transitions in the space of concentration parameter α from a uniform distribution defined over an interval of length 1 centred around the current value of α, resulting in the transition kernel ratio reducing to 1. We propose moves in the space of cluster membership indices $\boldsymbol{c}$ by using a cluster split-merge strategy. Any cluster of size 2 or more can be split in two disjoint clusters, corresponding to the clades supported by the children of the original cluster's root. We can merge any two neighbouring clusters, or in other words, any two clusters whose most recent common ancestor is at most one split above their respective roots. The transition kernel is a discrete uniform distribution over all split-merge transitions allowed by the topology from the current state. It follows that the transition kernel ratio is equal to the total number of potential moves from the current configuration divided by the total number of potential moves starting from the proposal. With the ratio of priors obtained from Equation 14 and the conventional likelihood ratio, we have all necessary components for computing the MH ratio.

3.5. Point estimates for cluster membership indices. We produce two kinds of estimates for cluster membership indices, the Maximum Posterior probability (MAP) estimate, and the *linkage-xx* estimate, which we obtain in three steps,

(1) Derive an *adjacency matrix* from each sampled cluster membership indices vector. An adjacency matrix is a symmetrical matrix with a 1 at position (i, j) if elements i and j co-cluster, and with a 0 otherwise.

(2) Average adjacency matrices computed in step 1 and apply a co-clustering frequency threshold of $xx\%$.

    The average adjacency matrix provides co-clustering frequencies. All frequencies higher than the threshold are rounded up to 1, while all others are rounded down to 0.

(3) Identify all *disjoint* sets, called *modules* or *components*, from the matrix obtained in step 2.

    Two sets of sequences are disjoint if no co-clustering exists between them. We use the walktrap algorithm [91] to detect disjoint sets, which leads to the cluster estimates.

We present a structured, step-by-step description of DM-PhyClus in Supplementary Material S1.

3.6. Simulation study.

3.6.1. *Data.* We simulate an HIV-1 sequence dataset of size 200 by going through the following steps:

(1) Sample the total number of clusters from a Poisson distribution with mean 50,
(2) Sample cluster assignment probabilities from a symmetric Dirichlet distribution with a concentration parameter generated from a normal distribution with mean 10 and standard deviation 2,
(3) Sample 200 values from a multinomial distribution with the obtained probability vector,
(4) Generate each within-cluster phylogeny by picking a topology at random, and by sampling branch lengths from an exponential distribution with mean equal to 0.003,
(5) Generate the between-cluster phylogeny by picking a topology at random, and by sampling branch lengths from a log-normal distribution with mean and standard deviation equal to 0.008,
(6) Let the *HXB2* sequence evolve along the simulated tree, with evolution rate matrix and limiting probabilities obtained from [92].

HXB2 is an HIV-1 subtype B sequence that serves as a reference for site position numbers in any HIV-1 sequence. In other words, the range of site indices in any HIV-1 sequence is found by aligning it with HXB2. We generate 50 datasets in total, and add to each of them an arbitrary subtype C outgroup (`http://www.hiv.lanl.gov/`, accession number: AB254141) for rooting the inferred phylogenies. We list parameters used for data generation in Supplementary Material S2.

3.6.2. *Scenarios.* Assessing sensitivity of the cluster estimates to the concentration parameter prior is vital, as it may be challenging to properly specify in practice. For each simulated dataset, we run DM-PhyClus under the assumption that the concentration parameter follows a gamma distribution with scale parameter 0.1, and, successively, with means 1, 10, and 100. The use of fixed estimates for the mutation rate matrix and limiting probabilities may also affect cluster recovery. To verify that such a restriction is not overly detrimental to cluster recovery, we use values for those parameters obtained from a separate analysis of a real HIV-1 sequence dataset, that we ensure are reasonably different from those used for data generation.

3.6.3. *Setup.* Given the synthetic nature of the problem, tuning priors for branch lengths is difficult and so, we opt for an empirical Bayes approach, where we use maximum likelihood phylogenetic estimates to derive mean branch lengths in the within- and between-cluster phylogenies. We then define a range around each of the obtained means with radius equal to 8% of the obtained mean. Finally, we select 20 equidistant points in each range, at which we compute transition probability matrices by sampling $100,000$ values from the log-normal distribution for between-cluster branch lengths, or the exponential distribution for within-cluster branch lengths.

We use RAxML [111] to obtain an estimate of the maximum likelihood phylogeny, and to perform 500 bootstrap iterations, producing the usual clade support estimates. We then get starting values for the cluster membership indices by running a depth-first search on the tree. We stop exploration along any path once we find a clade with bootstrap support greater than 70% and with patristic distances below a certain threshold, selected through maximisation of the Dunn index [35], a measure of clustering quality. In a first round of simulations, we

use that partition as a starting value for the chain, and the maximum likelihood topology to bound the space of cluster solutions.

In a second round of simulations, before launching the main chain, we explore the topological space around the maximum likelihood phylogeny, using nearest-neighbour interchange to find a configuration that improves posterior probability, and letting values for the concentration parameter and cluster membership indices vary as well. We start the MCMC run once a suitable topology is identified. We present an exhaustive list of the tuning parameter values used in the simulations in Supplementary Material S2.

3.6.4. *Chain configuration and point estimates comparison.* For each simulated dataset, we produce $55,000$ samples from the posterior distribution of the cluster membership indices vector. We apply a thinning ratio of 1 over 50, and take out the first $5,000$ iterations as a burn in, leaving us with 1000 samples. Once the MCMC run is complete, we obtain the MAP and linkage-xx cluster estimates, and measure overlap between the real and inferred clusters with the Adjusted Rand Index (ARI), a measure of similarity between two sets of clusters. It involves the ratio of pairs of elements that are similarly co-clustered or dissociated in both sets to the total number of pairs in the sample, combined with a numerical adjustment for chance. It is bounded above by 1, which indicates perfect correspondence. We compare those estimates to those we initially obtained from RAxML, which we refer to as the *Bootstrap-70* estimates, and to the estimates from the so-called *Gap procedure*, a quick distance-based genetic sequence clustering approach that requires minimal tuning [130]. The Bootstrap-70 estimate is a natural standard for comparison, since it is obtained by applying a conventional method for the clustering of HIV-1 sequencing data [41].

3.7. Real data analysis.

3.7.1. *Data.* The original sample consists of $3,537$ HIV-1 subtype B sequences collected for the Québec HIV genotyping program [19]. Each sequence is from a different male patient belonging to the Injection Drug User (IDU) or Men who have Sex with Men (MSM) risk category, and that has not yet started antiretroviral therapy, the standard treatment regimen for HIV-positive individuals. The dataset includes sites 10-297 of the Protease (PR) region, and 112-741 of the Reverse Transcriptase (RT) region, of the *pol* gene.

[20] obtained an initial set of clusters by partitioning the sample through inspection of the maximum likelihood tree, selecting clades with bootstrap support greater than 98% and whose patristic distances were below 0.01 nt/bp. They also looked for congruent polymorphisms and mutational motifs. Whenever new sequences entered the database, they updated their cluster estimates by re-inferring the tree, and attaching new sequences to previously-inferred clusters when the clade they belonged to had bootstrap support greater than 98%. They also used clinical and demographic information to exclude sequences from inferred clusters.

We focus on a subsample of 526 sequences, made up of 18 previously-inferred clusters of sizes ranging from 2 to 69, inclusively, as well as 12 singletons selected uniformly at random in the original sample. We add to the sample 3 subtype C outgroups from Zambia, downloaded from the Los Alamos HIV-1 sequencing database (`http://www.hiv.lanl.gov/`, accession numbers AB254141, AB254142, AB254143).

3.7.2. *Bootstrap analysis.* To evaluate sensitivity of DM-PhyClus to the input topology, we produce 100 bootstrap samples of the data by resampling site indices with replacement and re-assembling each sequence based on the sampled indices. We use maximum likelihood topological estimates and use the same strategy as in the simulations to obtain starting values for the chain. Each run also consists of $55,000$ iterations, with a burn-in of $5,000$ and a thinning ratio of $1/50$.

3.7.3. *Approximation of the fully Bayesian analysis.* Fixing the topological parameter in the chain results in the inference not being fully Bayesian. Such an approximation is acceptable only so long as we can establish that the results do not differ too much from those resulting from the fully Bayesian approach. To do so, we first use MrBayes [105], run under the default configuration, to sample 1.5 million phylogenies from the posterior distribution $P(\tau \mid \boldsymbol{y}, \dots)$, where $\dots$ represents the other parameters. We take out the first $375,000$ samples as a burn-in, and apply a thinning ratio of $1/500$. Of the remaining $2,250$ samples, we select 100 uniformly at random, which we use as input in 100 separate runs of DM-PhyClus. Each run produces samples from the conditional posterior distribution of the

cluster membership indices $P(\boldsymbol{c} \mid \tau_i, \ldots, \boldsymbol{y})$, $i = 1, \ldots, 100$. Noting that,

$$P(\boldsymbol{c} \mid \boldsymbol{y}) = E_\tau[P(\boldsymbol{c} \mid \tau, \boldsymbol{y})] \approx \sum_{i=1}^{100} P(\boldsymbol{c} \mid \tau_i, \boldsymbol{y})/100,$$

we see that high overlap between the maximum posterior probability cluster membership indices obtained from the 100 chains ensures that the peak of $P(\boldsymbol{c} \mid \boldsymbol{y})$ is found at a configuration similar to those obtained in each individual run, thus confirming the quality of the approximation resulting from the conditioning assumption.

3.7.4. *Main run.* We obtain starting values with the help of RAxML, under the assumption that genetic distance follows the GTR + $\Gamma(3)$ model. As in the simulations, we configure priors for branch lengths based on the maximum likelihood topology. We use limiting probabilities and nucleotide substitution rates previously inferred for HIV-1 subtype B [92]. We assume discrete gamma substitution rate variation with 3 categories. Finally, we fix the rate parameter for the Poisson distribution at 30, the number of clusters obtained in [20]. We run $220,000$ iterations, keeping one iteration out of 150 and taking out the first $70,000$ iterations as a burn-in. We then obtain point estimates for cluster membership indices as before. An exhaustive list of tuning parameter values used in all real data analyses is available in Supplementary Material S3.

3.8. Software. We present a technical description of the software in Supplementary Material S4. We implement the algorithm in R, with functions contained in the *phangorn*, *ape*, and *phytools* libraries [108, 99]. Likelihood evaluations rely on compiled C++ code integrated into the R script using the *Rcpp* and *RcppArmadillo* packages [38, 37]. We produce starting values with RAxML [111]. Finally, we also produce cluster estimates with the the GapProcedure package [130]. A package, *DMphyClus*, is available on Github (`https://github.com/villandre/DMphyClus`) and will be submitted to CRAN.

4. Results

4.1. Simulation study. On an Intel(R) Xeon(R) CPU E7-4820 v4 2.00GHz CPU, running $55,000$ iterations took on average a bit more than 2 hours. Log-posterior probability

Topology used	Alpha mean	Estimator	Min.	Max.	10th perc.	Median	90th perc.	Mean	SD	SE
GapProcedure	-	-	0.012	0.719	0.030	0.385	0.654	0.361	0.227	0.005
Bootstrap-70	-	-	0.074	0.882	0.256	0.483	0.771	0.504	0.221	0.004
ML topology	10	MAP	0.000	0.935	0.686	0.820	0.900	0.769	0.210	0.004
		Linkage-0.7	0.000	0.946	0.711	0.853	0.920	0.793	0.213	0.004
		Linkage-0.8	0.000	0.971	0.707	0.838	0.912	0.793	0.213	0.004
		Linkage-0.9	0.000	0.962	0.710	0.822	0.893	0.771	0.206	0.004
		Linkage-1	0.089	0.710	0.359	0.494	0.631	0.484	0.129	0.003
	1	MAP	0.098	0.862	0.328	0.619	0.833	0.601	0.199	0.004
		Linkage-0.7	0.012	0.939	0.381	0.725	0.861	0.653	0.218	0.004
		Linkage-0.8	0.011	0.959	0.394	0.760	0.865	0.680	0.207	0.004
		Linkage-0.9	0.053	0.937	0.466	0.776	0.885	0.712	0.191	0.004
		Linkage-1	0.159	0.716	0.397	0.470	0.646	0.491	0.103	0.002
	100	MAP	0.123	0.931	0.594	0.848	0.917	0.790	0.196	0.004
		Linkage-0.7	0.123	0.973	0.346	0.859	0.931	0.791	0.215	0.004
		Linkage-0.8	0.123	0.971	0.348	0.852	0.920	0.785	0.211	0.004
		Linkage-0.9	0.123	0.980	0.378	0.820	0.896	0.761	0.202	0.004
		Linkage-1	0.123	0.802	0.351	0.514	0.652	0.504	0.133	0.003
MAP topology	10	MAP	0.000	0.935	0.714	0.839	0.923	0.798	0.180	0.004
		Linkage-0.7	0.000	0.950	0.727	0.858	0.919	0.818	0.172	0.004
		Linkage-0.8	0.000	0.953	0.791	0.846	0.919	0.823	0.165	0.003
		Linkage-0.9	0.000	0.947	0.751	0.824	0.891	0.798	0.156	0.003
		Linkage-1	0.000	0.686	0.318	0.449	0.598	0.454	0.117	0.002
	1	MAP	0.011	0.870	0.329	0.623	0.832	0.598	0.203	0.004
		Linkage-0.7	0.162	0.930	0.321	0.738	0.848	0.649	0.212	0.004
		Linkage-0.8	0.170	0.931	0.384	0.746	0.872	0.671	0.201	0.004
		Linkage-0.9	0.175	0.911	0.437	0.764	0.852	0.693	0.178	0.004
		Linkage-1	0.341	0.745	0.396	0.516	0.660	0.524	0.093	0.002
	100	MAP	0.123	0.947	0.761	0.854	0.914	0.816	0.171	0.003
		Linkage-0.7	0.123	0.976	0.793	0.867	0.923	0.830	0.170	0.003
		Linkage-0.8	0.123	0.970	0.789	0.857	0.914	0.825	0.169	0.003
		Linkage-0.9	0.123	0.965	0.703	0.819	0.901	0.789	0.164	0.003
		Linkage-1	0.123	0.672	0.298	0.459	0.619	0.457	0.122	0.002

TABLE 2. **Summary statistics for adjusted Rand indices (ARI) for cluster membership estimates obtained from chains run on** 50 **datasets under different simulation scenarios.**

graphs show no obvious issue with autocorrelation or convergence, and indicate good mixing (see, for example, Supplementary Material S5). We show the obtained ARIs for the six scenarios in table 2. Overall, mean cluster recovery from DM-PhyClus was superior than

that from the conventional Bootstrap-70 approach and GapProcedure, both of which usually struggled to recover the clusters. We observe a noticeable drop in mean overlap when the concentration parameter has a prior whose mean is much smaller than that used for data generation, but not when it is larger.

The linkage-xx estimates performed comparably or slightly better than the MAP estimates when the linkage requirement was $0.7 - 0.8$ and the prior on the concentration parameter had mean equal or superior to the the true value. When the prior underestimated the true concentration parameter value however, the linkage estimates greatly improved recovery, sometimes as much as much as 10%, as long as the linkage requirement was not 1. Maximum observed recovery rates were also consistently superior for the linkage estimates.

The slightly better performance of DM-PhyClus when the concentration parameter has a mean greater than that used for data generation was unexpected. We observe it both when the MAP and ML topologies are used. When the concentration parameter prior had mean 10, two chains returned a MAP configuration with a single cluster, producing the 0 in the table, which explains at least part of the gap. The datasets analysed by those chains seem to imply a hard clustering problem, as evidenced by the low recovery rates from Bootstrap-70, 0.13 and 0.18. Overall, starting with the MAP configuration from a shorter preliminary run resulted in small increases in mean recovery rates. When the concentration prior mean was 10, the same two chains as before resulted in a MAP configuration with only 1 cluster, yielding ARI = 0. With median recovery around 0.87 in the better scenarios, we are not overly worried about the consequences of using fixed values for the limiting probabilities and mutation rate matrices, as long as they are selected reasonably.

4.2. Real data analysis.

4.2.1. *Bootstrap analysis.* We measured overlap within all pairs of MAP configurations produced in the bootstrap analysis. ARIs ranged from 0.10 to 0.98, with median 0.83 and mean 0.72, indicating reasonable robustness of the chain to the assumed topology. Unsurprisingly, linkage estimates led to essentially the same conclusion. For example, overlap between cluster configurations proposed under the linkage-70 estimate ranged from 0.11 to

0.98, with median 0.83 and mean 0.70. Moreover, concordance between MAP estimates from the bootstrap replicas and the MAP cluster configuration obtained from the full data was generally high, with median and mean ARI equal to 0.88 and 0.80, respectively.

4.2.2. *Approximation of the fully Bayesian analysis.* Estimates based on the 100 topologies sampled with MrBayes were overall very similar, leading to the conclusion that the DM-PhyClus estimates are reasonable approximations of those resulting from a fully Bayesian analysis. Indeed, concordance between the MAP estimates obtained from the 100 chains tended to be high: ARIs ranged from 0.38 to 1, with median and mean 0.89 and 0.86, respectively. Overlap with the usual MAP estimate, obtained conditional on the topology found to optimise joint posterior probability after a short exploration of the topological space, was also considerable, with median and mean 0.92 and 0.90, respectively.

4.2.3. *Full data analysis.* The MAP configuration obtained from DM-PhyClus revealed the existence of 16 clusters of size 2 or more, and 2 singletons. Linkage estimates were identical to the MAP estimate when the linkage requirement was 98% or below, indicating little uncertainty in the returned partition. The Gap Procedure returned a rather similar set of clusters (ARI = 0.87). We represent clusters from DM-PhyClus against those from the curated analysis in Figure 20. DM-PhyClus has a tendency to merge neighbouring clusters, as evidenced by the smaller number of singletons and the merger of clusters 43 and 83, which also absorbed sequence r132, and of clusters 27 and 49. The GapProcedure, on the other hand, proposed a configuration with 43 clusters of size 2 or more, and 14 singletons, splitting, for example, clusters 18 and 59 in 3 and 8 sets, respectively.

5. Discussion

In this paper, we introduced a phylogenetic clustering algorithm, DM-PhyClus, that integrates an original cluster definition into cluster inference, which results in more intuitive estimates, unlike conventional approaches, that rely instead on arbitrary cutpoints applied a posteriori to a phylogenetic estimate. Simulations indicate that the algorithm can accurately recover phylogenetic clusters, often outperforming more conventional approaches. Analysis

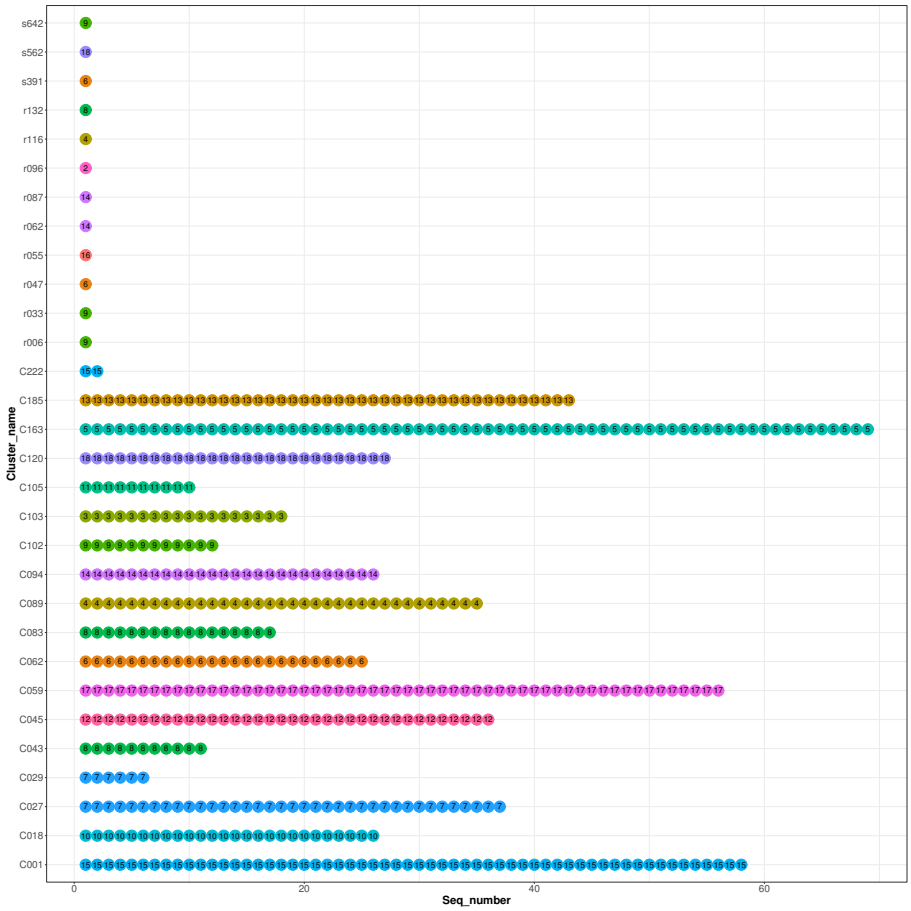

FIGURE 20. **Comparison of the DM-PhyClus cluster estimates with a proposed cluster configuration for the real dataset.** The coordinates on the vertical axis indicate cluster membership according to [20], and the colour and number of each dot, the cluster membership according to the maximum posterior probability (MAP) estimate of DM-PhyClus.

of a real dataset of HIV-1 subtype B sequences revealed a set of clusters largely similar to that from a previous analysis, but with more straightforward inference.

The study does have some limitations. Because of time constraints, we were only able to run short chains in the simulations. Log-posterior probability graphs for the simulated samples however did strongly suggest that the chains had converged, making us confident that increasing the number of iterations would not change our conclusions. We suspect that the apparent weakness of Bootstrap-70 might be in part attributable to the use of the Dunn index. For several simulated datasets, we noticed that it failed to identify the optimal

partition in terms of recovery. Comparing our results to that solution would have been unfair, however, since identifying it requires knowledge of the true clusters. For computational reasons and to ensure adequate mixing in the chain, we opted for a fixed topology, thus limiting the number of partitions the algorithm can propose and ignoring uncertainty in phylogenetic reconstruction. Although simulations and the real data analyses indicate that this simplification works well in practice, proposing an efficient transition kernel that jointly updates cluster membership indices and the phylogeny would be necessary.

Further DM-PhyClus rests on the assumption that cluster-specific phylogenies have a distinctive branch length distribution. Our goal was to reflect intuitive understanding of transmission clusters, but our branch length assumptions do remain simplistic. Phylogenies for HIV-1, for instance, are characterised by long external branches [72]. Moreover the exponential prior is known for producing overly long trees [131]. The assumption however is common in Bayesian phylogenetic inference [33], and leads to considerable computational simplifications. It is unclear whether more sophisticated, potentially dependent, branch length priors would improve cluster inference overall. Given the often high recovery rates observed in the simulations, we are confident that the simplification was not overly detrimental. Improvements to the code should also make it possible to apply DM-PhyClus to much larger datasets, such as those collected for major HIV-1 genotyping programs.

We contend DM-PhyClus is a worthwhile addition to existing methods used to detect transmission clusters. Understanding clustering in epidemics is crucial: in the case of HIV-1 among men who have sex with men for example, transmission clusters have been found to contribute overwhelmingly to incidence [20, 21]. Investigations into the reasons behind the existence of those clusters are likely to help in reducing transmission rates, and those studies will need to rely on methods based on cluster definitions that reflect clinical insight, like DM-PhyClus.

Ethics approval and consent to participate

Ethics approval for the Quebec HIV genotyping program was obtained from individual study sites, the Laboratoire de santé publique du Québec, and the Quebec Ministry of Health committee on confidentiality and access of information.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by a training award from the Fonds de recherche du Québec-Santé (FRQS), funding from the Centre de Recherches Mathématiques (CRM), a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant, and a Canadian Institutes of Health Research (CIHR) grant (CIHR HHP-126781).

Author's contributions

LV wrote the article, performed the simulations. LV, AL, and DAS jointly formulated the algorithm. AL and DAS suggested and reviewed analyses. BB and MR provided the HIV-1 sequences.

Data availability

All simulated data generated or analysed during this study are included in this published article or on Zenodo (DOI 10.5281/zenodo.839849). The Québec HIV genotyping program sequences cannot be made publicly available for confidentiality reasons. A small subset of sequences can be provided for verification purposes upon request.

Acknowledgements

We ran computations on the Guillimin and MP2 supercomputers, administered by McGill High-Performance Computing and Université de Sherbrooke, respectively, and managed by Calcul Québec and Compute Canada. The operation of these supercomputers is funded by the Canada Foundation for Innovation (CFI), ministère de l'Économie, de la Science et de l'Innovation du Québec (MESI) and the Fonds de recherche du Québec - Nature et technologies (FRQ-NT).

The Quebec HIV genotyping program is sponsored by the Ministère de la Santé et des Services sociaux (MSSS) du Québec and by the Fonds de recherche du Québec (FRQ-S) Réseau SIDA/MI.

Supplementary Material S1 - Algorithm description

Input:

(1) **Topology**: For example, the maximum likelihood topology,
(2) **Nucleotide transition rate matrix**: An empirical estimate, like the one in [92], or alternatively, one derived from the sample itself, with the help of RAxML or MrBayes for example,
(3) **Gamma shape parameter for among-loci mutation rate variation**: Assumed equal to the scale parameter, can be obtained in the same way as the nucleotide transition rate matrix. In the simulations, we use an estimate from [92],

(4) **Cluster membership indices prior**: Follows a Dirichlet-multinomial distribution, combined with a Poisson-distributed weight with a pre-determined rate parameter, e.g. the number of clusters resulting from a conventional bootstrap-maximum likelihood phylogenetic clustering analysis,

(5) **Poisson rate for the assumed number of clusters**,

(6) **Concentration parameter prior**: Assumed gamma-distributed with user-specified scale and shape parameters,

(7) **Shape and scale parameter values for the concentration parameter prior**: We set the scale parameter equal to 0.1 in all analyses, and changed the shape parameter to vary the distributional mean,

(8) **Transition kernel for the concentration parameter**: A uniform distribution with radius 0.5 centred at the current parameter value,

(9) **Transition kernel for the cluster membership indices**: A uniform distribution over all configurations reachable from the current state. A configuration is reachable if it can be obtained by splitting in two a cluster of size 2 or more, or merging two neighbouring clusters. Two clusters are considered neighbours if their respective MRCAs are siblings. Clusters are obtained by partitioning the sample into disjoint clades. It follows that each cluster can be represented, alternatively, by its MRCA. When a cluster is split in two, the MRCAs of the new clusters are the children nodes of the original cluster's MRCA. When two neighbouring clusters are merged, the new cluster's MRCA is the parent node of the selected two clusters' MRCAs.

(10) **Prior for branch lengths in the within-cluster phylogenies**: Assumed to follow the exponential distribution,

(11) **Prior for branch lengths in the within-cluster phylogeny**: Assumed to follow a log-normal distribution with equal mean and standard deviation, which implies a coefficient of variation of 1,

(12) **Prior for the transition probabilities along branches in the within-cluster phylogenies**: Represented by an array of 4×4 matrices. Each row of the array corresponds to a different assumed mean branch length, while each column corresponds to a different rate variation category,

(13) **Prior for the transition probabilities along branches in the between-cluster phylogeny**: Same as before,
(14) **Starting value for the cluster membership indices**: Must be a partition of the sample into clades found in the input topology,
(15) **Starting value for the Dirichlet-multinomial concentration parameter**,
(16) **Starting values for the between-cluster and within-cluster transition probabilities**,
(17) **Number of iterations**,
(18) **Burn-in size**,
(19) **Thinning ratio**.

Algorithm output:

(1) Values sampled from the posterior distribution of the cluster membership indices,
(2) Values sampled from the posterior distribution of the concentration parameter,
(3) A non-standardised joint log-posterior probability value for the parameter values at the end of each iteration.

A standard run.

Obtaining the topology. In each simulation run, we start by obtaining an estimate of the maximum likelihood topology from RAxML. We assume that genetic distances follow the GTR+$\Gamma(5)$ model and use a subtype C outgroup (`http://www.hiv.lanl.gov/`, accession number: AB254141). We then produce 500 bootstrap estimates of the tree, resulting in the usual clade support estimates. RAxML stores the best scoring tree in a file with the "bestTree" mention. More details on RAxML's tree optimisation and scoring methods can be found in [112].

Starting values for the cluster membership indices. We then use the topology to obtain initial cluster estimates. More specifically, we look for a partition of the sample into clades for which,

(1) Maximum patristic distance between any pair of elements within a clade is bounded above by an arbitrary value, e.g. 0.05 nt/bp,
(2) Bootstrap support for any clade is above a certain value, e.g. 70%.

We find such a partition by traversing the tree starting at the root. At the beginning, all sequences are assumed to be in one cluster. If the (trivial) clade supported by the root node meets the requirements above, no further move is required. If not, we move down to the two children nodes, and update the cluster membership vector to account for the creation of a new cluster after the split of the original cluster into two non-overlapping clusters. At each child, we repeat the checks performed at the root, moving down and splitting clusters until a set that meets the clustering criteria is encountered, or until we reach a tip.

In the analyses, we impose a confidence requirement of 70%, and find cluster configurations for maximum genetic distance requirements between 0.03 nt/bp and 0.12 nt/bp. For each distance requirement, we have a potentially different set of clusters, and for each of them, we calculate the Dunn index [35], deriving the distance matrix from the phylogenetic estimate. Finally, we pick the set that maximises that index as the starting value for the cluster membership indices.

Estimates of transition probabilities. Once we have an estimate of cluster membership indices, we use it to set up priors for transition probabilities along branches in the within-cluster and between-cluster phylogenies. In the within-cluster phylogenies, branch lengths have an exponential prior. We pick a range of values for the mean parameter by,

(1) Computing the average branch length across all within-cluster phylogenies obtained from the starting partition,
(2) Finding 20 equidistant points in a radius equal to 8% of the value computed previously.

For each point in the range, we simulate $100,000$ values from the corresponding exponential distribution. We then obtain the required transition probability matrices by computing,

$$P^{(r)} = \sum_{i=1}^{1e5} \exp(Qd_i l_r)/1e5, \quad r = 1, 2, 3,$$

where r indexes the rate variation category, d_i denotes a value generated previously, Q, a transition rate matrix estimate, and l_r, a distance scaling factor. We use a similar strategy to derive a prior distribution for transition probabilities along branches in the between-cluster phylogeny.

Running the chain and obtaining point estimates for cluster membership indices. Each iteration in the chain involves successive Metropolis-Hastings updates of the cluster membership indices, the between and within-cluster transition probabilities, and the concentration parameter. The algorithm produces a joint posterior probability value at the end of each iteration, which we use to identify the MAP estimate. To obtain the linkage-xx estimates, we compute an adjacency matrix from each sampled cluster membership vector, under the assumption that all sets of co-clustering sequences form fully-connected graphs, all disjoint from each other. We then average all adjacency matrices, and apply the xx threshold to the resulting matrix, rounding up to 1 all values in the matrix above the threshold, and down to 0 the other values. We then run the walktrap algorithm [91], using chains of 10 steps to detect disjoint sets, which correspond to the cluster membership indices estimate.

Supplementary Material S2 - Tuning parameters used in the simulations

Simulating datasets.

- Sample size: 200,
- Rate parameter for Poisson-distributed number of clusters: 50,
- Mean value for normally-distributed concentration parameter: 10,
- Standard deviation for normally-distributed concentration parameter: 2,
- Number of rate variation categories: 5,
- Shape and scale parameters for gamma-distributed rate variation: 0.7589,
- Number of datasets: 100,
- Root sequence: HXB2 sequence (`http://www.hiv.lanl.gov/`), sites 10-297 of the Protease (PR) region, and 112-741 of the Reverse Transcriptase (RT) region, of the *pol* gene.
- Limiting probabilities: $(A = 0.39, T = 0.22, C = 0.17, G = 0.22)$

- Rate matrix Q:

$$\begin{bmatrix} -0.8371 & 0.0432 & 0.1213 & 0.6726 \\ 0.0766 & -0.8255 & 0.6614 & 0.0876 \\ 0.2782 & 0.8559 & -1.1857 & 0.0516 \\ 1.1924 & 0.0876 & 0.0398 & -1.3198 \end{bmatrix}$$

- Mean parameter for exponentially-distributed branch lengths in within-cluster phylogenies: 0.003,
- Mean and standard deviation parameters for log-normal-distributed branch lengths in between-cluster phylogenies: 0.008.

Chain parameters.

- Number of discrete states for the within-cluster and between-cluster transition probability matrices: 20,
- Number of samples used to obtain transition probability matrices: $100,000$,
- Radius around mean within-cluster and between-cluster branch length estimates: 8%,
- Bootstrap confidence requirement for initial cluster estimate: 70%,
- Limiting probabilities: $(A = 0.4298969, T = 0.2227602, C = 0.1459, G = 0.2014428)$,
- Rate matrix Q:

$$\begin{bmatrix} -0.7963 & 0.0456 & 0.1085 & 0.6422 \\ 0.0880 & -0.7635 & 0.5919 & 0.0836 \\ 0.3198 & 0.9037 & -1.2727 & 0.0492 \\ 1.3705 & 0.0925 & 0.0357 & -1.4986 \end{bmatrix}$$

- Shape parameter for concentration parameter prior: 1000, 100, 10,
- Scale parameter for concentration parameter prior: 0.1,
- Poisson rate for weight applied to the cluster membership vector prior: 50,
- Number of iterations: $55,000$.

Supplementary Material S3 - Tuning parameters used in the real data analysis

Bootstrap analysis.

- Number of discrete states for the within-cluster and between-cluster transition probability matrices: 20,
- Number of samples used to obtain transition probability matrices: $100,000$,
- Radius around mean within-cluster and between-cluster branch length estimates: 8%,
- Discrete gamma distribution parameter: 0.7589,
- Bootstrap confidence requirement for initial cluster estimate: 70%,
- Limiting probabilities: $(A = 0.39, T = 0.22, C = 0.17, G = 0.22)$,
- Rate matrix Q:

$$\begin{bmatrix} -0.8371 & 0.0432 & 0.1213 & 0.6726 \\ 0.0766 & -0.8255 & 0.6614 & 0.0876 \\ 0.2782 & 0.8559 & -1.1857 & 0.0516 \\ 1.1924 & 0.0876 & 0.0398 & -1.3198 \end{bmatrix}$$

- Shape parameter for concentration parameter prior: 1000,
- Scale parameter for concentration parameter prior: 0.1,
- Poisson rate for weight applied to the cluster membership vector prior: 32,
- Number of iterations: $55,000$.

Approximation of the fully Bayesian analysis.

- Number of discrete states for the within-cluster and between-cluster transition probability matrices: 20,
- Number of samples used to obtain transition probability matrices: $100,000$,
- Radius around mean within-cluster and between-cluster branch length estimates: 8%,
- Discrete gamma distribution parameter: 0.4394492,
- Limiting probabilities: $(A = 0.4032,\ T = 0.2148,\ C = 0.1625,\ G = 0.2195)$,

- Rate matrix Q:

$$\begin{bmatrix} -0.8411 & 0.0592 & 0.1122 & 0.6697 \\ 0.1112 & -0.8053 & 0.6214 & 0.0727 \\ 0.2784 & 0.8211 & -1.1718 & 0.0723 \\ 1.2305 & 0.07116 & 0.0535 & -1.3552 \end{bmatrix}$$

- Shape parameter for concentration parameter prior: 1000,
- Scale parameter for concentration parameter prior: 0.1,
- Poisson rate for weight applied to the cluster membership vector prior: 32,
- Number of iterations: $55,000$.

Main run.

- Number of discrete states for the within-cluster and between-cluster transition probability matrices: 20,
- Number of samples used to obtain transition probability matrices: $100,000$,
- Radius around mean within-cluster and between-cluster branch length estimates: 8%,
- Discrete gamma distribution parameter: 0.7589,
- Bootstrap confidence requirement for initial cluster estimate: 70%,
- Limiting probabilities: $(A = 0.39, T = 0.22, C = 0.17, G = 0.22)$,
- Rate matrix Q:

$$\begin{bmatrix} -0.8371 & 0.0431 & 0.1213 & 0.6726 \\ 0.0766 & -0.8255 & 0.6614 & 0.0876 \\ 0.2782 & 0.8559 & -1.1857 & 0.0516 \\ 1.1924 & 0.0876 & 0.0398 & -1.3198 \end{bmatrix}$$

- Shape parameter for concentration parameter prior: 1000,
- Scale parameter for concentration parameter prior: 0.1,
- Poisson rate for weight applied to the cluster membership vector prior: 32,
- Number of iterations: $220,000$.

Supplementary Material S4 - Notes on the software

We implemented DM-PhyClus in R, with C++ modules to handle log-likelihood evaluations. In R, we use classes and functions defined in the *ape* and *phangorn* packages [108] to represent and manipulate phylogenies. The interface between R and C++ relies on features offered by the *Rcpp* and *RcppArmadillo* packages. [38, 37].

The C++ modules make extensive use of containers in the Standard Template Library (STL) and functionalities implemented in the C++11 standard. For now, the code still relies on the GNU Scientific Library (GSL) for random number generation, but we intend to change that in future versions in order to improve portability. Phylogenies are represented by a custom binary tree class, consisting of objects instanced from an input node class, representing the tips of the tree, and from an internal node class. Both classes inherit from an abstract class, standing in for a generic tree node.

We use Felsenstein's tree-pruning algorithm [42] to perform likelihood evaluations. Our implementation of the latter algorithm makes use of containers, functions, and operators defined in the Armadillo library [107]. To reduce the algorithm's memory footprint and improve performance, all intermediate solutions are saved in a map container, and the tree node objects store merely a pointer to the corresponding map elements. To ensure pointer validity, we opted for an ordered map. We use functions in the *boost* package in the generation of keys for map elements. The keys are obtained recursively by combining, among other things, keys computed for children nodes.

The size of the map tends to increase quickly for even moderately-sized datasets, eventually saturating the memory on most standard machines, and so, the software wipes the map periodically. That strategy is also beneficial from a computational standpoint: by eliminating configurations rarely visited by the algorithm, mean lookup time is reduced. Moreover, allowing very large maps is detrimental from a computational standpoint: once a map reaches a certain size, re-computing solutions turns out to be on average faster than doing a lookup.

We obtained a great boost in performance after defining a persistent pointer to the object used to represent the tree structure. Indeed, profiling had revealed that the software was being weighed down considerably by the memory allocation operations involved in building the tree structure, hence the vast improvement resulting from keeping the object in memory and updating it when required. More specifically, we implemented that strategy by passing a so-called *external pointer* to R, implemented by the XPtr class template in the Rcpp library. By trading the pointer between R and C++, we effectively prevent garbage collection of the tree object until the pointer goes out of scope.

We wrote a vignette that explains how the R package can be used to cluster an arbitrary dataset.

Supplementary Material S5 - Log-posterior probability graph

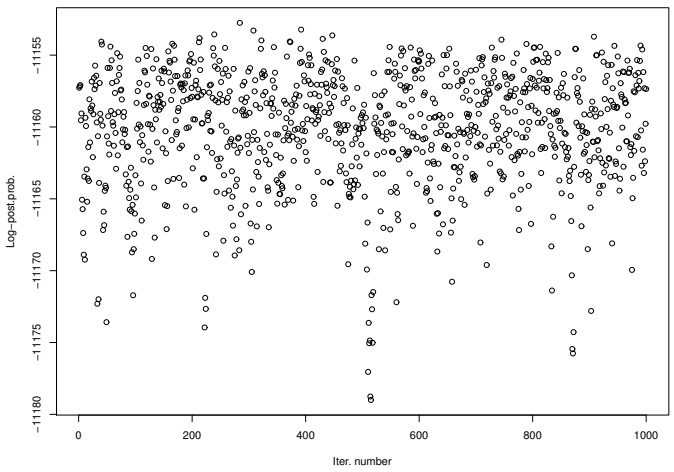

FIGURE 21. **Log-posterior probability graph for the thinned chain obtained from one of the simulated samples.**

6. Bridge between manuscript 2 and manuscript 3

Manuscript 2 described a new phylogenetic method grounded in a straightforward definition of phylogenetic clusters, yielding results more in line with clinical understanding. To demonstrate the capacity of the method to handle sequencing data in practice, we performed a clustering analysis of a real dataset, obtained by sampling potential clusters from the Québec HIV genotyping program database. That analysis, however, was mostly for illustrative purposes. It was short and limited and so, did not provide much insight into current clustering patterns in diagnosed HIV infections. Manuscript 3 tackles that question: it contains a thorough clustering analysis of a large sample of sequences obtained from MSMs in the province of Québec. We use both conventional methods, implemented in popular software solutions, and more recent methods, including DM-PhyClus, to produce updated cluster estimates. Manuscript 3 therefore strengthens the conclusions of manuscript 2, by highlighting the potential of the algorithm to produce reasonable inference for large sequence datasets, representative of those collected for HIV-1 genotyping programs around the world.

CHAPTER 5

Manuscript 3: "Characterizing HIV-1 transmission clusters among men who have sex with men in Quebec, Canada"

1. Preamble

[18] detail the contribution of clustering to the HIV-1 epidemic among MSMs in the province of Québec, Canada. They present estimates for the growth of large clusters in a period ranging from early 2002 to early 2012. Manuscript 3 aims to update and extend several analyses in that paper. Indeed, the Quebec HIV genotyping program database now includes sequences from more than $3,704$ ART-naive MSMs, diagnosed before February 1st, 2016. Up-to-date estimates of clusters in those data, with a focus on their expansion after January 2012, have not yet been published.

In this paper, we compare six competing approaches for the clustering of HIV-1 sequence data. Three of them use an estimate of the maximum likelihood phylogeny, incremented with measures of bootstrap support for its clades. All three involve a minimum bootstrap support requirement for a clade to be called a cluster, but differ with respect to the formulation of the within-cluster genetic distance requirement. One of them imposes a maximum on median [94] patristic distances, while another puts a cap on *maximum* patristic distances instead [19]. The third one ignores the within-cluster phylogenies altogether, by ensuring instead that within-cluster *p-distances* - the usual pairwise distance estimates presented in Section 4 - are bounded above by a sensible value [97].

The fourth method relies on conventional Bayesian phylogenetic inference. From the output of the MCMC algorithm, the so-called *majority-rule consensus tree*, which summarises sampled trees by collapsing into a polytomy any clade missing from more than half of the sampled trees [136], is constructed. The basic majority-rule consensus tree has no branch lengths, which is why we obtain clusters by applying to it the p-distance requirement. Finally,

we partition the sample with the help of the Gap Procedure and DM-PhyClus, introduced in Manuscript 2.

Studies focusing on HIV-1 transmission cluster inference typically opt for one method only, without properly justifying their choice. Manuscript 3 is a worthy addition to the literature because it avoids that pitfall: the use of a large array of clustering algorithms considerably strengthens the credibility of the proposed estimates. Further, we make explicit the reasons behind our selection of cutpoints, which is rarely done in the literature [26].

Manuscript 3 fulfils the last objective of the thesis. Although it does not focus on DM-PhyClus per se, by highlighting how its estimates compare to those resulting from a number of favoured approaches in a real-life analysis, it remains a crucial part of the thesis.

Characterizing HIV-1 transmission clusters among men who have sex with men in Quebec, Canada

Luc Villandré[1] Aurelie Labbe[2¶] Bluma Brenner[3¶] Michel Roger[4,5¶] David A. Stephens[6¶]

1 Department of Epidemiology, Biostatistics, and Occupational Health, McGill University

2 Department of Decision Science, HEC Montréal

3 McGill AIDS Centre, Lady Davis Institute, Jewish General Hospital

4 Centre de Recherche du Centre Hospitalier de l'Université de Montréal (CRCHUM)

5 Département de microbiologie, infectiologie et immunologie, Université de Montréal

6 Department of Mathematics and Statistics, McGill University

Abstract

Background. Several studies have used phylogenetics to investigate HIV-1 transmission among Men who have Sex With Men (MSMs) in the province of Québec, Canada, revealing many transmission clusters. The Québec HIV genotyping program sequence database now includes viral sequences from close to 4,000 HIV-positive individuals classified as MSMs. In this paper, we investigate clustering in those data by comparing results from several methods: the conventional Bayesian and maximum likelihood-bootstrap methods, and two more recent algorithms, DM-PhyClus, a Bayesian algorithm that produces a measure of uncertainty for proposed partitions, and the Gap Procedure, a fast distance-based approach. We estimate cluster growth by focusing on recent cases in the Primary HIV Infection (PHI) stage. **Results.** The analyses reveal considerable overlap between cluster estimates obtained from conventional methods. The Gap Procedure and DM-PhyClus rely on different cluster

definitions and as a result, suggest moderately different partitions. All estimates lead to similar conclusions about cluster expansion: several large clusters have experienced sizeable growth, and a few new transmission clusters are likely emerging. **Conclusions.** The lack of a gold standard measure for clustering quality makes picking a best estimate among those proposed difficult. Work aiming to refine clustering criteria would be required to improve estimates. Nevertheless, the results unanimously stress the role that clusters still play in promoting HIV incidence among MSMs.

2. Introduction

The genotyping of pathogens provide novel opportunities to improve understanding of epidemic dynamics, and as a result, phylogenetic models have become a common tool in the study of infectious disease transmission [34, 70, 46, 64]. Those models have been used extensively to study HIV epidemics [21, 18], mainly due to the availability of large sequence databases, collected mainly in the context of antiretroviral drug resistance testing [117, 1, 62]. The Québec HIV genotyping program database [19] for example, as of 2017, contains 27,487 sequences from HIV-positive individuals, living mostly in Montreal, Québec, Canada.

Men who have Sex With Men (MSMs) remain especially at risk of contracting HIV: in Montreal, prevalence in that risk group could be as high as 13% [95]. Phylogenetic analyses of sequences obtained from MSMs in the Québec HIV genotyping program database have revealed the existence of many large transmission clusters, and highlighted their association with incidence: 51% of newly-infected MSMs belonged to a large transmission cluster in 2011, compared to 25% in 2005 [18]. Highly Active Antiretroviral Therapy (HAART) has been successful in substantially suppressing viremia within the diagnosed population, making late transmission of the virus a lot less common, and consequently, early transmission has been increasingly driving the epidemic. Recently-infected individuals are much more likely to transmit because of their high viral load, potentially leading to *quick transmission chains*, consecutive transmission events happening in a short time span. Quick transmission chains tend to manifest themselves as clusters in the sequence data, and increased clustering may therefore point to a higher proportion of early transmissions. Quantifying the role of early

transmission in the epidemic is important from a public health standpoint, as it can help assess the extent to which programs are able to reach infected individuals early enough. This is the motivation behind the current study, in which we analyse a large sample of sequences collected via the Québec HIV genotyping program with a variety of clustering methods, comparing their estimates to shed light on the recent evolution of clustering in the epidemic.

2.1. Background. Phylogenetic studies of clustering in HIV-1 epidemics tend to rely on a number of *ad hoc* rules applied a posteriori to phylogenetic estimates. Availability of software like MEGA and PAUP* [121, 118] has led to widespread adoption of maximum likelihood phylogenetic reconstruction, coupled with the bootstrap to evaluate confidence in the inferred clades. In that context, cluster estimation relies on an arbitrary cutoff applied to bootstrap support estimates, usually between 70% and 95% [58, 60, 19]. Alternatively, software like BEAST and MrBayes [33, 105] have popularised Bayesian phylogenetic estimation. Both are based on versions of the Markov Chain Monte Carlo (MCMC) algorithm, that numerically approximate posterior distributions for a variety of evolutionary and phylogenetic parameters. They also provide posterior probability support for clades, a crucial measure for the identification of clusters, which in phylogenetic terms correspond to non-nested clades forming a partition of the sample. For example, many studies require posterior probability support of 1 to conclude in clustering [136].

In addition to clade confidence requirements, many studies also impose a within-cluster genetic distance requirement, usually between 0.01 nt/bp and 0.05 nt/bp [97]. Distance requirements may be applicable to mean [62], median [94], or maximum *patristic distances* [19], that is, distances calculated by summing branch lengths along the shortest path between any two tips in the phylogeny. The *ClusterPicker* algorithm [97] instead formulates that requirement in terms of maximum within-cluster p-distances, e.g. the Hamming distance.

Cutoffs are however hard to justify rigorously [26] and so, methods grounded in more explicit definitions of clusters have been published. [130] proposed the so-called *Gap Procedure*, a fast pure distance-based approach that requires minimal tuning. In a similar vein, [127] formulated DM-PhyClus, a Bayesian algorithm that aims to minimise reliance on thresholds while still offering rigorous inference for cluster membership.

The heavy computational burden of conventional phylogenetic inference is problematic in light of the fast increase in the size of sequence databases, and can therefore limit its applicability [131]. Thankfully, software designed to handle larger datasets is now available. RAxML [111] and FastTree [93], for example, make use of heuristics in phylogenetic optimisation to improve scalability of the maximum likelihood phylogenetic methods. Clustering of large datasets in a purely Bayesian paradigm is a computational challenge that has not yet been fully overcome, although vast progress has been made thanks in part to GPU computing [33, 105].

2.2. Objectives. This paper aims to provide up-to-date estimates of transmission clusters in the HIV epidemic among MSMs in the province of Québec and assess their temporal expansion. In doing so, we seek to improve previous assessments of the contribution of early transmission and quick transmission chains to the epidemic. We therefore perform an exhaustive clustering analysis of HIV-1 subtype B sequences in the most recent version of the Québec HIV genotyping database originating from MSM subjects. There is a lack of consensus as to how clustering of HIV-1 sequence data should be done, and different methods may produce equally valid, but conflicting results [21]. To assess sensitivity of cluster estimates to phylogenetic assumptions and cluster definitions, we compare results from a number of methods,

(1) Maximum likelihood phylogenetic reconstruction, coupled with a bootstrap support requirement for clades, which we refer to as the *ML-bootstrap* approach,
(2) Bayesian phylogenetic inference, coupled with a posterior probability support requirement for clades,
(3) DM-PhyClus [127],
(4) Gap Procedure [130].

3. Materials and methods

3.1. Data. The Québec HIV genotyping program database comprises $27,487$ sequences collected in the province of Québec. They cover sites 10-297 of the protease region (PR),

and 112-741 of the reverse transcriptase (RT) region, of the *pol* gene, for a total of 918 loci. Each of them comes with a time stamp, indicating when the blood sample was collected, and an indicator of infection status, either chronic treated, chronic untreated, or Primary HIV Infection (PHI). A case is considered a PHI if the sequence was obtained fewer than 6 months after seroconversion [19].

3.2. Methods.

3.2.1. *Conventional maximum likelihood.* We obtain the maximum likelihood (ML) phylogenetic estimate [42] with RAxML 8.2.10 [111], under the assumption that nucleotide evolution follows the GTR + I + $\Gamma(5)$ model. We produce $1,000$ bootstrap trees, and use them to evaluate confidence in clades present in the "best scoring" phylogeny, RAxML's estimate of the ML phylogeny. To conclude in clustering, we require, in turns, bootstrap support greater than 70%, 90%, or 95% and consider genetic distance requirements of 0.015 nt/bp, 0.03 nt/bp, 0.045 nt/bp, 0.068 nt/bp, and 0.077 nt/bp. More specifically, we apply, in turns, the maximum within-cluster Hamming distance requirement of ClusterPicker [97], the maximum median within-cluster patristic distance requirement of PhyloPart [94], and the maximum within-cluster patristic distance requirement of [19]. For the PhyloPart analysis, [94] recommend setting cutpoints based on percentiles of the total tree patristic distance distribution. In their real data analysis, they use the 15th and 30th percentiles as cutpoints, which we also try.

3.2.2. *Conventional Bayes.* We perform phylogenetic inference with MrBayes 3.2.6 [105] using default parameters, under the assumption that nucleotide evolution follows the GTR + I + $\Gamma(4)$ model. MrBayes uses the MCMC algorithm [56], more specifically the so-called *Metropolis-coupled MCMC*, or $(MC)^3$ [49], algorithm, to generate estimates for the posterior distribution of phylogenetic parameters. The MCMC algorithm lets us recursively obtain samples from the posterior distributions of interest. It starts off by setting all parameters at an arbitrary value. Then, in each iteration, updates to parameter values are proposed, conditional on their current values. Each proposal is randomly accepted with probability equal to the Metropolis-Hastings (MH) ratio, producing a move in the parameter space; else, no move is recorded. After a large number of iterations, parameter values generated

throughout the chain are used to empirically estimate the posteriors. We run three million iterations, burning in the first 50% and sampling one iteration out of 500. We derive the majority rule consensus tree from the remaining 3,000 trees, and produce cluster estimates by identifying clades with posterior probability support of 1.0. Once again, we use the ClusterPicker algorithm to obtain cluster estimates, under the requirement that within-cluster distance be, in turns, bounded above by 0.015 nt/bp, 0.03 nt/bp, and 0.045 nt/bp, 0.068 nt/bp, and 0.077 nt/bp.

3.2.3. *Cutpoint selection.* All conventional clustering approaches require selection of genetic distance and confidence cutpoints. Prior to the analyses, researchers involved directly in the Québec HIV genotyping program performed a preliminary clustering of the dataset, and identified from the results seven noteworthy sets of sequences, comprising 372 sequences in total, that they expect correspond with genuine transmission clusters. One of those sets, for instance, comprises 68 sequences and is characterised by more than half of its members harbouring the Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTI) mutation K103N. As in [94], we use that subsample as a reference set. We compare partitions obtained across a range of cutpoints with that set using the Adjusted Rand Index (ARI), a measure of similarity between two partitions, with the aim of maximising overlap. Greater ARI values are better, and the measure is bounded above by 1, indicating perfect correspondence. We describe the comparison scheme in more details in Supplementary Material S1.

3.2.4. *DM-PhyClus.* DM-PhyClus is a Bayesian phylogenetic algorithm that aims to estimate transmission clusters directly, by identifying sets of sequences supported by distinctive subtrees, thus avoiding the need to specify thresholds arbitrarily [127]. Unlike conventional methods, as a way to directly find sets of sequences resulting from quick transmission chains, it defines a cluster as a clade supported by a phylogeny with a distinctive branch length distribution, usually with a relatively small mean. Conditional on an input phylogeny – the maximum likelihood estimate in this study – it uses the MCMC approach to produce an estimate of the posterior distribution of cluster membership indices. As a result, it has the added benefit of providing a straightforward measure of uncertainty for cluster membership estimates, in the form of co-clustering frequencies across the chain. It requires specification of a number of other priors and evolutionary parameters, which we list in Supplementary

Material S3. We perform 220,000 iterations, discarding the first 20,000 as a burn-in and applying a thinning ratio of 1 over 200, leaving us with a sample of size 1,000. We identify the partition that maximises the joint posterior probability score, which we refer to as the Maximum Posterior probability (MAP) estimate.

3.2.5. *Gap Procedure.* The Gap Procedure is a pure distance-based clustering algorithm that requires minimal tuning, and avoids reliance on ad hoc cutpoints by partitioning sets of sequences into distinctive components without requiring phylogenetic estimation [130]. When the true clusters are compact and separable enough, the Gap Procedure can propose partitions that largely agree with conventional phylogenetic estimates, but in a fraction of the time normally required for such analyses, thus making the method ideal for handling large datasets. For example, in an analysis presented in [130], partitioning a dataset comprising 627 sequences of length 810 took 126 hours with MrBayes and less than a second with the Gap Procedure. The method takes as input a matrix of pairwise distances, which we obtain under the K80 model. We leave tuning parameters at their default values.

3.2.6. *Cluster growth evaluation.* To evaluate cluster growth properly, we would need to know seroconversion dates for all cases whose sequences were sampled. The dataset however contains instead an infection stage indicator, equivalent to a censored estimate of infection time, i.e. smaller (greater or equal) than six months prior to the sampling date for PHIs (chronic cases). Most HIV-positive individuals are diagnosed while already in the chronic stage, at which point seroconversion date estimates are very imprecise [72]. As a result, we use PHIs only to obtain a lower bound estimate for the growth of inferred clusters, since PHIs can be reliably associated with a short time window prior to sampling. We focus on a "recent" period, ranging from January 1, 2012 to February 1, 2016, during which 957 cases were added to the database, including 304 PHIs. For example, one of the methods may propose a cluster of size 20, with eight of its sequences having been obtained from cases diagnosed while in the PHI stage at some point in 2014. We can therefore be certain that those cases were infected after January 1, 2012 and so, we conclude that the cluster has accrued at least eight new cases in the selected period.

3.2.7. *Software.* Under the conventional methods, we get cluster estimates by importing phylogenetic estimates from RAxML or MrBayes into R v3.2.3 and analysing them with

functions in the *phangorn* and *ape* libraries [108]. We use functions in the *GapProcedure* and *DMphyClus* R libraries to obtain the other estimates.

4. Results

4.1. Sample characteristics. We retain 3,936 subtype B sequences, each obtained from a different individual who self-identified as MSM. Since the analyses focus on transmission clusters among MSMs only, we exclude 20 sequences obtained from women, leaving us with 3,916 sequences. To avoid potential artefacts resulting from selective pressure induced by antiretroviral therapy, we remove sequences for chronic treated patients and patients with missing infection status information as well, leaving us with 3,704 sequences. Of those, 1,402 are from PHIs, and 2,302 are from chronic untreated cases. The earliest sequence was collected on May 3rd 1996, and the latest, on January 12, 2016. Only 69 sequences were collected prior to 2002. The number of sequences added yearly to the sample follows an upward trend until 2008, with 359 sequences added that year, and then steadily decreases until it reaches 217 in 2015. The sample includes only one sequence collected in 2016. Finally, for rooting purposes, we add to the sample three subtype A outgroups from Zambia [1] (NCBI accession numbers AB254141, AB254142, AB254143).

4.2. Cutpoint selection. In all maximum likelihood analyses, the bootstrap support requirement of 70% resulted in greater overlap with the reference set. Under the maximum patristic distance scheme of [19], we found that a distance requirement of 0.077 nt/bp maximised the correspondence (ARI = 0.91). With ClusterPicker, requirements of either 0.068 nt/bp or 0.077 nt/bp were preferable (ARI = 0.91). In PhyloPart, a median within-cluster patristic distance requirement of 0.03 nt/bp resulted in the largest overlap with the reference (ARI = 0.98). Finally, in the Bayesian analysis, in addition to a posterior probability requirement of 1, we determined that a 0.068 nt/bp or 0.077 nt/bp requirement for maximum within-cluster Hamming distances were equivalent (ARI = 0.91). Except for [94], published clustering analyses tend to rely on more restrictive distance requirements and so, in cases where several distance requirements were equivalent, we picked the smallest one.

	ML + CP	ML + PhyloPart	ML + Max. pat. dist.	MrBayes + CP	Gap Procedure	DM-PhyClus
ML + CP	1.00	0.92	0.93	0.94	0.83	0.65
ML + PhyloPart	0.92	1.00	0.91	0.86	0.88	0.68
ML + Max. pat. dist.	0.93	0.91	1.00	0.88	0.83	0.66
MrBayes + CP	0.94	0.86	0.88	1.00	0.84	0.64
Gap Procedure	0.83	0.88	0.83	0.84	1.00	0.72
DM-PhyClus	0.65	0.68	0.66	0.64	0.72	1.00

TABLE 3. **Adjusted Rand index for the overlap between the cluster estimates obtained from the different methods.** CP stands for ClusterPicker.

It is no surprise that the proposed schemes resulted in similar choices of cutpoints. Cluster estimation based on the consensus tree computed from the Bayesian tree search relies on ClusterPicker, just like one of the ML-bootstrap approaches. Normally, clusters found through a Bayesian analysis agree substantially with those obtained through a ML-bootstrap approach. Also, ClusterPicker uses maximum within-cluster pairwise distances, which provide a rough approximation of patristic distances. It follows that tuning ClusterPicker to produce estimates in line with those from the method of [19] should be feasible.

4.3. Estimates comparison. We first compare optimal estimates from all methods with the ARI, cf. Table 3. We observed the largest overlap between the partitions resulting from the *ML + maximum patristic distance* and *ML+ClusterPicker* methods (ARI = 0.94). On the other hand, we obtained the smallest overlap between estimates suggested by the MrBayes+CP and DM-PhyClus methods (ARI = 0.64). DM-PhyClus produced the most distinctive set of clusters, with overlap with clusters from the other methods ranging from 0.64 and 0.72. The larger correspondence with the Gap Procedure estimate is not surprising, since both methods define clusters in terms of their separation from other clusters.

We represent graphically the correspondence between the different estimates in Figure 22. The heat map, showing the $2,938$ sequences found to co-cluster with at least one other sequence by at least one of the methods, reveals 11 moderately-sized clusters. The largest rectangle, marked "1" in the figure, matches roughly one of the reference clusters, and is of size ≈ 125. The earliest sequence in the cluster was collected on August 13, 2002 and was a PHI, and the latest sequence, also corresponding to a PHI, was obtained on December

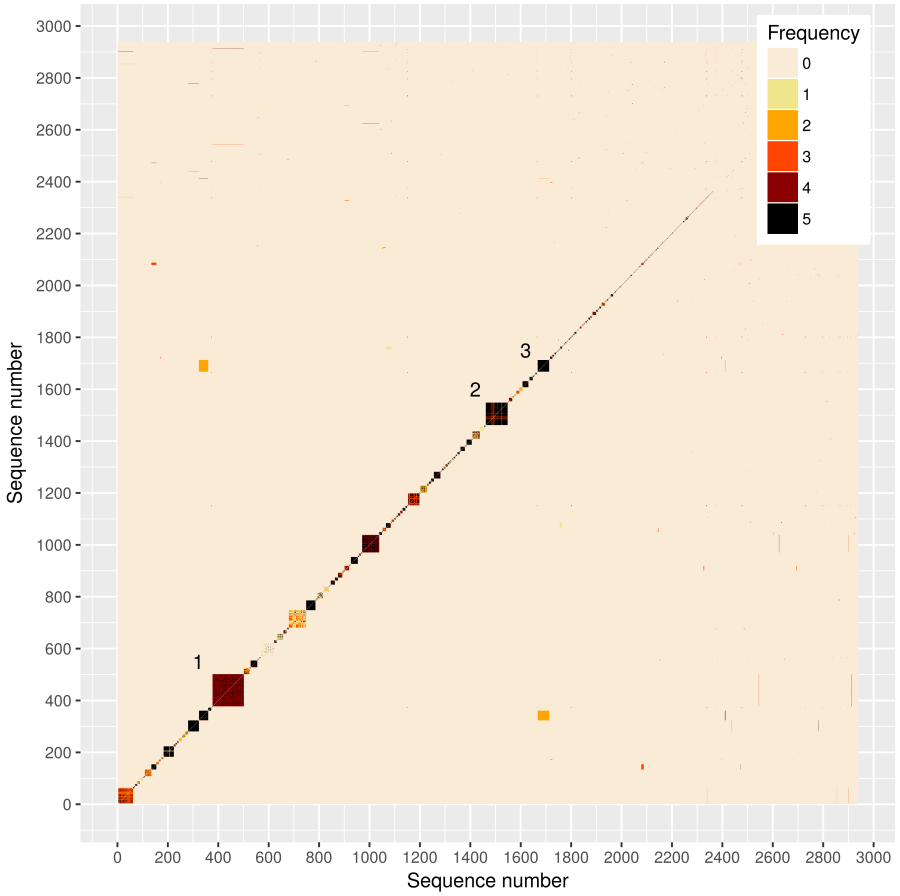

FIGURE 22. **Heat map showing the frequency at which sequences co-clustered across methods.** The x and y axis represent the $2,938$ sequences that were found to be non-singletons by at least one of the methods.

23, 2015. The MAP estimate of DM-PhyClus, on the other hand, split this cluster into 14 components, including three clusters of size 37, 36, and 14, respectively, and 7 singletons. Instead of the MAP estimate, we could have derived the so-called *linkage estimate* from the chain results [127]. Broadly speaking, the linkage estimate proposes clusters by partitioning the sample into subsets of sequences that co-cluster often across iterations in the chain. We present a more detailed description in Supplementary Material S4. The linkage estimate ends up more in line with the other estimates: it contains 12 components, with 3 large clusters of sizes 37, 36, and 30, and 7 singletons.

The second largest cluster, represented by the mostly black block on the right, marked "2" in the figure, comprises 87 sequences, and is also part of the reference set. All methods

	ML + ClusterPicker	ML + PhyloPart	ML + Max. pat. dist.	MrBayes + ClusterPicker	Gap Procedure	DM-PhyClus
Mean clus. size	2.29	2.08	2.19	2.48	2.33	2.11
Mean (no singletons)	6.01	5.53	5.62	5.96	4.80	5.62
Median (no singletons)	3.00	3.00	3.00	3.00	2.00	3.00
Max. clus. size	126	126	126	126	125	77
Num. singletons	1205	1353	1261	1051	1035	1330
Num. clus. size ≥ 2	1621	1779	1696	1497	1592	1753

TABLE 4. **Summary statistics for estimates returned by the different methods.**

agree more or less that it indeed represents a transmission cluster. It comprises sequences sampled from May 11, 2004 (chronic untreated) to December 14, 2015 (PHI), which highlights its durability. The moderately-sized black block to its immediate right, marked "3", also stands out. Its 45 sequences, also found in the reference set, co-cluster according to all the methods. Its first sequence was collected on January 11, 2012 and its last, on April 8, 2015. Two methods, *MrBayes + ClusterPicker* and *ML + ClusterPicker* added to that cluster an extra 38 sequences, as evidenced by the light orange rectangle underneath it.

Figure 23 presents truncated cluster size distributions derived from the preferred estimate from each method and Table 4 gives related summary statistics. Unsurprisingly, distributions obtained from the four conventional methods are very similar. Among those, the one for the conventional Bayesian estimate, labelled "MrBayes + ClusterPicker", stands out because of its thicker right tail. The distribution derived from the DM-PhyClus estimate is also distinctive, because of its much thinner right tail. Frequencies for singletons are not shown in the graphs for readability purposes. We found that ML + PhyloPart and DM-PhyClus had the highest proportions of singletons, each having approximately 36% of size 1 clusters. On the other hand, the Gap Procedure and the conventional Bayesian estimate had the fewest singletons, with 28% of clusters having a single member. The Gap Procedure estimate, however, had much more transmission pairs than the other methods.

4.4. Cluster growth assessment. A total of 957 cases in the MSM risk group were added to the database in the selected time period, including 304 PHIs. Of those PHIs, 254 were sampled after June 30, guaranteeing that the corresponding transmission events took place in 2012. According to the *ML + PhyloPart* estimate, 50 (20%) of those 254 PHIs are singletons, 23 (9%) are found in transmission pairs, and 153 (60%) belong to clusters of size

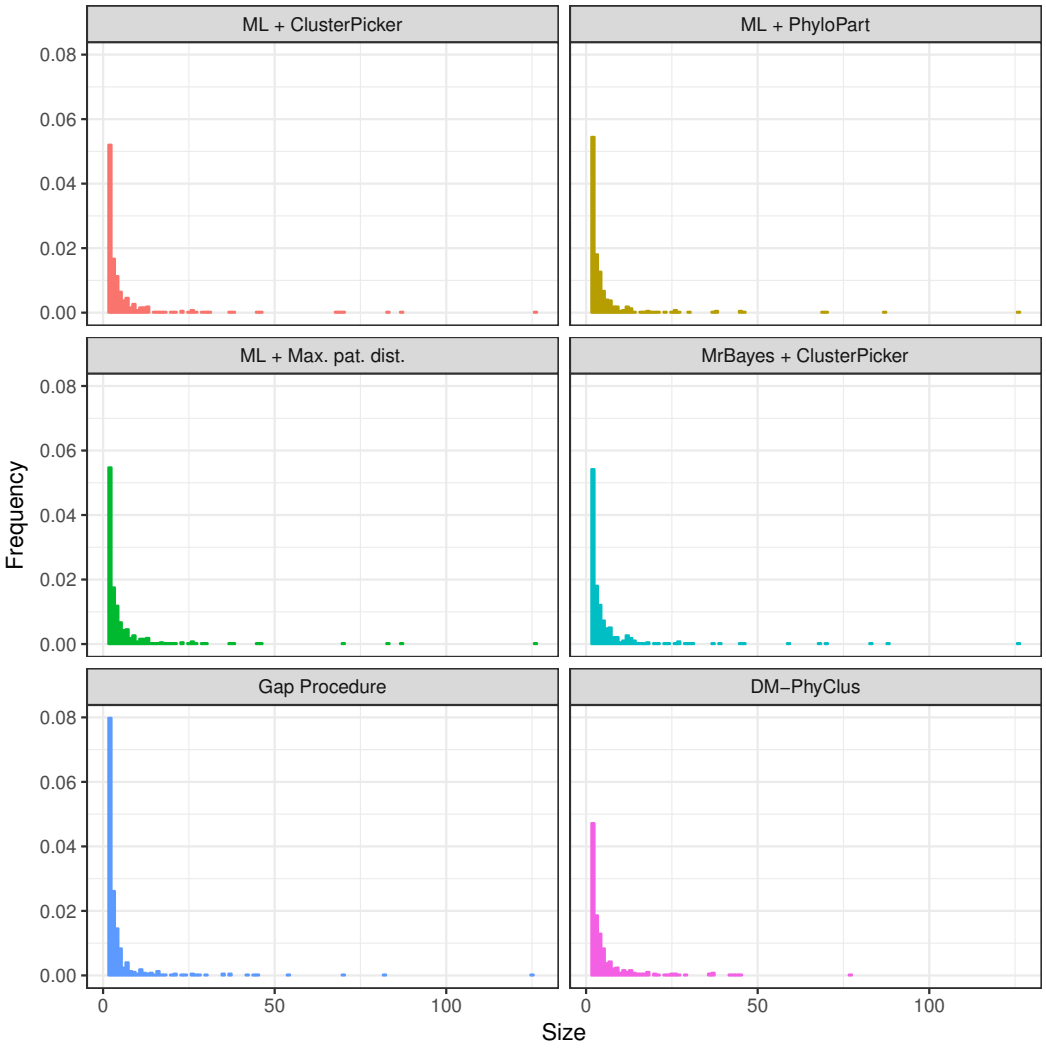

FIGURE 23. **Truncated cluster size distributions for the preferred estimate across methods.** We refer to the figure in section 4.3, that focuses on summary statistics for the obtained cluster estimates, in order to highlight their differences. To improve readability, we removed the bars corresponding to singletons.

five or more. In comparison, in the period ranging from July 1st 2008 to January 1st 2012, 319 MSM cases diagnosed in the PHI stage were added to the database. After excluding all sequences sampled after January 1st 2012, we find that of those PHIs, 83 (26%) form singletons, 34 (11%) belong to transmission pairs, and 159 (50%) are part of clusters of size five or more. If we do not exclude the more recent sequences, we find that 79 of the 319

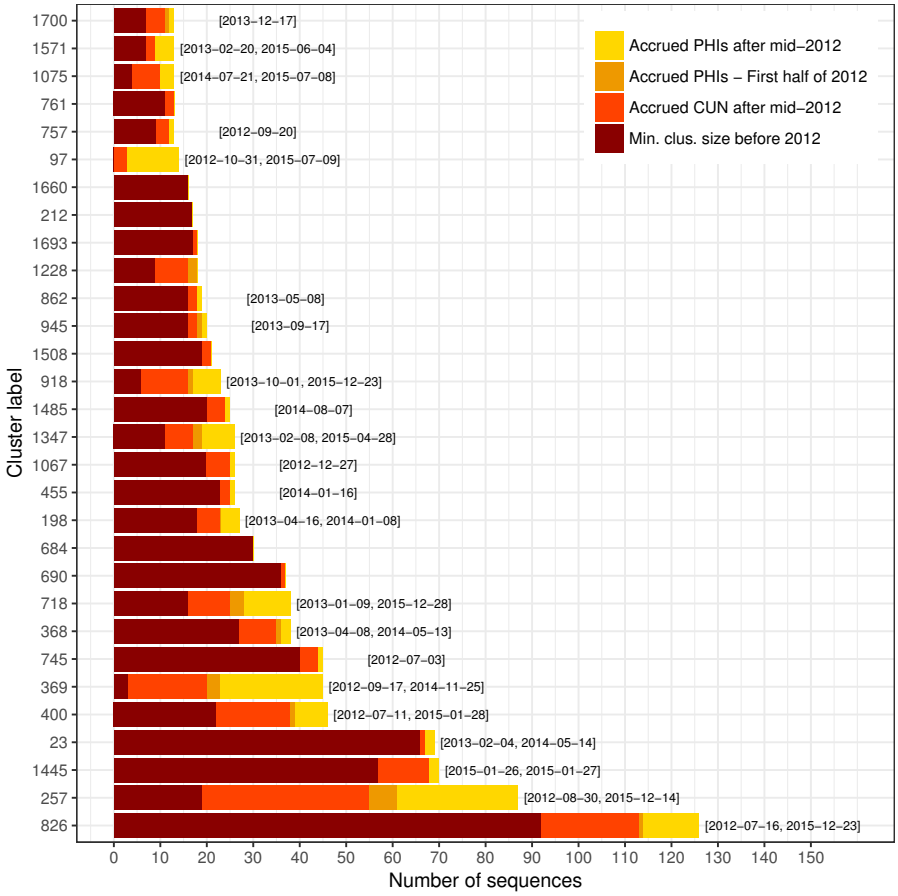

FIGURE 24. **Bar plot showing the breakdown in membership for the 30 largest clusters in the ML + PhyloPart estimate.** The labels at the end of each bar indicate the sequence collection dates for the first and last "recent" PHIs in the cluster, that is, recorded on or after July 1st, 2012. When there is only one such PHI, we display the corresponding collection date instead. We assume that all chronic cases recorded before July 1st, 2012 were infected prior to 2012. The dark red bar represents the "minimum cluster size before 2012" because several chronic cases diagnosed after July 1st 2012 were probably infected prior to 2012. Also, it is likely that several PHIs sampled in the first half of 2012 match with transmission events that occurred late in 2011.

cases (25%) are still singletons, which tends to indicate that the more recent PHIs tend to cluster more.

We represent the 30 largest clusters, according to the *ML + PhyloPart* estimate, in Figure 24. Those clusters include 126 recent PHIs, split between 22 clusters. Among the largest ten clusters, nine include at least one recent PHI. The largest cluster includes 12 recent PHIs, while the second and third include 26 and two, respectively. Cluster 369 is noteworthy: despite its small size prior to 2012, it has grown quickly, with the addition of 22 recent PHIs. Cluster 97, on the other hand, is still small, but has not been recorded before. Each of those two clusters has a PHI recorded as late as the second half of 2015, indicating that they may still be expanding. Other conventional estimates and the Gap Procedure lead to similar conclusions, as can be seen in Supplementary Material S5.

The partition produced by DM-PhyClus is different, but leads to similar conclusions, as shown in Figure 25. Of the 30 largest clusters, 20 include at least one recent PHI. The largest cluster overlaps largely with cluster 257 in Figure 24 and includes 26 recent cases, while the second and third largest include 22 and 1, respectively. The fifth largest cluster includes 10 recent PHIs, out of 42 members, also hinting at considerable growth.

5. Discussion

5.1. Summary. In this paper, we investigated clustering in a sample of 3,704 HIV-1 cases belonging to the men who have sex with men risk category. We compared estimates from six methods, four conventional approaches relying on a variety of cutpoints applied to phylogenetic estimates, and two additional recent approaches seeking to avoid cutpoints entirely, the Gap Procedure and DM-PhyClus. We found that estimates obtained from conventional methods were overall fairly similar. The estimate from DM-PhyClus involved a noticeably different, albeit not unreasonable, cluster size distribution. Unlike other methods however, DM-PhyClus provides a straightforward measure of co-clustering frequencies and so, we found that requiring a certain degree of co-clustering, through the *linkage-xx* estimate, could change estimates for certain clusters. All estimates however produced a similar assessment of cluster growth in the period ranging from January 1st, 2012 to February 1st, 2016: nine of the ten largest clusters had grown in the selected period, three of those having accrued at least ten new cases. Further, we observed several emerging clusters.

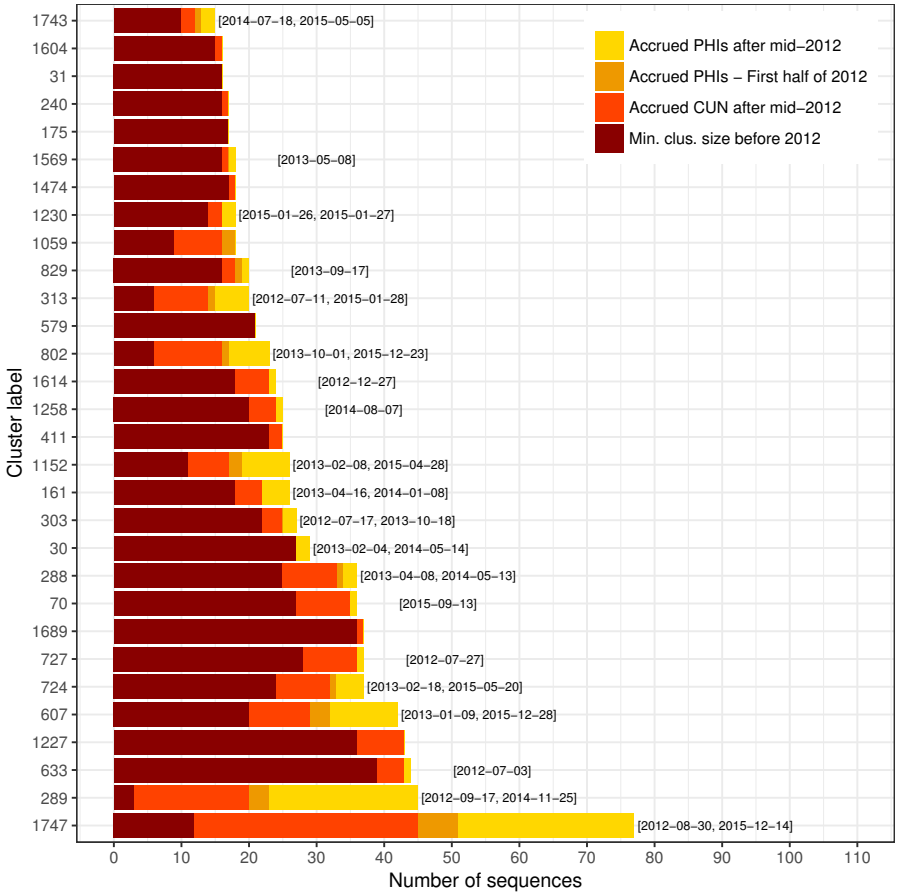

FIGURE 25. **Bar plot showing the breakdown in membership for the 30 largest clusters in the DM-PhyClus estimate.** The labels at the end of each bar indicate the sequence collection dates for the first and last "recent" PHIs in the cluster, that is, recorded on or after July 1st, 2012. When there is only one such PHI, we display the corresponding collection date instead. We assume that all chronic cases recorded before July 1st, 2012 were infected prior to 2012. The dark red bar represents the "minimum cluster size before 2012" because several chronic cases diagnosed after July 1st 2012 were probably infected prior to 2012. Also, it is likely that several PHIs sampled in the first half of 2012 match with transmission events that occurred late in 2011.

5.2. Limitations. The study has several limitations. Cutpoint selection remains inherently subjective. Indeed, choosing cutpoints as to maximise overlap with a reference set does not guarantee that other clusters will be estimated well. Moreover, identifying a suitable reference set can be difficult. In our study, researchers involved directly in the Québec HIV

genotyping program proposed the set based on a curated clustering analysis they conducted. A different reference set might have led to different cutpoints. Several of the approaches we used did manage to recover the reference set very closely though, which suggests that it is not unrealistic.

DM-PhyClus, being a Bayesian method, rests on a number of prior assumptions, which are all more or less informative, and it follows that prior calibration is key. [127] suggest that estimates are reasonably robust to some prior assumptions, but it remains possible that a combination of very poorly chosen priors may result in misleading cluster estimates.

Reliable infection date estimates for cases diagnosed while in the chronic stage are unavailable and so, we could only obtain a lower bound for cluster growth between January 1st, 2012 and February 1st, 2016. The average time between seroconversion and diagnostic is between 2 and 3 years [126], and it follows that several chronic cases diagnosed after June 30th might have been infected during the selected period. Estimating infection time from the fraction of ambiguous nucleotides in each sequence would have been possible [72], but the high standard deviation for such predictions would have limited their usefulness.

Because of the non-random sampling of cases, we cannot readily deduce from our estimates the population-level cluster size distribution. In the absence of covariate information, we cannot model the sampling process. If, for example, the probability for a case to be sampled correlates positively with cluster size, we might end up underestimating the size of smaller clusters and the number of singletons, and consequently, overestimating the contribution of clustering to the epidemic. Nevertheless, the results we presented provide good evidence of cluster growth, and that phenomenon alone warrants attention.

5.3. Selecting a best transmission cluster estimate. Determining which partition among the six proposed provides the most accurate representation of transmission clusters in the sample is difficult. The choice depends ultimately on our confidence in the assumptions of each approach, and on substantive knowledge. The agreement between estimates from the conventional approaches, although explained in great part by shared assumptions, is still a good sign. The moderately different partitions proposed by the Gap Procedure and DM-PhyClus are not erroneous: they result from the way the two methods define clusters. The

two approaches also have additional aims and benefits. [130] designed the Gap Procedure with scalability in mind, and [127] formulated DM-PhyClus in such a way that it could offer a straightforward measure of uncertainty around the returned clusters.

5.4. Conclusion. The existence of large transmission clusters is not only a feature of transmission of HIV-1 among MSMs in the province of Québec: it has been observed across Europe and other regions of North America as well [75, 78, 13, 11]. The increasing size of sequence databases represents a considerable computational challenge, especially in the Bayesian framework, and so, scalability should be an essential feature of future clustering algorithms [47]. We contend that methods that avoid cutpoint selection altogether are convenient and promising, and would benefit from further improvements. In addition to lightening their computational burden, adapting them to use time-stamp and covariate data, for example, would be a welcome extension. Further, methods designed to provide a clear measure of uncertainty for estimated partitions, like DM-PhyClus, would warrant more attention. Indeed, the strength of co-clustering between sequences within an inferred cluster may vary sizeably, and the separation between neighbouring clusters may not be very clear-cut. Such variability may be hard to measure rigorously under conventional phylogenetic clustering approaches.

Phylogenetic surveillance of HIV transmission among MSMs provides helpful clues for explaining the persistence of the epidemic. The portrait of clustering presented in this study suggests an ongoing contribution of quick transmission chains to incidence, a finding that should inform public health strategies to reduce transmission rates.

Ethics approval and consent to participate

The study has received ethics approval from the McGill Faculty of Medicine Institutional Review Board. Ethics approval for the Quebec HIV genotyping program was obtained from individual study sites, the Laboratoire de santé publique du Québec, and the Quebec Ministry of Health committee on confidentiality and access of information.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by a training award from the Fonds de recherche du Québec-Santé (FRQS), funding from the Centre de Recherches Mathématiques (CRM), a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant, and a Canadian Institutes of Health Research (CIHR) grant (CIHR HHP-126781).

Author's contributions

LV wrote the article, performed the analyses. LV, AL, and DAS jointly formulated the study plan. AL and DAS suggested and reviewed analyses. BB and MR provided the HIV-1 sequences.

Data availability

The Québec HIV genotyping program sequences cannot be made publicly available for confidentiality reasons. A small subset of sequences can be provided for verification purposes upon request.

Acknowledgements

The Quebec HIV genotyping program is sponsored by the Ministère de la Santé et des Services sociaux (MSSS) du Québec and by the Fonds de recherche du Québec (FRQ-S) Réseau SIDA/MI.

Supplementary Material S1: Cutpoint selection with a partial gold standard

The reference set includes only a small fraction of sequences in the dataset, and acts therefore as a *partial gold standard*. We select cutpoints for each method as to maximise overlap with that reference set. The lack of a reference solution for other sequences in the sample makes comparison with this standard non-straightforward. Let us assume we have a sample of size 10, and that sequences 1-3 and 4-6 form two confirmed clusters, labelled 1 and 2, respectively. A representation for cluster membership in the full gold standard would be [1, 1, 1, 2, 2, 2, Not 1 or 2, Not 1 or 2, Not 1 or 2, Not 1 or 2]. To best quantify overlap with the full gold standard, in all partitions we test, all sequences that do not co-cluster with any element in the reference set are given a membership index equal to (Number of clusters found among sequences in the reference set + 1). The full gold standard is reformulated in such a way that all sequences outside the reference set are given index (Number of true clusters in the reference set + 1). In the example, the gold standard would be reformulated [1, 1, 1, 2, 2, 2, 3, 3, 3, 3]. Let's say a clustering algorithm returns configuration [1, 1, 2, 3, 3, 3, 3, 4, 4, 5]. To obtain the correct ARI, we would need to transform it into [1, 1, 2, 3, 3, 3, 3, 4, 4, 4].

Supplementary Material S2: MrBayes script

begin mrbayes; set autoclose=yes nowarn=yes; execute brennerComplete-Data.nex; lset nst=6 rates=invgamma; outgroup AB254141; set beaglescaling=dynamic beaglesse=yes; mcmc nruns=2 nchains=4 ngen=3000000 samplefreq=500 diagnfreq=10000 printfreq=500 append=yes; sump relburnin=yes burninfrac=0.25; end;

Supplementary Material S3: Tuning parameters used in the DM-PhyClus and Gap Procedure analyses

DM-PhyClus.

- Number of discrete states for the within-cluster and between-cluster transition probability matrices: 20,
- Number of samples used to obtain transition probability matrices: $100,000$,
- Radius around mean within-cluster and between-cluster branch length estimates: 25%,
- Discrete gamma distribution parameter: 1,
- Bootstrap and distance requirements for initial cluster estimate: 90%, 0.045,
- Limiting probabilities: $(A = 0.38, T = 0.24, C = 0.16, G = 0.21)$,
- Rate matrix Q:

$$\begin{bmatrix} -0.8891 & 0.0659 & 0.1324 & 0.6908 \\ 0.1047 & -0.7205 & 0.5477 & 0.0681 \\ 0.3096 & 0.8069 & -1.1801 & 0.0636 \\ 1.2540 & 0.0779 & 0.0494 & -1.3812 \end{bmatrix}$$

- Shape parameter for concentration parameter prior: 500,
- Scale parameter for concentration parameter prior: 0.2,
- Poisson rate for weight applied to the cluster membership vector prior: 2368,
- Number of iterations: $220,000$.

Gap Procedure.

- Threshold for largest gap search: 90%.

Supplementary Material S4: The linkage estimate

We obtain the linkage estimate by first projecting each cluster membership vector produced by DM-PhyClus as an unweighted undirected network graph, where each sequence is represented by a vertex, and an edge between any two vertices implying co-clustering between the corresponding sequences. For example, cluster membership vector [1, 1, 1, 2, 2, 2] would translate as a graph with six vertices, split between two disjoint components, each of those being a *fully-connected* graph. In other words, all vertices within each component

are inter-connected. We can express an unweighted undirected network graph with an *adjacency matrix*, a symmetric matrix with as many rows and columns as vertices, with a 1 (0) at position (i, j) indicating a connection (no connection) between vertices i and j. Elements on the diagonal are set to 0.

Once we have adjacency matrices for all cluster membership states visited by the chain, we average all the matrices element-wise, resulting in an adjacency matrix for a *weighted* undirected network. Values in that matrix, all between 0 and 1, indicate the strength of the association between any two sequences. We then run the *walktrap algorithm* on the corresponding graph to identify *communities* [91]. Communities are sets of vertices that are a lot more interconnected than would be expected from chance alone. The walktrap algorithm works by performing a large number of short random walks on the graph. It starts at a random vertex, and jumps to neighbouring vertices a fixed number of times. It is based on the principle that a short random walk starting in a community is more likely to end up in the same community, because of the high degree of interconnectedness between its vertices. The algorithm then outputs an estimate of community structure in the form of a vector of arbitrary community labels, which corresponds to the desired linkage estimate.

Supplementary Material S5: Additional bar plots depicting cluster growth between January 1st, 2012 and February 1st, 2016

The bar plots in this section can be read like Figures 24 and 25. The labels at the end of each bar indicate the sequence collection dates for the first and last "recent" PHIs in the cluster, that is, recorded on or after July 1st, 2012. When there is only one such PHI, we display the corresponding collection date instead. We assume that all chronic cases recorded before July 1st, 2012 were infected prior to 2012. The dark red bar represents the "minimum cluster size before 2012" because several chronic cases diagnosed after July 1st 2012 were probably infected prior to 2012. Also, it is likely that several PHIs sampled in the first half of 2012 match with transmission events that occurred late in 2011.

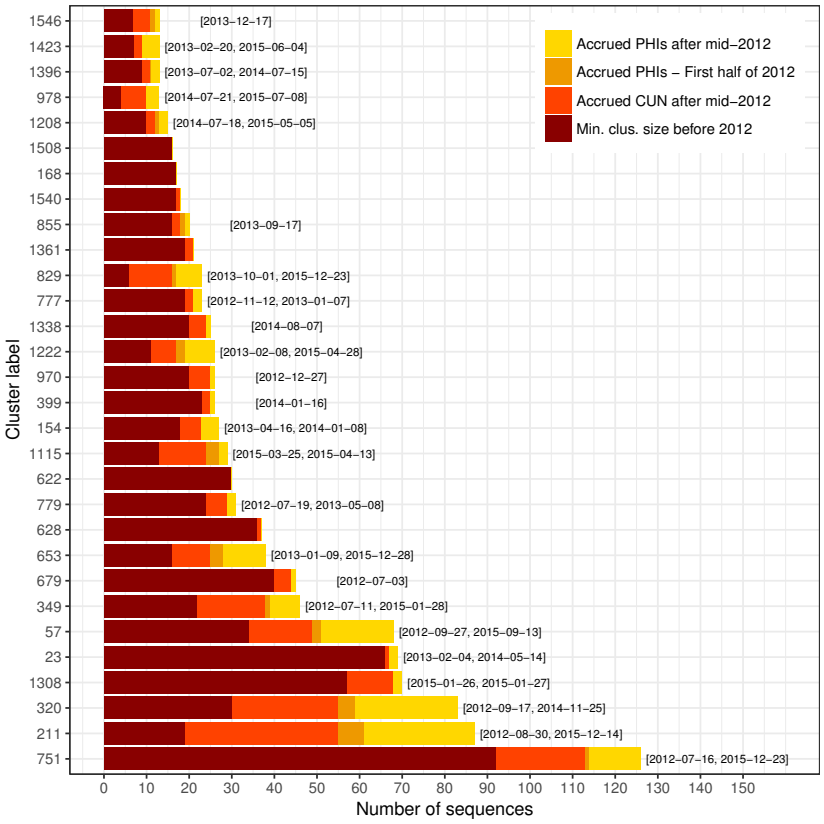

FIGURE 26. **Bar plot showing the breakdown in membership for the 30 largest clusters in the ML + ClusterPicker estimate.**

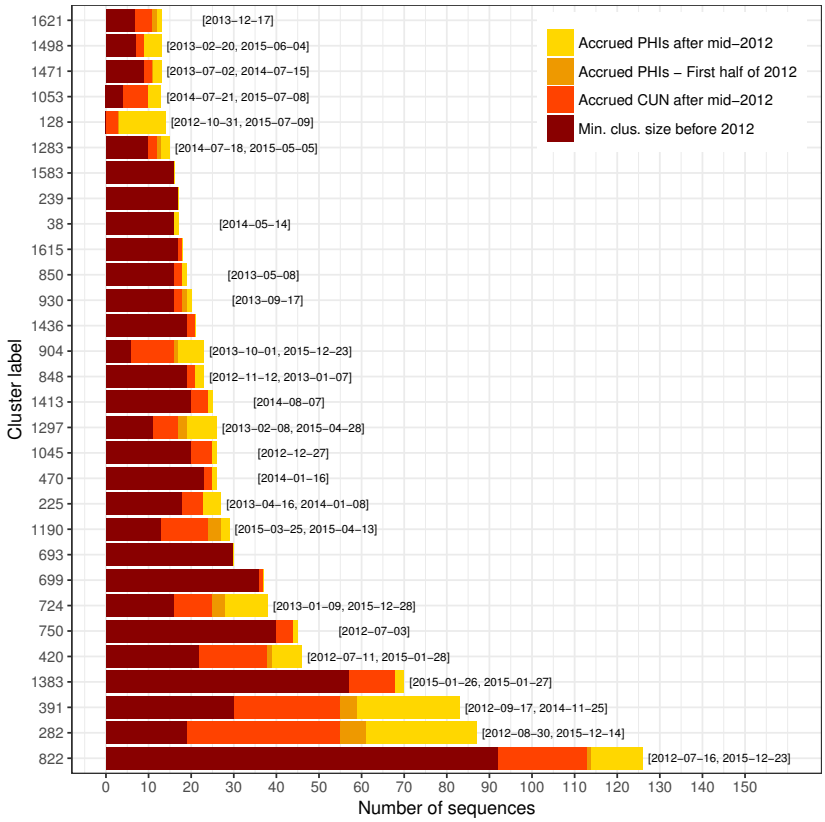

FIGURE 27. **Bar plot showing the breakdown in membership for the 30 largest clusters in the ML + maximum patristic distance estimate.**

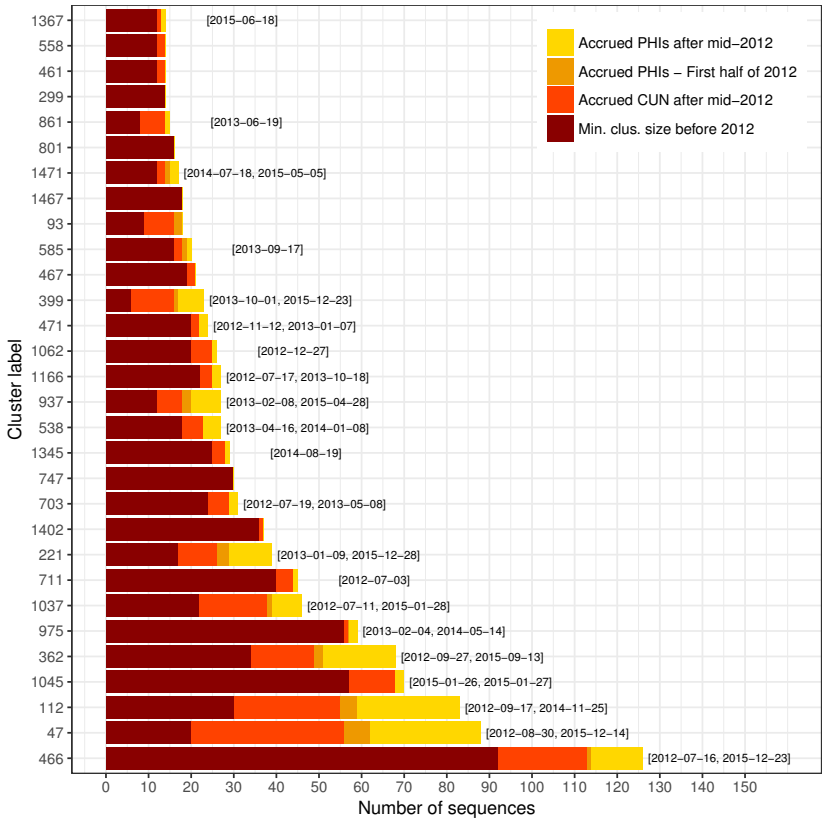

FIGURE 28. **Bar plot showing the breakdown in membership for the 30 largest clusters in the MrBayes+CP estimate.**

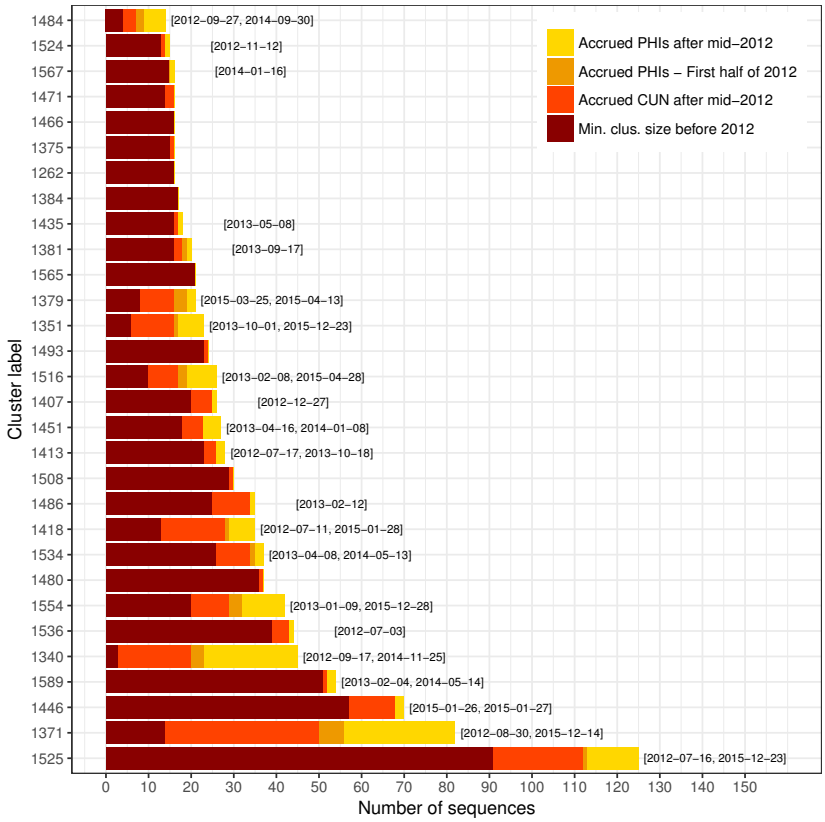

FIGURE 29. **Bar plot showing the breakdown in membership for the 30 largest clusters in the Gap Procedure estimate.**

CHAPTER 6

Discussion

1. Summary

In this thesis, we explored concepts related to phylogenetic clustering, with a substantive focus on HIV-1. In manuscript 1, we investigated through simulations a network interpretation of phylogenetic clusters, obtained under conventional definitions. Overall, we found limited correspondence between phylogenetic clusters and communities in networks, suggesting a limited potential for phylogenetic clusters to be readily used in community inference. Previous studies had found small effects of network characteristics on phylogenetic structure [133, 101], without however focusing on communities. Manuscript 1 complements the conclusions from those studies, by stressing that the link between communities and phylogenetic clusters may not be straightforward.

In manuscript 2, we presented DM-PhyClus, a new Bayesian phylogenetic clustering algorithm. After testing the method through simulations, that highlighted its potential to uncover clusters better than conventional methods, we applied it to a real HIV-1 sequence dataset, revealing a set of clusters largely similar to that inferred in a previous curated analysis [19]. DM-PhyClus partitions the sample by identifying distinctive subtrees, characterised by a different branch length distribution, thus removing the need for arbitrary genetic distance requirements, one major shortcoming of standard approaches. As an additional advantage, the approach also provides straightforward estimates of uncertainty around the obtained cluster estimates. The similarities between the results of conventional analyses and those from DM-PhyClus are not accidental: if clusters are defined in terms of maximum patristic distance, many of them will also be supported by subtrees with noticeably shorter branches.

In manuscript 3, we presented a thorough clustering analysis of a large sample of sequences, obtained from individuals belonging to the MSM risk category who consented to participate in the Québec HIV genotyping program. The manuscript first showed results produced by conventional methods, namely bootstrap-augmented maximum likelihood phylogenetic estimation and pure Bayesian phylogenetic estimation. We included an explicit justification for cutpoint selection, which is rarely seen in the literature. We then clustered the dataset with DM-PhyClus and the Gap Procedure, a fast distance-based approach that also aims to avoid cutpoint selection. Overall, estimates from all conventional approaches were fairly similar. DM-PhyClus proposed the most distinctive partition, its cluster size distribution having a noticeably thinner right tail. All methods however similarly attributed a large proportion of confirmed recent infections to existing or new clusters, a result concordant with [18]. The analyses therefore stress the persistent association between clustering and incidence, and incidentally, the increasing role of recently-infected MSMs as transmission vectors for the virus.

2. Future work

Although the work described in the first manuscript suggests that, in general, phylogenetic clusters and communities in a contact network should not be equated, we did find cases where such a connection may be warranted. It follows that the manuscript does not discredit the usefulness of sequence data for contact network inference. Nevertheless, what phylogenetic inference can tell us about an epidemic's underlying contact network is still, as far as we know, an open question. Sequencing data alone might not be enough to permit reliable inference of network structure or parameters in real-life applications. Combining those data with covariate and epidemiologic information, or partial contact or infection tracing data, might be necessary to improve understanding of the interplay between network structure and the evolution of pathogen populations. Efforts to model explicitly the association between data of those types and phylogenetic estimates would therefore be worthwhile.

To improve mixing in the chain, we opted for a "fixed topology" version of DM-PhyClus: after identifying a suitable topology, the algorithm explores solely the space of cluster configurations permitted by it. Although simulations indicate that such an approach is not

overly detrimental in practice, in view of the sometimes sizeable uncertainty in phylogenetic inference, it might result in an underestimation of the uncertainty around cluster estimates. Rigorously addressing the issue would involve formulating a new transition kernel that proposes joint moves in the space of phylogenies and of cluster membership indices. The main challenge resides in the interaction between the topology and the space of cluster membership indices: since clusters must correspond to clades, a minor transition in the topological space can have profound effects on the cluster membership states accessible from the current configuration. Because of the very large size of the topological space, conventional phylogenetic transition kernels such as nearest-neighbour interchange invariably cause the chain to get trapped in low posterior probability configurations. Escaping those regions requires the transition kernel to propose a very specific series of moves, which may take a very large number of iterations. Updating the algorithm to let the chain explore various topologies, while avoiding the stated problem, would therefore be a natural and necessary improvement.

In 2015, the Swiss HIV Cohort Study (SHCS) database contained information about $19,074$ HIV-positive individuals diagnosed in Switzerland [117], representing a majority of cases diagnosed in the country. Comparing cluster estimates for those data with those presented in manuscript 3 could help highlight the effect of the sampling proportion, higher in the SHCS than in the Québec HIV genotyping program, on cluster inference. The sheer size of the SHCS database remains a major challenge from a computational standpoint though. Work to reduce the computational burden of our algorithm is therefore an essential prerequisite. Improvements to the software, coupled with better transition kernels, will lead to better scalability, allowing DM-PhyClus to cope with the expanding scope of sequence databases. Finally, reworking and adding functionalities to the software's user interface would also be crucial to help the method gain widespread acceptance.

Bibliography

[1] Los Alamos HIV sequence database. http://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html.

[2] Stages of HIV infection, August 2015. https://www.aids.gov/hiv-aids-basics/just-diagnosed-with-hiv-aids/hiv-in-your-body/stages-of-hiv/.

[3] Summary: Estimates of HIV incidence, prevalence and proportion undiagnosed in Canada, 2014, November 2015.

[4] UNAIDS: Fact sheet 2016, 2016. http://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf.

[5] Ahumada-Ruiz, S., Flores-Figueroa, D., Toala-González, I., and Thomson, M. M. Analysis of HIV-1 pol sequences from Panama: identification of phylogenetic clusters within subtype B and detection of antiretroviral drug resistance mutations. *Infect Genet Evol 9*, 5 (Sep 2009), 933–940.

[6] Allen, L. J. An introduction to stochastic epidemic models. In *Mathematical epidemiology*. Springer, 2008, pp. 81–130.

[7] Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., and Ronquist, F. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics 20*, 3 (Feb 2004), 407–415.

[8] Ammassari, A., Murri, R., Pezzotti, P., Trotta, M. P., Ravasio, L., De Longis, P., Lo Caputo, S., Narciso, P., Pauluzzi, S., Carosi, G., Nappa, S., Piano, P., Izzo, C. M., Lichtner, M., Rezza, G., Monforte, A., Ippolito, G., d'Arminio Moroni, M., Wu, A. W., Antinori, A., and AdICONA Study Group. Self-reported symptoms and medication side effects influence adherence to highly active antiretroviral therapy in persons with HIV infection. *J Acquir Immune Defic Syndr 28*, 5 (Dec 2001), 445–449.

[9] Andrieu, C., Moulines, É., and Priouret, P. Stability of stochastic approximation under verifiable conditions. *SIAM J Control Optim 44*, 1 (2005), 283–312.

[10] Antoniak, C. E. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann Stat 2*, 6 (11 1974), 1152–1174.

[11] Avila, D., Keiser, O., Egger, M., Kouyos, R., Böni, J., Yerly, S., Klimkait, T., Vernazza, P. L., Aubert, V., Rauch, A., Bonhoeffer, S., Günthard, H. F., Stadler, T., Spycher, B. D., and , S. H. I. V. C. S. Social meets molecular: Combining phylogenetic and latent class analyses to understand HIV-1 transmission in Switzerland. *Am J Epidemiol 179*, 12 (Jun 2014), 1514–1525.

[12] Barabási, A.-L., and Albert, R. Emergence of scaling in random networks. *Science 286*, 5439 (1999), 509–512.

[13] Bezemer, D., van Sighem, A., Lukashov, V. V., van der Hoek, L., Back, N., Schuurman, R., Boucher, C. A. B., Claas, E. C. J., Boerlijst, M. C., Coutinho, R. A., de Wolf, F., and ATHENA observational cohort. Transmission networks of HIV-1 among men having sex with men in the Netherlands. *AIDS 24*, 2 (Jan 2010), 271–282.

[14] Blackwell, D., and MacQueen, J. B. Ferguson distributions via Polya urn schemes. *Ann Stat 1*, 2 (03 1973), 353–355.

[15] Bocour, A., Udeagu, C.-C. N., Renaud, T. C., Hadler, J. L., and Begier, E. M. Comparing HIV partner notification effectiveness between Blacks and Hispanics in New York City. *Sex Transm Dis 37*, 12 (Dec 2010), 784–788.

[16] Bouchard-Côté, A., Sankararaman, S., and Jordan, M. I. Phylogenetic inference via sequential Monte Carlo. *Syst Biol 61*, 4 (Jul 2012), 579–593.

[17] Brandley, M. C., Leaché, A. D., Warren, D. L., and McGuire, J. A. Are unequal clade priors problematic for Bayesian phylogenetics? *Syst Biol 55*, 1 (Feb 2006), 138–46; author reply 147–51.

[18] Brenner, B., Wainberg, M. A., and Roger, M. Phylogenetic inferences on HIV-1 transmission: implications for the design of prevention and treatment interventions. *AIDS 27*, 7 (Apr 2013), 1045–1057.

[19] Brenner, B. G., Roger, M., Routy, J.-P., Moisi, D., Ntemgwa, M., Matte, C., Baril, J.-G., Thomas, R., Rouleau, D., Bruneau, J., Leblanc, R., Legault, M., Tremblay, C., Charest, H., Wainberg, M. A., and Quebec Primary HIV Infection Study Group. High rates of forward transmission events

after acute/early HIV-1 infection. *J Infect Dis 195*, 7 (Apr 2007), 951–959.

[20] Brenner, B. G., Roger, M., Stephens, D. A., Moisi, D., Hardy, I., Weinberg, J., Turgel, R., Charest, H., Koopman, J., Wainberg, M. A., and Montreal PHI Cohort Study Group. Transmission clustering drives the onward spread of the HIV epidemic among men who have sex with men in Quebec. *J Infect Dis 204*, 7 (Oct 2011), 1115–1119.

[21] Brenner, B. G., and Wainberg, M. A. Future of phylogeny in HIV prevention. *J Acquir Immune Defic Syndr 63 Suppl 2* (Jul 2013), S248–S254.

[22] Bryant, D. A classification of consensus methods for phylogenetics. *DIMACS series in discrete mathematics and theoretical computer science 61* (2003), 163–184.

[23] Caliński, T., and Harabasz, J. A dendrite method for cluster analysis. *Commun Stat Theory 3*, 1 (1974), 1–27.

[24] Centers for Disease Control and Prevention. Estimated HIV incidence in the United States, 2007-2010. HIV surveillance supplemental report., December 2012.

[25] Chaix, M.-L., Descamps, D., Harzic, M., Schneider, V., Deveau, C., Tamalet, C., Pellegrin, I., Izopet, J., Ruffault, A., Masquelier, B., Meyer, L., Rouzioux, C., Brun-Vezinet, F., and Costagliola, D. Stable prevalence of genotypic drug resistance mutations but increase in non-B virus among patients with primary HIV-1 infection in France. *AIDS 17*, 18 (Dec 2003), 2635–2643.

[26] Chalmet, K., Staelens, D., Blot, S., Dinakis, S., Pelgrom, J., Plum, J., Vogelaers, D., Vandekerckhove, L., and Verhofstede, C. Epidemiological study of phylogenetic transmission clusters in a local HIV-1 epidemic reveals distinct differences between subtype B and non-B infections. *BMC Infect Dis 10* (2010), 262.

[27] Chen, H.-F. *Stochastic approximation and its applications*, vol. 64. Springer Science & Business Media, 2006.

[28] Cheon, S., and Liang, F. Phylogenetic tree construction using sequential stochastic approximation Monte Carlo. *Biosystems 91*, 1 (Jan 2008), 94–107.

[29] Cleveland, W. S. Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc 74*, 368 (1979), 829–836.

[30] Cowles, M. K., and Carlin, B. P. Markov chain Monte Carlo convergence diagnostics: a comparative review. *J Am Stat Assoc 91*, 434 (1996), 883–904.

[31] Delva, W., Leventhal, G. E., and Helleringer, S. Connecting the dots: network data and models in HIV epidemiology. *AIDS 30*, 13 (Aug 2016), 2009–2020.

[32] Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc B Met* (1977), 1–38.

[33] Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol 29*, 8 (Aug 2012), 1969–1973.

[34] Dudas, G., and Rambaut, A. Phylogenetic analysis of Guinea 2014 EBOV ebolavirus outbreak. *PLoS Currents 6* (2014).

[35] Dunn, J. C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J Cybernetics* (1973).

[36] Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[37] Eddelbuettel, D. *Seamless R and C++ Integration with Rcpp*. Springer, New York, 2013. ISBN 978-1-4614-6867-7.

[38] Eddelbuettel, D., and François, R. Rcpp: Seamless R and C++ integration. *J Stat Softw 40*, 8 (2011), 1–18.

[39] Efron, B., Halloran, E., and Holmes, S. Bootstrap confidence levels for phylogenetic trees. *Proc Natl Acad Sci U S A 93*, 23 (Nov 1996), 13429–13434.

[40] Erdős, P., and Rényi, A. On the strength of connectedness of a random graph. *Acta Math Hung 12*, 1-2 (1961), 261–267.

[41] Erixon, P., Svennblad, B., Britton, T., and Oxelman, B. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Syst Biol 52*, 5 (2003), pp. 665–673.

[42] Felsenstein, J. Evolutionary trees from dna sequences: A maximum likelihood approach. *J Mol Evol 17*, 6 (1981), 368–376.

[43] Felsenstein, J. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution 39*, 4 (1985), pp. 783–791.

[44] Felsenstein, J., and Kishino, H. Is there something wrong with the bootstrap on phylogenies? a reply to Hillis and Bull. *Syst Biol 42*, 2 (1993), pp. 193–200.

[45] Feng, X., Buell, D. A., Rose, J. R., and Waddell, P. J. Parallel algorithms for Bayesian phylogenetic inference. *J Parallel Distr Com 63*, 7 (2003), 707–718.

[46] Foley, B. T., Leitner, T. K., Apetrei, C., Hahn, B., Mizrachi, I., Mullins, J., Rambaut, A., Wolinsky, S., and Korber, B. T. M. HIV sequence compendium 2015. Tech. rep., Los Alamos National Lab (LANL), Los Alamos, NM (United States), 2015.

[47] Frost, S. D., Pybus, O. G., Gog, J. R., Viboud, C., Bonhoeffer, S., and Bedford, T. Eight challenges in phylodynamic inference. *Epidemics 10* (2015), 88–92.

[48] Gascuel, O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol 14*, 7 (Jul 1997), 685–695.

[49] Geyer, C. J. Markov chain Monte Carlo maximum likelihood. In *Proceedings of 23rd Symposium on the Interface* (Fairfax, Virginia, 1992), E. Karamigas and S. Kaufman, Eds., Interface Foundation of North America, pp. 156–163.

[50] Girvan, M., and Newman, M. E. Community structure in social and biological networks. *P Natl Acad Sci USA 99*, 12 (2002), 7821–7826.

[51] Görür, D., and Edward Rasmussen, C. Dirichlet process Gaussian mixture models: Choice of the base distribution. *J Comput Sci Technol 25*, 4 (2010), 653–664.

[52] Green, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*, 4 (1995), 711–732.

[53] Hamming, R. W. Error detecting and error correcting codes. *Bell Syst Tech J 29*, 2 (1950), 147–160.

[54] Hasegawa, M., Kishino, H., and Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol 22*, 2 (1985), 160–174.

[55] Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer series in statistics. Springer-Verlag New York, New York, 2009.

[56] Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*, 1 (1970), 97–109.

[57] Hemelaar, J., Gouws, E., Ghys, P. D., and Osmanov, S. Global trends in molecular epidemiology of HIV-1 during 2000-2007. *AIDS 25*, 5 (Mar 2011), 679–689.

[58] Hillis, D. M., and Bull, J. J. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst Biol 42*, 2 (1993), 182–192.

[59] Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. *Bayesian nonparametrics*, vol. 28. Cambridge University Press, 2010.

[60] Holder, M., and Lewis, P. O. Phylogeny estimation: traditional and Bayesian approaches. *Nat Rev Genet 4*, 4 (Apr 2003), 275–284.

[61] Holder, M. T., Sukumaran, J., and Lewis, P. O. A justification for reporting the majority-rule consensus tree in Bayesian phylogenetics. *Syst Biol 57*, 5 (2008), 814.

[62] Hué, S., Clewley, J. P., Cane, P. A., and Pillay, D. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS 18*, 5 (Mar 2004), 719–728.

[63] Hué, S., Clewley, J. P., Cane, P. A., and Pillay, D. Investigation of HIV-1 transmission events by phylogenetic methods: requirement for scientific rigour. *AIDS 19*, 4 (Mar 2005), 449–450.

[64] Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L. P., Marcet-Houben, M., and Gabaldón, T. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res 42*, Database issue (Jan 2014), D897–D902.

[65] Ibe, S., Hattori, J., Fujisaki, S., Shigemi, U., Fujisaki, S., Shimizu, K., Nakamura, K., Kazumi, T., Yokomaku, Y., Mamiya, N., Hamaguchi, M., and Kaneda, T. Trend of drug-resistant HIV type 1 emergence among therapy-naive patients in Nagoya, Japan: an 8-year surveillance from 1999 to 2006. *AIDS Res Hum Retroviruses 24*, 1 (Jan 2008), 7–14.

[66] Ishwaran, H., and James, L. Gibbs sampling methods for stick-breaking priors. *J Am Stat Assoc 96*, 453 (2001), 161–173.

[67] Joint United Nations Programme on HIV/AIDS and others. *Global AIDS Update: 2016*. UNAIDS, 2016.

[68] Jukes, T. H., and Cantor, C. R. Evolution of protein molecules. *Mammalian protein metabolism 3*, 21 (1969), 132.

[69] Keeling, M. J., and Eames, K. T. D. Networks and epidemic models. *J R Soc Interface 2*, 4 (Sep 2005), 295–307.

[70] Kenah, E., Britton, T., Halloran, M. E., and Longini, Jr, I. M. Molecular infectious disease epidemiology: Survival analysis and algorithms linking phylogenies to transmission trees. *PLOS Comput Biol 12*, 4 (04 2016), 1–29.

[71] Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol 16*, 2 (Dec 1980), 111–120.

[72] Kouyos, R. D., von Wyl, V., Yerly, S., Böni, J., Rieder, P., Joos, B., Taffé, P., Shah, C., Bürgisser, P., Klimkait, T., Weber, R., Hirschel, B., Cavassini, M., Rauch, A., Battegay, M., Vernazza, P. L., Bernasconi, E., Ledergerber, B., Bonhoeffer, S., Günthard, H. F., and Swiss H. I. V Cohort Study. Ambiguous nucleotide calls from population-based sequencing of HIV-1 are a marker for viral diversity and the age of infection. *Clin Infect Dis 52*, 4 (Feb 2011), 532–539.

[73] Kouyos, R. D., von Wyl, V., Yerly, S., Böni, J., Taffé, P., Shah, C., Bürgisser, P., Klimkait, T., Weber, R., Hirschel, B., Cavassini, M., Furrer, H., Battegay, M., Vernazza, P. L., Bernasconi, E., Rickenbach, M., Ledergerber, B., Bonhoeffer, S., and Günthard, H. F. Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. *J Infect Dis 201*, 10 (May 2010), 1488–1497.

[74] Larget, B., and Simon, D. L. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol Biol Evol 16* (1999), 750–759.

[75] Leigh Brown, A. J., Lycett, S. J., Weinert, L., Hughes, G. J., Fearnhill, E., Dunn, D. T., and the UK HIV Drug Resistance Collaboration. Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *J Infect Dis 204*, 9 (Nov 2011), 1463–1469.

[76] Leventhal, G. E., Günthard, H. F., Bonhoeffer, S., and Stadler, T. Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission. *Mol Biol Evol 31*, 1 (Jan 2014), 6–17.

[77] Leventhal, G. E., Kouyos, R., Stadler, T., von Wyl, V., Yerly, S., Böni, J., Cellerai, C., Klimkait, T., Günthard, H. F., and Bonhoeffer, S. Inferring epidemic contact structure from phylogenetic trees. *PLoS Comput Biol 8*, 3

(2012), e1002413.

[78] Lewis, F., Hughes, G. J., Rambaut, A., Pozniak, A., and Brown, A. J. L. Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med 5*, 3 (Mar 2008), e50.

[79] Liang, F., Liu, C., and Carroll, R. J. Stochastic approximation in Monte Carlo computation. *J Am Stat Assoc 102*, 477 (2007), 305–320.

[80] Liljeros, F., Edling, C. R., Amaral, L. A., Stanley, H. E., and Aberg, Y. The web of human sexual contacts. *Nature 411*, 6840 (Jun 2001), 907–908.

[81] Lindström, A., Ohlis, A., Huigen, M., Nijhuis, M., Berglund, T., Bratt, G., Sandström, E., and Albert, J. HIV-1 transmission cluster with M41L 'singleton' mutation and decreased transmission of resistance in newly diagnosed Swedish homosexual men. *Antivir Ther 11*, 8 (2006), 1031–1039.

[82] Makarenkov, V., Boc, A., Xie, J., Peres-Neto, P., Lapointe, F.-J., and Legendre, P. Weighted bootstrapping: a correction method for assessing the robustness of phylogenetic trees. *BMC Evol Biol 10* (2010), 250.

[83] Marcus, J. L., Chao, C. R., Leyden, W. A., Xu, L., Quesenberry, Jr, C. P., Klein, D. B., Towner, W. J., Horberg, M. A., and Silverberg, M. J. Narrowing the gap in life expectancy between HIV-infected and HIV-uninfected individuals with access to care. *J Acquir Immune Defic Syndr 73*, 1 (Sep 2016), 39–46.

[84] Mau, B., Newton, M. A., and Larget, B. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics 55*, 1 (Mar 1999), 1–12.

[85] McCloskey, R. M., Liang, R. H., and Poon, A. F. Reconstructing contact network parameters from viral phylogenies. *bioRxiv* (2016), 050435.

[86] Nei, M., and Kumar, S. *Molecular evolution and phylogenetics*. Oxford University Press, 2000.

[87] Newman, M. E. J., and Girvan, M. Finding and evaluating community structure in networks. *Phys Rev E 69* (Feb 2004), 026113.

[88] Paradis, E., Claude, J., and Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics 20* (2004), 289–290.

[89] Pickett, K. M., and Randle, C. P. Strange Bayes indeed: uniform topological priors imply non-uniform clade priors. *Mol Phylogenet Evol 34*, 1 (Jan 2005), 203–211.

[90] Pitman, J., and Yor, M. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann Probab 25*, 2 (1997), 855–900.

[91] Pons, P., and Latapy, M. Computing communities in large networks using random walks. *ArXiv Physics e-prints* (Dec. 2005).

[92] Posada, D., and Crandall, K. A. Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). *Mol Biol Evol 18*, 6 (Jun 2001), 897–906.

[93] Price, M. N., Dehal, P. S., and Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLOS One 5*, 3 (03 2010), 1–10.

[94] Prosperi, M. C. F., Ciccozzi, M., Fanti, I., Saladini, F., Pecorari, M., Borghi, V., Giambenedetto, S. D., Bruzzone, B., Capetti, A., Vivarelli, A., Rusconi, S., Re, M. C., Gismondo, M. R., Sighinolfi, L., Gray, R. R., Salemi, M., Zazzi, M., Luca, A. D., and the ARCA collaborative group. A novel methodology for large-scale phylogeny partition. *Nat Commun 2* (May 2011), 321.

[95] Public Health Agency of Canada. Population-specific HIV/AIDS status report: Gay, bisexual, two-spirit and other men who have sex with men, 2013.

[96] Raghavan, U. N., Albert, R., and Kumara, S. Near linear time algorithm to detect community structures in large-scale networks. *Phys Rev 76*, 3 (Sept. 2007), 036106.

[97] Ragonnet-Cronin, M., Hodcroft, E., Hué, S., Fearnhill, E., Delpech, V., Brown, A. J. L., Lycett, S., and the UK HIV Drug Resistance Database. Automated analysis of phylogenetic clusters. *BMC Bioinformatics 14* (2013), 317.

[98] Rannala, B., Zhu, T., and Yang, Z. Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. *Mol Biol Evol 29*, 1 (Jan 2012), 325–335.

[99] Revell, L. J. phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol 3* (2012), 217–223.

[100] Richardson, S., and Green, P. J. On Bayesian analysis of mixtures with an unknown number of components (with Discussion). *J Roy Stat Soc B 59*, 4 (1997), 731–792.

[101] Robinson, K., Fyson, N., Cohen, T., Fraser, C., and Colijn, C. How the dynamics and structure of sexual contact networks shape pathogen phylogenies. *PLoS Comput Biol 9*, 6 (2013), e1003105.

[102] Rodríguez, F., Oliver, J. L., Marín, A., and Medina, J. R. The general stochastic model of nucleotide substitution. *J Theor Biol 142*, 4 (Feb 1990), 485–501.

[103] Rokas, A., Krüger, D., and Carroll, S. B. Animal evolution and the molecular signature of radiations compressed in time. *Science 310*, 5756 (Dec 2005), 1933–1938.

[104] Ronquist, F., and Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics 19*, 12 (Aug 2003), 1572–1574.

[105] Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol 61*, 3 (2012), 539–542.

[106] Saitou, N., and Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol 4*, 4 (Jul 1987), 406–425.

[107] Sanderson, C., and Curtin, R. Armadillo: a template-based C++ library for linear algebra. *JOSS 1*, 2 (2016), 26–32.

[108] Schliep, K. phangorn: phylogenetic analysis in R. *Bioinformatics 27*, 4 (2011), 592–593.

[109] Schülter, E., Oette, M., Balduin, M., Reuter, S., Rockstroh, J., Fätkenheuer, G., Esser, S., Lengauer, T., Agacfidan, A., Pfister, H., Kaiser, R., and Akgül, B. HIV prevalence and route of transmission in Turkish immigrants living in North-Rhine Westphalia, Germany. *Med Microbiol Immunol* (Apr 2011).

[110] Soltis, P. S., and Soltis, D. E. Applying the bootstrap in phylogeny reconstruction. *Stat Sci 18*, 2 (2003), pp. 256–267.

[111] Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* (2014).

[112] Stamatakis, A., Ludwig, T., and Meier, H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics 21*, 4 (2005), 456.

[113] Steel, M., and Pickett, K. M. On the impossibility of uniform priors on clades. *Mol Phylogenet Evol 39*, 2 (May 2006), 585–586.

[114] Studier, J. A., Keppler, K. J., et al. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol Biol Evol 5*, 6 (1988), 729–731.

[115] Sullivan, P. S., Carballo-Diéguez, A., Coates, T., Goodreau, S. M., McGowan, I., Sanders, E. J., Smith, A., Goswami, P., and Sanchez, J. Successes and challenges of HIV prevention in men who have sex with men. *Lancet 380*, 9839 (Jul 2012), 388–399.

[116] Susko, E. Bootstrap support is not first-order correct. *Syst Biol 58*, 2 (Apr 2009), 211–223.

[117] Swiss HIV Cohort Study. Cohort profile: the Swiss HIV cohort study. *Int J Epidemiol 39*, 5 (Oct 2010), 1179–1189.

[118] Swofford, D. L. PAUP*: Phylogenetic analysis using parsimony (and other methods)., 2003.

[119] Tamura, K., and Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol 10*, 3 (May 1993), 512–526.

[120] Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol 28*, 10 (Oct 2011), 2731–2739.

[121] Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol 30*, 12 (Dec 2013), 2725–2729.

[122] Tierney, L. Markov chains for exploring posterior distributions. *Ann Stat* (1994), 1701–1728.

[123] Toivonen, R., Onnela, J.-P., Saramäki, J., Hyvönen, J., and Kaski, K. A model for social networks. *Physica A 371*, 2 (2006), 851–860.

[124] Travers, J., and Milgram, S. An experimental study of the small world problem. *Sociometry* (1969), 425–443.

[125] Van der Spoel van Dijk, A., Makhoahle, P. M., Rigouts, L., and Baba, K. Diverse molecular genotypes of Mycobacterium tuberculosis complex isolates circulating in the Free State, South Africa. *Int J Microbiol 2016* (2016), 6572165.

[126] van Sighem, A., Nakagawa, F., De Angelis, D., Quinten, C., Bezemer, D., de Coul, E. O., Egger, M., de Wolf, F., Fraser, C., and Phillips, A. Estimating HIV incidence, time to diagnosis, and the undiagnosed HIV epidemic using routine surveillance data. *Epidemiology 26*, 5 (2015), 653.

[127] Villandre, L., Labbe, A., Brenner, B., Roger, M., and Stephens, D. A. DM-PhyClus: A Bayesian phylogenetic algorithm for transmission cluster inference. *ArXiV* (2017).

[128] Villandre, L., Stephens, D. A., Labbe, A., Günthard, H. F., Kouyos, R., Stadler, T., and Swiss HIV Cohort Study. Assessment of overlap of phylogenetic transmission clusters and communities in simple sexual contact networks: Applications to HIV-1. *PLoS One 11*, 2 (2016), e0148459.

[129] Vinh, N. X., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning* (2009), ACM, pp. 1073–1080.

[130] Vrbik, I., Stephens, D. A., Roger, M., and Brenner, B. G. The Gap procedure: for the identification of phylogenetic clusters in HIV-1 sequence data. *BMC Bioinformatics 16* (2015), 355.

[131] Wang, Y., and Yang, Z. Priors in Bayesian phylogenetics. *Bayesian phylogenetics: methods, algorithms, and applications. Chapman and Hall/CRC* (2014), 5–23.

[132] Watts, D. J., and Strogatz, S. H. Collective dynamics of 'small-world'networks. *Nature 393*, 6684 (1998), 440–442.

[133] Welch, D. Is network clustering detectable in transmission trees? *Viruses 3*, 6 (Jun 2011), 659–676.

[134] Wensing, A. M., Calvez, V., Günthard, H. F., Johnson, V. A., Paredes, R., Pillay, D., Shafer, R. W., and Richman, D. D. 2015 update of the drug resistance mutations in HIV-1. *Top Antivir Med 23*, 4 (2015), 132–141.

[135] World Health Organization. Programmatic update: antiretroviral treatment as prevention (TASP) of HIV and TB: executive summary. Tech. rep., World Health

Organization, 2012.

[136] Yang, Z. *Computational Molecular Evolution*. Oxford Series in Ecology and Evolution. Oxford University Press, Oxford, 2006.

[137] Yang, Z., and Rannala, B. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Mol Biol Evol 14*, 7 (Jul 1997), 717–724.

[138] Yang, Z., and Rannala, B. Branch-length prior influences Bayesian posterior probability of phylogeny. *Syst Biol 54*, 3 (Jun 2005), 455–470.

[139] Yang, Z., and Rannala, B. Molecular phylogenetics: principles and practice. *Nat Rev Genet 13*, 5 (May 2012), 303–314.

[140] Zarrabi, N., Prosperi, M., Belleman, R. G., Colafigli, M., De Luca, A., and Sloot, P. M. A. Combining epidemiological and genetic networks signifies the importance of early treatment in HIV-1 transmission. *PLoS One 7*, 9 (2012), e46156.