ARTICLE IN PRESS

Epidemics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

# Comparison of cluster-based and source-attribution methods for estimating transmission risk using large HIV sequence databases

Stéphane Le Vu[a,*], Oliver Ratmann[b], Valerie Delpech[c], Alison E. Brown[c], O. Noel Gill[c], Anna Tostevin[d], Christophe Fraser[e], Erik M. Volz[a]

[a] Department of Infectious Disease Epidemiology and the NIHR HPRU on Modeling Methodology, Imperial College London, United Kingdom
[b] Department of Mathematics, Imperial College London, United Kingdom
[c] HIV and STI Department of Public Health England's Centre for Infectious Disease Surveillance and Control, London, United Kingdom
[d] Department of Infection and Population Health and the NIHR HPRU in Blood Borne and Sexually Transmitted Infections, University College London, United Kingdom
[e] Li Ka Shing Centre for Health Information and Discovery, Oxford University, United Kingdom

ARTICLE INFO

*Keywords:*
Phylogenetic analysis
Cluster analysis
Phylodynamics
HIV epidemiology
Computer simulation

## ABSTRACT

Phylogenetic clustering of HIV sequences from a random sample of patients can reveal epidemiological transmission patterns, but interpretation is hampered by limited theoretical support and statistical properties of clustering analysis remain poorly understood. Alternatively, source attribution methods allow fitting of HIV transmission models and thereby quantify aspects of disease transmission.

A simulation study was conducted to assess error rates of clustering methods for detecting transmission risk factors. We modeled HIV epidemics among men having sex with men and generated phylogenies comparable to those that can be obtained from HIV surveillance data in the UK. Clustering and source attribution approaches were applied to evaluate their ability to identify patient attributes as transmission risk factors.

We find that commonly used methods show a misleading association between cluster size or odds of clustering and covariates that are correlated with time since infection, regardless of their influence on transmission. Clustering methods usually have higher error rates and lower sensitivity than source attribution method for identifying transmission risk factors. But neither methods provide robust estimates of transmission risk ratios. Source attribution method can alleviate drawbacks from phylogenetic clustering but formal population genetic modeling may be required to estimate quantitative transmission risk factors.

## 1. Introduction

Phylogenetic clustering of HIV sequences has been commonly used to characterise transmission patterns (Lewis et al., 2008; Little et al., 2014; Poon et al., 2015). In developed countries, routine testing for drug resistance mutations has led to the development of large HIV sequence databases. Numerous previous investigations have leveraged such databases along with patients' clinical and demographic covariates to study clusters of patients with closely related HIV sequences. These clusters can be defined in numerous ways, such as using genetic, evolutionary (Aldous et al., 2012) or phylogeny-based distance criteria (Poon et al., 2015), or including measures of phylogenetic credibility (Hué et al., 2004). The central idea underlying cluster analysis is that patients with similar viruses are likely to be epidemiologically related, such as by direct transmission, or by being infected by a common source or by a short chain of transmissions with potentially unsampled intermediate members (Frost and Pillay, 2015). Consequently, individuals who are responsible for more transmissions would likely be in a cluster with more individuals, and patients who transmit at a higher rate would also more likely be in a cluster as opposed to isolated.

The majority of clustering analyses identify transmission risk factors by regressing odds of cluster membership (Brenner et al., 2011; Dennis et al., 2012) or sometimes cluster size or node degree (Little et al., 2014; Pines et al., 2016; Morgan et al., 2017) on patient covariates, although particular statistical models vary greatly. Because HIV evolves rapidly and the rate of mutation has been estimated within hosts, it is possible to quantify the probability of virus lineages diverging over a given range of time, such as the time of diagnosis of a putative donor and recipient of infection (Leitner and Albert, 1999). Using information about molecular clock rates and diagnosis dates can be used to refine clustering analyses by excluding pairs which are incompatible with clinical and behavioural histories (Ratmann et al., 2016).

Clustering analyses are common, because they are easy to implement and computationally cheap once a phylogeny is estimated.

* Corresponding author.
*E-mail address:* s.le-vu@imperial.ac.uk (S. Le Vu).

http://dx.doi.org/10.1016/j.epidem.2017.10.001
Received 31 July 2017; Received in revised form 12 October 2017; Accepted 17 October 2017

Clustering methods can be applied to sequence databases involving tens of thousands of patients. But despite a long history and numerous published examples, clustering analysis as a statistical methodology has several drawbacks. Most methods rely on a tuneable threshold, such as a cutoff for genetic distance below which samples are considered to be clustered (Grabowski and Redd, 2014). It is problematic to tune this threshold and most analyses use an ad-hoc threshold or evaluate sensitivity over a range of thresholds. If a panel of known transmission pairs is available, the threshold genetic distance used by clustering methods can be tuned to achieve a desired tradeoff in sensitivity and specificity in classifying transmission pairs (Rose et al., 2016). Note however, that a threshold in one setting may not be appropriate in all settings, since optimality will depend on the background genetic diversity of the sample and proportion of hosts sampled (Volz et al., 2012). Even with carefully calibrated thresholds to identify transmission pairs, clustering does not exclude the possibility that an unsampled individual is a common source of infection for closely related patients, and potentially informative links with distances above threshold are neglected.

The interpretation of clustering often makes an implicit assumption that clusters form a simple uniform random sample of transmission pairs over the recent epidemic history. But numerous factors influence the probability of a sample appearing in a cluster, foremost the time since infection at the time of sampling (Volz et al., 2012). Patients who are early in the course of their infection are likely to be closely related to their donor, and thus cluster membership is not necessarily related to the transmission risk of such patients. Any variable correlated with time since infection is likely to be found associated with clustering. That includes CD4 count, viral load, diagnosis status, treatment status, age of the patient and propensity for early testing (Frost and Pillay, 2015; Poon, 2016).

In this study, we use a recently-developed method which is computationally tractable and based on estimating the probability that a given sampled case is the source of infection for another case, called the *infector probability* (Volz and Frost, 2013). While conceptually similar to clustering, rather than dichotomising all pairs as ‘clustered’ or ‘not clustered’, the source attribution (SA) approach weights each pair in the phylogeny by the estimated probability that the putative donor infected the recipient. These probabilities account for additional epidemiological and clinical data that generic clustering methods neglect. The method can make use of variables that are informative about time since infection, such as CD4, viral load, or incidence assays to account for biased sampling. It also makes use of independent estimates of prevalence and incidence, which yield insight into the proportion of the population sampled and the probability that an unsampled individual is the source of infection. Additionally, the method obviates the need to define arbitrary clustering thresholds, so that patients sampled late in infection and who have correspondingly distant relations in the virus phylogeny can nevertheless be included in the analysis. Finally, the sum of infector probabilities for a given potential donor provides an intuitive statistic to examine individual factors influencing transmission rates.

The aim of our study is to assess how source attribution compares with generic clustering method in detecting heterogeneous transmission according to patients characteristics. This assessment was based on detailed simulations that aim to match on epidemiological and molecular data available in the United Kingdom. We particularly wanted to illustrate how outcomes from the two methods can be informative about transmission risk among men who have sex with men.

## 2. Materials and methods

In this section, we first describe the source attribution and generic clustering methods used to infer epidemiological quantities from labeled viral sequence data. We then present the simulation experiments used to generate epidemic trajectories and phylogenetic trees taken as input for the above methods. Finally, we describe the statistical tests used to evaluate the ability of both approaches to identify transmission risk factors and to correctly estimate transmission risk ratios assigned in two counterfactual scenarios in the simulations.

### 2.1. Source attribution method

We applied a phylogenetic source attribution (SA) method that infers the probability of potential transmission between each pair of individuals (infector probability) from a time-scaled phylogeny (Volz and Frost, 2013). The calculation of infector probabilities also uses as inputs additional epidemiological data such as incidence and prevalence of infection. These data inform the proportion of the population sampled, and thus influence the estimated probability that closely related patients have a common source of infection or an unsampled intermediary in a transmission chain. The SA method can also account for the time since infection at the time of sampling, using CD4 counts or incidence assay test results (i.e. in the form recently infected or not).

The calculation detailed in Volz and Frost (2013) is based on the following rationale: For a sampled individual $i$ to have infected a sampled individual $j$, the lineages ancestral to $i$ and $j$ must be in patients $i$ and $j$ around the time of transmission (assuming small within-host genetic diversity). This probability is modeled with the survivor function $\psi_i(t)$. Initially, $\psi_i(t_i) = 1$ at the time $t_i$ that $i$ is sampled. Going backwards in time denoted $s$ (towards the root of the phylogeny), the survivor function is modeled as

$$\frac{d}{ds}\psi_i(s) = -\psi_i(s)P(i\,\text{infected at}\,s|x_i,\,F(s),\,Y(s)), \tag{1}$$

where $x_i$ is a vector of covariates for patient $i$; $Y(s)$ is a potentially vector-valued function of time that denotes the total number infected in the population of different types corresponding to covariates $x_i$ (demes); and $F(s)$ is a matrix-valued function of time that describes the rate of transmission within and between demes. Note that $1 - \psi_i(s)$ is the probability that the lineage ancestral to patient $i$ is in a different host who may have been unsampled.

Secondly, conditional on both lineages being hosted by $i$ and $j$ between the time of their most recent common ancestor (MRCA) $s_{ij}$ and time of sampling, a transmission event must have taken place from a donor characterized by covariates $x_i$ and sampling time $t_i$ to a recipient characterized by $x_j$ and $t_j$, rather than the opposite. Combining these conditions gives the model for estimated infector probability $W_{ij}$ that patient $i$ infected $j$:

$$W_{\mathrm{ij}} = \psi_i(s_{\mathrm{ij}})\psi_j(s_{\mathrm{ij}})P(i \to j|i \to j\;\text{or}\;j \to i,\,x_i,\,t_i,\,x_j,\,t_j). \tag{2}$$

The SA method uses a continuous time Markov chain (CTMC) model to reconstruct the likely state of a lineage at the time of transmission given observed covariates at time of sampling. Rates of the CTMC are derived from an epidemic trajectory summarized by three processes: $Y(s)$, $F(s)$ and $G(s)$ which is a matrix valued function of time that describes migration between demes, including progression of infected hosts through different stages of infection. By solving ordinary differential equations, the model updates the probability $\psi(s)$ as a function of $F(s)$, $G(s)$ and $Y(s)$ that a lineage corresponds to the same host that was sampled while traversing the time-scaled phylogeny backwards in time.

This model is conceptually similar to the coalescent approach used to simulate the trees as described in Section 2.3.3. But to account for realistic lack of prior knowledge about the epidemic history, we used a misspecified model for lineage transition rates. The model only accounted for progression between CD4-stages, not on diagnosis, treatment, demographic age or other stages; the generic transmission risk factor was assumed to be unobserved; and the model deliberately misspecified transmission rates by stage of infection as constant. Furthermore, infector probabilities were computed under the approximation that incidence and prevalence were constant over the past 20 years of the epidemic history. With poor prior information about transmission patterns, the analysis procedure offered fewer chances to recover the simulation inputs. Thus we expected that the outcomes provide a conservative picture of the performance of the SA method, that could be improved if more refined surveillance estimates are available.

To estimate relative transmission risk, a summary statistic called *out-degree* was computed from infector probabilities and used for subsequent statistical analysis. The out-degree for individual *i* is defined as $d_i = \Sigma_{j \neq i} W_{ij}$, and represents the estimated cumulative number of transmissions that are included in the sample and originate from patient *i*. We used this quantity to provide estimates of relative transmission risk by stage of infection, accounting for their expected durations. The transmission rate for a patient sampled at a given stage was derived as his out-degree normalised by the cumulative duration of all previous stages. We estimated the difference and ratio in transmission rates between patients in first and last stage of infection.

To study assortative transmission patterns, we also computed the total number of transmissions between groups as $A_{\text{uv}} = \sum_{i \in S_u} \sum_{j \in S_v} W_{\text{ij}}$, where $S_u$ is the set of sampled hosts in state $u$ (defined by stage, demographic age, risk factor, and continuum of care).

The algorithm for calculating the matrix of infector probabilities in this study is implemented in function *phylo.source.attribution.hiv.msm* of the R package *phydynR* (Volz, 2016b).

### 2.2. Clustering algorithms

We used three hierarchical clustering algorithms. Firstly, we used the *hivclustering* software (Weaver and Kosakovsky Pond, 2016; Wertheim et al., 2014; Little et al., 2014) on pairwise patristic evolutionary distances derived from coalescent trees (described in Section 2.3.3). This approach links individuals within a set such that at least one other individual within the set has a distance less than a pre-specified threshold value (single-linkage algorithm) (Jain et al., 1999). Secondly, we computed *neighborhood* sizes, which we define as the number of individuals within a pre-specified threshold evolutionary distance (complete-linkage algorithm). For these two methods, we varied the thresholds of genetic distance under which cluster membership is defined with values 0.5%, 1.5% and 5.0% substitutions per site. Third clustering method (denoted here *tMRCA*) is another single-linkage algorithm but directly uses the branch lengths from time dated trees. It links two individuals whose nodes have both a time to their MRCA that is less than some threshold, indicating a limited amount of divergence between the respective viruses (Leigh Brown et al., 2011). We tested threshold values of 2, 5 and 10 years.

Networks were characterized by the odds of clustering, the sizes of clusters or neighborhoods, and assortativity (like-with-like composition of clusters) (Newman, 2003). Unless otherwise specified (particularly in Section 3.4), clustering results in Section 3 are those from *hivclustering* method. To compare transmission networks as described by the SA method and *hivclustering* method, we represented graphically how individuals from one same cluster selected at threshold 5% are connected by infector probabilities and genetic distance below 1.5% threshold using igraph (Csardi and Nepusz, 2006).

### 2.3. Simulations

Our motivation was to obtain a simple though realistic transmission history in a population that is comparable to men who have sex with men (MSM) in London. Simulations were designed to replicate the sampling proportion, clinical data, and age structure of London MSM in the UK HIV Drug Resistance database (UK HIV Drug Resistance Database, 2016). Epidemic simulation was based on a compartmental model which describe the dynamics of the number of infected hosts in different categories. Additionally, genealogical trees were simulated conditioning on the epidemic history, and trees were matched to the real data from the UK pertaining to the number of sequence samples, times of sampling and clinical stage of infection.

**Table 1**
Initial parameter values for simulated epidemic model.

| Notation | Parameter | Value |
|---|---|---|
| | *Age progression rate*[a] | |
| $\alpha_1$ | Group 1 [18–27) | $1/9/365\ \text{day}^{-1}$ |
| $\alpha_2$ | Group 2 [27–33) | $1/6/365\ \text{day}^{-1}$ |
| $\alpha_3$ | Group 3 [33–40) | $1/7/365\ \text{day}^{-1}$ |
| $\alpha_4$ | Group 4 [40–80.5) | $1/40.5/365\ \text{day}^{-1}$ |
| | *Stage progression rate*[b] | |
| $\gamma_1$ | Stage 2 (CD4 > 500 cells/mm$^3$) | $1/3.32/365\ \text{day}^{-1}$ |
| $\gamma_2$ | Stage 3 (350 < CD4 ≤ 500 cells/mm$^3$) | $1/2.7/365\ \text{day}^{-1}$ |
| $\gamma_3$ | Stage 4 (200 < CD4 ≤ 350 cells/mm$^3$) | $1/5.5/365\ \text{day}^{-1}$ |
| $\gamma_4$ | Stage 5 (CD4 ≤ 200 cells/mm$^3$) | $1/5.06/365\ \text{day}^{-1}$ |
| | *Fraction of individuals transitioning from*[b] | |
| $\pi_1$ | Stage 1 to stage 2 | 0.76 |
| $\pi_2$ | Stage 1 to stage 3 | 0.19 |
| $\pi_3$ | Stage 1 to stage 4 | 0.05 |
| $\pi_4$ | Stage 1 to stage 5 | 0 |
| $a$ | Age assortativity factor[c] | 0.5 |
| $p$ | Proportion of individuals in low-risk group | 0.8 |
| $m$ | Per lineage rate of migration to source compartment | $1/50/365\ \text{day}^{-1}$ |
| $g$ | Rate of growth of source compartment | $1/3/365\ \text{day}^{-1}$ |
| $s$ | Initial size of source compartment | 1000 |
| $i$ | Incidence scaling factor for London MSM[d] | 0.03 |
| | *Diagnosis rate* | |
| $d_{85}$ | Fixed rate prior to 1985 | $1/10\ \text{year}^{-1}$ |
| $\mu_d$ | Maximum value of logistic function after 1985 [d] | $1/3\ \text{year}^{-1}$ |
| $k_d$ | Steepness of logistic function after 1985[d] | $1/7\ \text{year}^{-1}$ |
| | *Treatment rate* | |
| $t_{95}$ | Fixed rate prior to 1995 | 0 |
| $\mu_t$ | Maximum value of logistic function after 1995 | 1 |
| $k_t$ | Steepness of logistic function after 1995 | 0.5 |
| $e$ | Treatment effectiveness | 0.95 |
| | *Transmission weight conferred to individuals in* | |
| $w_{s1}$ | Stage 1 | 1 |
| $w_{s2}$ to $w_{s4}$ | Stages 2 to 4[e] | 0.1 |
| $w_{s5}$ | Stage 5[e] | 0.3 |
| $w_{a1}$ to $w_{a4}$ | Age groups 1 to 4 | 1 |
| $w_{c1}$ | Care status 1 (undiagnosed) | 1 |
| $w_{c2}$ | Care status 2 (diagnosed and untreated) | 0.5 |
| $w_{c3}$ | Care status 3 (diagnosed and treated) | 0.05 |
| $w_{r1}$ | Risk status 1 (low risk) | 1 |
| $w_{r2}$ | Risk status 2 (high risk) | 10 |

[a] From quartiles of age of MSM diagnosed in London reported in UKDRDB.

[b] From Cori et al. (2015).

[c] Factor raised to the power of age class difference, in the form $a^{\text{age}_i} - \text{age}_j$.

[d] Initial value later calibrated to retrieve the observed number of diagnosed cases from surveillance data.

[e] Corresponding to baseline scenario where transmission varies by stage. In equal-rates simulation scenario weights $w_{s1}$ to $w_{s5}$ are all equal to 1.

#### 2.3.1. Compartmental epidemic model

The epidemic history representing London MSM was modeled with a compartmental model that captures disease progression by a system of ordinary differential equations determining transmission and transition through 5 stages of infection, 4 age groups and 3 diagnosis states (undiagnosed, diagnosed untreated and diagnosed under treatment). The 5 stages of infection corresponded to early HIV infection (stage 1) and 4 stages of declining CD4 as detailed in Table 1 (Cori et al., 2015). Individuals were further stratified in two risk categories influencing transmission. The population was thus structured in 120 states. Fig. S1 shows a subset of the transition flow between stages of infection and diagnosis states, omitting the transition between age groups that similarly affects all compartments and risk categories between which there is no transition. Furthermore, we modeled importation of infections into the population, which can have a dramatic effect on HIV genetic diversity.

### 2.3.2. Model parameters

Mean time of progressions to CD4 stages and proportion in each CD4 category after seroconversion were obtained from Cori et al. (2015). Transmission was allowed to vary according to weights provided by risk category, treatment status and according to age assortativity. A proportion of 20% of the population were deemed to be at high risk with a ten-fold increase in transmission than low risk counterparts. Relative to undiagnosed individuals, diagnosed and treated patients had a reduction in transmission by respectively a factor 2 and 20. An age assortativity parameter was introduced in the transmission matrix which caused transmission rates to decrease as a power law function of the difference in age (cf Table 1). Age groups were based on quantiles of observed age distribution of MSM diagnosed with HIV in London (HIV, 2014) and transmission rates were independent of age.

Two variations in this simulation were explored in terms of how transmission rate varies with time since infection in order to evaluate rates of false-positive identification of transmission risk factors. In a ‘baseline scenario’, we let infection stage influence probability of transmission in early HIV infection (ten-fold increase) and AIDS stage (three-fold increase) relative to chronic infection (stages 2–4). In an ‘equal-rates scenario’, transmission was independent of infection stage. Expressed mathematically, the total transmission rate of a patient with CD4 stage $i$, continuum of care status $j$, and generic risk factor $k$ is $\lambda_{ijk}(t) \propto r_i r_j r_k$, where $r_.$ are risk ratios for each category. Individual transmission rates are normalised so that total incidence is given by $\iota(t)$ based on a previous study (Phillips et al., 2013) assuming that dynamics of new infections in MSM was the same at the country level and in London.

Incidence and diagnosis rates were modeled as logistic functions of time and jointly calibrated to match the number of MSM living with diagnosed HIV in London in 2012 (Yin et al., 2014). Rates of treatment were modelled as zero before 1995 and then increase according to a logistic function with maximum 1 and steepness 0.5. Parameter values are summarized in Table 1.

### 2.3.3. Coalescent tree simulation

We simulated coalescent trees by conditioning on HIV epidemic histories using the approach described by Volz (2012). This method is implemented in the *phydynR* R package (Volz, 2016b). The simulated tree genealogy assumes that each infected patient corresponds to a single lineage of virus HIV-1 (Joseph et al., 2015), ignoring super-infection, and that the time at which two lineages coalesce corresponds to a transmission event. This approximation is reasonable if within-host evolution generates coalescence time considerably shorter than at the population epidemic level. Coalescent simulation is based on a CTMC model with time-dependent rates which describes the time evolution of the states of lineages. The rate of coalescence for a pair of lineages depends on reconstructed states and underlying transmission rates in the epidemic simulations. Further details can be found in Volz (2012) and Volz and Frost (2013). Coalescent simulations also condition on the times and states of sample lineages. These were chosen to match the times of sampling in the UK resistance database and associated ages and CD4-stages of patients (UK HIV Drug Resistance Database, 2016; Yin et al., 2014). Trees comprised 12,164 taxa, corresponding to the number of MSM patients diagnosed with HIV-1 subtype B between 1979 and the end of 2012 in London with at least one partial pol gene sequence available in the database. One hundred trees were simulated for both the baseline and equal-rates scenarios.

Branch lengths for coalescent trees are in calendar year. To apply clustering algorithms based on genetic distance, the number of nucleotide substitutions was simulated with a Langley-Fitch model (Langley and Fitch, 1974). Branch lengths in substitution per site were estimated as a Poisson distributed variable centred on branch length estimates in years multiplied by a substitution rate of $1.8 \times 10^{-3}$ per site per year. All code used to simulate epidemic histories and genealogical trees is available online (Volz, 2016a).

## 2.4. Statistical analysis

We used statistical models that have been commonly employed to study how phylogenetic cluster characteristics depend on one or more individuals covariates. We considered both univariate models and multivariate models that adjust for stage of infection at time of sampling using CD4 data. Non-parametric Wilcoxon test was used for univariate comparison of transmission by risk level. Linear regression models were used to examine the association between cluster size or out-degree (as dependent variable $Y$) and patient covariates in the form: $Y_i = \beta X_i$, with $X_i$ comprising age, risk level, and infection stage as both an independent variable and interaction term with age. Logistic regression models were used to examine how the probability of being into a cluster with at least two members vary by patient covariates, in the form $\text{logit}(p_i) = \beta X_i$. In regression models, out-degree and cluster sizes were standardized into dimensionless quantities by subtracting population mean and dividing by standard deviation. For each simulation scenario, we quantified the number of simulation replicates where the null hypothesis of no association between covariates and dependent variables was rejected.

The transmission model comprised 4 categories of age determined by quartiles of age at diagnosis of MSM in London with at least one available virus sequence. For each simulation replicate, an age mixing matrix $e_{ij}$ was computed by cumulating the number of common cluster or neighborhood pairwise memberships for each pair of age categories. For SA statistics, we calculated the sum of infector probabilities from donors of each age category to recipients of each category. Assortativity matrices by age were computed as the difference between these age matrices and a null expectation under random linking. Age assortativity was quantified by Newman's assortativity coefficient which summarizes the extent to which links between age groups differ from random mixing (Newman, 2003): $r = (\Sigma_i e_{ii} - \Sigma_i a_i b_i)/(1 - \Sigma_i a_i b_i)$, where $a_i = \Sigma_j e_{ij}$ and $b_j = \Sigma_i e_{ij}$.

Since age is often found associated with cluster characteristics, we studied the association between cluster sizes or out-degrees as dependent variable and age categories as independent variable of a linear regression. We also introduced a variable for stage of infection and its interaction term with age to test if an independent effect of age remained after adjustment. Note that in our simulation model, there is no association between age group and transmission rates. All above analyses were performed using R Statistical Software (R Core Team, 2015).

# 3. Results

## 3.1. Detecting the difference in transmission by risk level

We compared the ability of source attribution and clustering methods to detect the difference in transmission rates by risk level. Fig. 1a shows that out-degrees (i.e. estimated number of attributable transmissions) are significantly larger for the high risk category, whereas cluster size is not associated with level of risk. When testing for a difference on each baseline simulation replicate, we found that an univariate analysis would correctly detect significantly larger values of out-degree in 95% of experiments. Analysis of cluster sizes led to the corresponding figures of maximum 89% for the lowest 0.5% threshold and dropping to 17% at 1.5% threshold. These results for respective methods and distance thresholds are illustrated by the distribution of p-values for 100 experiments in Fig. 2 and percentage of errors in Table 2.

When controlling for the stage of infection, the proportion of tests correctly detecting the difference in transmission rates decreased for all methods. Specifically it was 52% when considering out-degrees and a maximum of 16% for cluster sizes at the lowest distance threshold.

In multivariate logistic regressions, cluster membership could be

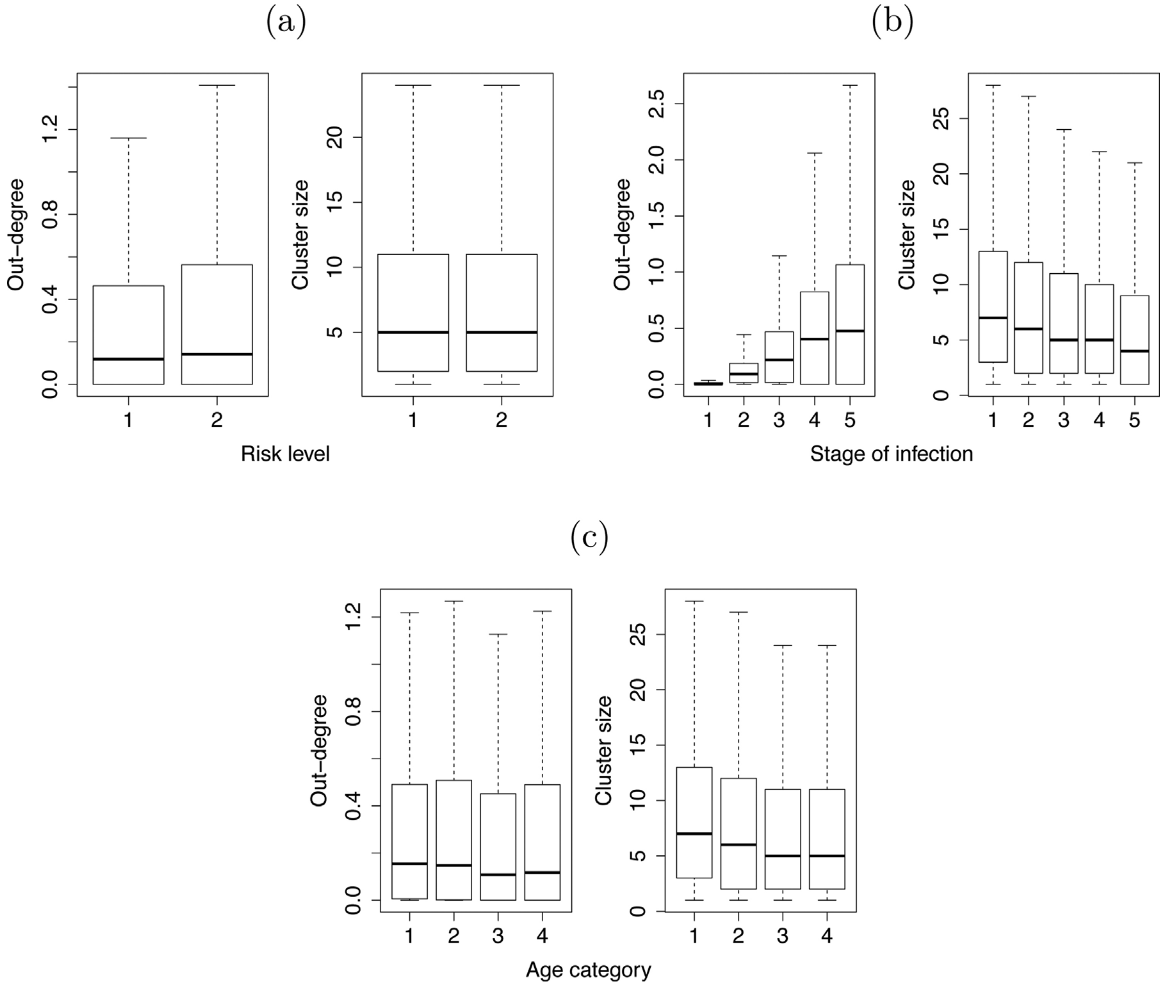

**Fig. 1.** Distribution of out-degrees and cluster sizes by (a) risk level: transmission rate was defined in the model as 10 times higher for risk level 2 relative to risk level 1; (b) stage of infection: relative transmission rates in the model were respectively 10, 1, 1, 1 and 3 for stages 1–5 of infection; (c) age category: transmission rates were equal for all 4 age categories in the model. Values are aggregated from 100 simulation replicates. Outliers are not shown. Distance threshold for clustering algorithm is 1.5%.

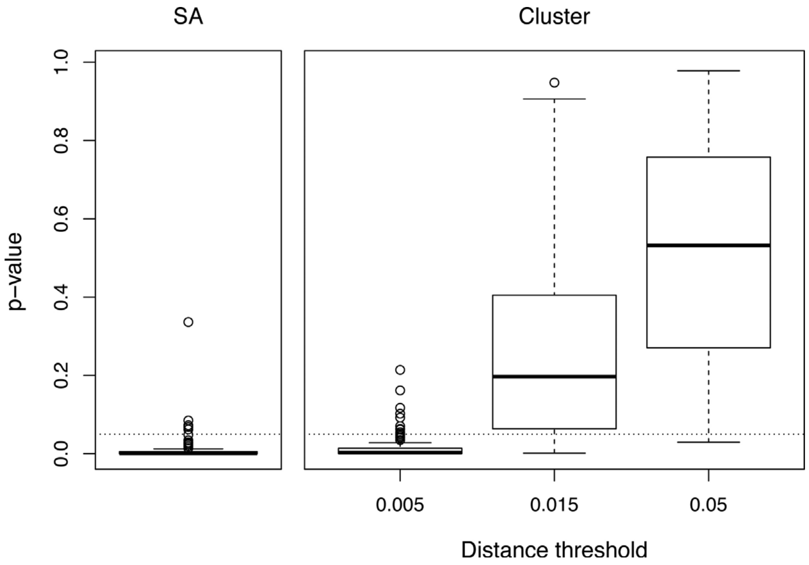

**Fig. 2.** Distribution of *p*-values of univariate test of difference in out-degree or cluster size by risk level. Values are aggregated from 100 simulation replicates. Dotted line indicates *p*-value = 0.05.

detected as an independent predictor of risk level in respectively 94, 56 and 15% of simulations with threshold 0.5, 1.5 and 5%, with an average odds-ratio of 1.19, 1.10 and 1.06.

**Table 2**
Percentage of error of source attribution (SA) and clustering methods at detecting heterogeneous transmission rates.

| | Type of error[a] | SA | Clustering | | |
|---|---|---|---|---|---|
| Analysis | | | 0.5% | 1.5% | 5.0% |
| Risk level[b] | | | | | |
| Unadjusted | II | 5 | 11 | 83 | 98 |
| Adjusted for stage of infection | II | 48 | 84 | 86 | 81 |
| Stage of infection[c] | | | | | |
| Equal-rate scenario | I | 17 | 100 | 100 | 100 |
| Baseline scenario | II | 19 | 0 | 0 | 0 |
| Age category[d] | | | | | |
| Unadjusted | I | 8 | 75 | 86 | 63 |
| Adjusted for stage of infection | I | 0 | 18 | 25 | 93 |

[a] Type I error corresponds here to falsely associating a variable to an increased transmission (false positive) and type II error corresponds to not detecting a true difference in transmission (false negative). Values reported are % of simulations leading to an erroneous outcome.

[b] Individuals allocated in high-risk category had a ten-fold increase in transmission rate. This allocation was completely random and had no dependance on stage of infection or other clinical variables. Values correspond to the analysis of transmission rate ratio (see Section 3.2 in the main text).

[c] In the equal-rate scenario, transmission was independent of infection stage. In the baseline scenario, transmission rates was increased ten-fold in early HIV infection and three-fold in AIDS stage.

[d] There was no association between age category and transmission rates.

### 3.2. *Inferring difference in transmission rates by stage of infection*

Next, we compared the outcomes of the two methods by stage of infection. For the SA method, Fig. 3 shows the 95% confidence intervals of relative difference and ratio in transmission rates between early and late stage of infection in 100 simulations of equal-rate scenario (left) and baseline scenario (right). We estimate that the SA method would fail to detect the heterogeneous transmission rates we introduced in the baseline scenario in 19% of the simulations (type II error) and would falsely detect a difference in the equal-rate scenario in 17% of the simulations (type I error) when estimating rate ratio and 19% when estimating rate difference. However, while SA method generally detected the difference in baseline scenario, in Fig. 3 (right) we see that it underestimates the actual transmission rate difference and ratio between early and late stage of infection.

For clustering method, since larger cluster sizes were always found in individuals at earlier stages of infection (cf. Fig. 1b), substituting out-degree by cluster size in all previous analyses led to the same association between earlier stages and increased transmission. This resulted in a 100% type I error in the equal rates scenario, and a 0% type II error in the baseline scenario.

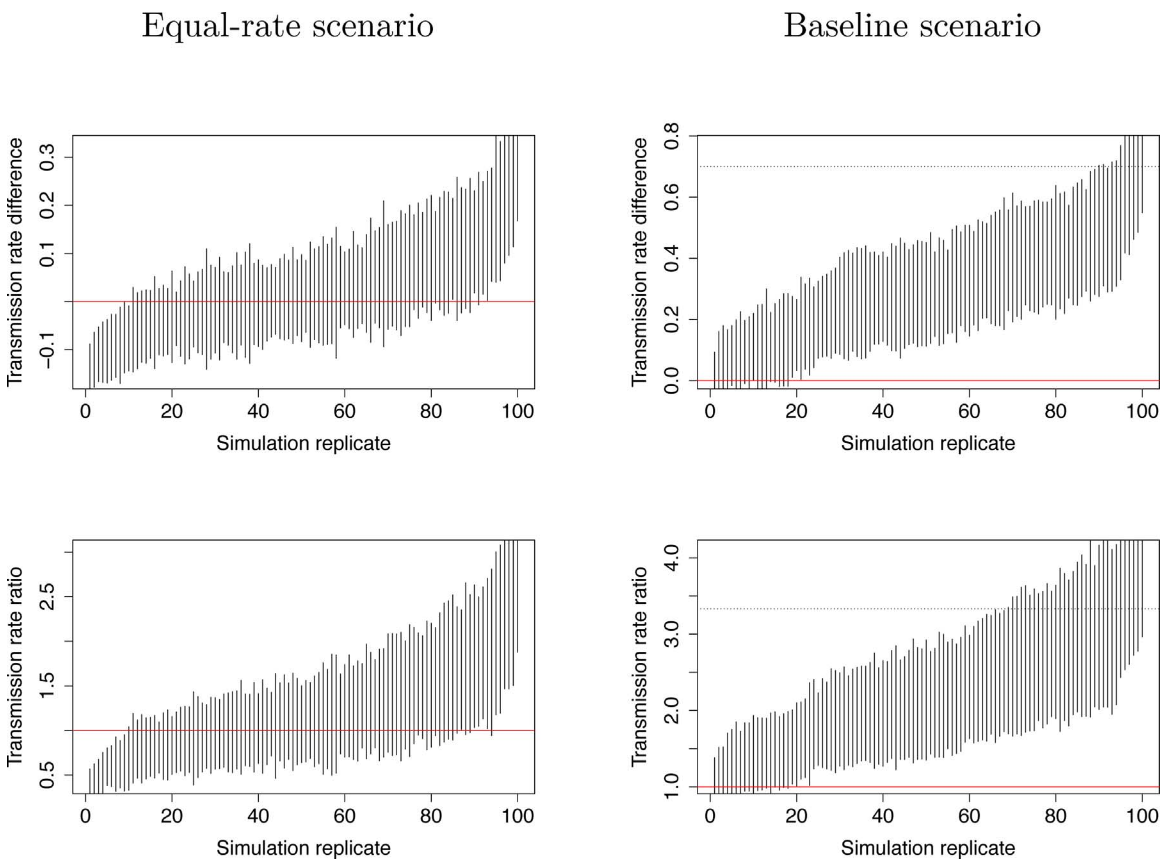

**Fig. 3.** Confidence intervals of difference and ratio of transmission rates between early and late stages of infection estimated by source attribution. The results of 100 simulation replicates are sorted in increasing order of the median of rate difference or ratio (*x*-axis). The plain red line corresponds to the null-hypothesis of no difference in transmission rate by stage. First row presents estimates of rate difference and second row rate ratio. Left column shows results for the 'equal-rate' scenario so that confidence intervals crossing the red line indicate true negative results. Right column shows results for the 'baseline' scenario where confidence intervals not crossing the red line correspond to true positive results. In this 'baseline' scenario true values of rate difference (top-right) and rate ratio (bottom-right) are indicated by a black dotted line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

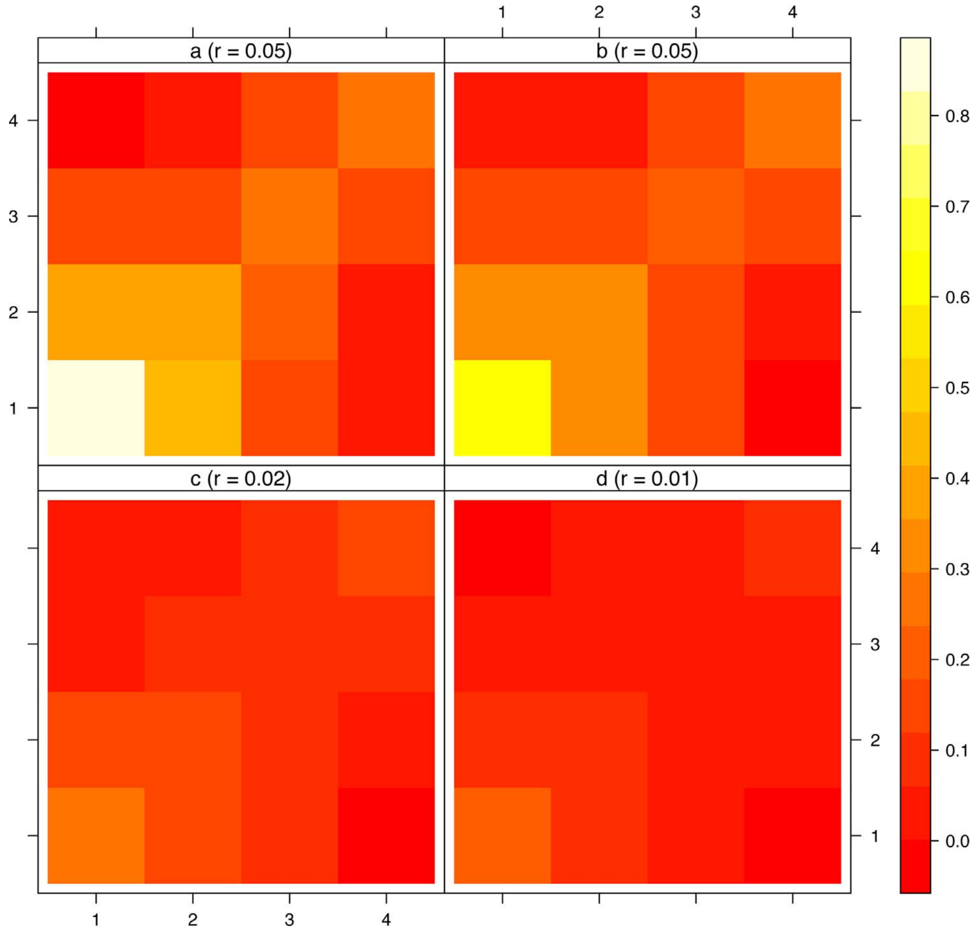

**Fig. 4.** Age assortativity matrices and coefficient from 100 simulations by method. Panel a: infector probabilities; panels b to d: cluster size at 0.5%, 1.5% and 5% thresholds. Labels of $x$ and $y$ axes represent age categories. $r$ values are Newman assortativity coefficients.

### 3.3. Age assortativity and relation to cluster characteristics and out-degrees

Next, we studied if outcomes of the two methods could reflect the preferential mixing by age and the independence between age and transmission rates. Assortativity matrices in Fig. 4 show some level of age assortativity both for clustering and source attribution methods. The larger assortativity is seen for the younger category of age. Estimated levels of assortativity are decreasing as increasing distance thresholds are chosen for the clustering method.

Fig. 5 shows the distribution of estimated age assortativity coefficient for respective methods and the true level of the Newman's coefficient ($r$ = 0.31) as a result of the parameterization in the transmission model. At 0.5% threshold, clustering method allows an estimation slightly closer to the true value and with significantly higher variance than SA. When testing increasingly lower thresholds, we found that estimates became largely imprecise and central value plateaued at

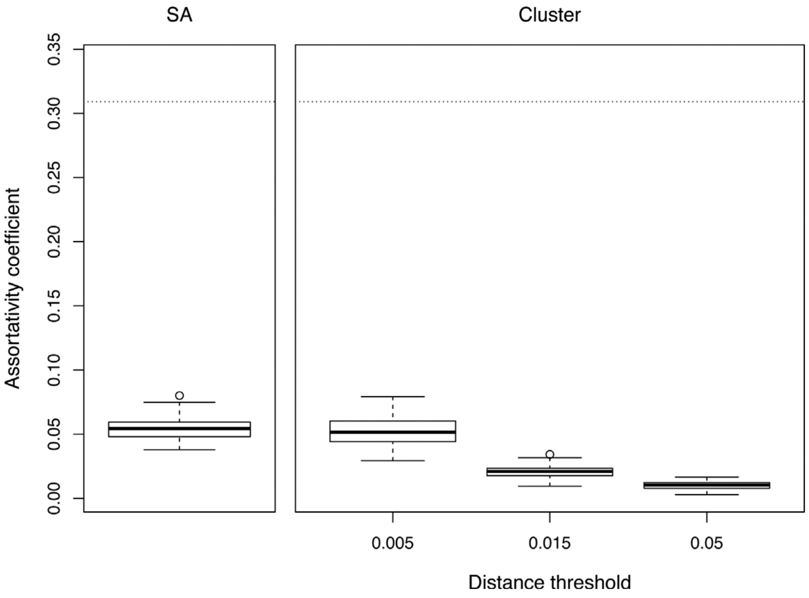

**Fig. 5.** Distributions of age assortativity coefficient by method. Values are aggregated from 100 simulation replicates. Dotted line indicates the true level of assortativity coefficient ($r$ = 0.31).

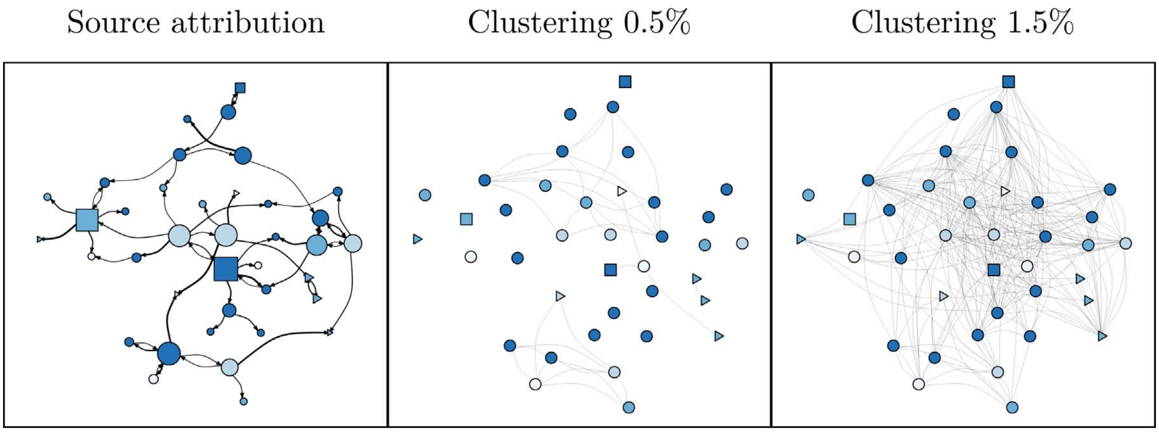

**Fig. 6.** Comparison of source attribution (left panel) and threshold distance clustering (middle and right panel) applied to one same cluster from simulated coalescent tree. The initial sample comprised 62 individuals forming a cluster at threshold 5%. The node positions, colours, and shapes are the same for all networks. For the SA graph, width of links is proportional to square root of the infector probability. Infector probabilities < 0.1% are not shown. Colour represents age category at time of diagnosis (darker colours represent older patients). Triangle nodes represent patients in early HIV infection, circles represent chronic stages and square nodes represent AIDS stage at time of diagnosis. Node size is proportional to square root of out-degree. For the clustering graphs, the single-linkage algorithm was re-applied to the sample with a genetic distance threshold of 0.5% or 1.5%.

$r = 0.12$ (not shown). Both methods greatly underestimate the true value of the assortativity coefficient.

Linear regressions between out-degree and age produced a statistically significant association (type I error) in 8% of analyses using the unadjusted model and in 0% when controlling for stage of infection (cf. Table 2). For a typical threshold of 1.5%, we found that cluster size decreased significantly with age in respectively 86% of the simulations for the unadjusted models and 25% for models controlling for the stage of infection.

### 3.4. Variation of clustering algorithms

In addition to clustering based on patristic distances (with *hivclustering*), the two other clustering algorithms (*neighborhood* and *tMRCA*) gave very similar results. The correlation coefficients between cluster sizes obtained by respective methods were between 77 to 87% but correlation with out-degrees from source attribution was only 9% (Fig. S2). As in Fig. 1, Figs. S3 and S4 show that the same associations between cluster sizes and individuals risk level, stage of infection and age were found for all clustering methods tested.

### 3.5. Network representations

Fig. 6 shows the respective transmission networks we obtained in applying source attribution (left panel) and clustering with a 0.5% (middle) and 1.5% threshold (right) to the same simulated sample of patients differentiated by age and stage of infection. This example illustrates the contrasting information provided when considering the probability of potential transmission to others (proportional to the width of the directional links) for any patient and the number of links he has in the cluster, that would correspond to his neighborhood size. The network from source attribution method is also showing that patients in later stages of infection have a larger number of attributable transmissions.

## 4. Discussion

Our simulation experiments show that detection of heterogeneous transmission is generally estimated with less precision using clustering methods than using source attribution. However, both methods underestimate the true level of assortative mixing.

Our clustering algorithms consistently produce a misleading result in associating younger age categories and cluster sizes when there is no difference in transmission by age. They also indicate a negative correlation between stage of infection and cluster sizes, even though cumulative number of transmissions of a patient is positively related to both its age and its progression in the course of infection. This is because these variables are correlated with the time since infection and clusters are more likely to be observed for recently infected patients. The results of multivariate analyses suggest that even when adjusting for a direct correlates of time since infection like CD4 staging, regression models of phylogenetic cluster characteristics would frequently lead to a false positive association with age. Nevertheless, including CD4 in multivariate analyses improved the performance of clustering analyses. Furthermore, for our purpose of detecting transmission risk factors, we found that much smaller clustering thresholds than are typically used (< 1.5%) minimized the type I error and increased detection power. But this could be only applicable when the sampling fraction of the infected population is sufficiently large to continue to observe related sequences as thresholds are decreased (Volz et al., 2012; Hassan et al., 2017).

By studying counterfactual scenarios of transmission variation by stage of infection, we confirm that inferring early stage infectivity from the characteristics of phylogenetic clusters is also potentially misleading (Volz et al., 2012). Our results also indicate that source attribution method has a greater power than clustering methods to detect that a characteristic of infected individuals is truly correlated with the risk of transmission. However, the stage of infection is confounding this relation and neither methods are able to capture the magnitude of the actual difference in transmission rates in the risk level variable. Our evaluation of clustering performance leads to the recommendation to always adjust for time since infection when testing for associations with transmission risk factors.

The clustering methods we used and their interpretation do not cover all the range of previous applications such as revealing sexual network structure (Leigh Brown et al., 2011; Grabowski and Redd, 2014) or detection of outbreaks (Poon et al., 2016). In transmission risk analyses, size of clusters has not been frequently used as a continuous value, but there are several published examples where a typology of small, intermediate or large clusters is used to interpret transmission networks (Brenner et al., 2011; Aldous et al., 2012; Junqueira et al., 2016). The use of such arbitrary complex cluster definition is questionable as it can force the interpretation of clustering patterns.

Although we used different clustering algorithms, our results may not generalize to all clustering methods. Our simulated genealogies were not obtained by inferring a phylogeny from sequence data, therefore we did not used bootstrap support as a cluster defining factor, as is common in the literature (Hassan et al., 2017). However, there is no reason why approaches starting by inferring a true genealogy would yield to different clustering patterns in relation to transmission. Indeed the simulation results from Poon (2016) indicate that various clustering methods generally failed to identify a subgroup with higher transmission rates. Moreover, methods harnessing bootstrap credibility in tree topology did not performed better than distance-based methods, including the 'patristic' method that is closely related to the method presented here.

Several alternatives to clustering analysis exist which are theoretically grounded and which have been shown to work well for transmission risk estimation, however their uptake is hampered by their additional complexity and computational cost. One approach is to make use of coalescent theory (e.g. Volz et al., 2013). These models provide a mathematical description of a phylogeny generated by a given epidemiological process. A related approach is the sampling birth-death model which can account for additional stochasticity in the epidemic history if sampling rates are known (Stadler, 2009). Both approaches can provide conditionally unbiased estimates of transmission rates given an exact time-scaled pathogen phylogeny and a correctly-specified epidemiological model. But, such approaches are also more difficult to implement and require additional effort to develop and compare epidemiological models.

The SA method presented has a computational burden similar to that of a tree-based clustering analysis but it accounts for incomplete sampling of infected cases, even with a weak prior epidemiological information. While we show that this SA method has acceptable properties for detecting transmission risk factors, neither clustering nor SA methods provide unbiased estimates of transmission risk ratios. Formal population genetic modeling (Volz et al., 2013) should be favoured if the aim is to estimate unbiased risk ratios for transmission risk factors.

Phylogenetic clustering analyses has served as a staple method for molecular epidemiological analysis of large pathogen sequence data due to its ease-of-use and computational tractability, but it has numerous shortcomings: clustering analyses rely on ad-hoc distance thresholds that must be chosen by the practitioner and are difficult to calibrate. Because there is no universal standard definition of a cluster, such analyses are prone to misinterpretation, and there is a danger that clustering thresholds will be chosen to demonstrate an effect rather than as a critical test of a hypothesis. Clustering analysis has extremely high type I error rates for any variable correlated with time since infection. And clustering analyses lose power by giving zero weight to all observations above the chosen genetic distance threshold. Recent advances in source attribution methods promise to alleviate these drawbacks, however further progress in this field is also required. Notably, within-host evolution is rarely taken into account, assuming coincidence of coalescent and transmission events. Advances in deep sequencing technologies promise to increase the fidelity of transmission pair identification (Romero-Severson et al., 2016) such as by using minority variants in a putative donor and recipient, but the standard form of data in resistance databases continues to be a single HIV partial *pol* sequence. To obtain robust estimates of transmission rates in the presence of incomplete sampling, there is currently no shortcut to doing formal population genetic modeling.

## Acknowledgements

We thank the UK Collaborative Group on HIV Drug Resistance for providing us with the surveillance data used to calibrate the simulated samples. We thank the Imperial College High Performance Computing Service (doi: 10.14469/hpc/2232). This work was supported by the National Institute for Health Research (NIHR) Health Protection Research Units in Modeling Methodology and Sexually Transmitted Infections (HPRU-2012-10080). E.M.V. is supported by the National Institutes of Health (R01AI087520). O.R. and C.F. are supported by Bill & Melinda Gates Foundation: Phylogenetics Networks to Address Transmission of HIV (OPP1084362). A.T. is supported by UK HIV Drug Resistance Database grant from the Medical Research Council (164587).

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.epidem.2017.10.001.

## References

Aldous, J.L., Pond, S.K., Poon, A., Jain, S., Qin, H., Kahn, J.S., Kitahata, M., Rodriguez, B., Dennis, A.M., Boswell, S.L., Haubrich, R., Smith, D.M., 2012. Characterizing HIV transmission networks across the United States. Clin. Infect. Dis. 55 (8), 1135–1143. http://dx.doi.org/10.1093/cid/cis612.

Brenner, B.G., Roger, M., Stephens, D., Moisi, D., Hardy, I., Weinberg, J., Turgel, R., Charest, H., Koopman, J., Wainberg, M.A., t. M. P. C. S. Group, 2011. Transmission clustering drives the onward spread of the HIV epidemic among men who have sex with men in Quebec. J. Infect. Dis. 204 (7), 1115–1119. http://dx.doi.org/10.1093/infdis/jir468.

Cori, A., Pickles, M., van Sighem, A., Gras, L., Bezemer, D., Reiss, P., Fraser, C., 2015. CD4+ cell dynamics in untreated HIV-1 infection: overall rates, and effects of age, viral load, sex and calendar time. AIDS (London, England) 29 (18), 2435–2446. http://dx.doi.org/10.1097/QAD.0000000000000854.

Csardi, G., Nepusz, T., 2006. The igraph software package for complex network research. InterJ. Complex Syst. 1695.

Dennis, A.M., Hué, S., Hurt, C.B., Napravnik, S., Sebastian, J., Pillay, D., Eron, J.J., 2012. Phylogenetic insights into regional HIV transmission. AIDS (London, England) 26 (14), 1813–1822. http://dx.doi.org/10.1097/QAD.0b013e3283573244.

Frost, S.D.W., Pillay, D., 2015. Understanding drivers of phylogenetic clustering in molecular epidemiological studies of HIV. J. Infect. Dis. 211 (6), 856–858. http://dx.doi.org/10.1093/infdis/jiu563.

Grabowski, M.K., Redd, A.D., 2014. Molecular tools for studying HIV transmission in sexual networks. Curr. Opin. HIV AIDS 9 (2), 126–133. http://dx.doi.org/10.1097/COH.0000000000000040.

Hassan, A.S., Pybus, O.G., Sanders, E.J., Albert, J., Esbjörnsson, J., 2017. Defining HIV-1 transmission clusters based on sequence data. AIDS 31 (9), 1211–1222. http://dx.doi.org/10.1097/QAD.0000000000001470.

HIV and STIs in men who have sex with men in London, Tech. rep., Public Health England. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/357451/2014_09_17_STIs_HIV_in_MSM_in_London_v1_0.pdf.

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

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Comput. Surv. (CSUR) 31 (3), 264–323.

Joseph, S.B., Swanstrom, R., Kashuba, A.D.M., Cohen, M.S., 2015. Bottlenecks in HIV-1 transmission: insights from the study of founder viruses. Nat. Rev. Microbiol. 13 (7), 414–425. http://dx.doi.org/10.1038/nrmicro3471.

Junqueira, D.M., de Medeiros, R.M., Gräf, T., Almeida, S.E., 2016. Short-term dynamic and local epidemiological trends in the South American HIV-1B epidemic. PLOS ONE 11 (6). http://dx.doi.org/10.1371/journal.pone.0156712.

Langley, C.H., Fitch, W.M., 1974. An examination of the constancy of the rate of molecular evolution. J. Mol. Evol. 3 (3), 161–177. http://dx.doi.org/10.1007/BF01797451.

Leigh Brown, A.J., Lycett, S.J., Weinert, L., Hughes, G.J., Fearnhill, E., Dunn, D.T., UK HIV Drug Resistance Collaboration, 2011. Transmission network parameters estimated from HIV sequences for a nationwide epidemic. J. Infect. Dis. 204 (9), 1463–1469. http://dx.doi.org/10.1093/infdis/jir550.

Leitner, T., Albert, J., 1999. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proc. Natl. Acad. Sci. U. S. A. 96 (19), 10752–10757. http://dx.doi.org/10.1073/pnas.96.19.10752.

Lewis, F., Hughes, G.J., Rambaut, A., Pozniak, A., Leigh Brown, A.J., 2008. Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med. 5 (3), e50. http://dx.doi.org/10.1371/journal.pmed.0050050.

Little, S.J., Kosakovsky Pond, S.L., Anderson, C.M., Young, J.A., Wertheim, J.O., Mehta, S.R., May, S., Smith, D.M., 2014. Using HIV networks to inform real time prevention interventions. PLOS ONE 9 (6), e98443. http://dx.doi.org/10.1371/journal.pone.0098443.

Morgan, E., Nyaku, A.N., D'Aquila, R.T., Schneider, J.A., 2017. Determinants of HIV phylogenetic clustering in Chicago among young black men who have sex with men from the uConnect Cohort. JAIDS J. Acquir. Immune Defic. Syndr. 75 (3), 265–270. http://dx.doi.org/10.1097/QAI.0000000000001379.

Newman, M.E.J., 2003. Mixing patterns in networks. Phys. Rev. E 67 (2), 026126. http://dx.doi.org/10.1103/PhysRevE.67.026126.

Phillips, A.N., Cambiano, V., Nakagawa, F., Brown, A.E., Lampe, F., Rodger, A., Miners, A., Elford, J., Hart, G., Johnson, A.M., Lundgren, J., Delpech, V.C., 2013. Increased HIV incidence in men who have sex with men despite high levels of ART-induced viral suppression: analysis of an extensively documented epidemic. PLOS ONE 8 (2), e55312. http://dx.doi.org/10.1371/journal.pone.0055312.

Pines, H.A., Wertheim, J.O., Liu, L., Garfein, R.S., Little, S.J., Karris, M.Y., 2016. Concurrency and HIV transmission network characteristics among MSM with recent HIV infection. AIDS (London, England) 30 (18), 2875–2883. http://dx.doi.org/10.1097/QAD.0000000000001256.

Poon, A.F.Y., Joy, J.B., Woods, C.K., Shurgold, S., Colley, G., Brumme, C.J., Hogg, R.S., Montaner, J.S.G., Harrigan, P.R., 2015. The impact of clinical, demographic and risk factors on rates of HIV transmission: a population-based phylogenetic analysis in British Columbia, Canada. J. Infect. Dis. 211 (6), 926–935. http://dx.doi.org/10.1093/infdis/jiu560.

Poon, A.F.Y., Gustafson, R., Daly, P., Zerr, L., Demlow, S.E., Wong, J., Woods, C.K., Hogg, R.S., Krajden, M., Moore, D., Kendall, P., Montaner, J.S.G., Harrigan, P.R., 2016. Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. Lancet HIV 3 (5), e231–e238. http://dx.doi.org/10.1016/S2352-3018(16)00046-1.

Poon, A.F.Y., 2016. Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks. Virus Evol. 2 (2), vew031. http://dx.doi.org/10.1093/ve/vew031.

R Core Team, 2015. R: A Language and Environment for Statistical Computing.

Ratmann, O., van Sighem, A., Bezemer, D., Gavryushkina, A., Jurriaans, S., Wensing, A., de Wolf, F., Reiss, P., Fraser, C., A. observational Cohort, 2016. Sources of HIV infection among men having sex with men and implications for prevention. Sci. Transl. Med. 8 (320), 320ra2. http://dx.doi.org/10.1126/scitranslmed.aad1863.

Romero-Severson, E.O., Bulla, I., Leitner, T., 2016. Phylogenetically resolving epidemiologic linkage. Proc. Natl. Acad. Sci. U. S. A. 201522930. http://dx.doi.org/10.1073/pnas.1522930113.

Rose, R., Lamers, S.L., Dollar, J.J., Grabowski, M.K., Hodcroft, E.B., Ragonnet-Cronin, M., Wertheim, J.O., Redd, A.D., German, D., Laeyendecker, O., 2016. Identifying transmission clusters with cluster picker and HIV-TRACE. AIDS Res. Hum. Retroviruses 33 (3), 211–218. http://dx.doi.org/10.1089/aid.2016.0205.

Stadler, T., 2009. On incomplete sampling under birth–death models and connections to the sampling-based coalescent. J. Theor. Biol. 261 (1), 58–66. http://dx.doi.org/10.1016/j.jtbi.2009.07.018.

UK HIV Drug Resistance Database. http://www.hivrdb.org.uk/.

Volz, E.M., Frost, S.D.W., 2013. Inferring the source of transmission with phylogenetic data. PLoS Comput. Biol. 9 (12). http://dx.doi.org/10.1371/journal.pcbi.1003397.

Volz, E.M., Koopman, J.S., Ward, M.J., Brown, A.L., Frost, S.D.W., 2012. Simple epidemiological dynamics explain phylogenetic clustering of HIV from patients with recent infection. PLoS Comput. Biol. 8 (6), e1002552. http://dx.doi.org/10.1371/journal.pcbi.1002552.

Volz, E.M., Koelle, K., Bedford, T., 2013. Viral phylodynamics. PLoS Comput. Biol. 9 (3), e1002947. http://dx.doi.org/10.1371/journal.pcbi.1002947.

Volz, E.M., 2012. Complex population dynamics and the coalescent under neutrality. Genetics 190 (1), 187–201. http://dx.doi.org/10.1534/genetics.111.134627.

Volz, E.M., 2016a. London MSM Tree Simulator. https://github.com/emvolz-phylodynamics/londonMSM_tree_simulator.

Volz, E.M., 2016b. PhydynR – Coalescent Simulation and Likelihood for Phylodynamic Inference. https://github.com/emvolz-phylodynamics/phydynR.

Weaver, S., Kosakovsky Pond, S., 2016. Hivclustering. https://github.com/veg/hivclustering.

Wertheim, J.O., Leigh Brown, A.J., Hepler, N.L., Mehta, S.R., Richman, D.D., Smith, D.M., Kosakovsky Pond, S.L., 2014. The global transmission network of HIV-1. J. Infect. Dis. 209 (2), 304–313. http://dx.doi.org/10.1093/infdis/jit524.

Yin, Z., Brown, A.E., Hughes, G., Nardone, A., Gill, O.N., Delpech, V.C., contributors, 2014. HIV in the United Kingdom 2014 Report: Data to end 2013, Tech. rep., Public Health England, London. . https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/401662/2014_PHE_HIV_annual_report_draft_Final_07-01-2015.pdf.