AIDS Research and Human Retroviruses

DOI: 10.1089/AID.2018.0236

Title: HIV-1 transmission patterns in men who have sex with men: insights from genetic source attribution analysis.

Running title: Patterns of HIV transmission among MSM.

Words main text: 3685

Authors: Stéphane Le Vu[1], Oliver Ratmann[2], Valerie Delpech[3], Alison E Brown[3], O Noel Gill[3], Anna Tostevin[4], David Dunn[4], Christophe Fraser[5], and Erik M Volz[1], on behalf of the UK HIV Drug Resistance Database.

Affiliations: [1]Department of Infectious Disease Epidemiology and the National Institute for Health Research Health Protection Research Unit on Modeling Methodology, Imperial College London; [2]Department of Mathematics, Imperial College London; [3]HIV and STI Department of Public Health England's Centre for Infectious Disease Surveillance and Control, London; [4]Institute for Global Health, University College London; [5]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom

Corresponding author: Stéphane Le Vu (s.levu@laposte.net) - Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London - St Mary's Campus, Norfolk Place, London W2 1PG

Keywords: Age-mixing; HIV epidemiology; Phylogenetic; Phylodynamics

Abstract

Background: Near 60% of new HIV infections in the United Kingdom are estimated to occur in men who have sex with men (MSM). Age-disassortative partnerships in MSM have been suggested to spread the HIV epidemics in many Western developed countries and to contribute to ethnic disparities in infection rates. Understanding these mixing patterns in transmission can help to determine which groups are at a greater risk and guide public health interventions.

Methods: We analyzed combined epidemiologic data and viral sequences from MSM diagnosed with HIV at the national level. We applied a phylodynamic source attribution model to infer patterns of transmission between groups of patients.

Results: From pair probabilities of transmission between 14 603 MSM patients, we found that potential transmitters of HIV subtype B were on average 8 months older than recipients. We also found a moderate overall assortativity of transmission by ethnic group and a stronger assortativity by region.

Conclusions: Our findings suggest that there is only a modest net flow of transmissions from older to young MSM in subtype B epidemics and that young MSM, both for Black or White groups, are more likely to be infected by one another than expected in a sexual network with random mixing.

Introduction

Men who have sex with men (MSM) account for forty percent of new HIV diagnoses in Europe [1]. In the United Kingdom (UK), nearly sixty percent of new infections are estimated to occur in MSM, although there is a recent sign of decline in diagnoses particularly recorded in London [2]. It has been estimated that the largest contribution to transmission in the UK is attributable to young HIV positive MSM [3]. More generally, since the early work from Morris et al. [4], young MSM having sex with older partners have been suggested to increase the risk of infection [5,6] and to represent a significant driver of the epidemic in North America [7]. This disassortative age mixing pattern is also considered in interaction with mixing by ethnicity [8,9]. Among MSM, black men appear to be more affected by HIV in both the UK and US contexts and age mixing patterns have been evaluated to illuminate this ethnic disparity in prevalence [10–12]. In addition to the question of transmission patterns by age and ethnicity, it is unclear whether the geographic variation in diagnosis rate for MSM is solely reflecting the demographic distribution of groups at greater risk in the country, or can also be explained by a varying extent of transmission between persons of different regions [13]. Assessing the primary sources of infection in these different demographic groups could prove helpful to design more effective intervention strategies.

Several studies have used phylogenetics to infer transmission patterns based on co-clustering of persons from different demographic or risk groups. For instance, occurrences of clustering observed between older and younger MSM is suggestive of a flow of transmission from old to young, as prevalence tends to increase with age [14,15].

However, there are several limitations to the interpretation of genetic clustering in terms of transmission. Clustering of genetically similar viruses is influenced by time since infection when patients are sampled, which is confounded by patients' age as well as CD4 and clinical stage of infection. Also the extent of clustering is dependent on the fraction of infected persons sampled, which makes direct inference of transmission patterns difficult using genetic clustering [16–18]. Particularly, the direction of putative transmission events cannot be resolved by pairwise genetic distance alone, and it is not possible to estimate flows of transmission between age groups based on clustering observations.

In this study, we applied a phylogenetic source attribution (SA) method that infers the probability of potential transmission (infector probability) between pairs of patients among approximately 15 000 MSM diagnosed in the UK with available genetic sequences [19]. Source attribution methods based on consensus pol-sequence data cannot be used to infer transmission pairs with high confidence, but can provide useful insights when studied in aggregate over thousands of putative transmission pairs. In general, direction of transmission cannot be inferred from consensus HIV sequence data, but in combination with clinical stage of infection at time of sequencing, directionality can be inferred probabilistically in some cases, as when for example a patient with chronic infection is linked to a patient with early infection. By combining phylogenetic analysis with stage of infection data and independent estimates of incidence and prevalence in the population, we are able to quantify potentially imbalanced transmission patterns between different risk groups. To this end , we used sequencing data routinely collected for drug resistance testing, patient-level data informative of the time since infection to account for biased sampling, and population estimates of background prevalence and incidence to account for potentially unsampled individuals that could be the sources of infection. In estimating transmission pair probabilities, our objective was to reveal patterns of transmission in men who have sex with men according to age, ethnicity, and geography. In particular, we searched for evidence of source-sink relationships in transmission patterns between age groups and examined the hypothesis that there is a net flow of transmissions from old to young MSM overall or by ethnicity.

Materials and Methods

Data

We used partial HIV-1 pol sequences collected in the UK HIV Drug Resistance Database [20] linked with characteristics of patients newly diagnosed with HIV from the UK Collaborative HIV Cohort study database and the national HIV/AIDS Reporting System database [21], as of end of August 2016. Among MSM diagnosed with HIV after 1997 in the UK, 58% had at least one sequence. The data were fully anonymised.

We analyzed adult patients reported as MSM; infected by HIV-1 subtype A1, B, C or CRF-02AG (the 4 most represented subtypes); and having a nucleotide sequence while treatment naive. The first sequence per patient with length > 950 nucleotides was included. CD4 count values closest to and within a maximum of 1 year of the date of sequence sampling were used to define 5 stages of infection, comprising early HIV infection (stage 1) and 4 stages of declining CD4 with thresholds at 500, 350 and 200 cells/mm3 [22]. In our sample, 81% of patients had a CD4 count. A positive result from the avidity-based recent infection testing algorithm (RITA) led to classifying a patient as at stage 1. Results of RITA at diagnosis were available as of 2009, and from this year were informed for 46% of patients.

Age of patients was categorized in quartiles of age at the date of resistance testing. Difference in age between patients was calculated relative to year of birth. Ethnicity categories were grouped in 7 classes: White; Black Caribbean; Black African; Other or unspecified black; Indian, Pakistani or Bangladeshi (South Asian); Other Asian or Oriental, Other and mixed. Regions of diagnosis were categorized in 5 classes: London; South of England; Midlands and East of England; North of England; Northern Ireland, Scotland and Wales. In analyses of assortativity, unknown category was treated as missing data.

Sequence processing

Partial HIV-1 pol sequences from the UK were sampled from 1997 to July 2015 with a majority obtained after 2009. Subtypes were determined with REGA version 3 [23]. To infer importation of viral lineages, a BLAST search [24] was performed for each UK sequence to identify the global sequence from the Los Alamos HIV sequence database (LANL)[25] with highest similarity. We retained 1780 unique matching global sequences, as more than one UK sequence may have the same BLAST match. Four reference alignments [26] per each subtype were also added to UK sequences to serve as outgroup for rooting the phylogenetic trees. All alignments were obtained with MAFFT version 7 [27]. Drug resistance mutation sites were stripped from the alignments [28].

Phylogenetic analysis

Phylogenetic trees were constructed with ExaML by maximum likelihood based inference with a gamma distribution model for rate heterogeneity among sites [29]. One hundred bootstrap replicates of each tree were computed to account for phylogenetic uncertainty.

We calculated root-to-tip distance and regressed distance by time from MRCA to sample. By iterations of Grubb's algorithm [30], we identified on overall 0.3% sequences as outliers in terms of divergence time and evolutionary rate. We applied least-square dating algorithm [31] on rooted trees and sampling times to estimate the substitution rate and dates of ancestral nodes.

We analyzed separately the 4 main subtypes to account for different evolutionary rates. Fitch algorithm was used to reconstruct ancestral host status (UK vs global) and determine distinct clades of virus transmitted in the UK [32]. The dated subtype B phylogeny comprised 18,484 taxa and for computational reasons was split into subtrees (clades) for further analyses. The tree splitting step consisted in iteratively testing thresholds of forward times (above the root) in order to slice [33] the large tree into clades with maximum size of 1000 taxa (viruses from UK patients). Thus for each of 100 bootstrap trees for subtype B, resulting clades were different.

Probabilistic source attribution

We applied a phylogenetic source attribution method that uses a population-genetic model to derive probabilities that a given individual (donor) is the source of infection for another individual (recipient) in the sample. These probabilities, termed *infector probabilities*, account for the epidemiological and sampling processes by incorporating into their calculation the time-scaled phylogeny, patient data on stage of infection, and population-level data on occurrence of infection [19]. The method was evaluated in a previous simulation study [18].

For population-level epidemic statistics, we used updated incidence estimates of CD4-based back-calculation method for MSM population and prevalence estimates of Bayesian multiparameter synthesis of surveillance data, as reported by Public Health England in

2017 [13]. To account for uncertainty in those input parameters, we randomly drew 5 pair values of incidence and prevalence per bootstrap replicates (2000 in total) from normal distributions inferred from the credible intervals of those estimates. Incidence and prevalence were assumed to be proportional across subtypes.

The source attribution method uses a continuous time Markov chain model to reconstruct the likely state of a lineage at the time of transmission given the CD4-stage of infection at time of sampling. The definition of stages of infection and progression rates were based on Cori et al. [22], as described in our previous analysis [18]. In case of missing CD4 count and missing RITA results at sampling, individuals were assigned a stage with probability relative to the average duration of respective stages. The method assumes that each infected patient corresponds to a single lineage of virus, ignoring multiple infections, and that internal nodes in the phylogeny correspond to a transmission event between hosts. To limit calculations to non-negligible pairing, only coalescent events within a limit of 20 years prior to sequence sampling were incorporated to compute infector probabilities.

Statistical procedures

Infector probabilities W_{ij} for each donor-recipient pair were averaged over all bootstrap replicates. To compare the mean age of donors and recipients we used a two-tailed paired weighted t-test on years of birth, with pair-level infector probabilities as weights.

To characterize transmission patterns by patients' covariates, we first computed a symmetric mixing matrix M as the normalized sum of infector probabilities representing aggregated number of transmissions between category k $(k = 1, \dots, m)$ of recipients and category l $(l = 1, \dots, m)$ of donors defined by age, ethnicity and region of diagnosis $(\sum_k^m \sum_l^m M_{kl} = 1)$. We then calculated 3 types of output matrices: (1) $R_{kl} = M_{kl} / \sum_{z=1}^m M_{kz}$, representing the conditional probability for a recipient in category k of being infected by a donor in category l; (2) $D_{kl} = M_{kl} / \sum_{z=1}^m M_{zl}$, representing the conditional probability for a donor in category l of having transmitted to a recipient in category k; and (3) $A = (M - E)/E$, the assortativity matrix representing excessive transmission between categories of donors and recipients relative to random allocation. The matrix E has elements $E_{kl} = \sum^k M_{kl} \otimes \sum^l M_{kl} / \sum^k \sum^l M_{kl}$, and represents the expected values in the

absence of preferential mixing [34]. Matrix E allows the calculation of Newman's assortativity coefficient $r = (Tr(M) - Tr(E))/(1 - Tr(E))$. The coefficient ranges from -1 to 1, where $r = 0$ when there is no assortative mixing, $r = 1$ when there is perfect assortativity (every link connects individuals of the same type), and some negative value $-1 \leq r < 0$ for a perfectly disassortative network (the lower bound depending on number of categories and density of subgraphs in each category). In all matrix-type figures, we represent transmission going from donors in columns to recipient in rows.

Code availability

The code used in this article is available as a R package: https://github.com/slevu/garel

Results

Characteristics of the study population

The demographic and geographic composition of the 19,847 HIV-1 partial pol sequences from treatment naive patients diagnosed in the UK is described in table 1. Most gay and bisexual men diagnosed in the UK were infected with subtype B (93%). Therefore the patterns of transmission inferred from reconstructed phylogeny of subtype B sequences are largely dominating that of all MSM patients. Patients infected with non-B subtype were on average sampled later (median year of 2008 for subtype B, 2009 for subtypes A1 and C, and 2011 for CRF02AG) and were on average younger (median age of 35 for subtype B, 34 for subtypes A and C, and 32 for CRF02AG).

In terms of ethnicity, the majority (84%) of patients were white persons. Patients infected with C or CRF02AG were more commonly of non-white ethnicity: Black-African for 11% and 16% and from other non-white ethnicity for 19% and 26% respectively.

In terms of geography, half of subtype B and 71% of subtype CRF02AG sequences were sampled in Greater London. Apart from London, subtype A was especially prevalent in North of England (27%).

Infector probabilities

Across 100 bootstrap tree replicates for each subtype, we computed infector probabilities for on average 554 514 potential transmission pairs involving 14 603 patients (cf. table 2). The remaining 5244 individuals from the initial sample, besides 250 outliers in tree reconstruction, could not be connected by a probability of transmission due to their isolation in distinct clades or the time limit imposed to coalescent event. Although the distribution of infector probabilities is varying across bootstrap replicates, almost all estimates are very small (figure S1). This confers a very low confidence in any particular pair and interpretations in terms of transmission are only applicable at group level. Given the n by n matrix of probabilities that a patient i transmitted to a patient j, the sum $\sum^{i} W_{ij}$ represents the probability that the infector of j is in the sample. This quantity, denoted 'in-degree' , indicates that on average 36.6% (95%CI[35.2 - 38.0]) of potential donors are included in our sampled population (cf. table 2). Our estimates of in-degrees were moderately influenced by the variation in inputs of background incidence and prevalence, with lower incidence (or higher prevalence) increasing average in-degrees as the probability of an unsampled intermediary transmitter is decreased (cf. figure S2).

Age difference between donors and recipients

Table 3 shows the mean difference in age between donors and recipients, weighted by infector probabilities. A significant difference is only detectable for subtype B, donors being on average less than 8 months older than recipients. For subtype B, most transmission pairs in our sample involved individuals less than 30 years old (figure 1M). The largest proportion (46%) of infection acquired by young individuals was attributable to individuals in the same age category (figure 1R). And a strong assortativity in transmission mixing is seen in this youngest age category, indicating that young MSM are preferentially infected by young MSM. This preferential mixing is also seen in among individuals over 44 years. The overall assortativity coefficient was moderate with $r = 0.16$. Similar transmission patterns between age groups were observed for subtypes A and C (table S1). However transmission of subtype CRF02AG was characterized by a strong assortativity mostly in the oldest age category but more intergenerational mixing between other

categories(figure S3A). Despite the lack of significant difference in average age of donors relative to recipient shown previously for subtype CRF02AG, the most probable infector for individuals from intermediate age quartiles (30-36 and 37-43) was younger (less than 30) (figure S3R).

Transmission by ethnicity

The vast majority (85%) of MSM infected with subtype B viruses were of white ethnicity. We estimated that 82% of all transmissions in our sample occurred between white individuals, and that recipients of all ethnicities had a majority of white donors. The probability of having been infected by a white individual was 92% for whites, 77% for Indian/Pakistani or Bengladeshi, 75% for other Asians, 55% for Black Africans and 54% for Blacks Caribbeans. Conversely, a majority of transmission originating from donors of any ethnic group was estimated to affect white recipients. Figure 2a shows the level of assortativity in transmission of subtype B viruses between ethnic groups. Inter-ethnic transmission (cumulated pair probabilities outside the diagonal) represented 17% on overall and 58% when excluding the white category. Overall assortativity was moderate ($r = 0.17$) but a preferential mixing was especially observed within and between all black ethnic groups and within the South Asian group.

We estimated the probability of transmission of subtype B viruses between young (<30) and older MSM (30+) either from white or black ethnicity (figure 3). The relative excess of transmission within age categories observed previously is observed for both white and black ethnicities, and overall assortativity by age was similar ($r = 0.25$ for white and 0.28 for black). However, for a given older MSM the probability of transmitting to a young MSM was higher in black (39%) than in white ethnic group (22%).

Transmission by geographical region

Analyses of transmission by region show the largest level of assortativity, indicating a overall strong spatial structure of the epidemics (see figure 2b). Assortativity coefficients were 0.56 for subtype B and 0.49 for subtype CRF02AG. For those two subtypes, figure 4 shows the probability for a donor in a given region to transmit to a recipient of each respective region. For subtype B (left), the majority of transmissions (at least 60%) occur

within the same region but donors from every region contributed to infections diagnosed in London (10% for North of England, Northern Ireland, Scotland and Wales, 20% for the Midlands and East England, and 30% for the South of England). For subtype CRF02AG, there was a higher probability for donors from North of England (60%) or Northern Ireland, Scotland and Wales (70%) to infect recipients in London than individuals within the same region.

Discussion

The objective of this study was to describe patterns of HIV transmission between age, ethnicity and geographical categories in the United Kingdom. We used a phylodynamic inference based on sequences collected among diagnosed MSM which accounts for incomplete sampling and stage of infection at sampling time. By modelling an epidemic process that is compatible with the evolution of transmitted viruses and epidemiological surveillance data, we characterized past transmission events among nearly 15 000 MSM patients at the national level. Pair probabilities averaged over phylogenies and aggregated by age groups indicated a modest overall net flow of transmission from older to young MSM. This result is compatible with other studies reporting co-clustering of young and older patients [14,15] as we do not observe pure assortative mixing, with probable transmission occurring in both directions across age groups. But our results indicate that on average, flow from old to young is mostly compensated by the transmission from young to old (cf. figure 1). And when the flow is imbalanced, as for transmission of subtype B viruses, the difference is small. We observed an overall preferential mixing in transmission by age with greater assortativity both in the youngest and oldest age groups and more random mixing in intermediate age groups. Understanding age mixing patterns in transmission can help to determine which groups are at a greater risk and potentially guide public health interventions [35]. Our findings confirm that young MSM infect one another more than expected by random mixing which supports the idea that prevention benefit could be enhanced by focusing on this small group [36]. This result also corroborates the observation of recent clusters of young MSM sustaining the epidemic in the Netherlands [37]. We showed an overall preferential pairing by ethnicity in conjunction

with an important mixing between white men and men from each other ethnicity. It can be explained by the overwhelming proportion of white men in the population. But in non-white groups, more than a half of transmission was inter-ethnic, revealing that a substantial amount of transmission has occurred between ethnic groups among MSM. A similar pattern for sexual partnership between ethnic groups was reported in Britain [10]. Although we found a relatively higher assortativity among black MSM in general and a non-negligible mixing between black ethnic groups from different origins (African, Caribbean and other), HIV transmission appears less assortative among black MSM in the UK than it is in the USA [38]. We assessed whether intergenerational transmission was different in white and black MSM and found a similar level of age assortativity in both groups. Therefore as others in the US context [9] we did not find support in our findings to explain a disparity in HIV prevalence by age mixing [7,8]. Finally, we found a strong geographical structure for the epidemics among MSM, with region of diagnosis as the variable associated with the highest level of assortativity. This implies that interventions in a particular location would take time to diffuse to a wider population. It should be noted that region of diagnosis can be different than the region of residency or of actual transmission, which may lead to an underestimation of the true level of geographical structure.

Several potential limitations of our study relate to the assumptions of the phylogenetic inference and source attribution method. First, as stated in methods section, the source attribution method neglects some effects of within-host evolution which can cause discordance between phylogenies and transmission trees [39]. This approximation is reasonable if within-host evolution generates coalescence time considerably shorter than between hosts at the population level. Secondly, we incorporated crude estimates of incidence and prevalence in the inference of infector probabilities. These were assumed constant over the period and proportional across subtypes. However variation of these inputs within credible limits had limited impact on average infector probabilities (figure S2). Third, the direction in transmission was derived from CD4 count and RITA result data that were partially complete.

Nevertheless, our analysis aimed to improve the use of phylogenetic information relative to genetic clustering in two ways. First, by providing a rough measure of transmission probability which unlike linkage into clusters can indicate a directionality and gives more weight to pairs with higher credibility. Notably, output matrices and patterns between groups would be symmetrical if based on clustering. Secondly, by correcting for biases stemming from incomplete sampling of the infected host population. Lastly, the source attribution method was fast to compute and scaled easily to phylogenies based on many thousands of sequences. The approach we take is generalizable to many different settings and has wider applicability to other large pathogen sequence databases.

Future directions for this work include applying the analysis to the heterosexual population, where phylogenetic information could contribute to assess age disparity in mixing across gender [40,41]. Another direction would be to use methods exploiting next-generation sequencing, that account for within-host evolution and enhance resolution in identifying transmission [39,42].

In conclusion, this study has leveraged available patients data and viral sequences to provide evidence of assortativity in HIV transmission by age, ethnicity and geography. Understanding these patterns of transmission is important to modelling the impact of intervention strategies.

Acknowledgements

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). We thank the Imperial College High Performance Computing Service (doi: 10.14469/hpc/2232).

Author contributions

S.L.V. designed the study, performed the analysis and wrote the manuscript; O.R. contributed to the phylogenetic analysis and writing the manuscript; V.D., A.E.B., O.N.G., A.T., D.D. contributed to data collection, molecular sequencing, data monitoring and manuscript evaluation. C.F. contributed to manuscript editing and project leading. E.M.V. designed the study, con tributed to manuscript review and editing and project leading.

Competing interests statement

The authors declare no competing interests.

References

[1] European Centre for Disease Prevention and Control, WHO Regional Office for Europe. HIV/AIDS Surveillance in Europe 2017 - 2016 Data. Stockholm: ECDC; 2017.

[2] Brown AE, Nash S, Connor N, Kirwan PD, Ogaz D, Croxford S, et al. Towards elimination of HIV transmission, AIDS and HIV-related deaths in the UK. HIV Med 2018. doi:10.1111/hiv.12617.

[3] Punyacharoensin N, Edmunds WJ, De Angelis D, Delpech V, Hart G, Elford J, et al. Modelling the HIV Epidemic among MSM in the United Kingdom: Quantifying the Contributions to HIV Transmission to Better Inform Prevention Initiatives. AIDS (London, England) 2015;29:339–49. doi:10.1097/QAD.0000000000000525.

[4] Morris M, Zavisca J, Dean L. 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 1995;7:24–35.

[5] Hurt CB, Matthews DD, Calabria MS, Green KA, Adimora AA, Golin CE, et al. Sex with Older Partners Is Associated With Primary HIV Infection Among Men Who Have Sex With Men in North Carolina: JAIDS Journal of Acquired Immune Deficiency Syndromes 2010:1. doi:10.1097/QAI.0b013e3181c99114.

[6] Jin F, Grulich AE, Mao L, Zablotska I, O'Dwyer M, Poynten M, et al. Sexual Partner's Age as a Risk Factor for HIV Seroconversion in a Cohort of HIV-Negative Homosexual Men in Sydney. AIDS and Behavior 2013;17:2426–9. doi:10.1007/s10461-012-0350-7.

[7] Coburn BJ, Blower S. A Major HIV Risk Factor for Young Men Who Have Sex With Men Is Sex With Older Partners: JAIDS Journal of Acquired Immune Deficiency Syndromes 2010:1. doi:10.1097/QAI.0b013e3181d43999.

[8] Berry M, Raymond HF, Mcfarland W. Same Race and Older Partner Selection May Explain Higher Hiv Prevalence among Black Men Who Have Sex with Men. Aids 2007;21:2349–50. doi:10.1097/QAD.0b013e3282f12f41.

[9] Grey JA, Rothenberg RB, Sullivan PS, Rosenberg ES. Disassortative Age-Mixing Does Not Explain Differences in HIV Prevalence between Young White and Black MSM: Findings from Four Studies. PLoS ONE 2015;10. doi:10.1371/journal.pone.0129877.

[10] Doerner R, McKeown E, Nelson S, Anderson J, Low N, Elford J. Sexual Mixing and HIV Risk Among Ethnic Minority MSM in Britain. AIDS and Behavior 2012;16:2033–41. doi:10.1007/s10461-012-0265-3.

[11] Hickson F, Melendez-Torres GJ, Reid D, Weatherburn P. HIV, sexual risk and ethnicity among gay and bisexual men in England: survey evidence for persisting health inequalities. Sex Transm Infect 2017;93:508–13. doi:10.1136/sextrans-2016-052800.

[12] Millett GA, Peterson JL, Flores SA, Hart TA, Jeffries WL, Wilson PA, et al. Comparisons of Disparities and Risks of HIV Infection in Black and Other Men Who Have Sex with Men in Canada, UK, and USA: A Meta-Analysis. The Lancet 2012;380:341–8. doi:10.1016/S0140-6736(12)60899-X.

[13] Brown AE, Kirwan P, Chau C, Khawam J, Gill ON, Delpech VC. Towards Elimination of HIV Transmission AIDS and HIV Related Deaths in the UK - 2017 Report. Public Health England; 2017.

[14] Whiteside YO, Song R, Wertheim JO, Oster AM. Molecular Analysis Allows Inference into HIV Transmission among Young Men Who Have Sex with Men in the United States. AIDS (London, England) 2015;29:2517–22. doi:10.1097/QAD.0000000000000852.

[15] Wolf E, Herbeck JT, Van Rompaey S, Kitahata M, Thomas K, Pepper G, et al. Phylogenetic Evidence of HIV-1 Transmission Between Adult and Adolescent Men Who Have Sex with Men. AIDS Research and Human Retroviruses 2016;33:318–22. doi:10.1089/aid.2016.0061.

[16] Volz EM, Koopman JS, Ward MJ, Brown AL, Frost SDW. Simple Epidemiological Dynamics Explain Phylogenetic Clustering of HIV from Patients with Recent Infection. PLoS Computational Biology 2012;8:e1002552. doi:10.1371/journal.pcbi.1002552.

[17] Poon AFY. Impacts and Shortcomings of Genetic Clustering Methods for Infectious Disease Outbreaks. Virus Evolution 2016;2:vew031. doi:10.1093/ve/vew031.

[18] Le Vu S, Ratmann O, Delpech V, Brown AE, Gill ON, Tostevin A, et al. Comparison of Cluster-Based and Source-Attribution Methods for Estimating Transmission Risk Using Large HIV Sequence Databases. Epidemics 2017. doi:10.1016/j.epidem.2017.10.001.

[19] Volz EM, Frost SDW. Inferring the Source of Transmission with Phylogenetic Data. PLoS Computational Biology 2013;9. doi:10.1371/journal.pcbi.1003397.

[20] HIVRDB. UK HIV Drug Resistance Database 2016. http://www.hivrdb.org.uk/ (accessed July 29, 2016).

[21] Public Health England. HIV and AIDS Reporting System 2018. https://www.gov.uk/government/collections/hiv-surveillance-data-and-management (accessed February 1, 2018).

[22] Cori A, Pickles M, van Sighem A, Gras L, Bezemer D, Reiss P, et al. CD4+ Cell Dynamics in Untreated HIV-1 Infection: Overall Rates, and Effects of Age, Viral Load, Sex and Calendar Time. AIDS (London, England) 2015;29:2435–46. doi:10.1097/QAD.0000000000000854.

[23] Pineda-Peña A-C, Faria NR, Imbrechts S, Libin P, Abecasis AB, Deforche K, et al. Automated Subtyping of HIV-1 Genetic Sequences for Clinical and Surveillance Purposes: Performance Evaluation of the New REGA Version 3 and Seven Other Tools. Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases 2013;19:337–48. doi:10.1016/j.meegid.2013.04.032.

[24] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. Journal of Molecular Biology 1990;215:403–10. doi:10.1016/S0022-2836(05)80360-2.

[25] Los Alamos National Laboratory. Main Search Interface of HIV Sequence Database 2017. http://www.hiv.lanl.gov/ (accessed February 1, 2018).

[26] Leitner T, Korber B, Daniels M, Calef C, Foley B. HIV-1 Subtype and Circulating Recombinant Form (CRF) Reference Sequences 2005. https://www.hiv.lanl.gov/content/sequence/HIV/REVIEWS/RefSeqs2005/RefSeqs05.html (accessed February 1, 2018).

[27] Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 2013;30:772–80. doi:10.1093/molbev/mst010.

[28] Wensing AM, Calvez V, Günthard HF, Johnson VA, Paredes R, Pillay D, et al. 2015 Update of the Drug Resistance Mutations in HIV-1. Top Antivir Med 2015;23:132–41.

[29] Kozlov AM, Aberer AJ, Stamatakis A. ExaML Version 3: A Tool for Phylogenomic Analyses on Supercomputers. Bioinformatics 2015;31:2577–9. doi:10.1093/bioinformatics/btv184.

[30] Komsta L. Outliers: Tests for Outliers 2011. https://cran.r-project.org/web/packages/outliers/index.html (accessed February 2, 2018).

[31] To T-H, Jung M, Lycett S, Gascuel O. Fast Dating Using Least-Squares Criteria and Algorithms. Systematic Biology 2016;65:82–97. doi:10.1093/sysbio/syv068.

[32] Fitch WM. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. Systematic Zoology 1971;20:406–16. doi:10.2307/2412116.

[33] Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 2012;3:217–23. doi:10.1111/j.2041-210X.2011.00169.x.

[34] Newman MEJ. Mixing Patterns in Networks. Physical Review E 2003;67:026126. doi:10.1103/PhysRevE.67.026126.

[35] Anema A, Marshall BDL, Stevenson B, Gurm J, Montaner G, Small W, et al. Intergenerational Sex as a Risk Factor for HIV among Young Men Who Have Sex with Men: A Scoping Review. Current HIV/AIDS Reports 2013;10:398–407. doi:10.1007/s11904-013-0187-3.

[36] Volz EM, Le Vu S, Ratmann O, Tostevin A, Dunn D, Orkin C, et al. Molecular Epidemiology of HIV-1 Subtype B Reveals Heterogeneous Transmission Risk: Implications for Intervention and Control. The Journal of Infectious Diseases 2018;217:1522–9. doi:10.1093/infdis/jiy044.

[37] Bezemer D, Cori A, Ratmann O, van Sighem A, Hermanides HS, Dutilh BE, et al. Dispersion of the HIV-1 Epidemic in Men Who Have Sex with Men in the Netherlands: A Combined Mathematical Model and Phylogenetic Analysis. PLoS Med 2015;12:e1001898. doi:10.1371/journal.pmed.1001898.

[38] Oster AM, Wertheim JO, Hernandez AL, Bañez Ocfemia MC, Saduvala N, Hall IH. Using Molecular HIV Surveillance Data to Understand Transmission between Subpopulations in the United States. Journal of Acquired Immune Deficiency Syndromes (1999) 2015. doi:10.1097/QAI.0000000000000809.

[39] Romero-Severson EO, Bulla I, Leitner T. Phylogenetically Resolving Epidemiologic Linkage. Proceedings of the National Academy of Sciences 2016:201522930. doi:10.1073/pnas.1522930113.

[40] Prah P, Copas AJ, Mercer CH, Nardone A, Johnson AM. Patterns of Sexual Mixing with Respect to Social, Health and Sexual Characteristics among Heterosexual Couples in England: Analyses of Probability Sample Survey Data. Epidemiology and Infection 2015;143:1500–10. doi:10.1017/S0950268814002155.

[41] de Oliveira T, Kharsany ABM, Gräf T, Cawood C, Khanyile D, Grobler A, et al. Transmission Networks and Risk of HIV Infection in KwaZulu-Natal, South Africa: A Community-Wide Phylogenetic Study. The Lancet HIV 2017;4:e41–50. doi:10.1016/S2352-3018(16)30186-2.

[42] Wymant C, Hall M, Ratmann O, Bonsall D, Golubchik T, de Cesare M, et al. PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity. Molecular Biology and Evolution 2018;35:719–33. doi:10.1093/molbev/msx304.

Table 1: Characteristics of the study population

Subtype		A		B		C		CRF02AG		All	
		n	(%)	n	(%)	n	(%)	n	(%)	n	(%)
Year of sampling	(-Inf,2002]	17	4	1867	10	18	3	5	2	1907	10
	(2002,2007]	128	29	6497	35	188	31	47	15	6860	35
	(2007,2012]	186	42	7652	41	303	50	171	53	8312	42
	(2012, Inf]	107	24	2468	13	94	16	99	31	2768	14
Age group	[16,30)	151	34	4946	27	196	33	145	45	5438	27
	[30,37)	95	22	5160	28	152	25	62	19	5469	28
	[37,44)	89	20	4303	23	129	21	57	18	4578	23
	[44,85]	103	24	4075	22	126	21	58	18	4362	22
Ethnicity	White	377	86	15664	85	417	69	177	55	16635	84
	Black-Caribbean	4	1	408	2	14	2	23	7	449	2
	Black-African	19	4	199	1	67	11	52	16	337	2
	Black-other/unspecified	3	1	186	1	20	3	12	4	221	1
	Indian/Pakistani/Bangladeshi	7	2	265	1	25	4	6	2	303	2

	Other Asian/Oriental	11	3	549	3	19	3	20	6	599	3
	Other/Mixed	12	3	625	3	24	4	20	6	681	3
	Other	2	0	285	2	12	2	3	1	302	2
	Not known	3	1	303	2	5	1	9	3	320	2
Region of birth	UK	247	56	9489	51	249	41	136	42	1012 1	51
	SS Africa	19	4	379	2	81	13	33	10	512	3
	Other	73	17	3207	17	107	18	91	28	3478	18
	Not known	99	23	5409	29	166	28	62	19	5736	29
Region of diagnosis	London	174	40	9417	51	269	45	229	71	1008 9	51
	ML_E_England	23	5	1892	10	59	10	31	10	2005	10
	N_England	117	27	2309	12	70	12	18	6	2514	13
	S_England	56	13	2559	14	113	19	30	9	2758	14
	NI_S_W	31	7	784	4	32	5	6	2	853	4
	Not_known	37	8	1523	8	60	10	8	2	1628	8
	All	438	100	1848 4	100	603	100	322	100	1984 7	100

Table 2: Phylogenetic reconstruction and source attribution results by subtype

Subtype	A	B	C	CRF02AG	All
Number of global sequences	199	831	612	138	1780
Number of sequence outliers	6	163	7	74	250
Median TMRCA (year)	1951	1966	1961	1975	NA
Number of UK patients either donors or recipients	337	13665	346	255	14603
Number of infector probabilities estimated between potential transmission pairs	19818	521811	6350	6535	554514
Mean in-degree (%)	39.4	36.7	32.6	28.9	36.6

Results are averaged over 100 bootstrap replicates. Global sequences are unique sequences from Los Alamos HIV sequence database matching UK sequences from a BLAST search. Outliers are UK sequences identified as outliers in root-to-tip regression. Mean in-degree represents the probability that the donor of a given recipient is included in the sample.

Table 3: Difference in year between age of donor and age of recipient

Subtype	A	B	C	CRF02AG
Age difference*	0.13 [-0.80; 0.60]	0.63 [0.53; 0.73]	0.20 [-0.39; 0.71]	0.33 [-0.34; 1.03]
Birth year of donor	1973.8 [1973.2; 1974.5]	1972.1 [1971.9; 1972.2]	1974.5 [1973.8; 1975.1]	1977.0 [1975.9; 1978.6]
Birth year of recipient	1974.0 [1972.4; 1974.7]	1972.7 [1972.6; 1972.8]	1974.7 [1974.2; 1975.0]	1977.3 [1976.4; 1978.8]
Positive difference in age** (n)	30	100	28	47
Negative difference in age** (n)	5	0	3	3
Age at sampling of donor	35.3 [34.7; 36.2]	36.3 [36.2; 36.4]	34.8 [34.2; 35.5]	33.4 [31.9; 34.6]
Age at sampling of recipient	35.4 [34.8; 37.2]	35.9 [35.8; 35.9]	34.8 [34.4; 35.3]	33.2 [31.8; 34.0]

Results are averaged across 100 bootstrap replicates and intervals are 2.5 and 97.5 percentiles. *: Age difference is calculated relative to year of birth. All results are in years except **: Number of p-values < 0.05 for two-tailed weighted t-test of the age difference, either positive (donor older than recipient) or negative (donor younger than recipient).

List of Figures for the article "HIV-1 transmission patterns in men who have sex with men: insights from genetic source attribution analysis."

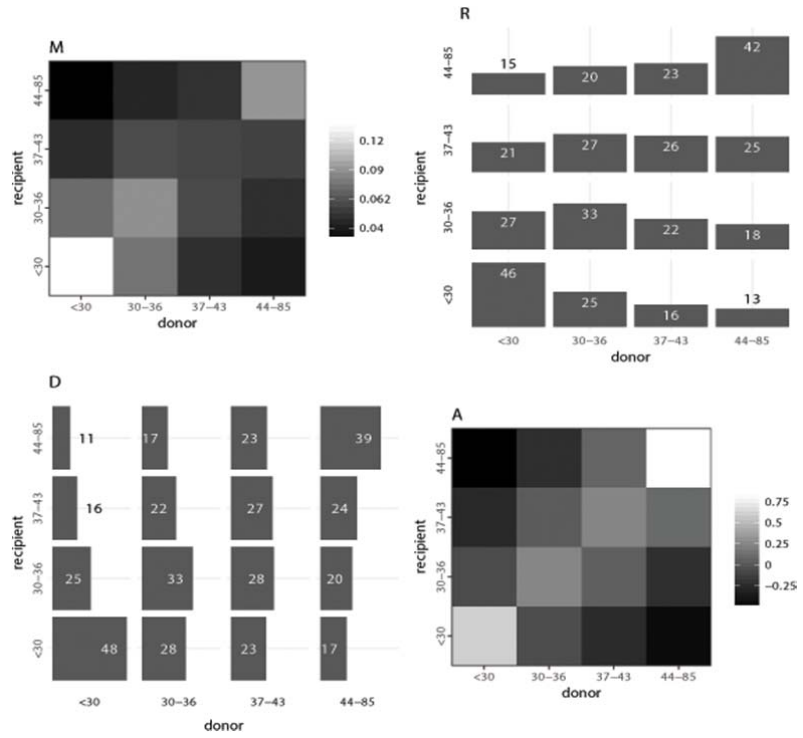

Figure 1: Patterns of transmission of HIV subtype B by age in quartiles. The four graphics depict transmission from donor categories in column to recipient categories in row (from x-axis to y-axis). Axes labels represent ranges of quartiles of age. **M**: Each cell represents the proportion of overall transmissions from one category to another, with higher proportion in lighter shade, *i.e.* the highest amount (14%) of transmissions involved donors and recipients both aged less than 30. **R**: Each row represents the probability distribution for a given age category of recipients of having been infected by donors by age, *i.e.* 25% of recipients less than 30 years old were infected by donor aged 30-36. **D**: Each column represents the probability distribution for a given age category of donors of having transmitted to recipients by age, *i.e.* 28% of donors aged 37-43 infected recipients aged 30-36. **A**: The assortativity matrix indicates that, relative to random mixing more transmissions occurred within the same age category, particularly for the oldest and youngest. Assortativity coefficient r = 0.16.

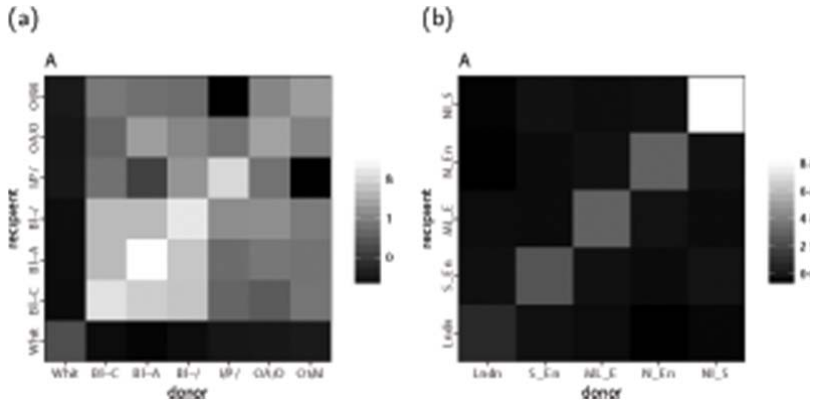

Figure 2: Assortativity in transmission of HIV-1 subtype B by ethnicity and region of diagnosis. Lighter shades represent higher assortativity. **(a)**: Ethnicities: White; Black Caribbean (Bl-C); Black African (Bl-A); Other or unspecified black (Bl-); Indian, Pakistani or Bangladeshi (I/P/); Other Asian or Oriental (OA/O), Other and mixed (Ot/M). Assortativity coefficient r = 0.17. **(b)**: Regions: London, South of England; Midlands and East of England; North of England; Northern Ireland, Scotland and Wales. Assortativity coefficient r = 0.56.

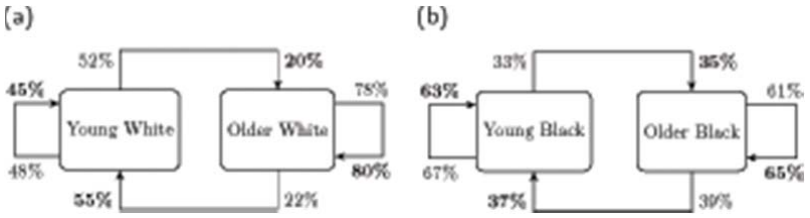

Figure 3: Patterns of transmission of HIV-1 subtype B between young MSM (less than 30) and older MSM by ethnicity: (a) White, (b) Black (including Black Africans, Black Carribean and other and unspecified Black). Percentages represent conditional probability of transmitting to recipient type per donor type (normal font) and of acquiring infection from donor type per recipient type (bold font).

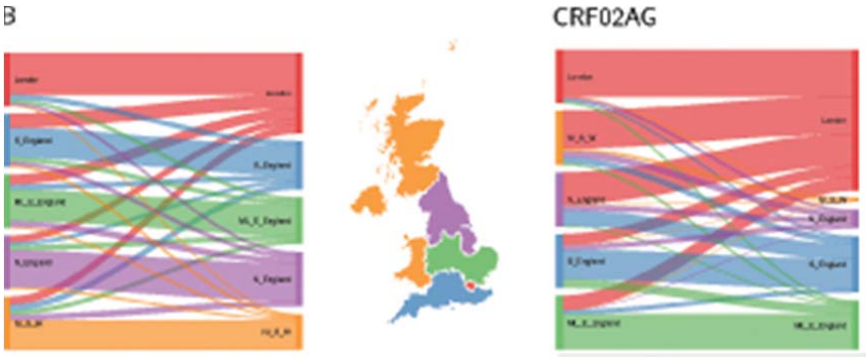

Figure 4: Patterns of transmission of HIV-1 subtype B (left) and CRF02AG (right), by geography. Each flow diagram, obtained from D matrix described in Methods section, has connections proportional to the probability of transmission from a donor given his region (left side) to recipients from respective regions (right side). The map is colored by groups of region of diagnosis: London, South of England (S_England); Midlands and East of England (ML_E_England); North of England (N_England); Northern Ireland, Scotland and Wales (NI_S_W).

Table S1: 95% confidence intervals for proportion of transmission by age quartiles.

		M				R				D			
Subtype	Age quartile	30	37	44	85	30	37	44	85	30	37	44	85
A	85	03;08	03;05	06;09	06;09	14;27	12;20	24;34	26;40	09;25	15;23	27;46	31;43
	44	03;04	02;04	04;07	06;09	16;23	11;20	21;34	33;45	08;12	10;18	20;29	28;39
	37	07;09	04;08	02;05	03;04	33;45	19;35	08;25	12;20	19;25	19;35	10;23	11;19
	30	15;23	07;11	04;06	03;05	47;60	19;28	10;16	07;15	44;61	35;48	19;28	13;21
B	85	03;03	04;05	05;05	09;09	14;16	20;21	22;23	41;43	11;12	16;17	22;24	38;40
	44	04;05	06;06	05;06	05;06	20;22	26;28	25;27	24;26	15;16	21;23	26;28	23;24
	37	07;07	09;09	06;06	05;05	26;28	32;34	21;23	17;19	24;25	32;34	27;29	20;21
	30	14;14	07;08	05;05	04;04	45;47	24;26	15;16	12;14	47;49	28;29	22;24	16;18
C	85	01;03	03;05	03;04	05;09	07;16	17;30	17;28	34;51	03;07	10;18	14;22	30;49
	44	05;07	05;07	07;10	03;06	20;27	21;30	28;38	13;22	13;19	19;27	36;47	19;31
	37	08;11	05;09	04;06	03;05	31;43	21;33	16;24	11;20	23;31	21;31	21;32	16;30
	30	17;21	08;13	02;04	01;03	50;61	25;37	06;11	04;09	48;57	32;44	10;19	08;17
CRF02AG	85	02;05	02;04	01;04	05;11	13;30	10;24	11;26	31;58	05;10	09;19	08;23	35;61
	44	05;08	01;04	02;04	01;04	33;53	09;23	16;30	10;25	09;18	07;19	13;26	10;27
	37	05;09	05;09	04;07	01;03	27;41	28;41	17;29	06;15	12;19	28;47	20;37	08;19
	30	25;34	05;10	05;09	02;04	55;69	11;23	10;19	04;09	56;68	29;49	29;46	13;28

M, R and D columns refer to the matrices described in methods section and represented in figure 1. Column labels for donors and row labels for recipients represent the upper bound of quartiles of age.

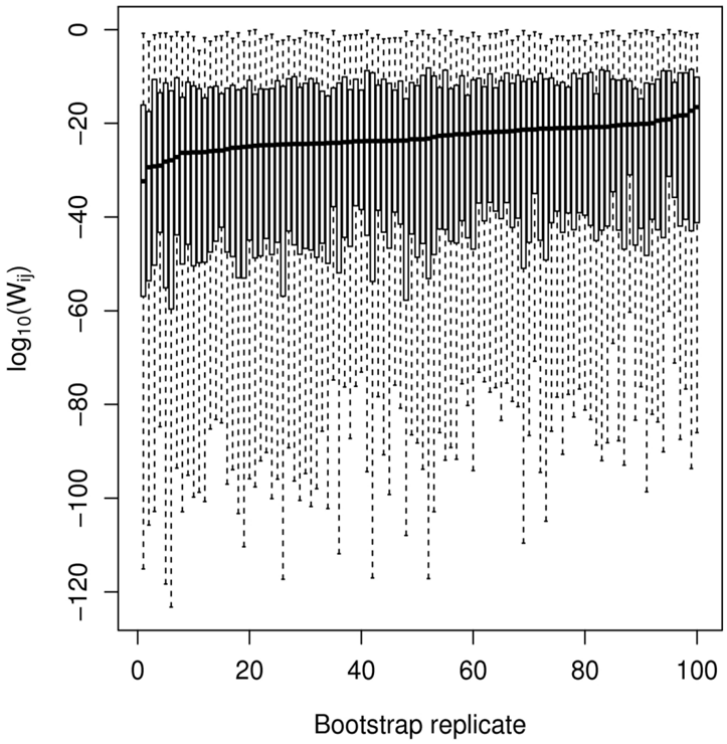

Figure S1: Distribution of infector probabilities across 100 bootstrap replicates.

To illustrate the variability of estimates between bootstrap replicates, a random sample of 100 probabilities (in logarithm) per replicate is represented. Results are sorted in increasing order of the median value. Outliers are not shown.

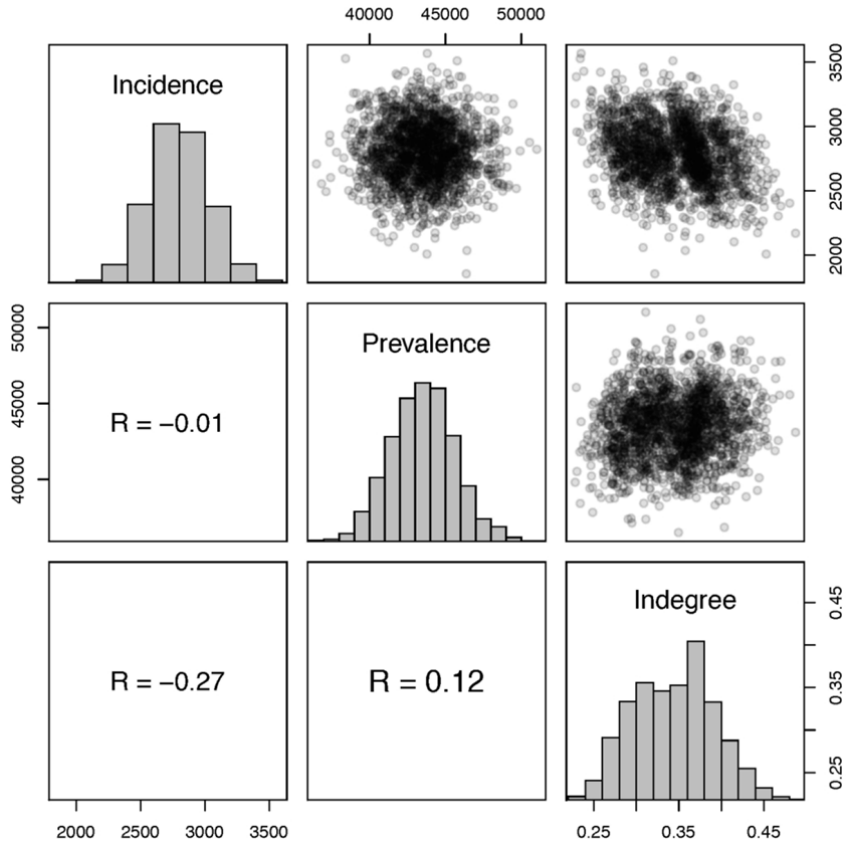

Figure S2: Influence of variation in incidence and prevalence inputs on mean in- degree. Incidence (number of new infections per year) and prevalence (number of persons living with HIV) were drawn independently 5 times per each of 100 phylogenetic trees for subtype B transmission. Under the source attribution model, the in-degree is the sum of infector probabilities incoming to an individual. It also represents the probability that its infector is in the sample. The mean in-degree tends to increase as prevalence is increased or as incidence is lowered, as it decreases the probability of an unsampled source case. Both variations have a limited impact on the average in-degree estimates as indicated by correlation coefficient R.

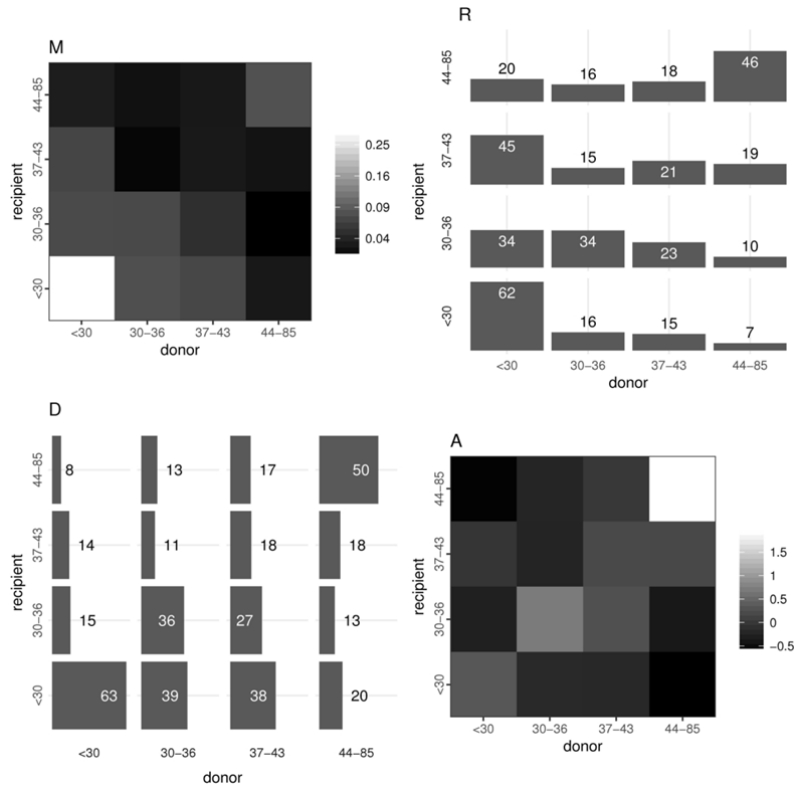

Figure S3: Patterns of transmission of HIV-1 subtype CRF02AG by age. See reading notes from figure 1 in main text.