Deep-sequence phylogenetics to quantify patterns of HIV transmission in the context of a universal testing and treatment trial – BCPP/ Ya Tsie trial

Lerato E. Magosi[1*], Yinfeng Zhang[2], Tanya Golubchick[3], Victor De Gruttola[4], Eric Tchetgen Tchetgen[5], Vlad Novitsky[6,10], Janet Moore[7], Pam Bachanas[7], Tebogo Segolodi[8], Refeletswe Lebelonyane[9], Molly Pretorius Holme[6], Sikhulile Moyo[10], Joseph Makhema[10], Shahin Lockman[6,10,11], Christophe Fraser[3], Max Essex[6,10] & Marc Lipsitch[1*] on behalf of the Botswana Combination Prevention Project and the PANGEA consortium[#]

1. Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard T.H. Chan School of Public Health, Harvard University, Boston, USA

2. Division of Molecular & Genomic Pathology, University of Pittsburgh Medical Center Presbyterian Shadyside, Pittsburgh, USA

3. Oxford Big Data Institute, Li Ka Shing Center for Health Information and Discovery, Nuffield Department of Medicine, Old Road Campus, University of Oxford, Oxford, UK

4. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, Boston, USA

5. Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, USA

6. Harvard T.H. Chan School of Public Health AIDS Initiative, Department of Immunology and Infectious Disease, Harvard T.H. Chan School of Public Health, Harvard University, Boston, USA

7. Division of Global HIV/AIDS and TB, Centers for Disease Control and Prevention, Atlanta, USA

8. HIV Prevention Research Unit, Centers for Disease Control and Prevention, Gaborone, Botswana

9. Ministry of Health, Republic of Botswana, Gaborone, Botswana

10. Botswana Harvard AIDS Institute Partnership, Gaborone, Botswana

11. Brigham and Women's Hospital, Division of Infectious Diseases, Boston, USA

Membership of the Botswana Combination Prevention Project and the PANGEA consortium is provided in Supplementary Note 7

*Corresponding author. Email: lmagosi@hsph.harvard.edu (Lerato E. Magosi, DPhil); mlipsitc@hsph.harvard.edu (Marc Lipsitch, DPhil)

Abstract

Mathematical models predict that community-wide access to HIV testing-and-treatment can rapidly and substantially reduce new HIV infections. Yet several large universal test-and-treat HIV prevention trials in high-prevalence epidemics demonstrated variable reduction in population-level incidence. To elucidate patterns of HIV spread in universal test-and-treat trials we quantified the contribution of geographic-location, gender, age and randomized-HIV-intervention to HIV transmissions in the 30-community Ya Tsie trial in Botswana (estimated trial population: 175,664). Deep-sequence phylogenetic analysis revealed that most inferred HIV transmissions within the trial occurred within the same or between neighboring communities, and between similarly-aged partners. Transmissions into intervention communities from control communities were more common than the reverse post-baseline (30% [12.2 – 56.7] versus 3% [0.1 – 27.3]) than at baseline (7% [1.5 – 25.3] versus 5% [0.9 – 22.9]) compatible with a benefit from treatment-as-prevention. Our findings suggest that population mobility patterns are fundamental to HIV transmission dynamics and to the impact of HIV control strategies.

Introduction

The global number of new infections with HIV-1, the virus that causes AIDS, has gradually declined since the peak in 1997, yet population-level HIV incidence remains high in East and Southern Africa, especially among young women [1]. Despite efforts spanning four decades there is still no successful HIV vaccine or widely administrable cure. This emphasizes the importance of curbing new infections in order to bring the HIV epidemic under control.

Four large cluster-randomized trials were recently conducted in East (Kenya and Uganda) and Southern Africa (Botswana, South Africa and Zambia) to evaluate the effect of universal HIV testing and treatment in reducing population-level HIV incidence [2-5]. The trials showed variable outcomes, ranging from modest to no reduction in the occurrence of new HIV infections [6]. The trials were motivated by evidence that early initiation of antiretroviral therapy substantially reduced onward transmission in people with HIV owing to suppressed virus levels [7] and by mathematical models predicting that widely expanded access to HIV testing and treatment would rapidly and substantially reduce occurrence of new infections [8].

The joint United Nations Programme on HIV/AIDS (UNAIDS) had set a global target of fewer than 500,000 new HIV infections by 2020 that was unmet. To stem the spread of new HIV infections we need to better understand patterns of HIV transmission in high-prevalence, generalized epidemics in sub-Saharan Africa. To obtain such understanding requires a better grasp of where and from whom those most at risk acquire infection and to whom they are likely to spread infection. Previous work in a high HIV burden setting in Kwa-Zulu Natal, South Africa

78 identified age-disparate sexual partnerships between younger women under 25 years and older
79 men as a primary contributor to HIV transmission and incidence in younger women [9].
80 We hypothesized that geographic proximity, gender, age and ready access to HIV care
81 (randomized-HIV-intervention) contribute to shaping patterns of HIV transmission. To test
82 these hypotheses, we performed a deep-sequence phylogenetic study of adults with HIV aged
83 16-64 years who consented to participate in the Botswana Ya Tsie HIV prevention trial to
84 quantify the contribution of the above factors to HIV transmission patterns in Botswana
85 adjusting for sampling heterogeneity.
86 Botswana is a sparsely populated, landlocked country in southern Africa, roughly the
87 size of France, 581,730 km^2 (224,610 square miles), with a population of about 2.3 million
88 people. The Kalahari Desert occupies approximately 70% of the country, and the population is
89 largely distributed along an eastern corridor bordered by South Africa, Zambia and Zimbabwe
90 [10]. It is a high middle-income country with a Gini index of 53.3 [11]. One in five adults in
91 Botswana aged 15 – 49 years were reported to be living with HIV in 2019, representing the third
92 highest HIV prevalence in the world after Lesotho and Eswatini [12].
93 Our results produced three key findings: First, most HIV transmissions inferred in the
94 trial population occurred between partners residing in the same community or neighboring
95 communities, and between partners of similar ages. Second, transmission events identified
96 among trial participants sampled one or more years post-baseline were consistent with a
97 greater flow of HIV transmissions into intervention communities from control communities
98 than vice versa during the trial, although dates of transmission were not identified. Third, men
99 and women sampled in the Botswana Ya Tsie trial whose sexual partners could be inferred

100 phylogenetically contributed similarly to the spread of HIV infection. Overall, our findings
101 suggest that population mobility patterns are central to understanding HIV transmission
102 dynamics and should be considered when designing and evaluating HIV control strategies.
103

Methods

Overview of trial population. The Ya Tsie trial or Botswana Combination Prevention Project (BCPP) was a pair-matched community randomized trial to evaluate whether an optimized combination of effective HIV treatment and prevention interventions would substantially reduce population-level HIV incidence over 29 months. The trial was conducted from 2013-2018 in 30 rural and peri-urban communities (average population size: 5,855) in the Central, South-East and North/North-East regions of Botswana, that represented an estimated total population of 175,664 (7.6% of the national population) (Fig 1). Among eligible community residents 13,131 adults aged 16-64 years gave informed consent to participate in the trial. See [3, 13] for detailed eligibility criteria. The estimated HIV prevalence among participating communities at the start of the trial was 25.7% (27,446 / 106,712) (S1 Table) (Supplementary Note 1). Communities were matched into 15 pairs according to population size and age structure, access to health services and geographic location relative to major urban centers; communities in each pair were then randomized into intervention and control arms. Intervention communities received expanded HIV testing, early initiation of antiretroviral treatment and strengthened linkage-to-care, for example, clinic referrals with appointment dates, text alert reminders of appointments and following-up with those who missed appointments. Intervention communities also received wider access to prevention of mother-to-child transmission services and male circumcision services, compared with control communities that received the standard of care.

126

127 **Ethics statement.** The trial was approved by the Botswana Health Research and
128 Development Committee, the institutional review board of the Centers for Disease Control and
129 Prevention and the Harvard School of Public Health Office of Regulatory Affairs and Research
130 Compliance; and was monitored by a data and safety monitoring board and Westat. Written
131 informed consent for enrollment in the trial and viral HIV genotyping was obtained from all
132 participants.

133

134 **Dataset of people with HIV in the Botswana -Ya Tsie trial**. To better understand HIV
135 transmission patterns among participating communities, residents with HIV-1 were invited to
136 provide a sample for viral phylogenetic analyses. This included 1) all people with HIV-1
137 identified during a baseline survey of 20% of households randomly sampled from each trial
138 community to establish an incidence follow-up cohort, 2) all people with HIV-1 identified
139 through annual household surveys conducted in this 20% household sample in all 30
140 communities during the trial, 3) all people with HIV-1 who were enrolled in a community-wide
141 survey in 100% of 6 communities , 3 in the intervention and 3 in the control arms (the "end of
142 study survey"), 4) all newly identified people with HIV-1 discovered during an initial community-
143 wide HIV testing and counselling campaign followed by more targeted HIV testing in the 15
144 intervention communities, 5) all people with HIV-1 in intervention communities who later
145 presented at health care facilities after being identified during community-wide testing
146 campaigns, and 6) all people with HIV-1 who were already receiving HIV care at health facilities
147 in the 15 intervention communities. Altogether, 5,114 participants consented to a blood draw

148 for HIV-1 viral genotyping and their viral genomes were successfully deep-sequenced,
149 representing the largest phylogenetics study conducted in Botswana (S1 Table).
150
151 **Deep-sequencing of HIV viral genomes for phylogenetic analyses.** Because most
152 participants were virally suppressed at the time of sampling, deep sequencing of HIV-1 viral
153 genomes was generally done on proviral DNA (81%, n = 4,142) [14] (Supplementary Note 2.1).
154 We deep-sequenced near-full length viral genomes to optimize resolution and power to detect
155 viral genetic clusters of similar HIV-1 infections [15] (S1 and S2 Figs). For each participant we
156 thus obtained mapped short reads generated with deep-sequencing that capture the HIV-1 viral
157 diversity in an individual and a corresponding consensus sequence that represents a summary
158 of the HIV-1 viral population in that individual. Consensus sequences were generated from
159 alignments of the mapped short reads by identifying the majority nucleotide call at each base
160 position along the HIV-1 genome [16].
161
162 **Criteria for inclusion in phylogenetic analyses**. Our analysis was restricted to samples
163 from 3,832 participants (approximately 14 percent of the estimated 27,446 individuals living
164 with HIV in trial communities) who had available demographic information and met minimum
165 criteria on genome missingness (≥ 6,300 nucleotides) for inclusion in phylogenetic analyses.
166 More precisely, individuals whose viral consensus genomes had fewer than 30% of the bases
167 missing beyond the first 1,000 nucleotides were retained for analysis. Sampling density is
168 defined as the estimated proportion of individuals living with HIV in a trial community that were
169 included in our sample. Intervention communities were sampled more densely, 16% (2,281 /

170 14,263) than control communities, 12% (1,551 / 13,183); and sampling densities among all 30
171 trial communities ranged from 2.7% - 36.2% (S3 Fig, S2 Table) (Supplementary Note 1). Based
172 on the level of sexual mixing predicted for the trial [17] -- that 21% (standard error: 2.6%) of
173 relationships would be out-of-community partnerships (assumed for this calculation to be with
174 communities outside the sampled ones) -- and a genotyping density of 14% we expected our
175 dataset to contain the viral transmitter of a genotyped individual in about one out of every nine
176 cases of HIV-1 transmission i.e. $0.14 \times (1 - 0.21)$. This estimate reflects simplifying
177 assumptions that: (i) for a pair of individuals whose viral sequences are genetically linked, the
178 probability of sampling one member of the pair is independent from that of sampling the other
179 member of the pair such that the probability of observing a pair that is genetically linked is the
180 product of the individual sampling probabilities and (ii) the sampling probabilities were equal
181 for men and women with HIV.
182
183 **Consensus sequence phylogenetics to identify clusters of participants with genetically**
184 **similar HIV-1 infections**. Two clustering algorithms, HIV Transmission Cluster Engine (HIV-
185 TRACE) v0.4.4 [18, 19] and cluster picker v1.2.3 [20], were used to identify clusters of
186 individuals whose HIV-1 viral consensus sequences were genetically similar -- suggesting they
187 might be members of a transmission chain [21, 22] (Supplementary Note 2.3).
188
189
190

191 **Deep-sequence phylogenetics to infer the probable order of transmission events**
192 **within identified clusters of genetically similar HIV-1 infections.** To resolve the probable order
193 of transmission events within identified genetic clusters we performed parsimony-based
194 ancestral host state reconstruction of the mapped deep-sequencing short reads of participants
195 in clusters of genetically similar HIV-1 infections with the phyloscanner software v1.8.0 [23-25].
196 More precisely, we identified probable transmission pairs within each genetic cluster and the
197 probable direction of transmission between them (Supplementary Notes 2.4, 2.5).
198
199 **Adjustment for variable sampling rates across different demographic groups or**
200 **randomized-HIV-interventions.** To estimate HIV transmission flows in the trial population
201 within and between different age-gender groups and locations (trial communities, trial arms
202 and geographic regions), accounting for variable rates of sampling, we used the method
203 described by Carnegie and colleagues [26] (Supplementary Note 4). The method by Carnegie
204 and colleagues uses a frequentist approach, alternatively a Bayesian approach could be
205 employed as described by Ratmann and colleagues [27].
206
207 **Computation of confidence intervals for estimated HIV transmission flows in the trial**
208 **population.** Simultaneous confidence intervals for estimated HIV transmission flows in the trial
209 population were computed at the 5% significance level with the Goodman method [28]
210 (Supplementary Note 4). A continuity correction factor was implemented for the Goodman
211 method to optimize performance in the presence of small sample sizes [29].
212

213 ## Results
214

215 **Consensus sequence phylogenetics to identify clusters of participants with genetically similar**
216 **HIV-1 infections**

217

218 **Most participants with HIV at time of enrollment in the trial were on anti-retroviral**
219 **treatment and virally suppressed.** Of the 5,114 participants who provided a sample for HIV
220 viral genotyping, 3,178 were sampled at baseline and 1,936 were sampled post-baseline, that
221 is, about a year or more after the end of baseline survey activities in their communities. At the
222 time of identification, most (86%, n = 4,410) individuals with HIV-1 infection were on anti-
223 retroviral therapy and virally suppressed (≤ 400 copies per mL of blood plasma). Compared to
224 the 2011 Botswana census population of trial communities, our data set underrepresents men
225 (n = 1,475 *versus* 3,639) and individuals aged 16-24 years (S1 Table).

226

227 **Most of the HIV viral genomes sampled were from prevalent infections, and typical**
228 **genetic distances were large.** Of the 5,114 participants whose HIV-1 viral genomes were deep-
229 sequenced, 3,832 participants met minimum criteria for inclusion in phylogenetic analyses (see
230 methods); and of those 3,832 participants, 2,465 were sampled at baseline and 1,367 sampled
231 post-baseline. The HIV-incidence prospective cohort of the Botswana / Ya Tsie trial comprised a
232 random sample of 20% of households in each community to evaluate incidence of HIV-infection
233 and adoption of the intervention during the trial. The 3,832 participants included in
234 phylogenetic analyses comprised 85% (124 / 146) of the seroconverters in the HIV-incidence
235 cohort; seroconverters in the HIV-incidence cohort were defined as individuals that acquired

236 HIV infection during the trial (S2 Table). Seven of the 2,465 baseline samples included in
237 phylogenetic analyses were seroconverters in the HIV-incidence cohort (i.e. individuals with a
238 negative-HIV test and a subsequent positive-HIV test sample over the course of less than a year,
239 and thus defined as baseline) compared with 117 of the 1,367 post-baseline samples. The mean
240 $\pm$ standard deviation pairwise genetic distance between HIV-1 viral whole genome consensus
241 sequences of participants included in phylogenetic analyses was, $12.5 \pm 3.0\%$ and the
242 maximum pairwise distance was 40% substitutions per site (S4 Fig) (Supplementary Note 2.2);
243 we used a threshold of 4.5% substitutions per site to identify potentially closely related HIV-1
244 infections. The large distances between sequences possibly reflects sparse sampling and/or the
245 sampling of a large percentage of trial participants relatively late in infection. HIV-1 viral
246 populations typically diverge over time owing in part to within-host evolution.
247
248 **About 1 in 7 genotyped participants included in consensus sequence phylogenetic**
249 **analyses were assigned to genetic similarity clusters of people with closely related HIV-1**
250 **infections.** We identified 236 genetic similarity clusters comprising HIV-1 viral consensus
251 sequences from 14% (525 / 3,832) of participants (Fig 2). This represents the union of clusters
252 detected with HIV-TRACE (122 clusters, comprising 283 consensus sequences) and Cluster
253 Picker (231 clusters, comprising 484 consensus sequences) algorithms (S7 Fig). There was
254 substantial overlap between the two clustering algorithms, 96% (117 / 122) of genetic clusters
255 found with HIV-TRACE were also detected with cluster picker (S7 Fig). Only 4% (21 / 525) of the
256 participants assigned to genetic similarity clusters were seroconverters from the HIV-incidence
257 cohort.

258

259 **Most genetic clusters had few members and limited geographic spread.** The sizes of
260 genetic clusters ranged from 2 to 12 members, with 97% (230 / 236) of clusters comprising 2 or
261 3 members (S7 Fig). The two largest clusters had 9 and 12 members respectively, none of whom
262 were seroconverters from the HIV-incidence cohort; their members resided in communities in
263 all three geographic regions. Overall, however, genetic clusters were generally concentrated
264 within geographic regions (Central-East, North/North-East and South-East) than spread out
265 across them (Fig 2), with 42% (100 / 236) of clusters localized to two communities (S8 Fig) and
266 52% (124 / 236) of the clusters entirely within a single community (S8 Fig). Moreover, genetic
267 clusters were relatively evenly spread across trial arms; 38% (90 / 236) were unique to
268 intervention communities; 33% (78 / 236) to control communities; and 29% (68 / 236) had
269 genetic cluster members from both intervention and control communities (Fig 2). The number
270 of genetic clusters identified within trial communities generally increased with sampling
271 density; for example, communities such as: Shakawe, Ramokgonami, Mmathethe, Gumare,
272 Maunatlala, and Tati Siding which had larger numbers of unique and shared clusters compared
273 to other trial communities were also among the most densely sampled (Fig 3, S2 Table). The
274 abundance of small-sized genetic similarity clusters may reflect a moderate (14%) sampling
275 fraction [30].

276

277

278

279

280 **Characteristics of participants in clusters.** About 70% (366 / 525) of the HIV-1 viral
281 consensus sequences in clusters were from women (Fig 4). This result was consistent with the
282 overrepresentation of women among sequenced infections (S1 Table). The ages of clustered
283 men and women were similar with women having a median [lower – upper quartile] age of 37.8
284 [29.6 – 45.1] years and men 43.0 [34.9 – 51.0] years at time of sampling. There were few
285 clusters (4% (10 / 236)) where all members had HIV-1 infections with unsuppressed virus (S4
286 Table). 54% (286 / 525) of the participants in clusters of genetically similar HIV-1 infections
287 were also part of a baseline survey of 20% of households randomly sampled from each trial
288 community to establish an incidence follow-up cohort and gather information on different
289 socio-demographic variables including sexual risk behavior; see [3, 13] for details. Among them,
290 10% (30 / 286) of participants self-reported having multiple partners in the past twelve months;
291 73% (208 / 286) of participants self-reported a single partner in that period, 15% (44 / 286) of
292 participants self-reported zero partners in that period and the rest did not disclose their
293 number of partners. Of the 30 participants involved in multiple relationships, 90% (27 / 30)
294 disclosed having one other sexual partner, and 10% (3 / 30) reported having two other sexual
295 partners in addition to their primary sexual partner.
296

297 **Deep-sequence phylogenetics to infer the probable order of transmission events within**
298 **identified clusters of genetically similar HIV-1 infections**
299
300 We identified 153 highly supported probable source-recipient pairs within 236 clusters
301 of genetically similar HIV-1 infections. Seventy-one were same-sex linkages between women (n
302 = 65) or men (n = 6), and the remaining were male-female pairs (n = 82). Considering that the
303 predominant mode of HIV-1 transmission in Botswana and most of southern Africa is through
304 heterosexual contact, that direct transmission between women is rare [31], same-sex pairs
305 were presumed to be members of transmission chains with unsampled intermediates. We used
306 the probability of inferring a phylogenetically-linked directed male-male pair in the sample to
307 calibrate an upper bound on the false-discovery rate of inferring direct transmission between
308 males and females in phylogenetically-linked male-female pairs (Supplementary Note 3); the
309 estimated number of linked male-female pairs in the sample with unsampled intermediates
310 was approximately 30, corresponding to a false discovery rate of 36% (30 / 82). This estimated
311 false positive rate is likely inflated given that two individuals would need to be missing from the
312 sequenced sample to incorrectly infer transmission in a male-female pair, whereas only a single
313 female would need to be missing to erroneously infer transmission in a male-male pair. As
314 indicated earlier (see methods section on criteria for inclusion in phylogenetic analyses), we
315 expected our dataset to contain the HIV-1 viral sequence of the transmitter for 11% (0.14 x (1 –
316 0.21)) of sequenced participants included in phylogenetic analyses (n = 3,832), compared with
317 the 82 male-female probable transmission pairs that we identified thus suggesting that
318 consensus sequence phylogenetic analysis missed some viral genetic linkages between

319 participants with chronic infections; and deep-sequence phylogenetic analysis identified only a
320 subset of HIV-1 transmission events between participants assigned to genetic clusters. We
321 restricted further analyses on inference of HIV transmission patterns to the highly supported 82
322 male-female probable transmission pairs. Analyses are presented first without an adjustment
323 for sampling variability to illuminate patterns of viral transmission within the deep-sequenced
324 sample. Thereafter an adjustment for variable sampling rates by demographic group (age,
325 gender, trial community, geographical region) or randomized-HIV-intervention (trial arm) is
326 made to estimate the flow of HIV transmissions within the trial population (Supplementary
327 Notes 1 and 4).
328
329 **Age and sex distribution of sources and recipients in inferred transmission pairs.** The
330 age distribution of transmitters and recipients, by gender, for the 82 male-female probable
331 transmission pairs is shown in Fig 5. Inferred male-female and female-male HIV transmissions
332 had similar age distributions for the inferred source of infection. Males were a mean $\pm$
333 standard deviation of 3.5 ± 9.6 years older than females in probable transmission pairs overall,
334 with a difference of 4.9 ± 7.5 in male-to-female transmission events, and 1.8 ± 11.6 years in
335 female-to-male transmission events. Adjusted for variability in sampling of the trial population
336 by gender and 5-year age group (Supplementary Note 5), these differences were somewhat
337 reduced: males in probable transmission pairs were a mean $\pm$ standard deviation of 1.3 ± 10.4
338 years older than females overall; with a difference of 3.8 ± 6.9 years in male-to-female
339 transmission events, and -1.3 ± 12.7 years in female-to-male transmission events. The

difference between the unadjusted and the weighted mean age gap between males and females in probable transmission pairs is consistent with the under-sampling of younger males.

Most HIV transmissions occurred between similarly aged partners. S5 Table shows the proportions (unadjusted) of viral transmission events within and between 5-year age groups among the 82 probable male-female pairs inferred from the deep-sequenced sample. After adjusting for differential sampling (Supplementary Notes 1 and 4) by 5-year age group, the predicted proportions of HIV-1 transmissions in the trial population suggested preferential sexual mixing between similarly-aged partners (Fig 6 and S5 Table). An adjustment by both 5-year age group and gender revealed broadly similar transmission patterns in the trial population (S9 Fig and S6 Table).

Similar proportions of men and women as inferred sources of transmission. 55% [95% Confidence Interval 39.8 – 69.1] of the 82 inferred male-female transmission events originated from men with the remainder occurring from women, 45% [30.9 – 60.2] (Fig 7). An adjustment for variable sampling by gender revealed similar transmission patterns in the trial population (S7 Table).

361 **Most inferred HIV transmissions occurred within communities or between**
362 **neighboring communities.** Trial communities could be broadly grouped into three geographical
363 areas: in the North/North-East, South-East, and Central-East region of Botswana, which borders
364 South Africa and Zimbabwe (Fig 1). The proportion of inferred transmission events of HIV-1
365 infection in the deep-sequenced sample between members of the same trial community was
366 69% (57 / 82) compared with 21% (17 / 82) for members of different trial communities in the
367 same region, and 10% (8 / 82) for members of trial communities in different regions (Figs 7 and
368 8). Of the 74 (74 / 82) transmission events within a single region (including those within a single
369 community), 22 were in the Central-East, 28 in the North/North-East and 24 in the South-East
370 (Fig 7). For pairs where transmission events linked people in different trial communities the
371 median [lower – upper quartile] driving distance between the trial communities was 161 km
372 [108 - 420 km] or 100 mi [67 – 261 mi] and the median drive time was 1.86 h [1.41 – 4.68 h].
373 Furthermore, a comparison of mean drive times between pairs of trial communities for which
374 transmission events were identified (n = 22) with those of all other possible pairs of
375 participating trial communities excluding same community pairs (n = 900 – 22 – 30 = 848)
376 revealed a shorter drive time on average between genetically linked trial communities
377 compared with what would be expected under random sexual mixing $(\text{permutation-test}, p <$
378 $0.01)$ (Supplementary Note 6). These results suggest that out-of-community sexual
379 partnerships were more likely to form between partners residing in nearby communities
380 compared with more distant communities.
381 After adjusting for differential sampling by community (Supplementary Note 4), an
382 estimated 24% of transmissions to a resident of a trial community originated from another trial

383 community (S8 Table). This is consistent with the high levels of mobility in Botswana, where
384 people typically have family ties in ancestral villages and work in peri-urban or urban areas [10].
385 Furthermore, trial communities were distributed along two principal highways that might have
386 contributed to increased mobility: the A1 highway that connects the Northern and Southern
387 parts of the country and the A3 highway that branches from the A1 highway connecting the
388 North-Eastern and North-Western regions.
389 The estimated proportion of HIV transmissions in the trial population, after adjusting for
390 differential sampling by geographic region (Supplementary Note 4), that occurred within the
391 South-East region, 50% [95% Confidence Intervals 29.6 – 70.5] was higher compared to that in
392 the Central-East, 23% [9.7 – 45.4] and North-East regions, 18% [6.5 – 39.8], respectively (Fig 9
393 and S9 Table). For comparison, the proportion of trial participants with viral genomes that met
394 criteria for inclusion in phylogenetic analyses were 26% (1,000 / 3,832) in the South-East region,
395 42% (1,630 / 3,832) in the North/North-East region and 31% (1,202 / 3,832) in the Central-East
396 region (S2 Table). The South-East region includes Gaborone city, the economic and
397 administrative capital of Botswana; conversely, economic activity in the North- East and
398 Central- East regions is centered around mining and agriculture.
399
400 **More HIV transmission events to residents of intervention communities originated**
401 **from residents of control communities than vice versa. Overall analysis:** The proportion of
402 inferred transmission events of HIV-1 infection in the deep-sequenced sample occurring within
403 intervention (n = 33) or within control communities (n = 32) was 79% (65 / 82) compared with
404 21% (17 / 82) across intervention and control communities (Figs 7 and 8). Of the thirty-three

405 transmission events inferred within intervention communities, seven occurred between
406 partners residing in different communities; and for those inferred within control communities,
407 one of the thirty-two transmission events occurred between partners from different
408 communities. There were more transmission events identified in intervention communities that
409 originated from control communities than the reverse (n = 13 *versus* n = 4).

410 After adjusting for differential sampling by trial arm (Supplementary Note 4), the
411 estimated proportions of HIV transmissions in the trial population within control communities,
412 51% [35.9 – 66.5] was about twice that predicted in intervention communities, 29% [16.6 –
413 44.6] (Fig 10 and S10 Table); and the estimated proportion of HIV transmissions in the trial
414 population flowing into intervention communities from control communities, 15% [6.9 – 30.1]
415 was about three times higher than the reverse, 5% [1.0 – 16.9]. Comparatively, the proportions
416 of trial participants that met criteria for inclusion in phylogenetic analyses from control
417 communities were 40% (1,551 / 3,832) and from intervention communities, 60% (2,281 /
418 3,832).

419 **Baseline analysis:** Of the 82 male-female probable transmission pairs identified in the
420 deep-sequenced sample, 51 involved a recipient that was first sampled during the period of
421 baseline household survey activities in their community, that is, a transmission that could not
422 have been affected by the intervention because it occurred before the start of the intervention.
423 Among these (n = 51) the proportion of inferred transmission events that occurred within
424 intervention (n = 19) or within control communities (n = 25) was 86% (44 / 51) compared with
425 14% (7 / 51) across intervention and control communities (S11 Table). Of the nineteen
426 transmission events identified within intervention communities, four were between partners

427 living in different communities; compared with one out of twenty-five transmission events
428 identified within control communities. The number of transmission events identified in
429 intervention communities that originated from control communities was similar to the reverse
430 (n = 4 *versus* n = 3).

431 Adjusted for sampling variability among trial arms (Supplementary Note 4) the flow of
432 HIV transmissions in the trial population within control communities, 62% [41.6 – 78.8] was
433 more than twice that in intervention communities, 25% [11.8 – 45.9] (S11 Table); and the flow
434 of HIV transmissions into intervention communities from control communities, 7% [1.5 – 25.3]
435 was similar to the reverse, 5% [0.9 – 22.9].

436 **Post-baseline analysis:** By contrast, 31 of the 82 male-female probable transmission
437 pairs inferred in the deep-sequenced sample involved a recipient that was first sampled about a
438 year or more after the end of baseline household survey activities in their community; this
439 subset reflects transmissions that could have in principle been affected by the intervention, as
440 they may have occurred after the intervention began. Among these (n = 31) the proportion of
441 inferred transmission events that occurred within intervention (n = 14) or within control
442 communities (n = 7) was 68% (21 / 31) compared with 32% (10 / 31) across intervention and
443 control communities (S12 Table). Moreover, three of the fourteen transmission events
444 identified within intervention communities were between partners residing in different
445 communities, and all seven transmission events identified within control communities were
446 between partners living in the same community. We identified more transmission events in
447 intervention communities that originated from control communities than vice versa, (n = 9
448 *versus* n = 1).

449 Adjusted for variability in sampling among trial arms (Supplementary Note 4) the flow of
450 HIV transmissions within control communities, 32% [13.3 – 58.3] was similar to that within
451 intervention communities, 34% [15.0 – 60.9]; and HIV transmissions into intervention
452 communities from control communities, 30% [12.2 – 56.7] were more common than the
453 reverse, 3% [0.1 – 27.3] (S12 Table). This result is consistent with a predicted benefit of the
454 treatment-as-prevention intervention in reducing HIV transmission in the Botswana - Ya Tsie
455 trial, albeit precise dates of transmission were not identified.
456

457 **Discussion**

458

459 HIV incidence remains high in East and Southern Africa especially among young women.
460 Promising HIV prevention intervention packages that combine population-level HIV testing
461 coupled with strengthened linkage-to-care and early initiation of anti-retroviral therapy, jointly
462 termed universal test-and-treat, have yielded lower than expected reductions in the occurrence
463 of new HIV-1 infections[2-6]. This makes it more challenging to attain the UNAIDS goal of
464 ending the HIV/AIDS epidemic as a public health emergency by 2030. An emerging hypothesis
465 among four of the largest universal test-and-treat HIV prevention trials ever conducted in East
466 and Southern Africa in Kenya, Uganda, Botswana, South Africa and Zambia[6], comprising tens
467 of thousands of participants, is that the intervention impact to reduce population-level HIV
468 incidence was dampened by population mobility; in particular, sexual partnerships formed 1)
469 between individuals in communities randomized to different HIV-intervention conditions and 2)
470 with individuals in communities outside the trial population.

471

472 In this study, we focused on one of the four cluster/community-randomized trials, the
473 Ya Tsie trial conducted in Botswana[3], to quantify the extent of sexual mixing within and
474 between: trial communities, trial arms and geographic regions. Furthermore, we quantified the
475 contribution of age-difference and gender to the spread of HIV infections in Eastern Botswana.
476 We first used HIV-1 whole genome viral consensus sequences to identify clusters of trial
477 participants with genetically similar HIV-1 infections; and thereafter, employed deep-sequence
478 phylogenetics to resolve the probable order of HIV transmission events within each identified

479 cluster. Identified transmission events in the deep-sequenced sample were then used to
480 estimate the flow of HIV transmissions in the trial population adjusting for sampling variability.
481
482 We found that HIV transmissions in the trial were more likely to occur within
483 communities or between neighboring communities than between distant trial communities, as
484 well as between similarly-aged partners. One of the most striking findings of our analysis was
485 that an estimated 24% of transmissions involving residents of trial communities occurred
486 between different communities in the trial. Given that the trial covered only 7.6% of the
487 national population, this suggests that a large fraction of all transmissions involved a partner
488 outside the trial.
489 Furthermore, there was substantial sexual mixing between intervention and control
490 communities. Overall, about one in every five HIV transmissions in the trial population involved
491 sexual partnerships between residents of intervention communities and control communities.
492 After baseline, that is, about a year or more after the introduction of the intervention one in
493 three HIV transmissions occurred between intervention and control communities. HIV
494 transmissions into intervention communities from control communities were similar to the
495 reverse at baseline, and ten times more common post-baseline, concordant with a predicted
496 benefit of a universal test-and-treat HIV prevention intervention, though dates of transmission
497 were not identified.
498 The extent of sexual mixing between residents of intervention and control communities
499 highlights high mobility between the two arms of the Botswana - Ya Tsie trial. For the primary
500 trial endpoint comparing HIV incidence in intervention versus control communities, this kind of

501 mixing will tend to dilute the effect since some of the prevention value of the intervention
502 would appear in control communities, while intervention communities would have incidence
503 that comes from control communities and from communities outside the trial thus is not
504 reachable by the effect of the intervention. Both sources of incidence, from control
505 communities and from outside the trial, could dilute the intervention effect observed in the
506 trial. Although we could not evaluate it in this study, it is possible that out-of-trial index cases
507 would dilute the effect even more strongly than control-arm ones because the control arms
508 comprise only about 5% of Botswana's population, while about 90% of the population lives
509 outside the trial.

510 These findings might be relevant for the interpretation of the results of the other three
511 universal test-and-treat HIV prevention trials conducted in Kenya, Uganda, South Africa and
512 Zambia. An important next step would be to quantify the extent to which the size of the
513 observed effect in the Botswana – Ya Tsie trial was reduced by imported infections between
514 partners in communities randomized to different HIV-intervention conditions and partners in
515 communities outside the trial population.

516

517 There was a deliberate effort to enroll men and young individuals in the Botswana -Ya
518 Tsie study as these groups are less likely to be diagnosed with HIV and engaged in care.
519 Following an adjustment for variable sampling by age and gender (S9 Fig and Supplementary
520 Note 4) we found that men and women for whom sexual partners could be inferred
521 phylogenetically contributed similarly to viral transmission of HIV-1 infection in the Botswana /
522 Ya Tsie trial. There would also be a contribution to HIV incidence from individuals whose source

523 partner resided in a community outside the trial, though not evaluated in this analysis as HIV
524 viral sequences from communities outside the trial population were not available.
525
526 Prior to this study, a recent consensus sequence phylogenetic study assessing age-
527 disparate sexual partnerships in Kwa-Zulu Natal, South Africa identified men older than 25
528 years as an important source of viral transmission to younger women aged 15-25 years. Age
529 disparate partnerships were defined as ones with an age-gap of more than 5 years. The Kwa-
530 Zulu Natal study was based on a community sampling fraction of 4% that was not specifically
531 targeted to people with HIV and assumed that the direction of viral transmission within clusters
532 would be from older partners to younger ones [9]. Phylogenetic analyses in the Botswana / Ya
533 Tsie trial were based on an overall sampling fraction of 14% of people aged 16 – 64 years living
534 with HIV in trial communities; the sampling fraction ranged from 2.7% – 36.2% among the 30
535 trial communities. In the phylogenetic analyses of the Botswana / Ya Tsie study there was little
536 genetic evidence from analyses of deep-sequence data to suggest that older men were a
537 substantial source of HIV-1 infection to younger women in Botswana. This finding is relevant to
538 programs aimed at reducing HIV incidence in young women in sub-Saharan Africa such as the
539 Determined, Resilient, Empowered, AIDS-free, Mentored, and Safe women (DREAMS)
540 partnership. Consistent with our results, the Vaginal and Oral Interventions to Control the
541 Epidemic (VOICE) trial, a placebo-controlled randomized study of pre-exposure prophylaxis to
542 test the efficacy of providing oral and vaginal gel tenofovir to women in Southern Africa to
543 prevent HIV infection [32, 33], found no association between age-disparate relationships and
544 risk of HIV-1 infection in young women under 25 years residing in South Africa in the cities of

545 Durban, Johannesburg and Klerksdorp who had male partners that were > 5 years older at
546 enrollment.
547
548 Taken together, these findings might be relevant for understanding heterogeneity of
549 HIV transmission in similarly designed phylogenetic studies in sub-Saharan Africa based on the
550 sample size, demographic population and adjustments for variable sampling.
551
552 Participants with genetically similar HIV-1 infections in the Botswana – Ya Tsie trial
553 tended to aggregate in small-sized (two or three-person) clusters. This is consistent with
554 phylogenetic clustering studies in generalized epidemics in Southern Africa but contrasts with
555 larger clusters typically found in concentrated HIV epidemics in Europe and North America [23].
556 The small-sized genetic similarity clusters identified in trial communities might reflect under-
557 sampling of key populations, for example men and younger individuals, in the HIV transmission
558 network. Alternatively, the predominance of smaller clusters in generalized epidemics
559 compared with larger clusters in concentrated epidemics might reflect differences in the
560 manner in which sexual networks under the two settings evolve over time.
561
562 Our findings provide insight on HIV transmission patterns in Botswana over a five-year
563 period between 2013 and 2018 during which the Botswana- Ya Tsie trial was conducted. An
564 important next step would be the integration of the large database of HIV viral genomes
565 assembled by the Botswana Ya Tsie trial into a real-time HIV genomic surveillance program to
566 highlight emerging hubs of HIV spread and inform targeted studies of HIV prevention. Such an

567 effort would be facilitated by samples routinely collected for HIV viral load monitoring and the
568 continuously decreasing cost of high-throughput sequencing.
569
570 Our study had several limitations. First, viral genomes from an estimated 14% of all
571 people with HIV aged 16-64 years were sampled among participating trial communities.
572 Comparatively, the sampling fraction of phylogenetic studies based on African genomes is
573 commonly under 10% [23]. Although our sampling fraction was a modest improvement over
574 previous studies it is a relatively small sample of HIV viral genomes in the trial population. To
575 estimate HIV transmission flows in the trial population based on probable viral transmission
576 events identified in the deep-sequenced sample we used the method of Carnegie and
577 colleagues in which undetected HIV transmissions were assumed to be missing at random
578 conditional on group membership [26] (Supplementary Note 4). Here group membership refers
579 to demographic groups (age-group, gender, trial community, geographical region) or
580 randomized-HIV interventions.
581 Of the 82 probable male-female transmission pairs identified in the deep-sequenced
582 sample there was a single individual involved in multiple transmission events; a male who
583 transmitted to two females each residing in different communities from his own. This
584 suggested that only a small degree of bias in estimation of confidence intervals would arise
585 from a simplifying assumption of independence between inferred transmission events; and
586 allowed us to estimate confidence intervals using methods for multinomial proportions
587 (Supplementary Note 4).
588

589 Second, although our estimates of HIV transmission flows in the trial population were
590 adjusted for differential sampling we cannot exclude the possibility that unmeasured factors
591 with influence on HIV transmission flows may have impacted our results. Most participants with
592 HIV in this trial were virally suppressed, this might have biased our results in the sense that
593 there might be some differences between the HIV transmission patterns we found and those
594 that exist among viral genetic clusters largely populated by people with unsuppressed virus.
595 Nevertheless, this study was based on a relatively representative sample of adults with HIV
596 living in rural and peri-urban villages across Eastern Botswana.
597
598 In sum, we identified several key factors in the Botswana -Ya Tsie trial that impacted HIV
599 transmission dynamics with potential relevance to similar studies done in sub-Saharan Africa.
600 We found that most HIV transmissions in the Botswana Ya Tsie trial occurred between similarly-
601 aged partners within the same trial community or between trial communities in close
602 proximity. Moreover, there was a greater flow of HIV transmissions into intervention
603 communities from control communities than vice versa potentially reducing the observed
604 effect-size of the trial, as would transmissions averted by the intervention that were not picked
605 up because the recipient was not in a trial community.
606
607 We recommend widely distributed and easily accessible HIV testing (e.g. universal HIV
608 testing campaigns), treatment and linkage-to-care to support people as they intermix within
609 and across communities; augmented by targeted programs that might offer pre-exposure
610 prophylaxis to younger women and the option of self-testing for men. Population-level genomic

surveillance programs to identify communities with high flows of viral transmission within and between them would be helpful in spotting emerging hubs of transmission that can be prioritized for intervention. Together, these measures may reduce new HIV infections and shorten the time to epidemic control.

Data availability: All relevant data are within the paper and the Supplemental Notes, figures and tables. HIV-1 reads are available on reasonable request through the PANGEA consortium (www.pangea-hiv.org)

Code availability: Algorithms to estimate HIV transmission flows within and between population groups accounting for sampling variability and corresponding confidence intervals have been implemented as an R package, *bumblebee* that will be made available at the following URL: https://magosil86.github.io/bumblebee

625 ## References

626

627 1. UNAIDS, *Global AIDS update 2019 — Communities at the centre*. 2019.

628 2. Hayes, R.J., et al., *Effect of Universal Testing and Treatment on HIV Incidence — HPTN*
629 *071 (PopART).* New England Journal of Medicine, 2019. **381**(3): p. 207-218.

630 3. Makhema, J., et al., *Universal Testing, Expanded Treatment, and Incidence of HIV*
631 *Infection in Botswana.* New England Journal of Medicine, 2019. **381**(3): p. 230-242.

632 4. Havlir, D.V., et al., *HIV Testing and Treatment with the Use of a Community Health*
633 *Approach in Rural Africa.* New England Journal of Medicine, 2019. **381**(3): p. 219-229.

634 5. Iwuji, C.C., et al., *Universal test and treat and the HIV epidemic in rural South Africa: a*
635 *phase 4, open-label, community cluster randomised trial.* Lancet HIV, 2018. **5**(3): p.
636 e116-e125.

637 6. Abdool Karim, S.S., *HIV-1 Epidemic Control — Insights from Test-and-Treat Trials.* New
638 England Journal of Medicine, 2019. **381**(3): p. 286-288.

639 7. Cohen, M.S., et al., *Antiretroviral Therapy for the Prevention of HIV-1 Transmission.* New
640 England Journal of Medicine, 2016. **375**(9): p. 830-839.

641 8. Granich, R.M., et al., *Universal voluntary HIV testing with immediate antiretroviral*
642 *therapy as a strategy for elimination of HIV transmission: a mathematical model.* Lancet,
643 2009. **373**(9657): p. 48-57.

644 9. de Oliveira, T., et al., *Transmission networks and risk of HIV infection in KwaZulu-Natal,*
645 *South Africa: a community-wide phylogenetic study.* Lancet HIV, 2017. **4**(1): p. e41-e50.

646 10. Essex, M., J. Makhema, and S. Lockman, *Reaching 90-90-90 in Botswana.* Current
647 Opinion in HIV and AIDS, 2019. **14**(6): p. 442-448.

648 11. The World Bank, *GINI index - Botswana.* World Development Indicators, 2015.

649 12. UNAIDS, *HIV estimates with uncertainty bounds 1990-2019.* Global AIDS update 2020,
650 2020.

651 13. Gaolathe, T., et al., *Botswana's progress toward achieving the 2020 UNAIDS 90-90-90*
652 *antiretroviral therapy and virological suppression goals: a population-based survey.*
653 Lancet HIV, 2016. **3**(5): p. e221-30.

654 14. Novitsky, V., et al., *Long-Range HIV Genotyping Using Viral RNA and Proviral DNA for*
655 *Analysis of HIV Drug Resistance and HIV Clustering.* Journal of Clinical Microbiology,
656 2015. **53**(8): p. 2581-92.

657 15. Dennis, A.M., et al., *Phylogenetic studies of transmission dynamics in generalized HIV*
658 *epidemics: an essential tool where the burden is greatest?* Journal of Acquired Immune
659 Deficiency Syndromes, 2014. **67**(2): p. 181-95.

660 16. Wymant, C., et al., *Easy and accurate reconstruction of whole HIV genomes from short-*
661 *read sequence data with shiver.* Virus Evolution, 2018. **4**(1): p. vey007.

662 17. Novitsky, V., et al., *Phylogenetic Relatedness of Circulating HIV-1C Variants in Mochudi,*
663 *Botswana.* PLOS ONE, 2013. **8**(12): p. e80589.

664 18. Wertheim, J.O., et al., *The global transmission network of HIV-1.* Journal of Infectious
665 Diseases, 2014. **209**(2): p. 304-13.

666 19. Kosakovsky Pond, S.L., et al., *HIV-TRACE (TRAnsmission Cluster Engine): a Tool for Large*
667 *Scale Molecular Epidemiology of HIV-1 and Other Rapidly Evolving Pathogens.* Molecular
668 biology and evolution, 2018. **35**(7): p. 1812-1819.

669 20. Ragonnet-Cronin, M., et al., *Automated analysis of phylogenetic clusters.* BMC
670 Bioinformatics, 2013. **14**(1): p. 317.

671 21. Poon, A.F., *Impacts and shortcomings of genetic clustering methods for infectious*
672 *disease outbreaks.* Virus Evolution, 2016. **2**(2): p. vew031.

673 22. Rose, R., et al., *Identifying Transmission Clusters with Cluster Picker and HIV-TRACE.* AIDS
674 Research and Human Retroviruses, 2017. **33**(3): p. 211-218.

675 23. Grabowski, M.K., J.T. Herbeck, and A.F.Y. Poon, *Genetic Cluster Analysis for HIV*
676 *Prevention.* Current HIV/AIDS Reports, 2018. **15**(2): p. 182-189.

677 24. Ratmann, O., et al., *Inferring HIV-1 transmission networks and sources of epidemic*
678 *spread in Africa with deep-sequence phylogenetic analysis.* Nature Communications,
679 2019. **10**(1): p. 1411.

680 25. Wymant, C., et al., *PHYLOSCANNER: Inferring Transmission from Within- and Between-*
681 *Host Pathogen Genetic Diversity.* Molecular Biology and Evolution, 2017. **35**(3): p. 719-
682 733.

683 26. Carnegie, N.B., et al., *Linkage of viral sequences among HIV-infected village residents in*
684 *Botswana: estimation of linkage rates in the presence of missing data.* PLoS
685 Computational Biology, 2014. **10**(1): p. e1003430.

686 27. Ratmann, O., et al., *Quantifying HIV transmission flow between high-prevalence*
687 *hotspots and surrounding communities: a population-based study in Rakai, Uganda.*
688 Lancet HIV, 2020. **7**(3): p. e173-e183.

689 28. Goodman, L.A., *On Simultaneous Confidence Intervals for Multinomial Proportions.*
690 Technometrics, 1965. **7**(2): p. 247-254.

691 29. Cherry, S., *A Comparison of Confidence Interval Methods for Habitat Use-Availability*
692 *Studies.* The Journal of Wildlife Management, 1996. **60**(3): p. 653-658.

693 30. Murray, M. and D. Alland, *Methodological Problems in the Molecular Epidemiology of*
694 *Tuberculosis.* American Journal of Epidemiology, 2002. **155**(6): p. 565-571.

695 31. Chan, S.K., et al., *Likely female-to-female sexual transmission of HIV--Texas, 2012.*
696 MMWR. Morbidity and mortality weekly report, 2014. **63**(10): p. 209-212.

697 32. Saag, M.S., *Preventing HIV in Women — Still Trying to Find Their VOICE.* New England
698 Journal of Medicine, 2015. **372**(6): p. 564-566.

699 33. Balkus, J.E., et al., *Age-Disparate Partnerships and Risk of HIV-1 Acquisition Among*
700 *South African Women Participating in the VOICE Trial.* Journal of Acquired Immune
701 Deficiency Syndromes, 2015. **70**(2): p. 212-7.
702

703
704

705 **List of figures**
706
707 **Figure 1. Location of 30 rural and peri-urban communities that participated in the Botswana –**
708 **Ya Tsie trial. Panel A.** A map of Botswana showing the spatial distribution of intervention and
709 control communities in the Botswana -Ya Tsie trial. Intervention communities are denoted by
710 filled blue circles and control communities are represented by filled red triangles. Trial
711 communities are shown in the context of level-2 administrative sub-divisions (solid black lines)
712 and 2011 census enumeration areas (solid grey lines). Major urban areas are symbolized as
713 filled orange squares. **Panel B.** Spatial distribution of the estimated population density of
714 persons per square kilometer in Botswana in 2011. Filled blue circles represent the location of
715 intervention communities and filled green triangles denote control communities in the
716 Botswana -Ya Tsie trial.
717
718 **Figure 2. A dendogram showing clusters of genetically similar HIV-1 infections in the**
719 **Botswana -Ya Tsie trial.** Genetic clusters are summarized by randomized-HIV-intervention
720 condition, geographical region and occurrence within and between trial communities. Clusters
721 were identified from 3,832 HIV-1 viral whole genome consensus sequences as two or more
722 sequences separated by a genetic distance not exceeding 0.045 (4.5%) substitutions per site
723 and a bootstrap support threshold of at least 80%. More genetic clusters occurred within trial
724 communities and geographical regions than between them.
725
726 **Figure 3. A barplot summarizing clusters of genetically similar HIV-1 infections by trial**
727 **community.** Bars represent the total number of genetic clusters identified in each trial
728 community partitioned into clusters that are specific to a community (lighter shade) and those
729 that are shared with at least one other trial community (darker shade). Genetic clusters in
730 intervention communities are shown in blue and those in control communities are represented
731 in green. The number of genetic clusters identified among intervention and control
732 communities varied.
733
734 **Figure 4. Density plots summarizing trial participants in clusters of genetically similar HIV-1**
735 **infections by age and gender.** The bottom two plots show the age-gender distribution of trial
736 participants whose HIV-1 viral whole genome consensus sequences were in genetic clusters
737 identified with HIV-TRACE (n = 283) and Cluster Picker (n = 484) clustering algorithms. The top
738 two panels show the age-gender distribution of trial participants whose HIV-1 viral whole
739 genome consensus sequences were in the union (n = 525) and overlap (n = 281) of genetic
740 clusters identified with HIV-TRACE and Cluster Picker. The raw data is shown as a rug plot (red
741 tick marks) on the x-axis. Men and women in genetic clusters had similar ages.
742
743 **Figure 5. Age distribution of 82 male-female probable transmission pairs sampled in the**
744 **Botswana -Ya Tsie trial.** Forty-five male-to-female and 37 female-to-male transmission events
745 were identified from the deep-sequenced viral whole genomes of 525 trial participants in
746 clusters of genetically similar HIV-1 infections. Men were typically older than women in both
747 male-to-female and female-to-male transmission events.
748

749 **Figure 6. Estimated transmission flows of HIV-1 infection within and between 5 year age**
750 **groups in the Botswana -Ya Tsie trial population.** Transmission flows of HIV-1 infection in the
751 trial population were estimated from 82 male-female probable transmission pairs identified
752 from the deep-sequenced HIV-1 viral whole genomes of trial participants (n = 525) in clusters (n
753 = 236) of genetically similar HIV-1 infections. Furthermore, transmission flows were adjusted for
754 differential sampling among age groups (Supplementary Note 4). Most viral transmission events
755 occurred between similarly aged partners.
756
757 **Figure 7. An alluvial plot showing counts of male-female probable transmission pairs sampled**
758 **in the Botswana -Ya Tsie trial by geographical location.** The 82 probable male-female infector -
759 infectee pairs that were identified from clusters of genetically similar HIV-1 infections (n = 236)
760 in the Botswana -Ya Tsie trial are first summarized into those with a male versus a female
761 source of viral transmission. Male-to-female and female-to-male transmission events are then
762 further partitioned according to whether partners in a transmission pair resided in the same or
763 different trial communities and if those trial communities belonged to the same or different
764 trial arms and geographical regions. Male-to-female transmission events are shown in green
765 and female-to-male transmission events are represented in yellow. About 69% (n = 57) of
766 sampled transmission events in the Botswana -Ya Tsie trial occurred within the same trial
767 community.
768
769 **Figure 8. A transmission matrix summarizing sampled transmission events within and**
770 **between Botswana -Ya Tsie trial communities.** Numbered tiles represent counts of
771 transmission events identified between a pair of trial communities. There were 32 transmission
772 events identified within control communities (magenta), 33 within intervention communities
773 (cyan), 13 from control-to-intervention communities (green) and 4 from intervention-to-control
774 communities (red). Most transmission events were identified within trial communities. The
775 number of sampled transmission events into intervention communities from control
776 communities was higher than the reverse (n = 13 vs 4).
777
778 **Figure 9. Estimated transmission flows of HIV-1 infection in the Botswana -Ya Tsie trial**
779 **population within and between geographical regions.** Communities in the Botswana -Ya Tsie
780 trial can be broadly grouped into three geographical regions in the Central-East, North/North-
781 East and South-Eastern parts of the country. Transmission flows were estimated as described in
782 Figure 6. Moreover, transmission flows were adjusted for sampling heterogeneity among
783 geographic regions. The flow of HIV-1 transmissions in the trial population was highest in the
784 South-East region (50.1%) compared with the Central-East (23.1%) and North/North-East
785 (17.9%) regions respectively.
786
787 **Figure 10. Estimated transmission flows of HIV-1 infection within and between intervention**
788 **and control communities in the Botswana -Ya Tsie trial population.** Transmission flows were
789 estimated as described in Figure 6. Additionally, transmission flows were adjusted for
790 differential sampling among trial arms. Most transmissions of HIV-1 infection occurred within
791 the same trial arm and the flow of viral transmissions into intervention communities from
792 control communities (15.3%) was about three times higher than the reverse (4.7%).

793 **Acknowledgements**

794 We are grateful to participants and collaborators from the Botswana Combination Prevention
795 Project and PANGEA consortium for their support during this work. Additionally, we thank
796 Susan Eshleman for contributing data from the HPTN 052 study, and Roger Shapiro for his
797 helpful comments on early versions of this manuscript. This study was supported by the
798 National Institute of General Medical Sciences (U54GM088558); the Fogarty International
799 Center (FIC) of the U.S. National Institutes of Health (D43 TW009610); and the President's
800 Emergency Plan for AIDS Relief through the Centers for Disease Control and Prevention (CDC)
801 (Cooperative agreements U01 GH000447 and U2G GH001911).

802

803 **Disclaimer**

804 The findings and conclusions in this report are those of the author(s) and do not necessarily
805 represent the official position of the funding agencies.

806

807 **Author contributions**

808 L.E.M., M.L., M.E. conceived the study; M.E., S.L., J.M., C.F., M.L., M.P., P.B., J.M., T.S., S.M.,
809 R.L., V.N., Y.Z. facilitated data acquisition; V.N., C.F., T.G. supervised viral genotyping and
810 performed data curation; L.E.M. performed statistical and computational analyses; M.L., V.D.,
811 C.F., E.T. evaluated statistical and computational analyses; S.L., M.P. administered the project;
812 M.L., M.E. supervised the project; L.E.M., M.L., V.D. wrote the original draft of the manuscript;
813 all authors reviewed and approved the statistical and computational analyses and final version
814 of the manuscript.

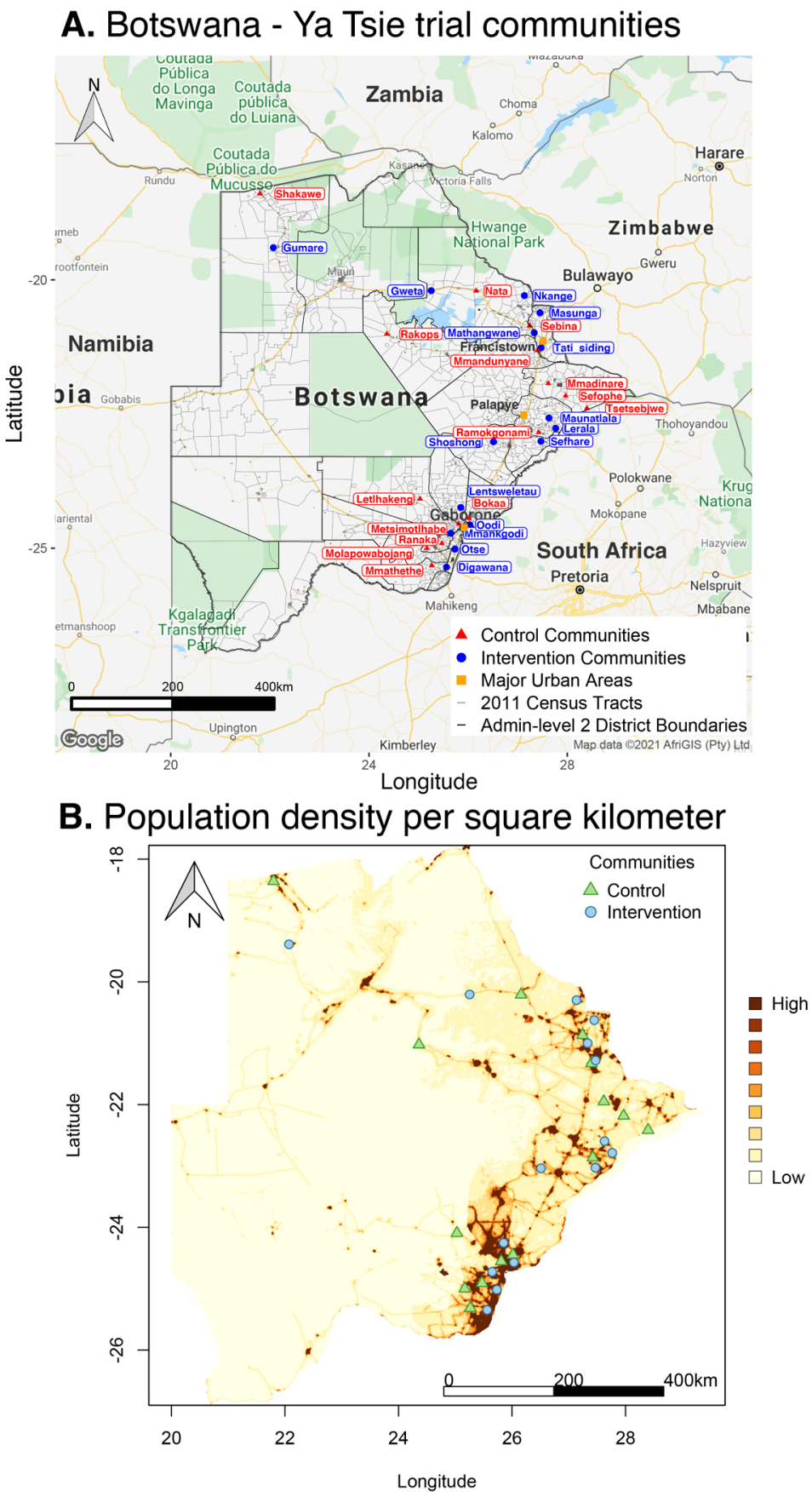


Figure 1. Location of 30 rural and peri-urban communities that participated in the Botswana – Ya Tsie trial. Panel A. A map of Botswana showing the spatial distribution of intervention and control communities in the Botswana -Ya Tsie trial. Intervention communities are denoted by filled blue circles and control communities are represented by filled red triangles. Trial communities are shown in the context of level-2 administrative sub-divisions (solid black lines) and 2011 census enumeration areas (solid grey lines). Major urban areas are symbolized as filled orange squares. Panel B. Spatial distribution of the estimated population density of persons per square kilometer in Botswana in 2011. Filled blue circles represent the location of intervention communities and filled green triangles denote control communities in the Botswana -Ya Tsie trial.

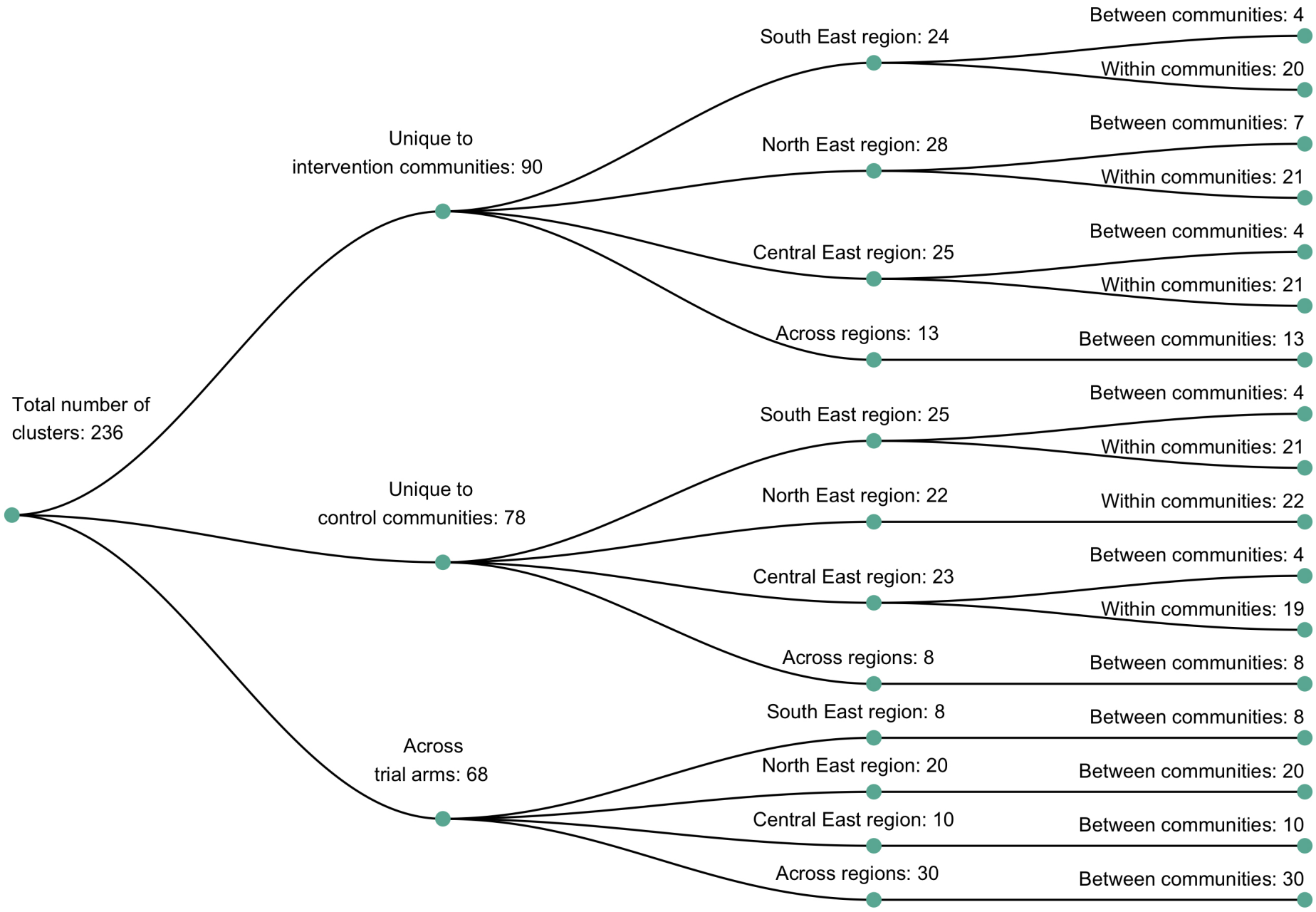


Figure 2. A dendogram showing clusters of genetically similar HIV-1 infections in the Botswana -Ya Tsie trial. Genetic clusters are summarized by randomized-HIV-intervention condition, geographical region and occurrence within and between trial communities. Clusters were identified from 3,832 HIV-1 viral whole genome consensus sequences as two or more sequences separated by a genetic distance not exceeding 0.045 (4.5%) substitutions per site and a bootstrap support threshold of at least 80%. Most genetic clusters occurred within trial communities and geographical regions than between them.

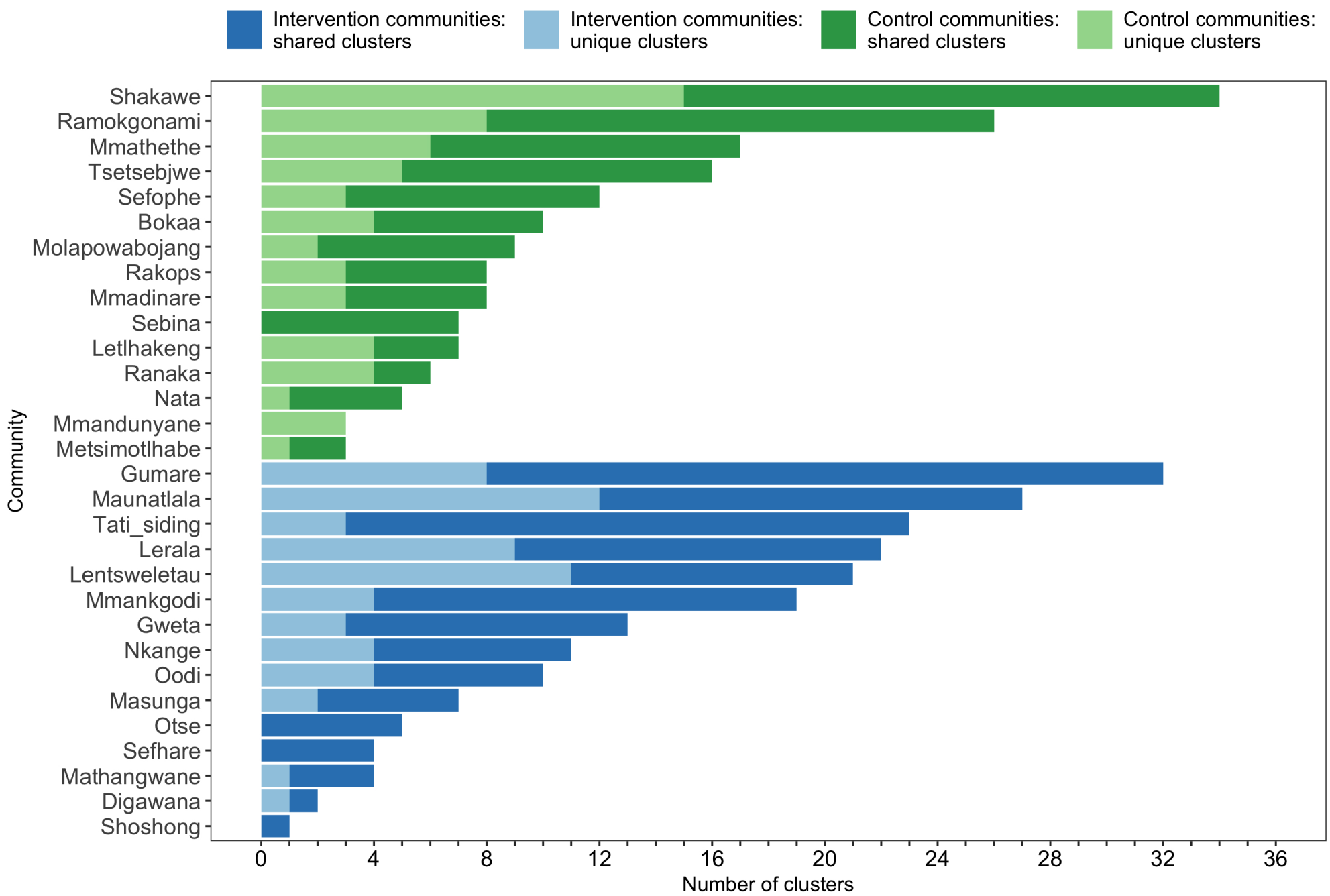


Figure 3. A barplot summarizing clusters of genetically similar HIV-1 infections by trial community. Bars represent the total number of genetic clusters identified in each trial community partitioned into clusters that are specific to a community (lighter shade) and those that are shared with at least one other trial community (darker shade). Genetic clusters in intervention communities are shown in blue and those in control communities are represented in green. The number of genetic clusters identified among intervention and control communities varied.

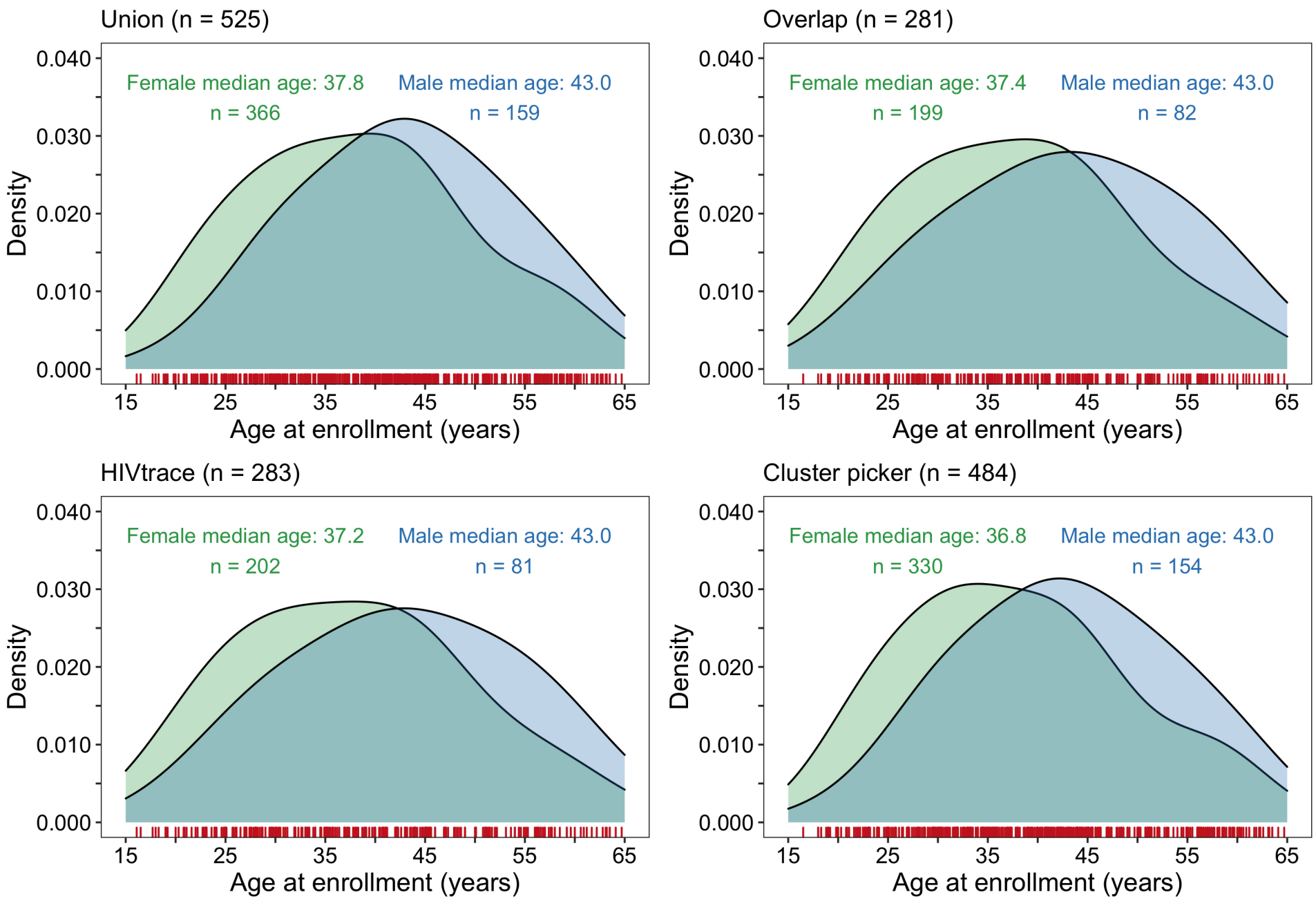


Figure 4. Density plots summarizing trial participants in clusters of genetically similar HIV-1 infections by age and gender. The bottom two plots show the age-gender distribution of trial participants whose HIV-1 viral whole genome consensus sequences were in genetic clusters identified with HIV-TRACE (n = 283) and Cluster Picker (n = 484) clustering algorithms. The top two panels show the age-gender distribution of trial participants whose HIV-1 viral whole genome consensus sequences were in the union (n = 525) and overlap (n = 281) of genetic clusters identified with HIV-TRACE and Cluster Picker. The raw data is shown as a rug plot (red tick marks) on the x-axis. Men and women in genetic clusters had similar ages.

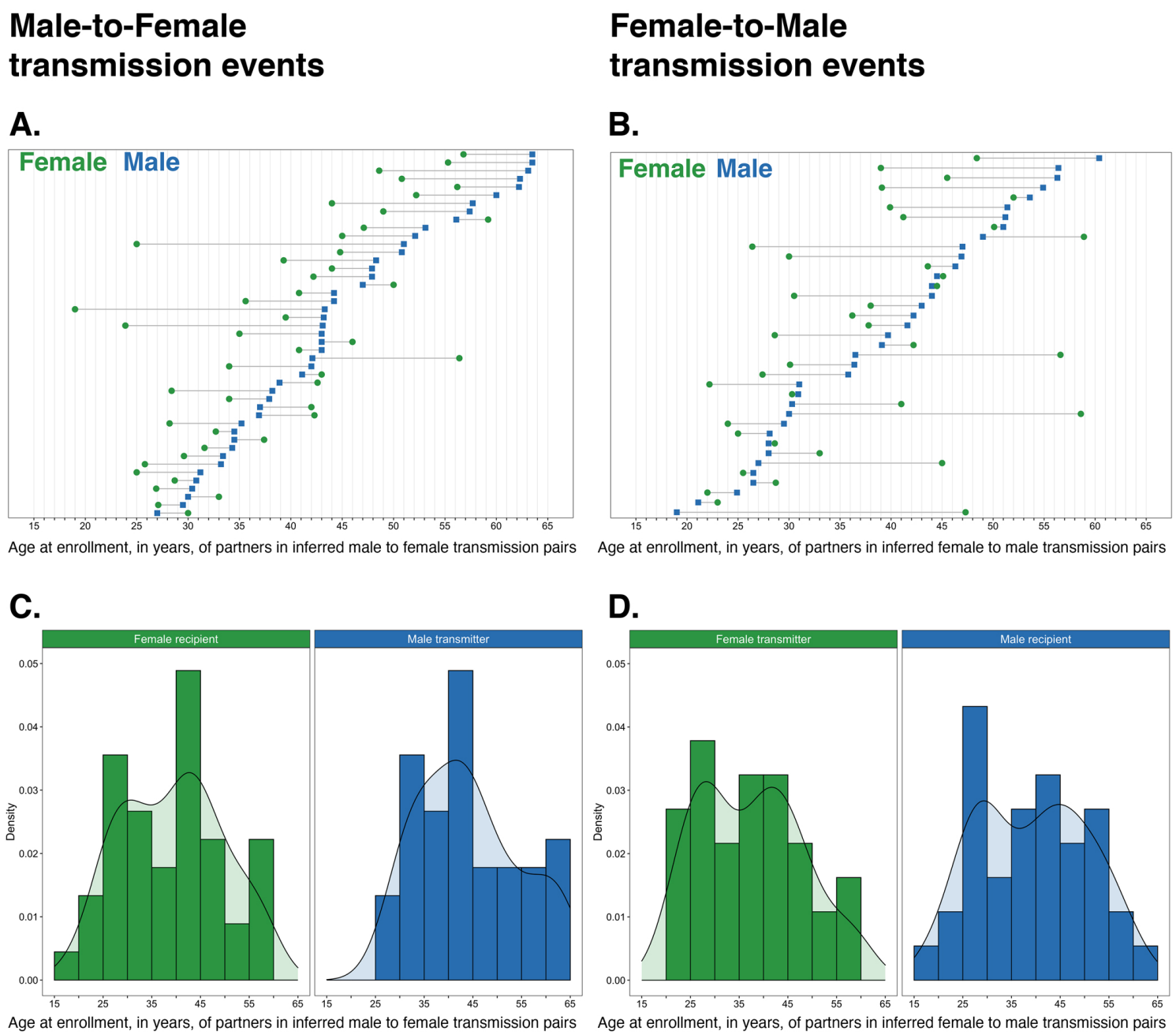


Figure 5. Age distribution of 82 male-female probable transmission pairs sampled in the Botswana -Ya Tsie trial. Forty-five male-to-female and 37 female-to-male transmission events were identified from the deep-sequenced viral whole genomes of 525 trial participants in clusters of genetically similar HIV-1 infections. Men were typically older than women in both male-to-female and female-to-male transmission events.

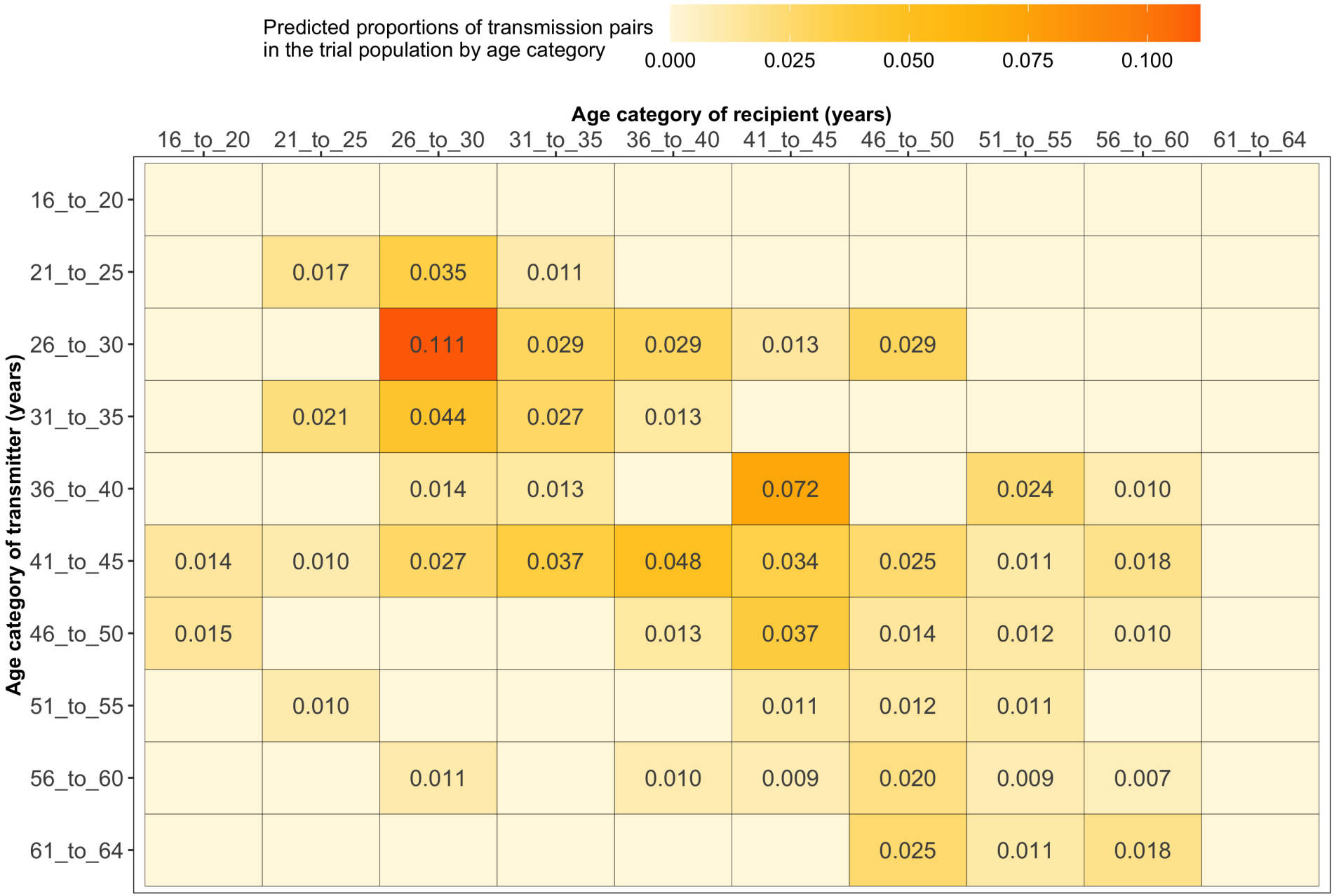


Figure 6. Estimated transmission flows of HIV-1 infection within and between 5 year age groups in the Botswana -Ya Tsie trial population. Transmission flows of HIV-1 infection in the trial population were estimated from 82 male-female probable transmission pairs identified from the deep-sequenced HIV-1 viral whole genomes of trial participants (n = 525) in clusters (n = 236) of genetically similar HIV-1 infections. Furthermore, transmission flows were adjusted for differential sampling among age groups (Supplementary Note 4). Most viral transmission events occurred between similarly aged partners.

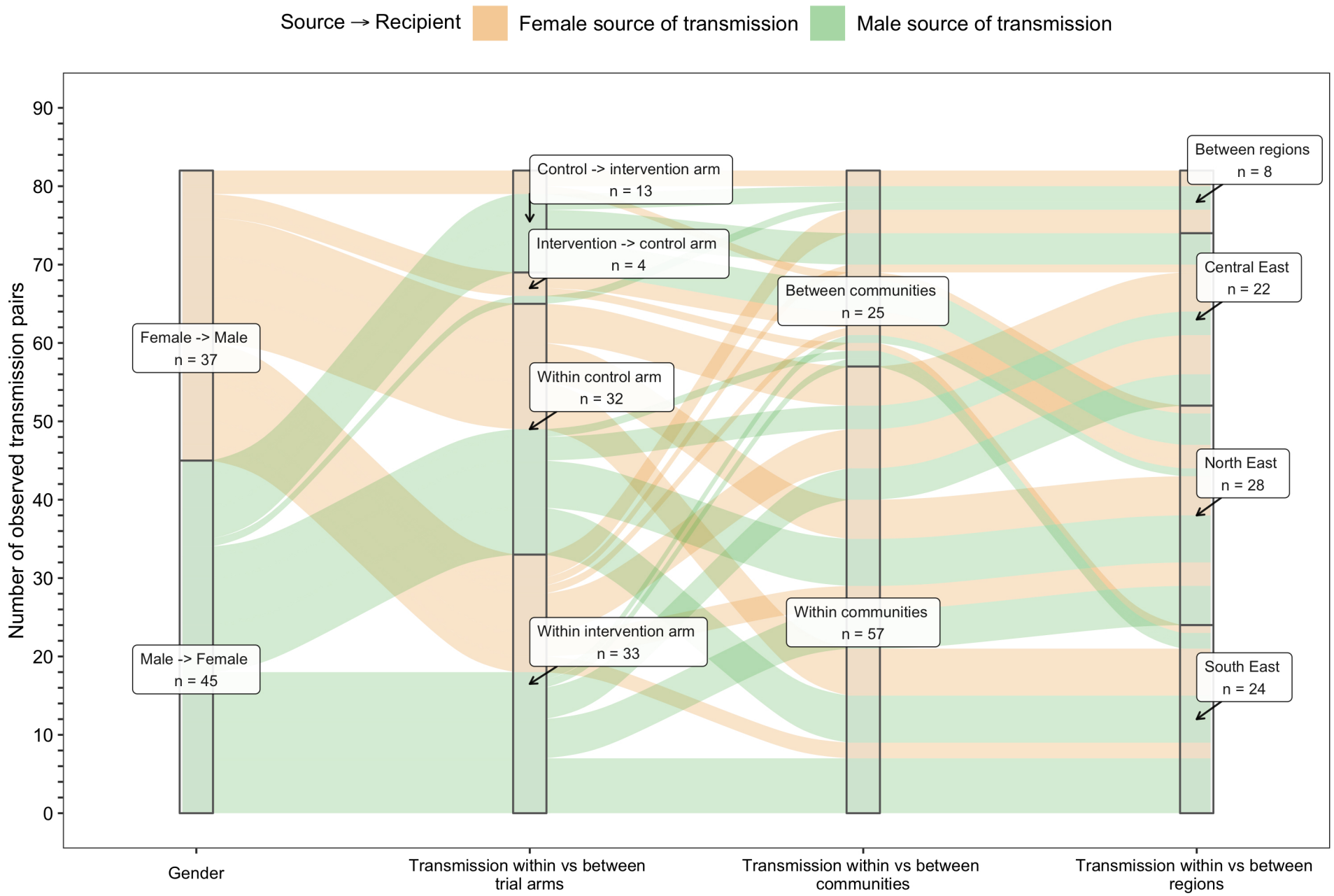


Figure 7. An alluvial plot showing counts of male-female probable transmission pairs sampled in the Botswana -Ya Tsie trial by geographical location. The 82 probable male-female infector -infectee pairs that were identified from clusters of genetically similar HIV-1 infections (n = 236) in the Botswana -Ya Tsie trial are first summarized into those with a male versus a female source of viral transmission. Male-to-female and female-to-male transmission events are then further partitioned according to whether partners in a transmission pair resided in the same or different trial communities and if those trial communities belonged to the same or different trial arms and geographical regions. Male-to-female transmission events are shown in green and female-to-male transmission events are represented in yellow. About 69% (n = 57) of sampled transmission events in the Botswana -Ya Tsie trial occurred within the same trial community.

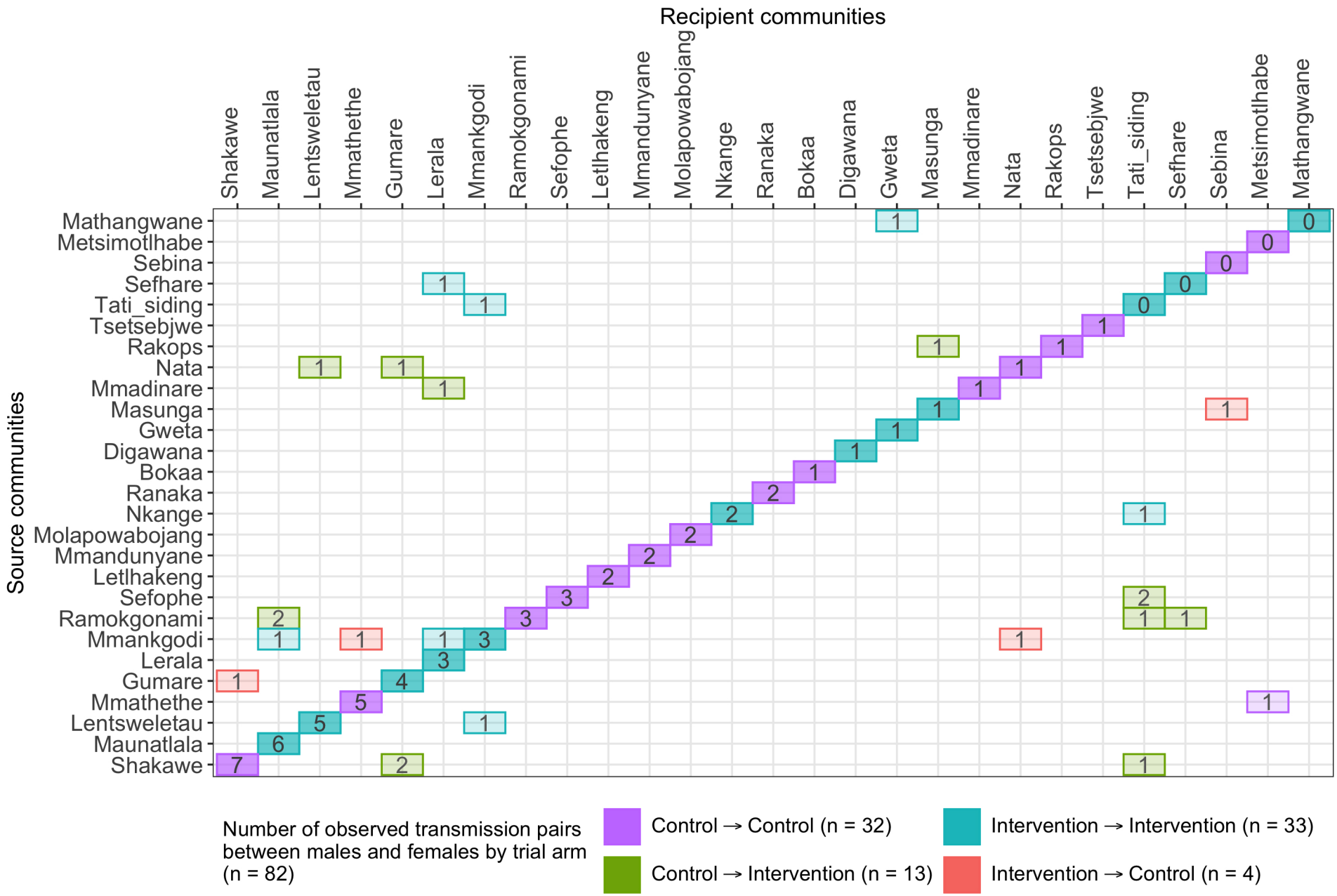


Figure 8. A transmission matrix summarizing sampled transmission events within and between Botswana -Ya Tsie trial communities. Numbered tiles represent counts of transmission events identified between a pair of trial communities. There were 32 transmission events identified within control communities (magenta), 33 within intervention communities (cyan), 13 from control-to-intervention communities (green) and 4 from intervention-to-control communities (red). Most transmission events were identified within trial communities. The number of sampled transmission events into intervention communities from control communities was higher than the reverse (n = 13 vs 4).

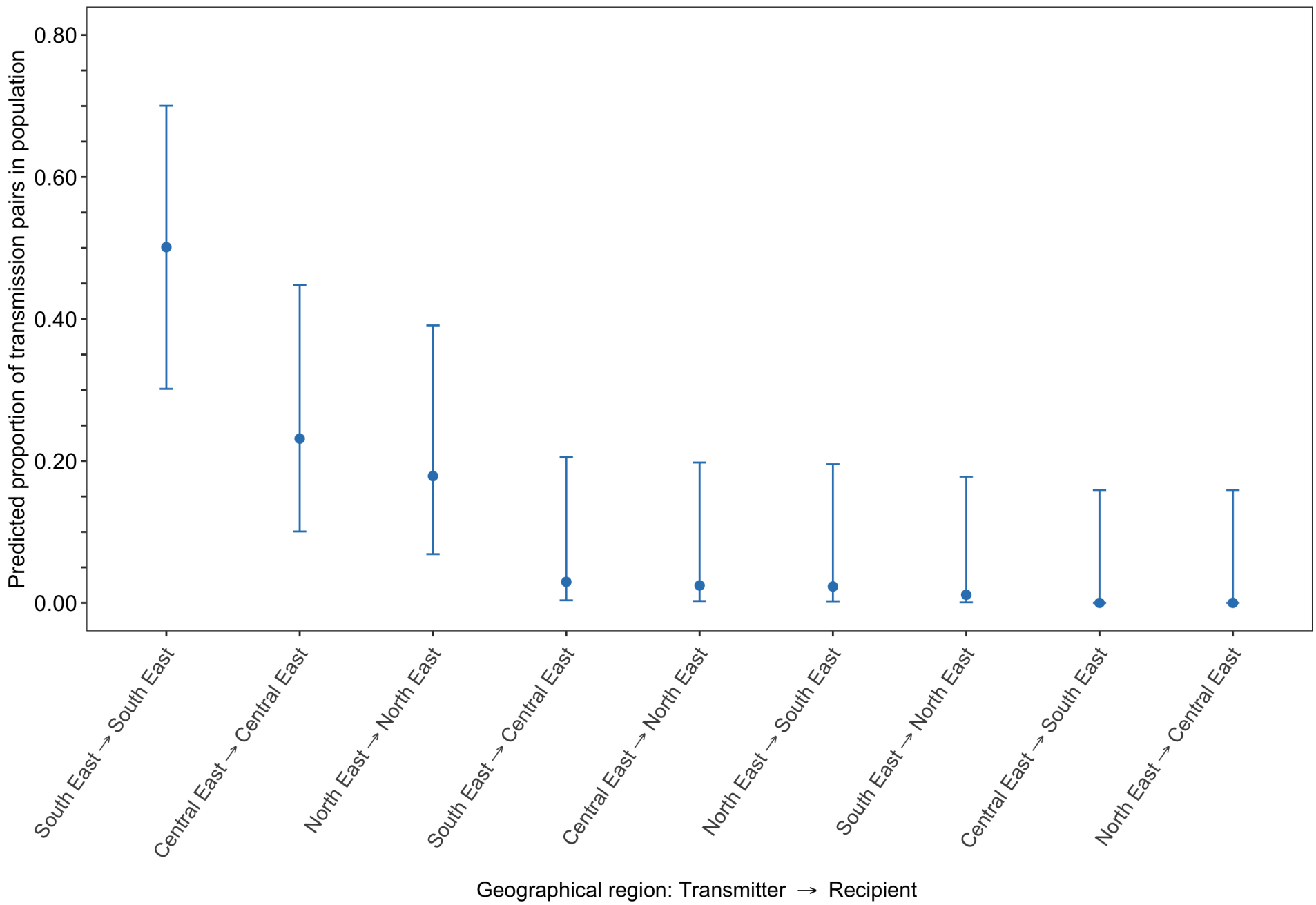


Figure 9. Estimated transmission flows of HIV-1 infection in the Botswana -Ya Tsie trial population within and between geographical regions. Communities in the Botswana -Ya Tsie trial can be broadly grouped into three geographical regions in the Central-East, North/North-East and South-Eastern parts of the country. Transmission flows were estimated as described in Figure 6. Moreover, transmission flows were adjusted for sampling heterogeneity among geographic regions. The flow of HIV-1 transmissions in the trial population was highest in the South-East region (50.1%) compared with the Central-East (23.1%) and North/North-East (17.9%) regions respectively.

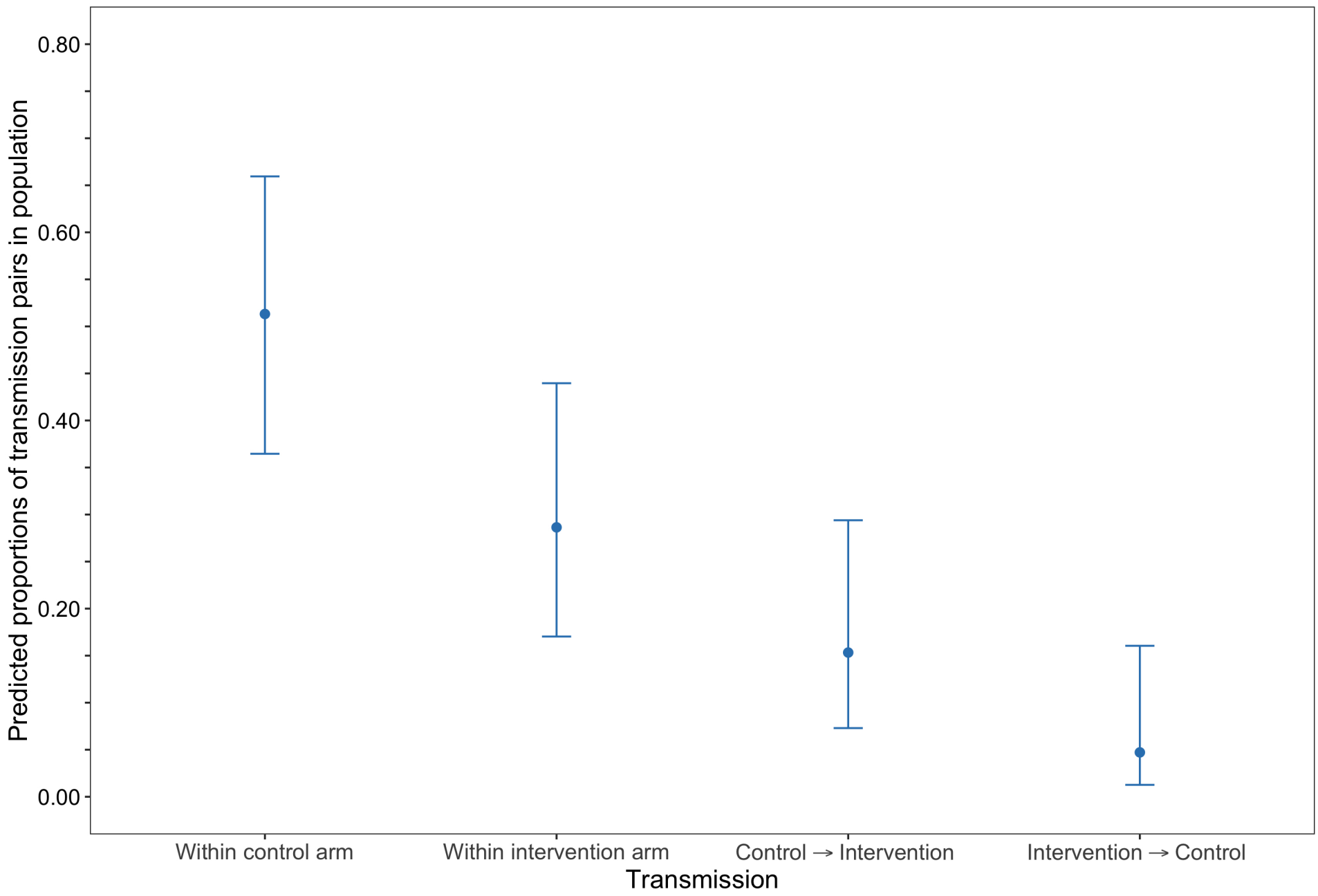


Figure 10. Estimated transmission flows of HIV-1 infection within and between intervention and control communities in the Botswana -Ya Tsie trial population. Transmission flows were estimated as described in Figure 6. Additionally, transmission flows were adjusted for differential sampling among trial arms. Most transmissions of HIV-1 infection occurred within the same trial arm and the flow of viral transmissions into intervention communities from control communities (15.3%) was about three times higher than the reverse (4.7%).

Supplemental information includes seven supplementary notes, nine supplementary figures and twelve supplementary tables.

Supplementary Note 1. Computing age-gender estimates of the number of people with HIV in each trial community

Age-gender HIV prevalence estimates in each trial community. To obtain HIV prevalence estimates we first fit a random-intercept Poisson regression model of HIV status, with age and gender as fixed-effects, using non-parametric maximum likelihood estimation in the gllamm program [1, 2] in Stata 13.1. The model was fit to data indicating the HIV status of 12,570 participants that had consented to an HIV test and were part of a baseline survey of 20% of households randomly sampled from each of the 30 trial communities to establish an incidence follow-up cohort; the fraction of positive tests for HIV-1 infection was, 28.6% (3,596 / 12,570). Ages of participants were grouped in two ways, as 5-year age-categories and four age-categories: 16-24, 25-34, 35-49 and 50-64 years. Furthermore, the counts of people with HIV for a specific age-gender grouping, i, in community, j were assumed to have a Poisson distribution with mean, u_{ij} so that

$$\ln(u_{ij}) = \ln(e_{ij}) + \beta_1 + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \zeta_j \ . \tag{1}$$

Accordingly, β_1 denotes the intercept, ζ_j the predicted random-intercept representing unobserved heterogeneity across trial communities, $\beta_2 x_{2ij}$ and $\beta_3 x_{3ij}$ fixed-effect covariates for age-group and gender respectively, and $\ln(e_{ij})$ an offset where e_{ij} represents the number of people tested for HIV in a specific age-gender grouping, i, in community, j (i.e. exposed) so that,

$$\ln(u_{ij}) - \ln(e_{ij}) = \beta_1 + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \zeta_j \ . \tag{2}$$

Computing empirical Bayes estimates of HIV prevalence. Next, to account for heterogeneity in HIV prevalence across trial communities we computed empirical Bayes estimates (or posterior means) for the age-gender specific HIV prevalence in each community. The empirical Bayes estimate of HIV prevalence for the jth community-age-gender combination is

(3) $$\hat{\zeta}_j^{EB} = \hat{R}_j \hat{\zeta}_j^{ML},$$

where, $\hat{R}_j = \frac{\hat{\psi}_j}{\hat{\psi}_j + \frac{\hat{\theta}}{\hat{n}_j}}$.

Here, $\hat{R}_j$ denotes a shrinkage factor that describes the proportion of the total variation, $\hat{\psi}_j + \frac{\hat{\theta}}{\hat{n}_j}$ attributed to differences in the age-gender specific HIV prevalence among trial communities, $\hat{\psi}_j$. Furthermore, $\hat{\zeta}_j^{ML}$ specifies the maximum likelihood estimate of ζ_j . Empirical Bayes estimates were obtained with the gllapred program [1, 2] in Stata 13.1.

Estimates of size of population with HIV in trial communities. We thereafter estimated the size of the population with HIV for each age-gender category in each trial community. Empirical Bayes estimates of the age-gender specific HIV prevalence for each trial community were standardized to the proportion of participants in the 2011 Botswana population and housing census that were in the same age and gender categories of these communities. Then for a specific community-age-gender combination, for example, females aged 16-24 years in community j , the estimated number of people with HIV, $\hat{N}_{HIV+\,[Female,\ 16-24]}$ is computed as

45 (4) $\widehat{N}_{HIV+\,[Female,\ 16-24]} = \widehat{N}_{total}\ x\ E\ x\ \{\ Pr_{census}[\ \text{Female}, 16-24\ \text{years}\ ]\ x\ Pr[\ \text{HIV}+$

46 $|\ \text{Female}, 16-24\ \text{years}\ ]\ \}$

47 where,

48 $\widehat{N}_{total}$: Estimated population size for the jth trial community based on plot and household

49 enumeration data [3],

50 E: Proportion of enumerated household members that were eligible to participate in the trial,

51 $\text{Pr}_{census}[\text{Female}, 16\ -\ 24\ \text{years}]$: Joint probability denoting the proportion of census

52 participants who are female and aged 16-24 years within community j

53 $\text{Pr}[\text{HIV}+\ |\ \text{Female}, 16-24\ \text{years}]$: Empirical Bayes estimate of HIV prevalence for females

54 aged 16-24 years that participated in a baseline survey of 20% of households randomly sampled

55 from community j.

56 The age-gender estimates of the number of people with HIV in each trial community were used

57 to adjust for sampling heterogeneity when estimating the proportions of HIV transmissions in

58 the trial population (see supplementary note 4). Population-size estimates of people with HIV

59 were computed in R version 3.5.2 [4].

60 **Supplementary Note 2.1. Paired-end deep-sequencing of HIV viral genomes for phylogenetic**
61 **analyses.** Paired-end deep-sequencing was performed with Illumina MiSeq and HiSeq
62 instruments at two locations, the BioPolymers sequencing core facility at Harvard Medical
63 School, Boston, United States of America (n = 4,246 participants) (S1 Fig), and at the Wellcome
64 Trust Sanger Institute, Hinxton, United Kingdom (n = 868 participants) (S2 Fig) through the
65 PANGEA consortium [5]. The sequencing success rate was relatively high, 75-80% for first round
66 amplicons and 95-98% for second round amplicons. The shiver sequence assembly software [6]
67 was used to assemble and map each participant's deep-sequencing short reads to a de-novo
68 reference sequence tailored to the participant's viral population.
69
70 **Supplementary Note 2.2. Comparing genetic distances between HIV-1 viral consensus**
71 **sequences of trial participants included in phylogenetic analyses.** We compared the distances
72 between all possible pairs (n = 7,340,196) of the 3,832 HIV-1 viral whole genome consensus
73 sequences available for phylogenetic analyses (S4 Fig). Sequences were codon-aligned to HXB2,
74 a standard HIV reference, with MAFFT v7.407 [7, 8] and HIVAlign
75 (https://www.hiv.lanl.gov/content/sequence/VIRALIGN/viralign.html); genetic distances were
76 then computed under the Tamura-Nei-1993 nucleotide substitution model [9] with TN93 v1.0.6
77 (github.com/veg/tn93).
78
79
80

Supplementary Note 2.3. Consensus sequence phylogenetics to infer clusters of genetically similar HIV-1 infections. HIV Transmission Cluster Engine (HIV-TRACE) v0.4.4 [10, 11] and cluster picker v1.2.3 [12] algorithms, were used to identify clusters of individuals whose HIV-1 viral whole genome consensus sequences were genetically similar -- suggesting they were probably members of a transmission chain [13, 14]. HIV-TRACE defines clusters based on pairwise genetic distances only; comparatively, cluster picker identifies clusters using pairwise genetic distances with the guidance of a phylogenetic tree. A codon-based multiple sequence alignment as described in Supplementary Note 2.2 was provided as input to HIV-TRACE and cluster picker. Additionally, for cluster picker, a corresponding phylogenetic tree inferred with FastTree2 v2.1.10 and boot-strap support values approximated with the Shimodaira-Hasegawa test [15] were provided as inputs. We defined genetic similarity clusters as groups of two or more participants whose viral consensus sequences were separated by a genetic distance at or smaller than 4.5% nucleotide substitutions per site--and, for cluster picker, a bootstrap support value of at least 80%. The genetic distance threshold of 4.5% nucleotide substitutions per site was motivated by the distribution of genetic distances separating HIV-1 subtype C viral genomes of epidemiologically-linked couples in the HIV Prevention Trials Network (HPTN) 052 trial (S5 Fig) [16, 17].

Supplementary Note 2.4. Deep-sequence phylogenetics to infer the probable order of transmission events within identified clusters of genetically similar HIV-1 infections. We performed parsimony-based ancestral host state reconstruction of the mapped deep-sequencing short reads of participants in clusters of genetically similar HIV-1 infections (n = 525) with the phyloscanner software v1.8.0 [18-20] to identify probable transmission pairs within clusters and the probable direction of transmission between them. Phyloscanner identifies potential transmission pairs by inferring whether the viral population in an individual is ancestral to or descendent from that of another individual, and crucially addresses contamination by excluding duplicates and phylogenetic outliers. To infer phylogenetic linkage and direction of transmission for a set of individuals phyloscanner aligns submitted mapped reads in sliding windows along the genome that are matched across individuals; and at each window, infers a phylogeny for ancestral state reconstruction to identify probable ancestral relationships between all possible pairs of individuals in the set. Pairs of individuals are then classified as phylogenetically-linked or unlinked based on the distance, adjacency and topology of their subgraphs. A subgraph in phyloscanner refers to all tips and internal nodes of a phylogeny assigned to an individual through parsimony-based ancestral state reconstruction. A pair of individuals, i and j is considered to be phylogenetically-linked, more specifically, phylogenetic linkage is not excluded when the minimum patristic distance between their subgraphs, Δ_{ij} is below a set threshold, and i and j are adjacent meaning that the shortest path connecting their subgraphs in a phylogeny does not pass through a third individual. Tree topology is used to infer the probable direction of transmission between a pair of individuals and refers to the number of subgraphs belonging to each individual in the pair that are

ancestral-to or descendant-from those of the other individual. We classified pairs of individuals separated by a subgraph distance below 0.035 substitutions per site as phylogenetically-linked based on previous work by Ratmann and colleagues [19]. The subgraph distance distribution of all pairs identified with phyloscanner within genetic similarity clusters is shown in S6 Fig. A listing of command-line parameters used for deep-sequence phylogenetics with phyloscanner is provided in S3 Table. Identified genetic similarity clusters were analyzed with phyloscanner in parallel for computational efficiency using shell scripts that are available upon request.

131 **Supplementary Note 2.5. Identifying probable source-recipient pairs with strong phylogenetic**
132 **evidence for linkage and direction of transmission**. Probable transmission pairs identified with
133 phyloscanner were further classified on the strength of evidence for phylogenetic linkage, $\hat{\lambda}_{ij}$ to
134 separate pairs with strong support for linkage and direction of transmission from other
135 potentially linked pairs; accounting for the extent of overlap in the read alignments from which
136 the deep-sequence phylogenies were inferred. The strength of phylogenetic evidence or
137 linkage score is described as, $\hat{\lambda}_{ij} = \frac{k_L}{n}$ where n is the number of deep-sequence phylogenies
138 inferred in windows along the genome for which individuals i and j had sufficient deep-
139 sequence reads for phylogenetic inference; and k_L is the number of deep-sequence
140 phylogenies along the genome that support a specific phylogenetic relationship type between
141 individuals i and j, for example, that the subgraphs of individual i are ancestral-to, descendent-
142 from, sibling-to or intermingled-with those of individual j; sibling and intermingled topologies
143 are ones where there is insufficient evidence to make inferences about the probable direction
144 of transmission [19]. We used a linkage and direction of transmission score threshold of 57%
145 $(\hat{\lambda}_{ij} > c;\ c = 0.57)$ or on average at least 24 out of 42 windows supporting linkage and
146 direction of transmission for a highly supported transmission pair. Our phyloscanner analyses
147 were based on windows or mapped read alignments that were 200 bp long, affording n = 42
148 (8,400 / 200) non-overlapping alignments and deep-sequence trees along the genome; we
149 excluded the first 1,000 nucleotides of the HIV-1 genome from analysis as they are typically
150 poorly sequenced corresponding to a genome length of 8,400 bp. The linkage and direction of
151 transmission score threshold, $c \in (0,1)$ was selected such that the posterior probability for

$\hat{\lambda}_{ij} > 50\%$, at least half of the windows along the genome supporting an ancestral relationship type between a pair of individuals, i and j exceeds 80%, $p(\hat{\lambda}_{ij} > 0.5 \mid k_L, n) > \alpha$, where $\alpha = 80\%$ and $k_L \sim \mathrm{Binomial}(n, \hat{\lambda}_{ij})$. The linkage and direction of transmission score threshold was computed with the Bayesian binomial test function in the BayesianFirstAid R package v0.1.

160 **Supplementary Note 3. Estimating error rates in phylogenetic inference of direct transmission**

161 **between sampled males and females**

162

163 HIV-1 infections in southern Africa are more commonly transmitted through heterosexual

164 contact between males and females, compared with sexual contact between same-sex couples.

165 Thus, we can use the probability of inferring a phylogenetically-linked male-male pair to

166 calibrate an upper bound on the false discovery rate of inferring a phylogenetically-linked male-

167 female pair, if we assume the probability of erroneously inferring direct transmission between a

168 male-female pair is the same as the probability of inferring a phylogenetically-linked male-male

169 pair [19].

170

171 **Calibrating a false discovery rate for inferring direct transmission.** Consider a deep-sequence

172 phylogenetic analysis where L_{mm} male-male pairs and L_{mf} male-female pairs were

173 phylogenetically inferred in the sequenced sample. For S_m successfully sequenced males and

174 S_f successfully sequenced females, let the probability of inferring a phylogenetically-linked

175 male-male pair in the sample be

176

177 (5) $$\frac{L_{mm}}{S_m \, x \, (S_m - 1)/2},$$

178

179 where L_{mm} denotes the number of phylogenetically-linked male-male pairs that were

180 identified and $S_m \, x \, (S_m - 1)/2$ represents the number of distinct possible male-male pairs

that could have been identified in the sample. The estimated number of phylogenetically-linked male-female pairs between whom direct transmission did not occur would then be

$$\hat{F}^{c}_{mf} = \frac{L_{mm}}{S_m \, x \, (S_m - 1)/2} \, x \, S_m \, x \, S_f \, . \tag{6}$$

Thereby affording a false discovery rate estimate of

$$\hat{\rho}^{c}_{mf} = \frac{\hat{F}^{c}_{mf}}{L_{mf}} \, . \tag{7}$$

Supplementary Note 4. Estimation of the proportions of HIV transmission flows in the trial population adjusting for sampling heterogeneity.

Dataset and sampling probability. Consider a population of individuals with HIV, Ω of finite size, N partitioned into $u = 1, \cdots, G$ disjoint groups so that, N_u represents the number of individuals with HIV in group, u. Groups might represent communities (or geographic locations), age-categories, genders or trial arms. Let us suppose that we would like to estimate the proportions of HIV transmissions that occurred within and between intervention and control communities in the Botswana / Ya Tsie trial so that group, u are individuals residing in intervention communities and group, v are individuals residing in control communities. We denote n_u and n_v as the number of individuals randomly sampled from groups, u and v respectively whose HIV-1 viral genomes were successfully sequenced; yielding a sampling probability, $s_u = n_u/N_u$ for group, u and similarly, $s_v = n_v/N_v$, for group v. The vector

$$z = (z_{uu}, z_{uv}, z_{vu}, z_{vv}) \tag{8}$$

describes the counts of directed (source-recipient) transmission pairs identified in the sequenced sample within and between intervention and control communities, where z_{uu} denotes transmission pairs within intervention communities; z_{vv} transmission pairs within control communities; z_{uv} transmission pairs for whom a member of an intervention community was the inferred source of transmission to a control community member and vice-versa, z_{vu}. We assume that HIV transmission pairs that are absent from the sequenced sample are missing

212 at random conditional on group membership; meaning that the identified probable source-
213 recipient transmission pairs are a random sample of the source-recipient transmission pairs
214 present in the trial population. Let
215
216 (9) $$n_{uv} = \begin{cases} n_u \; x \; n_v, & u \neq v \\ n_u(n_u - 1) \; / \; 2, & u = v \end{cases}$$
217
218 be the number of distinct possible transmission pairs among individuals sampled from u and v;
219 similarly, let $N_{uv} = N_u \; x \; N_v$ (or $N_u(N_u - 1) \; / \; 2 \; if \; u = v$) be the number of distinct
220 transmission pairs available between trial population groups, u and v.
221
222 **Probability of viral linkage between a pair of sampled individuals.** Estimates of
223 probabilities of linkage between two individuals randomly sampled from within demographic
224 groups are observed using the method of Carnegie and colleagues [21]. The method assumes
225 that samples are missing at random from within demographic groups and that relative increase
226 in weight of individuals from a given group between the observed and true sample is the same
227 as relative increase in weight between sub-samples of the observed data—obtained from the
228 observed sampling rate—and the observed data.
229 In accordance with [21] and accounting for the probable direction of transmission, we
230 estimate the probability of viral linkage between two individuals randomly sampled from their
231 respective groups for each group-pairing as
232
233 (10) $$\hat{p} = (\hat{p}_{uu}, \hat{p}_{uv}, \hat{p}_{vu}, \hat{p}_{vv})$$

where $\hat{p}_{ij} = \frac{z_{ij}}{n_{ij}}$, $i = u, v$ and $j = u, v$.

Proportion of HIV transmissions in the trial population adjusted for sampling heterogeneity. We next estimate the relative probability that an HIV transmission event in the trial population occurred from sexual contact between partners residing within intervention communities or within control communities; compared with partners where one resides in an intervention community and the other a control community or vice versa. Accordingly, we estimate the conditional probability, $\hat{\theta}_{ij}$ that a pair of individuals are from a specific group-pairing given that their HIV-1 viral genomes are linked as,

$$\hat{\theta}_{ij} = P(\text{pair is from groups, } i \text{ and } j \mid \text{pair is linked}) \tag{11}$$

$$= \frac{N_{ij}\hat{p}_{ij}}{\sum_m \sum_{n \geq m} N_{mn}\hat{p}_{mn}} \text{ , } i = u, v \text{ and } j = u, v \text{ .}$$

Here, $N_{ij}\hat{p}_{ij}$ denotes the estimated number of HIV transmissions in the trial population attributed to groups, i and j. Similarly, $\sum_m \sum_{n \geq m} N_{mn}\, \hat{p}_{mn}$ represents the number of HIV transmissions in the trial population for all groups-pairings. Thus, the vector

$$\hat{\theta} = (\hat{\theta}_{uu}, \hat{\theta}_{uv}, \hat{\theta}_{vu}, \hat{\theta}_{vv}) \tag{12}$$

254 describes estimated proportions of HIV transmissions in the trial population within and
255 between intervention and control communities adjusted for differential sampling by trial arm.
256
257 **Estimation of confidence intervals around proportions of HIV transmissions in trial**
258 **population.** We use the Goodman method (1965) [22] to compute simultaneous confidence
259 intervals at the 5% significance level for the estimated proportions of HIV transmissions within
260 and between population groups (or strata). A continuity correction factor is implemented for
261 the Goodman's confidence intervals to account for small sample sizes [23]. The Goodman
262 method assumes proportions for which confidence intervals are computed are multinomially
263 distributed and thus independence of observations. Among the 82 probable male-female
264 transmission pairs identified from the deep-sequenced sample there was a single individual
265 involved in multiple transmission events; a man who was the probable source of viral
266 transmission for two women each residing in different communities from his own. This
267 suggested that only a small degree of bias in estimation of confidence intervals would arise
268 from a simplifying assumption of independence between inferred transmission events; and
269 allowed us to estimate confidence intervals using methods for multinomial proportions.
270
271 To estimate Goodman's confidence intervals the estimated number of HIV transmission
272 pairs in the trial population within and between intervention and control communities are
273 scaled such that the sum of the weights is equal to the total number of probable male-female
274 transmission pairs identified in the deep-sequenced sample (n = 82), and the weights are
275 treated as known; broadly similar to calibration of weights in survey sampling. For example, the

weighted counts of probable male-female transmission pairs within and between intervention and control communities are given by the vector,

$$\delta = z^{+}\, x\left(\hat{\theta}_{uu}, \hat{\theta}_{uv}, \hat{\theta}_{vu}, \hat{\theta}_{vv}\right), \tag{13}$$

$$= \left(z^{+}\hat{\theta}_{uu}, z^{+}\hat{\theta}_{uv}, z^{+}\hat{\theta}_{vu}, z^{+}\hat{\theta}_{vv}\right)$$

where $z^{+} = \sum_{ij} z_{ij}$, $i = u, v$ and $j = u, v$.

This ensures that the estimated Goodman's confidence intervals of HIV transmission flows within and between groups/strata in the trial population reflect that amount of information in the sample.

Let the weighted counts of the probable male-female transmission pairs identified within and between intervention and control communities, $\left(z^{+}\hat{\theta}_{uu}, z^{+}\hat{\theta}_{uv}, z^{+}\hat{\theta}_{vu}, z^{+}\hat{\theta}_{vv}\right)$ denote the observed cell frequencies, $\delta_1, \delta_2, \cdots, \delta_k$ of a sample of size, z^{+} from a multinomial distribution with population parameters, $\pi_1, \pi_2, \cdots, \pi_k$ and $l = 1, 2, \cdots, k$ population strata.

For example, δ_l represents the weighted counts of probable male-female transmission pairs identified in the lth stratum, and π_l the corresponding probability that a probable male-female transmission pair falls within the lth stratum. In this case, we have $k = 4$ population strata that represent HIV transmission pairs within intervention communities; within control communities; transmission pairs for whom a member of an intervention community was the inferred source of transmission to a control community member and vice-versa. The point

estimate for, π_l or estimated proportion of HIV transmissions in the trial population in the lth stratum is $\hat{\pi}_l = \frac{\delta_l}{z^+}$. The set of k simultaneous confidence intervals at the 5% significance level for the k population-level HIV transmission flows, $\pi_1, \pi_2, \cdots, \pi_k$ are given by [23]

$$\hat{\pi}_l^- \leq \pi_l \leq \hat{\pi}_l^+ \text{ for } (l = 1, 2, \cdots, k),$$

where the lower confidence bounds are described by,

$$\hat{\pi}_l^- = \left[B + 2(\delta_l - 0.5) - \sqrt{B(B + 4(\delta_l - 0.5)(z^+ - \delta_l + 0.5)/z^+)}\right] / 2(z^+ + B)$$

and the upper confidence bounds are described by,

$$\hat{\pi}_l^+ = \left[B + 2(\delta_l + 0.5) + \sqrt{B(B + 4(\delta_l + 0.5)(z^+ - \delta_l - 0.5)/z^+)}\right] / 2(z^+ + B).$$

Here, B denotes the 95th percentile of a chi-squared distribution with $k - 1$ degrees of freedom. Note that $\hat{\pi}_l^- = 0 \; if \; \delta_l = 0$, and $\hat{\pi}_l^+ = 1 \; if \; \delta_l = z^+$.

We have implemented the algorithms to estimate HIV transmission flows within and between population groups accounting for sampling heterogeneity; and the corresponding confidence intervals as an R package, *bumblebee* that will be made available at the following url: https://magosil86.github.io/bumblebee.

Supplementary Note 5. Calculation of the weighted mean age gap between males and females in inferred probable transmission pairs accounting for variability in sampling of the trial population by gender and 5-year age-group.

Overview: To compare the ages of males and females in inferred male-female probable transmission pairs we computed weighted (arithmetic) mean age-gaps between males and females in 1) all inferred probable transmission pairs (n = 82), 2) male-to-female transmission events (n = 45) and 3) female-to-male transmission events (n = 37).

Dataset and sampling probabilities: Consider a deep-sequence phylogenetic analysis where L_{mf} male-female probable transmission pairs/events were phylogenetically inferred in the sequenced sample. Here the sequenced sample refers to individuals whose HIV-1 viral genomes were successfully sequenced and met minimum criteria for inclusion in phylogenetic analysis. Let individuals in the sequenced sample be grouped by gender into 5-year age-categories according to their age at enrollment in the study so that the number of successfully sequenced males in a specific 5-year age category is denoted by, $n_{m(\text{5-year age group})}$ and the number of successfully sequenced females, $n_{f(\text{5-year age group})}$. Similarly, let $N_{m(\text{5-year age group})}$ and $N_{f(\text{5-year age group})}$ denote the estimated number of males and females respectively with HIV in the trial population in a specific 5-year age category. So that the sampling probabilities of male and female partners in a probable transmission pair are given by, $s_{m(\text{5-year age group})} = n_{m(\text{5-year age group})}/N_{m(\text{5-year age group})}$ for the male partner and, $s_{f(\text{5-year age group})} = n_{f(\text{5-year age group})}/N_{f(\text{5-year age group})}$ for the female partner.

Computation of mean weighted age gap: Let g_i denote the age gap in years between male and female partners in the ith transmission pair (i.e. male age – female age) and w_i represent the weight assigned to the ith transmission pair. The weight of a male-female probable transmission pair is computed as the inverse of the product of the sampling probabilities of the male and female partners in the pair,

(1) $$w_i = \frac{1}{s_{m(\text{5-year age group})}\ x\ s_{f(\text{5-year age group})}}.$$

Then the weighted mean age gap is given by

(2) $$\bar{g} = \frac{\sum_{i=1}^{n} w_i\ x\ g_i}{\sum_{i=1}^{n} w_i},$$

which can be expanded as:

(3) $$\bar{g} = \frac{w_1 g_1 + w_2 g_2 + \cdots + w_n g_n}{w_1 + w_2 + \cdots + w_n}.$$

Estimation of the weighted age gap variance: We compute the variance of the age-gaps between males and females in probable transmission pairs as:

(4) $$Var(\text{weighted age gap}) = \frac{\sum_{i=1}^{n} w_i\,(g_i - \bar{g})^2}{V_1 - \left({}^{V_2}/_{V_1}\right)},$$

where $V_1 = \sum_{i=1}^{n} w_i$ represents the sum of the weights, and $V_2 = \sum_{i=1}^{n} w_i^2$ the sum of the squared weights.

356 **Supplementary Note 6. Permutation test to evaluate whether there was preferential sexual**
357 **mixing among trial communities by geographic proximity**
358
359 To evaluate whether people were more likely to form out-of-community sexual partnerships
360 with partners from nearby communities than communities further away, we performed a
361 permutation test under the null hypothesis that the mean travel distance (kilometers) and drive
362 time (hours) between communities where HIV transmission events were identified was no
363 different from that between any pair of randomly selected trial communities. We obtained null
364 distributions for the mean travel distance and drive time, respectively by permuting the order
365 of community pairs over 10,000 iterations. The underlying correlation structure of viral genetic
366 linkage was preserved, for example, transmission events that occurred between members of
367 two communities or between members of a community and two or more other communities. A
368 one-sided p-value was obtained as the fraction of iterations where the mean travel distance
369 between permuted community pairs was smaller than that for community pairs where HIV
370 transmission events were identified; similarly, a one-sided p-value was derived for travel time.
371 Travel distances and drive times were sourced from the google distance matrix API (application
372 programming interface) with the mapsapi R package v0.4.2. and will be made available as an R
373 dataset. The permutation-test was performed using the R statistical software version 3.5.2.
374

375 **Supplementary Note 7. Membership and affiliations of the Botswana Combination**
376 **Prevention Project and PANGEA consortium**

377

378 **Botswana Combination Prevention Project**
379

380 Myron Essex[1], Joseph Makhema[1], Shahin Lockman[1], Eric Tchetgen Tchetgen[1], Molly Pretorius

381 Holme[1], Ingrid Bassett[1], Kara Bennett[1], Victor DeGruttola[1], Scott Dryden-Peterson[1], Kenneth

382 Freedberg[1], Vladimir Novitsky[1], Kathleen Powis[1], Roger Shapiro[1], Rochelle Walensky[1], Rui

383 Wang[1], Kathleen Wirth[1], Marina Anderson[2], Shenaaz El-Halabi[2], Ernest B. Fetogang[2], Refeletswe

384 Lebelonyane[2], Madisa Mine[2], Tendani Gaolathe[3], Simani Gaseitsiwe[3], Poloko Kebaabetswe[3],

385 Mompati Mmalane[3], Mosepele Mosepele[3], Sikhulile Moyo[3], Rosemary Musonda[3], Erik van

386 Widenfelt[3], Pam Bachanas[4], Ebi Bile[4], Arielle Lasry[4], Tafireyi Marukutira[4], Lisa A. Mills[4], Janet

387 Moore[4], Sherri Pals[4], Elliot Raizes[4], Chunfu Yang[4]

388

389 Affiliations

390 1. Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA

391 2. Ministry of Health, Republic of Botswana, Gaborone, Botswana

392 3. Botswana Harvard AIDS Institute Partnership, Gaborone, Botswana

393 4. Division of Global HIV/AIDS and TB, Centers for Disease Control and Prevention, Atlanta,

394 Georgia, USA

395

396

397

398 **PANGEA Consortium:**

399

400 Helen Ayles[1], Rory Bowden[2], Vincent Calvez[3], Myron Cohen[4], Tulio D'Oliveira[5], Ann Dennis[6],

401 Max Essex[7], Sarah Fidler[8], Dan Frampton[9], Richard Hayes[10], Josh Herbeck[11], Pontiano Kaleebu[12],

402 Cissy Kityo[13], Jairam Lingappa[14], Vladimir Novitsky[15], Nick Paton[16], Andrew Rambaut[17], Janet

403 Seeley[18], Deogratius Ssemwanga[19], Frank Tanser[20]

404

405

406 Affiliations

407

408 1. London School of Hygiene & Tropical Medicine, London, UK,

409 2. University of Oxford, Oxford, UK

410 3. Ecole Normale Supérieure de Lyon, Lyon, France

411 4. University of North Carolina, Chapel Hill, NC, USA

412 5. University of KwaZulu-Natal, Durban, KwaZulu-Natal, South Africa

413 6. University of North Carolina, Chapel Hill, NC, USA

414 7. Harvard University, Boston, MA, USA

415 8. Imperial College London, London, UK

416 9. University College London, London, UK

417 10. London School of Hygiene & Tropical Medicine, London, UK

418 11. University of Washington, Seattle, WA, USA

419 12. MRC/UVRI Uganda, Kampala, Uganda

420 13. Joint Clinical Research Centre, Kampala, Uganda

421 14. University of Washington, Seattle, WA, USA

422 15. Africa Health Research Institute, Durban, KwaZulu-Natal, South Africa and University College
423 London, London, UK
424 16. Yong Loo Lin School of Medicine, Singapore, Singapore
425 17. University of Edinburgh, Edinburgh, UK
426 18. MRC/UVRI Uganda, Kampala, Uganda and London School of Hygiene & Tropical Medicine
427 London, UK
428 19. MRC/UVRI Uganda, Kampala, Uganda
429 20. Africa Health Research Institute, Durban, KwaZulu-Natal, South Africa
430

431 ## References

432

433 1. Sophia, R.-H., *GLLAMM: Stata program to fit generalised linear latent and mixed models*.
434 1999, Boston College Department of Economics.

435 2. Rabe-Hesketh, S. and A. Skrondal, *Multilevel and Longitudinal Modeling Using Stata, 3rd*
436 *Edition*. 2012: StataCorp LP.

437 3. Makhema, J., et al., *Universal Testing, Expanded Treatment, and Incidence of HIV*
438 *Infection in Botswana.* New England Journal of Medicine, 2019. **381**(3): p. 230-242.

439 4. R Core Team, *R: A language and environment for statistical computing.* 2020, R
440 Foundation for Statistical Computing: Vienna, Austria.

441 5. Pillay, D., et al., *PANGEA-HIV: phylogenetics for generalised epidemics in Africa.* The
442 Lancet. Infectious diseases, 2015. **15**(3): p. 259-261.

443 6. Wymant, C., et al., *Easy and accurate reconstruction of whole HIV genomes from short-*
444 *read sequence data with shiver.* Virus Evolution, 2018. **4**(1): p. vey007.

445 7. Katoh, K., et al., *MAFFT: a novel method for rapid multiple sequence alignment based on*
446 *fast Fourier transform.* Nucleic Acids Res, 2002. **30**(14): p. 3059-66.

447 8. Katoh, K. and D.M. Standley, *MAFFT multiple sequence alignment software version 7:*
448 *improvements in performance and usability.* Mol Biol Evol, 2013. **30**(4): p. 772-80.

449 9. Tamura, K. and M. Nei, *Estimation of the number of nucleotide substitutions in the*
450 *control region of mitochondrial DNA in humans and chimpanzees.* Mol Biol Evol, 1993.
451 **10**(3): p. 512-26.

452 10. Wertheim, J.O., et al., *The global transmission network of HIV-1.* Journal of Infectious
453 Diseases, 2014. **209**(2): p. 304-13.

454 11. Kosakovsky Pond, S.L., et al., *HIV-TRACE (TRAnsmission Cluster Engine): a Tool for Large*
455 *Scale Molecular Epidemiology of HIV-1 and Other Rapidly Evolving Pathogens.* Molecular
456 biology and evolution, 2018. **35**(7): p. 1812-1819.

457 12. Ragonnet-Cronin, M., et al., *Automated analysis of phylogenetic clusters.* BMC
458 Bioinformatics, 2013. **14**(1): p. 317.

459 13. Poon, A.F., *Impacts and shortcomings of genetic clustering methods for infectious*
460 *disease outbreaks.* Virus Evolution, 2016. **2**(2): p. vew031.

461 14. Rose, R., et al., *Identifying Transmission Clusters with Cluster Picker and HIV-TRACE.* AIDS
462 Research and Human Retroviruses, 2017. **33**(3): p. 211-218.

463 15. Shimodaira, H. and M. Hasegawa, *Multiple Comparisons of Log-Likelihoods with*
464 *Applications to Phylogenetic Inference.* Molecular Biology and Evolution, 1999. **16**(8): p.
465 1114-1114.

466 16. Cohen, M.S., et al., *Antiretroviral Therapy for the Prevention of HIV-1 Transmission.* New
467 England Journal of Medicine, 2016. **375**(9): p. 830-839.

468 17. Eshleman, S.H., et al., *Analysis of genetic linkage of HIV from couples enrolled in the HIV*
469 *Prevention Trials Network 052 trial.* J Infect Dis, 2011. **204**(12): p. 1918-26.

470 18. Grabowski, M.K., J.T. Herbeck, and A.F.Y. Poon, *Genetic Cluster Analysis for HIV*
471 *Prevention.* Current HIV/AIDS Reports, 2018. **15**(2): p. 182-189.

472 19. Ratmann, O., et al., *Inferring HIV-1 transmission networks and sources of epidemic*
473 *spread in Africa with deep-sequence phylogenetic analysis.* Nature Communications,
474 2019. **10**(1): p. 1411.

475 20. Wymant, C., et al., *PHYLOSCANNER: Inferring Transmission from Within- and Between-*
476 *Host Pathogen Genetic Diversity.* Molecular Biology and Evolution, 2017. **35**(3): p. 719-
477 733.

478 21. Carnegie, N.B., et al., *Linkage of viral sequences among HIV-infected village residents in*
479 *Botswana: estimation of linkage rates in the presence of missing data.* PLoS
480 Computational Biology, 2014. **10**(1): p. e1003430.

481 22. Goodman, L.A., *On Simultaneous Confidence Intervals for Multinomial Proportions.*
482 Technometrics, 1965. **7**(2): p. 247-254.

483 23. Cherry, S., *A Comparison of Confidence Interval Methods for Habitat Use-Availability*
484 *Studies.* The Journal of Wildlife Management, 1996. **60**(3): p. 653-658.

485 24. Novitsky, V., et al., *Long-Range HIV Genotyping Using Viral RNA and Proviral DNA for*
486 *Analysis of HIV Drug Resistance and HIV Clustering.* Journal of Clinical Microbiology,
487 2015. **53**(8): p. 2581-92.
488
489

490 **List of supplementary figures and tables**
491
492 **S1 Fig. Characterization of deep-sequenced HIV-1 viral whole genomes of 4,246 trial**
493 **participants (88% had viral genomes sequenced from pro-viral DNA) sequenced at the**
494 **Harvard biopolymers sequencing core facility.** Trial participants that consented to viral
495 genotyping (n = 5,114) were sequenced at two locations, the Harvard biopolymers sequencing
496 core facility in the United States (n = 4,246) and the Wellcome Trust Sanger Institute in the
497 United Kingdom (n = 868). Most trial participants were virally suppressed at the time of
498 sampling; thus deep-sequencing of HIV-1 viral genomes was generally done on pro-viral DNA
499 [24]. Deep sequencing of viral genomes was performed with Illumina Miseq and Hiseq
500 instruments. The resulting deep-sequencing reads were then assembled and mapped with the
501 shiver sequence assembly software [6] against de-novo reference sequences tailored to each
502 participant's viral population. **Panel A.** shows the length of the HIV-1 viral genome covered on
503 average by deep-sequencing reads of varying minimum read-lengths and sequencing depths;
504 genome coverage dropped progressively beyond 200 nucleotides. **Panel B.** The proportion of
505 participants with deep-sequenced mapped reads covering at least 750 nucleotides of the HIV-1
506 viral genome decreased with increasing minimum read-length and depth of sequencing. **Panel**
507 **C.** Read-length distribution of deep-sequenced mapped reads within an individual. **Panel D.**
508 Distribution of participants with deep-sequencing mapped-reads at each position of the HIV-1
509 viral genome. Most participants sequenced to a depth of at least 30X with mapped reads that
510 were at least 200 bp long and covered at least 750 nucleotides of the HIV-1 genome had reads
511 distributed relatively uniformly along the HIV-1 genome.
512
513 **S2 Fig. Characterization of deep-sequenced HIV-1 viral whole genomes of 868 trial**
514 **participants (45% had viral genomes sequenced from pro-viral DNA) sequenced at the**
515 **Wellcome Trust Sanger Institute.** Trial participants were sequenced as described in S1 Fig.
516 **Panel A.** Breadth of coverage of the HIV-1 viral genome for varying minimum read-lengths and
517 depths of sequencing. There was a marked decrease in average coverage of the HIV-1 genome
518 at or above a sequencing depth of 10X. Furthermore, there was a progressive decline in average
519 viral genome coverage beyond 200 nucleotides. **Panel B.** Proportion of participants with deep-
520 sequencing mapped reads covering over 750 nucleotides of the HIV-1 genome decreased
521 markedly with increasing minimum read-lengths and sequencing depths. **Panel C.** Proportion of
522 deep-sequenced mapped reads in an individual by read length. **Panel D.** Distribution of
523 participants with deep-sequencing mapped-reads at each genome position. Most participants
524 sequenced to a minimum depth of 30X and minimum read-length of 200 bp with a genome
525 coverage of at least 750 nucleotides had mapped reads covering the HIV-1 group-specific
526 antigen (*GAG)* gene.
527
528 **S3 Fig. A plot comparing sampling densities of HIV-1 viral genomes in intervention and control**
529 **communities in the Botswana -Ya Tsie trial.** Intervention communities (blue curve) were
530 sampled more densely than control communities (green curve). Raw data points for the density
531 plots are displayed as a rug-plot (red) on the x-axis. The overall genotyping density across all 30
532 trial communities was 14% (3,832 / 27,446).
533

S4 Fig. A histogram showing the distribution of genetic distances between HIV-1 viral whole genome consensus sequences of trial participants included in phylogenetic analyses. Of the 5,114 trial participants that consented to viral genotyping 3,832 met criteria for inclusion in phylogenetic analyses; these were individuals whose viral genomes had fewer than 30% of the bases missing beyond the first 1,000 nucleotides and a minimum genome length of 6300 nucleotides. Tamura-Nei 93 genetic distances were computed between HIV-1 viral whole genome consensus sequences of all possible pairs (n = 7,340,196) of the 3,832 trial participants available for phylogenetic analyses. To improve visibility pairwise genetic distances below 0.05 (5%) nucleotide substitutions per site are also shown in a zoomed-in plot. The mean ($\pm$ standard deviation) pairwise distance separating sequences was, $12.5 \pm 3.0\%$ and the maximum pairwise distance was, 40.0% substitutions per site.

S5 Fig. A histogram showing the distribution of genetic distances between viral whole genome consensus sequences of epidemiologically linked HIV-1 subtype-C couples in the HIV Prevention Trials Network 052 study. Tamura-Nei 93 genetic distances were computed between viral whole genome consensus sequences of all possible pairs (n = 6,670) of the 116 epidemiologically linked couples with HIV-1 subtype-C infections in the HIV Prevention Trials Network 052 study. Pairwise comparisons between epidemiologically linked sequences (yellow) and ones between epidemiologically unlinked sequences (blue) are highlighted in color. The mean ($\pm$ standard deviation) pairwise distance separating the sequences was, $12.0\% \pm 4.1\%$ and the maximum pairwise distance was, 19.6% substitutions per site. Genetic distances separating viral sequences of most epidemiology-linked couples were below 5.0% substitutions per site.

S6 Fig. Barplots showing subgraph distance distributions between deep-sequenced HIV-1 viral whole genomes of trial participants in clusters of genetically similar HIV-1 infections. Within each genetic cluster parsimony-based ancestral host state reconstruction was done with phyloscanner [18-20] to identify probable transmission pairs and resolve the probable order of transmission events. Thereafter, within each identified genetic cluster, the median subgraph distance and most frequent subgraph topology between each pair of individuals was determined across all deep-sequence phylogenies along the genome where the pair had sufficient mapped reads for phylogenetic inference. A subgraph refers to all tips and internal nodes of a phylogeny assigned to an individual through parsimony-based ancestral state reconstruction. Subgraph distances were standardized to the mean rate of evolution for HIV-1 group-specific antigen (*GAG)* and polymerase *(POL)* genes to account for variation in mutation rates along the genome. The three panels show median subgraph distance distributions of pairs in the union (panel A) of genetic clusters detected with HIV-TRACE (panel B) and Cluster Picker (panel C) algorithms. Within each panel, phylogenetically-linked pairs with an inferred direction of transmission are summarized into same-sex (male-male and female-female) and opposite-sex pairs (male-female). The subgraphs of deep-sequenced HIV-1 viral genomes of most phylogenetically-linked directed pairs were separated by a patristic distance below 0.035 (3.5%). Same-sex pairs were presumed to be members of transmission chains with missing intermediates.

S7 Fig. Barplots showing the size distribution of clusters of genetically similar HIV-1 infections in the Botswana -Ya Tsie trial. The bottom two plots show the size distribution of genetic clusters identified with HIV-TRACE (n = 122) and Cluster Picker algorithms (n = 231), respectively; and the top two plots show the size distribution of the union (n = 236) and overlap (n = 117) of genetic clusters detected with HIV-TRACE and Cluster Picker. Most clusters were small-sized comprising two or three members.

S8 Fig. Barplots showing the spread of genetic clusters across trial communities. The bottom two plots show the number of genetic clusters identified with HIV-TRACE (n = 122) and Cluster Picker algorithms (n = 231), respectively; and the top two plots show the union (n = 236) and overlap (n = 117) of genetic clusters found with HIV-TRACE and Cluster Picker. Most lineages localized to one or two trial communities; with few clusters having members spread out across five or more trial communities.

S9 Fig. Estimated transmission flows of HIV-1 infection by gender within and between 5 year age groups in the Botswana -Ya Tsie trial population. Transmission flows were estimated as described in Figure 6. Additionally, transmission flows were adjusted for differential sampling by gender and age group.

S1 Table. Characteristics of the Botswana -Ya Tsie trial population living with HIV

S2 Table. HIV-1 viral genotyping densities of communities in the Botswana Ya Tsie HIV trial

S3 Table: A listing of command-line parameters used for deep-sequence phylogenetics with phyloscanner

S4 Table. Summary of viral suppression among trial participants in clusters of genetically similar HIV-1 infections

S5 Table. Observed and predicted proportions of HIV-1 transmissions in the trial population by age group

S6 Table. Observed and predicted proportions of HIV-1 transmissions in the trial population by gender and age group

S7 Table. Proportions of inferred HIV-1 transmissions among trial participants by gender

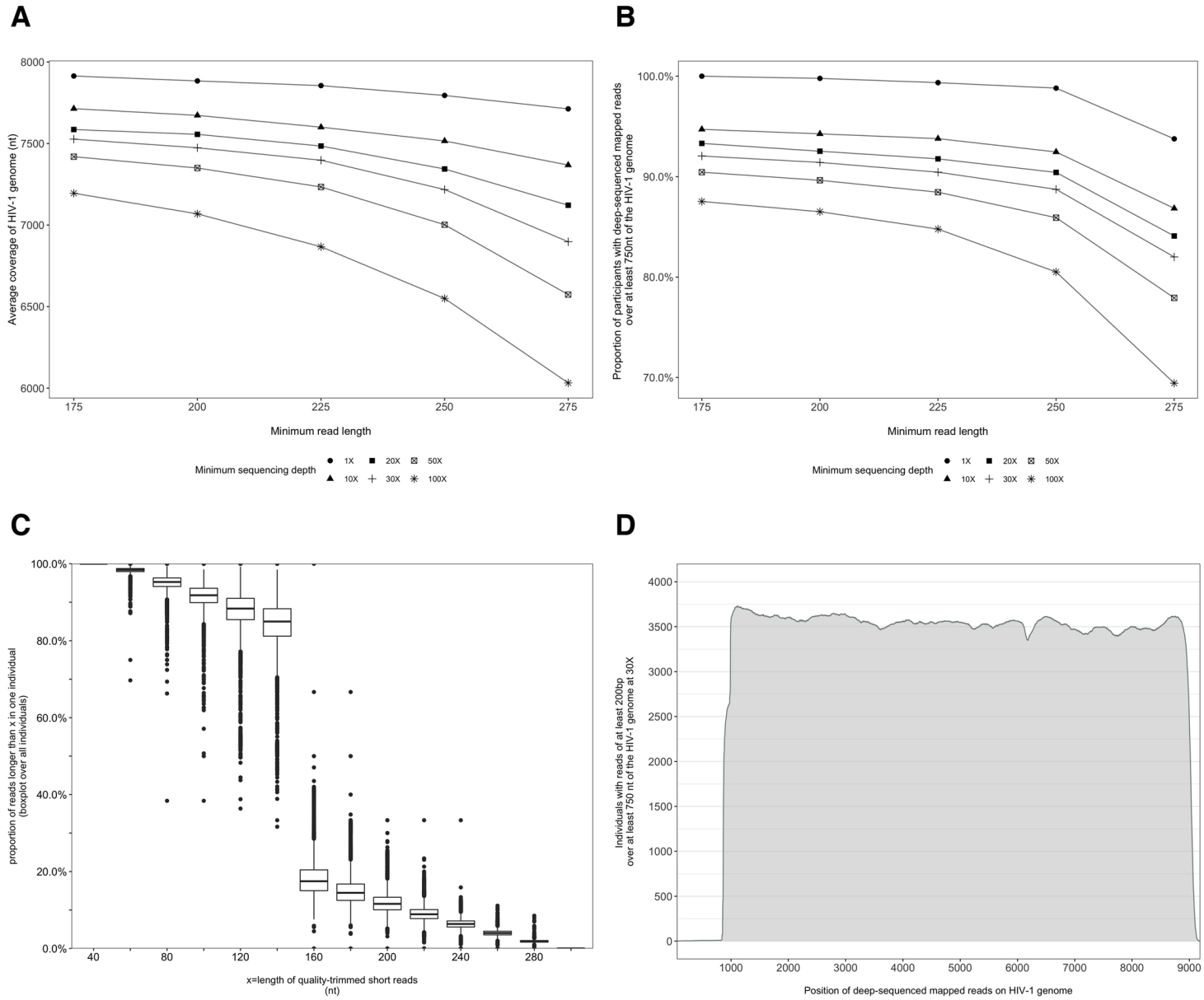
S8 Table. Observed and predicted proportions of HIV-1 transmissions in the trial population by community

S9 Table. Observed and predicted proportions of HIV-1 transmissions in the trial population by geographical region

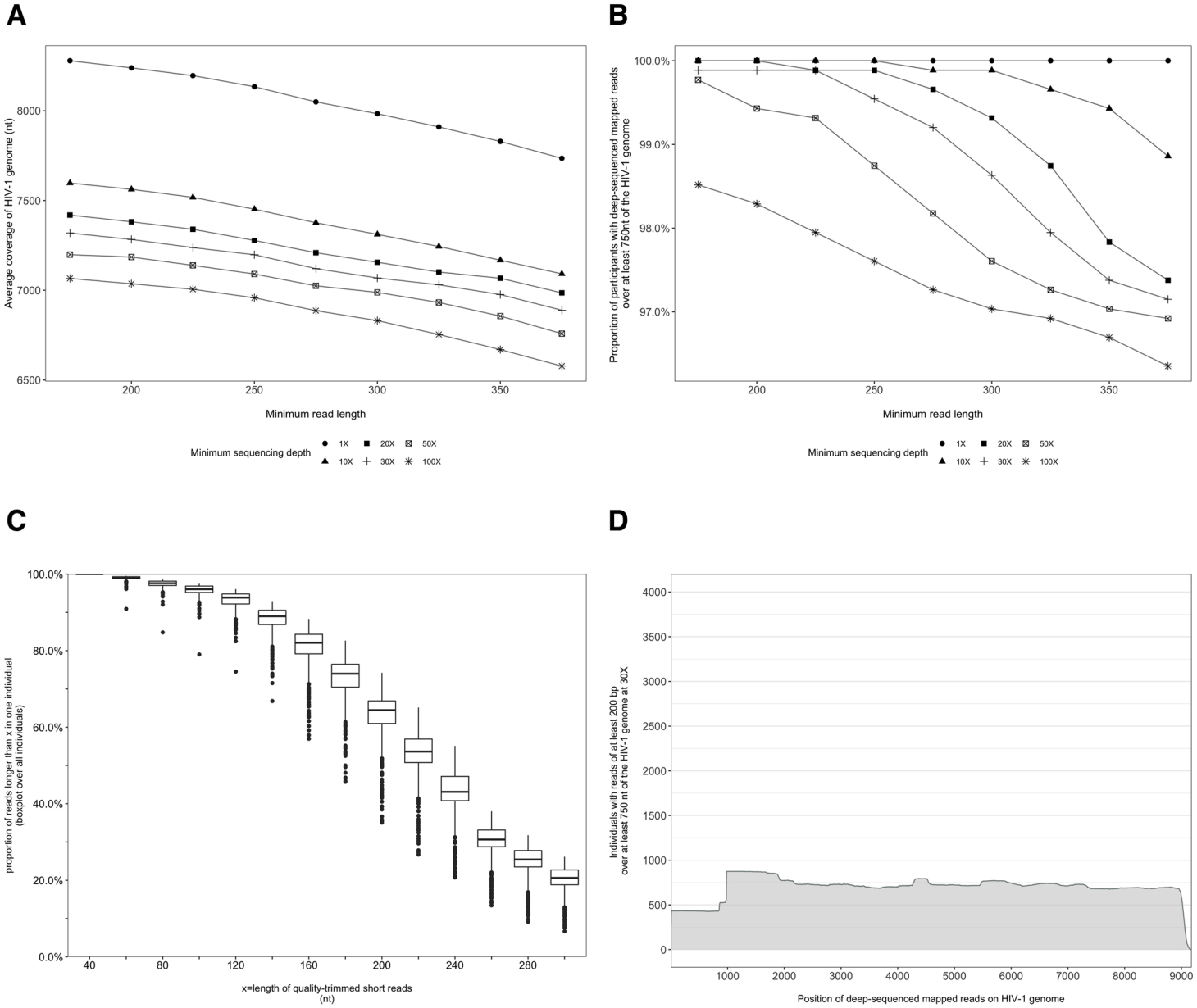
S10 Table. Observed and predicted proportions of HIV-1 transmissions in the trial population by randomization-intervention condition

S11 Table. Observed and predicted proportions of HIV-1 transmissions in the trial population at baseline by randomization-intervention condition

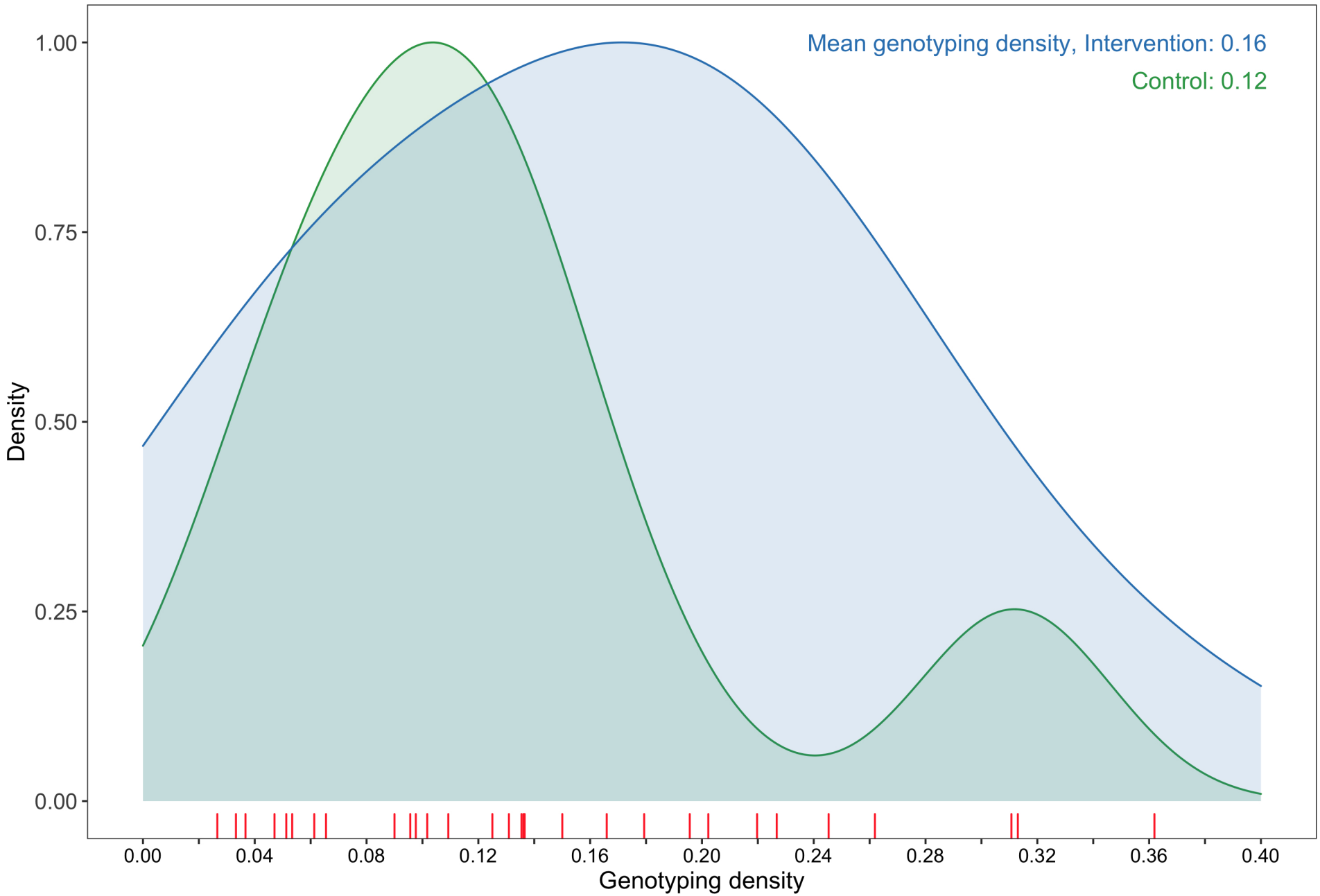
S12 Table. Observed and predicted proportions of HIV-1 transmissions in the trial population after baseline by randomization-intervention condition



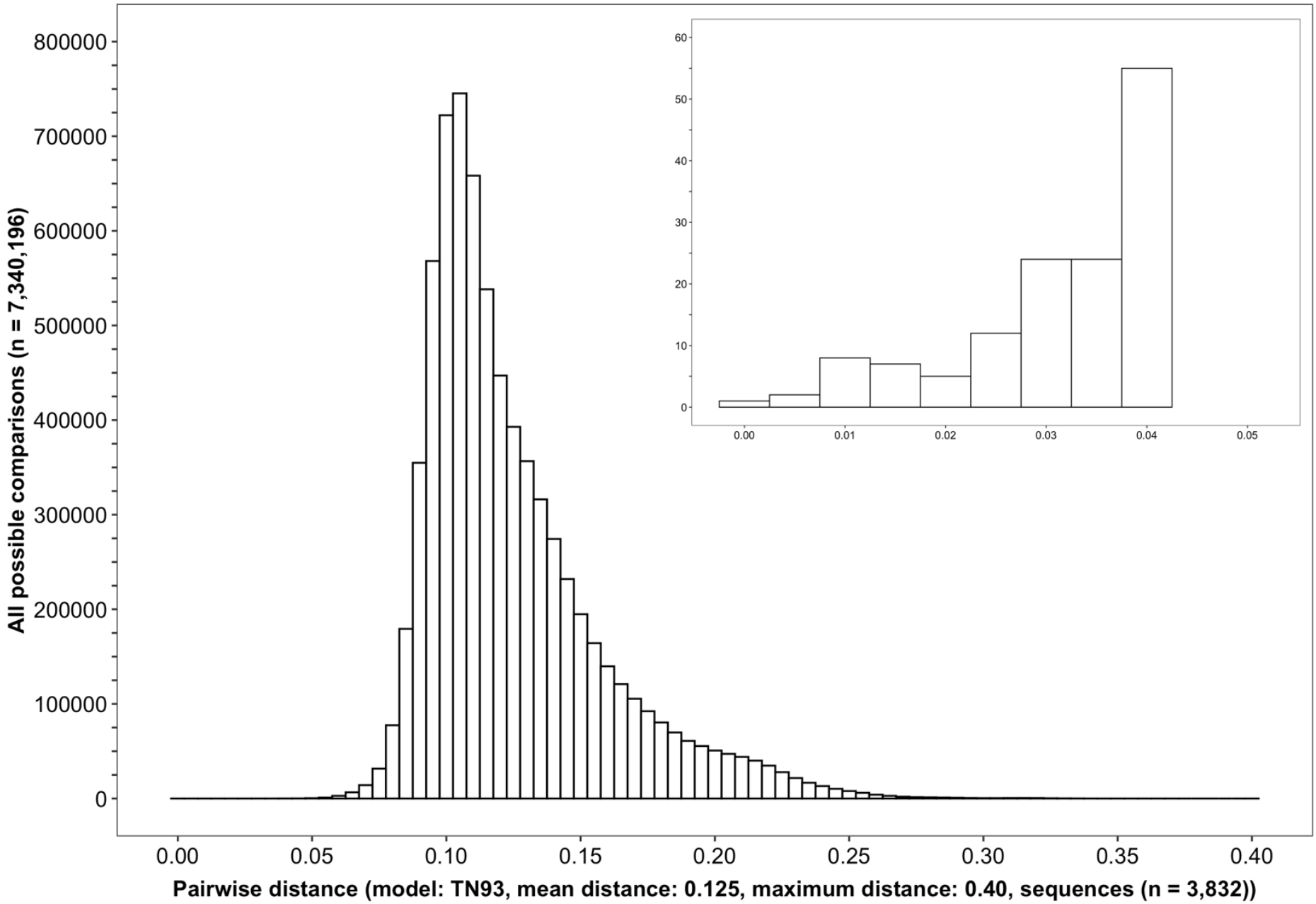
S1 Fig. Characterization of deep-sequenced HIV-1 viral whole genomes of 4,246 trial participants (88% had viral genomes sequenced from pro-viral DNA) sequenced at the Harvard biopolymers sequencing core facility. Trial participants that consented to viral genotyping (n = 5,114) were sequenced at two locations, the Harvard biopolymers sequencing core facility in the United States (n = 4,246) and the Wellcome Trust Sanger Institute in the United Kingdom (n = 868). Most trial participants were virally suppressed at the time of sampling; thus deep-sequencing of HIV-1 viral genomes was generally done on pro-viral DNA [24]. Deep sequencing of viral genomes was performed with Illumina Miseq and Hiseq instruments. The resulting deep-sequencing reads were then assembled and mapped with the shiver sequence assembly software [6] against de-novo reference sequences tailored to each participant's viral population. Panel A. shows the length of the HIV-1 viral genome covered on average by deep-sequencing reads of varying minimum read-lengths and sequencing depths; genome coverage dropped progressively beyond 200 nucleotides. Panel B. The proportion of participants with deep-sequenced mapped reads covering at least 750 nucleotides of the HIV-1 viral genome decreased with increasing minimum read-length and depth of sequencing. Panel C. Read-length distribution of deep-sequenced mapped reads within an individual. Panel D. Distribution of participants with deep-sequencing mapped-reads at each position of the HIV-1 viral genome. Most participants sequenced to a depth of at least 30X with mapped reads that were at least 200 bp long and covered at least 750 nucleotides of the HIV-1 genome had reads distributed relatively uniformly along the HIV-1 genome.



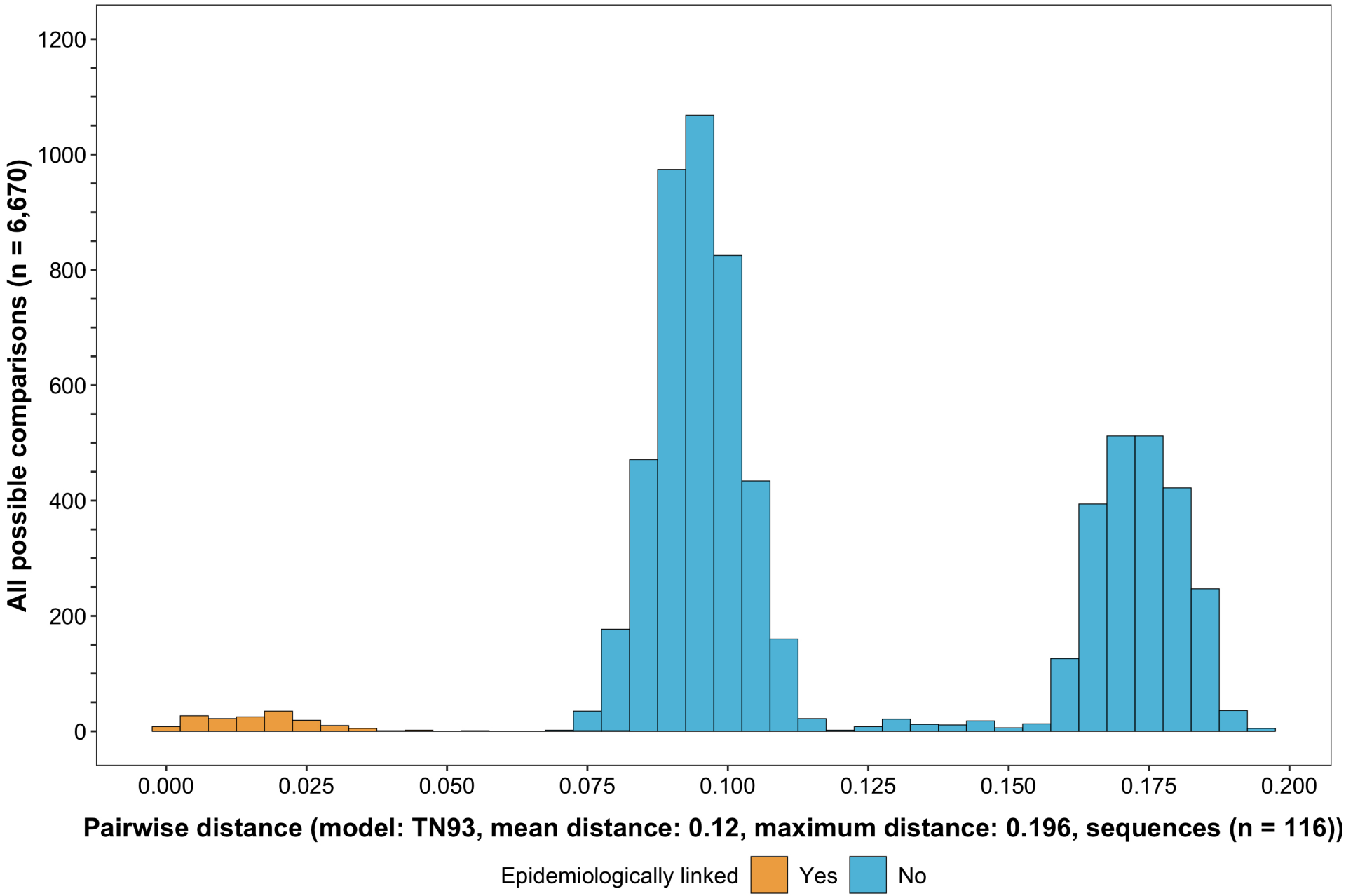
S2 Fig. Characterization of deep-sequenced HIV-1 viral whole genomes of 868 trial participants (45% had viral genomes sequenced from pro-viral DNA) sequenced at the Wellcome Trust Sanger Institute. Trial participants were sequenced as described in S1 Fig. Panel A. Breadth of coverage of the HIV-1 viral genome for varying minimum read-lengths and depths of sequencing. There was a marked decrease in average coverage of the HIV-1 genome at or above a sequencing depth of 10X. Furthermore, there was a progressive decline in average viral genome coverage beyond 200 nucleotides. Panel B. Proportion of participants with deep-sequencing mapped reads covering over 750 nucleotides of the HIV-1 genome decreased markedly with increasing minimum read-lengths and sequencing depths. Panel C. Proportion of deep-sequenced mapped reads in an individual by read length. Panel D. Distribution of participants with deep-sequencing mapped-reads at each genome position. Most participants sequenced to a minimum depth of 30X and minimum read-length of 200 bp with a genome coverage of at least 750 nucleotides had mapped reads covering the HIV-1 group-specific antigen (GAG) gene.



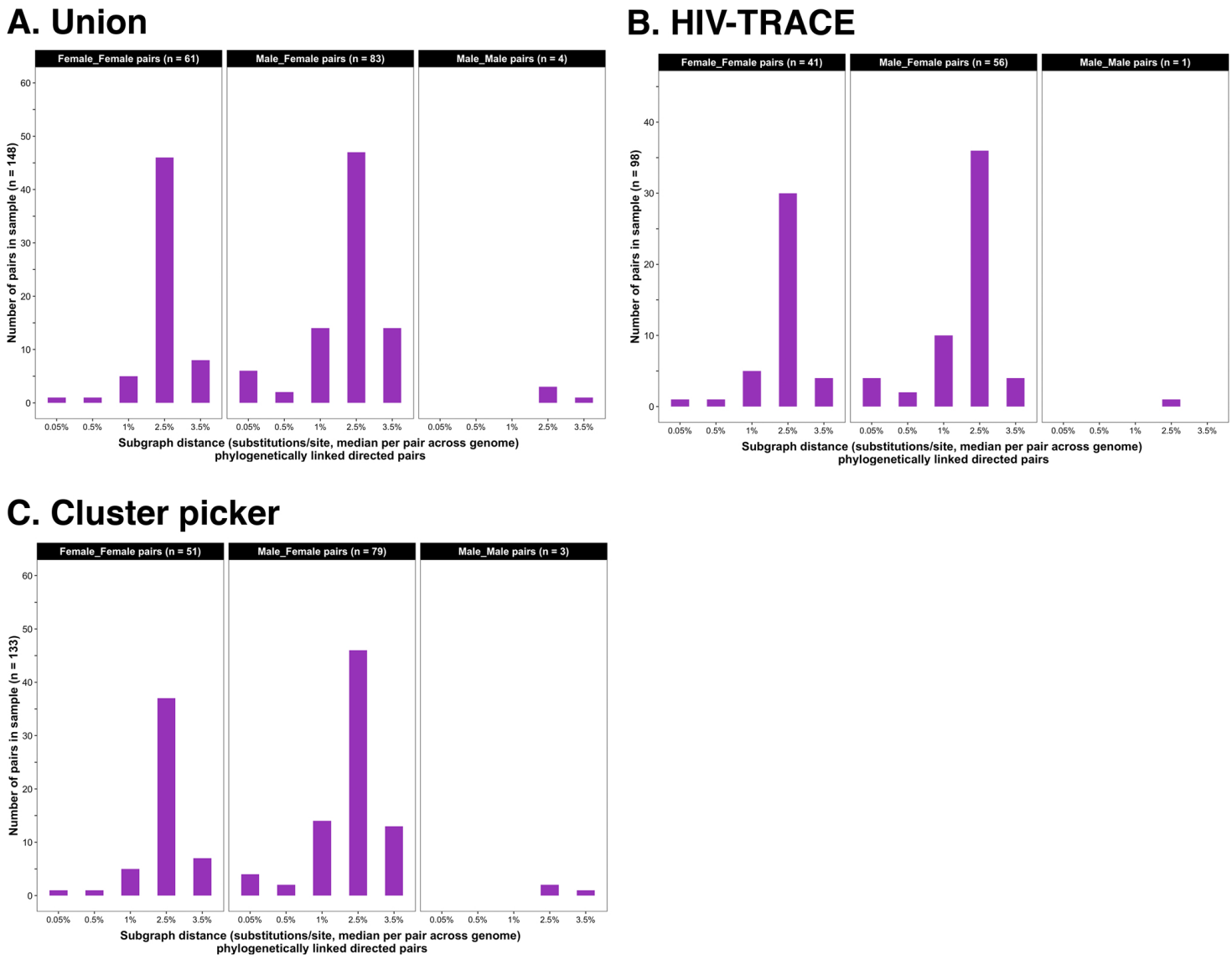
S3 Fig. A plot comparing sampling densities of HIV-1 viral genomes in intervention and control communities in the Botswana -Ya Tsie trial. Intervention communities (blue curve) were sampled more densely than control communities (green curve). Raw data points for the density plots are displayed as a rug-plot (red) on the x-axis. The overall genotyping density across all 30 trial communities was 14% (3,832 / 27,446).



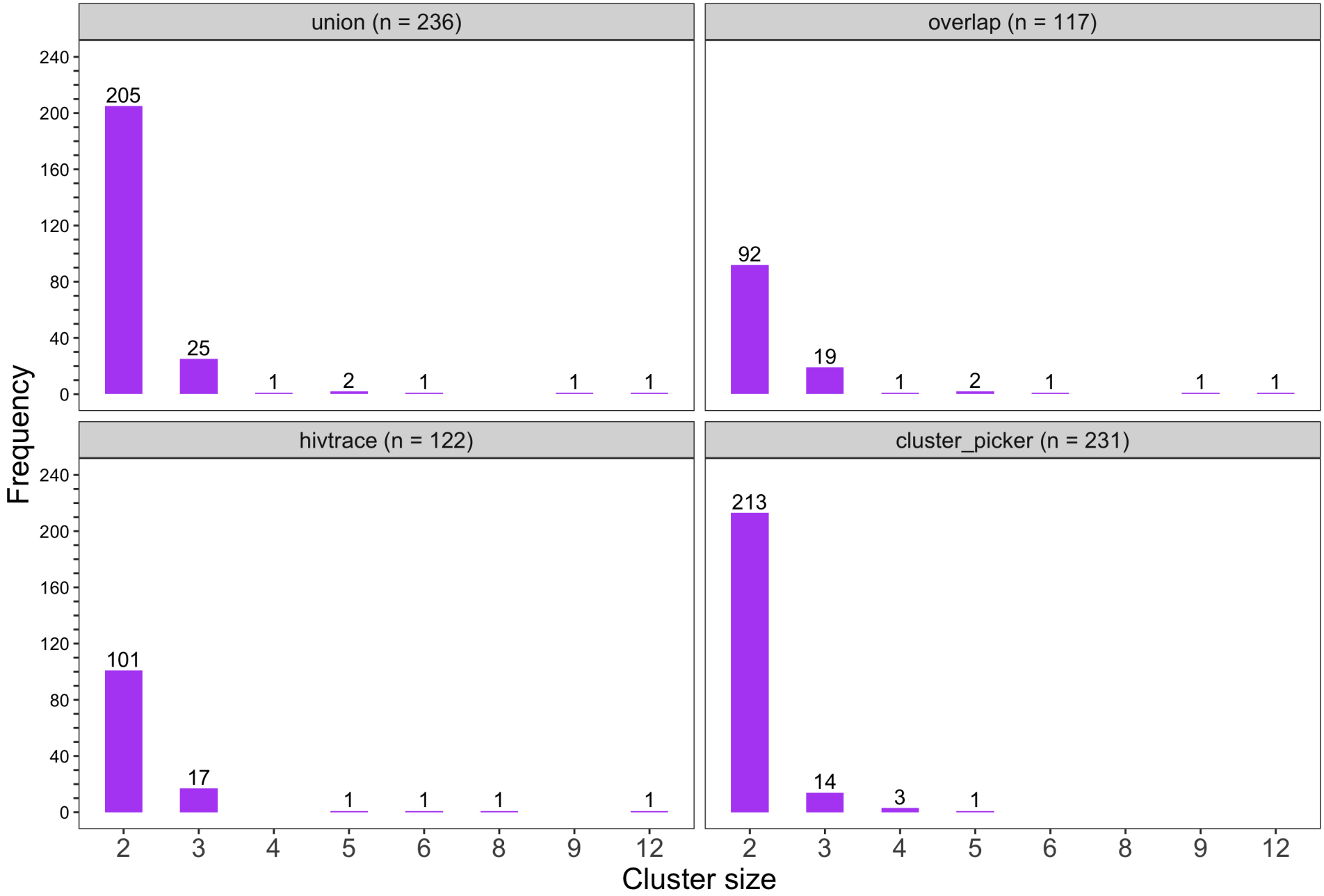
S4 Fig. A histogram showing the distribution of genetic distances between HIV-1 viral whole genome consensus sequences of trial participants included in phylogenetic analyses. Of the 5,114 trial participants that consented to viral genotyping 3,832 met criteria for inclusion in phylogenetic analyses; these were individuals whose viral genomes had fewer than 30% of the bases missing beyond the first 1,000 nucleotides and a minimum genome length of 6300 nucleotides. Tamura-Nei 93 genetic distances were computed between HIV-1 viral whole genome consensus sequences of all possible pairs (n = 7,340,196) of the 3,832 trial participants available for phylogenetic analyses. To improve visibility pairwise genetic distances below 0.05 (5%) nucleotide substitutions per site are also shown in a zoomed-in plot. The mean (± standard deviation) pairwise distance separating sequences was, 12.5±3.0% and the maximum pairwise distance was, 40.0% substitutions per site.



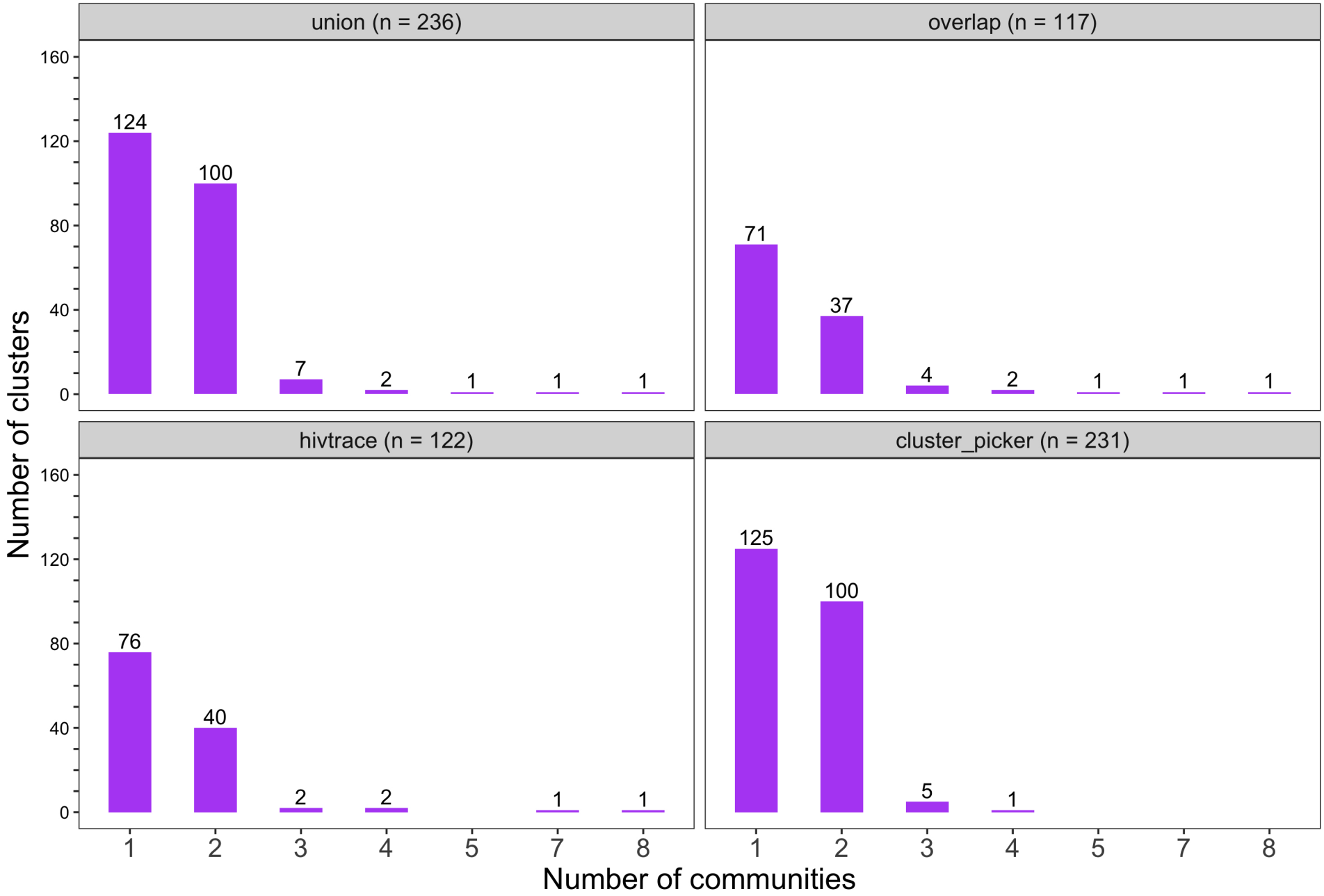
S5 Fig. A histogram showing the distribution of genetic distances between viral whole genome consensus sequences of epidemiologically linked HIV-1 subtype-C couples in the HIV Prevention Trials Network 052 study. Tamura-Nei 93 genetic distances were computed between viral whole genome consensus sequences of all possible pairs (n = 6,670) of the 116 epidemiologically linked couples with HIV-1 subtype-C infections in the HIV Prevention Trials Network 052 study. Pairwise comparisons between epidemiologically linked sequences (yellow) and ones between epidemiologically unlinked sequences (blue) are highlighted in color. The mean (± standard deviation) pairwise distance separating the sequences was, 12.0% ±4.1% and the maximum pairwise distance was, 19.6% substitutions per site. Genetic distances separating viral sequences of most epidemiology-linked couples were below 5.0% substitutions per site.



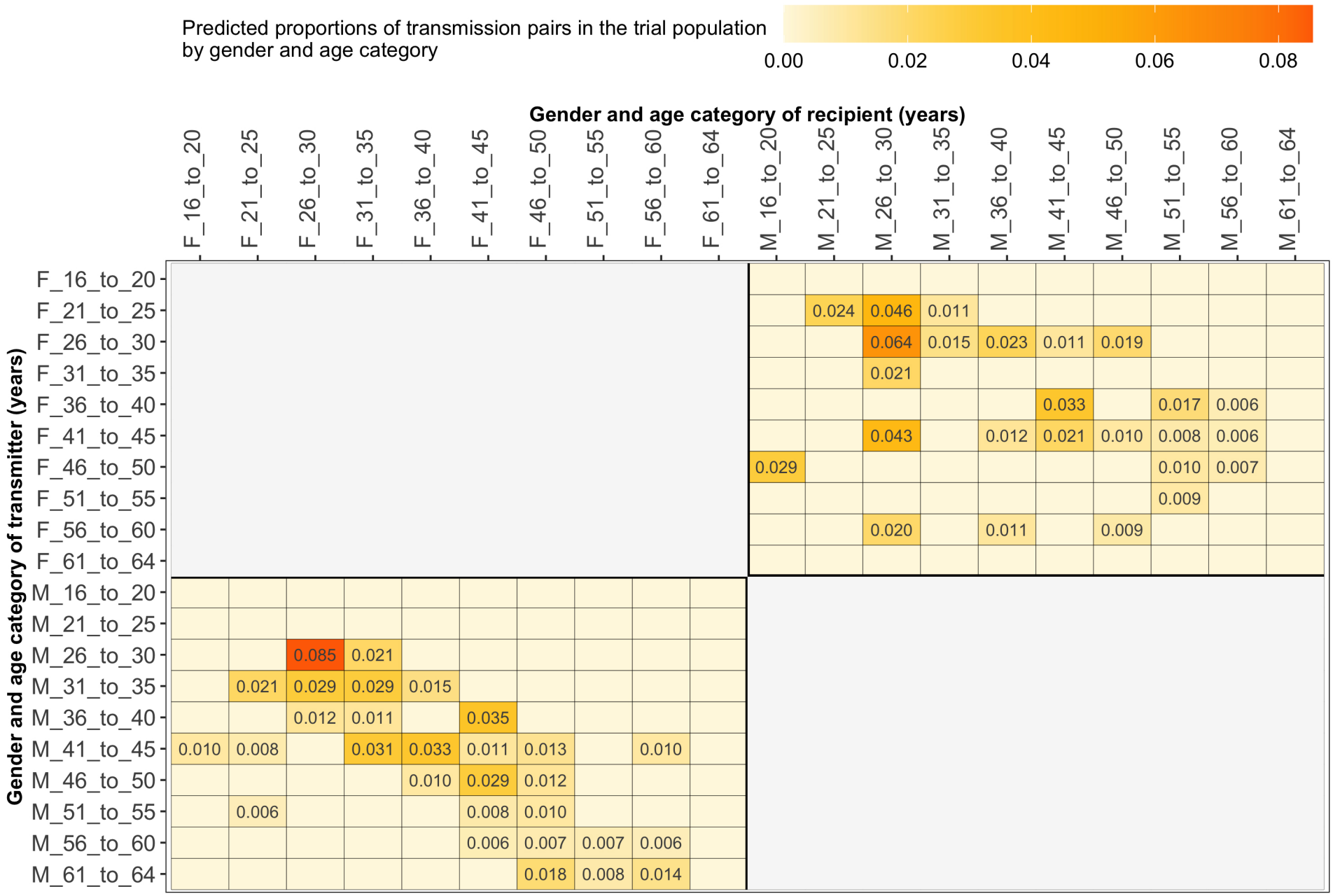
S6 Fig. Barplots showing subgraph distance distributions between deep-sequenced HIV-1 viral whole genomes of trial participants in clusters of genetically similar HIV-1 infections. Within each genetic cluster parsimony-based ancestral host state reconstruction was done with phyloscanner [18-20] to identify probable transmission pairs and resolve the probable order of transmission events. Thereafter, within each identified genetic cluster, the median subgraph distance and most frequent subgraph topology between each pair of individuals was determined across all deep-sequence phylogenies along the genome where the pair had sufficient mapped reads for phylogenetic inference. A subgraph refers to all tips and internal nodes of a phylogeny assigned to an individual through parsimony-based ancestral state reconstruction. Subgraph distances were standardized to the mean rate of evolution for HIV-1 group-specific antigen (*GAG*) and polymerase (*POL*) genes to account for variation in mutation rates along the genome. The three panels show median subgraph distance distributions of pairs in the union (panel A) of genetic clusters detected with HIV-TRACE (panel B) and Cluster Picker (panel C) algorithms. Within each panel, phylogenetically-linked pairs with an inferred direction of transmission are summarized into same-sex (male-male and female-female) and opposite-sex pairs (male-female). The subgraphs of deep-sequenced HIV-1 viral genomes of most phylogenetically-linked directed pairs were separated by a patristic distance below 0.035 (3.5%). Same-sex pairs were presumed to be members of transmission chains with missing intermediates.



S7 Fig. Barplots showing the size distribution of clusters of genetically similar HIV-1 infections in the Botswana -Ya Tsie trial. The bottom two plots show the size distribution of genetic clusters identified with HIV-TRACE (n = 122) and Cluster Picker algorithms (n = 231), respectively; and the top two plots show the size distribution of the union (n = 236) and overlap (n = 117) of genetic clusters detected with HIV-TRACE and Cluster Picker. Most clusters were small-sized comprising two or three members.



S8 Fig. Barplots showing the spread of genetic clusters across trial communities. The bottom two plots show the number of genetic clusters identified with HIV-TRACE (n = 122) and Cluster Picker algorithms (n = 231), respectively; and the top two plots show the union (n = 236) and overlap (n = 117) of genetic clusters found with HIV-TRACE and Cluster Picker. Most lineages localized to one or two trial communities; with few clusters having members spread out across five or more trial communities.



S9 Fig. Estimated transmission flows of HIV-1 infection by gender within and between 5 year age groups in the Botswana -Ya Tsie trial population. Transmission flows were estimated as described in Figure 6. Additionally, transmission flows were adjusted for differential sampling by gender and age group.