View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Serveur académique lausannois

VIRUS EVOLUTION

Virus Evolution, 2018, 4(2): vey024

doi: 10.1093/ve/vey024

Research article

Inferring the age difference in HIV transmission pairs by applying phylogenetic methods on the HIV transmission network of the Swiss HIV Cohort Study

Katharina Kusejko,[1,2,*,†] Claus Kadelka,[1,2] Alex Marzel,[1,2] Manuel Battegay,[3] Enos Bernasconi,[4] Alexandra Calmy,[5] Matthias Cavassini,[6] Matthias Hoffmann,[7] Jürg Böni,[2] Sabine Yerly,[5] Thomas Klimkait,[8] Matthieu Perreau,[6] Andri Rauch,[9] Huldrych F. Günthard,[1,2] Roger D. Kouyos,[1,2,*,§] and the Swiss HIV Cohort Study,[‡]

[1]Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zürich, Rämistrasse 100, CH-8091 Zürich, Switzerland, [2]Institute of Medical Virology, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland, [3]Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, Petersgraben 4, CH-4031 Basel; University of Basel, Petersplatz 1, CH-4001 Basel, Switzerland, [4]Division of Infectious Diseases, Regional Hospital Lugano, Via Tesserete 46, CH-6900 Lugano, Switzerland, [5]Laboratory of Virology and Division of Infectious Diseases, Genève University Hospital, Rue Gabrielle-Perret-Gentil 4, CH-1205 Genève; University of Genève, 24 rue du Général-Dufour, CH-1211 Genève, Switzerland, [6]Division of Infectious Diseases, Lausanne University Hospital, Rue du Bugnon 46, CH-1011 Lausanne, Switzerland, [7]Division of Infectious Diseases, Cantonal Hospital St Gallen, Rorschacher Strasse 95, CH-9007 St. Gallen, Switzerland, [8]Molecular Virology, Department of Biomedicine, University of Basel, Petersplatz 10, CH-4051 Basel, Switzerland and [9]Clinic for Infectious Diseases, Bern University Hospital, Freiburgstrasse 18, CH-3010 Bern; University of Bern, Hochschulstrasse 6, CH-3012 Bern, Switzerland

*Corresponding authors: E-mails: katharina.kusejko@usz.ch (K.K.); roger.kouyos@usz.ch (R.D.K.)

†https://orcid.org/0000-0002-4638-1940

§https://orcid.org/0000-0002-9220-8348

‡The members of the Swiss HIV Cohort Study are listed in Acknowledgements section.

Abstract

Age-mixing patterns are of key importance for understanding the dynamics of human immunodeficiency virus (HIV)-epidemics and target public health interventions. We use the densely sampled Swiss HIV Cohort Study (SHCS) resistance database to study the age difference at infection in HIV transmission pairs using phylogenetic methods. In addition, we investigate whether the mean age difference of pairs in the phylogenetic tree is influenced by sampling as well as by additional distance thresholds for including pairs. HIV-1 *pol*-sequences of 11,922 SHCS patients and approximately 240,000 Los Alamos background sequences were used to build a phylogenetic tree. Using this tree, 100 per cent down to 1 per cent of the tips were sampled repeatedly to generate pruned trees ($N = 500$ for each sample proportion), of which pairs of SHCS patients were extracted.

© The Author(s) 2018. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

The mean of the absolute age differences of the pairs, measured as the absolute difference of the birth years, was analyzed with respect to this sample proportion and a distance criterion for inclusion of the pairs. In addition, the transmission groups men having sex with men (MSM), intravenous drug users (IDU), and heterosexuals (HET) were analyzed separately. Considering the tree with all 11,922 SHCS patients, 2,991 pairs could be extracted, with 954 (31.9 per cent) MSM-pairs, 635 (21.2 per cent) HET-pairs, 414 (13.8 per cent) IDU-pairs, and 352 (11.8 per cent) HET/IDU-pairs. For all transmission groups, the age difference at infection was significantly ($P < 0.001$) smaller for pairs in the tree compared with randomly assigned pairs, meaning that patients of similar age are more likely to be pairs. The mean age difference in the phylogenetic analysis, using a fixed distance of 0.05, was 9.2, 9.0, 7.3 and 5.6 years for MSM-, HET-, HET/IDU-, and IDU-pairs, respectively. Decreasing the cophenetic distance threshold from 0.05 to 0.01 significantly decreased the mean age difference. Similarly, repeated sampling of 100 per cent down to 1 per cent of the tips revealed an increased age difference at lower sample proportions. HIV-transmission is age-assortative, but the age difference of transmission pairs detected by phylogenetic analyses depends on both sampling proportion and distance criterion. The mean age difference decreases when using more conservative distance thresholds, implying an underestimation of age-assortativity when using liberal distance criteria. Similarly, overestimation of the mean age difference occurs for pairs from sparsely sampled trees, as it is often the case in sub-Saharan Africa.

Key words: HIV; phylogenies; age structure; sampling; cophenetic distance.

1. Introduction

Human immunodeficiency virus (HIV) infection and acquired immunodeficiency syndrome (AIDS) is a major health threat with approximately 1.8 million new HIV infections and 1 million AIDS-related deaths worldwide in 2016 (UNAIDS) (Fact sheet—Latest statistics on the status of the AIDS epidemic, 2017). Targeting public health interventions for the prevention of new infections in subpopulations at risk is therefore crucial to curb the epidemic. In this context, age mixing and its impact on HIV transmission was studied in different settings in the past. For example, in sub-Saharan Africa, the region that carries the highest burden of the HIV epidemic, there is evidence that older men infecting younger women drive the HIV epidemic (Ott et al., 2011; Oliveira et al., 2017; Schaefer et al., 2017). Many public health interventions therefore target young women, e.g., by teaching them cautiousness around so-called 'sugar daddies', i.e., older men. In the USA, African Americans carry a disproportionate burden of the HIV epidemic, in particular young men who have sex with men (MSM). Age patterns, in particular differences between black and white MSM, were analyzed by Grey et al. (2015). Black MSM exhibited a slightly more disassortative age mixing compared with white MSM, but this difference was too weak to explain the higher HIV prevalence of black MSM in their model. Doherty, Schoenbach, and Adimora (2009) investigated sexual mixing of heterosexual African Americans and found an overall strong assortativity with respect to illicit drug use and assortative mixing with respect to education and incarceration primarily for males. In Hurt et al. (2010), the age differences of the three most recent sexual partners of young MSM in the USA were used to quantify how the odds of acquiring HIV increase with the age of the sexual partners. In addition, a modeling study by Wilson (2009) showed that for Australian MSM, despite the increasing mean age of HIV-infected MSM, the epidemic is likely to be sustained due to frequent age-disparate mixing. Another Australian study, by Chow et al. (2016), shows that sexual mixing is assortative with respect to age and condom use in MSM and heterosexual relationships.

Most of the above-mentioned studies were based on questionnaires about the age of sexual partners of the study participants. Such studies heavily rely on correct reporting by the study participants, but also on the ability of estimating the age of the sexual partners correctly. Phylogenetic analysis of HIV sequences can overcome these potential biases introduced by incorrect reports. The underlying assumption of analyses of phylogenetic trees is that two patients whose HIV sequences are clustered in a tree share a social network or even form an HIV transmission pair. Calculating the mean of the absolute age differences in birth years of patients clustered in the tree gives hence information about the age differences at infection in the HIV transmission network. This method is, however, sensitive to the choice of certain parameters. Several drawbacks of phylogenetic cluster methods, such as the potential bias introduced by the time since infection, were pointed out in a simulation study by Le Vu et al. (2018). Novitsky et al. (2014) studied the impact of sample density on the proportion of HIV sequences in phylogenetic clusters. The performance of different phylogenetic methods in challenging, i.e., poorly sampled, settings was analyzed by Ratmann et al. (2017) based on simulated HIV-1 epidemics with a focus on recent transmission dynamics. Phylogenetic analyses of demographic and social patterns depend on the sample proportion of the whole population of people living with HIV (PLWH), on distance thresholds used for inclusion of clusters and of course the scientific question of interest itself. It is expected that any significant pattern detected in HIV transmission networks, e.g., clustering of patients of similar age or same ethnicity, will be underestimated if only few patients are sampled. With a small sample proportion, the phylogeny might not reflect the HIV transmission network well and the chances of obtaining HIV transmission pairs in the phylogeny are small. Inclusion of a large number of pairs, which form a pair in the phylogenetic tree only because the intermediate links of the transmission chain are not sampled, will therefore underestimate how strong patterns are pronounced in the HIV transmission network (see Fig. 1 for the underlying idea).

In our study, we use the Swiss HIV Cohort Study (SHCS) resistance database to analyze the age difference in pairs of patients in the HIV transmission network. In particular, we study the age difference at infection for the three most frequent transmission groups of HIV, namely in MSM, heterosexuals (HET), and intravenous drug users (IDU). Moreover, we use this dataset as an example to better understand the impact of sample proportion and distance thresholds on the age difference of pairs, measured by the difference in birth years, in the phylogenetic tree.

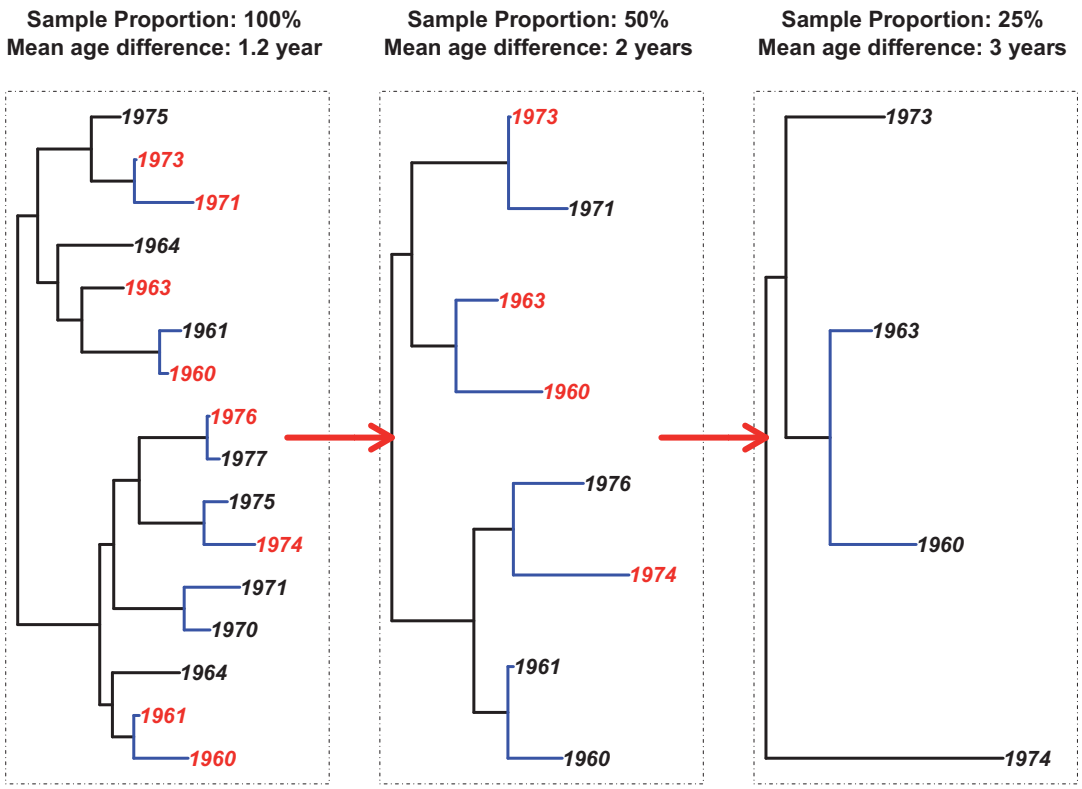

Figure 1. 'Heuristic' example of the possible impact of the sample proportion: we start with sixteen patients: the tips are labeled with the birth year. The left tree has sixteen tips and six pairs (in blue). For the middle tree, eight tips are randomly sampled from the left tree (the red tips). The middle tree has three pairs (in blue). For the right tree, four tips are randomly sampled from the middle tree (the red tips). The right tree has two pairs (in blue). For each tree, the mean age difference of the pairs is calculated: 1.2 years for the left tree, 2 years for the middle tree and 3 years for the right tree.

2. Methods

2.1 Swiss HIV Cohort Study

The SHCS is a prospective multicenter study including PLWH at the age of sixteen year or older in Switzerland and was launched in 1988. It is estimated that the SHCS covers at least 45 per cent of all PLWH and 69 per cent of all AIDS patients in Switzerland (Schoeni-Affolter et al. 2010). Baseline demographic information, such as birth year, gender, most likely route of HIV infection and ethnicity, is collected at study entry. Clinical and laboratory information, such as CD4 cell counts and HIV viral load, is collected in two to four follow-up visits per year. The genotypic-resistance-test database of the SHCS contains HIV-1 *pol*-sequences of 11,922 patients, which is 60 per cent of all patients enrolled up to 2016. Considering only patients enrolled between 1996 and 2016, the database contains at least one sequence for 77 per cent of the patients, due to considerable retrospective sequencing based on the bio bank. Combining the sample proportion of the SHCS of at least 45 per cent of the whole Swiss epidemic and the 60 per cent of SHCS patients with at least one sequence in the database, we can deduce that the sequences available in the SHCS cover at least 27 per cent of the whole Swiss HIV epidemic, again with considerably higher coverage for recent years. The SHCS further contains sequences for an estimated 69 per cent of MSM diagnosed between 1996 and 2009 in Switzerland (Drescher et al. 2014), and a recent study by Shilaih et al. (2016) showed a good coverage of hard-to-reach subpopulations suggesting no systematic exclusion of marginalized populations neither from the cohort nor from the sequence database.

2.2 The phylogenetic tree

For the construction of the maximum-likelihood phylogenetic tree, we included HIV-1 *pol*-sequences stored in the SHCS database. Sequencing was routinely performed for the *pol* region from the nucleotide positions 2,253–3,870 in the HIV genome. Only sequences with a minimal length of 250 nucleotides in the protease and a minimal length of 500 nucleotides in the reverse transcriptase were included into our analysis. If more than one sequence per patient was available, the earliest sequence was considered. In a first step, the sequences were aligned to the reference genome HXB2 (accession number: K03455.1). In addition, the SHCS sequences were compared with approximately 240,000 sequences from the Los Alamos database by using *Basic Local Alignment Search tool* (BLAST) (https://blast.ncbi.nlm.nih.gov/Blast.cgi). Los Alamos sequences with at least 90 per cent identity to an SHCS sequence, called hits, were included, but at most the 10 closest hits per SHCS sequence. These criteria led to the inclusion of 11,922 SHCS sequences and 11,390 Los Alamos sequences. The median coverage of the protease was 297 nucleotides and of the reverse transcriptase 1,005 nucleotides. The phylogenetic tree including the SHCS sequences and the Los Alamos hits was built with *FastTree* (Price, Dehal, and Arkin 2009), by using the generalized time-reversible model of nucleotide evolution and the CAT approximation for rate variation across sites. This approach of building a tree was

already verified and used in other SHCS projects (Bachmann et al. 2017; Turk et al. 2017).

2.3 Sampling from the tree

We used the phylogenetic tree containing sequences of the SHCS with included Los Alamos hits and constructed new trees by keeping only a certain percentage of the tips of the original tree and dropping the other tips. We call the resulting trees pruned trees, i.e., the trees after dropping tips as well as the corresponding edges that connected these tips to the rest of the tree. These pruned trees keep the topology of the original tree, with the desired tips removed, as illustrated in Fig. 1. The pruning procedure was performed by using the *drop.tip* function in the R-package *Analyses of Phylogenetics and Evolution* (ape) (Paradis, Claude, and Strimmer 2004). In total, we generated repeatedly 1, 2, 3, 4, and 5 per cent of the tips and steps of 5 thereafter, i.e. 10, 15, ... , 95 per cent, to generate pruned trees. For each fixed sample percentage, 500 pruned trees were sampled, resulting in $23 \times 500 = 11{,}500$ pruned trees, plus the original tree. For each tree, all clusters of size 2 (pairs) with both patients being SHCS patients were extracted and saved together with the corresponding cophenetic distance. For each sub-analysis, pairs with a cophenetic distance below the threshold of interest were used. The birth year, sex, and most likely route of HIV transmission (referred to as transmission group) of the patients were mapped on the tips of the tree. In addition, pairs were grouped by transmission, namely pairs with both patients being MSM (MSM-pairs), both patients being IDU (IDU-pairs), both patients being HET (HET-pairs) and pairs with one patient being HET and one patient being IDU (HET/IDU-pairs).

2.4 Cophenetic distance

The cophenetic distance between two tips from a phylogenetic tree is the sum of the branch lengths connecting the two tips (Sokal and Rohlf 1962). Different thresholds ranging from 0.01 to 0.05 of this distance were considered for transmission pairs. Pairs that exceeded the distance threshold were not included in the respective analysis.

2.5 Calculating the mean age difference

We used the absolute age difference measured as the age difference of the birth years of the patients in all cases. For each fixed sample proportion and each fixed distance threshold, the mean of the absolute age differences of the pairs of interest, i.e., either all or transmission group specific, was calculated separately for each of the 500 corresponding pruned trees and then averaged. We used the age difference by birth year, since this measure stays constant over time and is independent of sample date or age at infection of the patients. If, for example, Patient A was born in 1960, but diagnosed and sequenced in 2000 and Patient B born in 1970, diagnosed in 2005 and sequenced in 2008, their age difference will be $1970 - 1960 = 10$ years at any time point.

2.6 Analysis of the age difference at infection

To assess whether HIV transmission is more likely in pairs of similar age, we used random pairs as a comparison. For that, we randomly assigned pairs of patients and computed the resulting mean age difference of all pairs. This process was repeated one hundred times and the mean of the resulting mean age differences was used as the reference value. In addition, we wanted to assess by how much the age difference of pairs was overestimated when varying the distance threshold and the sample density, denoted by 'assortativity lost'. For that, we used the minimum age difference (min) and the maximum age difference (max) obtained by varying the distance threshold and sample percentage. In particular, we compare 'min' and 'max' with the mean age difference obtained by random pairing (random). The assortativity lost is then defined as: $1 - (\text{random} - \text{max})/(\text{random} - \text{min})$.

3. Results

3.1 Study population

Sequences of 11,922 SHCS patients were included in the phylogenetic tree. Of them, 8,554 (71.75 per cent) were males and 3,368 (28.25 per cent) were females. Moreover, 4,738 (39.74 per cent) of the patients were MSM, 4,246 (35.61 per cent) HET and 2,430 (20.38 per cent) IDU. The whole tree with all 11,922 patients contained 2,991 potential SHCS transmission pairs, i.e., clusters of size 2 with both tips belonging to SHCS patients. Of these, 954 (31.9 per cent) were MSM-pairs, 635 (21.23 per cent) HET-pairs, 414 (13.84 per cent) IDU-pairs, and 352 (11.77 per cent) HET/IDU-pairs. In addition, there were 310 (10.4 per cent) pairs with one patient being MSM and one patient being HET, 105 (3.5 per cent) pairs with one patient being MSM and the other patient being IDU and 221 (7.4 per cent) with at least one patient not belonging to one of the three main transmission groups MSM, HET, or IDU. For further analyses, we either looked at all pairs together or concentrated on the four epidemiologically most relevant categories of pairs, namely MSM-, HET-, IDU-, and HET/IDU-pairs.

3.2 Age structure and age difference at infection

The median birth year of all included patients was 1965 with a standard deviation of 11.3 years. MSM had the median birth year 1965 (SD =12.2 years), HET the year 1966 (SD=11.9 years), and IDU the year 1963 (SD=6.4 years) (see Fig. 2A). Without including a distance threshold, the mean age difference between all pairs was 9.1 years with a median of 7 years. IDU-pairs had the smallest age difference with a mean of 5.6 years (median=4 years), followed by HET/IDU-pairs with a mean of 7.2 years (median=5 years), MSM-pairs with a mean of 9.3 years (median=7 years), and HET-pairs with a mean of 9.6 years (median=7 years), visualized in Fig. 2B. The age difference observed in pairs in the tree was indeed significantly smaller compared with the average age difference of two patients on the tree, namely 9.1 years compared with 12.2 years ($P < 0.001$). Random reassignment of MSM-pairs led to a mean age difference of 13.2 years (13.1 years for HET-pairs, 7.1 years for IDU-pairs and 10.4 years for HET/IDU-pairs). For each considered category, the mean age difference was significantly smaller ($P < 0.001$) compared with randomly assigned pairs, indicating that HIV transmission occurs between people of similar age, for each transmission group. In Fig. 2C, we showed the number of pairs for each combination of birth years (grouped by 5 years) and in Fig. 2D we normalized the number of pairs for each combination of birth years by the number of patients born in these categories. With Fig. 2D, we could also visualize that the age difference in pairs in the phylogenetic tree is smaller compared with random pairs, i.e., the diagonal is darker than the off-diagonal.

3.3 Impact of distance criterion

The impact of the distance criterion on the mean age difference of the pairs was analyzed by only including pairs with a

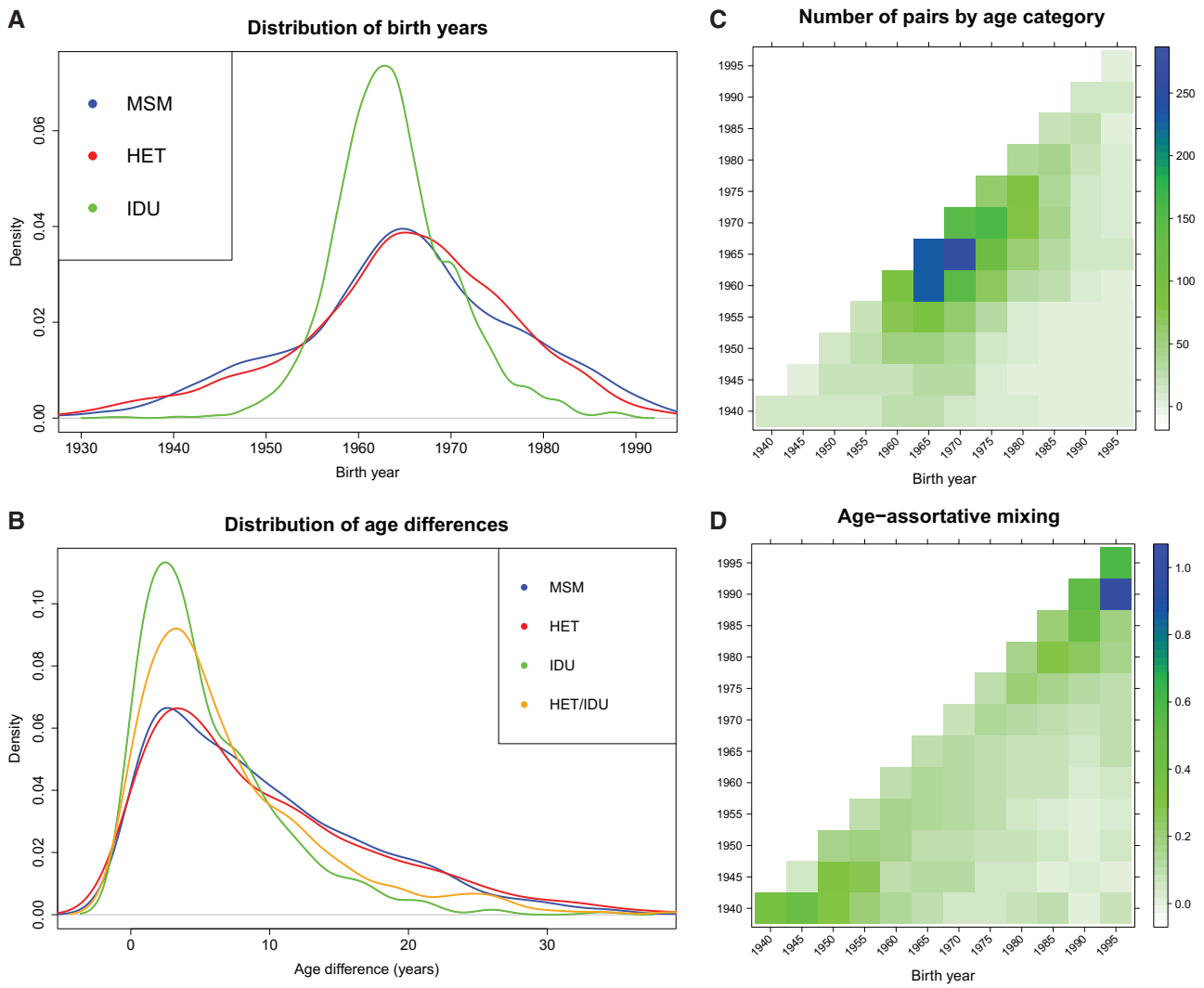

Figure 2. (A) Distribution of the birth years by route of HIV transmission: MSM, HET, and IDU; (B) distribution of age differences (in years) for the four categories of pairs: MSM-, HET-, IDU-, and HET/IDU-pairs; (C) number of pairs by birth year; (D) number of pairs by birth year, normalized by the number of patients in each birth year.

distance smaller than a given threshold, which varied from 0.01 to 0.05 (see Fig. 3). We observed a strong effect for HET-pairs, where the mean age difference ranged from 8.3 to 9 years. This presented a significant increase ($P < 0.001$) in age difference with a regression slope of 0.013 years per 0.01 increase of distance criterion. Similarly, the mean age difference for HET/IDU-pairs ranged from 6.1 to 7.4 years with a slope of 0.019 years per 0.01 increase of distance criterion ($P < 0.001$). For MSM-pairs, the mean age difference ranged from 8.7 to 9.2 with a significant ($P < 0.001$) slope of 0.011 years per 0.01 increase of distance criterion. For IDU-pairs, no significant correlation between the mean age difference and the distance criterion was observed (see Fig. 3).

3.4 Impact of sampling

The impact of sampling on the mean age difference of the pairs was analyzed by generating 500 pruned trees for various sample percentages between 1 and 100 per cent, and a liberal distance threshold of 0.05. Strong effects of the sample proportion were observed for HET-pairs with the mean age difference increasing from 9 to 10.3 years ($P < 0.001$), for MSM-pairs with an increase from 9.2 to 10.1 years ($P < 0.001$) and for IDU-pairs with an increase from 5.6 to 6.1 years ($P < 0.001$) (Fig. 4).

For HET/IDU-pairs, the mean age difference increased significantly ($P < 0.001$) from 7.3 to 7.9 years. The total difference in the mean age difference was rather small, e.g., 9.2 compared with 10.1 years for MSM-pairs. A relative comparison to the randomly expected mean age difference revealed, however, a noteworthy underestimation of the clustering of patient with similar age for too low sample proportions. In Table 1, we compared the mean age difference between two patients in the tree with the mean age difference obtained from pairs in the trees. We showed that the clustering of patient with similar age was underestimated by up to 45 per cent: for the full tree with a distance threshold of 0.01, we found that HET/IDU-pairs are 4.3 years younger compared with randomly assigned HET/IDU-pairs, but for the distance threshold of 0.05 in the 4 per cent pruned tree only 2.4 years younger.

Moreover, in Fig. 5, we show the impact of sampling given for varying distance thresholds. We see a trend that sampling has more impact, i.e., a larger difference in age, for more liberal distance thresholds.

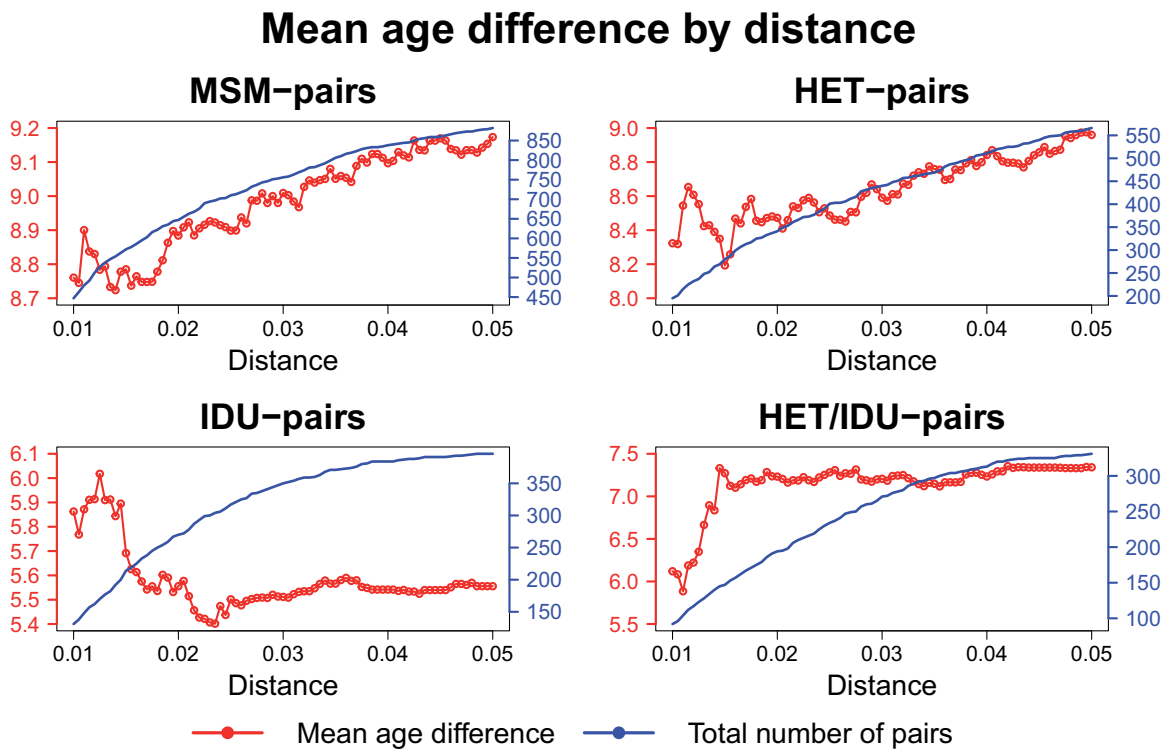

Figure 3. Analysis of the observed mean age difference of MSM-, IDU-, HET-, and HET/IDU-pairs for various distance thresholds. For each distance threshold, only pairs with a smaller distance are included. The number of pairs for each distance threshold is shown as well (in blue).

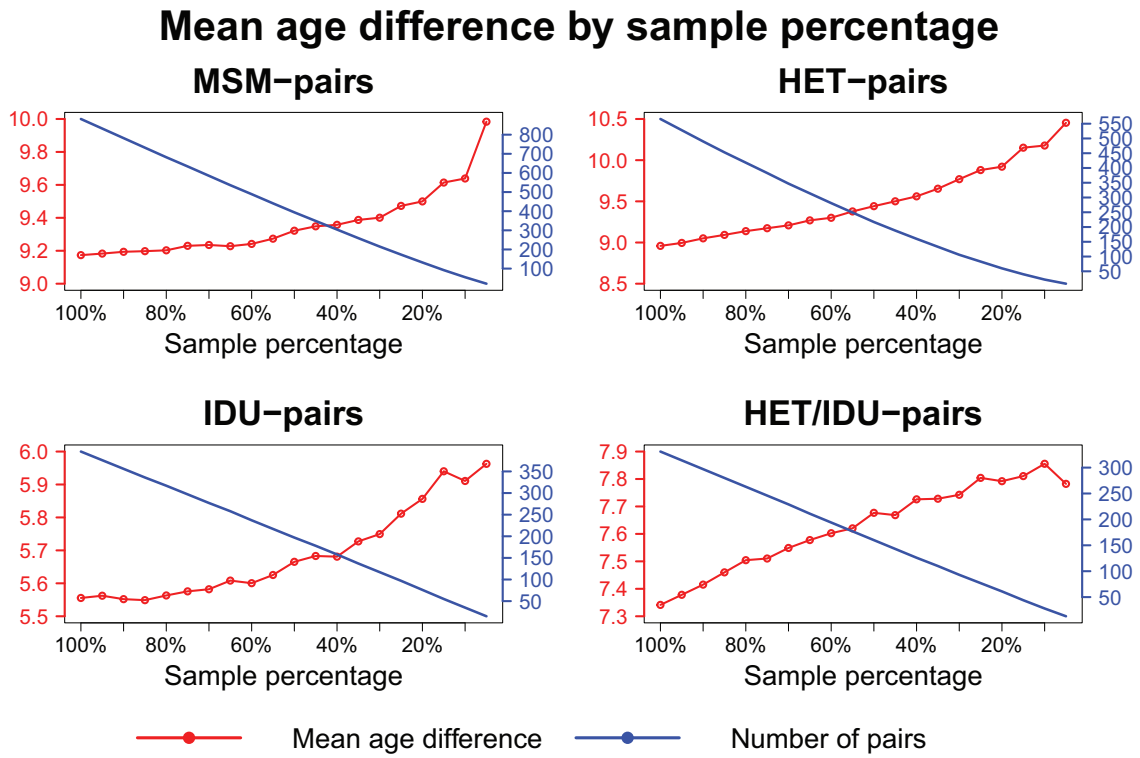

Figure 4. Analysis of the mean age difference of MSM-, IDU-, HET-, and HET/IDU-pairs for different sample proportions, averaging over 500 pruned trees for each sample proportion. The number of pairs for each sample proportion (again, averaged over 500 pruned trees) is shown as well (in blue).

Table 1. Comparison of the mean age difference (in years) for varying distance threshold and varying sample percentage, for all pairs, as well as stratified by transmission groups, including the expected mean age difference for randomly assigned pairs.

	ALL	MSM	HET	IDU	HET/IDU
Random	12.21	13.24	13.11	7.08	10.36
100 per cent sampling, distance					
≤0.05	8.7	9.2	9	5.6	7.3
≤0.04	8.6	9.1	8.8	5.5	7.2
≤0.03	8.4	9	8.6	5.5	7.2
≤0.02	8.4	8.9	8.5	5.6	7.2
≤0.01	8.3	8.8	8.3	5.9	6.1
Distance ≤0.05, sampling					
100 per cent	8.7	9.2	9	5.6	7.3
80 per cent	8.8	9.2	9.1	5.6	7.5
60 per cent	8.8	9.3	9.3	5.6	7.6
40 per cent	8.9	9.4	9.6	5.7	7.7
20 per cent	9	9.5	9.8	5.8	7.8
10 per cent	9.1	9.7	10.3	5.9	7.9
5 per cent	9.2	10	10.4	5.9	7.7
4 per cent	9.3	10.1	10.4	6	8
3 per cent	8.8	9.6	10.3	5.7	7.1
2 per cent	8.8	10	10.9	6	7.8
1 per cent	8.8	11.6	9.5	5.6	7.2
Assortativity lost	26 per cent	20 per cent	31 per cent	32 per cent	45 per cent

3.5 A closer look at HET-pairs

In the above analysis, we treated HET-pairs in the same way as MSM- and IDU-pairs in the sense that we considered two patients in the pair, regardless of their gender, to be interchangeable. While MSM-pairs are by definition pairs with both patients being male, all combinations of gender, i.e., male–male, male–female, and female–female, were considered in the analysis of IDU- and HET-pairs. In contrast to HET-pairs, all combinations of gender are plausible in IDU-pairs due to needle sharing. In the fully sampled tree without including a distance criterion, 414 (65.1 per cent) of the HET-pairs were male–female, but 140 (22 per cent) were female–female and 81 (12.8 per cent) were male–male pairs. In the male–female HET-pairs, the median birth year of male patients was 1963, 5 years earlier than the median birth year of female patients that was 1968. In 277 (66.9 per cent) of the male–female HET-pairs the male patient was older, in 116 (28.0 per cent) pairs the female patient was older and in 21 (5.1 per cent) pairs the two patients had the same birth year. While male–male HET-pairs could still be true transmission pairs due to incorrect report of sexual preference, sexual transmission of HIV between two female persons is very rare. Therefore, a large amount of female–female HET-pairs are most likely not real transmission pairs. The impact of the distance threshold and the sample density on the percentage of female–female HET-pairs is shown in Fig. 6. As expected, the percentage of female–female HET-pairs decreases with higher sample proportion and with stricter distance criterion.

4. Discussion

Phylogenetic analysis of HIV transmission is an efficient tool for obtaining a better understanding of the dynamics of the epidemic, but it needs to be executed with caution. In this study, we aimed to highlight the influence of the sample proportion of PLWH and distance threshold for genetically linked pairs on the mean age difference of pairs defined by the birth years of the patients in the pairs. Similar effects are, however, expected when extracting other traits by the same method, as could, for example, be done for the assortativity by ethnicity, the body-mass index or behavioral aspects such as smoking. All these scientific questions addressed by phylogenetic methods face the same underlying problem: the lower the sample proportion, the lower the probability of obtaining real transmission pairs or at least two patients who share indeed a social network. We use the SHCS resistance database, which is densely sampled with at least 27 per cent of sequences of the whole Swiss epidemic and even better coverage for recent years, to point out problems associated with phylogenetic analyses of demographic traits within HIV transmission networks. Exemplary, we use the mean age difference of pairs to demonstrate changes in the observed age difference at infection obtained by varying sample proportion and distance threshold of including pairs in the analysis.

In all transmission groups considered, i.e., MSM-, HET-, IDU-, and HET/IDU-pairs, we find that the age difference in pairs in the tree is slightly, but significantly, smaller compared with randomly assigned pairs. The mean age difference in MSM-pairs is around 9 years, which might be unexpectedly high, but is not implausible as studies by Hurt et al. (2010) and Morris, Zavisca, and Dean (1995) on young MSM in the USA identified young men who have sex with older men as the drivers of the epidemic. In particular, partners of young, primary HIV-infected MSM were on average 6 years older than partners of uninfected MSM of the same age class. One simple explanation of the high age difference in pairs of PLWH is certainly that the whole population of PLWH is ageing. Because of that, the odds of acquiring HIV when having an older partner is higher compared with having a younger partner (Morris, Zavisca, and Dean 1995; Hurt et al. 2010). On the other hand, some HIV transmissions might occur due to prostitution of young MSM with the client being much older, or the young MSM coming from a high prevalence country. We want to emphasize that the age difference calculated in this project reflects the age difference in MSM-pairs where an HIV transmission event happened, but does not necessarily reflect the sexual contact network in general. Moreover,

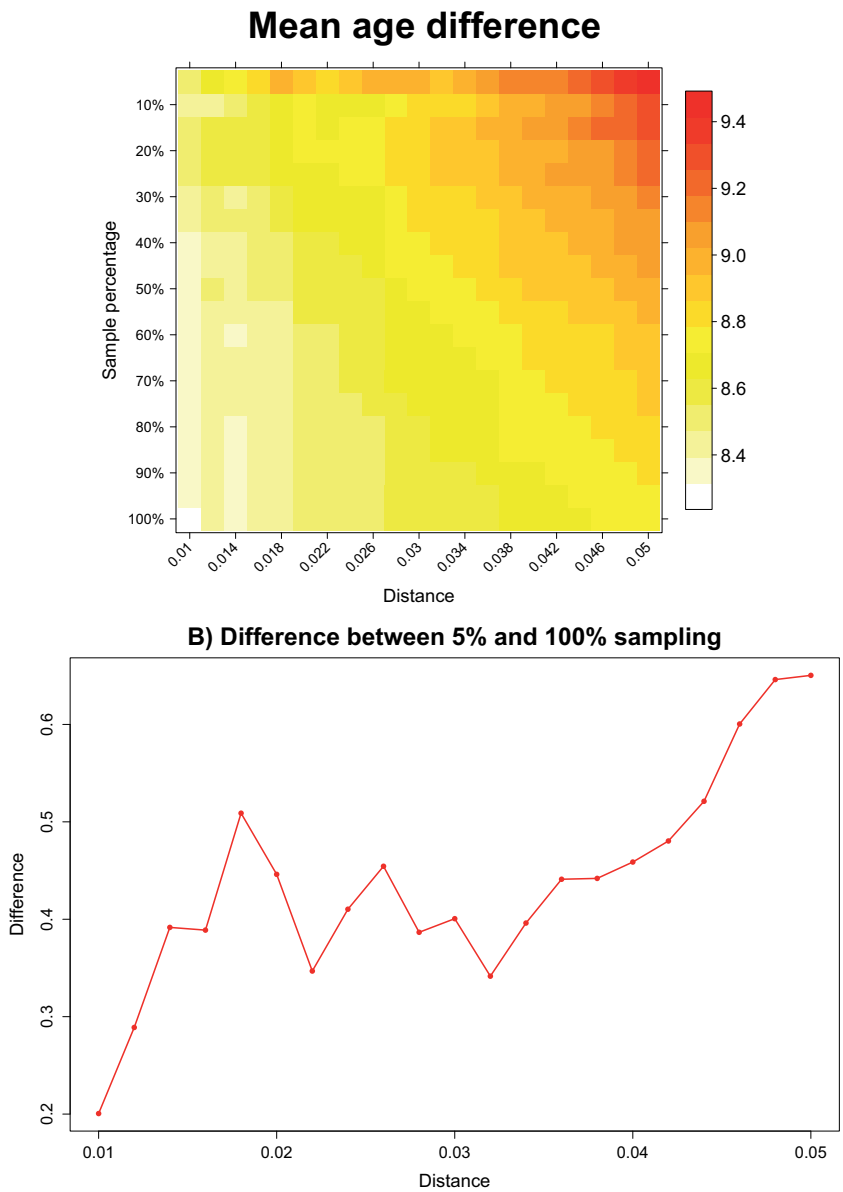

Figure 5. Top: mean age difference of all pairs in years (by color, scale bar on the right) derived from the phylogenetic tree, stratified by distance threshold (ranging from 0.01 to 0.05) and sample percentage (ranging from 5 to 100 per cent). Bottom: for each fixed distance threshold above, we show the difference in 'mean age difference' between 100 and 5 per cent sampling.

it should be noted that we only consider the age difference of patients in pairs. Although it might be potentially relevant, we do not distinguish different absolute ages leading to the same age difference, e.g., we do not distinguish between pairs with ages 18 and 25 years or 38 and 45 at infection (both correspond to an age difference 7 years). The main hurdle for distinguishing between different absolute ages leading to the same age difference is that the exact date of infection is unknown for most transmission events.

The same holds for the large age difference of more than 9 years observed for HET-pairs: not the age difference of heterosexual pairs in Switzerland is reflected, but the absolute age difference in pairs for which an HIV transmission event happened, with the most likely route being heterosexual contacts. Cases of HET-pairs with a large age difference might be pairs of patients with mixed ethnicity, as shown by Marzel et al. (2017). In their study, among thirty-three validated transmission pairs, twelve pairs were of mixed ethnicity with a large median age difference of 17.5 years. In addition, the mean age difference might still be an overestimation: we show that the mean age difference of pairs clearly increases when using a smaller sample proportion and as a result would most likely observe a decrease when performing the analysis with an even higher sample proportion than provided by the SHCS.

This raises a problem concerning phylogenetic studies performed on sparsely sampled populations, as, for example, by Oliveira et al. (2017). In this study, the authors wanted to highlight the 'sugar daddy'-phenomenon for KwaZulu-Natal, South Africa, by understanding the age structure of transmission pairs using phylogenetic analysis. Their study area, a part of the uMgungundlovu district of KwaZulu-Natal, has about 445,000 inhabitants, of which around 40 per cent are infected with HIV. Using a genetic distance threshold of 4.5 per cent, they identified 90 phylogenetic clusters in a tree of 1,589 sequences, all

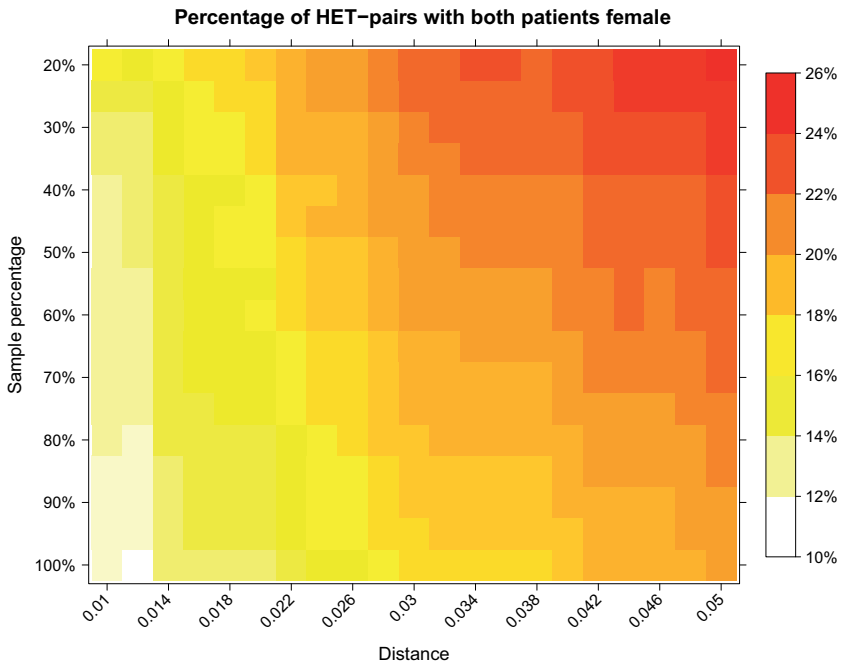

Figure 6. Percentage of HET-pairs (indicated by color, scale bar on the right) where both patients are female for various distance thresholds (ranging from 0.01 to 0.05) and sample proportions (ranging from 20 to 100 per cent).

sampled of individuals not virally suppressed at the time the study was undertaken. They found high age differences within these clusters. Our study showcases that their low sample proportion may (at least partially) explain these high age differences. Of course, there are many other studies that do not use phylogenetic methods, which report that high age differences in Sub-Saharan Africa, i.e., young women and older men, drive the HIV epidemics (Schaefer et al. 2017). Also, even with only 1 per cent sample density of the SHCS sequences—the lowest sample percentage we analyzed—and a liberal distance threshold, we still observed a lower age difference in the pairs in the trees compared with the average age difference of patients in the tree. However, the magnitude of the age difference reported by Oliveira et al. (2017) might be overestimated due to the low sampling density and a liberal genetic distance threshold.

Kiwuwa-Muyingo et al. (2017) found a high within-household and within-community HIV transmission in five fishing communities with approximately 15,000 inhabitants at Lake Victoria, Uganda, by using 238 sequences. To compensate for this low sample proportion, an extensive sensitivity analysis was performed by using genetic distance thresholds ranging from 0.5 to 5 per cent. Moreover, they used not only HIV-1 *pol*-sequences but also *env*-sequences for their analysis. An important question to investigate is therefore, whether a conservative distance criterion for including pairs or clusters in phylogenetic analyses can compensate for sparse sampling. For that, it is important to mention that we usually do not see a random sample of the population of interest, but a biased sample for several reasons: some transmission groups are more likely to be sampled earlier in their HIV infection, as is, for example, the case for MSM in Switzerland. This leads to a problem when choosing a distance threshold, as transmission pairs that are sequenced during the first few months of the HIV infection have a much smaller distance compared with pairs sampled many years after infection. With a too conservative distance threshold, recent transmission pairs are preferentially selected, introducing another bias into the phylogenetic analysis, as, for example, shown by Marzel et al. (2016). Another bias concerning the distance threshold could be introduced by different in-host evolution, as, for example, due to HIV infection by multiple founder viruses, HIV super infection or simply different host- or viral-genetic factors. These factors are hardly studied with regard to their impact on phylogenies, making it difficult to deduce the ideal distance criterion. For low sample proportions, there may not even be an ideal distance criterion; the mean age difference of pairs in the transmission tree is overestimated for all criteria at low sample proportions (Fig. 5).

To conclude, a good way of dealing with the problem of a sparsely sampled HIV population is certainly to perform extensive sensitivity analysis on the distance criterion, but also resampling the available sequences and understanding the impact of sample density on the specific scientific question. Moreover, combining phylogenetic analysis of different regions of the genome, as done for *env* and *pol* by Kiwuwa-Muyingo et al. (2017) could be a promising method, which needs, however, further investigation. It is important to realize, depending on the scientific question of interest, whether working with true transmission pairs is crucial or transmission pairs reflecting the underlying social network suffice to understand the underlying dynamics. Our results show that pairs defined by a conservative distance threshold are more robust to sparse sampling of sequences from a patient population. This suggests that if the sampling proportion is low and if it is important that phylogenetic pairs reflect true pairs (as might be the case for quantifying the prevalence of transmission in pairs with large age differences), the pitfalls induced by the low sample proportion can be alleviated by choosing a strict distance threshold. In addition, measures to estimate the magnitude of the pairs that are most likely no real transmission pairs could help to determine suitable parameters for the phylogenetic analysis. One example for such a measure is the percentage of pairs with both patients being female heterosexual (see Fig. 6). Of course, female–female

HET-pairs could still be true transmission pairs due to wrong classification of transmission group, i.e., they could have shared needles, or reflect rare events of sexual female-to-female HIV transmission (Chan et al. 2014). Nevertheless, in the case of a high number of female–female HET-pairs, further investigation on the chosen parameters should be considered. Finally, combining phylogenetic analysis with other clinical and demographic properties, such as, for example, done by Marzel et al. (2017) by looking at shared visits in the clinic for detecting true transmission pairs, could increase the credibility of phylogenetic studies.

Summary

The representativeness of the SHCS resistance database, which covers at least 27 per cent of the whole Swiss HIV epidemic, allowed us to analyze the impact of the sample proportion and the distance threshold on the age difference of observed pairs in the phylogenetic tree. Both factors proved to influence the mean age difference of HIV transmission pairs, which was measured by the absolute difference in birth years. The age difference decreased almost monotonically both with a stricter, i.e., smaller, distance threshold, and a higher sample proportion. Especially for low sample proportions, deriving the age difference of pairs in the phylogenetic tree, or similar quantitative measures, can be misleading and requires extensive sensitivity analysis (see Fig. 5).

Ethical approval and consent to participate

The SHCS was approved by the ethics committees of the participating institutions (Kantonale Ethikkommission Bern, Ethikkommission des Kantons St. Gallen, Comite Departemental d'Ethique des Specialites Medicales et de Medicine Communataire et de Premier Recours, Kantonale Ethikkommission Zürich, Repubblica et Cantone Ticino–Comitato Ethico Cantonale, Commission Cantonale d'Étique de la Recherche sur l'Être Humain, Ethikkommission beider Basel for the SHCS and Kantonale Ethikkommission Zürich for the ZPHI), and written informed consent was obtained from all participants.

Data availability

All clinical, demographical, and genetic data were obtained as part of the Swiss HIV Cohort Study (SHCS). Due to privacy reasons, the sensitivities associated with HIV infections, and the representativeness of the dataset, a deposition of the sequence data in an open database is not possible (our data would in principle allow the reconstruction of transmission events and could thereby endanger the patients' privacy. This is especially problematic because HIV-1 sequences have been frequently used in court cases). From a scientific point of view, the consequences of an open and uncontrolled access to such densely sampled sequences could jeopardize the future publication (and thus the investigation) of similarly complete data-sets and thereby be contra-productive even from an 'open-data' perspective. However, data can be made available for checking the results on a confidential basis and a sub-sample of 10 per cent sequences from the SHCS has been uploaded to Genbank in the context of a previous publications. Moreover, all data in the SHCS can be used for well-defined projects that are in accordance with the guidelines of the SHCS, if a corresponding project proposal is approved by the SHCS scientific board.

Acknowledgements

We thank the patients who participate in the Swiss HIV Cohort Study; the physicians and study nurses, for excellent patient care; the resistance laboratories, for high-quality genotyping drug resistance testing; SmartGene (Zug, Switzerland), for technical support; Alexandra Scherrer, Susanne Wild, Anna Traytel from the SHCS data center for data management, Danièle Perraudin and Mirjam Minichiello for administration. The members of the Swiss HIV Cohort Study include the following: Anagnostopoulos A, Battegay M, Bernasconi E, Böni J, Braun DL, Bucher HC, Calmy A, Cavassini M, Ciuffi A, Dollenmaier G, Egger M, Elzi L, Fehr J, Fellay J, Furrer H (Chairman of the Clinical and Laboratory Committee), Fux CA, Günthard HF (President of the SHCS), Haerry D (deputy of 'Positive Council'), Hasse B, Hirsch HH, Hoffmann M, Hösli I, Huber M, Kahlert C, Kaiser L, Keiser O, Klimkait T, Kouyos RD, Kovari H, Ledergerber B, Martinetti G, Martinez de Tejada B, Marzolini C, Metzner KJ, Müller N, Nicca D, Paioni P, Pantaleo G, Perreau M, Rauch A (Chairman of the Scientific Board), Rudin C (Chairman of the Mother & Child Substudy), Scherrer AU (Head of Data Centre), Schmid P, Speck R, Stöckle M, Tarr P, Trkola A, Vernazza P, Wandeler G, Weber R, Yerly S. Morover, we want to thank the anonymous referees for giving valuable input on an earlier version of this article.

Funding

This work was supported by the Swiss National Science Foundation (Grant number BSSGI0_155851). HFG was supported by the Swiss National Science Foundation (Grant number 179571). Furthermore, this study has been financed within the framework of the Swiss HIV Cohort Study, supported by the Swiss National Science Foundation (Grant number 148522), by the SHCS research foundation by the Yvonne Jacob Foundation (to HFG), by the clinical research priority program of the University of Zurich Viral infectious diseases, ZPHI (to HFG). HFG has received an unrestricted research Grant from Gilead to the SHCS Research Foundation.

Conflicts of interest: H.F.G. has received unrestricted research grants from Gilead Sciences and Roche; fees for data and safety monitoring board membership from Merck; consulting/advisory board membership fees from Gilead Sciences, Sandoz and Mepha; and travel reimbursement from Gilead. MB has received research or educational grants by Abb Vie AG, Gilead Sciences Switzerland Sàrl, Janssen-Cilag AG, MSD Merck Sharp & Dohme AG and ViiV Healthcare GmbH. EB has received fees for his institution for participation to advisory board from MSD, Gilead Sciences, ViiV Healthcare, Abbvie and Janssen. M.C. has received research and travel grants for his institution from ViiV and Gilead. A.C. has received unrestricted educational and research grants from MSD, Gilead and ViiV. A.R. reports support to his institution for advisory boards and/or travel grants from Janssen-Cilag, MSD, Gilead Sciences and Abbvie, and an unrestricted research grant from Gilead Sciences. All remuneration went to his home institution and not to AR personally, and all remuneration was provided outside the submitted work.

References

Bachmann, N. et al. (2017) 'Parent-Offspring Regression to Estimate the Heritability of an HIV-1 Trait in a Realistic Setup', *Retrovirology*, 14: 33.

Chan, S. K. et al.; Centers for Disease Control and Prevention (CDC) (2014) 'Likely female-to-female sexual transmission of HIV–Texas, 2012', *Morbidity and Mortality Weekly Report*, 63, 209–12.

Chow, E. P. F. et al. (2016) 'Assortative Sexual Mixing Patterns in Male-Female and Male-Male Partnerships in Melbourne, Australia: Implications for HIV and Sexually Transmissible Infection Transmission', *Sexual Health*, 13: 451–6.

Doherty, I. A., Schoenbach, V. J., and Adimora, A. A. (2009) 'Sexual Mixing Patterns and Heterosexual HIV Transmission Among African Americans in the Southeastern United States', *Journal of Acquired Immune Deficiency Syndromes*, 52: 114–20.

Drescher, S. M. et al.; Swiss HIV Cohort Study (2014) 'Treatment-Naive Individuals Are the Major Source of Transmitted HIV-1 Drug Resistance in Men Who Have Sex with Men in the Swiss HIV Cohort Study', *Clinical Infectious Diseases*, 58: 285–94.

Fact sheet—Latest statistics on the status of the AIDS epidemic (2017) [WWW Document], http://www.unaids.org/en/resources/fact-sheet accessed 30 March 2018.

Grey, J. A. et al. (2015) 'Disassortative Age-Mixing Does Not Explain Differences in HIV Prevalence between Young White and Black MSM: Findings from Four Studies', *PLos One*, 10: e0129877.

Hurt, C. B. et al. (2010) 'Sex with Older Partners Is Associated with Primary HIV Infection among Men Who Have Sex with Men in North Carolina', *Journal of Acquired Immune Deficiency Syndromes*, 54: 1–190.

Kiwuwa-Muyingo, S. et al. (2017) 'HIV-1 Transmission Networks in High Risk Fishing Communities on the Shores of Lake Victoria in Uganda: A Phylogenetic and Epidemiological Approach', *PLoS One*, 12: e0185818.

Le Vu, S. et al. (2018) 'Comparison of Cluster-Based and Source-Attribution Methods for Estimating Transmission Risk Using Large HIV Sequence Databases', *Epidemics*, 23: 1.

Marzel, A. et al. (2016) 'HIV-1 Transmission during Recent Infection and during Treatment Interruptions as Major Drivers of New Infections in the Swiss HIV Cohort Study', *Clinical Infectious Diseases*, 62: 115–22.

—— et al.; Swiss HIV Cohort Study (SHCS) (2017). Mining for pairs: shared clinic visit dates identify steady HIV-positive partnerships. *HIV Medicine*, 18: 667–76.

Morris, M., Zavisca, J., and Dean, L. (1995) 'Social and Sexual Networks: Their Role in the Spread of HIV/AIDS among Young Gay Men', *AIDS Education and Prevention: Official Publication of the International Society for Aids Education*, 7: 24–35.

Novitsky, V. et al. (2014) 'Impact of Sampling Density on the Extent of HIV Clustering', *AIDS Res. Hum. Retroviruses*, 30: 1226–35.

de Oliveira, T. et al. (2017) 'Transmission Networks and Risk of HIV Infection in KwaZulu-Natal, South Africa: A Community-Wide Phylogenetic Study', *Lancet HIV*, 4: e41–50.

Ott, M. Q. et al. (2011) 'Age-Gaps in Sexual Partnerships: Seeing beyond 'Sugar Daddies'', *AIDS (London, England)*, 25: 861–3.

Paradis, E., Claude, J., and Strimmer, K. (2004) 'APE: Analyses of Phylogenetics and Evolution in R Language', *Bioinformatics*, 20: 289–90.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009) 'FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix', *Molecular Biology and Evolution*, 26: 1641–50.

Ratmann, O. et al. (2017) 'Phylogenetic Tools for Generalized HIV-1 Epidemics: Findings from the PANGEA-HIV Methods Comparison', *Molecular Biology and Evolution*, 34: 185–203.

Schaefer, R. et al. (2017) 'Age-Disparate Relationships and HIV Incidence in Adolescent Girls and Young Women: Evidence from Zimbabwe', *AIDS (London, England)*, 31: 1461–70.

Schoeni-Affolter, F. et al. (2010) 'Cohort Profile: The Swiss HIV Cohort Study', *International Journal of Epidemiology*, 39: 1179–89.

Shilaih, M. et al.; Swiss HIV Cohort Study (2016) 'Genotypic Resistance Tests Sequences Reveal the Role of Marginalized Populations in HIV-1 Transmission in Switzerland', *Scientific Reports*, 6: 27580.

Sokal, R. R., and Rohlf, F. J. (1962) 'The Comparison of Dendrograms by Objective Methods', *Taxon*, 11: 33–40.

Turk, T. et al. (2017) 'Assessing the Danger of Self-Sustained HIV Epidemics in Heterosexuals by Population Based Phylogenetic Cluster Analysis', *eLife*, 6, https://doi.org/10.7554/eLife.28721.

Wilson, D. P. (2009) 'Modelling Based on Australian HIV Notifications Data Suggests Homosexual Age Mixing Is Primarily Assortative', *Journal of Acquired Immune Deficiency Syndromes*, 51: 356–60.