Longitudinal population-level HIV epidemiological and genomic surveillance highlights growing gender disparity of HIV transmission in Uganda

Mélodie Monod[1†], Andrea Brizzi[1†], Ronald M Galiwango[2†], Robert Ssekubugu[2†], Yu Chen[1†], Xiaoyue Xi[1†], Edward Nelson Kankaka[3,4†], Victor Ssempijja[5,6†], Lucie Abeler Dörner[7], Adam Akullian[8], Alexandra Blenkinsop[1], David Bonsall[9,10], Larry W Chang[3,2,11], Shozen Dan[1], Christophe Fraser[7,10], Tanya Golubchik[12,7], Ronald H Gray[13], Matthew Hall[7], Jade C Jackson[14], Godfrey Kigozi[2], Oliver Laeyendecker[15,16], Lisa A. Mills[17], Thomas C Quinn[14,15,16], Steven J. Reynolds[2,15,16], John Santelli[18], Nelson K. Sewankambo[19], Simon EF Spencer[20], Joseph Ssekasanvu[11], Laura Thomson[7], Maria J Wawer[2,11], David Serwadda[2,19], Peter Godfrey-Faussett[21†], Joseph Kagaayi[2†], M Kate Grabowski[2,14,11†], Oliver Ratmann[1†], Rakai Health Sciences Program and PANGEA-HIV consortium

[1]Department of Mathematics, Imperial College London, London, United Kingdom.
[2]Rakai Health Sciences Program, Kalisizo, Uganda.
[3]Division of Infectious diseases, Johns Hopkins School of Medicine, Baltimore, Maryland, United States.
[4]Research Department, Rakai Health Sciences Program, Rakai, Uganda.
[5]Clinical Monitoring Research Program Directorate, Frederick National Laboratory for Cancer Research, Frederick,Maryland, United States.
[6]Statistics Department, Rakai Health Sciences Program, Rakai, Uganda.
[7]Big Data Institute, University of Oxford, Oxford, United Kingdom.

[8]Bill and Melinda Gates Foundation, Seattle, Washington, United States.
[9]Wellcome Centre for Human Genomics, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom.
[10]Pandemic Sciences Institute, University of Oxford, Oxford, United Kingdom.
[11]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States.
[12]Sydney Infectious Diseases Institute, School of Medical Sciences, Faculty of Medicine and Health, University of Sydney, Sydney, Australia.
[13]Professor Emeritus, Department of Epidemiology, Johns Hopkins University, Bloomberg School of Public Health, Baltimore, Maryland, United States.
[14]Department of Pathology, Johns Hopkins School of Medicine, Baltimore, Maryland, United States.
[15]Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, United States.
[16]Department of Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, United States.
[17]Division of Global HIV and TB, U. S. Centers for Disease Control and Prevention, Kampala, Uganda.
[18]Population and Family Health and Pediatrics, Columbia Mailman School of Public Health, New York, United States.
[19]College of Health Sciences, School of Medicine, Makerere University, Kampala, Uganda.
[20]Department of Statistics, University of Warwick, Coventry, United Kingdom.
[21]Department of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom.

Corresponding authors jkagayi@rhsp.org; mgrabow2@jhu.edu; oliver.ratmann@imperial.ac.uk;
[†]These authors contributed equally to this work.

Abstract

HIV incidence in eastern and southern Africa has historically been concentrated among girls and women aged 15-24 years. As new cases decline with

HIV interventions, population-level infection dynamics may shift by age and gender. Here, we integrated population-based surveillance of 38,749 participants in the Rakai Community Cohort Study and longitudinal deep sequence viral phylogenetics to assess how HIV incidence and population groups driving transmission have changed from 2003 to 2018 in Uganda. We observed 1,117 individuals in the incidence cohort and 1,978 individuals in the transmission cohort. HIV viral suppression increased more rapidly in women than men, however incidence declined more slowly in women than men. We found that age-specific transmission flows shifted, while HIV transmission to girls and women (aged 15-24 years) from older men declined by about one third, transmission to women (aged 25-34 years) from men that were 0-6 years older increased by half in 2003 to 2018. Based on changes in transmission flows, we estimated that closing the gender gap in viral suppression could have reduced HIV incidence in women by half in 2018. This study suggests that HIV programs to increase HIV suppression in men are critical to reduce incidence in women, close gender gaps in infection burden and improve men's health in Africa.

Main text

Introduction

Despite the widespread availability of HIV prevention and treatment interventions, there were 1.5 million new HIV infections and 680,000 HIV-associated deaths in 2020 [1]. More than half of these new cases and deaths were concentrated in the eastern and southern regions of the African continent, where incidence rates have historically been highest in adolescent girls and young women, aged 15-24 years [2,3,4,5]. While HIV incidence has declined by 43% in eastern and southern Africa since 2010, current HIV service programs are failing to reduce new cases rapidly enough to meet United Nations health targets for HIV epidemic control [1]. With rising levels of HIV drug resistance [6,7] and flatlined global investment in HIV control [8], the African HIV epidemic has reached a critical inflection point [9].

Over the last decade, African HIV control programs, including the United States President's Emergency Plan for AIDS Relief (PEPFAR), have focused on expanding treatment coverage in people with HIV and reducing HIV infections among adolescent girls and young women [10,11]. However, recent data from Africa indicate that the mean age of infection is shifting [12,13] and incidence rates are declining faster in men than in women [14,15], suggesting that the age and gender structure of the African HIV epidemic is evolving. Here, we integrate 15 years of data on HIV incidence and onward transmission to show how the drivers of the heterosexual African HIV epidemic are changing by age and gender. We focus on a study population aged 15 to 49 years with an HIV risk profile typical across eastern and southern Africa [16,17], living in 36 semi-urban and rural agrarian communities that are part of the population-based Rakai Community Cohort Study (RCCS) in south-central Uganda [18] (Fig. 1a). We followed individuals in the RCCS who were HIV seronegative and documented new infection events. We also deep-sequenced HIV virus longitudinally from persons

with viremic HIV. This enabled us to infer directed transmission networks across age and gender[19,20], and focus on the time trends in infection dynamics and transmission networks during mass scale-up of HIV services in Africa[1].

Results

HIV incidence is declining faster in men than women

From September 23, 2003 to May 22, 2018, 38,749 participants were enrolled in the Rakai Community Cohort Study[14]. Of these participants, 22,724 tested HIV seronegative at first survey, and contributed an estimated 127,217 person-years of follow-up (Fig. 1b, Supplementary Tables S1- S2). Study participants were enrolled following population census, household enumeration, and informed consent in 9 survey rounds of approximately 18 months duration, herein denoted as survey rounds 10-18 (see Methods and Extended Data Fig. 1).

In total, we observed 1,117 incident HIV infections (Supplementary Tables S3-S4 and Extended Data Fig. 2). Fig. 1c shows that incidence rates among men in inland communities fell rapidly from 1.05 [1.03-1.08] per 100 person-years (PY) in 2003 (survey round 10) by 67.8% [66.2-69.2] to 0.34 [0.33-0.35] per 100 PY in 2018 (survey round 18), with no substantial shift in the median age of male incident infection (blue triangles in Fig. 1c). In young women aged 15 to 24 years, incidence rates fell similarly rapidly from 1.42 [1.35-1.5] per 100 PY in 2003 by 74.5% [71.6-77.1] to 0.36 [0.33-0.4] per 100 PY in 2018. However, among women aged 25-34, declines in HIV incidence were substantially slower (from 1.51 [1.45-1.57] per 100 PY in 2003 by 43.9% [40.5-47.4] to 0.84 [0.8-0.89] per 100 PY in 2018), and similarly in women aged 35-49 (from 0.9 [0.85-0.94] per 100 PY in 2003 by 37.4% [31.9-42.6] to 0.56 [0.52-0.6] per 100 PY in 2018), resulting in a progressive, substantial shift in the median age of infection in women from 23.4 [22.6-24.1] in 2003 to 28.2 [27.1-29.2] in 2018 (Fig. 1c-d). Progress in reducing HIV incidence thus continues to be substantially slower in women[14,21], especially among those aged 25 years and above.

The proportion of transmission from men is increasing

To characterize the population transmission flows by age and gender underlying observed shifts in incidence, we deep-sequenced virus from 1,978 participants with HIV (Supplementary Table S5[19]). By embedding genomic surveillance into a population-based cohort study, deep-sequence sampling coverage was high relative to typical pathogen sequencing studies, which is essential for reconstructing transmission events[20,22,23,24,25]. We characterized the phylogenetic ordering between multiple viral variants from individuals and estimated the direction of transmission with *phyloscanner* (Methods)[22,26]. We identified 236 heterosexual source-recipient pairs that were phylogenetically close and exhibited, in combination with data on last negative and first positive tests, strongly consistent evidence of the direction of transmission (Methods and Extended Data Fig. 3). We further estimated the likely infection date from deep-sequence data[27], which enabled us to place the source-recipient pairs in calendar time and consider their age at the time of infection (Extended Data Fig. 4).

Of the 236 heterosexual source-recipient pairs, we retained in total 227 pairs in whom transmission was estimated to have occurred during the study period.

Deep-sequence phylogenetics cannot prove direction of transmission between two persons[22], but in aggregate these data are able to capture heterosexual HIV transmission flows at a population level[20,28]. We estimated population-level transmission flows adjusting for detection probabilities with semi-parametric Poisson flow regression models[29], and under the constraint that the transmission flows needed to closely match the age- and gender-specific incidence dynamics shown in Fig. 1 (Methods, Extended Data Fig. 5, and Supplementary Table S6). The fitted model was consistent with all the available data (Extended Data Fig. 6). Fig. 2a shows the age profile of the estimated male and female sources of infection, such that the male plus the female sources sum to 100% for each survey round. Overall, we found that the contribution of men to onward transmission increased progressively from 57.9% [56.2-59.6] in 2003 to 62.8% [60.2-65.2] in 2018, indicating that HIV transmission is now more disproportionately driven by men than has been the case previously.

Transmissions from men are shifting to older ages

The age profile of the population-level sources of infection characterizes the major age groups that sustain transmission[30]. We find that the age of transmitting male partners progressively increased from a median age of 28.5 [27.1-30.1] years in 2003 to 33.5 [31.0-36.0] years in 2018 (Table 1 and Fig. 2a), and this increase in the age of transmitting male partners was largest in transmissions to women aged 20-24 (Fig. 2b). In contrast, the median age of female transmitting partners remained similar (from 25.0 [23.0-27.0] years in 2003 to 26.0 [24.0-28.0] years in 2018), corresponding to our earlier observations that the age of male incident infections also remained similar during the observation period.

Over time, substantially fewer infections occurred in adolescent girls and young women aged 15-24 years. In 2003 the largest transmission flows were to women aged 15-24 years from male partners 0-6 years older (15.5% [12.3-18.9]) and from male partners 6+ years older (16.0% [12.7-19.2]) (Supplementary Table S7). By 2018, these transmission flows declined by approximately one third, with 8.1% [5.6-11.0], to women aged 15-24 years from male partners aged 0-6 years older, and 12.1% [9.3-15.2] to women aged 15-24 years from male partners aged 6+ years older. In those infections in adolescent girls and young women that occurred in 2018, the median age difference between incident infections in adolescent girls and young women and their transmitting male partners were 9.0 [7.0-12.0] years (Fig. 2b and Supplementary Table S7), similarly as in a phylogenetic study from KwaZulu-Natal in South Africa[31]. This prompted us to estimate for comparison age-specific sexual contact patterns within RCCS communities (Methods and Supplementary Table S8). In 2018, the median age difference between adolescent girls and young women and their male sexual partners was 3.6 [3.5-3.9] years. Our data thus indicate that the main transmission flow into adolescent girls and young women is through contacts with considerably older men as compared to their typical sexual contacts[31,32], and that while this transmission flow has weakened overall, it remains the predominant mode of infection in adolescent girls and young women.

By 2018, the largest share of transmission flows shifted to women aged 25-34 years, from male partners 0-6 years older. In 2003, transmissions to women 25-34 years from these transmitting partners accounted for 7.7% [6.2-9.3] of all transmissions, and by 2018 the share of these flows increased by half to 12.0% [9.1-15.0] (Supplementary Table S7). We also find that the transmission flows to women aged 35 years and above increased (Supplementary Table S7, also indicated by wider boxplots in Fig. 2b).

Our data suggest further deviations in age-specific transmission flows from the typical sexual contact patterns within study communities. For all women aged 30 years and older, we estimate their male transmitting partners were of similar age with a posterior interquartile age range of 30.3-38.0 years in 2018, whereas for comparison the typical sexual contact partners of these women were older with a posterior interquartile age range of 40.0-42.7. These findings explain the unexpected age profile of male transmitting partners (Fig. 3c) that concentrates in men aged 25 to 40 instead of extending to progressively older men (Extended Data Fig. 7). Our observations are in line with recent studies from Zambia[20] and South Africa[33] that show having a male partner aged 25-40 years rather than the age gap between partners is associated with increased transmission risk.

The transmission flows into men remained similar over time (Fig. 2b). In 2018, the largest transmission flow was to men aged 25-34 years from transmitting female partners of similar age that were 0-6 years older (10.6% [8.9-12.3]).

Gender gaps in viral suppression are increasing

We next placed the reconstructed shifts in transmission dynamics into the wider context of rapidly expanding HIV treatment during the observation period[14]. We measured viral load from 2011 (survey round 15) among almost all participants with HIV (Supplementary Tables S1 and Extended Data Fig. 8)[34]. Following WHO criteria[35], individuals with viral load measurements below 1,000 copies/millilitre (mL) plasma were considered viremic (Methods and Supplementary Table S9). By 2018, we find that the proportion of men and women who were viremic was entirely decoupled from HIV prevalence in that while the proportion of women with HIV was substantially higher than in men, the proportion of viremic women was similar or lower than in men (Fig. 3a). We quantified these trends with the male-to-female ratio of the proportion of viremic individuals relative to 2003 levels, which has been progressively increasing in all age groups (Fig. 3b). This suggests[36] that faster rises in female HIV suppression could explain in part the faster declines in male incidence rates as higher rates of ART uptake and virus suppression in women mean that male partners are less likely to become infected, whereas men's higher rates of unsuppressed virus mean they are more likely to transmit to female partners (Extended Data Fig. 9). These trends have by 2018 accumulated to a substantial gap in suppression levels in men compared to women (Table 1 and Fig. 3c).

Men contribute disproportionally to transmission

Combining phylogenetics with the virus suppression data also allowed us to compare transmission with population-level infectiousness as measured through viremic individuals (Table 1 and Fig. 2c). In 2018, the contribution of men to viremic individuals was (49.2% [44.3-54.1]). For the same time period we found that the contribution of men to transmission was consistently higher (62.8% [60.2-65.2]), indicating that men contribute more to transmsission than population-viral load suggests. These findings are compatible with generally higher viral load in men than women[34,37] that are expected to lead to higher transmission rates per sex act from men than women, heterogeneous contact patterns[38], higher biological susceptibility of women to HIV infection in heterosexual contacts[39,40], but also lower susceptibility of men to HIV infection following voluntary medical male circumcision[41].

Policy implications

It has been previously demonstrated that people with HIV who are on ART and maintain suppressed virus do not transmit HIV[42,43]. On this basis, we quantified the impact that closing the gap in male-female virus suppression levels could have had on HIV transmission flows. Specifically, we parameterised the transmission flow model in terms of HIV seronegative individuals who are susceptible to infection and individuals with unsuppressed HIV who remain infectious. Thus, we were able to use the fitted model to estimate the impact of fewer individuals with unsuppressed HIV on evolving HIV transmission in counterfactual, modelled intervention scenarios (see Methods). We considered the impact of three hypothetical scenarios: first, the impact of reducing by half the gap in the proportion of men with suppressed virus as compared to women ("closing half the suppression gap in men") at the end of the observation period in 2018 (Fig. 3c); second, the impact of achieving the same virus suppression levels in men with HIV as in women in 2018 ("closing the suppression gap in men"); and third—for reference—achieving the UNAIDS 95-95-95 target that 86% of men ($0.95 * 0.95 * 0.95$) with HIV reach viral suppression in all age groups in 2018[44]. Table 1 and Fig. 4a describe the age-specific male counterfactual viral suppression targets of each scenario, and place these into the context of prevalence, suppression, and transmission. Overall, we found slightly older men would have reached suppression in the scenarios closing the suppression gap as compared to the UNAIDS 95-95-95 scenario. We predict that in the UNAIDS 95-95-95 scenario, an additional 172.6 [136.8-210.0] men with HIV would have reached viral suppression in 2018 (Fig. 4b) and this would have resulted in a 58.4% [54.9-61.7] additional reduction in HIV incidence in women in 2018 (Fig. 4c), which is in good agreement with the contribution of 95-95-95 interventions to projected incidence reductions for all of Eastern and Southern Africa under the mathematical models used to inform the global HIV prevention strategy[45]. In the scenario closing half the suppression gap in men, an additional 75.1 [53.9-96.0] men with HIV would have reached viral suppression in 2018 and resulted in a 25.1% [24.2-26.2] additional reduction in HIV incidence in women in 2018. In the scenario closing the entire suppression gap

in men, an additional 150.2 [107.8-193.0] men with HIV would have reached viral suppression in 2018 and resulted in a 50.6% [48.6-52.8] additional reduction in HIV incidence in women in 2018 (Fig. 4b-c). Thus, all three intervention scenarios involved reaching a small additional number of men compared to the thousands of women with higher risk of HIV acquisition in the same rural and semi-urban study areas[46]. We predict that closing the suppression gap in men would have changed the female-to-male incidence rate ratio from 1.59 [1.38-1.82] to 0.78 [0.69-0.87] in 2018 (Fig. 4d), entirely closing the growing gender disparity in HIV incidence.

Discussion

Effective HIV interventions and services are essential to bring most African countries on track to end AIDS as a public health threat by 2030 and accelerate progress towards the vision of the UNAIDS “three Zeros" target: zero new HIV infections, zero discrimination, and zero AIDS-related deaths[45,47]. Gender inequalities are among the main reasons why global targets on mass scale-up of HIV testing, biomedical interventions and on incidence reductions have not been achieved[48]. Here, we combined population-based incidence with deep-sequence viral phylogenetic surveillance data to characterize how HIV incidence and heterosexual transmission sources have been changing by age and gender in a typical rural and semi-urban African setting. We show that along with increasing availability of HIV services, there have been consistently faster increases in viral suppression in women than men and an increasing majority of new infections are arising from men. We also document substantial age shifts in HIV incidence and transmission sources, with the primary burden of incidence shifting to older women aged 25-34 years, the primary burden of transmission shifting to male partners aged 30-39 years, and the relative contribution of transmission flows to adolescent girls and young women from older men reducing by one third. Modeling counterfactual improvements in HIV outcomes for men on the inferred transmission flows during the last survey round in 2016-2018, we find that closing the male gender gap in viral suppression rates could have reduced incident female infections by half in that time period and brought about gender equality in HIV infection burden.

This study evaluated data from one longitudinal surveillance cohort in southern Uganda, but the increasing gender disparities and shifts in age-specific transmission are not unique. Incidence data published over the last decade documents widespread declining incidence across the African continent[17], greater differences in rates of new infections between men and women over calendar time, and rising average age of infection in women[17]. Data from population surveillance studies and HIV treatment and prevention trials shows higher levels of viremia among men compared to women with HIV[49,50], and phylogenetic studies from Botswana[51] and Zambia[20] also report gender disparities in HIV transmission. Together, these observations suggest that the principal characteristics of the evolving HIV epidemic likely apply more broadly in similar rural and semi-urban populations across Eastern and Southern Africa.

Given that the African HIV epidemic has historically been concentrated among adolescent girls and young women[4,5], programs and policies rightfully have concentrated on reducing HIV risk in this demographic. Here, we document that most

heterosexual transmission is driven by men and that — as incidence is declining — the contribution of men to onward heterosexual transmission is growing, likely due to slower population-level declines in HIV viremia in men. While there are emerging efforts to design male-centered HIV interventions[52,53], African men continue to be overlooked in the design of programmatic services[54,55]. Many factors, including gender norms, mobility, and lack of targeted programming to men contribute to lower uptake of HIV services by men[53]. Case finding of men with HIV might be difficult but could be strengthened by expanding access to HIV testing services most likely to reach them, such as through self-testing or assisted partner notification and other social network strategies[54,56,57]. Retention of men with HIV in treatment and care programs could be improved through male-centered differentiated service delivery. It is well-established that improving male engagement in HIV services leads to better health for men[58,59]. We expect additional interventions such as voluntary medical male circumcision, condom promotion, or pre-exposure prophylaxis would lead to further reductions in new cases[60].

Our findings are grounded in fifteen years of consecutive population-based epidemiologic and molecular surveillance in southern Uganda, enabling us to measure changes in HIV incidence and transmission during a critical period of HIV service scale-up. Though it is typically assumed that age-specific patterns in onward HIV transmission correspond to those of viremia or follow typical sexual contact patterns, we find that this is not always the case. First, men contributed disproportionally more to onward heterosexual transmission than to viremia across all survey rounds during which viremia were measured (Fig. 2c and Extended Data Fig. 7a). Second, older women contributed less to transmission than viremia suggest, an observation that was consistent with attenuating sexual activity of women from age 25 onwards (Extended Data Fig. 7a). Third, young women and young men tended to be infected by transmitting partners who were substantially older than the typical sexual partners of the same population age group (Fig. 2b and Extended Data Fig. 7b). These findings illustrate the central utility of pathogen genomics to track and understand patterns of transmission, especially when interpreted in the context of population-based surveillance data, and when implemented at high enough sequence coverage to reconstruct directed transmission networks.

This study has important limitations. First, not all census-eligible individuals participated in the survey, primarily due to absence for work or school (Extended Data Fig. 1)[14]. We used data from first-time participants as proxies of non-participants, but we cannot rule out that non-participants include disproportionally larger populations of people with HIV and/or with different risk profiles. In this case, sensitivity analyses (Supplementary Table S10) indicate that more viremic men would have to be reached in all intervention scenarios for similar HIV incidence reductions in women as in Fig. 4. Second, we were only able to deep-sequence a fraction of all transmission events, and these may not be representative of all transmissions. We characterized sampling probabilities under the assumption that individuals were ever deep-sequenced at random within age- and gender strata, and found that the sampling probabilities did not differ substantially between strata in each round (Extended

Data Fig. 5), so that the estimated transmission flows were not sensitive to our sampling probability adjustments (Supplementary Table S10). Of course, these sampling adjustments are modeled and it remains possible that missing data could bias our findings. Third, our error analyses indicate that deep-sequence phylogenetics are not a perfect marker of direction of transmission, with estimated false discovery rates of 16.3% [8.8-28.3%] in this cohort.[22] Fourth, over time some communities were added and others left the Rakai Community Cohort Study (Supplementary Table S2). We repeated our analysis on the subset of 28 continuously surveyed communities, and found similar incidence and transmission dynamics (Supplementary Table S10). Fifth, our findings on rural and semi-urban populations may not extend to populations with different demographics, risk profiles or healthcare access, and this includes populations in urban or metropolitan areas or key populations.

This study demonstrates shifting patterns in HIV incidence and in the drivers of HIV infection in communities typical of rural and semi-urban East Africa, providing key data for evidence-informed policy making. We find incidence rates have dropped substantially in women aged 15-24 years from 2003 to 2018, and incidence rates now peak among women aged 25-35 years, consistent with cross-sectional national surveillance data from Uganda[61]. Shifts in women's incidence are the result of an increase in the age of transmitting male partners, and the primary contribution to HIV transmission lies now in men aged 30 and above. The growing contribution of men to heterosexual transmission is associated with substantially slower declines of unsuppressed viremia in men than women. We predict successful interventions centered on men that bring suppression rates in men on par with those in women could reduce incidence in women by half and close the gender gap in new infections. These findings reinforce calls for HIV prevention programming and services to give greater priority to reach and retain in care men with HIV as this will improve male health, substantially reduce incidence in women, and close gender gaps in infection burden.

Methods

Rakai Community Cohort Study

Longitudinal surveillance. Between September 2003 and May 2018, nine consecutive survey rounds of the Rakai Community Cohort Study (RCCS) were conducted in 36 inland communities in south-central Uganda (Fig. 1, Supplementary Tables S1-S2, and Supplementary Fig. S1). The results presented in this paper derive from data collected through these surveys, including the population census, the RCCS survey participants, the incidence cohort, and the phylogenetic transmission cohort.

RCCS survey methods have been reported previously[14,18]. In brief, for each survey round, the RCCS did a household census, and subsequently invited all individuals that were of age 15-49 years and residents for at least 1 month to participate in the open, longitudinal RCCS survey; and so data collection was not randomized, and data collection was blind relative to previous interactions with individuals or any personal characteristics apart from age and residency status, and any research questions. Eligible individuals first attended group consent procedures, and individual consent was

obtained privately by a trained RCCS interviewer. Following consent, participants reported in a private location, typically a tent at the survey hub, on demographics, behavior, health, and health service use. All participants were offered free voluntary counseling and HIV testing as part of the survey. Rapid tests at the time of the survey and confirmatory enzyme immunoassays were performed to determine HIV status. All participants were provided with pre-test and post-test counseling, and referrals of individuals who were HIV-positive for ART. Additionally, all consenting participants, irrespective of HIV status, were offered a venous blood sample for storage/future testing, including viral phylogenetic studies. Supplementary Table S1 summarises the characteristics of the RCCS participants and HIV-positive participants by age and gender. For the purpose of our analyses, we combined data from three pairs of geographically close areas in peri-urban settings into three communities, and 28 of 36 communities were continuously surveyed over all rounds (Supplementary Table S2). All epidemiologic data collected through RCCS are stored in a database running Microsoft SQL server 2019 and Microsoft Access version 2016.

Ethics declarations. The study was independently reviewed and approved by the Ugandan Virus Research Institute, Scientific Research and Ethics Committee, protocol GC/127/13/01/16; the Ugandan National Council of Science and Technology; and the Western Institutional Review Board, protocol 200313317. All study participants provided written informed consent at baseline and follow-up visits using institutional review board approved forms. This project was reviewed in accordance with CDC human research protection procedures and was determined to be research, but CDC investigators did not interact with human subjects or have access to identifiable data or specimens for research purposes. Participants in the RCCS received 10,000UGX (approximately 2.50USD) in compensation for the baseline and follow-up surveys.

Population size estimates. To characterize changes in population demography, individual-level data on the census-eligible individuals that were obtained during each census were aggregated by gender, 1-year age band (between 15 and 49 years) and survey round (Extended Data Fig. 1a-b, bars). The age reported by household heads in the census surveys tended to reflect grouping patterns towards multiples of 5, suggesting that household heads reported ages only approximately. For this reason, we smoothed population sizes across ages independently for every gender and survey round, using locally weighted running line smoother (LOESS) regression methods that fit multiple polynomial regressions in local neighborhoods as implemented in the `R` package `stats` version 3.6.2 with `span` argument set to 0.5 (Extended Data Fig. 1a-b, line). Model fit was assessed visually without a formal test, suggesting that the data met the assumptions of the statistical model.

Participation rates. To characterize participation rates, we calculated the proportion of RCCS participants in the census-eligible population by gender, 1-year age band and survey round (Extended Data Fig. 1c-d, bars). Following consent, participants reported either their birth date or current age themselves, and accompanying documentary evidence was requested. There were no obvious age grouping patterns of multiple of 5 among participants. Overall, participation rates were lower in men than women (63% vs. 75%). Participation rates also increased with age for both men

and women, and were very similar across survey rounds. Considering the grouping patterns by age in the population count data, we again smoothed the participation rates across ages independently for every gender and survey round using LOESS regression as specified above for population size estimation (Extended Data Fig. 1c-d, line). Model fit was assessed visually without a formal test, suggesting that the data met the assumptions of the statistical model.

HIV status and prevalence. All RCCS participants were offered free HIV testing. Prior to October 2011, HIV testing was performed through enzyme immunoassays (EIAs) with confirmation via Western Blot and DNA PCR. After October 2011, testing was performed through a combination of three rapid tests with confirmation of positives, weakly positives and discordant results by at least two EIAs and Western Blot or DNA PCR[62]. Overall, 99.7% participants took up the test offer across survey rounds, and Supplementary Table S1 documents the number of participants with HIV. From these survey data, we estimated HIV prevalence (i.e., probability for a participant to have HIV) with a non-parametric Bayesian model over the age of participants independently for both genders and survey round. Specifically, we used a binomial likelihood on the number of participants with HIV parameterized by the number of participants and HIV prevalence in each 1-year age band. The HIV prevalence parameter was modeled on the logit scale by the sum of a baseline term and a zero-mean Gaussian Process on the age space. The prior on the baseline was set to a zero-mean normal distribution with a standard deviation of 10. The covariance matrix of the Gaussian Process was defined with a squared exponential kernel, using a zero-mean half-normal prior with a standard deviation of 2 on the scale parameter of the squared exponential kernel and a zero-mean half-normal prior with a standard deviation of 11.3 $((49-15)/3)$ on the lengthscale of the squared exponential kernel. The model was fitted with `Rstan` release 2.21.0 using `Stan`'s adaptive Hamiltonian Monte Carlo (HMC) sampler[63] with 10,000 iterations, including warm-up 500 iterations. Convergence and mixing were good, with highest R-hat value of 1.0029, and lowest effective sample size of 830). The model represented the data well, with 98.57% of data points inside 95% posterior predictive intervals, indicating that the data met the assumptions of the statistical model. For the mathematical modeling of transmission flows, we assumed that age- and gender-specific HIV prevalence were the same in non-participants in the RCCS communities as in the participants in these communities.

ART use. The RCCS measures ART use through participant reports since survey round 11. Self-reported ART use reflected viral suppression with high specificity and a sensitivity around 70% in the study population (Supplementary Table S9). We took the following pre-processing steps. For survey round 10, we assumed self-reported ART use to have been on the same levels as in round 11. Next, the ART use field was adjusted to "yes" for the participants with HIV who did not report ART use but who had a viral load measurement below 1,000 copies per milliliter (mL) plasma blood. Further, we considered it likely that with increasingly comprehensive care and changing treatment guidelines[14,64], ART use in individuals with HIV who did not participate increased substantively over time, and this prompted us to consider as proxy of ART use in non-participants the observed ART use in first-time participants

with HIV. Overall, first-time participants represented between 15.26% to 39.87% of all participants. Extended Data Fig. 8a-b exemplifies the self-reported ART use data in male participants and male first-time participants, along with the combined estimate of individuals with HIV in the study population who report ART use, summing over participants and non-participants. These estimates were obtained using the same Bayesian non-parametric model as for HIV prevalence. Convergence and mixing were good, with highest R-hat value of 1.0025 and lowest effective sample size of 978 for the participants and 1.0027, 521 respectively for first-time participants. The model represented the data well, with 99.67% of data points for the participants inside the corresponding 95% posterior predictive intervals, and 99.24% for the first-time participants, indicating that the data met the assumptions of the statistical model. The resulting, estimated ART use rates in infected men and women are shown in Extended Data Fig. 8c.

Viral suppression. Since survey round 15, HIV-1 viral load was measured on stored serum/plasma specimens from infected participants using the Abbott real-time m2000 assay (Abbott Laboratories, IL, USA), which is able to detect a minimum of 40 copies/mL. Viral suppression was defined as a viral load measurement below 1,000 copies/mL plasma blood following recommendations of the World Health Organisation (WHO)[35]. To estimate virus suppression levels in the infected non-participants, we considered again as proxy data on infected first-time participants. Overall, viral load measurements were obtained from 19.3% of participants with HIV in survey round 15 and nearly all (>97.71%) participants with HIV since survey round 16[65,66,67]. From these data we estimated the proportion of individuals in the study population with HIV who had suppressed virus, summing over participants and non-participants, using the same Bayesian non-parametric model as for HIV prevalence and ART use. Convergence and mixing were good with lowest R-hat value of 1.0016 and lowest effective sample size of 461 for the participants and 1.0052, 844 respectively for the first-time participants. The model represented the data well, with 98.19% of data points inside 95% posterior predictive intervals and 97.99% for the first-time participants, indicating that the data met the assumptions of the statistical model. For the purpose of mathematical modeling of transmission flows, we next considered the earlier survey rounds 10 to 14, for which viral load measurements were not available. On average, 93% of individuals reporting ART use also had suppressed virus (Supplementary Table S9), leading us to estimate the number of individuals with suppressed virus before 2011 from corresponding ART use data. Specifically, we estimated the proportion of the study population with HIV that was virally suppressed by adjusting the estimated ART use data with the sensitivity of being virally suppressed given self-reported ART use and the specificity of being virally suppressed given self-reported no ART use estimated from round 15 where available, and otherwise from round 16 (Supplementary Table S9). Specificity and sensitivity values by 1-year age bands were linearly interpolated between the midpoints of the age brackets in Supplementary Table S9. The resulting, estimated virus suppression levels in men and women with HIV are shown in Extended Data Fig. 8d, illustrating that the gap in virus suppression levels increased over time.

Sexual behavior. RCCS participants reported to interviewers in each round on aspects of sexual behavior, including the number of sexual partners in the past 12 months within the same community, the number of partners outside the community, and in round 15 also demographic characteristics of up to four partners (Supplementary Table S8). To interpret HIV transmission flows in the context of typical sexual contact networks, we focused on the detailed behavior data collected in round 15 and estimated sexual contact intensities between men and women by 1-year age band, defined as the expected number of sexual contacts of one individual of gender g and age a with the population of the opposite gender h and age b in the same community. Estimates were obtained with the Bayesian rate consistency model, version 1.0.0, using default prior specifications[68]. We noted along with previous work[69,70,71,72] that women tended to report considerably fewer contacts than men (Supplementary Table S8), prompting us to include in the linear predictor of contact rates additional age-specific random effects to capture under-reporting behavior in women. Further, community-specific baseline parameters were added to allow for variation in the average level of contact rates in each community, but the age-specific structure of contact rates was assumed to be identical across communities. The resulting model was fitted to all data pertaining to within-community sexual contacts in the last year, including reports of within-community contacts for which information on the partners remained unreported. Contacts reported with partners from outside the same community were excluded, because male-female contacts have to add up to female-male contacts only in the same population denominator, and hence under-reporting could only be adjusted for when within-community contacts are considered. The model was fitted with `CmdstanR` version 0.5.1[73] using Stan's adaptive HMC sampler[63] with 4 chains, where each chain runs 2800 iterations, including 300 warm-up iterations. Convergence and mixing were good, with highest R-hat value of 1.003, and lowest effective sample size of 1,745. The model represented the data well, with $> 99\%$ of data points inside 95% posterior predictive intervals, indicating that the data met the assumptions of the statistical model. Supplementary Table S8 reports the estimated sexual contact intensities from men and women in survey round 15, and shows that the estimated, under-reporting adjusted sexual contact intensities in women were considerably higher than those directly reported. The table also shows that the estimated number of sexual contacts from men to women equal those from women to men, and the estimated age distribution of sexual contacts is shown in Fig. 2 and Extended Data Fig. 7.

Longitudinal HIV incidence cohort

Data and outcomes from the incidence cohort. The RCCS encompasses both a full census of the study communities and a population-based survey in each surveillance round, which enables identification and follow up of unique individuals over time, and thus provides a comprehensive sampling frame to measure HIV incidence. The RCCS incidence cohort comprises of all RCCS study participants who were HIV-negative at their first visit (baseline) and had at least one subsequent follow-up visit (Supplementary Fig. S1). Individuals in the incidence cohort were considered to be at risk of acquiring HIV after their first visit, and stopped accruing risk at the date

of HIV acquisition or the date of last visit. Exposure times were estimated from data collected at survey visit times similarly as in [14]. Individuals in the incidence cohort who remained negative until the last survey round contributed their time between the first and last survey visit to their exposure period. Individuals in the incidence cohort who were found to have acquired HIV must have done so between the visit date of the last round in which they were negative and the visit date of the current round, and the infection date was imputed at random between the two dates. This included incident cases who had no missed visit between the last negative and current visit (type 1) or one missed visit (type 2) as in [14], but also cases who had more than one missed visit (type 3). Unknown dates were imputed at random 50 times, and individual exposure periods and incident cases were then attributed to each survey round, summed over the cohort, and then averaged over imputations. Supplementary Table S3 and Extended Data Fig. 2 illustrate the age- and gender-specific exposure times and incidence events in each survey round. In sensitivity analyses, we considered only those individuals in the incidence cohort who resided in one of the 28 inland communities that were continuously surveyed across all rounds 10 to 18, and found similar incidence dynamics with slightly faster declines in incidence rates in younger men, although this difference was not statistically significant. No statistical methods were used to pre-determine sample sizes but our sample sizes are similar to those reported in previous publications [14].

Modelling and analysis. The primary statistical objective was to estimate longitudinal age-specific HIV incidence rates by 1-year age bands across (discrete) survey rounds, separately for each gender. We used a log-link mixed-effects Poisson regression model, with individual-level exposure times specified as offset on the log scale, common baseline fixed effect, and further random effects. The random effects comprised a one-dimensional smooth function on the age space, a one-dimensional smooth function on the survey round space, and an interaction term between age and survey round. The functions were specified as one-dimensional Gaussian processes, similar as in the model for estimating HIV prevalence. Alternative specifications, including two-dimensional functions over the participant's age and survey round, and without interaction terms between age and survey rounds were also tried. We did not consider incidence trends in continuous calendar time because study communities were surveyed in turn, and so the incidence data within each round are structured by communities, which would require further modeling assumptions to account for. Due to the large number of individual observations, models were fitted using maximum-likelihood estimation (MLE) with the `R` package `mgcv` version 1.8-38 in the `R` language [74], to each of the 50 data sets with imputed exposure times for each gender. Numerical convergence was examined with the `gam.check` function. Within and between sample uncertainties in parameter estimates, from the variability of the estimation procedure and the data imputation procedure, were incorporated in the age-, gender- and survey round-specific incidence rate estimates by drawing 1,000 replicate incidence rate estimates from the MLE model parameters and associated standard deviation obtained on each of the 50 imputation data sets, and then calculating median estimates and 95% prediction intervals over the $1,000 \times 50$ Monte

Carlo estimates (Fig. 1c). Model fits were evaluated by comparing predicted HIV incidence infections estimates to the empirical data. To assess model fit, incident cases were predicted using the Poisson model parameterised by replicate MLE incidence estimates. Overall, model fit was very good, with 98.80% [98.10-99.49] data points inside the 95% prediction intervals across all 50 imputed data sets and the fitted model was consistent with all the available data (Extended Data Fig. 6), indicating that the data met the assumptions of the statistical model. The Akaike information criterion was used to identify the best model for each gender, and the best model was as described above (Supplementary Table S4).

Longitudinal viral phylogenetic transmission cohort

Data from the transmission cohort. Within the RCCS, we also performed population-based HIV deep-sequencing spanning a period of more than 6 years, from January 2010 to April 2018. The primary purpose of viral deep sequencing was to reconstruct transmission networks and identify the population-level sources of infections, thus complementing the data collected through the incidence cohort.

The RCCS viral phylogenetic transmission cohort comprises of all participants with HIV for whom at least one HIV deep sequence sample satisfying minimum quality criteria for deep-sequence phylogenetic analysis is available (Supplementary Fig. S1). For survey rounds 14 to 16 (PANGEA-HIV 1), viral sequencing was performed on plasma samples from participants with HIV who had no viral load measurement and self-reported being ART-naïve at the time of the survey, or who had a viral load measurement above 1,000 copies/mL plasma. We used this criterion because viral deep sequencing was not possible within our protocol on samples with virus less than 1,000 copies/mL plasma, and because self-reported ART use was in this population found to be a proxy of virus suppression with reasonable specificity and sensitivity[14,22]. Plasma samples were shipped to University College London Hospital, London, United Kingdom, for automated RNA sample extraction on QIAsymphony SP workstations with the QIAsymphony DSP Virus/ Pathogen Kit (Cat. No. 937036, 937055; Qiagen, Hilden, Germany), followed by one-step reverse transcription polymerase chain reaction (RT-PCR)[75]. Amplification was assessed through gel electrophoresis on a fraction of samples, and samples were shipped to the Wellcome Trust Sanger Institute, Hinxton, United Kingdom for HIV deep-sequencing on Illumina MiSeq and HiSeq platforms in the DNA pipelines core facility. Primers are publicly available[75]. For survey rounds 17 to 18 (PANGEA-HIV 2), viral load measurements were available for all infected participants and viral sequencing was performed on plasma samples of individuals who had not yet been sequenced and who had a viral load measurement above 1,000 copies/mL plasma. Plasma samples were shipped to the Oxford Genomics Centre, Oxford, United Kingdom, for automated RNA sample extraction on QIAsymphony SP workstations with the QIAsymphony DSP Virus/ Pathogen Kit (Cat. No. 937036, 937055; Qiagen, Hilden, Germany), followed by library preparation with the SMARTer Stranded Total RNA-Seq kit v2 - Pico Input Mammalian (Clontech, TaKaRa Bio), size selection on the captured pool to eliminate fragments shorter than 400 nucleotides (nt) with streptavidin-conjugated beads[76] to enrich the library with fragments desirable for

deep-sequence phylogenetic analysis, PCR amplification of the captured fragments, and purification with Agencourt AMPure XP (Beckman Coulter), as described in the veSEQ-HIV protocol[77]. Sequencing was performed on the Illumina NovaSeq 6000 platform at the Oxford Genomics Centre, generating 350 to 600 base pair (bp) paired-end reads. Sequencing probes are publicy available[78]. A subset of samples from survey rounds 14 to 16 with low quality read output under the PANGEA-HIV 1 procedure was re-sequenced with the veSEQ-HIV protocol. To enhance the genetic background used in our analyses, additional samples from the spatially neighboring MRC/UVRI/LSHTM surveillance cohorts and other RCCS communities were also included. For sequencing, the following software were used, QuantStudio Real-Time PCR System v1.3, Agilent TapeStation Software Analysis 4.1.1, Clarity Version 4.2.23.287, FreezerPro 7.4.0-r14598, and LabArchives Electronic Lab Notebook 2023. We restricted our analysis to samples from 2,172 individuals that satisfied minimum criteria on read length and depth for phylogeny reconstruction and subsequent inferences. Specifically, deep-sequencing reads were assembled with the `shiver` sequence assembly software, version 1.5.7[79]. Next, `phyloscanner` version 1.8.1[26] was used to merge paired-end reads, and only merged reads of at least 250 bp in length were retained in order to generate 250bp deep-sequence alignments as established in earlier work[22].

Deep-sequencing was performed from 2010 (survey round 14) onwards, but because sequences provide information on past and present transmission events, we also obtained information on transmission in earlier rounds and calculated sequence coverage in participants that were ever deep-sequenced at minimum quality criteria for phylogenetic analysis. Specifically, we required that individuals had a depth of ≥ 30 reads over at least 3 non-overlapping 250bp genomic windows. Individuals who did not have sequencing output meeting these criteria were excluded from further analysis, and these were largely individuals sequenced only in PANGEA-HIV 1, and were primarily associated with low viral load samples[77,80]. In total, we deep-sequenced virus from 1,978 participants with HIV of who 559 were also in the incidence cohort. Supplementary Table S5 characterizes HIV deep-sequencing outcomes in more detail. No statistical methods were used to pre-determine sample sizes but our sample sizes are similar to those reported in previous publications[20,28,51].

Reconstruction of transmission networks and source-recipient pairs. The HIV deep-sequencing pipeline provided sequence fragments that capture viral diversity within individuals, which enables phylogenetic inference into the direction of transmission from sequence data alone[22,79,81]. First, potential transmission networks were identified, and in the second step transmission networks were confirmed and the transmission directions in the networks were characterized as possible. In this study, the first step was modified from previous protocols[22] to ease computational burden, while the second step was as before performed with `phyloscanner`, using version 1.8.1.

In the first step[82], to identify potential transmission networks, HIV consensus sequences were generated as the most common nucleotide in the aligned deep-sequence fragments that were derived for each sample. We then calculated similarity scores between all possible combinations of consensus sequences in consecutive 500 bp

genomic windows rather than the entire genome to account for the possibility of recombination events and divergent virus in parts of the genome. Similarity score thresholds to identify putative, genetically close pairs were derived from data of long-term sexual partners enrolled in the RCCS cohort similarly as in [22,82], and then applied to the population-based sample of all possible combinations of successfully sequenced individuals. Overall, 2525 putative, genetically close individuals were identified, and these formed 305 potential transmission networks.

In the second step, we confirmed the potential transmission networks in phylogenetic deep-sequence analyses. We updated the background sequence alignment used in `phyloscanner` to a new sequence data set that included 113 representatives of all HIV subtypes and circulating recombinant forms and 200 near full-genome sequences from Kenya, Uganda, and Tanzania, obtained from the Los Alamos National Laboratory HIV Sequence Database (http://www.hiv.lanl.gov/). The deep-sequence alignment options were updated to using `MAFFT` version 7.475 with iterative refinement [83], and additional iterative re-alignment using consistency scores in case a large proportion of gap-like columns in the first alignment was detected. Deep-sequence phylogeny reconstruction was updated to using `IQ-TREE` version 2.0.3 with GTR+F+R6 substitution model, resolving the previously documented deep-sequence phylogenetics branch length artefact [20,84]. Confirmatory analyses of the potential transmission networks were updated to using `phyloscanner` version 1.8.1 with input argument *zeroLengthAdjustment* set to TRUE. From `phyloscanner` output, we calculated pairwise linkage scores that summarise how frequently viral phylogenetic subgraphs of two individuals were adjacent and phylogenetically close in the deep-sequence phylogenies corresponding to all 250bp genomic windows that contained viral variants from both individuals [22,26]. Similarly we calculated pairwise direction scores that summarise how frequently viral phylogenetic subgraphs of one individual were ancestral to the subgraphs of the other individual in the deep-sequence phylogenies corresponding to all 250bp genomic windows that contained viral variants from both individuals and in which subgraphs had either ancestral or descendant relationships [22,26]. Phylogenetically likely source-recipient pairs with linkage scores ≥ 0.5 and direction scores ≥ 0.5 were extracted, and only the most likely source-recipient pair with highest linkage score was retained if multiple likely sources were identified for a particular recipient. The resulting source-recipient pairs were checked further against sero-history data from both individuals where available. If sero-history data indicated the opposite direction of transmission, the estimated likely direction of transmission was set to that indicated by sero-history data.

Infection time estimates. The shape and depth of an individual's subgraph in deep-sequence phylogenies also provide information on the time since infection, and since the sequence sampling date is known thus also on the infection time [85] and the age of both individuals at the time of the infection event. We used the `phyloTSI` random forest estimation routine with default options, which was trained on HIV seroconverter data from the RCCS and other cohorts, and uses as input the output of the `phyloscanner` software [27]. Individual-level time since infection estimates were associated with wide uncertainty (Extended Data Fig. 4a), and for this reason we refined estimates for the phylogenetically likely recipient in source-recipient pairs

using the inferred transmission direction, age data, and where available longitudinal sero-history data. Specifically, we refined plausible infection ranges as indicated in the schema in Supplementary Fig. S2. Here, the dotted red rectangle illustrates the 2.5% and 97.5% quantiles of the `phyloTSI` infection time estimates for the phylogenetically likely recipient (x-axis) and transmitting partner (y-axis). We incorporated evidence on the direction of transmission by requiring that the date of infection of the phylogenetically likely recipient is after that of the transmitting partner (filled red triangle). Sero-history and demographic data were incorporated as follows. For both the recipient and the transmitting partner, the upper bound of the infection date was set as the 30th day prior to the first positive test of the participant[86]. The lower bound of the infection date was set to the largest of the following dates, the date of last negative test if available, the 15th birthday, or the date corresponding to 15 years prior the upper bound[87]. The refined uncertainty range of the infection time estimates of the phylogenetically likely transmitting partner and recipient are illustrated as the purple triangle in the schema above, and obtained as follows. Firstly, we defined individual-level plausible ranges, by intersecting the range of dates consistent with the `phyloTSI` predictions and sero-history data. If the intersection was empty, we discarded the `phyloTSI` estimates. Then we intersected the rectangle given by the cartesian product of the plausible intervals for source and recipient with the half-plane consistent with the direction of transmission. Finally, infection dates were sampled at random from the refined uncertainty range, so that the median infection date estimates correspond to the center of gravity of the triangle (cross). In sensitivity analyses, we further integrated estimates of transmission risk by stage of infection[88], though this had limited impact on the estimates (see Sensitivity analyses section below). In cases where the likely transmitting partner in one heterosexual pair was the recipient partner in another heterosexual pair, the above infection date refinement algorithm was applied recursively so that the refined infection date estimates were consistent across pairs. Finally, the transmission events captured by each source-recipient pair were attributed to the survey round into which the posterior median infection time estimate of the recipient fell, and in cases where the median estimate fell after the start time of a round and the end time of the preceding round, the event was attributed to the preceding round.

In total, we identified 539 source-recipient pairs that involved participants from the 36 survey communities and further individuals from the background data set. In 13 of the 539 source-recipient pairs, available dates of last negative tests indicated that only the opposite transmission direction was possible and in these cases the inferred direction of transmission was set to the opposite direction. The resulting pairs included 501 unique recipient partners, and for reach we retained the most likely transmitting partner. To identify pairs capturing transmission events within the RCCS inland communities, we restricted analysis initially to 236 heterosexual source-recipient pairs in whom both individuals were ever resident in the 36 survey communities. Of these, 142 pairs were from men to women and 94 from women to men. Infection times were estimated for all sampled individuals and refined for the recipient partners in the 236 heterosexual source-recipient pairs. For 4 recipient

partners, the `phyloTSI` estimates were ignored as they were incompatible with inferred transmission direction and survey data, and was based on sero-history data only. The phylogenetically most likely location of both individuals at time of transmission was estimated as their location at the RCCS visit date that was closest to the posterior median infection time estimate. Using this location estimate, 233 of the 236 heterosexual source-recipient pairs were estimated to capture transmission events in RCCS inland communities and were retained for further analysis. A further 6 recipient partners had posterior median infection time estimates outside the observation period from September 2003 to May 2018 and were excluded, leaving for analysis 227 heterosexual source-recipient pairs that captured transmission events in RCCS inland communities during the observation period. This excluded 88 potential source-recipient pairs from our study due to ethical considerations and prior analyses suggesting these pairs most likely represent partially sampled transmission chains (i.e., "false positives")[22].

Transmission flow analysis

Statistical framework. We next estimated the sources of the inferred population-level HIV incidence dynamics from the dated, source-recipient pairs in the viral phylogenetic transmission cohort. Overall, inference was done in a Bayesian framework using a semi-parametric Poisson flow model similar to Xi, X. *et al.*[29], that was fitted to observed counts of transmission flows $Y_{p,i,j}^{g\to h}$ with transmission direction $g \to h$ (male-to-female or female-to-male), time period p (R10-R15 and R16-R18) in which the recipient was likely infected, and 1-year age bands i, j of the source and recipient populations respectively, where

$$i, j \in \mathcal{A} = \{15, 16, \ldots, 48, 49\} \tag{1a}$$

$$(g \to h) \in \mathcal{D} = \{\text{male-to-female}, \text{female-to-male}\}. \tag{1b}$$

The target quantity of the model is the expected number of HIV transmissions in the study population in transmission direction $g \to h$ (male-to-female or female-to-male), survey round r (R10 to R18) in which infection occurred, and 1-year age bands i, j of the source and recipient populations respectively, which we denote by $\lambda_{r,i,j}^{g\to h}$. We considered that the expected number of HIV transmissions in the study population is characterized by transmission risk and modulated by the number of infectious and susceptible individuals, which prompted us to express $\lambda_{r,i,j}^{g\to h}$ in the form of a standard discrete-time susceptible-infected (SI) model,

$$\lambda_{r,i,j}^{g\to h} = \beta_{r,i,j}^{g\to h} \times S_{r,j}^{h} \times I_{r,i}^{g} \times \left|\left(t_r^{\text{end}} - t_r^{\text{start}}\right)\right|, \tag{2}$$

where $\beta_{r,i,j}^{g\to h} > 0$ is the transmission rate exerted by one infected, virally unsuppressed individual of gender g and age i on one person in the uninfected ("susceptible") population of the opposite gender h and age j in a standardized unit of time in round r. With model (2), we express expected transmission flows with a population-level mechanism of how transmission rates from individuals with unsuppressed HIV act on the susceptible population, and we preferred model (2) over a

purely phenomenological model of the $\lambda_{r,i,j}^{g\to h}$ for the generalizing insights it provides. The main simplifying approximations in (2) are that all quantities on the right-hand side of (2) are in discrete time and constant in each round, meaning we approximate over changes in population size, HIV prevalence, and viral suppression at a temporally finer scale, and assume further that one generation of transmissions occurs from individuals with unsuppressed HIV in each round. Importantly, in this framework, we can then relate the expected transmission flows to the HIV incidence dynamics and the data from the longitudinal incidence cohort by summing in (2) over the sources of infections,

$$\sum_i \lambda_{r,i,j}^{g\to h} = \left(\sum_i \beta_{r,i,j}^{g\to h} \times I_{r,i}^{g}\right) \times S_{r,j}^{h} \times \left|\left(t_r^{\text{end}} - t_r^{\text{start}}\right)\right| \tag{3a}$$

$$=: \kappa_{r,j}^{h} \times S_{r,j}^{h} \times \left|\left(t_r^{\text{end}} - t_r^{\text{start}}\right)\right|, \tag{3b}$$

where $\kappa_{r,j}^{h}$ is the incidence rate per census-eligible, susceptible person of gender h and age j in round r ($S_{r,j}^{h}$) and per unit time ($|(t_r^{\text{end}} - t_r^{\text{start}})|$). Estimates of $\kappa_{r,j}^{h}$ were calculated in units of 100 person-years as described above and shown in Fig. 1c, and we will constrain the semi-parametric Poisson flow model using these estimates. From the model output, we are primarily interested in the transmission flows and transmission sources during each round as quantities out of 100%, defined respectively by

$$\pi_{r,i,j}^{g\to h} = \lambda_{r,i,j}^{g\to h} \Big/ \Big(\sum_{i,j\in\mathcal{A},(g\to h)\in\mathcal{D}} \lambda_{r,i,j}^{g\to h}\Big) \tag{4a}$$

$$\delta_{r,i,j}^{g\to h} = \pi_{r,i,j}^{g\to h} \Big/ \Big(\sum_{k\in\mathcal{A}} \pi_{r,k,j}^{g\to h}\Big) \tag{4b}$$

$$\delta_{r,i}^{g\to h} = \sum_{j\in\mathcal{A}} \pi_{r,i,j}^{g\to h}\,. \tag{4c}$$

In words, (4b) quantifies the sources of infection in individuals of gender h and age j in round r such that the sum of $\delta_{r,i,j}^{g\to h}$ over i equals one, and (4c) quantifies the sources of infection in the entire population in round r that originate from the group of individuals of gender g and age i such that the sum of $\delta_{r,i}^{g\to h}$ over g and i equals one. The width of the boxplots in Fig. 2b shows (4b) and Fig. 2a, c show (4c).

Specification of susceptible and infected individuals. The number $S_{r,j}^{h}$ of the susceptible population of gender h and age j was calculated by multiplying the smoothed estimate $N_{r,j}^{g}$ of the census-eligible population of gender h and age j (shown in Extended Data Fig. 1a-b) with 1 minus the posterior median estimate of HIV prevalence $\rho_{r,j}^{h}$ in census-eligible individuals of gender h and age j of round r (calculated as described further above). To specify the number $I_{r,i}^{g}$ of individuals with unsuppressed HIV of gender g and age i, we multiplied the smoothed estimate $N_{r,i}^{g}$ of the census-eligible population of gender g and age i of round r (shown in Extended

Data Fig. 1a-b) with the posterior median estimate of HIV prevalence in the census-eligible population of gender g and age i ($\rho^g_{r,i}$) with 1 minus the posterior median estimate $\nu^g_{r,i}$ of the proportion of census-eligible individuals of gender g and age i in round r that have suppressed HIV (calculated as described further above and shown in Extended Data Fig. 8d). The start and end times of each survey round, t_r^{start} and t_r^{end} were set as shown in Fig. 1b and specified in units of years, so that the transmission intensity is also expressed in units of years.

Bayesian model. We first present the likelihood of the observed counts of transmission flows $Y^{g\to h}_{p,i,j}$ under the semi-parametric Poisson flow model that is parameterized in terms of (2). The phylogenetically reconstructed source-recipient pairs capture only a subset of incidence events, and so it is important to characterize the sampling frame. As in Xi, X. *et al.*[29], we consider the unknown transmission events $Z^{g\to h}_{r,i,j}$ in round r and assume these are sampled at random within each strata with probabilities that factorise into sampling probabilities of sources of age i and gender g and sampling probabilities of recipients of age j and gender h, $Y^{g\to h}_{r,i,j} \sim \text{Binomial}(Z^{g\to h}_{r,i,j}, \xi^1_{r,g,i}\xi^2_{r,h,j})$. Using (4a), we also let $Z^{g\to h}_{r,i,j} \sim \text{Multinomial}(Z_r, \pi^{g\to h}_{r,i,j})$, where Z_r is the total number of infection events in round r.

Because we have data from both the transmission and incidence cohorts, we are able to constrain the sampling problem with the detection probabilities of incidence events. Specifically, setting $Y^h_{r,j} = \sum_{i\in\mathcal{A}} Y^{g\to h}_{r,i,j}$ and $Z^h_{r,j} = \sum_{i\in\mathcal{A}} Z^{g\to h}_{r,i,j}$, we let $Y^h_{r,j} \sim \text{Binomial}(Z^h_{r,j}, \zeta^h_{r,j})$ and set the detection probability to the proportion of the expected number of incident cases of gender h and age j that could be phylogenetically reconstructed in time period p,

$$\zeta^h_{r,j} = \zeta^h_{p,j} = \Big(\sum_{r\in p, i\in\mathcal{A}} Y^{g\to h}_{r,i,j}\Big) \Big/ \Big(\sum_{r\in p} \kappa^h_{r,j} \times S^h_{r,j} \times |(t_r^{\text{end}} - t_r^{\text{start}})|\Big), \tag{5}$$

for all rounds r in the two time periods R10-R15 and R16-R18. We focused in (5) on time periods due to the limited phylogenetic count data. The advantage in constraining the transmission model with the detection probabilities (5) is that the estimates of the transmission model will be consistent with the incidence dynamics that we already estimated with data from the incidence cohort. Re-arranging terms between Binomial and Multinomial models, we obtain

$$Y^{g\to h}_{r,i,j} \sim \text{Multinomial}\Big(Z_r,\ \xi^1_{r,g,i}\,\xi^2_{r,h,j}\,\pi^{g\to h}_{r,i,j}\Big) \tag{6a}$$

$$\xi^2_{r,h,j} = \frac{\sum_{i\in\mathcal{A}} \pi^{g\to h}_{r,i,j}}{\sum_{i\in\mathcal{A}} \xi^1_{r,g,i}\pi^{g\to h}_{r,i,j}}\,\zeta^h_{r,j}, \tag{6b}$$

which shows that the sampling probabilities of recipients $\xi^2_{r,h,j}$ can be expressed in terms of the detection probability of infection events, weighted by the relative contribution and sampling of source-specific transmission events to the same incidence group. We still need to specify $\xi^1_{r,h,i}$ to complete the sampling model. Here, we approximated the sampling probability of sources with the proportion of individuals of

age i and gender h with unsuppressed virus in round r that were ever deep-sequenced. Note that the sampling model (6) will alter the posterior mean transmission flows $\pi^{g\to h}_{r,i,j}$ only when the sampling probabilities $\xi^1_{r,h,i}$ and $\zeta^h_{r,j}$ differ between age and gender strata in the same round. Extended Data Fig. 5 visualizes our specifications of $\zeta^h_{r,j}$ and $\xi^1_{r,h,j}$, and shows that the sampling differences between age and gender groups are relatively modest in any given round, which suggests that the adjustments on the inferred transmission flows based on our modelled sampling probabilities will be modest.

In the semi-parametric Poisson flow model of Xi, X. *et al.*[29], the sampling model (6) can be analytically integrated out based on standard thinning properties, which in turn allows us to express the likelihood of observing the phylogenetic data with

$$Y^{g\to h}_{p,i,j} \sim \text{Poisson}\left(\sum_{r\in p} \xi^1_{r,g,i}\, \xi^2_{r,h,j} \lambda^{g\to h}_{r,i,j}\right) \tag{7a}$$

$$\lambda^{g\to h}_{r,i,j} = \beta^{g\to h}_{r,i,j} \times S^h_{r,j} \times I^g_{r,i} \times \left|\left(t^{\text{end}}_r - t^{\text{start}}_r\right)\right| \tag{7b}$$

$$\log \beta^{g\to h}_{r,i,j} = \hat{\boldsymbol{c}}^{g\to h}(i,j) + \gamma_0 + \gamma_g + \gamma_r + \gamma_{p(r)} + \boldsymbol{f}^{g\to h}_0(i,j) + \boldsymbol{f}^{g\to h}_r(j) + \boldsymbol{f}^{g\to h}_{p(r)}(i) \tag{7c}$$

$$\xi^2_{r,h,j} = \frac{\sum_{i\in\mathcal{A}} \lambda^{g\to h}_{r,i,j}}{\sum_{i\in\mathcal{A}} \xi^1_{r,g,i} \lambda^{g\to h}_{r,i,j}}\, \zeta^h_{r,j}, \tag{7d}$$

where $\hat{\boldsymbol{c}}^{g\to h}(i,j)$ is the posterior median estimate of the log rate of sexual contacts within communities in one year between one person of age i and gender g and one person of age j and gender h that we estimated from the sexual behavior data, and the remaining terms quantify the transmission probability per sexual contact on the log scale. The model is designed in such a way that the log sexual contact rates describe a fixed age-specific non-zero mean surface, and the remaining parameters describe age-specific random deviations around the mean surface. With this approach, any inferred deviations in transmission rates relative to sexual contact rates are informed by the phylogenetic data and robust to prior specifications on the random deviations. Specifically, γ_0 is the baseline parameter characterizing overall transmission risk per sexual contact, γ_g is a gender-specific offset which is set to zero in the female-to-male direction and a real value in male-to-female direction, γ_r a round-specific offset which is set to zero for the first survey round 10, and γ_p is a time period specific offset which is set to zero for the first time period. We assume the age-specific structure of transmission rates in terms of the transmitting partners (denoted by i) and recipients (denoted by j) are similar across similar ages, and so we can exploit regularising prior densities[29] to learn smooth, latent transmission rate surfaces from the sparse data shown in Extended Data Fig. 3. In detail, we modeled the age-specific structure of transmission rates non-parametrically with 2 time-invariant random functions $\boldsymbol{f}^{g\to h}_0$ with two-dimensional inputs on the domain $[15, 50] \times [15, 50]$ that characterize age-age interactions in transmission risk for each gender, 2×8 random functions

$\boldsymbol{f}_r^{g\to h}$ with one-dimensional inputs that characterize time trends in the age of recipients for each gender for survey rounds after round 10, and 2 random functions $\boldsymbol{f}_p^{g\to h}$ with one-dimensional inputs that characterize time trends in the age of transmitting partners for each gender for the second time period. We attach to each of these random functions computationally efficient B-splines projected Gaussian process (GP) priors[89], which we constructed by describing the random functions with cubic B-splines over equidistant knots and modeling the prior relationship of the B-splines parameters with GPs with squared exponential kernels with variance and length-scale hyper-parameters, denoted respectively by σ^2 and ℓ. The prior densities of our Bayesian model are

$$\gamma_0 \sim \mathcal{N}(0, 10^2) \tag{8a}$$
$$\gamma_{\text{male}} \sim \mathcal{N}(0, 1) \tag{8b}$$
$$\gamma_r \sim \mathcal{N}(0, 1) \qquad \text{for } r > \text{R10} \tag{8c}$$
$$\gamma_p \sim \mathcal{N}(0, 1) \qquad \text{for } p = \text{R16-R18} \tag{8d}$$
$$\boldsymbol{f}_0^{g\to h} \sim \text{2D-B-splines-GP}(\sigma_0^{g\to h}, \ell_{0,i}^{g\to h}, \ell_{0,j}^{g\to h}) \tag{8e}$$
$$\boldsymbol{f}_r^{g\to h} \sim \text{1D-B-splines-GP}(\tilde{\sigma}_r^{g\to h}, \tilde{\ell}_r^{g\to h}) \qquad \text{for } r > \text{R10} \tag{8f}$$
$$\boldsymbol{f}_p^{g\to h} \sim \text{1D-B-splines-GP}(\breve{\sigma}^{g\to h}, \breve{\ell}^{g\to h}) \qquad \text{for } p = \text{R16-R18} \tag{8g}$$
$$\sigma_{0,i}^{g\to h}, \sigma_{0,j}^{g\to h}, \tilde{\sigma}^{g\to h}, \breve{\sigma}^{g\to h} \sim \text{Half-Cauchy}(0, 1) \tag{8h}$$
$$\ell_{0,i}^{g\to h}, \ell_{0,j}^{g\to h}, \tilde{\ell}^{g\to h}, \breve{\ell}^{g\to h} \sim \text{Inv-Gamma}(2, 2), \tag{8i}$$

where the 2×8 recipient-specific time-varying 1D B-splines GPs each have squared exponential kernels with hyper-parameters $\tilde{\sigma}_r^{g\to h}$, $\tilde{\ell}^{g\to h}$, the 2 source-specific time-varying 1D B-splines GPs each have squared exponential kernels with hyper-parameters $\breve{\sigma}^{g\to h}$, $\breve{\ell}^{g\to h}$, and the 2 time-invariant 2D B-splines GPs each have squared exponential kernels with hyper-parameters $\sigma_{0,i}^{g\to h}$, $\ell_{0,i}^{g\to h}$ and $\ell_{0,j}^{g\to h}$ decomposed as follows,

$$k_0^{g\to h}\Big((i,j),(i',j')\Big) = (\sigma_0^{g\to h})^2 \exp\left(-\frac{(i-i')^2}{2(\ell_{0,i}^{g\to h})^2}\right) \exp\left(-\frac{(j-j')^2}{2(\ell_{0,j}^{g\to h})^2}\right). \tag{9}$$

We constrain the model further with a pseudo-likelihood term so that the model's implied incidence rate $\kappa_{r,j}^h$ in (3b) is around the MLE incidence rate estimate obtained from the incidence cohort. We took this approach in lieu of fitting the model to both the source-recipient and individual-level incidence exposure data to bypass extreme computational runtimes[12], and in the context that the source-recipient data are not informative of incidence dynamics[90]. Specifically, we fitted log-normal distributions to the $1,000 \times 50$ Monte Carlo replicate rate estimates for individuals of gender h and age j in round r (see above) using the `lognorm` R package version 0.1.6[91], and

then set

$$\frac{\sum_i \lambda^{g\to h}_{r,i,j}}{S^h_{r,j} \times \left|\left(t^{\text{end}}_r - t^{\text{start}}_r\right)\right|} \sim \text{LogNormal}\left(\text{mean} - \hat{\kappa}^h_{r,j}, \text{var} - \hat{\kappa}^h_{r,j}\right), \tag{10}$$

where mean-$\hat{\kappa}^h_{r,j}$ and var-$\hat{\kappa}^h_{r,j}$ denote respectively the parameters of the fitted log-normal distributions, and the left-hand side is calculated from (7b) and matches the model's incidence rate $\kappa^h_{r,j}$ in (3b).

Computational inference. Model (7-10) was fitted with `Rstan` version 2.21.0, using Stan's adaptive HMC sampler[63] with 4 chains for 3,500 iterations including 500 warm-up iterations. Convergence and mixing were good, with highest Rhat value of 1.0027 and lowest effective sample sizes of 1444. The model presented the data well, with 99.63% data point inside 95% posterior predictive intervals and the fitted model was consistent with all the available data (Extended Data Fig. 6), indicating that the data met the assumptions of the statistical model. There were no divergent transitions, suggesting non-pathological posterior topologies.

Counterfactual interventions

We investigated —given the inferred transmission flows— the hypothetical impact of targeted counterfactual intervention scenarios c on predicted incidence reductions in women in the most recent survey round 18. In the model, counterfactual interventions were implemented by calculating the expected number of transmission flows (2) into women under counterfactual c that fewer men of age i had remained with unsuppressed HIV in survey round 18, which we denote by $\tilde{I}^M_{\text{R18},i,c}$. We obtained the expected number of incident cases in women of age j in round 18 in counterfactual c via

$$\tilde{\lambda}^{M\to F}_{\text{R18},j,c} = \int \sum_i \hat{\beta}^{M\to F}_{\text{R18},i,j} \times \tilde{I}^M_{\text{R18},i,c} \times S^F_{\text{R18},j} \times \left|\left(t^{\text{end}}_{\text{R18}} - t^{\text{start}}_{\text{R18}}\right)\right| d\hat{\beta}^{M\to F}_{\text{R18},i,j}, \tag{11}$$

where uncertainty in the posterior age-specific transmission rates after fitting model (7-10) is integrated out. The predicted incidence rate reductions were based on comparing the counterfactuals (11) to the inferred cases in women in the corresponding age group (3b), $1 - \left(\sum_j \tilde{\lambda}^{M\to F}_{\text{R18},j,c}\right) / \left(\sum_j \hat{\lambda}^{M\to F}_{\text{R18},j}\right)$.

Closing half the gap in viral suppression rates in men relative to women. In this scenario, we considered the impact of reducing by half the gap in the proportion of men with unsuppressed HIV compared to the same proportion in women. To this end, we first calculated for each 1-year age band the average of the estimated proportion of census-eligible infected men in round 18 with suppressed virus and the same proportion in women, $\tilde{\nu}^M_{\text{R18},i} = (\nu^M_{\text{R18},i} + \nu^F_{\text{R18},i})/2$. Next, we set $\tilde{I}^M_{\text{R18},i,1}$ to the smoothed estimate of census-eligible men of age i in round R18 multiplied with the posterior median estimate of HIV prevalence in census-eligible men of age i, and with $1 - \tilde{\nu}^M_{\text{R18},i}$.

Closing the gap in viral suppression rates in men relative to women. In this scenario, we considered the impact of achieving the same proportions of men with unsuppressed HIV as in women. To this end, we set $\tilde{I}^{M}_{\text{R18},i,2}$ to the smoothed estimate of census-eligible men of age i in round R18 multiplied with the posterior median estimate of HIV prevalence in census-eligible men of age i, and with $1 - \nu^{F}_{\text{R18},i}$.

95-95-95 in men. In this scenario, we considered the impact of achieving viral suppression in 85.7% ($0.95 \times 0.95 \times 0.95$) in each 1-year age group of men with HIV. The number of remaining men with unsuppressed HIV in round 18, $\tilde{I}^{M}_{\text{R18},i,3}$, was calculated by multiplying the smoothed estimate of the census-eligible men of age i in round R18 with the posterior median estimate of HIV prevalence in the census-eligible men of age i, and with $1 - 0.857$.

Sensitivity analyses

Sensitivity in incidence rate estimates to the GAM incidence model specification. The longitudinal age-specific HIV incidence rates of the central analysis were estimated with a log-link generalized additive effects Poisson regression model with a linear predictor comprising relatively simple main and interaction effects by age and survey round, fitted to individual-level 0/1 incidence outcomes and exposure times specified as offset on the log scale. To assess sensitivity against the relatively simple linear predictor, we considered a more complex mean specification comprising independent LOESS smoothers to capture age-specific incidence trends in each survey round, and fitted this mean model for computational reasons to crude HIV incidence rates. Specifically, we fitted LOESS regressions as implemented in the `R` package `stats` version 3.6.2 with span argument set to 0.7 to the crude age-, gender- and round-specific HIV incidence rates in all 50 imputation data sets, and weighted by the corresponding, group-level aggregated exposure times. The HIV incidence rate estimates under the LOESS model had as expected a smaller mean absolute error against the crude estimates as compared against the GAM model (0.0048 [0.0046-0.0051] versus 0.0053 [0.0051-0.0056]) (Supplementary Fig. S3). Overall, the contribution of men to incidence was more variable across rounds while the shifts in the median age at infection were similar in the central and this sensitivity analysis (Supplementary Table S10).

Sensitivity in incidence rate and transmission flow estimates to limited communities. Over time some communities were added and others left the RCCS (see Supplementary Table S2). We repeated our analysis on the subset of 28 consecutively surveyed communities. We found similar incidence rates with slightly faster declines in male new infections and larger gender disparities (Supplementary Fig. S4). All other primary findings remained insensitive (Supplementary Table S10).

Sensitivity in estimating transmission flows to uncertainty in infection time estimates. In the central analysis, `phyloTSI` infection time estimates associated to source-recipient pairs were refined using the inferred transmission direction, age, and sero-history data. To assess sensitivity to the infection time estimates used, we inferred transmission flows on the basis of the raw `phyloTSI` infection time estimates as

long as they were compatible with the inferred transmission direction, and otherwise on the basis of the refined estimates. Overall, we found source-recipient pairs were potentially allocated to earlier or later time periods reflecting the wide uncertainty in infection time estimates, though across the sample the age distribution of sources and recipients was remarkably stable (Extended Data Fig. 4). All primary findings were insensitive to using the raw infection time estimates (Supplementary Table S10).

Sensitivity in time since infection estimates to higher transmissibility during acute infection. In the central analysis, transmission flows were estimated using the centre of gravity of the uncertainty region associated with the refined infection time estimates. To account for higher transmission rates during acute infection of the transmitting partner[88], we assumed that the transmission hazard was 5 times higher in the first 2 months after infection of the transmitting partner as compared to the following period, and obtained the resulting mean infection time estimate under this assumption by generalizing our Monte Carlo approach used in the central analysis to an importance sampling approach under piecewise linear transmission hazards. The primary results were insensitive to these changes as less than 5% of source-recipient pairs were attributed to different survey rounds (Supplementary Table S10).

Sensitivity in estimating transmission flows to right censoring of likely transmission pairs. The RCCS transmission cohort was defined retrospectively and so it is possible that some transmission events, especially in later rounds, remain as of yet unseen because the corresponding individuals are not yet in the survey or do not yet have virus deep-sequenced. To assess sensitivity to right censoring, we excluded from analysis those source-recipient pairs for which virus of the source or the recipient was deep-sequenced only after rounds 17, 16 and 15. The primary findings were insensitive to these analyses because the probabilities of detecting infection events in the phylogenetic data changed accordingly (Supplementary Table S10 and Supplementary Fig. S5).

Sensitivity in estimated transmission flows to limited sample size of likely transmission pairs. The number of observed infection events in the incidence cohort was $\approx$4 times larger than the number of reconstructed transmission events, prompting us to explore the effect of sampling uncertainty on the transmission flow estimates. We bootstrap sampled source-recipient pairs at random with replacement three times, and repeated inferences on these bootstrap samples. Our primary findings remained insensitive (Supplementary Table S10).

Sensitivity in estimated transmission flows to modelled sampling estimates. The sampling adjustments in (6) require assumptions including that sampling is independent of infection and transmission, independent between source and recipient, at random within strata, and well approximated by approximating sources with individuals with unsuppressed virus. We repeated flow inferences without any adjustments and without adjustments for potentially unequal sampling of sources. Our primary findings were insensitive across these analyses (Supplementary Table S10).

Sensitivity in transmission flow estimates to the phylo-SI model specification. In the central analysis, the log transmission rates that underpin the estimated transmission

flows were estimated using the linear predictor in (7c), and this model specification was associated with overall smallest mean absolute error and posterior predictive coverage as shown in Supplementary Table S6 against the following alternative models,

$$\log \beta^{g\to h}_{r,i,j} = \hat{\boldsymbol{c}}^{g\to h}(i,j) + \gamma_0 + \gamma_g + \gamma_r + \gamma_{p(r)} + \boldsymbol{f}^{g\to h}_0(i,j) + \boldsymbol{f}^{g\to h}_{p(r)}(i), \tag{12a}$$

$$\log \beta^{g\to h}_{r,i,j} = \hat{\boldsymbol{c}}^{g\to h}(i,j) + \gamma_0 + \gamma_g + \gamma_r + \gamma_{p(r)} + \boldsymbol{f}^{g\to h}_0(i,j) + \boldsymbol{f}^{g\to h}_{p(r)}(j), \tag{12b}$$

$$\log \beta^{g\to h}_{r,i,j} = \hat{\boldsymbol{c}}^{g\to h}(i,j) + \gamma_0 + \gamma_g + \gamma_r + \gamma_{p(r)} + \boldsymbol{f}^{g\to h}_0(i,j) + \boldsymbol{f}^{g\to h}_{p(r)}(i,j), \tag{12c}$$

$$\log \beta^{g\to h}_{r,i,j} = \hat{\boldsymbol{c}}^{g\to h}(i,j) + \gamma_0 + \gamma_g + \gamma_r + \gamma_{p(r)} + \boldsymbol{f}^{g\to h}_0(i,j) + \boldsymbol{f}^{g\to h}_r(j) \tag{12d}$$

$$\log \beta^{g\to h}_{r,i,j} = \hat{\boldsymbol{c}}^{g\to h}(i,j) + \gamma_0 + \gamma_g + \gamma_r + \gamma_{p(r)} + \boldsymbol{f}^{g\to h}_0(i,j) + \boldsymbol{f}^{g\to h}_r(j) + \boldsymbol{f}^{g\to h}_{p(r)}(j), \tag{12e}$$

$$\log \beta^{g\to h}_{r,i,j} = \hat{\boldsymbol{c}}^{g\to h}(i,j) + \gamma_0 + \gamma_g + \gamma_r + \gamma_{p(r)} + \boldsymbol{f}^{g\to h}_0(i,j) + \boldsymbol{f}^{g\to h}_r(j) + \boldsymbol{f}^{g\to h}_{p(r)}(i,j). \tag{12f}$$

Models specifying transmission rates without a round-specific random function on the age of infected individuals, (12a)-(12c), did not fit the data well (Supplementary Table S6). The remaining models, (12d)-(12f) performed as well as the model used in the central analysis (Supplementary Table S6) and our primary findings remained insensitive (Supplementary Table S10).

Sensitivity in counterfactual intervention impacts to assumptions on viral suppression levels in non-participants. Infection and viremia in the non-participant census-eligible population remained unknown and in the central analysis, we considered as proxy of virus suppression levels among non-participants data from first-time participants. We performed two sensitivity analyses, assuming first that all non-participants with HIV were also viremic across all rounds, and assuming second that virus suppression was identical among non-participants and participants of the same age, gender and survey round. Together, the two scenarios likely encompass the true, unknown viral suppression levels in non-participants. These scenarios were implemented by updating the number of individuals with viremia in (2), and refitting the model. The sensitivity analysis assuming all non-participants with HIV were viremic resulted in larger predicted incidence reductions in women around 75%, while the sensitivity analysis assuming virus suppression was the same among non-participants as among participants of the same age, gender and survey round resulted in similar predicted incidence reductions in women than in the central analysis (Supplementary Table S10).

Sensitivity in counterfactual intervention impacts to potentially higher HIV prevalence in non-participants. In the central analysis, we assumed that HIV prevalence was the same in participants and non-participants of the same age, gender and survey round. We considered three sensitivity analyses, assuming first that prevalence was 25% higher in male non-participants compared to male participants of the same age, gender and survey round, assuming second that prevalence was 25% higher in female non-participants compared to female participants of the same age, gender and survey round, and assuming third that prevalence was 25% higher in female and male non-participants compared to female and male participants of the same age, gender and survey round respectively. These scenarios were implemented by updating the number of virally unsuppressed individuals in (2), and refitting the model. Our primary findings remained insensitive (Supplementary Table S10).

Sensitivity in counterfactual intervention impacts to lower viral suppression thresholds. Different definitions of HIV suppression are currently operational, and we considered the effect of lower thresholds to define viral suppression (<200 copies/mL) than in the central analysis (<1,000 copies/mL). This scenario was implemented by re-estimating the age- and gender-specific proportions of individuals with HIV in the study population who had suppressed virus at the lower threshold, re-calculating gaps in viral suppression levels in men relative to women, and re-calculating the additional number of men needed to reach and maintain viral suppression in the counterfactual intervention scenarios. We found slightly smaller gender gaps in viral suppression at the lower threshold and the predicted incidence reduction in women in the counterfactual that assessed closing the suppression gap in men was around 45%, and all other findings remained insensitive (Supplementary Table S10).

Data Availability

Pseudo-anonymised data from the RCCS incidence and transmission cohort as well as pseudo-anonymised deep-sequence phylogenies to reproduce all analyses are available from Zenodo (https:/zenodo.org/record/8412741) as open-access data set under the CC-BY-4.0 license[92]. HIV consensus sequences have been deposited to Genbank under the accession number xxx.

Additional deep-sequence HIV-1 reads can be requested from PANGEA-HIV under a managed access policy due to privacy and ethical reasons, which aligns with UNAIDS ethical guidelines. The process for accessing data, the PANGEA-HIV Data Sharing Policy and a detailed description of what data are available is laid out in full at (https://www.pangea-hiv.org/join-us). Briefly, applicants can apply to receive additional data by submitting a concept sheet proposal in which they explain the research question and how they will mitigate potential risks to participant privacy. In line with requirements for PANGEA members, applicants will be asked to present proof of human subject research training and comply with PANGEA-internal publication agreements. PANGEA encourages external applicants to collaborate with the researchers who generated the data. For more information contact PANGEA project

manager Lucie Abeler-Dörner (lucie.abeler-dorner@bdi.ox.ac.uk). The time frame for a response to requests is 2-4 weeks.

Additional cohort data can be requested from RHSP. Because HIV transmission is criminalized in Uganda and due to further privacy considerations, RHSP maintains a controlled access data policy for corresponding epidemiological metadata and corresponding data collection tools. In brief, RHSP policy requires individuals to submit an RSHSP data request form (available upon request from info@rhsp.org or gkigozi@rhsp.org) and a brief concept note (1-2 pages) detailing their research questions and methods. In addition, researchers are asked to provide a curriculum vitae/resume along with proof of human subjects research training. Concept sheets can be submitted to Dr. Godfrey Kigozi (gkigozi@rhsp.org), executive director of the RHSP. Only individuals named on the original data request and who provide the request, CV/resume and HSR training, are permitted access to the data. Released data are not to be reused for other purposes outside of approved concepts. The time frame for a response to requests is 2-4 weeks.

Code availability

Code to reproduce all analyses is freely available on GitHub version 1.1.2 under the GNU General Public License version 3.0 at the repository (https://github.com/MLGlobalHealth/phyloSI-RakaiAgeGender).

Acknowledgements

We thank all contributors, program staff and participants to the Rakai Community Cohort Study; all members of the PANGEA-HIV consortium, the Rakai Health Sciences Program, and CDC Uganda for comments on an earlier version of the manuscript; the Imperial College Research Computing Service (https://doi.org/10.14469/hpc/2232) and the Biomedical Research Computing Cluster at the University of Oxford for providing the computational resources to perform this study; the Office of Cyberinfrastructure and Computational Biology at the National Institute for Allergy and Infectious Diseases for data management support; and Zulip for sponsoring team communications through the Zulip Cloud Standard chat app. This study was supported by the Bill & Melinda Gates Foundation (OPP1175094 to CF, OPP1084362 to Prof Deenan Pillay); the National Institute of Allergy and Infectious Diseases (U01AI051171 to RHG, U01AI075115 to RHG, UM1AI069530-16 to Dr Mary Glenn Fowler, Dr Philippa Musoke, and Dr Aaron Tobian, R01AI087409 to RHG, U01AI100031 to RHG, R01AI110324 to RHG, R01AI114438 to MJW, K25AI114461 to Dr Xiangrong Kong, R01AI123002 to Dr Cindy Liu, K01AI125086 to MKG, R01AI128779 to Dr Aaron Tobian, R01AI143333 to LWC, R21AI145682 to Dr Caitlin Kennedy, R01AI155080 to MKG, ZIAAI001040 to TCQ); the National Institute of Mental Health (F31MH095649 to Dr Jennifer Wagman, R01MH099733 to Ned Sacktor and MJW, R01MH107275 to LWC, R01MH115799 to MJW and LWC, U19MH110001 to Dr Mary McKay and Dr Fred Ssewamala); the National Institute of Child Health and Development (R01HD038883 to RHG, R01HD050180 to MJW, R01HD070769 to MJW, R01HD091003 to JS); the Division of Intramural Research of the National Institute for Allergy and Infectious Diseases (K01AA024068 to Dr Jennifer Wagman); the National Heart, Lung, and Blood Institute (R01HL152813 to LWC); the Fogarty International Center (D43TW009578 to RHG, D43TW010557 to LWC); the Doris Duke Charitable Foundation to Dr Aaron Tobian; the Johns Hopkins University Center for AIDS Research (P30AI094189 to Dr Richard Chaisson); the U.S. President's Emergency Plan for AIDS Relief (PEPFAR) through the Centers for Disease Control and Prevention (NU2GGH000817 to RHSP); the Engineering and Physical Sciences Research Council through the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning at Imperial and Oxford (EP/S023151/1 to Prof Axel Gandy); and the Imperial College London President's PhD Scholarship fund to YC. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. The findings and conclusions in this report are those of the author(s) and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Author contributions

OR and MKG designed the study. OR, MKG, JK, PGF, DS oversaw the study. SA, RG, RS, EK, VS, LAD, DB, LWC, CF, TG, GJG, JJ, GK, LM, LN, OL, TQ, SJR, JS, LT, MJW, DS, JK, MKG oversaw and performed data collection. MM, ABr, XX, EK, VS, AA, ABl, YC, SD, TG, MH, SS, LT, MKG, OR contributed to the analysis. MM,

ABr, XX, ABl, YC, MKG, OR wrote the first draft. None supplied materials. All authors discussed the results and contributed to the revision of the final manuscript.

Competing interests

OR, MKG, CF, DB report grants from the Bill & Melinda Gates Foundation during the conduct of this study. MKG, MJW, RHG, LCW report grants from the National Institutes of Health during the conduct of this study. ABr, YC, XX report an EPSRC PhD studentship during the conduct of this study. Dr. Wawer and Dr. Gray are paid consultants to the Rakai Health Sciences Program and serves on its Board of Directors. This arrangement has been reviewed and approved by the Johns Hopkins University in accordance with its conflict of interest policies.

Tables

Age	HIV prevalence in men	Men with HIV who have unsuppressed virus	Male-female difference in proportion of individuals with HIV who have unsuppressed virus	Contribution of age group to all men with unsuppressed virus	Contribution of age group to all transmitting male partners	Closing half the suppression gap: Contribution of age group to additional number of men with unsuppressed virus in counterfactual	Closing half the suppression gap: Predicted reduction in incidence in women in round 18	Closing the suppression gap: Contribution of age group to additional number of men with unsuppressed virus in counterfactual	Closing the suppression gap: Predicted reduction in incidence in women in round 18	95-95-95 in men: Contribution of age group to additional number of men with unsuppressed virus in counterfactual	95-95-95 in men: Predicted reduction in incidence in women in round 18
	(% in age bracket)	(% in age bracket)	(difference)	(%)	(%)	(%)	(% of actual incidence)	(%)	(% of actual incidence)	(%)	(% of actual incidence)
15-19	0.8 [0.4-1.3]	73.2 [56.6-86.2]	36.8 [16.8-54.2]	5.3 [3.0-8.8]	1.4 [0.4-3.5]	5.3 [2.5-8.7]	21.9 [21.2-22.8]	5.3 [2.5-8.7]	43.7 [42.4-45.6]	7.4 [5.2-9.9]	72.2 [68.8-74.6]
20-24	2.1 [1.6-2.7]	66.5 [55.5-76.6]	26.6 [14.0-38.2]	9.0 [6.5-11.8]	12.7 [7.5-19.2]	7.1 [3.7-10.9]	24.7 [23.2-26.2]	7.1 [3.7-10.9]	49.6 [46.6-52.7]	12.2 [9.4-15.6]	60.4 [55.6-65.0]
25-29	6.2 [5.2-7.3]	53.5 [45.3-61.4]	23.2 [13.7-32.6]	17.9 [14.6-21.5]	20.2 [13.8-27.5]	15.0 [9.1-21.5]	25.1 [23.9-26.6]	15.0 [9.1-21.5]	50.5 [47.9-53.5]	22.7 [18.2-28.1]	58.4 [54.0-62.4]
30-34	12.5 [10.8-14.2]	39.6 [33.6-45.8]	19.2 [12.1-26.1]	24.3 [20.8-28.0]	19.4 [13.3-27.3]	21.3 [13.7-27.7]	25.2 [23.6-27.0]	21.3 [13.7-27.7]	50.7 [47.4-54.3]	26.9 [22.2-31.5]	58.7 [53.0-64.0]
35-39	16.4 [14.6-18.2]	28.9 [23.3-35.0]	17.6 [11.2-24.3]	21.3 [17.9-24.8]	25.8 [18.4-34.9]	24.0 [16.9-31.1]	26.9 [24.8-28.7]	24.0 [16.9-31.1]	54.3 [49.8-58.1]	18.6 [13.5-23.6]	52.9 [46.6-60.0]
40-44	16.8 [14.8-18.9]	21.9 [16.3-28.2]	14.6 [8.5-21.5]	13.8 [10.6-17.1]	14.9 [9.2-21.8]	17.7 [10.9-23.9]	28.3 [25.8-30.0]	17.7 [10.9-23.9]	57.7 [52.4-61.1]	8.3 [2.6-13.4]	42.5 [35.2-51.9]
45-49	16.4 [14.1-19.1]	19.4 [13.0-27.0]	12.2 [4.5-20.5]	8.2 [5.5-11.3]	4.4 [1.7-8.5]	9.6 [2.7-15.8]	27.1 [24.6-29.0]	9.6 [2.7-15.8]	56.1 [50.1-59.7]	3.7 [0.0-8.5]	40.3 [29.4-55.1]
Total	8.0 [7.4-8.6]	33.9 [29.7-38.3]	14.8 [10.0-19.6]	100	100	100	25.1 [24.2-26.2]	100	50.6 [48.6-52.8]	100	58.4 [54.9-61.7]

Table 1: **HIV prevalence, viral suppression, transmission sources, and impact of counterfactual interventions focussed on closing the suppression gap in men by age of male partner, round 18, Oct 2016-Apr 2018.**

Figures

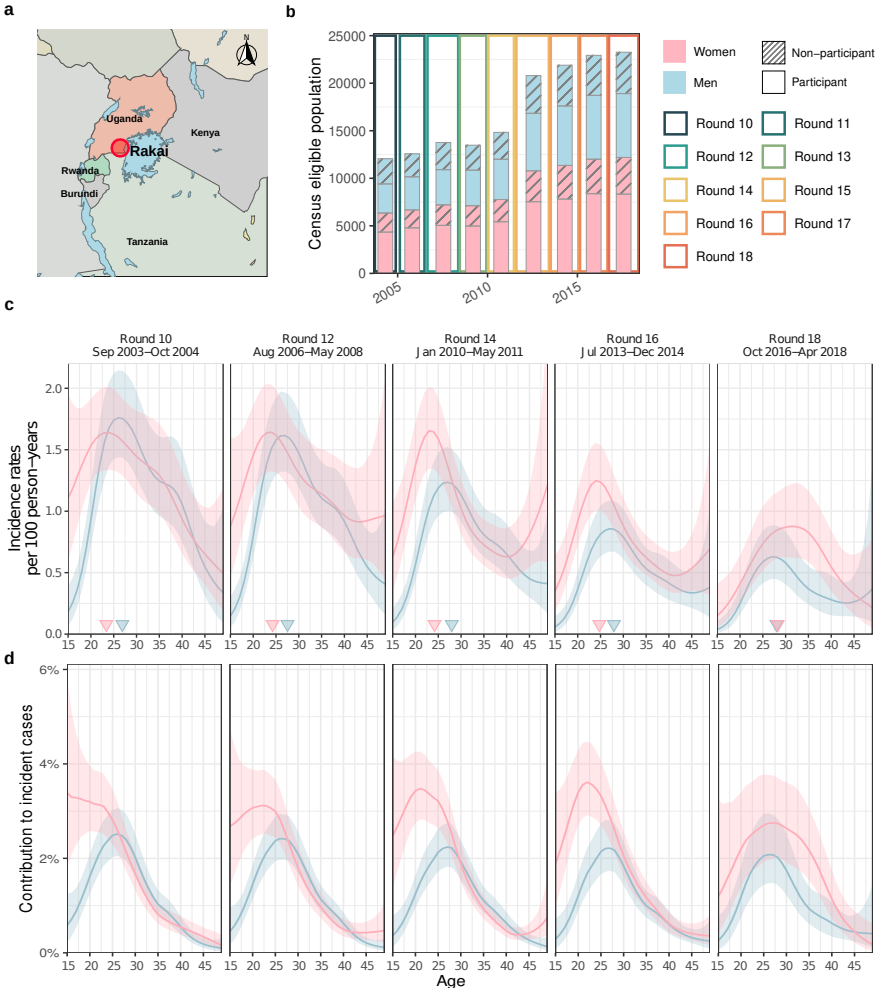


Fig. 1: **Time trends in age-specific HIV incidence rates for men and women in Rakai, Uganda.** (**a**) Location of the Rakai Community Cohort Study (RCCS) in south-central Uganda. Study outcomes are reported for all RCCS communities located inland to Lake Victoria across nine survey rounds. Sources of the map data: OpenStreetMap open data by OpenStreetMap contributors, see https://www.openstreetmap.org/copyright. (**b**) Number of RCCS participants in the census-eligible population of age 15 to 49 by survey round. (**c**) Estimated mean HIV incidence rates per 100 person-years of exposure in uninfected individuals (line) by 1-year age band, gender and survey round, along with 95% confidence intervals (ribbon), and median age of incident cases (cross). (**d**) Estimated median contribution to incidence cases in the study population (line) by 1-year age band, gender and survey round, along with 95% confidence intervals (ribbon). Throughout all subfigures, incidence estimates are based on n =1,117 individuals in the incidence cohort.

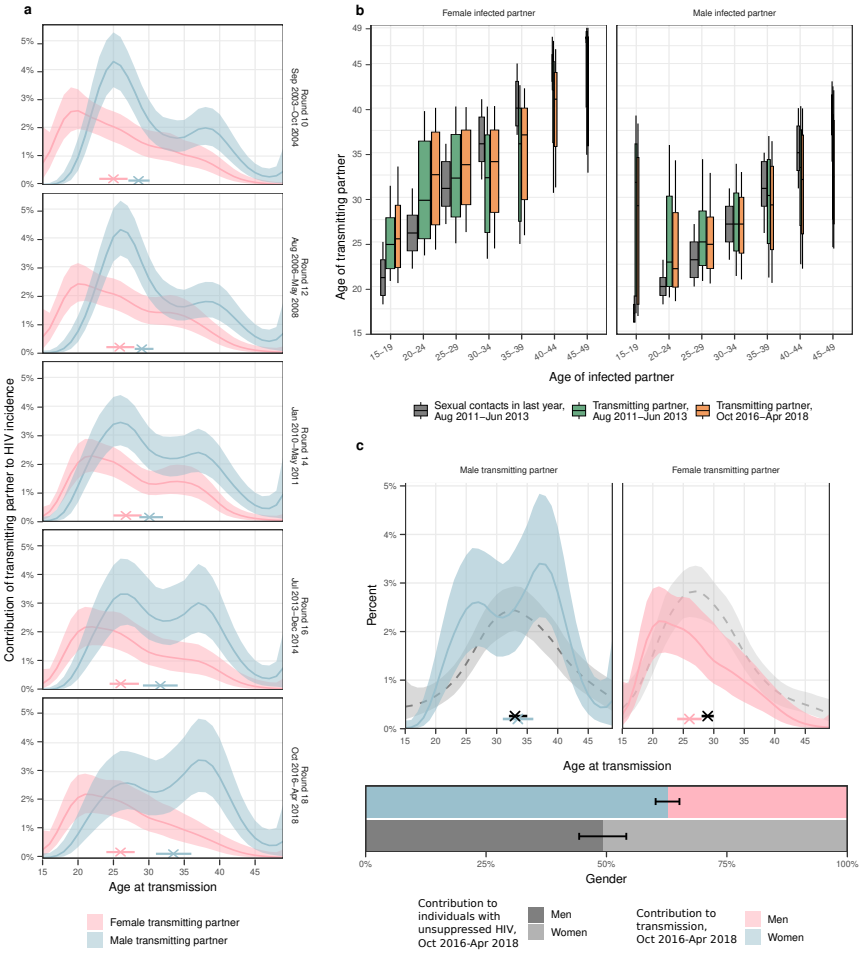


Fig. 2: **Time trends in age-specific sources of HIV infections in women and men.** (**a**) Estimated age distributions of transmitting partners (posterior median: line, 95% credible interval: ribbon), along with the median age at transmission (posterior median: cross, 95% credible interval: linebar). Age contributions sum to 100% for each round, summing over men and women. (**b**) Estimated age distributions of transmitting partners by 5-year age bracket of infected partners (posterior median: thick black bar in boxplots, 50% interquartile range: height of box, 80% credible intervals: whiskers in boxplots). The width of the boxplots is proportional to the total infections in each recipient group. For reference, posterior estimates of the age distributions of sexual contact partners of men and women by 5-year age bands in the past 12 months in the same communities are shown in dark grey (estimates visualized in the same manner). (**c**) Comparison of the age contributions to transmitting partners (color) to the age contributions to men and women with unsuppressed HIV (posterior median: dashed black line, 95% credible interval: ribbon), along with median age (posterior median: cross, 95% credible interval: linebar). Age contributions sum to 100% for men and women combined. Throughout all subfigures, transmission flow estimates are based on $n = 227$ heterosexual source-recipient pairs identified among $n =$1,978 individuals in the transmission cohort and $n =$1,117 individuals in the incidence cohort.

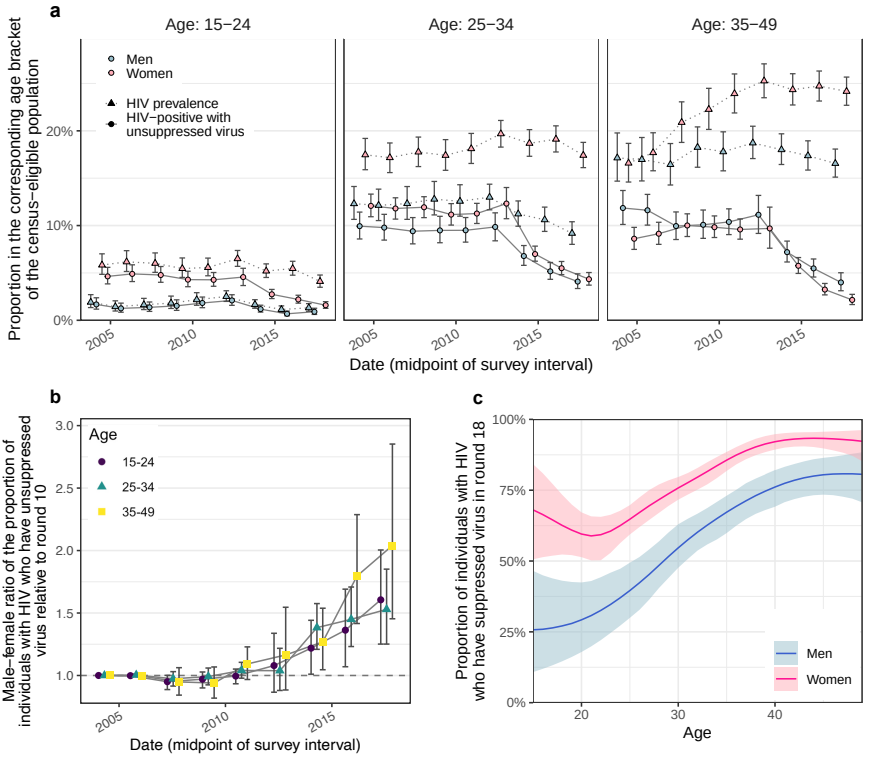


Fig. 3: **Changes in population-level suppression of HIV viral load.** (**a**) Estimated trends in HIV prevalence and the proportion of census-eligible individuals in three age brackets that remain virally unsuppressed, defined as viral load above 1,000 copies/mL blood (posterior median: dots, 95% credible interval: errorbars), combining data from participants and from first-time participants as proxy of non-participants. (**b**) Male-to-female ratio in changes in population-level viral load suppression relative to round 10 (posterior median: dots, 95% credible interval: errorbars). (**c**) Estimated viral suppression rates by 1-year age band (x-axis) and gender (color) for survey round 18 (posterior median: dots, 95% credible interval: errorbars). Throughout all subfigures, estimates are based on data from $n = 38,749$ participants including $n = 3,265$ participants with HIV and with measured viral load. First-time participants were used as proxies of individuals who did not participate in the survey.

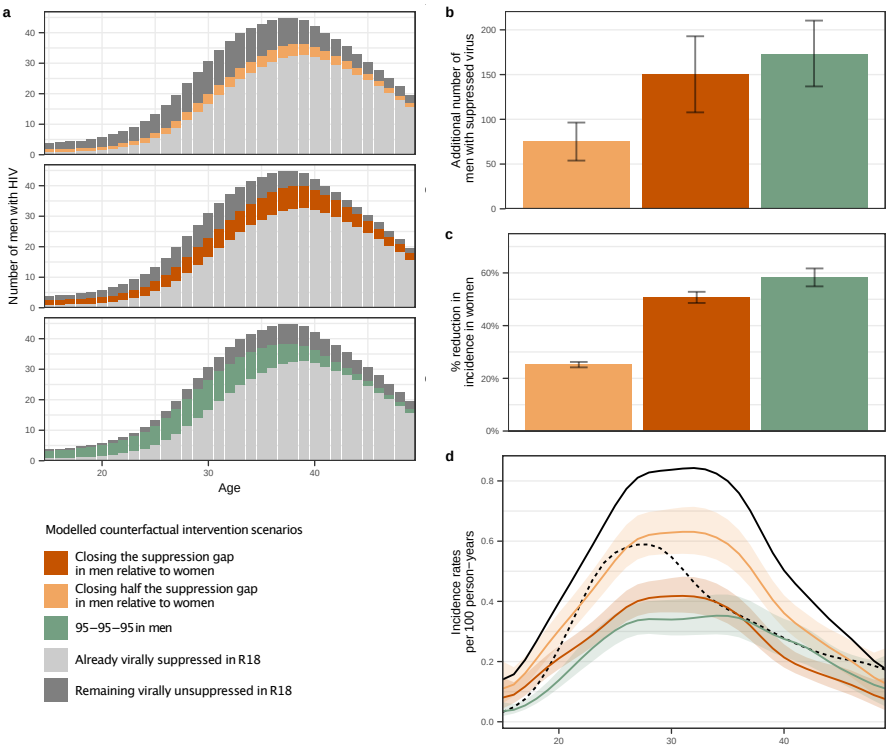


Fig. 4: **Counterfactual modeling scenarios predicting the impact of interventions to increase HIV suppression in men on incidence reductions in women.** (**a**-**b**) Estimated additional number of men with HIV in the census-eligible population in round 18 that already had suppressed virus (light grey), those who would have achieved viral suppression in the counterfactual intervention scenarios (color), and those who would have remained with unsuppressed virus in the counterfactuals (dark grey). Posterior median: bars, 95% credible interval: errorbars. (**c**) Percent reduction in incidence in women of the census-eligible population in round 18 under the counterfactual targeted scenarios. Posterior median: bars, 95% credible interval: errorbars. (**d**) Estimated incidence rates among women in the census-eligible population in round 18 (black solid line) and the counterfactual scenarios (color), with incidence rates among men in round 18 shown as reference (black dashed line). Posterior median: lines, 95% credible intervals: ribbons. Throughout all subfigures, estimates are based on data from $n = 15,053$ participants in survey round 18, including $n = 110$ seroconverts in the incidence cohort in round 18, $n = 432$ individuals with HIV and with measured viral load in round 18, and $n = 61$ heterosexual source-recipient pairs in rounds 16-18, and information inferred through hierarchical models from all individuals in earlier rounds.

References

[1] UNAIDS. 2021 UNAIDS global AIDS update — confronting inequalities — lessons for pandemic responses from 40 years of AIDS (2021). Available at https://www.unaids.org/sites/default/files/media_asset/2021-global-aids-update_en.pdf.

[2] UNAIDS. Women and HIV — a spotlight on adolescent girls and young women (2019). Available at https://www.unaids.org/sites/default/files/media_asset/2019_women-and-hiv_en.pdf.

[3] Goga, A. *et al.* Centring adolescent girls and young women in the HIV and COVID-19 responses. *Lancet* **396**, 1864–1866 (2020) .

[4] UNAIDS. Dangerous inequalities: world AIDS day report 2022 (2022). Available at https://www.unaids.org/sites/default/files/media_asset/dangerous-inequalities_en.pdf.

[5] UNAIDS. Full report — in danger: UNAIDS global AIDS update 2022 (2022). Available at https://www.unaids.org/sites/default/files/media_asset/2022-global-aids-update_en.pdf.

[6] Gupta, R. K. *et al.* HIV-1 drug resistance before initiation or re-initiation of first-line antiretroviral therapy in low-income and middle-income countries: a systematic review and meta-regression analysis. *Lancet Infectious Diseases* **18**, 346–355 (2018) .

[7] World Health Organization. HIV drug resistance report 2021 (2021). Available at https://www.who.int/publications-detail-redirect/9789240038608.

[8] UNAIDS. Resources and financing for the AIDS response (2022). Available at https://www.unaids.org/en/topic/resources.

[9] UNAIDS. Agenda item 2: Report of the 48th PCB meeting (2021). Available at https://www.unaids.org/sites/default/files/media_asset/PCBSS_Report_48th_PCB_EN_rev3.pdf.

[10] United States Department of State. Deams partnership (2022). Available at https://www.state.gov/wp-content/uploads/2022/09/PEPFAR-Strategic-Direction_FINAL.pdf.

[11] United States Department of State. Reimagining PEPFAR at 20 to end the HIV/AIDS pandemic by 2030 (2022). Available at https://www.state.gov/reimagining-pepfar-at-20-to-end-the-hiv-aids-pandemic-by-2030/.

[12] Risher, K. A. *et al.* Age patterns of HIV incidence in eastern and southern Africa: a modelling analysis of observational population-based cohort studies. *Lancet HIV* **8**, e429–e439 (2021) .

[13] Akullian, A. *et al.* Large age shifts in HIV-1 incidence patterns in KwaZulu-Natal, South Africa. *Proceedings of the National Academy of Sciences of the United States of America* **118**, e2013164118 (2021) .

[14] Grabowski, M. K. *et al.* HIV prevention efforts and incidence of HIV in Uganda. *New England Journal of Medicine* **377**, 2154–2166 (2017) .

[15] Vandormael, A., Akullian, A., Siedner, M., de Oliveira, T., Bärnighausen, T. & Tanser, F. Declines in HIV incidence among men and women in a South African population-based cohort. *Nature communications* **10**, 1–10 (2019) .

[16] Grabowski, M. K. *et al.* The role of viral introductions in sustaining community-based HIV epidemics in rural Uganda: evidence from spatial clustering, phylogenetics, and egocentric transmission models. *PLoS medicine* **11**, e1001610 (2014) .

[17] Joshi, K. *et al.* Declining HIV incidence in sub-Saharan Africa: a systematic review and meta-analysis of empiric data. *Journal of the International AIDS Society* **24**, e25818 (2021) .

[18] Chang, L. W. *et al.* Heterogeneity of the HIV epidemic in agrarian, trading, and fishing communities in Rakai, Uganda: an observational epidemiological study. *Lancet HIV* **3**, e388–e396 (2016) .

[19] Abeler-Dörner, L., Grabowski, M. K., Rambaut, A., Pillay, D. & Fraser, C. PANGEA-HIV 2: phylogenetics and networks for generalised epidemics in africa. *Current Opinion in HIV and AIDS* **14**, 173 (2019) .

[20] Hall, M. *et al.* Demographic characteristics of sources of HIV-1 transmission in the era of test and treat. *medRxiv* (2022). Preprint at https://www.medrxiv.org/content/early/2022/10/13/2021.10.04.21263560.full.pdf .

[21] Vandormael, A., Akullian, A., Siedner, M., de Oliveira, T., Bärnighausen, T. & Tanser, F. Declines in HIV incidence among men and women in a South African population-based cohort. *Nature Communications* **10**, 5482 (2019) .

[22] Ratmann, O. *et al.* Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis. *Nature Communications* **10**, 1411 (2019) .

[23] Fisher, M. *et al.* Determinants of HIV-1 transmission in men who have sex with men: a combined clinical, epidemiological and phylogenetic approach. *Aids* **24**, 1739–1747 (2010) .

[24] Ratmann, O. *et al.* Sources of HIV infection among men having sex with men and implications for prevention. *Science translational medicine* **8**, 320ra2 (2016) .

[25] Poon, A. F. *et al.* Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. *Lancet HIV* **3**, e231–e238 (2016) .

[26] Wymant, C. *et al.* PHYLOSCANNER: inferring transmission from within-and between-host pathogen genetic diversity. *Molecular Biology and Evolution* **35**, 719–733 (2018) .

[27] Golubchik, T. *et al.* HIV-phyloTSI: Subtype-independent estimation of time since HIV-1 infection for cross-sectional measures of population incidence using deep sequence data. *medRxiv* (2022). Preprint at https://www.medrxiv.org/content/early/2022/05/16/2022.05.15.22275117.full.pdf .

[28] Ratmann, O. *et al.* Quantifying HIV transmission flow between high-prevalence hotspots and surrounding communities: a population-based study in Rakai, Uganda. *Lancet HIV* **7**, e173–e183 (2020) .

[29] Xi, X. *et al.* Inferring the sources of HIV infection in Africa from deep-sequence data with semi-parametric Bayesian Poisson flow models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **71**, 517–540 (2022) .

[30] Wilson, D. & Halperin, D. T. "Know your epidemic, know your response": a useful approach, if we get it right. *Lancet* **372**, 423–426 (2008) .

[31] Oliveira, T. d. *et al.* Transmission networks and risk of HIV infection in KwaZulu-Natal, South Africa: a community-wide phylogenetic study. *Lancet HIV* **4**, e41–e50 (2017) .

[32] Evans, M. *et al.* Age-disparate sex and HIV risk for young women from 2002 to 2012 in South Africa. *Journal of the International AIDS Society* **19**, 21310 (2016) .

[33] Akullian, A., Bershteyn, A., Klein, D., Vandormael, A., Bärnighausen, T. & Tanser, F. Sexual partnership age pairings and risk of HIV acquisition in rural South Africa. *AIDS (London, England)* **31**, 1755–1764 (2017) .

[34] Kyle, I. Population level HIV viral load varies by gender, age, and location in Rakai, Uganda (2020). Available at https://www.croiconference.org/wp-content/uploads/sites/2/posters/2020/1430_3_Quinn_00865.pdf.

[35] World Health Organization. *Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach* (World Health Organization, 2016).

[36] Tanser, F. *et al.* Effect of population viral load on prospective HIV incidence in a hyperendemic rural African community. *Science Translational Medicine* **9**, eaam8012 (2017) .

[37] Donnelly, C. *et al.* Gender difference in HIV-1 RNA viral loads. *HIV medicine* **6**, 170–178 (2005) .

[38] Reniers, G., Armbruster, B. & Lucas, A. Sexual networks, partnership mixing, and the female-to-male ratio of HIV infections in generalized epidemics: An agent-based simulation study. *Demographic research* **33**, 425–450 (2015) .

[39] Quinn, T. C. & Overbaugh, J. HIV/AIDS in women: an expanding epidemic. *Science* **308**, 1582–1583 (2005) .

[40] Glynn, J. R. *et al.* Why do young women have a much higher prevalence of HIV than young men? A study in Kisumu, Kenya and Ndola, Zambia. *Aids* **15**, S51–S60 (2001) .

[41] Loevinsohn, G. *et al.* Effectiveness of voluntary medical male circumcision for human immunodeficiency virus prevention in Rakai, Uganda. *Clinical Infectious Diseases* **73**, e1946–e1953 (2021) .

[42] Rodger, A. J. *et al.* Risk of HIV transmission through condomless sex in serodifferent gay couples with the HIV-positive partner taking suppressive antiretroviral therapy (PARTNER): final results of a multicentre, prospective, observational study. *Lancet* **393**, 2428–2438 (2019) .

[43] Cohen, M. S. *et al.* Antiretroviral therapy for the prevention of HIV-1 transmission. *New England Journal of Medicine* **375**, 830–839 (2016) .

[44] UNAIDS. Global AIDS strategy 2021-2026 - end inequalities end AIDS. (2021). Available at https://www.unaids.org/sites/default/files/media_asset/global-AIDS-strategy-2021-2026_en.pdf.

[45] Stover, J. *et al.* Modeling the epidemiological impact of the UNAIDS 2025 targets to end AIDS as a public health threat by 2030. *PLOS Medicine* **18**, e1003831 (2021) .

[46] Ssempijja, V. *et al.* High rates of pre-exposure prophylaxis eligibility and associated HIV incidence in a population with a generalized HIV epidemic in Rakai, Uganda. *Journal of acquired immune deficiency syndromes (1999)* **90**, 291–299 (2022) .

[47] Godfrey-Faussett, P., Frescura, L., Karim, Q. A., Clayton, M., Ghys, P. D. & On Behalf of the 2025 Prevention Targets Working. HIV prevention for the next decade: Appropriate, person-centred, prioritised, effective, combination prevention. *PLOS Medicine* **19**, e1004102 (2022) .

[48] The Lancet HIV. Addressing inequalities still key to ending HIV/AIDS. *Lancet HIV* **10**, e1 (2023) .

[49] Havlir, D. *et al.* What do the universal test and treat trials tell us about the path to HIV epidemic control? *Journal of the International AIDS Society* **23**, e25455 (2020) .

[50] The PHIA Project. PHIA data manager (2022). Available at https://phia-data.icap.columbia.edu/.

[51] Magosi, L. E. *et al.* Deep-sequence phylogenetics to quantify patterns of HIV transmission in the context of a universal testing and treatment trial - BCPP/Ya Tsie trial. *eLife* **11**, e72657 (2022) .

[52] The RISE consortium. Engaging men in HIV testing, linkage, and retention in care (2020). Available at https://www.jhpiego.org/wp-content/uploads/2021/01/8-RISE-Engaging-Men-Brief-1.pdf.

[53] Colvin, C. J. Strategies for engaging men in HIV services. *Lancet HIV* **6**, e191–e200 (2019) .

[54] Sithole, N. *et al.* Implementation of HIV self-testing to reach men in rural uMkhanyakude, KwaZulu-Natal, South Africa. A DO-ART trial sub study. *Frontiers in Public Health* **9**, 652887 (2021) .

[55] United States Department of State. The United States president's emergency plan for AIDS relief (2022). Available at https://www.state.gov/pepfar/.

[56] Farquhar, C., Masyuko, S. & Mugo, P. Social network–based strategies to improve uptake of HIV testing and linkage to care among men who have sex with men in Sub-Saharan Africa. *JAMA Network Open* **5**, e220155 (2022) .

[57] De Cock, K. M., Barker, J. L., Baggaley, R. & El Sadr, W. M. Where are the positives? HIV testing in sub-Saharan Africa in the era of test and treat. *AIDS* **33**, 349 (2019) .

[58] Mugavero, M. J. *et al.* Beyond core indicators of retention in HIV care: missed clinic visits are independently associated with all-cause mortality. *Clinical Infectious Diseases* **59**, 1471–1479 (2014) .

[59] Mukumbang, F. C. Leaving no man behind: how differentiated service delivery models increase men's engagement in HIV care. *International Journal of Health Policy and Management* **10**, 129–140 (2021) .

[60] Kripke, K., Eakle, R., Cheng, A., Rana, S., Torjesen, K. & Stover, J. The case for prevention – Primary HIV prevention in the era of universal test and treat: A mathematical modeling study. *eClinicalMedicine* **46**, 101347 (2022) .

[61] Rosenberg, N. E. *et al.* Adult HIV-1 incidence across 15 high-burden countries in sub-saharan africa from 2015 to 2019: a pooled analysis of nationally

representative data. *The Lancet HIV* **10**, e175–e185 (2023) .

[62] Galiwango, R. M. *et al.* Evaluation of current rapid HIV test algorithms in Rakai, Uganda. *Journal of Virological Methods* **192**, 25–27 (2013) .

[63] Carpenter, B. *et al.* Stan: A probabilistic programming language. *Journal of Statistical Software* **76**, 1–32 (2017) .

[64] Kong, X. *et al.* Association of medical male circumcision and antiretroviral therapy scale-up with community HIV incidence in Rakai, Uganda. *JAMA* **316**, 182–190 (2016) .

[65] Ssempijja, V. *et al.* Results of early virologic monitoring may facilitate differentiated care monitoring strategies for clients on ART, Rakai, Uganda. *Open Forum Infectious Diseases* **5**, ofy212 (2018) .

[66] Ssempijja, V. *et al.* Adaptive viral load monitoring frequency to facilitate differentiated care: A modeling study from Rakai, Uganda. *Clinical Infectious Diseases* **71**, 1017–1021 (2019) .

[67] Grabowski, M. K. *et al.* Prevalence and predictors of persistent human immunodeficiency virus viremia and viral rebound after universal test and treat: A population-based study. *Journal of Infectious Diseases* **223**, 1150–1160 (2021) .

[68] Dan, S. *et al.* Estimating fine age structure and time trends in human contact patterns from coarse contact data: the bayesian rate consistency model. *arXiv* (2022). Preprint at https://arxiv.org/abs/2210.11358 .

[69] Lewontin, R. C. Sex, lies, and social science. *New York Review of Books* **42**, 24–29 (1995) .

[70] Weinhardt, L. S., Forsyth, A. D., Carey, M. P., Jaworski, B. C. & Durant, L. E. Reliability and validity of self-report measures of HIV-related sexual behavior: progress since 1990 and recommendations for research and practice. *Archives of sexual behavior* **27**, 155–180 (1998) .

[71] Gregson, S., Zhuwau, T., Ndlovu, J. & Nyamukapa, C. A. Methods to reduce social desirability bias in sex surveys in low-development settings: experience in Zimbabwe. *Sexually transmitted diseases* **29**, 568–575 (2002) .

[72] Kelly, C. A. *et al.* Using biomarkers to assess the validity of sexual behavior reporting across interview modes among young women in Kampala, Uganda. *Studies in family planning* **45**, 43–58 (2014) .

[73] Gabry, J. & Češnovar, R. cmdstanr: R interface to 'CmdStan' (2020). Available at https://mc-stan.org/users/interfaces/cmdstan.

[74] Hastie, T. Generalized additive models R package (2020). Available at https://cran.r-project.org/web/packages/gam/gam.pdf.

[75] Gall, A. *et al.* Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. *Journal of Clinical Microbiology* **50**, 3838–3844 (2012) .

[76] Bonsall, D. *et al.* ve-SEQ: Robust, unbiased enrichment for streamlined detection and whole-genome sequencing of HCV and other highly diverse pathogens. *F1000Research* **4**, 1062 (2015) .

[77] Bonsall, D. *et al.* A comprehensive genomics solution for hiv surveillance and clinical monitoring in low-income settings. *Journal of Clinical Microbiology* **58**, e00382–20 (2020) .

[78] Jenkins, F. *et al.* Validation of a hiv whole genome sequencing method for hiv drug resistance testing in an australian clinical microbiology laboratory. *medRxiv* (2023). https://arxiv.org/abs/https://www.medrxiv.org/content/early/2023/07/06/2023.07.05.23292232.full.pdf .

[79] Wymant, C. *et al.* Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with shiver. *Virus Evolution* **4**, vey007 (2018) .

[80] Ratmann, O. *et al.* HIV-1 full-genome phylogenetics of generalized epidemics in Sub-Saharan Africa: Impact of missing nucleotide characters in next-generation sequences. *AIDS Research and Human Retroviruses* **33**, 1083–1098 (2017) .

[81] Zhang, Y. *et al.* Evaluation of phylogenetic methods for inferring the direction of human immunodeficiency virus (HIV) transmission: HIV prevention trials network (HPTN) 052. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* **72**, 30–37 (2020) .

[82] Xi, X. *Bayesian methods for source attribution using HIV deep sequence data*. Ph.D. thesis, Imperial College London (2022).

[83] Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research* **30**, 3059–3066 (2002) .

[84] Nguyen, L., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**, 268–274 (2014) .

[85] Poon, A. F. *et al.* Dates of HIV infection can be estimated for seroprevalent patients by coalescent analysis of serial next-generation sequencing data. *AIDS* **25**, 2019–2026 (2011) .

[86] Cohen, M. S., Gay, C. L., Busch, M. P. & Hecht, F. M. The detection of acute HIV infection. *Journal of Infectious Diseases* **202**, S270–S277 (2010) .

[87] Pantazis, N. *et al.* Discriminating between premigration and postmigration HIV acquisition using surveillance data. *Journal of Acquired Immune Deficiency Syndromes* **88**, 117–124 (2021) .

[88] Bellan, S. E., Dushoff, J., Galvani, A. P. & Meyers, L. A. Reassessment of HIV-1 acute phase infectivity: Accounting for heterogeneity and study design with simulated cohorts. *PLOS Medicine* **12**, e1001801 (2015) .

[89] Monod, M. *et al.* Regularised b-splines projected gaussian process priors to estimate time-trends in age-specific COVID-19 deaths. *Bayesian Analysis* **18** (2023) .

[90] Jacob, P. E., Murray, L. M., Holmes, C. C. & Robert, C. P. Better together? statistical learning in models made of modules. *arXiv* (2017). Preprint at https://arxiv.org/abs/1708.08719 .

[91] Wutzler, T. lognorm: Functions for the lognormal distribution. *R package version 0.1* **6** (2019) .

[92] Monod, M. *et al.* Phylogenetic and epidemiologic data relating to age-specific HIV incidence and transmission in Rakai, Uganda, 2003-2018. (2023). Data at https://doi.org/10.5281/zenodo.8412741.

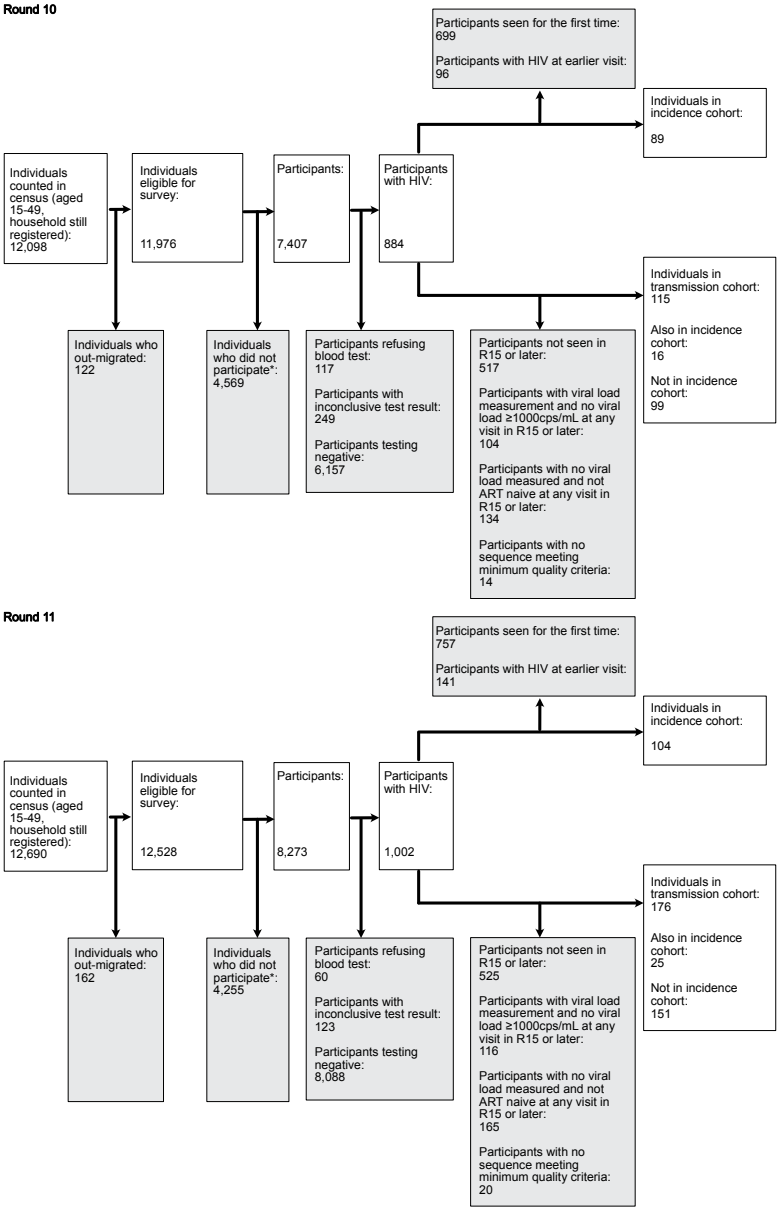
Rakai Health Sciences Program consortium authors

Larry W Chang[3,2,11], Ronald M Galiwango[2], M Kate Grabowski[2,14,11], Ronald H Gray[13], Jade C Jackson[14], Joseph Kagaayi[2], Edward Nelson Kankaka[3,4], Godfrey Kigozi[2], Oliver Laeyendecker[15,16], Thomas C Quinn[14,15,16], Steven J. Reynolds[2,15,16], John Santelli[18], David Serwadda[2,19], Nelson K. Sewankambo[19], Joseph Ssekasanvu[11], Robert Ssekubugu[2], Victor Ssempijja[5,6], Maria J Wawer[2,11]

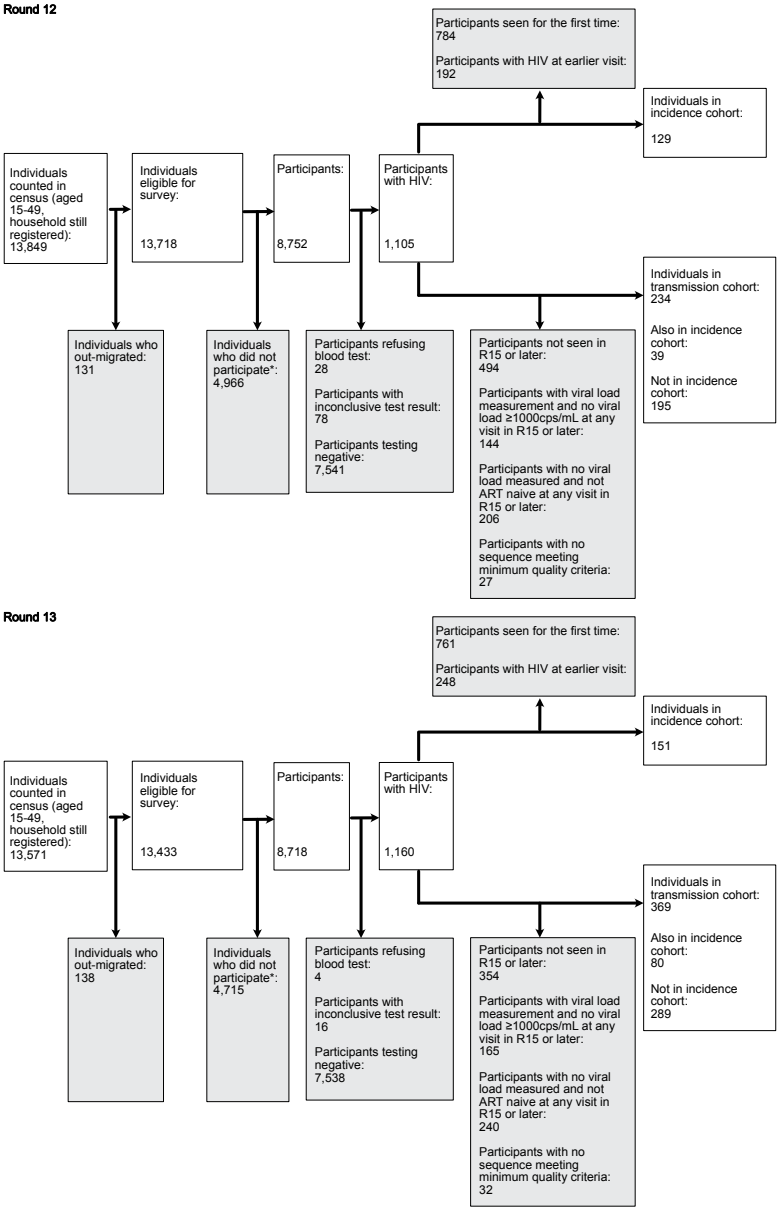
PANGEA-HIV consortium authors

Lucie Abeler Dörner[7], David Bonsall[9,10], Christophe Fraser[10,7], Tanya Golubchik[12,7], M Kate Grabowski[2,14,11], Joseph Kagaayi[2], Thomas C Quinn[14,15,16], Oliver Ratmann[1], Maria J Wawer[2,11]

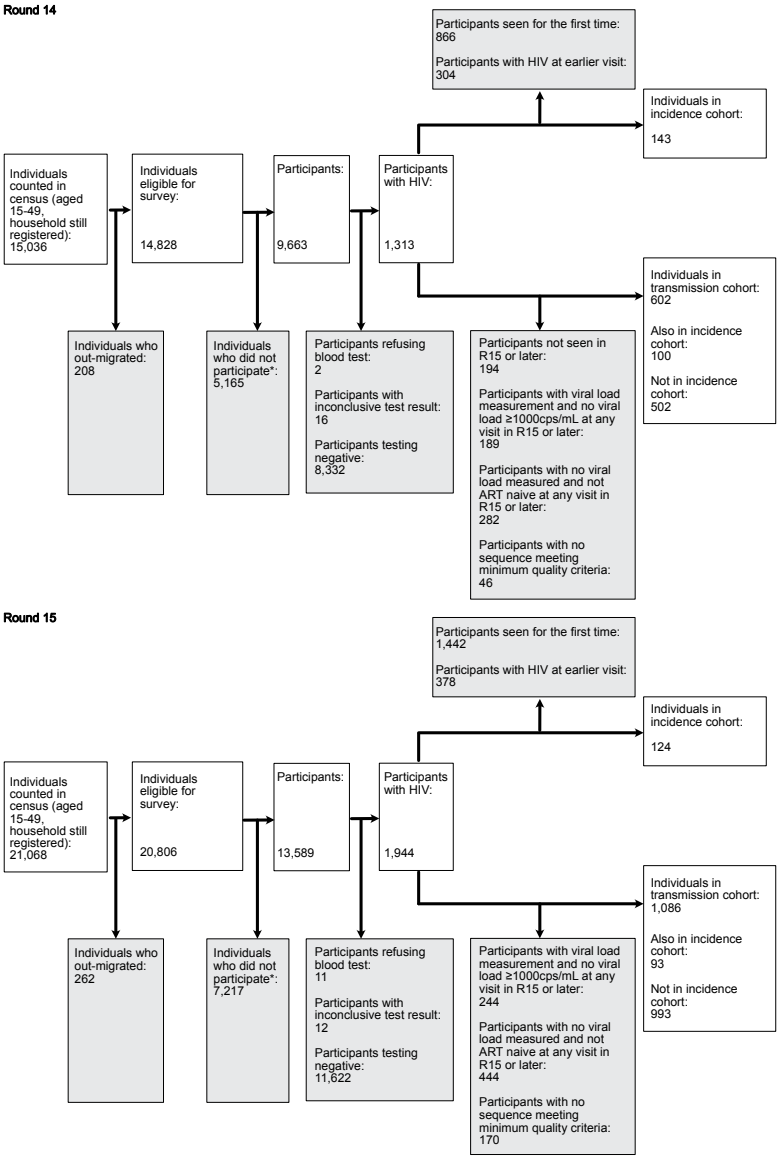
S1 Supplementary Figures



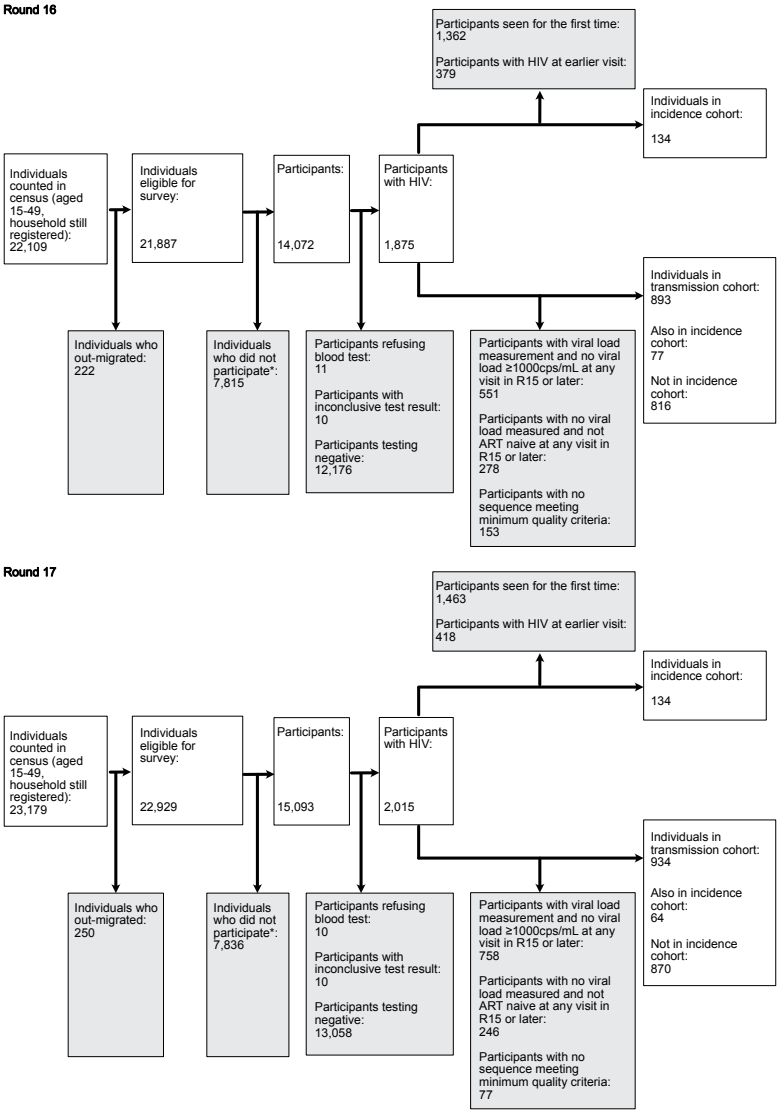
Supplementary Fig. S1: Flowchart of census eligible individuals through to individuals in the incidence and transmission cohorts.



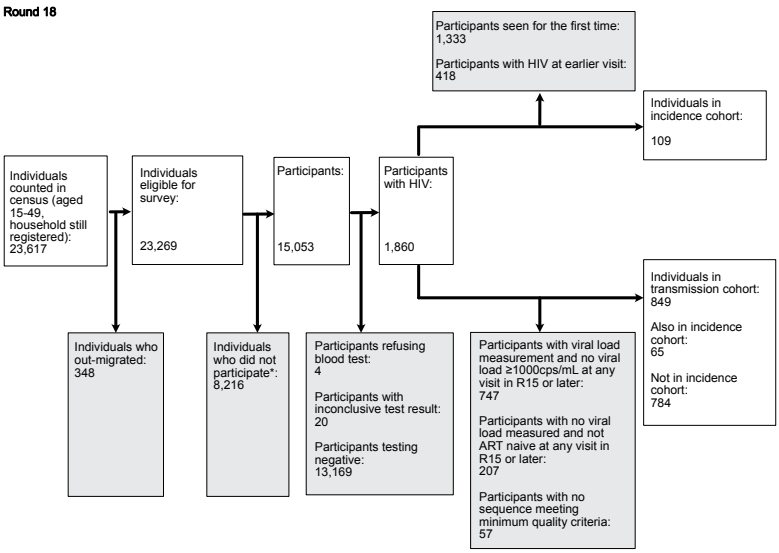
Supplementary Fig. S1: **Flowchart of census eligible individuals through to individuals in the incidence and transmission cohorts.** (cont.)



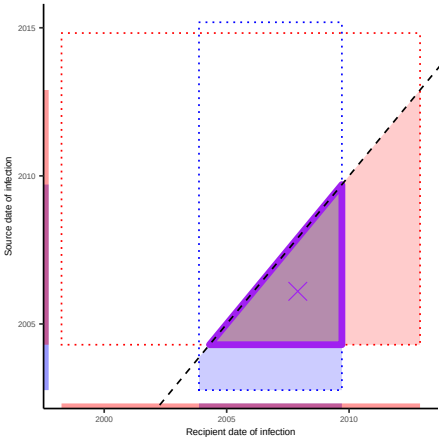
Supplementary Fig. S1: **Flowchart of census eligible individuals through to individuals in the incidence and transmission cohorts.** (cont.)



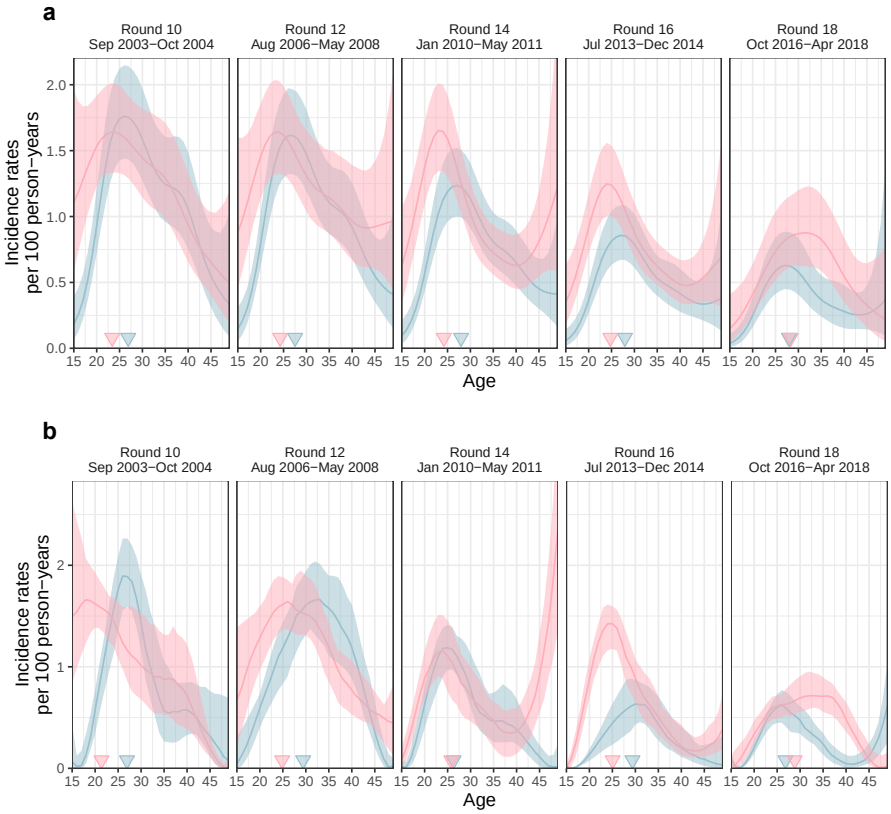
Supplementary Fig. S1: **Flowchart of census eligible individuals through to individuals in the incidence and transmission cohorts.** (cont.)



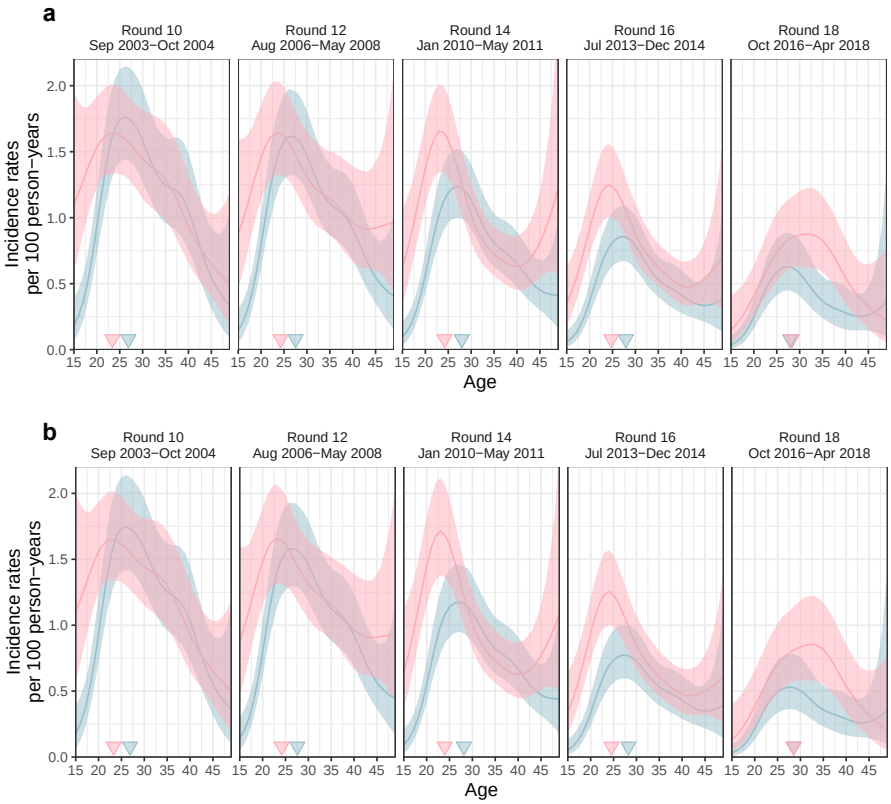
Supplementary Fig. S1: **Flowchart of census eligible individuals through to individuals in the incidence and transmission cohorts.** (cont.)



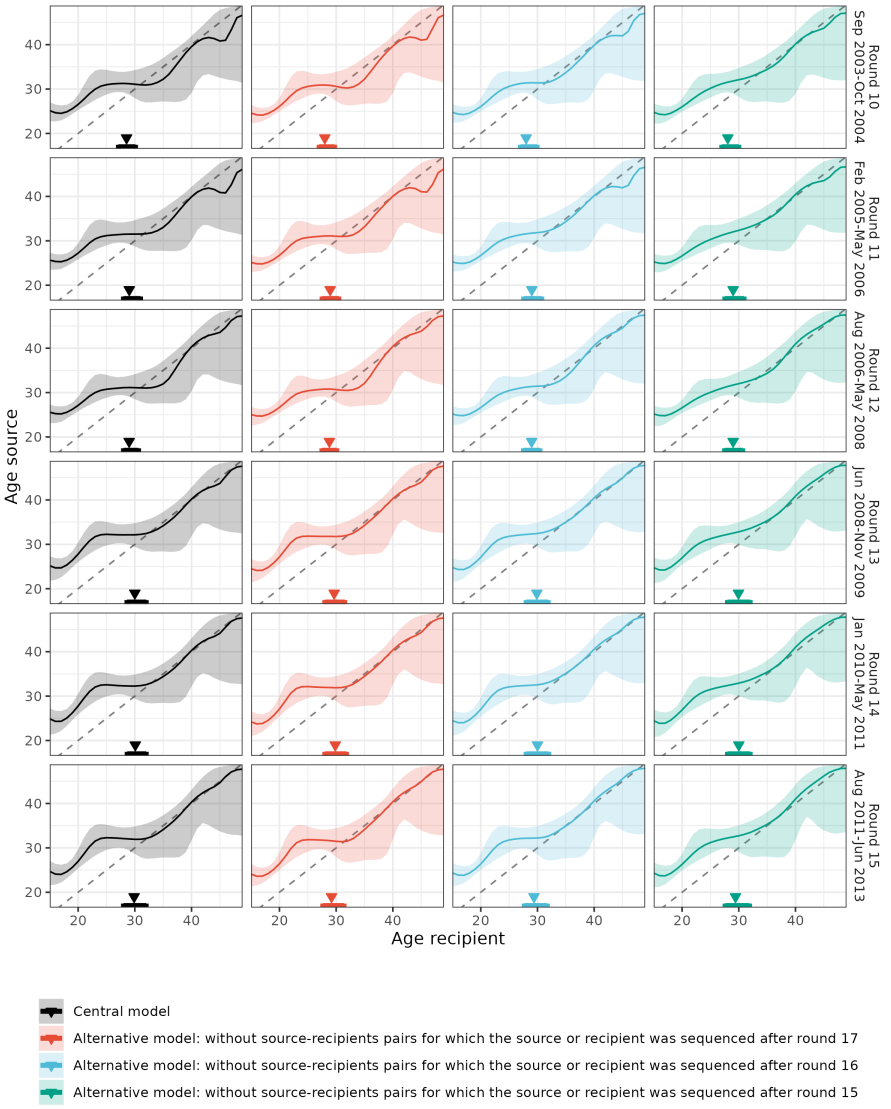
Supplementary Fig. S2: **Schema illustrating the refinement of phylogenetic time since infection estimates.**



Supplementary Fig. S3: **Comparison of incidence rate estimates under an individual-level additive effects Poisson regression model and a population-level LOESS model with independent age effects in each survey round.** (**a**) Mean and 95% uncertainty ranges of longitudinal age-specific incidence rates obtained with the individual-level additive effects Poisson regression model used in the central analysis (**b**) Same using a population-level LOESS model with independent age effects in each survey round.



Supplementary Fig. S4: **Comparison of incidence rate estimated on data containing all communities and data subset to 28 continuously surveyed communities** (**a**) Mean and 95% uncertainty ranges of longitudinal age-specific incidence rates estimated on data from all communities surveyed (**b**) Same using data subset to 28 continuously surveyed communities.



Supplementary Fig. S5: **Sensitivity in estimating the age of transmitting partners to right censoring of likely transmission pairs.** Posterior median (line) and 95% credible interval (ribbon) of the age of male transmitting partners by the age of the infected female (x-axis) by survey round (row facet) for the central and sensitivity analyses (column facet). Median and 95% credible interval of the age of male transmitting partners across the age of the infected female is indicated with a triangle and an error bar.

S2 Supplementary Tables

	Census-eligible individuals	Participants	Participants with HIV	Participants with HIV and with measured viral load	Participants with HIV reporting to be ART naïve†	Participants with HIV and with unsuppressed virus†	Participants with HIV and with virus ever deep-sequenced‡
Round 10, September 26, 2003 - November 23, 2004; 28 communities surveyed							
Total	11,976	7,407	884	–	–	–	115
Female	6,299	4,341	575	–	–	–	60
Age							
15-24	3,118	1,768	131	–	–	–	17
25-34	1,916	1,538	280	–	–	–	27
35-49	1,265	1,035	164	–	–	–	16
Male	5,677	3,066	309	–	–	–	55
Age							
15-24	2,672	1,186	38	–	–	–	9
25-34	1,845	1,132	145	–	–	–	27
35-49	1,160	748	126	–	–	–	19
Round 11, February 15, 2005 - June 30, 2006; 28 communities surveyed							
Total	12,528	8,273	1,002	884	–	–	176
Female	6,644	4,786	658	568	–	–	97
Age							
15-24	3,146	1,818	141	138	–	–	26
25-34	2,175	1,842	323	286	–	–	50
35-49	1,323	1,126	194	144	–	–	21
Male	5,884	3,487	344	316	–	–	79
Age							
15-24	2,670	1,293	30	30	–	–	6
25-34	1,956	1,290	160	153	–	–	40
35-49	1,258	904	154	133	–	–	33
Round 12, August 30, 2006 - June 06, 2008; 28 communities surveyed							
Total	13,718	8,752	1,105	912	–	–	234
Female	7,185	5,047	746	610	–	–	140
Age							
15-24	3,331	1,903	151	149	–	–	37
25-34	2,416	1,958	354	297	–	–	67
35-49	1,438	1,186	241	164	–	–	36
Male	6,533	3,705	359	302	–	–	94
Age							
15-24	2,866	1,426	26	25	–	–	8
25-34	2,200	1,305	168	156	–	–	50
35-49	1,467	974	165	121	–	–	36
Round 13, June 17, 2008 - July 12, 2009; 28 communities surveyed							
Total	13,433	8,718	1,160	900	–	–	369
Female	7,086	4,975	760	580	–	–	204
Age							
15-24	3,160	1,736	128	124	–	–	45
25-34	2,379	1,946	347	278	–	–	99
35-49	1,547	1,293	285	178	–	–	60
Male	6,347	3,743	400	320	–	–	165
Age							
15-24	2,749	1,397	32	31	–	–	19
25-34	2,042	1,275	177	160	–	–	82
35-49	1,556	1,071	191	129	–	–	64
Round 14, January 18, 2010 - June 21, 2011; 28 communities surveyed							
Total	14,828	9,663	1,313	964	–	–	602
Female	7,766	5,430	869	615	–	–	341
Age							
15-24	3,376	1,877	134	125	–	–	71
25-34	2,633	2,084	379	290	–	–	167
35-49	1,757	1,469	356	200	–	–	103
Male	7,062	4,233	444	349	–	–	261
Age							
15-24	2,963	1,617	40	38	–	–	31
25-34	2,276	1,398	185	163	–	–	120
35-49	1,823	1,218	219	148	–	–	110

† Unsuppressed virus was defined as a plasma viral load measurement above 1000 copies/mL plasma blood. In R10, participants were not asked about ART status and viral loads were not measured. In R11-R14, participants reported their ART status and viral loads were not measured. In R15, participants reported both their ART status and a subset of viral loads were measured. In R16-R18, participants reported both their ART status and viral loads were measured comprehensively in participants with HIV. ‡ Samples were selected for deep-sequencing from participants who had no viral load measured and reported being ART-naive or participants with viral load above 1,000 copies/mL plasma. Individuals participated across rounds, so for individuals participating in a given round, samples for sequencing could also be obtained in other rounds and we tabulate the proportion of participants ever deep-sequenced. Individuals with virus ever deep-sequenced were defined as HIV-positive individuals with deep-sequence output meeting minimum quality criteria, see Methods.

Supplementary Table S1: Characteristics of the RCCS study population.

	Census-eligible individuals	Participants	Participants with HIV	Participants with HIV and with measured viral load	Participants with HIV reporting to be ART naïve†	Participants with HIV and with unsuppressed virus†	Participants with HIV and with virus ever deep-sequenced‡
Round 15, August 10, 2011 - July 05, 2013; 33 communities surveyed							
Total	20,806	13,589	1,944	1,331	207	367	1,086
Female	10,782	7,538	1,287	844	122	232	637
Age							
15-24	4,751	2,742	217	186	23	31	157
25-34	3,631	2,825	568	405	64	101	307
35-49	2,400	1,971	502	253	35	100	173
Male	10,024	6,051	657	487	85	135	449
Age							
15-24	4,150	2,368	68	58	10	11	54
25-34	3,243	1,955	260	218	41	57	208
35-49	2,631	1,728	329	211	34	67	187
Round 16, July 08, 2013 - January 30, 2015; 35 communities surveyed							
Total	21,887	14,072	1,875	868	671	1,829	893
Female	11,346	7,816	1,255	537	390	1,224	521
Age							
15-24	5,089	2,891	194	129	97	189	83
25-34	3,547	2,669	502	238	175	486	249
35-49	2,710	2,256	559	170	118	549	189
Male	10,541	6,256	620	331	281	605	372
Age							
15-24	4,436	2,462	50	40	34	47	35
25-34	3,241	1,883	219	141	123	212	155
35-49	2,864	1,911	351	150	124	346	182
Round 17, February 23, 2015 - September 02, 2016; 35 communities surveyed							
Total	22,929	15,093	2,015	646	514	2,004	934
Female	11,990	8,377	1,390	408	304	1,384	554
Age							
15-24	5,393	3,035	205	94	84	204	97
25-34	3,544	2,723	529	194	147	525	250
35-49	3,053	2,619	656	120	73	655	207
Male	10,939	6,716	625	238	210	620	380
Age							
15-24	4,677	2,662	41	28	26	40	31
25-34	3,121	1,912	208	102	91	206	139
35-49	3,141	2,142	376	108	93	374	210
Round 18, October 03, 2016 - May 22, 2018; 35 communities surveyed							
Total	23,269	15,053	1,860	432	375	1,850	849
Female	12,193	8,331	1,275	263	206	1,271	492
Age							
15-24	5,484	3,049	158	72	63	158	80
25-34	3,472	2,592	461	117	95	457	208
35-49	3,237	2,690	656	74	48	656	204
Male	11,076	6,722	585	169	169	579	357
Age							
15-24	4,739	2,671	38	22	24	36	27
25-34	3,077	1,850	183	79	78	183	128
35-49	3,260	2,201	364	68	67	360	202

† Unsuppressed virus was defined as a plasma viral load measurement above 1000 copies/mL plasma blood. In R10, participants were not asked about ART status and viral loads were not measured. In R11-R14, participants reported their ART status and viral loads were not measured. In R15, participants reported both their ART status and a subset of viral loads were measured. In R16-R18, participants reported both their ART status and viral loads were measured comprehensively in participants with HIV. ‡ Samples were selected for deep-sequencing from participants who had no viral load measured and reported being ART-naive or participants with viral load above 1,000 copies/mL plasma. Individuals participated across rounds, so for individuals participating in a given round, samples for sequencing could also be obtained in other rounds and we tabulate the proportion of participants ever deep-sequenced. Individuals with virus ever deep-sequenced were defined as HIV-positive individuals with deep-sequence output meeting minimum quality criteria, see Methods.

Supplementary Table S1: **Characteristics of the RCCS study population (continued).**

Community Identifier†	Part of RCCS								
	Round 10	Round 11	Round 12	Round 13	Round 14	Round 15	Round 16	Round 17	Round 18
i-01	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-02	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-03	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-04	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-05	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-06	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-07	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-08	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-09	No	No	No	No	No	Yes	Yes	Yes	Yes
i-10	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-11	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-12	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-13	No	No	No	No	No	No	Yes	Yes	Yes
i-14	No	No	No	No	No	Yes	Yes	Yes	Yes
i-15	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-16	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-17	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-18	No	No	No	No	No	No	Yes	Yes	Yes
i-19	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-20	No	No	No	No	No	Yes	No	No	No
i-21	No	No	No	No	No	No	Yes	Yes	Yes
i-22	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-23	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-24	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-25	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-26	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-27	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-28	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-29	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-30	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-31	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-32	No	No	No	No	No	Yes	Yes	Yes	Yes
i-33	No	No	No	No	No	Yes	Yes	Yes	Yes
i-34	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-35	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i-36	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

† Three pairs of geographically close areas in peri-urban settings were merged into three communities.

Supplementary Table S2: **Communities surveyed by RCCS in rounds 10-18.**

	Incidence cohort†	Person-years‡	Incidence events§	Incidence rate estimate per 100 PY¶
Round 10, September 26, 2003 - November 23, 2004; 28 communities surveyed				
Total	7,372	9,464.33 [9,448.40-9,481.02]	122.0 [112.22-130.77]	1.32 [1.27-1.37]
Female	4,055	5,213.53 [5,201.59-5,224.77]	71.0 [61.22-77.00]	1.37 [1.30-1.45]
Age				
15-24	1,706	1,938.12 [1,928.82-1,944.71]	32.0 [28.22-37.77]	1.53 [1.40-1.68]
25-34	1,440	2,025.25 [2,015.02-2,032.45]	26.0 [21.00-34.77]	1.50 [1.38-1.63]
35-49	909	1,251.31 [1,247.69-1,255.27]	11.0 [8.00-13.77]	0.90 [0.81-1.01]
Male	3,317	4,252.08 [4,237.69-4,264.78]	51.0 [45.22-57.32]	1.26 [1.19-1.33]
Age				
15-24	1,328	1,522.78 [1,514.43-1,527.77]	12.0 [7.22-14.77]	1.04 [0.94-1.15]
25-34	1,254	1,718.06 [1,708.34-1,725.71]	29.0 [24.23-34.77]	1.61 [1.48-1.75]
35-49	735	1,011.40 [1,006.45-1,015.40]	10.0 [8.00-13.00]	1.00 [0.89-1.11]
Round 11, February 15, 2005 - June 30, 2006; 28 communities surveyed				
Total	7,787	11,484.46 [11,465.55-11,505.89]	144.0 [131.45-154.77]	1.29 [1.23-1.34]
Female	4,291	6,261.90 [6,247.27-6,278.78]	84.0 [76.22-91.00]	1.36 [1.27-1.44]
Age				
15-24	1,646	2,088.35 [2,078.11-2,095.10]	31.0 [25.45-37.00]	1.48 [1.34-1.64]
25-34	1,667	2,654.38 [2,644.12-2,664.53]	39.0 [34.00-43.77]	1.46 [1.34-1.59]
35-49	978	1,519.83 [1,515.14-1,526.38]	13.0 [11.00-17.00]	1.00 [0.90-1.11]
Male	3,496	5,222.93 [5,209.14-5,243.01]	60.0 [51.45-65.00]	1.20 [1.14-1.27]
Age				
15-24	1,323	1,781.07 [1,774.24-1,787.04]	17.0 [12.00-20.00]	0.97 [0.88-1.06]
25-34	1,356	2,145.73 [2,135.78-2,156.34]	31.0 [26.23-36.77]	1.55 [1.43-1.69]
35-49	817	1,296.59 [1,291.00-1,302.85]	11.0 [8.00-14.77]	0.95 [0.86-1.06]
Round 12, August 30, 2006 - June 06, 2008; 28 communities surveyed				
Total	8,480	12,396.23 [12,369.28-12,422.54]	168.0 [151.12-177.33]	1.21 [1.16-1.28]
Female	4,598	6,648.49 [6,632.13-6,668.15]	95.0 [84.67-101.00]	1.31 [1.24-1.43]
Age				
15-24	1,669	2,100.25 [2,091.86-2,108.29]	31.0 [25.00-36.77]	1.44 [1.29-1.62]
25-34	1,869	2,883.98 [2,869.79-2,897.39]	45.0 [39.23-52.77]	1.39 [1.27-1.54]
35-49	1,060	1,666.57 [1,659.44-1,673.22]	19.0 [16.00-22.00]	1.02 [0.90-1.18]
Male	3,882	5,746.59 [5,732.78-5,759.56]	72.0 [65.22-79.00]	1.09 [1.02-1.17]
Age				
15-24	1,460	1,990.10 [1,984.61-1,996.47]	15.0 [10.22-17.00]	0.83 [0.75-0.92]
25-34	1,474	2,246.22 [2,235.56-2,252.48]	38.0 [32.00-44.00]	1.46 [1.34-1.59]
35-49	948	1,511.44 [1,503.26-1,516.12]	19.0 [16.23-23.77]	0.88 [0.80-1.00]
Round 13, June 17, 2008 - July 12, 2009; 28 communities surveyed				
Total	8,770	11,823.39 [11,802.83-11,845.07]	136.0 [125.00-145.55]	1.08 [1.04-1.15]
Female	4,728	6,331.90 [6,313.25-6,348.15]	83.0 [73.45-89.00]	1.21 [1.15-1.33]
Age				
15-24	1,624	1,942.24 [1,932.52-1,949.30]	29.0 [25.00-35.77]	1.38 [1.26-1.54]
25-34	1,948	2,723.50 [2,708.26-2,732.27]	37.0 [32.00-43.55]	1.27 [1.17-1.41]
35-49	1,156	1,667.46 [1,661.12-1,673.50]	16.0 [12.00-21.77]	0.90 [0.81-1.06]
Male	4,042	5,490.33 [5,477.08-5,500.21]	52.0 [47.23-59.55]	0.94 [0.89-1.02]
Age				
15-24	1,491	1,900.09 [1,893.07-1,905.26]	17.0 [13.23-21.55]	0.69 [0.63-0.77]
25-34	1,475	2,004.64 [1,996.54-2,012.04]	23.0 [18.00-27.77]	1.30 [1.20-1.43]
35-49	1,076	1,586.25 [1,578.36-1,593.72]	13.0 [10.00-16.77]	0.78 [0.70-0.89]
Round 14, January 18, 2010 - June 21, 2011; 28 communities surveyed				
Total	9,290	12,359.17 [12,344.41-12,374.39]	107.5 [97.45-118.00]	0.93 [0.89-0.97]
Female	4,963	6,624.63 [6,608.63-6,638.01]	63.0 [55.00-71.78]	1.07 [1.00-1.13]
Age				
15-24	1,706	1,998.64 [1,991.11-2,007.00]	23.0 [19.00-30.00]	1.30 [1.17-1.43]
25-34	1,999	2,766.97 [2,761.12-2,775.18]	24.0 [15.68-30.55]	1.13 [1.03-1.22]
35-49	1,258	1,857.69 [1,850.32-1,863.26]	15.0 [11.00-19.00]	0.74 [0.66-0.82]
Male	4,327	5,734.81 [5,725.76-5,744.10]	46.0 [39.23-50.00]	0.77 [0.73-0.82]
Age				
15-24	1,642	1,988.89 [1,983.58-1,992.83]	14.0 [10.00-16.77]	0.55 [0.50-0.61]
25-34	1,487	1,999.30 [1,992.87-2,005.68]	22.0 [19.00-26.77]	1.11 [1.02-1.21]
35-49	1,198	1,747.16 [1,742.21-1,752.55]	9.0 [6.22-12.77]	0.65 [0.58-0.73]

† Number of RCCS study participants who were HIV-negative at their first visit and had at least one subsequent follow-up visit. ‡ Number of person-years of HIV acquisition risk. § Number of incidence events. The infection date was imputed at random to have occurred between the last negative and first positive survey visit dates, and the incidence event was attributed to the corresponding survey round 50 times. The range of the person-years and incidence events across the 50 data sets with imputed exposure times are presented. ¶ Estimated incidence rate per 100 person-years. The confidence interval of the estimated incidence rate incorporates both the variability of the estimation procedure and the data imputation procedure.

Supplementary Table S3: **Characteristics of the longitudinal HIV incidence cohort.**

	Incidence cohort†	Person-years‡	Incidence events§	Incidence rate estimate per 100 PY¶
Round 15, August 10, 2011 - July 05, 2013; 33 communities surveyed				
Total	10,441	17,621.81 [17,596.06-17,643.04]	140.0 [129.45-148.78]	0.79 [0.76-0.83]
Female	5,520	9,227.87 [9,204.36-9,242.47]	87.0 [79.22-94.77]	0.94 [0.88-0.99]
Age				
15-24	1,892	2,742.21 [2,728.96-2,752.62]	37.0 [31.23-43.77]	1.17 [1.05-1.30]
25-34	2,184	3,728.50 [3,713.89-3,735.50]	38.0 [34.00-42.77]	1.02 [0.92-1.10]
35-49	1,444	2,757.15 [2,750.51-2,765.25]	12.0 [9.23-15.77]	0.61 [0.54-0.68]
Male	4,921	8,395.89 [8,383.12-8,406.96]	52.0 [47.23-60.00]	0.64 [0.60-0.67]
Age				
15-24	1,848	2,842.07 [2,836.70-2,847.92]	11.0 [8.00-14.00]	0.45 [0.41-0.50]
25-34	1,657	2,865.12 [2,856.30-2,874.43]	31.0 [26.23-35.00]	0.92 [0.84-1.01]
35-49	1,416	2,687.98 [2,679.81-2,695.84]	11.0 [6.22-14.00]	0.52 [0.46-0.59]
Round 16, July 08, 2013 - January 30, 2015; 35 communities surveyed				
Total	12,142	16,633.57 [16,621.16-16,648.28]	108.5 [98.45-116.78]	0.66 [0.63-0.70]
Female	6,380	8,745.06 [8,737.02-8,758.26]	72.5 [64.22-80.78]	0.80 [0.75-0.86]
Age				
15-24	2,236	2,699.66 [2,693.50-2,703.90]	24.5 [21.23-31.55]	0.89 [0.80-0.99]
25-34	2,328	3,202.15 [3,195.15-3,209.33]	33.0 [27.00-38.77]	0.94 [0.85-1.04]
35-49	1,816	2,843.90 [2,839.95-2,847.65]	15.0 [11.22-18.00]	0.55 [0.49-0.62]
Male	5,762	7,888.21 [7,881.14-7,895.54]	35.0 [31.00-39.00]	0.51 [0.48-0.55]
Age				
15-24	2,206	2,803.63 [2,801.36-2,806.94]	8.0 [7.00-10.00]	0.37 [0.32-0.41]
25-34	1,813	2,501.71 [2,496.99-2,507.33]	17.0 [13.00-20.00]	0.77 [0.68-0.84]
35-49	1,743	2,582.08 [2,578.79-2,588.18]	9.0 [6.22-14.00]	0.43 [0.37-0.49]
Round 17, February 23, 2015 - September 02, 2016; 35 communities surveyed				
Total	12,738	17,437.70 [17,422.40-17,448.35]	89.5 [80.22-95.78]	0.56 [0.53-0.59]
Female	6,680	9,116.75 [9,106.85-9,127.51]	57.0 [48.45-61.77]	0.68 [0.64-0.72]
Age				
15-24	2,327	2,796.00 [2,790.86-2,799.37]	11.0 [8.00-13.77]	0.62 [0.56-0.70]
25-34	2,286	3,187.45 [3,182.16-3,194.41]	28.0 [23.23-32.00]	0.87 [0.80-0.95]
35-49	2,067	3,133.05 [3,127.18-3,138.08]	17.0 [15.00-21.77]	0.53 [0.48-0.59]
Male	6,058	8,321.01 [8,312.47-8,328.62]	32.0 [27.45-36.00]	0.43 [0.40-0.46]
Age				
15-24	2,353	3,012.95 [3,009.30-3,015.97]	9.0 [8.00-11.00]	0.30 [0.27-0.35]
25-34	1,796	2,485.06 [2,479.65-2,490.25]	14.0 [10.22-18.00]	0.65 [0.58-0.73]
35-49	1,909	2,823.11 [2,818.49-2,830.23]	9.0 [5.22-12.00]	0.36 [0.30-0.42]
Round 18, October 03, 2016 - May 22, 2018; 35 communities surveyed				
Total	12,217	17,992.52 [17,982.46-18,005.50]	89.0 [83.00-97.78]	0.50 [0.47-0.54]
Female	6,425	9,624.65 [9,617.33-9,633.49]	57.0 [53.00-65.00]	0.62 [0.56-0.68]
Age				
15-24	2,174	2,703.74 [2,699.61-2,706.79]	12.0 [10.00-13.77]	0.42 [0.35-0.51]
25-34	2,125	3,249.56 [3,241.74-3,255.03]	26.0 [24.00-30.77]	0.85 [0.75-0.96]
35-49	2,126	3,671.67 [3,665.44-3,676.22]	19.0 [16.23-23.00]	0.56 [0.47-0.65]
Male	5,792	8,368.03 [8,361.41-8,377.69]	32.0 [30.00-35.00]	0.37 [0.34-0.40]
Age				
15-24	2,229	2,895.16 [2,891.38-2,899.31]	10.0 [8.00-12.00]	0.26 [0.22-0.31]
25-34	1,664	2,496.56 [2,493.55-2,501.84]	14.0 [12.00-17.00]	0.56 [0.49-0.64]
35-49	1,899	2,976.37 [2,972.27-2,980.31]	8.0 [6.00-11.00]	0.31 [0.25-0.37]

† Number of RCCS study participants who were HIV-negative at their first visit and had at least one subsequent follow-up visit. ‡ Number of person-years of HIV acquisition risk. § Number of incidence events. The infection date was imputed at random to have occurred between the last negative and first positive survey visit dates, and the incidence event was attributed to the corresponding survey round 50 times. The range of the person-years and incidence events across the 50 data sets with imputed exposure times are presented. ¶ Estimated incidence rate per 100 person-years. The confidence interval of the estimated incidence rate incorporates both the variability of the estimation procedure and the data imputation procedure.

Supplementary Table S3: **Characteristics of the longitudinal HIV incidence cohort (continued).**

	Akaike information criterion (AIC)		% observations within 95% prediction intervals		
	Men	Women	Men	Women	All
Central model					
	8,032	**11,579**	98.77%	**98.82%**	**98.80%**
	[7,937-8,140]	[11,508-11,688]	[97.78-99.68]	[97.78-99.68]	[98.10-99.49]
Alternative models					
with 2D GP over age	8,033	11,580	**98.84%**	93.32%	96.08%
and survey round	[7,938-8,141]	[11,511-11,690]	[98.10-99.68]	[91.18-95.10]	[94.96-96.95]
without interaction term between	8,033	11,592	98.79%	93.83%	96.31%
age and survey round	[7,938-8,142]	[11,521-11,706]	[97.78-99.68]	[92.06-95.24]	[95.27-97.23]
with 2D GP over age					
and survey round and	8,035	11,590	98.82%	93.45%	96.13%
without interaction term between	[7,939-8,143]	[11,517-11,701]	[97.78-99.68]	[90.94-95.24]	[94.99-97.23]
age and survey round					

Supplementary Table S4: **Model comparison for estimating longitudinal, age-specific incidence rates.**

	Participants with HIV	Participants with HIV >1,000 cps/mL or or reporting no ART use if viral load was not measured	Participants with HIV and with virus ever deep-sequenced with Illumina MiSeq in PANGEA-HIV 1 †	Participants with HIV and with virus ever deep-sequenced with Illumina HiSeq in PANGEA-HIV 1 ‡	Participants with HIV and with virus ever deep-sequenced with Illumina NovaSeq in PANGEA-HIV 2 §	Participants with HIV and with virus ever deep-sequenced	Sequence sampling coverage of participants with HIV
	(n)	(n)	(n)	(n)	(n)	(n)	(%)
Round 10, September 26, 2003 - November 23, 2004; 28 communities surveyed							
Total	884	884	54	3	58	115	13.01
Female	575	575	25	2	33	60	10.43
Age							
15-24	131	131	8	1	8	17	12.98
25-34	280	280	9	0	18	27	9.64
35-49	164	164	8	1	7	16	9.76
Male	309	309	29	1	25	55	17.8
Age							
15-24	38	38	6	0	3	9	23.68
25-34	145	145	12	1	14	27	18.62
35-49	126	126	11	0	8	19	15.08
Round 11, February 15, 2005 - June 30, 2006; 28 communities surveyed							
Total	1002	884	80	3	93	176	17.56
Female	658	568	41	2	54	97	14.74
Age							
15-24	141	138	8	1	17	26	18.44
25-34	323	286	22	0	28	50	15.48
35-49	194	144	11	1	9	21	10.82
Male	344	316	39	1	39	79	22.97
Age							
15-24	30	30	4	0	2	6	20
25-34	160	153	20	1	19	40	25
35-49	154	133	15	0	18	33	21.43
Round 12, August 30, 2006 - June 06, 2008; 28 communities surveyed							
Total	1105	912	117	3	114	234	21.18
Female	746	610	63	2	75	140	18.77
Age							
15-24	151	149	16	1	20	37	24.5
25-34	354	297	31	0	36	67	18.93
35-49	241	164	16	1	19	36	14.94
Male	359	302	54	1	39	94	26.18
Age							
15-24	26	25	6	0	2	8	30.77
25-34	168	156	28	1	21	50	29.76
35-49	165	121	20	0	16	36	21.82
Round 13, June 17, 2008 - July 12, 2009; 28 communities surveyed							
Total	1160	900	179	3	187	369	31.81
Female	760	580	93	2	109	204	26.84
Age							
15-24	128	124	22	1	22	45	35.16
25-34	347	278	44	0	55	99	28.53
35-49	285	178	27	1	32	60	21.05
Male	400	320	86	1	78	165	41.25
Age							
15-24	32	31	14	0	5	19	59.38
25-34	177	160	41	1	40	82	46.33
35-49	191	129	31	0	33	64	33.51
Round 14, January 18, 2010 - June 21, 2011; 28 communities surveyed							
Total	1313	964	305	3	294	602	45.85
Female	869	615	166	2	173	341	39.24
Age							
15-24	134	125	40	0	31	71	52.99
25-34	379	290	81	1	85	167	44.06
35-49	356	200	45	1	57	103	28.93
Male	444	349	139	1	121	261	58.78
Age							
15-24	40	38	20	0	11	31	77.5
25-34	185	163	58	1	61	120	64.86
35-49	219	148	61	0	49	110	50.23

† RNA samples were sequenced using the protocol of[75] at the Wellcome Trust Sanger Institute, Hinxton, UK on Illumina MiSeq platforms. Deep-sequences reported satisfied minimum quality criteria for deep-sequence phylogenetic analysis, see Methods. ‡ As for previous column, on Illumina HiSeq platforms. § RNA samples were sequenced using the protocol of[77] at the Oxford Genomics Centre, Oxford, UK on Illumina NovaSeq 6000 platforms. Deep-sequences reported satisfied minimum quality criteria for deep-sequence phylogenetic analysis, see Methods.

Supplementary Table S5: **Longitudinal HIV deep-sequencing.**

	Participants with HIV (n)	Participants with HIV >1,000 cps/mL or reporting no ART use if viral load was not measured (n)	Participants with HIV and with virus ever deep-sequenced with Illumina MiSeq in PANGEA-HIV 1 † (n)	Participants with HIV and with virus ever deep-sequenced with Illumina HiSeq in PANGEA-HIV 1 ‡ (n)	Participants with HIV and with virus ever deep-sequenced with Illumina NovaSeq in PANGEA-HIV 2 § (n)	Participants with HIV and with virus ever deep-sequenced (n)	Sequence sampling coverage of participants with HIV (%)
Round 15, August 10, 2012 - July 05, 2013; 33 communities surveyed							
Total	1901	1298	282	2	802	1086	57.13
Female	1264	827	152	1	484	637	50.4
Age							
15-24	209	178	23	0	134	157	75.12
25-34	557	398	85	1	221	307	55.12
35-49	498	251	44	0	129	173	34.74
Male	637	471	130	1	318	449	70.49
Age							
15-24	67	57	17	0	37	54	80.6
25-34	249	208	55	0	153	208	83.53
35-49	321	206	58	1	128	187	58.26
Round 16, July 08, 2013 - January 30, 2015; 35 communities surveyed							
Total	1874	869	383	3	506	892	47.6
Female	1254	536	212	1	307	520	41.47
Age							
15-24	194	129	36	0	47	83	42.78
25-34	502	238	108	1	140	249	49.6
35-49	558	169	68	0	120	188	33.69
Male	620	333	171	2	199	372	60
Age							
15-24	50	40	21	0	14	35	70
25-34	219	141	75	0	80	155	70.78
35-49	351	152	75	2	105	182	51.85
Round 17, February 23, 2015 - September 02, 2016; 35 communities surveyed							
Total	2015	639	604	4	326	934	46.35
Female	1390	402	348	2	204	554	39.86
Age							
15-24	205	91	82	0	15	97	47.32
25-34	529	190	163	2	85	250	47.26
35-49	656	121	103	0	104	207	31.55
Male	625	237	256	2	122	380	60.8
Age							
15-24	41	28	28	0	3	31	75.61
25-34	208	102	101	0	38	139	66.83
35-49	376	107	127	2	81	210	55.85
Round 18, October 03, 2016 - May 22, 2018; 35 communities surveyed							
Total	1860	416	565	2	282	849	45.65
Female	1275	255	315	1	176	492	38.59
Age							
15-24	158	71	72	0	8	80	50.63
25-34	461	111	135	1	72	208	45.12
35-49	656	73	108	0	96	204	31.1
Male	585	161	250	1	106	357	61.03
Age							
15-24	38	22	26	0	1	27	71.05
25-34	183	76	101	0	27	128	69.95
35-49	364	63	123	1	78	202	55.49

† RNA samples were sequenced using the protocol of[75] at the Wellcome Trust Sanger Institute, Hinxton, UK on Illumina MiSeq platforms. Deep-sequences reported satisfied minimum quality criteria for deep-sequence phylogenetic analysis, see Methods. ‡ As for previous column, on Illumina HiSeq platforms. § RNA samples were sequenced using the protocol of[77] at the Oxford Genomics Centre, Oxford, UK on Illumina NovaSeq 6000 platforms. Deep-sequences reported satisfied minimum quality criteria for deep-sequence phylogenetic analysis, see Methods.

Supplementary Table S5: **Longitudinal HIV deep-sequencing (continued).**

Observed transmission events within 95% prediction interval (%)	**Observed transmission events vs. predicted transmission events** (MAE)†	**Incidence rate prior mean within 95% posterior range** (%)	**Incidence rate prior mean vs. incidence rate posterior median** (MAE)†
Central model			
$\log \beta^{g\to h}_{r,i,j} = \hat{\boldsymbol{c}}^{g\to h}(i,j) + \gamma_0 + \gamma_g + \gamma_r + \gamma_{p(r)} + \boldsymbol{f}^{g\to h}_0(i,j) + \boldsymbol{f}^{g\to h}_r(j) + \boldsymbol{f}^{g\to h}_{p(r)}(i)$, (7c)			
99.63	0.0459	97.14	0.00032
Alternative models			
$\log \beta^{g\to h}_{r,i,j} = \hat{\boldsymbol{c}}^{g\to h}(i,j) + \gamma_0 + \gamma_g + \gamma_r + \gamma_{p(r)} + \boldsymbol{f}^{g\to h}_0(i,j) + \boldsymbol{f}^{g\to h}_{p(r)}(i)$, (12a)			
99.59	0.0473	67.78	0.00057
$\log \beta^{g\to h}_{r,i,j} = \hat{\boldsymbol{c}}^{g\to h}(i,j) + \gamma_0 + \gamma_g + \gamma_r + \gamma_{p(r)} + \boldsymbol{f}^{g\to h}_0(i,j) + \boldsymbol{f}^{g\to h}_{p(r)}(j)$, (12b)			
99.61	0.0467	67.62	0.00058
$\log \beta^{g\to h}_{r,i,j} = \hat{\boldsymbol{c}}^{g\to h}(i,j) + \gamma_0 + \gamma_g + \gamma_r + \gamma_{p(r)} + \boldsymbol{f}^{g\to h}_0(i,j) + \boldsymbol{f}^{g\to h}_{p(r)}(i,j)$, (12c)			
99.57	0.0471	68.89	0.00056
$\log \beta^{g\to h}_{r,i,j} = \hat{\boldsymbol{c}}^{g\to h}(i,j) + \gamma_0 + \gamma_g + \gamma_r + \gamma_{p(r)} + \boldsymbol{f}^{g\to h}_0(i,j) + \boldsymbol{f}^{g\to h}_r(j)$, (12d)			
99.57	0.0457	96.35	0.00033
$\log \beta^{g\to h}_{r,i,j} = \hat{\boldsymbol{c}}^{g\to h}(i,j) + \gamma_0 + \gamma_g + \gamma_r + \gamma_{p(r)} + \boldsymbol{f}^{g\to h}_0(i,j) + \boldsymbol{f}^{g\to h}_r(j) + \boldsymbol{f}^{g\to h}_{p(r)}(j)$, (12e)			
99.53	0.0459	97.94	0.00031
$\log \beta^{g\to h}_{r,i,j} = \hat{\boldsymbol{c}}^{g\to h}(i,j) + \gamma_0 + \gamma_g + \gamma_r + \gamma_{p(r)} + \boldsymbol{f}^{g\to h}_0(i,j) + \boldsymbol{f}^{g\to h}_r(j) + \boldsymbol{f}^{g\to h}_{p(r)}(i,j)$, (12f)			
99.61	0.0459	97.14	0.00031

† MAE: Mean absolute error.

Supplementary Table S6: Model comparison for estimating longitudinal, age-specific transmission flows.

Transmission direction	Male-female difference in age at transmission	Infected partner by age at transmission			
		15-24 years (%)†	25-34 years (%)†	35-49 years (%)†	Total (%)†
Round 10, September 26, 2003 - November 23, 2004; 28 communities surveyed					
Male to female	Total	31.9% [30.2-33.5]	18.8% [17.9-19.7]	7.3% [6.7-7.9]	57.9% [56.2-59.6]
	<0 years	0.4% [0.2-0.6]	5.6% [3.9-7.4]	4.0% [2.6-5.5]	10.0% [7.5-12.5]
	0-6 years	15.5% [12.3-18.9]	7.7% [6.2-9.3]	3.0% [1.8-4.4]	26.3% [22.4-30.4]
	>6 years	16.0% [12.7-19.2]	5.4% [3.9-7.3]	0.2% [0.0-0.5]	21.6% [17.6-25.7]
Female to male	Total	14.8% [13.9-15.8]	20.6% [19.7-21.6]	6.6% [6.2-7.1]	42.1% [40.4-43.8]
	<0 years	6.6% [4.9-8.3]	4.7% [3.2-6.5]	0.4% [0.2-0.8]	11.7% [8.8-14.9]
	0-6 years	8.2% [6.2-10.1]	11.8% [9.9-13.4]	2.6% [1.8-3.4]	22.5% [19.4-25.6]
	>6 years	0.1% [0.0-0.2]	4.1% [2.7-5.9]	3.6% [2.6-4.7]	7.8% [5.7-10.1]
Total		46.7% [45.3-48.1]	39.4% [38.3-40.6]	13.9% [13.2-14.6]	100%
Round 15, August 10, 2011 - July 05, 2013; 33 communities surveyed					
Male to female	Total	32.2% [30.2-34.3]	22.0% [20.7-23.4]	7.7% [7.0-8.5]	61.9% [60.2-63.7]
	<0 years	0.5% [0.2-0.8]	6.0% [4.1-8.1]	3.8% [2.3-5.4]	10.3% [7.6-13.1]
	0-6 years	15.4% [12.2-19.0]	9.0% [7.1-11.0]	3.6% [2.2-5.0]	28.0% [23.8-32.4]
	>6 years	16.2% [12.8-19.7]	7.0% [5.1-9.1]	0.3% [0.1-0.8]	23.6% [19.2-28.0]
Female to male	Total	11.5% [10.6-12.4]	18.8% [17.8-19.9]	7.8% [7.2-8.4]	38.1% [36.3-39.8]
	<0 years	6.2% [4.8-7.7]	4.6% [3.2-6.4]	0.6% [0.2-1.1]	11.4% [8.8-14.3]
	0-6 years	5.2% [3.9-6.6]	11.3% [9.6-12.8]	3.2% [2.2-4.3]	19.7% [16.9-22.4]
	>6 years	0.0% [0.0-0.0]	2.9% [1.9-4.1]	3.9% [2.8-5.2]	6.9% [5.1-8.8]
Total		43.6% [41.8-45.5]	40.9% [39.3-42.4]	15.5% [14.5-16.4]	100%
Round 18, October 03, 2016 - May 22, 2018; 35 communities surveyed					
Male to female	Total	20.6% [18.1-23.4]	27.3% [25.2-29.5]	14.7% [13.3-16.3]	62.8% [60.2-65.2]
	<0 years	0.3% [0.1-0.6]	6.7% [3.9-10.1]	7.0% [4.5-9.7]	14.0% [9.8-18.8]
	0-6 years	8.1% [5.6-11.0]	12.0% [9.1-15.0]	7.1% [4.7-9.7]	27.3% [23.1-31.8]
	>6 years	12.1% [9.3-15.2]	8.5% [5.8-11.9]	0.5% [0.1-1.5]	21.3% [16.8-26.3]
Female to male	Total	11.2% [9.9-12.6]	17.4% [15.9-19.1]	8.6% [7.6-9.7]	37.2% [34.8-39.8]
	<0 years	5.5% [3.9-7.3]	3.8% [2.5-5.5]	0.5% [0.2-1.1]	9.8% [7.2-13.0]
	0-6 years	5.7% [3.9-7.4]	10.6% [8.9-12.3]	3.2% [2.0-4.5]	19.5% [16.6-22.4]
	>6 years	0.0% [0.0-0.1]	2.9% [1.9-4.3]	4.9% [3.5-6.5]	7.9% [5.9-10.1]
Total		31.8% [29.4-34.5]	44.8% [42.5-47.0]	23.4% [21.6-25.2]	100%

† Posterior median flow estimates and 95% credible intervals in each survey round.

Supplementary Table S7: **Longitudinal HIV transmission flows by age and gender.**

	Participants (n)	**Contacts with reported partner characteristics** (%)	**Reported contacts per participant** (n)	**Estimated contacts per person** (median, 95% CrI)	**Estimated reporting bias** (median, 95% CrI)	**Reported contacts scaled to population** (n)	**Estimated contacts scaled to population** (median, 95% CrI)
Total	13,277	85.1	0.74	0.84 [0.76, 0.95]	0.1 [0.02, 0.21]	16,025	18,183 [16,450, 20,613]
Female	7,375	87.69	0.64	0.81 [0.74, 0.91]	0.17 [0.10, 0.27]	7,189	9,092 [8,284, 10,238]
Age							
15-19	1,296	84.20	0.34	0.48 [0.44, 0.54]	0.14 [0.09, 0.20]	844	1,187 [1,067, 1,321]
20-24	1,378	91.06	0.84	1.17 [1.09, 1.25]	0.33 [0.25, 0.41]	1,787	2,487 [2,324, 2,662]
25-29	1,432	85.99	0.90	1.18 [1.10, 1.26]	0.27 [0.20, 0.36]	1,704	2,221 [2,074, 2,381]
30-34	1,323	87.64	0.84	0.99 [0.92, 1.08]	0.15 [0.07, 0.24]	1,334	1,569 [1,451, 1,705]
35-39	1,007	87.60	0.75	0.83 [0.75, 0.95]	0.08 [0.00, 0.20]	849	942 [847, 1,075]
40-44	562	90.03	0.60	0.65 [0.55, 0.81]	0.05 [-0.05, 0.21]	436	472 [398, 588]
45-49	377	83.73	0.49	0.34 [0.21, 0.61]	-0.15 [-0.28, 0.12]	236	164 [102, 293]
50-54	0	-	-	0.13 [0.06, 0.36]	- -	-	43 [20, 124]
55-59	0	-	-	0.01 [0.00, 0.18]	- -	-	4 [1, 45]
60-64	0	-	-	0.01 [0.00, 0.14]	- -	-	1 [0, 24]
65-69	0	-	-	0.01 [0.00, 0.17]	- -	-	1 [0, 20]
Male	5,902	82.58	0.85	0.88 [0.79, 1.00]	0.02 [-0.06, 0.15]	8,836	9,091 [8,166, 10,374]
Age							
15-19	1,295	66.42	0.20	0.17 [0.14, 0.20]	-0.04 [-0.06, -0.01]	444	363 [306, 431]
20-24	1,001	75.50	0.84	0.79 [0.72, 0.87]	-0.04 [-0.11, 0.03]	1,528	1,447 [1,321, 1,585]
25-29	1,001	82.29	1.17	1.15 [1.07, 1.24]	-0.02 [-0.10, 0.07]	1,928	1,902 [1,763, 2,049]
30-34	913	84.05	1.26	1.28 [1.19, 1.37]	0.02 [-0.08, 0.11]	1,858	1,881 [1,747, 2,022]
35-39	796	83.82	1.36	1.31 [1.21, 1.41]	-0.05 [-0.14, 0.05]	1,587	1,530 [1,418, 1,648]
40-44	554	88.94	1.20	1.22 [1.11, 1.33]	0.01 [-0.09, 0.12]	990	999 [913, 1,089]
45-49	342	91.35	0.97	1.12 [0.98, 1.27]	0.15 [0.01, 0.30]	502	580 [509, 656]
50-54	0	-	-	0.79 [0.47, 1.31]	- -	-	251 [151, 417]
55-59	0	-	-	0.48 [0.15, 1.43]	- -	-	98 [30, 290]
60-64	0	-	-	0.26 [0.06, 1.14]	- -	-	33 [7, 142]
65-69	0	-	-	0.10 [0.02, 0.62]	- -	-	7 [1, 45]

Supplementary Table S8: **Sexual behaviour characteristics in RCCS participants, round 15, October 08 2011 - July 05 2013.**

	Participants reporting no ART use and who have suppressed virus	Participants reporting no ART use and who have unsuppressed virus	Participants reporting ART use and who have suppressed virus	Participants reporting ART use and who have unsuppressed virus	Sensitivity	Specificity
Round 15, August 10, 2011 - July 05, 2013; 33 communities surveyed						
Total	65	202	95	5	95.0% [88.5- 98.1]	75.7% [70.2- 80.4]
Female	44	118	66	4	94.3% [85.8- 98.2]	72.8% [65.5- 79.1]
Age						
15-24	5	22	3	1	75.0% [28.9- 96.6]	81.5% [62.8- 92.3]
25-34	19	63	18	1	94.7% [73.5-100.0]	76.8% [66.5- 84.7]
35-49	20	33	45	2	95.7% [85.0- 99.6]	62.3% [48.8- 74.1]
Male	21	84	29	1	96.7% [81.9-100.0]	80.0% [71.3- 86.6]
Age						
15-24	1	10	0	0		90.9% [60.1-100.0]
25-34	8	41	8	0	100.0% [62.8-100.0]	83.7% [70.7- 91.8]
35-49	12	33	21	1	95.5% [76.5-100.0]	73.3% [58.8- 84.2]
Round 16, July 08, 2013 - January 30, 2015; 35 communities surveyed						
Total	235	596	923	75	92.5% [90.7- 94.0]	71.7% [68.6- 74.7]
Female	171	342	663	48	93.2% [91.1- 94.9]	66.7% [62.5- 70.6]
Age						
15-24	37	87	55	10	84.6% [73.7- 91.6]	70.2% [61.6- 77.5]
25-34	72	152	239	23	91.2% [87.1- 94.1]	67.9% [61.5- 73.6]
35-49	62	103	369	15	96.1% [93.6- 97.7]	62.4% [54.8- 69.5]
Male	64	254	260	27	90.6% [86.6- 93.5]	79.9% [75.1- 83.9]
Age						
15-24	5	32	8	2	80.0% [47.9- 95.4]	86.5% [71.5- 94.6]
25-34	19	115	70	8	89.7% [80.8- 94.9]	85.8% [78.8- 90.8]
35-49	40	107	182	17	91.5% [86.7- 94.7]	72.8% [65.1- 79.4]
Round 17, February 23, 2015 - September 02, 2016; 35 communities surveyed						
Total	221	421	1269	93	93.2% [91.7- 94.4]	65.6% [61.8- 69.2]
Female	165	241	915	63	93.6% [91.8- 94.9]	59.4% [54.5- 64.0]
Age						
15-24	28	66	92	18	83.6% [75.5- 89.5]	70.2% [60.3- 78.5]
25-34	73	119	305	28	91.6% [88.1- 94.2]	62.0% [54.9- 68.6]
35-49	64	56	518	17	96.8% [94.9- 98.0]	46.7% [38.0- 55.6]
Male	56	180	354	30	92.2% [89.0- 94.5]	76.3% [70.4- 81.3]
Age						
15-24	3	24	11	2	84.6% [56.5- 96.9]	88.9% [71.1- 97.0]
25-34	19	82	96	9	91.4% [84.3- 95.6]	81.2% [72.4- 87.7]
35-49	34	74	247	19	92.9% [89.1- 95.4]	68.5% [59.2- 76.5]
Round 18, October 03, 2016 - May 22, 2018; 35 communities surveyed						
Total	141	288	1334	87	93.9% [92.5- 95.0]	67.1% [62.6- 71.4]
Female	109	153	956	53	94.7% [93.2- 96.0]	58.4% [52.3- 64.2]
Age						
15-24	20	52	75	11	87.2% [78.4- 92.9]	72.2% [60.9- 81.3]
25-34	48	68	314	27	92.1% [88.7- 94.5]	58.6% [49.5- 67.2]
35-49	41	33	567	15	97.4% [95.8- 98.5]	44.6% [33.8- 55.9]
Male	32	135	378	34	91.7% [88.7- 94.1]	80.8% [74.2- 86.1]
Age						
15-24	1	20	11	4	73.3% [47.6- 89.5]	95.2% [75.6-100.0]
25-34	15	64	90	14	86.5% [78.5- 91.9]	81.0% [70.9- 88.3]
35-49	16	51	277	16	94.5% [91.3- 96.7]	76.1% [64.6- 84.8]

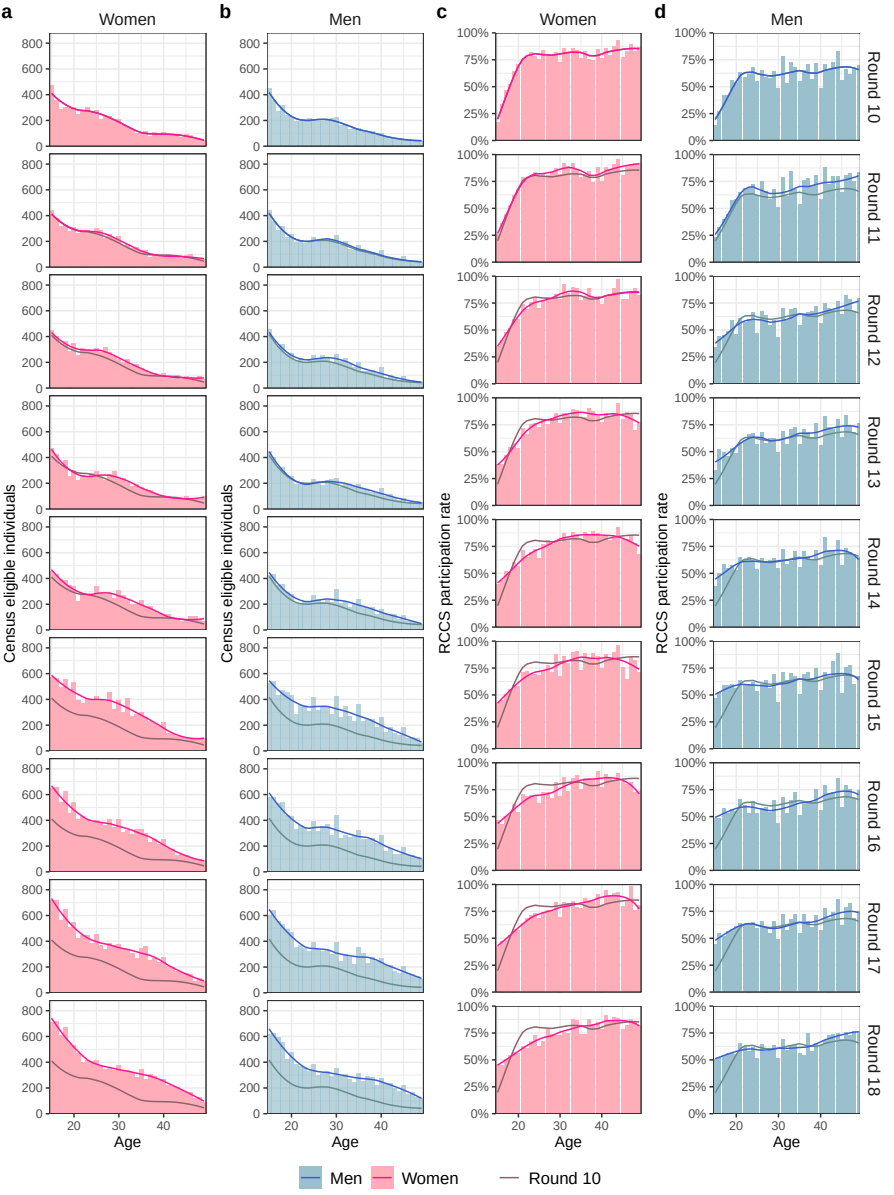
Supplementary Table S9: **Self-reported ART use and viral suppression in RCCS participants with HIV.**

Contribution from male sources to incidence			Median age of male sources			Median age of female sources			Counterfactual additional number of men suppressed			Counterfactual reduction in incidence in female		
Round 10	Round 14	Round 18	Round 10	Round 14	Round 18	Round 10	Round 14	Round 18	Closing half the suppression gap	Closing the suppression gap	95-95-95 in men	Closing half the suppression gap	Closing the suppression gap	95-95-95 in men
Central analysis														
57.9% [56.2-59.6]	61.4% [59.8-63.1]	62.8% [60.2-65.2]	28.5 [22.8-40.2]	30.1 [22.6-41.0]	33.5 [23.6-41.6]	25.0 [18.0-36.2]	26.8 [19.7-37.2]	26.0 [19.0-36.4]	75.1 [53.9-96.4]	150.2 [107.8-192.8]	172.6 [136.8-210.3]	25.1% [24.2-26.2]	50.6% [48.6-52.8]	58.4% [54.9-61.7]
Sensitivity analyses														
Using incidence rates estimated with LOESS regression														
61.5% [59.5-63.5]	57.5% [55.5-59.5]	62.1% [60.4-63.9]	27.7 [22.3-38.8]	31.8 [23.1-42.8]	34.0 [23.8-41.6]	24.0 [18.0-34.8]	25.0 [18.9-36.0]	26.0 [18.3-35.9]	75.1 [53.9-96.4]	150.2 [107.8-192.8]	172.6 [136.8-210.3]	25.3% [24.1-26.6]	50.9% [48.5-53.6]	58.1% [53.6-62.1]
Using incidence rates estimated on a data subset to 28 continuously surveyed communities														
58.2% [56.5-59.8]	62.3% [60.6-64.0]	64.3% [61.6-66.9]	29.5 [23.0-41.1]	31.0 [23.0-42.1]	34.0 [24.2-43.0]	25.0 [18.0-36.4]	27.7 [19.4-37.2]	27.0 [19.0-37.0]	75.1 [53.9-96.4]	150.2 [107.8-192.8]	172.6 [136.8-210.3]	25.5% [24.3-26.9]	51.5% [48.9-54.3]	56.4% [52.0-60.6]
Using non-refined infection time estimates														
57.9% [56.1-59.6]	61.4% [59.8-63.0]	62.8% [60.3-65.2]	28.1 [22.6-40.2]	30.0 [22.3-41.3]	33.4 [23.7-42.0]	24.5 [18.0-36.3]	26.0 [19.6-37.4]	25.8 [19.0-36.5]	75.1 [53.9-96.4]	150.2 [107.8-192.8]	172.6 [136.8-210.3]	25.1% [24.0-26.3]	50.6% [48.3-53.0]	58.3% [54.6-62.0]
Without source-recipients pairs for which the source or recipient was sequenced after round 17														
58.0% [56.2-59.7]	61.4% [59.8-63.1]	62.8% [60.2-65.2]	28.0 [22.7-40.0]	29.8 [22.4-40.9]	32.9 [23.0-41.5]	25.2 [18.1-36.7]	27.0 [20.0-37.6]	26.0 [19.7-36.7]	75.1 [53.9-96.4]	150.2 [107.8-192.8]	172.6 [136.8-210.3]	24.8% [23.7-26.0]	50.0% [47.7-52.4]	59.4% [55.6-63.1]
Without source-recipients pairs for which the source or recipient was sequenced after round 16														
58.0% [56.3-59.7]	61.4% [59.8-63.0]	62.7% [60.2-65.2]	28.0 [22.8-39.9]	30.0 [22.6-40.8]	33.0 [23.9-41.1]	24.7 [18.0-36.5]	26.0 [19.8-37.6]	25.0 [19.0-36.2]	75.1 [53.9-96.4]	150.2 [107.8-192.8]	172.6 [136.8-210.3]	24.9% [23.6-26.2]	50.1% [47.5-52.8]	59.0% [55.0-63.1]
Without source-recipients pairs for which the source or recipient was sequenced after round 15														
58.0% [56.2-59.7]	61.4% [59.7-63.0]	62.8% [60.3-65.2]	28.1 [22.8-39.6]	30.0 [22.5-40.7]	33.4 [24.0-41.2]	25.0 [18.0-37.0]	26.9 [19.6-38.0]	25.6 [19.0-36.8]	75.1 [53.9-96.4]	150.2 [107.8-192.8]	172.6 [136.8-210.3]	24.9% [23.6-26.3]	50.2% [47.5-52.9]	58.7% [54.5-63.0]
Using a bootstrap sample of the source-recipient pairs (first draw)														
58.0% [56.2-59.7]	61.4% [59.8-63.0]	62.8% [60.2-65.2]	29.2 [23.0-40.3]	31.0 [23.0-41.1]	34.0 [24.1-42.0]	25.4 [18.8-36.1]	27.1 [19.8-37.0]	29.0 [20.0-36.0]	75.1 [53.9-96.4]	150.2 [107.8-192.8]	172.6 [136.8-210.3]	25.4% [24.5-26.4]	51.2% [49.2-53.2]	57.0% [53.9-60.2]
Using a bootstrap sample of the source-recipient pairs (second draw)														
57.8% [56.1-59.5]	61.4% [59.7-63.0]	62.7% [60.3-65.2]	29.5 [23.0-40.0]	31.0 [23.0-40.6]	31.4 [23.0-41.1]	24.0 [18.0-36.8]	26.0 [19.4-37.8]	25.0 [19.0-35.3]	75.1 [53.9-96.4]	150.2 [107.8-192.8]	172.6 [136.8-210.3]	24.4% [23.4-25.3]	49.1% [47.1-51.0]	60.8% [57.5-64.1]
Using a bootstrap sample of the source-recipient pairs (third draw)														
57.9% [56.1-59.5]	61.4% [59.7-63.1]	62.7% [60.3-65.2]	29.0 [22.7-40.5]	31.0 [22.4-41.3]	34.0 [23.0-42.0]	24.6 [18.0-36.9]	26.4 [19.4-37.9]	25.0 [19.0-36.7]	75.1 [53.9-96.4]	150.2 [107.8-192.8]	172.6 [136.8-210.3]	25.5% [24.4-26.7]	51.4% [49.2-53.8]	57.3% [53.4-61.0]

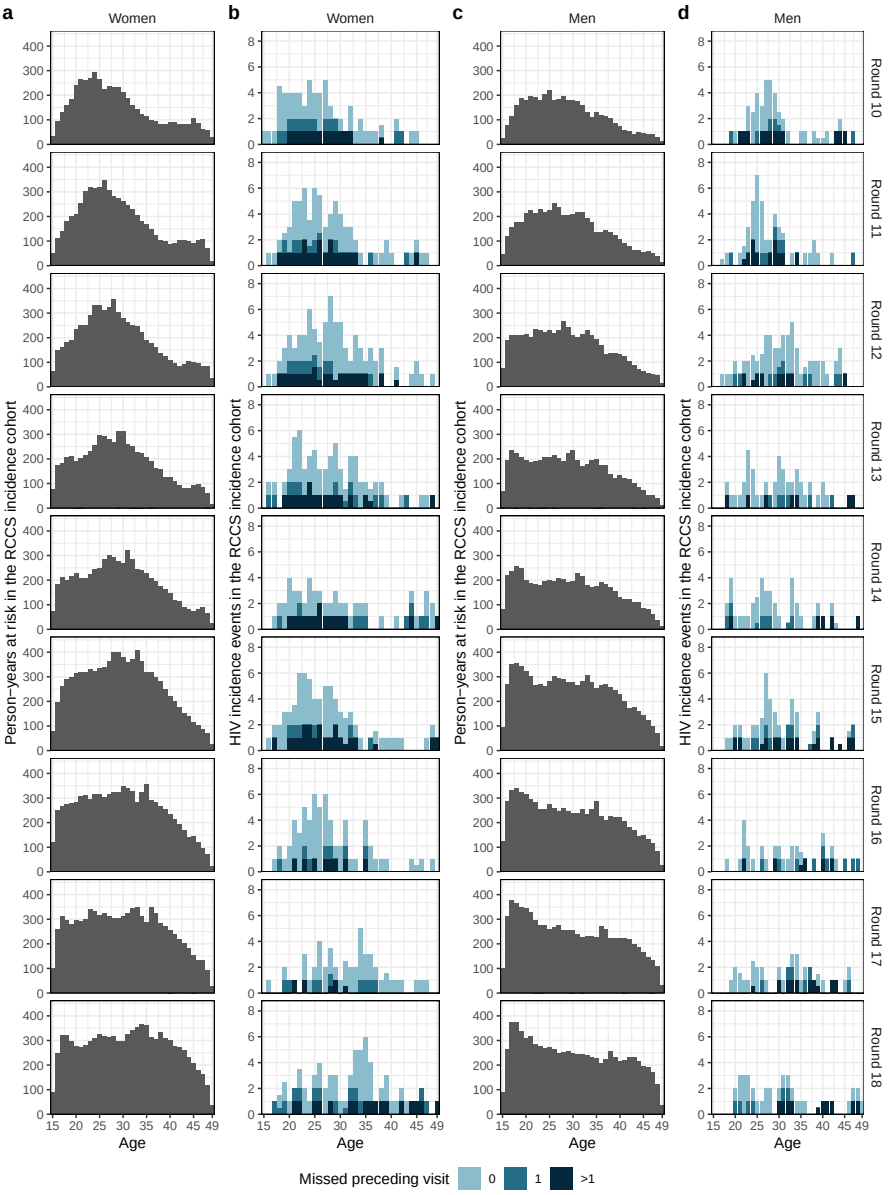
Supplementary Table S10: **Sensitivity analyses.**

Contribution from male sources to incidence			Median age of male sources			Median age of female sources			Counterfactual additional number of men suppressed			Counterfactual reduction in incidence in female		
Round 10	Round 14	Round 18	Round 10	Round 14	Round 18	Round 10	Round 14	Round 18	Closing half the suppression gap	Closing the suppression gap	95-95-95 in men	Closing half the suppression gap	Closing the suppression gap	95-95-95 in men
Assuming an alternative form of the transmission rate (12a))														
60.3%	60.7%	64.3%	29.0	31.7	35.0	24.6	26.4	26.0	75.1	150.2	172.6	25.7%	52.0%	55.3%
[59.3-61.3]	[59.8-61.6]	[63.0-65.9]	[23.0-44.9]	[23.0-48.6]	[24.1-43.4]	[17.9-36.3]	[19.3-37.4]	[19.0-36.7]	[53.9-96.4]	[107.8-192.8]	[136.8-210.3]	[24.8-26.6]	[50.2-53.7]	[52.3-59.5]
Assuming an alternative form of the transmission rate (12b))														
60.5%	60.6%	63.8%	30.0	32.0	33.0	24.5	26.3	25.2	75.1	150.2	172.6	24.9%	50.3%	57.0%
[59.5-61.4]	[59.7-61.5]	[62.5-65.1]	[23.0-42.5]	[23.0-46.4]	[23.0-46.5]	[17.8-36.4]	[19.3-37.5]	[18.8-36.7]	[53.9-96.4]	[107.8-192.8]	[136.8-210.3]	[24.2-25.6]	[48.8-51.8]	[54.0-60.5]
Assuming an alternative form of the transmission rate (12c))														
60.4%	60.6%	64.0%	29.7	32.0	33.5	24.6	26.3	26.0	75.1	150.2	172.6	25.0%	50.6%	56.9%
[59.4-61.3]	[59.7-61.5]	[62.7-65.4]	[23.0-43.8]	[23.0-48.0]	[23.3-44.8]	[17.9-36.4]	[19.4-37.5]	[19.0-36.8]	[53.9-96.4]	[107.8-192.8]	[136.8-210.3]	[24.1-26.1]	[48.7-52.7]	[53.4-60.5]
Assuming an alternative form of the transmission rate (12d))														
57.9%	61.4%	62.7%	29.0	30.5	32.9	25.0	26.5	25.8	75.1	150.2	172.6	24.9%	50.2%	59.1%
[56.2-59.6]	[59.8-63.0]	[60.2-65.2]	[23.0-40.1]	[22.9-40.9]	[23.0-42.0]	[18.0-36.1]	[19.7-37.1]	[19.0-36.5]	[53.9-96.4]	[107.8-192.8]	[136.8-210.3]	[24.1-25.7]	[48.6-51.9]	[56.2-61.9]
Assuming an alternative form of the transmission rate (12e))														
58.0%	61.4%	62.8%	29.0	30.5	33.0	25.0	26.7	25.9	75.1	150.2	172.6	24.9%	50.2%	58.9%
[56.3-59.7]	[59.7-63.0]	[60.2-65.2]	[23.0-40.2]	[22.9-41.0]	[23.0-42.0]	[18.0-36.2]	[19.8-37.2]	[19.0-36.6]	[53.9-96.4]	[107.8-192.8]	[136.8-210.3]	[24.1-25.7]	[48.6-51.9]	[56.0-61.8]
Assuming an alternative form of the transmission rate (12f))														
57.9%	61.4%	62.8%	28.9	30.4	33.0	25.0	26.6	26.0	75.1	150.2	172.6	25.0%	50.3%	58.9%
[56.2-59.6]	[59.8-63.1]	[60.3-65.2]	[23.0-40.1]	[22.8-40.9]	[23.0-41.9]	[18.0-36.2]	[19.7-37.2]	[19.0-36.6]	[53.9-96.4]	[107.8-192.8]	[136.8-210.3]	[24.1-25.9]	[48.4-52.3]	[55.7-62.1]
Assuming the same proportion of viral suppression among non-participants as among participants of the same age, gender, and survey round														
57.9%	61.4%	62.8%	28.6	30.0	33.0	25.0	26.7	25.9	71.7	143.3	143.5	26.7%	53.6%	52.2%
[56.2-59.6]	[59.7-63.0]	[60.2-65.2]	[22.8-40.2]	[22.6-40.9]	[23.6-41.5]	[18.0-36.3]	[19.6-37.2]	[19.0-36.5]	[54.6-89.5]	[109.3-179.0]	[114.1-175.7]	[26.1-27.2]	[52.4-54.7]	[47.0-57.1]
Assuming that non-participants are not suppressed														
58.0%	61.4%	62.7%	28.5	30.1	34.0	25.0	26.8	26.3	254.7	329.9	351.9	52.3%	68.1%	74.6%
[56.3-59.7]	[59.8-63.0]	[60.2-65.2]	[22.8-40.1]	[22.5-41.1]	[24.0-42.0]	[18.0-36.2]	[19.8-37.2]	[19.0-37.9]	[232.7-275.5]	[300.0-358.6]	[333.4-372.5]	[50.0-54.6]	[65.8-70.5]	[73.4-75.8]
Assuming that prevalence in non-participants is 25% higher than in participants														
58.0%	61.5%	62.7%	28.6	30.1	33.5	25.0	26.8	26.0	81.9	163.9	189.3	25.2%	50.7%	58.3%
[56.3-59.7]	[59.8-63.1]	[60.2-65.1]	[22.8-40.2]	[22.6-41.0]	[23.7-41.6]	[18.0-36.3]	[19.7-37.2]	[19.0-36.4]	[58.9-105.1]	[117.7-210.1]	[150.4-230.4]	[24.2-26.2]	[48.6-52.9]	[54.8-61.7]
Assuming that prevalence in men non-participants is 25% higher than in men participants														
58.2%	61.6%	62.9%	28.6	30.1	33.5	25.0	26.7	26.0	81.9	163.9	189.3	25.1%	50.6%	58.4%
[56.5-59.8]	[60.0-63.3]	[60.4-65.4]	[22.8-40.2]	[22.6-41.0]	[23.6-41.6]	[18.0-36.2]	[19.6-37.2]	[19.0-36.4]	[58.9-105.1]	[117.7-210.1]	[150.4-230.4]	[24.1-26.2]	[48.5-52.9]	[54.9-61.8]
Assuming that prevalence in women non-participants is 25% higher than in women participants														
57.8%	61.2%	62.5%	28.5	30.0	33.4	25.0	26.8	26.0	75.1	150.2	172.6	25.1%	50.6%	58.4%
[56.0-59.4]	[59.6-62.8]	[60.0-64.9]	[22.8-40.2]	[22.6-41.0]	[23.6-41.6]	[18.0-36.3]	[19.7-37.2]	[19.0-36.4]	[53.9-96.4]	[107.8-192.8]	[136.8-210.3]	[24.1-26.2]	[48.5-52.8]	[55.0-61.7]
Defining viral suppression as a viral load measurement below 200 copies/mL plasma blood														
57.9%	61.4%	62.8%	28.6	30.2	33.1	25.0	27.0	26.0	73.2	146.4	197.2	22.8%	46.0%	61.6%
[56.2-59.6]	[59.7-63.0]	[60.3-65.2]	[22.9-40.2]	[22.6-41.0]	[23.6-41.5]	[18.0-36.3]	[19.7-37.3]	[19.1-36.6]	[51.7-94.5]	[103.5-189.1]	[161.7-234.6]	[21.7-24.0]	[43.7-48.5]	[58.1-64.9]
Without adjustments for potentially unequal sampling of sources														
57.9%	61.4%	62.8%	29.0	30.6	33.7	25.0	26.3	26.0	75.1	150.2	172.6	25.2%	50.6%	58.4%
[56.1-59.6]	[59.8-63.1]	[60.2-65.2]	[22.7-40.0]	[22.5-40.8]	[23.5-41.4]	[18.0-36.1]	[19.7-37.2]	[19.0-36.2]	[53.9-96.4]	[107.8-192.8]	[136.8-210.3]	[24.2-26.2]	[48.7-52.7]	[55.1-61.6]

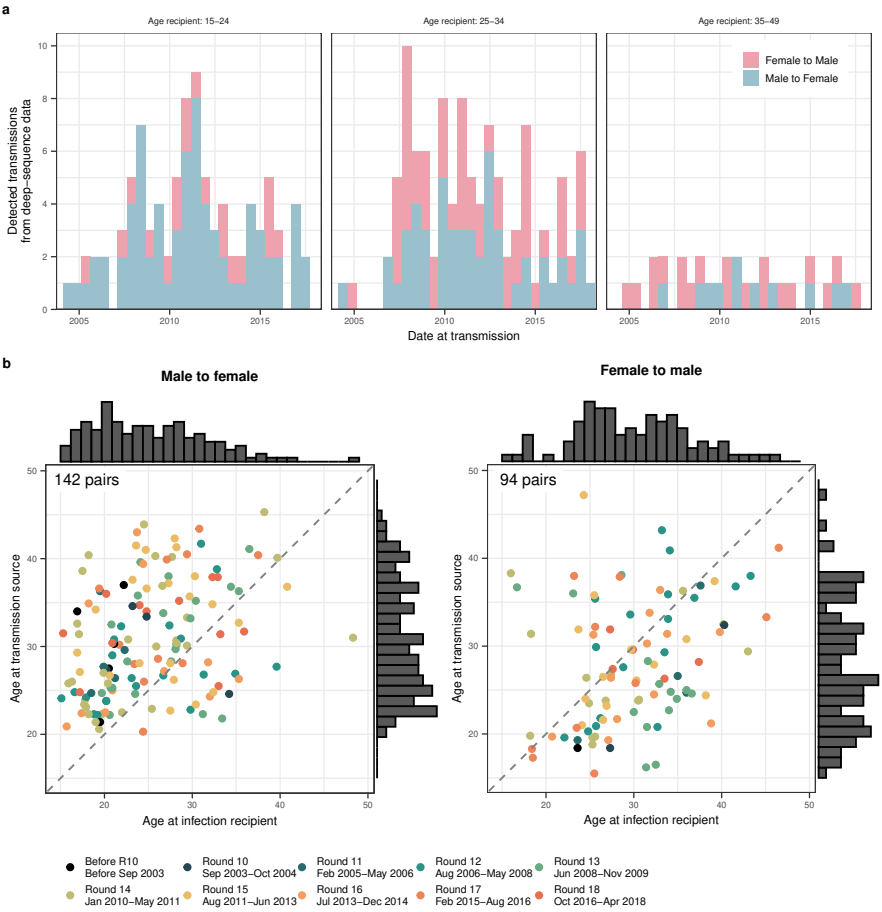
Supplementary Table S10: **Sensitivity analyses (continued).**



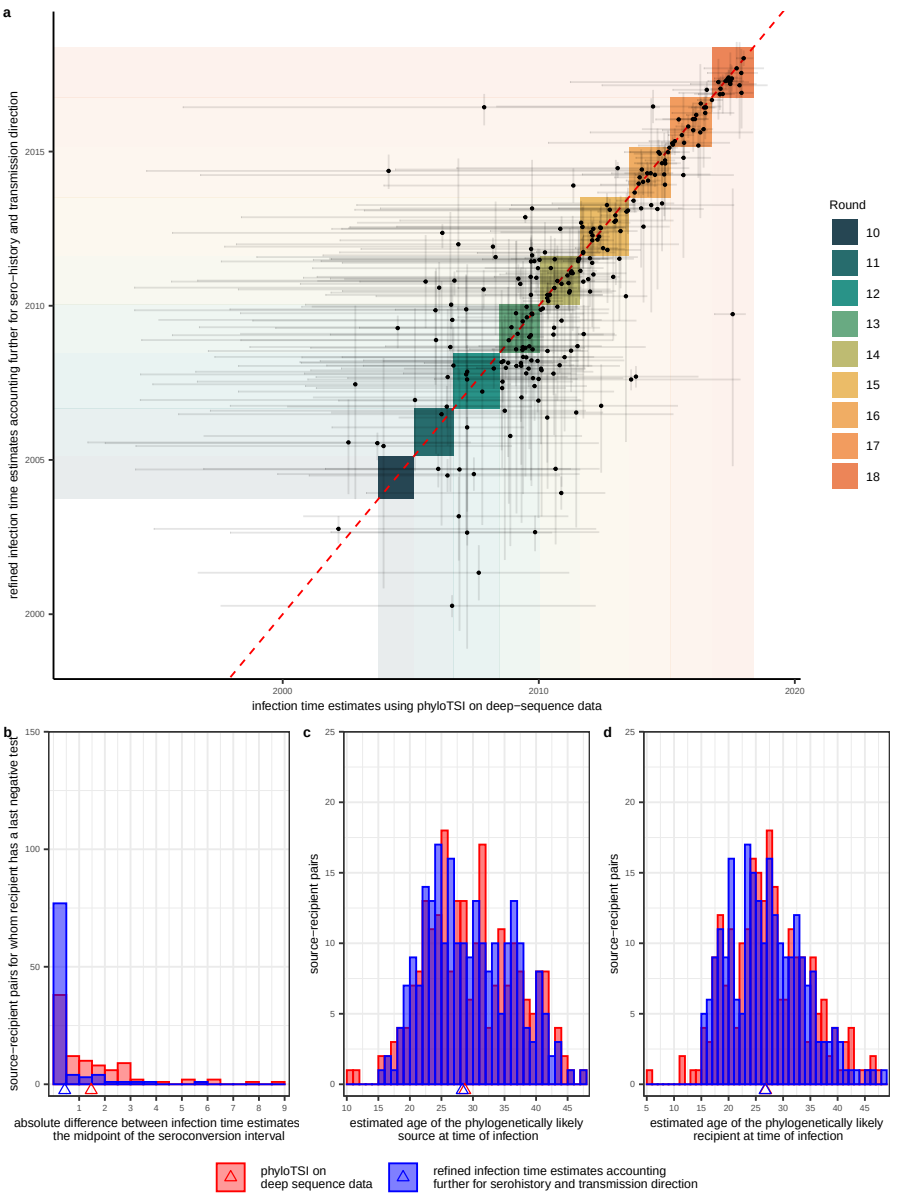
Extended Data Fig. 1: **Characteristics of the RCCS study population by age, gender, and time.** (**a**-**b**) Population size. Counts of the aggregated individual-level census data by 1-year age group (bars) are shown along LOESS smoothed population size estimates (line) for men and women (see text). (**c**-**d**) RCCS participation rates. Rates relative to the aggregated census data by 1-year age band (bars) are shown along LOESS smoothed participation rates (line) for both men and women. For reference, round 10 values are indicated in each subsequent plot in darker colors. The timeline of the survey rounds is shown in Figure 1b.



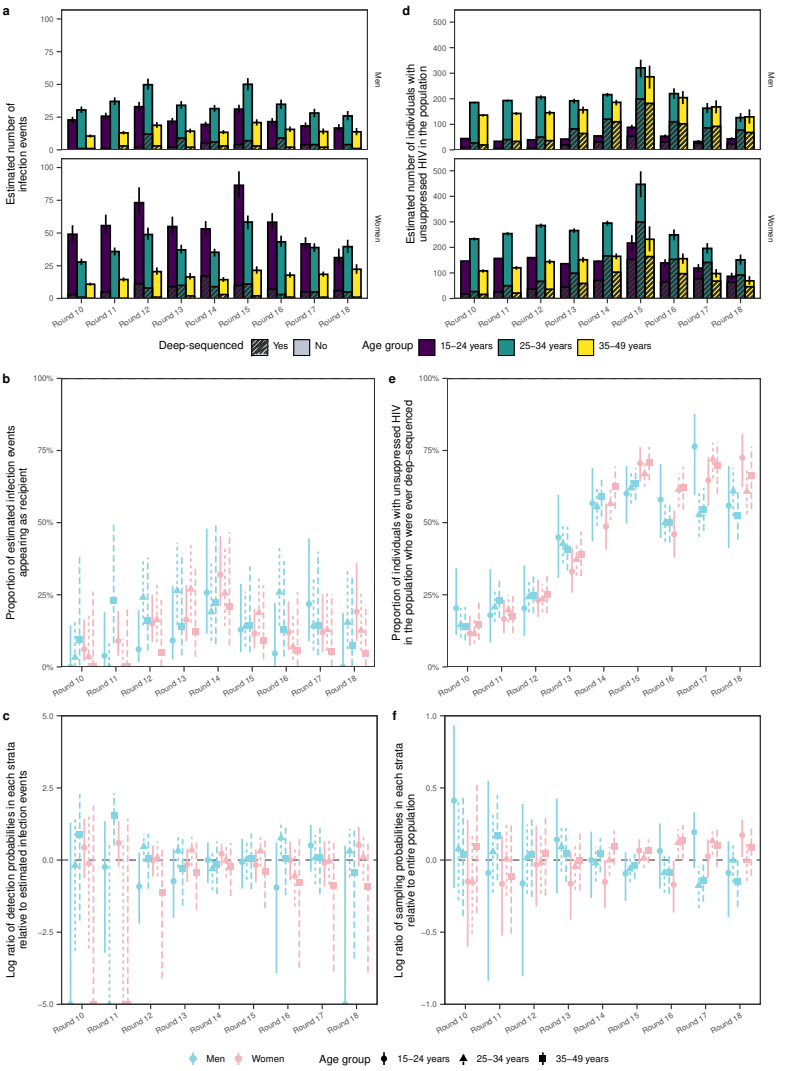
Extended Data Fig. 2: **Age- and gender-specific person-years at risk and HIV incidence events in the RCCS incidence cohort.** Person-years at risk in the RCCS incidence cohort among (**a**) women and (**c**) men. HIV incidence events in the RCCS incidence cohort among (**b**) women and (**d**) men.



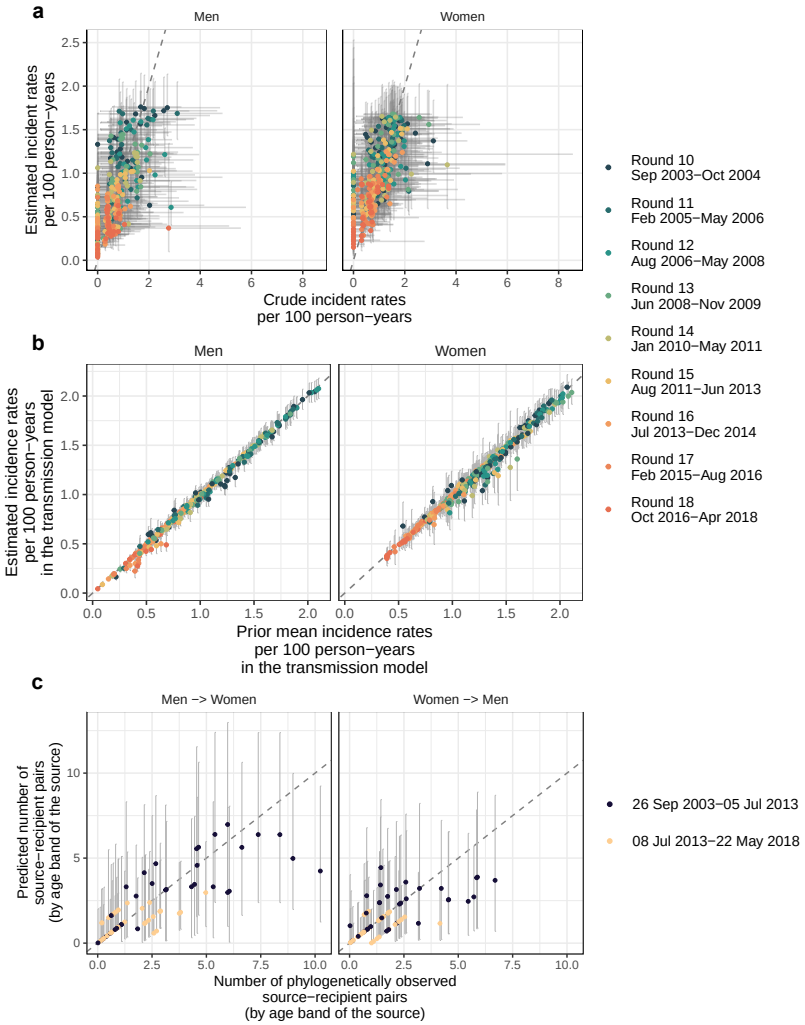
Extended Data Fig. 3: **Phylogenetically reconstructed source-recipient pairs.** (**a**) Number of heterosexual source-recipient pairs by the date of infection of the recipient (x-axis), the age of the recipient at infection (panel), and transmission direction (color). (**b**) Heterosexual source-recipient pairs by the age of the recipient (x-axis) and the age of the source (y-axis) at the median infection time estimate by the round (color) in which transmission was estimated to have occurred. The number of phylogenetically reconstructed source-recipient pairs is indicated in the top-left corner.



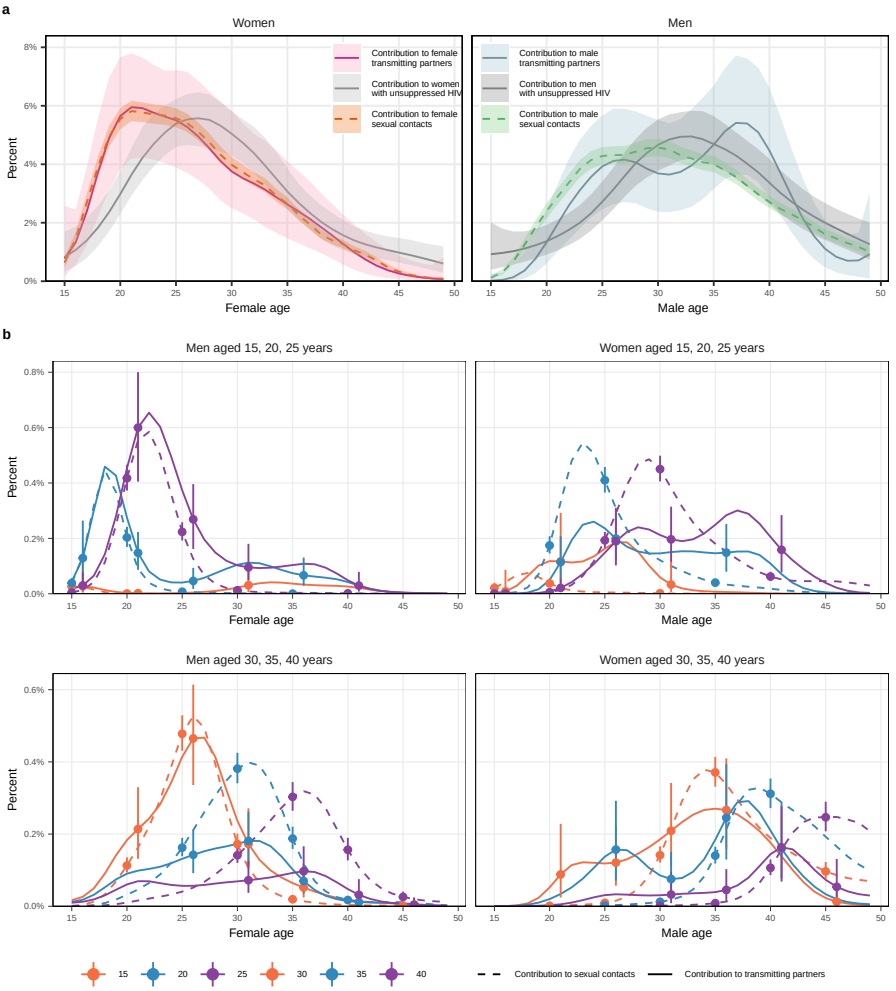
Extended Data Fig. 4: **Comparison of estimated infection dates in phylogenetically reconstructed source-recipient pairs.** (**a**) Estimated infection times of the recipient in the $n = 227$ phylogenetically reconstructed source-recipient pairs from `phyloTSI` based on deep-sequence data alone (x-axis) against refined estimates accounting for serohistory and inferred direction of transmission (y-axis). Median estimates (dots) are shown along 95% uncertainty ranges (lines). (**b**) Histogram of absolute difference (bars) and mean absolute differences (triangle) between infection time estimates and the midpoint of seroconversion intervals in 98 source-recipient pairs in which the recipient had a last negative test, across the two methods (color). (**c**) Histogram (bars) and median (triangle) age of the phylogenetically likely recipient, across the two methods (colors). (**d**) Histogram (bars) and median (triangle) age of the phylogenetically likely source, across the two methods (colors).



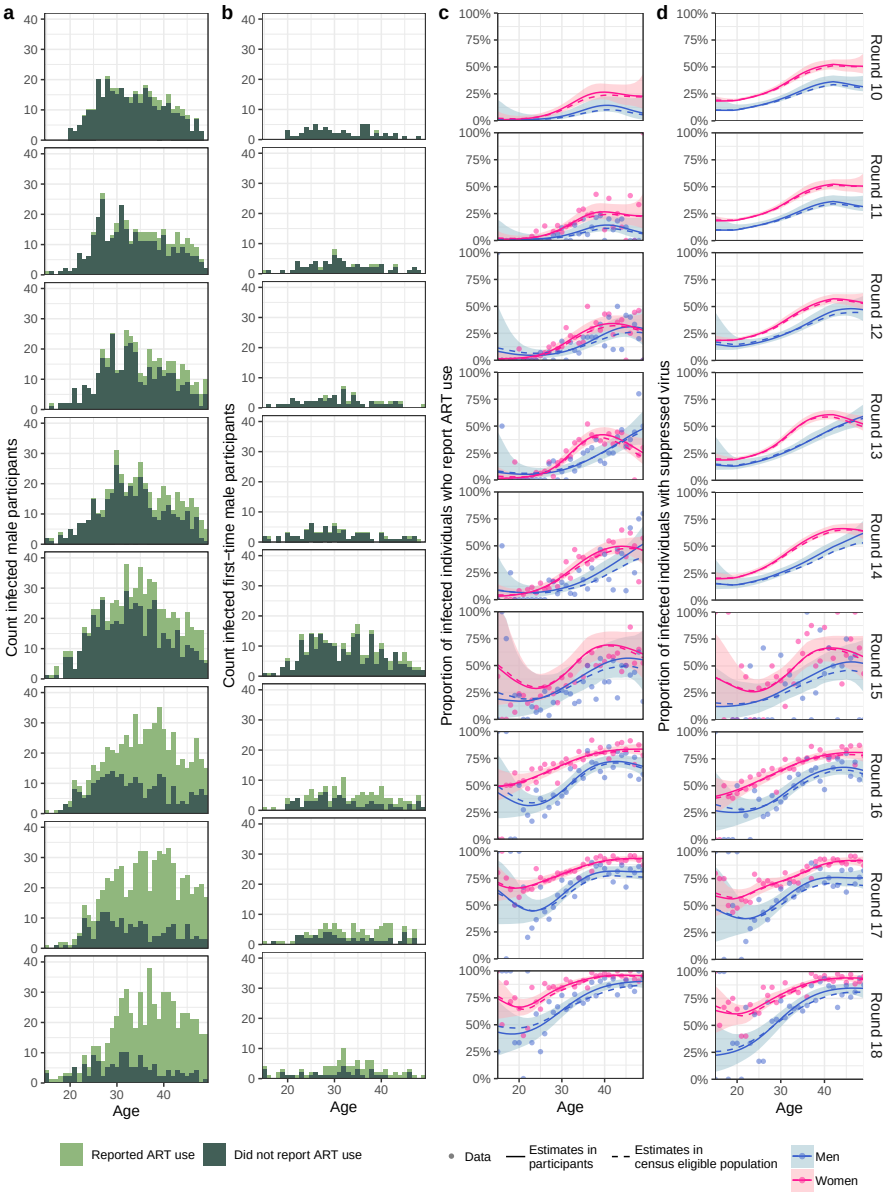
Extended Data Fig. 5: **Sampling estimates of transmission events and sources of infections.** (**a**-**c**) The sampling cascade of transmission events was modeled by comparing the number of phylogenetically reconstructed source-recipient events to the estimated number of infection events under the incidence model, by gender, age band and survey round of infected individuals. Throughout, shown are the number of sampled and unsampled transmission events, the estimated proportion of transmission events that were ever deep-sequenced, and log ratios of estimated proportion of transmission events that were ever deep-sequenced in any strata relative to the overall average across strata (point estimates: dots, 95% confidence intervals: linebars). Estimates are based on $n = 227$ source-recipient pairs and $n = 1{,}117$ observed incidence events. (**d**-**f**) Additional differences in source sampling were modelled by considering unsuppressed individuals as potential sources of infection, and calculating the number of unsuppressed individuals in a round that were ever deep-sequenced. Throughout, shown are the number of sampled and unsampled possible transmission sources, the estimated proportion of possible sources that were ever deep-sequenced, and log ratios of estimated proportion of possible sources that were ever deep-sequenced in any strata relative to the overall average across strata (point estimates:



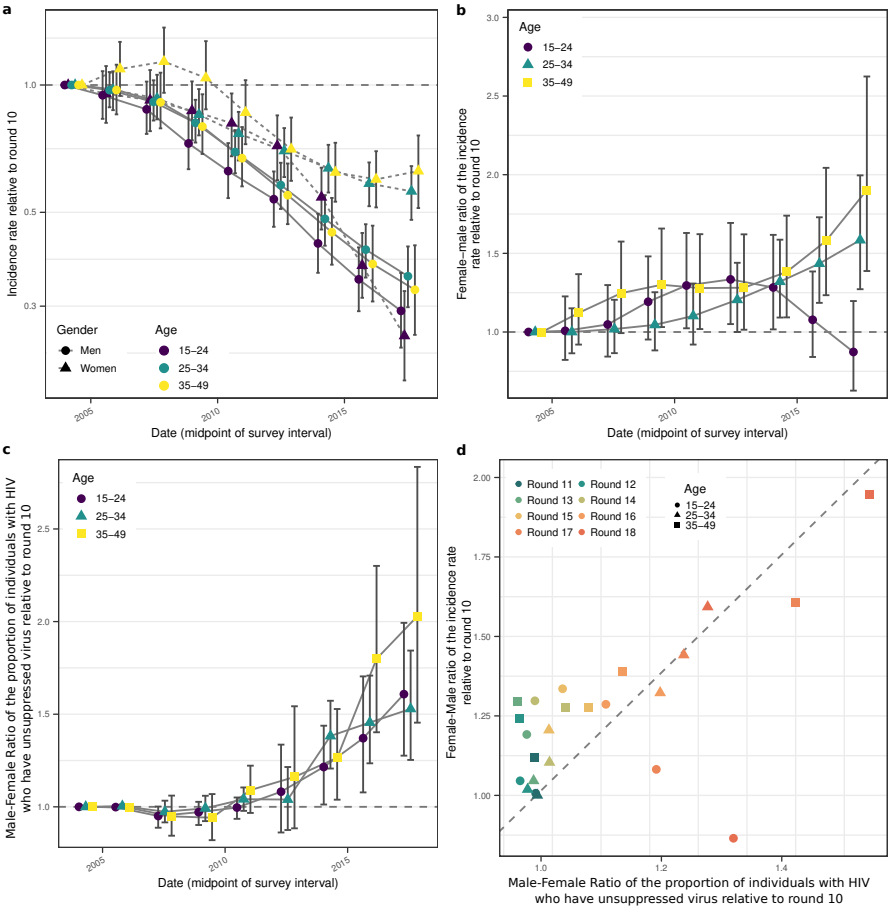
Extended Data Fig. 6: **Validation of the incidence rate and transmission flow models.** (**a**) Empirical HIV incidence rates were obtained for each of the 50 data sets with imputed exposure times and compared to the estimated HIV incidence rates under the Poisson model. The median (point) and 95% range (horizontal error bars) of the crude HIV incidence rates are plotted against the posterior median (point) and 95% range (vertical error bars) of estimated HIV incidence rates for each gender, age and round. (**b**) Prior incidence rates as specified according to the outputs of the incidence rate and used in the transmission model are compared versus the posterior incidence rates obtained with the transmission model. Shown are medians (point) and 95% credible intervals (error bars) by gender, age and round. (**c**) Observed transmission flow counts are compared to posterior predictive estimates under the transmission model. Shown are medians (point) and 95% credible intervals (error bars) by direction of transmission, time period, and age of the phylogenetically likely source. Throughout all subfigures, empirical and modelled incidence estimates are based on n =1,117 individuals in the incidence cohort and n = 227 source-recipient pairs among n =1,978 individuals in the transmission cohort.



Extended Data Fig. 7: **Age contributions to sexual contacts, viral suppression and transmission.** (**a**) Estimated age contributions from women to men of all ages (left) and from men to women of all ages (right) to sexual contacts in round 15, viral suppression in round 18, and transmission in round 18 (posterior median: line, 95% credible interval: ribbon). Age contributions sum to 100% separately for women and men. (**b**) Estimated age contributions from women to men of specific ages (left) and from men to women of specific ages (right) to sexual contacts in round 15, and transmission in round 18 (posterior median: line, 95% credible interval: errorbars). Age contributions sum to 100% for women and men combined. Throughout all subfigures, empirical and modelled incidence estimates are based on n =1,117 individuals in the incidence cohort and $n = 227$ source-recipient pairs among n =1,978 individuals in the transmission cohort.



Extended Data Fig. 8: **ART use and virus suppression in the RCCS study population by age, gender, and time.** (**a**) HIV-positive male participants, by whether they reported ART use (color), by 1-year age band (x-axis) and survey round (rows). (**b**) HIV-positive male first-time participants, by whether they reported ART use (color), by 1-year age band (x-axis) and survey round (panel). (**c**) Estimates of ART use in men (blue) and women (pink) in the study population by 1-year of age. Data from participants (dots) are shown along smoothed posterior median estimates (solid line) and 95% credible intervals (ribbon) in participants, and along posterior median estimates in the census-eligible population (dashed line), using data from first-time participants as proxy of ART use in non-participants (see text). (**d**) Estimates of virus suppression, defined as a viral load measurement below 1,000 copies of HIV per milliliter plasma blood, in men (blue) and women (pink) in the study population by 1-year of age. Data from participants (dots) are shown along smoothed posterior median estimates (solid



Extended Data Fig. 9: **Longitudinal changes in viral suppression and incidence rates in the RCCS study population since 2003.** (**a**) Changes in incidence rates relative to round 10, i.e. Sep 2003 to Oct 2004 (posterior median: dots, 95% confidence interval: errorbars). (**b**) Female-to-male ratio in changes in incidence rates relative to round 10 (posterior median: dots, 95% credible interval: errorbars). (**c**) Male-to-female ratio in changes in the proportion of individuals with HIV who have unsuppressed virus relative to round 10 (posterior median: dots, 95% credible interval: errorbars). (**d**) Scatter plot between the female-to-male ratio in changes in incidence rates as shown in (b) and the male-to-female ratio in changes in the proportion of individuals with HIV who have unsuppressed virus relative to round 10 as shown in (c). Throughout all subfigures, estimates are based on n =1,117 individuals in the incidence cohort and $n = 3,265$ participants with HIV and measured viral load.