

1 Unpacking sources of transmission in HIV prevention trials with deep-sequence pathogen data

2 – BCPP/ Ya Tsie study

3

4 Lerato E. Magosi<sup>1,2\*</sup>, Eric Tchetgen Tchetgen<sup>3</sup>, Vlad Novitsky<sup>4,5</sup>, Molly Pretorius Holme<sup>4</sup>, Janet  
5 Moore<sup>6</sup>, Pam Bachanas<sup>6</sup>, Refeletswe Lebelonyane<sup>7</sup>, Christophe Fraser<sup>8</sup>, Sikhulile Moyo<sup>5</sup>,  
6 Kathleen E. Hurwitz<sup>9</sup>, Tendani Gaolathe<sup>5</sup>, Ravi Goyal<sup>10</sup>, Joseph Makhema<sup>5</sup>, Shahin Lockman<sup>4,5,11†</sup>,  
7 Max Essex<sup>4,5†</sup>, Victor De Gruttola<sup>9†</sup>, & Marc Lipsitch<sup>1†\*</sup>

8

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

11 2. Wellcome Sanger Institute, Cambridge, United Kingdom

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

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

15 5. Botswana Harvard AIDS Institute Partnership, Gaborone, Botswana

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

17 7. Ministry of Health, Republic of Botswana, Gaborone, Botswana

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

20 9. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University,  
21 Boston, USA

22 10. Division of Infectious Diseases and Global Public Health, University of California San Diego,  
23 La Jolla CA, USA

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

25

26 †These authors contributed equally to this work and are co-senior authors

27

28 \*Corresponding author. Email: [lmagosi@hsph.harvard.edu](mailto:lmagosi@hsph.harvard.edu) (Lerato E. Magosi, DPhil);

29 [mlipsitc@hsph.harvard.edu](mailto:mlipsitc@hsph.harvard.edu) (Marc Lipsitch, DPhil)

30 **Abstract**

31 To develop effective HIV prevention strategies that can guide public health policy it is important  
32 to identify the main sources of infection in HIV prevention studies. Accordingly, we devised a  
33 statistical approach that leverages deep- (or next generation) sequenced pathogen data to  
34 estimate the relative contribution of different sources of infection in community-randomized  
35 trials of infectious disease prevention. We applied this approach to the Botswana Combination  
36 Prevention Project (BCPP) and estimated that 90% [95% Confidence Interval (CI): 81 – 93] of  
37 new infections that occurred in individuals in communities that received combination  
38 prevention (including universal HIV test-and-treat) originated from individuals residing in  
39 communities outside of the trial area. We estimate that the relative impact of the intervention  
40 was greater in rural geographically isolated communities with limited opportunity for imported  
41 infections compared to communities neighboring major urban centers. Treating people with  
42 HIV limits the spread of infection to uninfected individuals; accordingly, counterfactual  
43 modeling scenarios estimated that a nationwide application of the intervention could have  
44 reduced transmissions to recipients in trial communities by 59% [3 – 87], much higher than the  
45 observed 30% reduction. Our results suggest that the impact of the BCPP trial intervention was  
46 substantially limited by sources of transmission outside the trial area, and that the impact of  
47 the intervention could be considerably larger if applied nationally. We recommend that studies  
48 of infectious disease prevention consider the impact of sources of transmission beyond the  
49 reach of the intervention when designing and evaluating interventions to inform public health  
50 programs.

51



## 53 Introduction

54

55 Why did the landmark community-randomized universal HIV test-and-treat trials in sub-  
56 Saharan Africa - BCPP/ Ya Tsie [1], HPTN 071/ PopART [2], SEARCH [3] and ANRS 12249/TasP [4]  
57 - show variable reductions in the occurrence of new HIV infections in trial communities that  
58 received the intervention compared to control communities (0% – 30%) despite substantial  
59 gains in viral suppression [5]? This is one of the most important questions in the HIV policy  
60 world today because HIV ‘test-and-treat’ was thought to hold great potential to bring the HIV  
61 epidemic under control in the absence of a successful vaccine or functional cure. Some of the  
62 variation in the incidence reductions observed is thought to be due to a change in national HIV  
63 treatment guidelines to universal treatment part-way through the trials effectively reducing the  
64 difference between intervention and control communities. Another complementary hypothesis  
65 is that HIV transmissions to residents of intervention communities from individuals in non-  
66 intervention communities in the trial (control communities) and from communities not taking  
67 part in the trial (non-trial communities) limited the size of effect observed in the trials, but it is  
68 unknown to what degree. A large dilution of the intervention effect in the trials by transmission  
69 from non-intervention communities could suggest a larger impact of the intervention than  
70 originally envisaged [6-8].

71

72 We test this hypothesis in one of the four trials, the Botswana Combination Prevention Project.  
73 Specifically, we developed a statistical modeling approach that uses directed sexual contacts  
74 inferred from deep-sequenced HIV virus to estimate the relative extent to which transmissions

75 in trial communities occurred from individuals in the same community; different communities  
76 in the same trial arm; different communities in the opposite trial arm; and non-trial  
77 communities. In addition, to estimate what the impact of a nationwide intervention would have  
78 been on recipients in trial communities, we apply our statistical approach in “counterfactual”  
79 modeling scenarios that estimate the relative contribution of the different sources of infection  
80 mentioned above in the presence and absence of the intervention. Note that in a nationwide  
81 intervention all communities nationally would receive combination prevention (including  
82 universal HIV test-and-treat).

83

84

85

## 86 **Materials and Methods**

87

### 88 **BCPP Study Description and Data**

89 *BCPP study description.* The Botswana Combination Prevention Project (BCPP, also known as  
90 the Ya Tsie trial) was a pair-matched community-randomized trial to evaluate the effect of  
91 universal HIV testing and treatment on HIV incidence reduction. The trial was conducted in 30  
92 rural and peri-urban communities across Botswana from 2013-2018 [1]. Trial participants were  
93 adults aged 16-64 years and the average population size eligible to participate in each trial  
94 community was 3,820 people. Communities were matched into 15 pairs based on three criteria:  
95 geographical proximity to major urban areas (Gaborone city, Palapye and Francistown city),  
96 population size and age structure, and access to health services; then within each pair,  
97 communities were randomized into the intervention and control arms of the trial. The 15  
98 intervention communities in the trial received expanded access to universal HIV testing (with  
99 attempt to test all willing adult residents who did not have documented positive HIV status),  
100 strengthened linkage-to-care for early treatment, and expanded treatment availability. After a  
101 period of community sensitization through door-to-door canvassing, community leadership  
102 engagement and public loudspeaker announcements, mobile and home-based HIV testing  
103 campaigns were conducted within each intervention community over approximately two  
104 consecutive months [9]. Routine testing in intervention community health facilities was  
105 reinforced to diagnose all people with HIV and avail them early treatment. An additional effort  
106 was made to offer HIV testing to men and youth where they work and socialize, for example: at  
107 bars and community football (soccer) matches. To strengthen linkage-to-care, people with HIV

108 who were not on treatment were assisted to schedule an appointment at a local clinic,  
109 provided text alerts prior to the appointment and followed-up to re-schedule in the case of a  
110 missed appointment. Access to services for safe male circumcision and prevention of mother-  
111 to-child transmission was also expanded in intervention communities. By comparison, control  
112 communities received the standard-of-care, which before 2016 meant that people with HIV  
113 qualified to start antiretroviral treatment when their CD4 cell count was below 350 cells per  
114 microliter. Beginning June 2016, the national HIV treatment policy was changed to universal  
115 treatment meaning that immediate antiretroviral treatment was now available in both arms of  
116 the BCPP trial. To evaluate HIV incidence reduction, an HIV incidence follow-up cohort was  
117 established through a baseline household survey of a random sample of 20% of households in  
118 each trial community. Annual household surveys with retesting for HIV (in persons who were  
119 HIV-negative) were then conducted in the same 20% household sample in all 30 communities  
120 during the trial. The BCPP trial comprised 7.6% (175,664) of the national population and  
121 showed a 30% reduction in the occurrence of new HIV infections in intervention communities  
122 compared to control communities over an average of 29 months [1]. In addition, the BCPP trial  
123 conducted an end-of-study survey of 100% of households in 3 intervention communities and 3  
124 control communities to assess progress on the 90-90-90 UNAIDS targets. Trial participants with  
125 HIV were invited to provide a sample for viral phylogenetic analysis. This included all people  
126 with HIV from (1) the baseline household survey, (2) annual household surveys, (3) end-of-study  
127 survey, as well as (4) all people with HIV (but not yet on ART) who were referred for treatment  
128 during community-wide testing and counseling campaigns, (5) all people with HIV that later

129 presented at health care facilities in intervention communities and (6) all people with HIV who  
130 were already receiving HIV care at health facilities in intervention communities.

131

132 *Deep-sequence phylogenetics data.* Near full-length genome sequences were obtained using  
133 predominantly proviral DNA (as the majority of study participants were virally suppressed on  
134 ART) or RNA. The HIV-1 viral consensus whole genomes of individuals that met minimum  
135 criteria for inclusion in phylogenetic analyses were ones that had fewer than 30% of bases  
136 missing beyond the first 1,000 nucleotides [10]. To efficiently use computational resources, viral  
137 consensus whole genomes were used to identify groups (or clusters) of trial participants with  
138 genetically similar HIV-1 infections as a filtering step to exclude distantly related sequences  
139 from deep-sequence phylogenetic analysis [10]. A detailed description of the deep-sequence  
140 phylogenetic analysis is published in [10]. Briefly, we performed ancestral host-state  
141 reconstruction with the phyloscanner software [11, 12] to identify pairs of trial participants  
142 with genetically similar HIV-1 infections and the probable direction of transmission between  
143 them (female-to-male or male-to-female). For brevity, we refer to the directed opposite-sex  
144 transmission pairs as source-recipient pairs. For each identified transmission pair there was  
145 accompanying metadata on the names of the communities in which the source and recipient  
146 partners reside and the randomization-condition to which their communities were assigned.

147

148 *Pairwise drive distance data.* Pairwise drive distances between ordered pairs of 488  
149 communities in the 2011 Botswana population and housing census were successfully sourced  
150 from the google distance matrix application programming interface (API) with the mapsapi



151 package v0.5.0 in R v4.1.1 [13]. The 488 census communities included all 30 communities in the  
152 BCPP trial. Therefore, of the possible 488 x 488 ordered community pairs between the 488  
153 census communities we sourced 488 x 30 ordered community pairs that had any of the 488  
154 census communities as a source (origin) community and any of the 30 trial communities as a  
155 recipient (destination) community.

156

157 *Population-size and HIV prevalence estimates.* Population-size estimates of 488 communities in  
158 the 2011 Botswana population and housing census were sourced from [14], and district-level  
159 HIV prevalence estimates were obtained from the 2013 Botswana AIDS Impact Survey (BAIS  
160 2013) [15].

161

## 162 **Estimating Transmissions to Recipients in BCPP Trial Communities**

163

164 To estimate transmissions that occurred to recipients in trial communities from different  
165 sources of infection nationally, we first inferred directed (opposite-sex) transmission events  
166 between the 30 BCPP trial communities using deep-sequence phylogenetics. Then we  
167 statistically modeled the risk of transmission between trial communities using a negative-  
168 binomial regression framework; with the inferred transmission events as the response variable  
169 and the following variables as predictors: (1) the pairwise drive distance separating the source  
170 and recipient communities, (2) whether the source community was randomized to receive the  
171 intervention, and (3) whether the source community and the recipient community were the  
172 same (within-community transmission) or different (between-community transmission). After

173 that we used the pairwise drive distances between the 488 communities in the 2011 Botswana  
174 population and housing census as input to the model of the risk of transmission between trial  
175 communities to predict the risk of transmission (expected probability of viral genetic-linkage  
176 had the cases been sequenced) between communities nationally. Finally, to estimate the  
177 number of transmissions into trial communities from all communities nationally, estimates of  
178 the risk of transmission to recipients in trial communities from communities nationally were  
179 combined with population-size estimates from the 2011 Botswana population and housing  
180 census [14] and district-level HIV prevalence estimates from the 2013 Botswana AIDS Impact  
181 Survey (BAIS 2013) [15]. The supplementary appendix provides a detailed account of the  
182 statistical approach used to estimate the relative contribution of different sources of infection  
183 in: 1) the same community, 2) different communities in the same trial arm, 3) different  
184 communities in the opposite trial arm and 4) in communities outside the trial area.  
185

## 186 Results

187

188 Of the 5,114 trial participants who consented to a blood draw for viral genotyping and whose  
189 HIV viral whole genomes were successfully deep-sequenced [1, 10], 3,832 met inclusion criteria  
190 for phylogenetic analysis, and from those, we identified 82 directed opposite-sex transmission  
191 pairs between ordered pairs of the 30 communities in the BCPP trial (Supplementary Figure 1)  
192 [10]. Of the 82 source-recipient pairs, 51 (21 female-to-male, 30 male-to-female) were  
193 identified between HIV viral genomes sampled during the baseline period of the trial compared  
194 to 31 (16 female-to-male, 15 male-to-female) where the recipient's genome was sampled post-  
195 baseline. We defined the post-baseline period as at least one year after baseline household  
196 survey activities had concluded in a community such that the intervention could have taken  
197 effect.

198

199 *Relationship between the drive distance separating pairs of communities and the risk of*  
200 *transmission between them*

201

202 We first demonstrate a relationship between the drive distance separating communities  
203 in the BCPP trial and the risk of HIV-1 transmission between them (Table 1 and Figure 1). We  
204 define the risk of transmission as the expected probability of viral-linkage between deep-  
205 sequenced HIV viruses of individuals with HIV randomly sampled from their respective  
206 communities. Figure 1 and Table 1 show that the risk of transmission decreases as the drive  
207 distance separating community pairs increases, specifically by 27% [95% Confidence Interval

208 (CI): 3 - 45] per 100 kilometers. Note that the decrease in risk of transmission per 100  
209 kilometers is computed from Table 1 as  $(27\% = 100\% * [1 - \exp(100 \text{ km} * -0.0031)])$ . Beyond  
210 the effect of distance, the risk of transmission between individuals who reside within the same  
211 community was approximately 35-fold [13 - 98] higher at baseline and 8-fold [2 - 26] higher  
212 post-baseline compared to that between individuals residing in different communities. Note  
213 that the fold-change in the risk of transmission is computed from Table 1 as  $35 = \exp(3.56)$  at  
214 baseline and  $8 = \exp(2.05)$  post-baseline.

215

216 *Estimating the relative contribution of different sources of infection residing inside versus*  
217 *outside the trial area*

218

219 Next, using this model to estimate transmissions from communities that were not in the  
220 trial, we estimate proportions of transmissions into intervention communities and control  
221 communities of the BCPP trial that occurred from individuals in the same community; different  
222 communities in the same trial arm; different communities in the opposite trial arm; and non-  
223 trial communities (see “Supplementary Appendix sections S1.1 and S1.2”). We define non-trial  
224 communities as communities outside of the 30 communities that participated in the BCPP trial.  
225 We estimated that individuals in non-trial communities accounted for most of the transmissions  
226 that occurred to recipients in trial communities, with point estimates ranging from 84% to 92%  
227 in intervention communities and 73% to 92% in control communities (Figure 2). On average,  
228 90% [95% Confidence Interval (CI): 81 – 93] of transmissions to recipients in intervention  
229 communities and 86% [74 – 90] of transmissions to recipients in control communities were

230 estimated to have sources who lived in non-trial communities (Figure 3). This finding is  
231 consistent with communities in the BCPP trial being densely surrounded by communities  
232 outside the trial area and aligns with the fact that the BCPP trial participants represented a  
233 relatively small (7.6%) proportion of the national population.

234

235 Proximity to urban centers

236

237 Communities in the BCPP trial are distributed around three major urban areas that each have  
238 relatively high numbers of people with HIV; these are Gaborone city in the South-East, Palapye  
239 in the Central-East and Francistown city in the North/North-East (Figure 2 and Supplementary  
240 Figure 2). Figure 2 shows that sexual partners in the same community had a greater impact on  
241 transmission in rural communities that are geographically isolated compared to in communities  
242 that closely neighbor major urban centers. For example, Gumare intervention community and  
243 Shakawe control community in the Northern region of Botswana received an estimated 9% [2 –  
244 30] and 22% [8 - 55] of transmissions respectively from individuals in the same community; this  
245 estimated percentage was lower for communities on the periphery of densely populated urban  
246 areas such as Oodi intervention community 2% [0.4 – 7] and Bokaa control community 5% [2 –  
247 15] in the South-East region. Furthermore, we found that the proportions of transmissions to  
248 recipients in intervention communities from individuals in the same trial arm were similar  
249 across the three major urban areas (Central-East: 5% [2 – 14], North/North-East: 6% [2 – 16],  
250 South-East: 4% [1 – 14]) ( $\chi^2 = 0.8, df = 2, P = 0.7$ ).

251

252 Impact of communities in the opposite trial arm

253

254 Individuals in control communities contributed a higher proportion of transmissions to

255 intervention communities than the reverse. For example, the proportions of transmissions to

256 recipients in intervention communities from individuals in control communities ranged with

257 point estimates from 4.2% to 5.6%, compared to those to recipients in control communities

258 from individuals in intervention communities that ranged from 1.9% to 2.4% (Figure 2). On

259 average, 5.0% [4.5 – 5.2] of transmissions to recipients in intervention communities occurred

260 from individuals in control communities compared to 2.2% [0.7 – 5.2] of transmissions to

261 recipients in control communities that occurred from individuals in intervention communities,

262 consistent with a benefit of treatment-as-prevention (Figure 3). Furthermore, Figure 3 shows

263 that on average 2.9% [0.8 – 10.4] of transmissions to recipients in intervention communities

264 occurred from individuals in the same community and 1.9% [0.6 – 4.4] of transmissions

265 occurred from individuals in other intervention communities. In comparison, 8.2% [7.6 – 24.9]

266 of transmissions to recipients in control communities occurred from individuals in the same

267 community and 4.0% [3.1 – 4.3] of transmissions occurred from individuals in other control

268 communities.

269

270

271 *Impact of a nationwide intervention*

272

273 To evaluate what the impact of a national rollout of our HIV combination prevention  
274 intervention would have been on recipients in trial communities, we estimated what the  
275 number of transmissions would have been to recipients in trial communities if all (compared to  
276 none) of the non-trial communities nationwide had also received the intervention (see  
277 Supplementary Appendix section S1.2 “Counterfactual estimates”). We found that  
278 transmissions to recipients in trial communities could have been reduced by 59% [3 – 87] if the  
279 intervention had been applied nationally (Figure 4). Note that the proportion of transmissions  
280 in trial communities that could have been averted with a national rollout is computed from  
281 modeled estimates of the total number of transmissions to recipients in trial communities from  
282 sources inside trial communities as well as from sources in communities outside the trial area  
283 (averted transmissions =  $100\% * [ (\# \text{ transmissions into trial communities from all sources}$   
284  $\text{without intervention} - \# \text{ transmissions into trial communities from all sources with intervention})$   
285  $/ \# \text{ transmissions into trial communities from all sources without intervention}]$ ). Furthermore,  
286 this finding adds evidence that the impact of the BCPP trial intervention could be substantially  
287 larger than that observed in the trial if applied nationally.

## 288 **Discussion**

289

290 Global targets set by the Joint United Nations Programme on HIV/AIDS (UNAIDS) to have fewer  
291 than 500,000 new infections by the year 2020 on a path to reach epidemic control by the year  
292 2030 were missed. Relevant to this, several large community-randomized universal HIV test-  
293 and-treat trials that were at the center of HIV prevention efforts in East and Southern Africa  
294 showed mixed results [1-5]. To aid interpretation of the complex trial results and inform public  
295 health policy decisions about effective HIV prevention strategies, we developed a statistical  
296 modeling approach that uses directed sexual contacts inferred from deep-sequenced HIV virus  
297 to quantify the relative contribution of different sources of infection, that might be inside trial  
298 communities or outside the trial area. Briefly, to demonstrate the relative extent to which  
299 transmissions in intervention communities and control communities of the BCPP trial in  
300 Botswana occurred from individuals in the same community; different communities in the same  
301 trial arm; different communities in the opposite trial arm; and communities outside the trial  
302 area we first inferred directed opposite-sex transmission events between trial communities  
303 using deep-sequence phylogenetics. Then we used the directed opposite-sex transmission  
304 events that were inferred from deep-sequence phylogenetics together with the pairwise drive  
305 distances between trial communities and the intervention status of source communities to  
306 statistically model the risk of transmission between trial communities. After that we provided  
307 pairwise drive distances between communities that participated in the 2011 Botswana  
308 population and housing census to the model as input to estimate the risk of transmission  
309 (expected probability of viral genetic-linkage had the cases been sequenced) between



310 communities nationally. Then, to estimate the number of transmissions into trial communities  
311 from all communities nationally we combined estimates of the risk of transmission to recipients  
312 in trial communities from communities nationally with population-size estimates from the 2011  
313 Botswana population and housing census [14] and district-level HIV prevalence estimates from  
314 the 2013 Botswana AIDS Impact Survey (BAIS 2013) [15].

315  
316 Power analyses and model predictions for the primary endpoints of the BCPP trial in Botswana  
317 and the PopART trial in South Africa and Zambia assumed that 20% [95% CI: 15 – 25] and 5% of  
318 sexual partnerships would involve a partner outside one’s own community, respectively [2, 8,  
319 16]. Strikingly, we found that individuals in non-intervention communities accounted for most  
320 of the transmissions that occurred to recipients in intervention communities; with an estimated  
321 90% [81 – 93] of transmissions attributable to individuals from non-trial communities and 5.0%  
322 [4.5 – 5.2] of transmissions attributable to individuals from control communities. For context, a  
323 phylogenetic study that used consensus sequences of the HIV-1 *POL* (polymerase) gene to  
324 estimate the relative contribution of local transmission versus external introductions to HIV-1  
325 incidence in the Africa Health Research Institute (AHRI) study population, a rural and peri-urban  
326 population located immediately adjacent to the TasP trial study area in KwaZulu-Natal, South  
327 Africa, estimated that 35% [20 – 60] of new infections in the study population were external  
328 introductions that occurred from sexual partners outside the study area [5, 17]. Most of the  
329 external introductions in the AHRI phylogenetics study were estimated to be from sources  
330 within the national borders of South Africa with few cross-border external introductions from  
331 Botswana, Malawi, Mozambique, Zambia and Zimbabwe (see Figure 2B in [17]). BCPP trial

332 communities closely neighbor three major urban areas in Botswana (Gaborone city in the  
333 South-East, Palapye in the Central-East and Francistown city in the North/North-East), and  
334 people tend to be fairly mobile in Botswana. By comparison, the AHRI study area is relatively  
335 distantly located from the major urban area in the KwaZulu-Natal province of South Africa (200  
336 kilometers north of Durban city), therefore, it is unsurprising that there would be more external  
337 introductions to BCPP trial communities compared to the AHRI study population. In line with  
338 the finding from the AHRI phylogenetics study, a clustering analysis conducted by the PANGEA-  
339 HIV consortium on HIV-1 viral consensus sequences from the AHRI study population in South  
340 Africa, BCPP trial in Botswana, MRC study population in Uganda, PopART study population in  
341 Zambia and Rakai study population in Uganda found few clusters including cohorts from  
342 different countries. The limited cross-border external introductions into Botswana suggest that  
343 extending the BCPP trial intervention to all communities nationally to target more sources  
344 could effectively reduce the occurrence of new infections.

345  
346 Accordingly, we estimated in a counterfactual modeling scenario to demonstrate the impact of  
347 a national application of the intervention that applying the BCPP trial intervention in all  
348 communities nationwide (versus none) could have reduced transmissions to recipients in trial  
349 communities by 59% [3 – 87]) at the time (compared with the 30% reduction that was observed  
350 in the BCPP HIV incidence cohort). This was done under an assumption that the intervention  
351 effect that was observed in intervention communities would be similar when extended to a  
352 larger geographical area and would reduce the risk of transmission between any two individuals  
353 residing in Botswana, rather than (as in the trial) any two individuals residing in intervention

354 communities. In practice, the intervention effect could vary owing to differences in  
355 transmission patterns of different population sub-groups and geographical locations. These  
356 findings suggest that substantial reductions in transmission could be achieved if the  
357 intervention is applied nationwide and that estimating the relative contribution of various  
358 sources of transmission (attributable fraction of cases) could help to guide targeted applications  
359 of the intervention where resources are limited. The timing of the implementation of a  
360 universal test-and-treat intervention could be crucial. Fast roll-out could limit the spread of  
361 infection and shorten the time to reach epidemic control. Furthermore, the estimated  
362 reductions in transmissions with a nationwide intervention suggest that the universal HIV test-  
363 and-treat intervention could be used as a foundation for incidence reduction upon which other  
364 interventions could be layered to close the gap to reach epidemic control.

365  
366 A key strength of our statistical approach is that we demonstrate how deep-sequence pathogen  
367 genomics can be used at scale to assess interventions in cluster-randomized trials of infectious  
368 disease prevention. Our analysis is based on the central assumptions that transmission patterns  
369 in communities randomized to the control arm of the trial are representative of those found in  
370 non-trial communities, and that, the population-size and HIV prevalence of communities are  
371 known; and that the HIV prevalence in administrative districts is representative of that in  
372 communities (see Supplementary Appendix section S1.1 “Population-based molecular source  
373 attribution model”). There are some limitations to our analysis: First, our statistical approach is  
374 informed by pairwise drive distances separating pairs of communities and could be improved  
375 with mobile phone data to gain insight on daily commutes and seasonal migration for work (for

376 example: farming and mining) and holidays. Second, HIV viral sequences of cases were  
377 collected only in trial communities. However, Figure 1 and Table 1 show a relationship between  
378 the drive distance separating pairs of communities in the BCPP trial and the estimated risk of  
379 transmission between them. This relationship allows us to use the drive distances separating  
380 trial communities and non-trial communities to estimate the expected probability of viral  
381 linkage to source cases in non-trial communities had cases in non-trial communities also been  
382 sequenced. Third, even though we did not explicitly model the impact of community size on  
383 risk of HIV-1 transmission to- and from- communities we found that there was generally a  
384 positive correlation between the number of transmissions to recipients in trial communities  
385 predicted by the post-baseline model in Table 1 and the opportunity for transmission to  
386 recipients in trial communities. The opportunity for transmission to recipients in a community is  
387 defined as the maximum distinct possible opposite-sex transmission pairs that could involve  
388 recipients in that community and is based on the number of people with HIV in the source and  
389 recipient communities (Supplementary Figure 4). To broaden insights, our statistical modeling  
390 approach could be applied to estimate the relative contribution of various sources of infection  
391 by age and sex in the other community-randomized universal HIV test-and-treat trials that have  
392 assembled deep-sequence genomic data, for example the PopART trial in South Africa and  
393 Zambia.

394  
395 Our findings have implications for public health policy and for the design of effective HIV  
396 prevention strategies. By deconstructing the relative contribution of different sources of  
397 infection in intervention communities versus control communities this work aids interpretation

398 of the complex universal HIV test-and-treat trials in which the intervention is administered on  
399 one group (people with HIV) and the outcome (reduction in number of new cases) is measured  
400 on another group (people in the same community without HIV). For example, our findings  
401 elucidate the potential impact of a nationwide intervention and provide insight on the extent to  
402 which the BCPP intervention was diluted by spillover infections from control communities and  
403 from communities outside the trial area. Furthermore, our findings inform on-going public  
404 health policy discussions on whether the HIV testing component in national HIV prevention  
405 programs should be centered on facility-based testing at clinics and index-based testing of  
406 family and sexual contacts of people with HIV or anchored on intensive universal household-  
407 based HIV testing as was done in the trials. For example, this study shows how the combination  
408 of universal household-based HIV testing and routine HIV testing in health facilities - as was  
409 done in the combination prevention intervention in the BCPP trial - allows us to infer  
410 transmission patterns within and between communities to guide HIV prevention strategies.

411  
412 In sum, this work shows that individuals residing in communities outside the BCPP trial area  
413 accounted for most of the transmissions to recipients in intervention communities, limiting the  
414 impact of the BCPP trial intervention. Furthermore, substantial gains in reducing transmission  
415 could be made with a nationwide application of the intervention. With the introduction of  
416 interventions at the community-level (universal test-and-treat) and individual-level (pre-  
417 exposure prophylaxis and self-testing) our analysis suggests that genomic surveillance could  
418 provide a crucial platform to assess interventions allowing us to track how pathogens spread  
419 and evolve overtime in response to different interventions. For example, samples collected for

420 routine viral load testing could also be sequenced to track the directional spread of infection  
421 within and between communities and between age-sex population sub-groups. This could help  
422 to identify population sub-groups and communities in which HIV prevention interventions need  
423 to be further strengthened to reduce transmission. Pairing genomic information with  
424 information from studies that explicitly quantify the impact of social behavioral change on  
425 interventions could aid the interpretation of HIV prevention studies and evidence-based policy  
426 design. Based on our findings, we recommend that studies of infectious disease prevention  
427 consider the impact of sources of transmission beyond the reach of the intervention when  
428 evaluating interventions to inform public health programs.

429

430

431 **Supplementary Appendix S1 Statistical approach to estimate the relative contribution of**  
432 **various sources of infection in cluster-randomized trials of infectious disease prevention**

433

434

435 S1.1 Population-based molecular source attribution model

436

437 In cluster-randomized trials of HIV prevention, where the randomization unit is communities,  
438 new infections can arise from individuals: in the same community, in different communities in  
439 the same trial arm, in different communities in the opposite trial arm and in non-trial  
440 communities. The aim of this analysis is to estimate the proportions of transmissions in trial  
441 communities that are attributable to the above-mentioned sources of infection. Suppose we  
442 would like to estimate the number of transmissions in the population,  $Z_{ij}$  that occurred to  
443 individuals in community  $i$  from individuals of the opposite-sex in community  $j$ . We refer to  
444 community  $i$  as a recipient community and community  $j$  as a source community. Notice that,  $i =$   
445  $j$  when estimating the number of transmissions that occurred from individuals in the same  
446 community. We estimate the number of transmissions into a recipient community  $i$  from a  
447 source community  $j$  at period  $t$  using two quantities,  $NH_{ij}(t)$  which is treated as known and  
448 represents the maximum number of distinct possible (opposite-sex) transmission pairs in the  
449 population between the two communities at period  $t$  and,  $\pi_{ij}(t)$  the risk of transmission  
450 between the two communities at period  $t$ , that is, the expected probability of viral-linkage  
451 between deep-sequenced HIV viruses of individuals randomly sampled from the respective  
452 communities. More precisely, we consider the model

453

454 
$$z_{ij}(t) = NH_{ij}\pi_{ij}(t), \tag{1a}$$

455 where  $\pi_{ij}(t) = \frac{E(x_{ij}^{pairs}(t))}{n_{ij}^{seq}(t)}$  and  $x_{ij}^{pairs}(t) \sim \text{Negbin}(E(x_{ij}^{pairs}(t)), \theta)$ , (1b)

456 and where  $\ln\left(\frac{E(x_{ij}^{pairs}(t))}{n_{ij}^{seq}(t)}\right) = \beta_0 + \beta_1 c_{ij} + \beta_2 s_{ij} + \beta_3 d_{ij}$ , (1c)

457 or equivalently  $\ln(E(x_{ij}^{pairs}(t))) = \ln(n_{ij}^{seq}(t)) + \beta_0 + \beta_1 c_{ij} + \beta_2 s_{ij} + \beta_3 d_{ij}$  (1d)

458

459 with data and quantities

460	$N_i$	population-size of community $i$
461	$N_{f_i}$	female population-size of community $i$
462	$N_{m_i}$	male population-size of community $i$
463	$N_j$	population-size of community $j$
464	$N_{f_j}$	female population-size of community $j$
465	$N_{m_j}$	male population-size of community $j$
466	$H_i$	HIV prevalence of community $i$
467	$H_{f_i}$	female HIV prevalence of community $i$
468	$H_{m_i}$	male HIV prevalence of community $i$
469	$H_j$	HIV prevalence of community $j$
470	$H_{f_j}$	female HIV prevalence of community $j$
471	$H_{m_j}$	male HIV prevalence of community $j$
472	$NH_i = N_i * H_i$	number of people with HIV in community $i$



473	$NH_{f_i} = N_{f_i} * H_{f_i}$	number of females with HIV in community $i$
474	$NH_{m_i} = N_{m_i} * H_{m_i}$	number of males with HIV in community $i$
475	$NH_j = N_j * H_j$	number of people with HIV in community $j$
476	$NH_{f_j} = N_{f_j} * H_{f_j}$	number of females with HIV in community $j$
477	$NH_{m_j} = N_{m_j} * H_{m_j}$	number of males with HIV in community $j$
478	$n_i^{seq}(t)$	number of individuals randomly sampled from community $i$ at
479		period $t$ whose viral whole genomes were successfully deep-
480		sequenced
481	$n_{f_i}^{seq}(t)$	number of females randomly sampled from community $i$ at
482		period $t$ whose viral whole genomes were successfully deep-
483		sequenced
484	$n_{m_i}^{seq}(t)$	number of males randomly sampled from community $i$ at
485		period $t$ whose viral whole genomes were successfully deep-
486		sequenced
487	$n_j^{seq}(t)$	number of individuals randomly sampled from community $j$ at
488		period $t$ whose viral whole genomes were successfully deep-
489		sequenced
490	$n_{f_j}^{seq}(t)$	number of females randomly sampled from community $j$ at
491		period $t$ whose viral whole genomes were successfully deep-
492		sequenced
493	$n_{m_j}^{seq}(t)$	number of males randomly sampled from community $j$ at

494 period  $t$  whose viral whole genomes were successfully deep-  
495 sequenced

496  $n_{ij}^{seq}(t) = n_{m_i}^{seq}(t) * n_{f_j}^{seq}(t) + n_{f_i}^{seq}(t) * n_{m_j}^{seq}(t)$  maximum number of distinct possible  
497 (opposite-sex) transmission pairs at period  $t$  between  
498 individuals randomly sampled from communities  $i$  and  $j$  and  
499 whose viral whole genomes were successfully deep-  
500 sequenced

501  $NH_{ij}(t) = NH_{m_i} * n_{f_j}^{seq}(t) + NH_{f_i} * n_{m_j}^{seq}(t)$  maximum number of distinct possible  
502 (opposite-sex) transmission pairs in the population at period  $t$   
503 between communities  $i$  and  $j$

504  $x_{ij}^{pairs}(t)$  number of transmission pairs identified from the deep-  
505 sequenced HIV virus of individuals randomly sampled from  
506 communities  $i$  and  $j$  at period  $t$

507  $c_{ij}$  source community is a nonintervention community, that is, a  
508 control community or non-trial community (yes = 1, no = 0)

509  $s_{ij}$  source community and recipient community are the same, that  
510 is, same community transmission (yes = 1, no = 0)

511  $d_{ij}$  distance in kilometers separating the source community and  
512 recipient community

513  
514 and estimated parameters

515  $\pi_{ij}(t) = \frac{E(x_{ij}^{pairs}(t))}{n_{ij}^{seq}(t)}$  risk of HIV transmission between communities  $i$  and  $j$  at period  $t$

516  $\theta$  overdispersion parameter

517  $\beta_0, \beta_1, \beta_2, \beta_3$  fixed effects regression parameters

518

519 For clarity, the maximum number of distinct possible (opposite-sex) transmission pairs between

520 individuals randomly sampled from communities  $i$  and  $j$  during the baseline period of the trial

521 is,  $n_{ij}^{seq}(t_{baseline}) = n_{m_i}^{seq}(t_{baseline}) * n_{f_j}^{seq}(t_{baseline}) + n_{f_i}^{seq}(t_{baseline}) * n_{m_j}^{seq}(t_{baseline})$ . By

522 comparison, the maximum number of distinct possible (opposite-sex) transmission pairs

523 between individuals randomly sampled from communities  $i$  and  $j$  during the post-baseline

524 period of the trial is,  $n_{ij}^{seq}(t_{post-baseline}) = n_{m_i}^{seq}(t_{baseline} + t_{post-baseline}) * n_{f_j}^{seq}(t_{post-baseline}) + n_{f_i}^{seq}(t_{baseline} + t_{post-baseline}) * n_{m_j}^{seq}(t_{post-baseline})$ . The central

525 assumptions of this model are that the transmission pairs identified between samples from

526 communities  $i$  and  $j$  are independent, and that transmission patterns in communities

527 randomized to the control arm of the trial are representative of those found in non-trial

528 communities. The simplifying assumption of independence is appropriate when identified

529 transmission events mostly comprise small two-person clusters as found in the BCPP data and

530 could be less-suited for large clusters typical of super spreader events. Furthermore, we assume

531 that population-size and HIV prevalence of communities  $i$  and  $j$  are known, and that the HIV

532 prevalence in administrative districts is representative of that in communities. We acknowledge

533 that community-level HIV prevalence estimates can be obtained directly with the methods of

534 [19] and reserve such computation for future study.

535

536

537 S1.2 Application to BCPP study data

538

539 We applied the population-based molecular source attribution model in section S1.1 to  
540 estimate the relative contribution of sources of infection (inside versus outside the trial area) to  
541 transmissions that occurred to recipients in BCPP trial communities. Because viral genotyping  
542 was only done in trial communities, we used census data to combine information from trial  
543 communities with that from communities outside the trial area. First, we estimated the  
544 maximum number of distinct possible (opposite-sex) transmission pairs in the population  
545 between ordered pairs (488 x 29) of the 488 communities that participated in the 2011  
546 Botswana population and housing census (see methods "*Pairwise drive distance data*" and  
547 "*Population-size and HIV prevalence estimates*"). Because there were no individuals sampled  
548 from "Digawana" intervention community during the post-baseline period whose HIV-1 virus  
549 was successfully deep-sequenced and met inclusion criteria for phylogenetic analysis, we  
550 excluded ordered community pairs that had "Digawana" as a recipient (destination)  
551 community. Afterwards, as described in equations 1b to 1d, we used directed opposite-sex  
552 transmission pairs identified between ordered pairs of the 30 communities in the BCPP trial  
553 (Supplementary Figure 1) to estimate the risk of transmission, that is, expected probability of  
554 viral linkage between those ordered community pairs during the baseline and post-baseline  
555 periods of the trial (see Table 1, Figure 1 and methods "*Deep-sequence phylogenetics data*" and  
556 "*Pairwise drive distance data*"). Estimates were obtained using parametric maximum likelihood  
557 estimation with the nbreg module in Stata 13.1 and the glm.nb function in the MASS package

558 v7.3-54 in R v4.1.2 [13]. We excluded ordered community pairs that had "Digawana" as a  
559 recipient (destination) community from the input datasets used to model the risk of  
560 transmission for the same reasons as described above, in particular, there were no individuals  
561 sampled from "Digawana" intervention community during the post-baseline period whose HIV-  
562 1 virus was successfully deep-sequenced and met inclusion criteria for phylogenetic analysis.  
563 Consequently, the input datasets used to fit the baseline and post-baseline models to estimate  
564 the risk of transmission between communities in the BCPP trial comprised 870 distinct  
565 observations ( $870 = 30 \times 30 - 30$ ) that each contained six pieces of information: 1) ordered  
566 community pair, that is, source community and recipient community, 2) non-intervention  
567 community status, 3) same community transmission status, 4) drive distance in kilometers  
568 separating the source community and recipient community, 5) number of transmission pairs  
569 identified (observed) between individuals randomly sampled from the source community and  
570 recipient community during the relevant time period, and 6) the maximum number of distinct  
571 possible (opposite-sex) transmission pairs between individuals randomly sampled from the  
572 source community and recipient community during the relevant time period. Next, we used the  
573 post-baseline model of the risk of transmission between ordered pairs of communities in the  
574 BCPP trial to predict the risk of transmission to recipients in intervention communities and  
575 control communities from the 488 census communities. The risk prediction dataset comprised  
576 ( $14,152 = 488 \times 29$ ) distinct observations that each contained five pieces of information. These  
577 were: the same first four pieces of information as those in the input dataset for the post-  
578 baseline model of the risk of transmission, and for the fifth piece of information, we set the  
579 natural-log of the maximum number of distinct possible (opposite-sex) transmission pairs

580 between individuals randomly sampled from the source community and recipient community  
581 during the relevant time period to zero, that is,  $\ln(n_{ij}^{seq}(t)) = 0$ . We set  $\ln(n_{ij}^{seq}(t)) = 0$  to  
582 predict the risk of transmission (expected probability of viral linkage) instead of the expected  
583 transmission counts. In each ordered community pair, we used the origin community in an  
584 origin-destination pairing as a surrogate for the source community and the destination  
585 community as a surrogate the recipient community. Predictions were made with the predict  
586 function in the stats package in R v4.1.2. Then, as described in equation 1a, we estimated the  
587 number of transmissions between each ordered pair of communities as the product of the  
588 maximum number of distinct possible (opposite-sex) transmission pairs in the population  
589 between the ordered community pairs and the estimated risk of transmission between them.

590

591 After estimating the number of transmissions that occurred at period  $t$  between all ordered  
592 community pairs in the population that have community  $i$  as a recipient community, we denote  
593 the total estimated number of transmissions to recipients in community  $i$  at period  $t$  as the  
594 vector

$$595 \quad \hat{z}_i(t) = (\hat{z}_{ia}(t), \hat{z}_{ib}(t), \hat{z}_{ic}(t), \hat{z}_{id}(t)), \quad (2)$$

596

597 where  $\hat{z}_{ia}(t)$  is the estimated number of transmissions from individuals within the same  
598 community, that is, community  $i$ ,  $\hat{z}_{ib}(t)$  is the estimated number of transmissions from  
599 individuals in other communities that are in the same trial arm as community  $i$ ,  $\hat{z}_{ic}(t)$  is the  
600 estimated number of transmissions from individuals in communities that are in the opposite  
601 trial arm to community  $i$  and  $\hat{z}_{id}(t)$  is the estimated number of transmissions from individuals

602 in non-trial communities. Then we estimate the proportions of transmissions to recipients in  
603 community  $i$  from individuals (sources of infection) in different types of communities as

604

605 
$$\hat{\zeta}_i^{same\ community}(t) = \frac{\hat{z}_{ia}(t)}{(\hat{z}_{ia}(t) + \hat{z}_{ib}(t) + \hat{z}_{ic}(t) + \hat{z}_{id}(t))} \quad (3a)$$

606

607 
$$\hat{\zeta}_i^{same\ trial\ arm}(t) = \frac{\hat{z}_{ib}(t)}{(\hat{z}_{ia}(t) + \hat{z}_{ib}(t) + \hat{z}_{ic}(t) + \hat{z}_{id}(t))} \quad (3b)$$

608

609 
$$\hat{\zeta}_i^{opposite\ trial\ arm}(t) = \frac{\hat{z}_{ic}(t)}{(\hat{z}_{ia}(t) + \hat{z}_{ib}(t) + \hat{z}_{ic}(t) + \hat{z}_{id}(t))} \quad (3c)$$

610

611 
$$\hat{\zeta}_i^{non-trial\ community}(t) = \frac{\hat{z}_{id}(t)}{(\hat{z}_{ia}(t) + \hat{z}_{ib}(t) + \hat{z}_{ic}(t) + \hat{z}_{id}(t))}. \quad (3d)$$

612

613

614 Suppose community  $i$  was randomized to receive the intervention then the mean estimate of  
615 the proportion of transmissions to recipients in intervention communities from individuals in  
616 the same community is

617

618 
$$\hat{\zeta}_i^{same\ community}(t) = \frac{\sum \hat{z}_{ia}(t)}{\sum \hat{z}_i(t)}. \quad (4)$$

619

620 The same principle follows for the three other sources of infection: same trial arm, opposite  
621 trial arm and non-trial community; wherein the numerator for the mean estimate is the sum of  
622 the numerators of the individual community estimates and similarly, the denominator for the

623 mean estimate is the sum of the denominators of the individual community estimates (Figures  
624 2 and 3).

625

626 Counterfactual estimates

627 To model the impact of a nationwide intervention (Figure 4) we estimated what the number of

628 transmissions to recipients in trial communities would have been if all 488 census communities

629 had received the intervention ( $c_{ij} = 0$ ) compared to if none of the communities had received

630 the intervention ( $c_{ij} = 1$ ). This was done under an assumption that the observed effect of the

631 BCPP trial intervention in intervention communities would be similar when extended to a larger

632 geographical area. However, there could be variation in the effect of the intervention that is

633 observed due to heterogeneous transmission patterns across different population sub-groups

634 and geographical locations. Because there were no individuals sampled from "Digawana"

635 intervention community during the post-baseline period whose HIV-1 virus was successfully

636 deep-sequenced and met inclusion criteria for phylogenetic analysis we excluded ordered

637 community pairs that had "Digawana" as a recipient (destination) community. We estimate that

638 a nationwide application of the intervention could have reduced transmissions to recipients in

639 trial community  $i$  by,

640

641 
$$\hat{\zeta}_i^{\text{averted transmissions}}(t) = \frac{\hat{z}^{\text{None}}_i(t) - \hat{z}^{\text{All}}_i(t)}{\hat{z}^{\text{None}}_i(t)} \quad (5)$$

642

643

644 and that on average transmissions across all trial communities could have been reduced by,



645

646 
$$\hat{\zeta}_i^{\text{averted transmissions}}(t) = \frac{\sum(\hat{z}^{\text{None}}_i(t) - \hat{z}^{\text{All}}_i(t))}{\sum \hat{z}^{\text{None}}_i(t)}. \quad (6)$$

647

648 Alternative models

649 We considered alternative versions of the post-baseline model for the risk of transmission  
650 between ordered pairs of communities in the BCPP trial that used transforms of the pairwise  
651 drive distance separating communities (Supplementary Table 1). The linear model shown in  
652 supplementary Table 1 and described in equations 1b through 1d was selected as the best  
653 model based on Aikake information criterion (AIC) and parsimony. We noted that sampling of  
654 trial participants through clinics in the BCPP trial was only done in intervention communities  
655 but not in control communities resulting in an asymmetry between the intervention and control  
656 arms of the trial. Therefore, we modeled the risk of transmission between ordered pairs of  
657 communities in the BCPP trial excluding transmission pairs where both individuals were  
658 sampled at a clinic in an intervention community during the same trial period, that is, both  
659 individuals sampled at baseline or post-baseline (Supplementary Table 2). We found similar  
660 patterns to those observed in the baseline and post-baseline models in Table 1. The negative-  
661 binomial regression models used to estimate the risk of transmission between ordered pairs of  
662 communities in the BCPP trial were fit with parametric maximum likelihood estimation with the  
663 nbreg module in Stata 13.1 and the glm.nb function in the MASS package v7.3-54 in R v4.1.2  
664 [13].

665

666

667 Model diagnostics

668 We performed several diagnostics to assess the fit of the post-baseline model in Table 1 to the  
669 directed transmission pairs identified between ordered pairs of communities in the BCPP trial.  
670 First, we assessed if the model converged to the maximum likelihood of the data using a  
671 likelihood grid search wherein the mean number of transmission pair counts predicted by the  
672 post-baseline model in Table 1 was adjusted upwards and downwards by 10% and 20%. We  
673 found that adjusting the predicted number of counts upwards or downwards did not improve  
674 the log-likelihood suggesting that the model had converged on the maximum likelihood of the  
675 data. Second, we compared the number of transmission pairs identified between ordered pairs  
676 of communities in the BCPP trial with those that would be expected under the post-baseline  
677 model in Table 1. There was little evidence to suggest that the observed counts differed  
678 substantially from those expected under the model (Fisher exact  $P = 0.792$ ) (Supplementary  
679 Table 3). Third, we also used a simulation-based approach to compare the distribution of  
680 observed quantile residuals with that expected under the post-baseline model in Table 1 and  
681 found little appreciable difference between the observed and expected distributions  
682 (Supplementary Figure 3). The simulation-based approach was performed using the DHARMA  
683 package v0.4.6 in R v4.1.2.

684

685

686 Confidence intervals

687 We used an empirical bootstrap approach to compute 95% confidence intervals for the  
688 estimated proportions of transmissions attributable to different sources of infection wherein

689 each bootstrap procedure was performed with 1,000 replicates. The lower and upper bounds of  
690 the confidence intervals represent the 2.5% and 97.5% quantiles, respectively.

691

692

693 **Data and Code availability:** All relevant data are within the paper, figures and tables. A code

694 repository has been made available at the following URL:

695 <https://github.com/magosil86/spillover-infections>

696

697

698

699

700

## References

- 701  
702
- 703 1. Makhema, J., et al., *Universal Testing, Expanded Treatment, and Incidence of HIV*  
704 *Infection in Botswana*. New England Journal of Medicine, 2019. **381**(3): p. 230-242.
- 705 2. Hayes, R.J., et al., *Effect of Universal Testing and Treatment on HIV Incidence — HPTN*  
706 *071 (PopART)*. New England Journal of Medicine, 2019. **381**(3): p. 207-218.
- 707 3. Havlir, D.V., et al., *HIV Testing and Treatment with the Use of a Community Health*  
708 *Approach in Rural Africa*. New England Journal of Medicine, 2019. **381**(3): p. 219-229.
- 709 4. Iwuji, C.C., et al., *Universal test and treat and the HIV epidemic in rural South Africa: a*  
710 *phase 4, open-label, community cluster randomised trial*. Lancet HIV, 2018. **5**(3): p.  
711 e116-e125.
- 712 5. Abdool Karim, S.S., *HIV-1 Epidemic Control — Insights from Test-and-Treat Trials*. New  
713 England Journal of Medicine, 2019. **381**(3): p. 286-288.
- 714 6. Carnegie, N.B., R. Wang, and V.D. Gruttola, *Estimation of the Overall Treatment Effect in*  
715 *the Presence of Interference in Cluster-Randomized Trials of Infectious Disease*  
716 *Prevention*. Epidemiologic Methods, 2016. **5**(1): p. 57-68.
- 717 7. Halloran, M.E., et al., *Direct and indirect effects in vaccine efficacy and effectiveness*.  
718 American Journal of Epidemiology, 1991. **133**(4): p. 323-31.
- 719 8. Wang, R., et al., *Sample size considerations in the design of cluster randomized trials of*  
720 *combination HIV prevention*. Clinical Trials, 2014. **11**(3): p. 309-318.

- 721 9. Wirth, K.E., et al., *Population uptake of HIV testing, treatment, viral suppression, and*  
722 *male circumcision following a community-based intervention in Botswana (Ya*  
723 *Tsie/BCPP): a cluster-randomised trial*. *Lancet HIV*, 2020. **7**(6): p. e422-e433.
- 724 10. Magosi, L.E., et al., *Deep-sequence phylogenetics to quantify patterns of HIV*  
725 *transmission in the context of a universal testing and treatment trial - BCPP/Ya Tsie trial*.  
726 *eLife*, 2022. **11**.
- 727 11. Ratmann, O., et al., *Inferring HIV-1 transmission networks and sources of epidemic*  
728 *spread in Africa with deep-sequence phylogenetic analysis*. *Nature Communications*,  
729 2019. **10**(1): p. 1411.
- 730 12. Wymant, C., et al., *PHYLOSCANNER: Inferring Transmission from Within- and Between-*  
731 *Host Pathogen Genetic Diversity*. *Molecular Biology and Evolution*, 2017. **35**(3): p. 719-  
732 733.
- 733 13. R Core Team, *R: A language and environment for statistical computing*. 2021, R  
734 Foundation for Statistical Computing: Vienna, Austria.
- 735 14. U.S. Census Bureau Population Division, I.D.B., September 2018 release, available at  
736 <https://www.census.gov/data-tools/demo/idb/informationGateway.php>, accessed in  
737 July 2019, *Botswana Annual Five-Year Age Group Population Estimates by Sex for 2000*  
738 *to 2025: National and First-, Second-, and Third-Order Administrative Divisions*. 2018.
- 739 15. Statistics Botswana, *Table 6: HIV Prevalence Rate by District. BOTSWANA AIDS IMPACT*  
740 *SURVEY IV 2013. Statistical report*. 2013.

- 741 16. Novitsky, V., et al., *Phylogenetic relatedness of circulating HIV-1C variants in Mochudi,*  
742 *Botswana.* PLoS One, 2013. **8**(12): p. e80589.
- 743 17. Rasmussen, D.A., et al., *Tracking external introductions of HIV using phylodynamics*  
744 *reveals a major source of infections in rural KwaZulu-Natal, South Africa.* Virus Evolution,  
745 2018. **4**(2).
- 746 18. Havlir, D., et al., *What do the Universal Test and Treat trials tell us about the path to HIV*  
747 *epidemic control?* Journal of the International AIDS Society, 2020. **23**(2): p. e25455.
- 748 19. Dwyer-Lindgren, L., et al., *Mapping HIV prevalence in sub-Saharan Africa between 2000*  
749 *and 2017.* Nature, 2019. **570**(7760): p. 189-193.
- 750
- 751

752 **Acknowledgements and Funding**

753 We are grateful to participants and collaborators from the Botswana Combination Prevention  
754 Project for their support during this work. We also thank the following colleagues for their  
755 helpful suggestions: Stephanie Marie Davis, Carol A. Ciesielski, Anindya De and Stacie Greby.  
756 This study was supported by the National Institute of General Medical Sciences  
757 (U54GM088558); the Fogarty International Center (FIC) of the U.S. National Institutes of Health  
758 (D43 TW009610); and the President’s Emergency Plan for AIDS Relief through the Centers for  
759 Disease Control and Prevention (CDC) (Cooperative agreements U01 GH000447 and U2G  
760 GH001911), NIH K24 AI131928 as well as the Morris-Singer Fund, the VK Fund for the Harvard  
761 Center for Communicable Disease Dynamics and the Bill and Melinda Gates Foundation.

762

763 **Disclaimer**

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

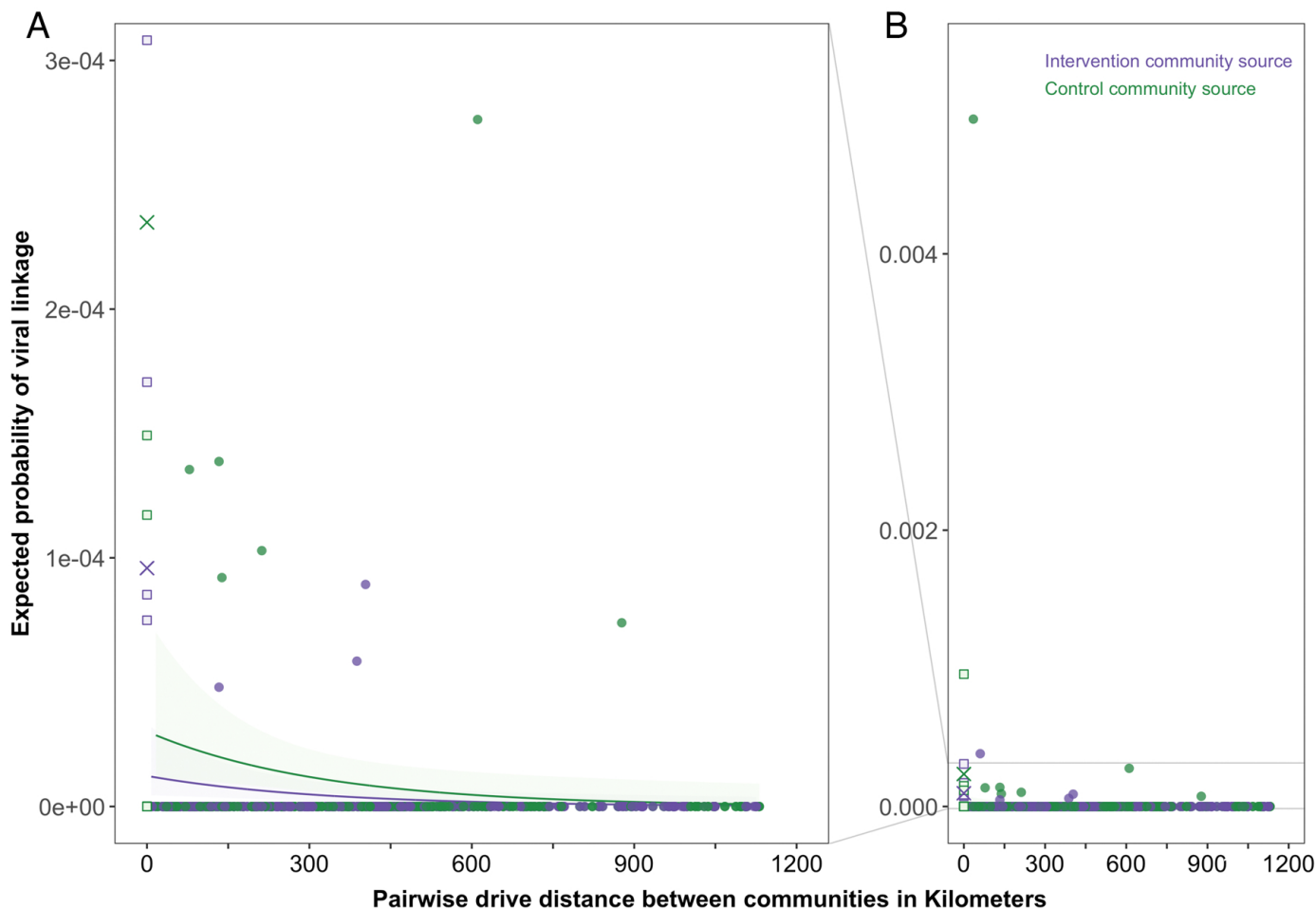
**Table 1:** Negative-binomial regression models describing the expected probability of viral linkage between a pair of individuals randomly sampled from their respective communities in the BCPP trial.

Variable	Coefficient	Standard Error	95% Conf. Interval	P value
<b>Baseline model: Before the intervention had taken effect</b>				
Intercept	-11.59	0.54	-12.65 to -10.53	< 0.001
Transmission source: <i>control community</i>	0.63	0.42	-0.20 to 1.45	0.14
Transmission type: <i>same community</i>	3.56	0.52	2.54 to 4.59	< 0.001
Drive distance between communities in kilometers	-0.0025	0.0012	-0.0047 to -0.0002	0.03
	<b>AIC</b>	215.98		
	<b>N</b>	870		
<b>Post baseline model: After the intervention had taken effect</b>				
Intercept	-11.31	0.54	-12.37 to -10.24	< 0.001
Transmission source: <i>control community</i>	0.90	0.51	-0.11 to 1.90	0.08
Transmission type: <i>same community</i>	2.05	0.61	0.86 to 3.25	0.001
Drive distance between communities in kilometers	-0.0031	0.0014	-0.0059 to -0.0003	0.03
	<b>AIC</b>	137.26		
	<b>N</b>	870		

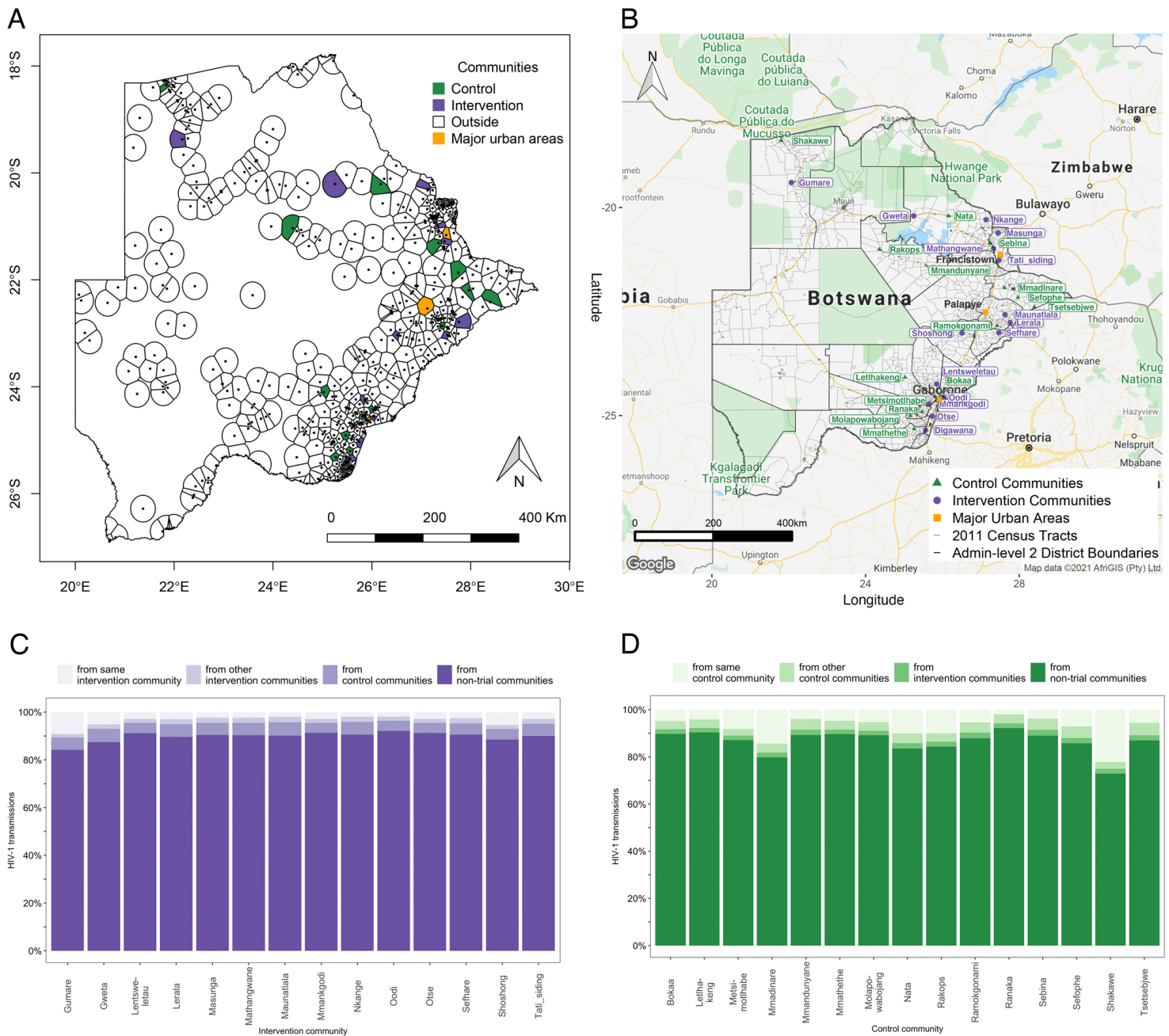
**Notes:**

1. Negative binomial regression models were fit to directed opposite-sex HIV-1 transmission pairs identified between ordered pairs of communities in the BCPP trial during the baseline ( $n = 51$ ) and post-baseline ( $n = 31$ ) periods (see methods "Deep-sequence phylogenetics data" and supplementary appendix sections S1.1 and S1.2).
2. Post-baseline is described as at least one year after baseline household survey activities had concluded in a community (see methods "BCPP study description").
3. The reference category is directed opposite-sex HIV-1 transmission pairs where individuals were randomly sampled from different communities in the BCPP trial and the source of transmission resides in an intervention community.
4. The intercept denotes the risk of HIV-1 transmission (i.e. expected probability of viral linkage) in the reference category.
5. Coefficients, standard errors and confidence bound values are presented on the linear scale and  $p$ -values are 2-sided.
6. **Transmission source: *control community*** denotes the effect on the risk of HIV-1 transmission when the source of transmission resides in a control community.
7. **Transmission type: *same community*** denotes the effect on the risk of HIV-1 transmission when both the source and recipient reside within the same community.
8. **Drive distance** between communities in kilometers denotes the effect on the risk of HIV-1 transmission for a 1 kilometer increase in the drive distance separating a pair of communities.

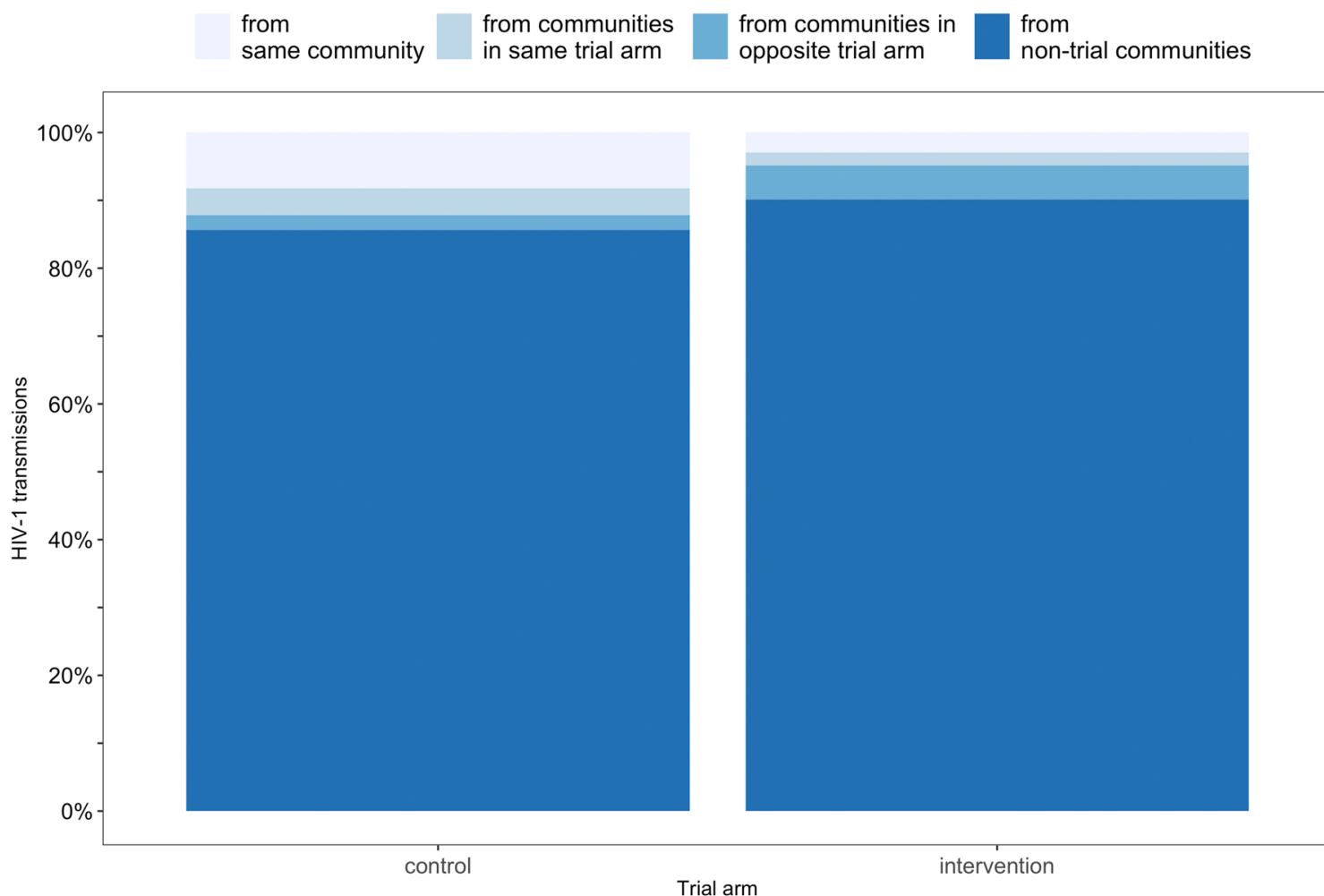




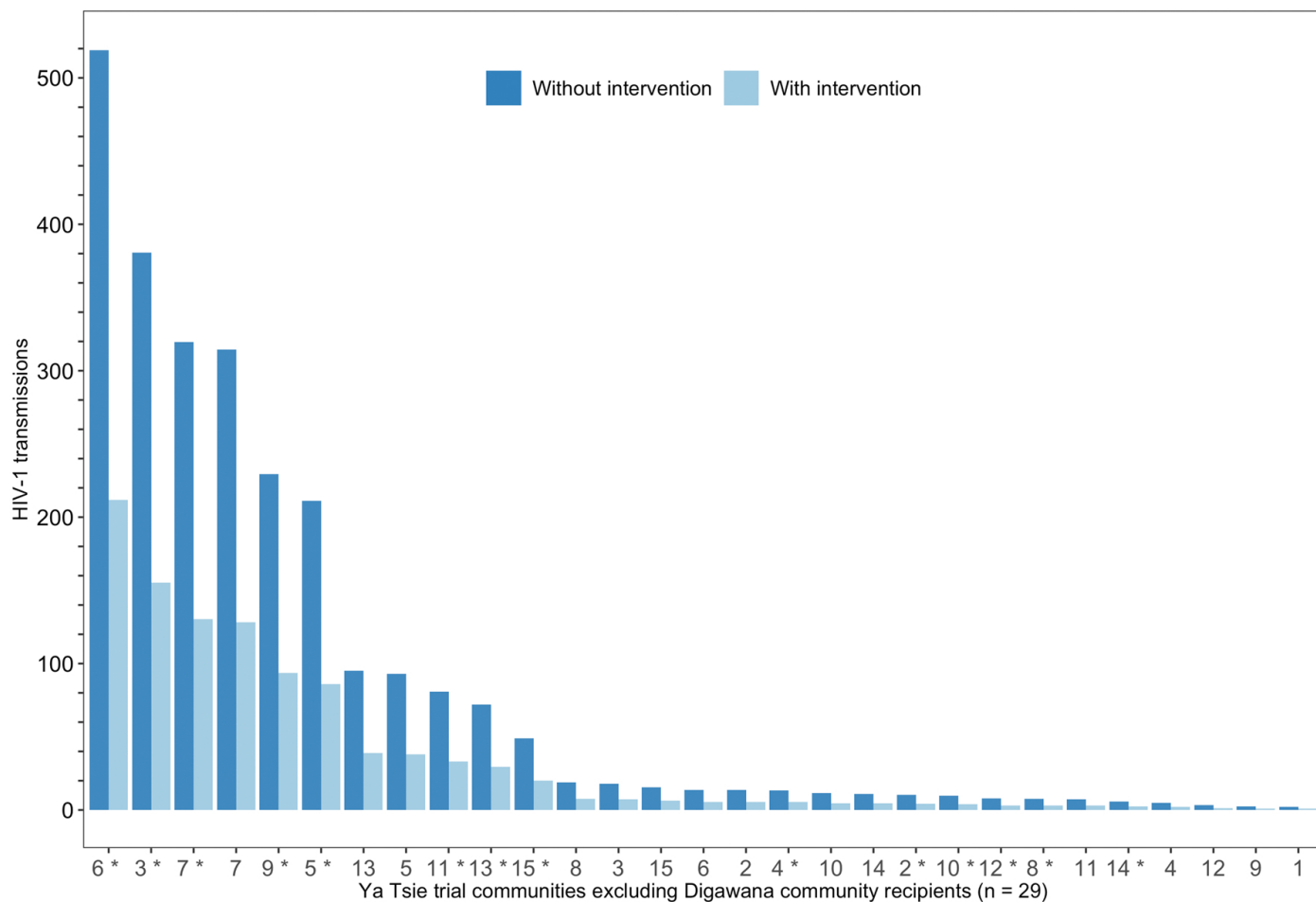
**Figure 1. The risk of HIV-1 transmission between communities in the BCPP trial decreases as the drive distance separating them increases.** The plot shows the expected probability of viral linkage, that is, risk of transmission between a pair of individuals randomly sampled from their respective communities in the BCPP trial. The expected probability of viral linkage was predicted with the post-baseline model in Table 1. To improve visibility **panel A** is a zoomed-in plot of the plot in **panel B**. Estimates for intervention community sources are shown in purple and those for control community sources are depicted in green. The solid curves and ribbons show the risk of transmission predicted by the post-baseline model between different communities in the BCPP trial and the associated uncertainty in the estimates. By comparison, the solid crosses depict the risk of transmission predicted by the post-baseline model within the same community. The squares and filled circles show the raw data for the 870 ordered community pairs of the 30 BCPP trial communities that were used to predict the expected probability of viral linkage within the same community (squares) or between different communities (filled circles). There were 15 intervention communities and 15 control communities in the BCPP trial. Because the "Digawana" intervention community had no participants with successfully sequenced samples during the post baseline period, community pairs with "Digawana" as a destination (recipient) community were excluded from the model ( $870 = 30 \times 30 - 30$ ). Among the 870 ordered community pairs, 14 were same community pairs with an intervention community as the origin (source) of transmission; 15 were same community pairs with a control community as the source of transmission, 421 were different community pairs with an intervention community as the source of transmission and 420 were different community pairs with a control community as the source of transmission. For each of the 870 ordered community pairs, the probability of viral linkage was computed from the raw data as the proportion of directed opposite-sex transmission pairs identified out of the total possible distinct opposite-sex transmission pairs among sampled participants.



**Figure 2. Estimated HIV-1 transmissions into communities in the BCPP trial from different sources of infection. Panel A.** A voronoi tessellation map of communities (n = 488) in the 2011 Botswana population and housing census showing that communities in the BCPP trial are densely surrounded by communities outside the trial area i.e. non-trial communities. To complement Panel A, **Panel B** shows the names of the intervention communities and control communities in the BCPP trial within the context of administrative districts and census tracts. **Panel C** shows, in increasing shades of purple, the estimated proportions of HIV-1 transmissions to recipients in intervention communities from individuals in: the same community, other intervention communities, control communities and from non-trial communities. Note that "Digawana" intervention community is omitted from panel C because there were no successfully sequenced post-baseline samples in the community. **Panel D** shows, in increasing shades of green, the same for recipients in control communities. Most of the transmissions to recipients in BCPP trial communities originated from individuals in non-trial communities.



**Figure 3. Mean estimates of HIV-1 transmissions that occurred to recipients in intervention communities and control communities in the BCPP trial from different sources of infection.** The barplots show, in increasing shades of blue, the estimated proportions of HIV-1 transmissions to recipients in intervention communities and control communities from individuals in the same community (intervention: 2.9% [95% CI: 0.8– 10.4], control: 8.2% [7.6– 24.9]), communities in the same trial arm (intervention: 1.9% [0.6– 4.4], control: 4.0% [3.1– 4.3]), communities in the opposite trial arm (intervention: 5.0% [4.5– 5.2], control: 2.2% [0.7– 5.2]), and from non-trial communities (intervention: 90.1% [81.1– 93.1], control: 85.6% [73.6– 90.5]). The mean estimate of the proportion of HIV-1 transmissions to recipients in intervention communities from intervention sources, that is, from individuals in the same intervention community and from individuals in other intervention communities was 4.9% [95% CI: 1.7– 14.4].



**Figure 4. Counterfactual estimates of HIV-1 transmissions into BCPP trial communities showing the impact of a nationwide intervention.** The grouped barplot shows the estimated number of transmissions to recipients in trial communities in the presence and absence of a nationwide intervention. Among the BCPP trial communities shown intervention communities are distinguished from control communities with an asterisk. The BCPP trial matched communities into 15 pairs based on geographical proximity to major urban areas, population-size and age structure, and access to health services. On average, a nationwide intervention could have reduced transmissions to recipients in trial communities by 59% [95% CI: 3– 87]. "Digawana" intervention community was excluded because there were no successfully sequenced post-baseline samples in the community that met inclusion criteria for phylogenetic analysis.

**Supplementary Table 1:** A comparison of three negative-binomial regression models that describe the expected probability of viral linkage between a pair of individuals randomly sampled from their respective communities in the BCPP trial. The models are fit (*with and without*) a transformation of the drive distance between communities.

Variable	Coefficient	Standard Error	95% Conf. Interval	P value
<b>Linear model</b>				
Intercept	-11.31	0.54	-12.37 to -10.24	< 0.001
Transmission source: <i>control community</i>	0.90	0.51	-0.11 to 1.90	0.08
Transmission type: <i>same community</i>	2.05	0.61	0.86 to 3.25	0.001
Drive distance between communities in kilometers	-0.0031	0.0014	-0.0059 to -0.0003	0.03
	<b>AIC</b>	137.26		
	<b>N</b>	870		
<b>Model with log transformed drive distance</b>				
Intercept	-8.23	1.36	-10.90 to -5.57	< 0.001
Transmission source: <i>control community</i>	0.96	0.53	-0.08 to 2.01	0.07
Transmission type: <i>same community</i>	-1.04	1.41	-3.80 to 1.73	0.46
Log <sub>e</sub> (drive distance between communities in kilometers)	-0.77	0.26	-1.28 to -0.25	0.004
	<b>AIC</b>	137.98		
	<b>N</b>	870		
<b>Model with squared drive distance</b>				
Intercept	-11.11	0.63	-12.35 to -9.88	< 0.001
Transmission source: <i>control community</i>	0.89	0.51	-0.11 to 1.90	0.08
Transmission type: <i>same community</i>	1.86	0.70	0.48 to 3.24	0.008
Drive distance between communities in kilometers	-0.0046	0.0033	-0.0110 to 0.0017	0.15
Squared (drive distance between communities in kilometers)	1.70E - 06	2.75E - 06	-3.69E - 06 to 7.08E - 06	0.54
	<b>AIC</b>	139.03		
	<b>N</b>	870		

**Notes:**

- All models were fit to directed opposite-sex HIV-1 transmission pairs (n = 31) identified between ordered pairs of communities in the BCPP trial during the post-baseline period (see methods "Deep-sequence phylogenetics data" and supplementary appendix sections S1.1 and S1.2).
- The linear model is the same as the post-baseline model in Table 1.
- The reference category is directed opposite-sex HIV-1 transmission pairs where individuals were randomly sampled from different communities in the BCPP trial and the source of transmission resides in an intervention community.
- The intercept denotes the risk of HIV-1 transmission (i.e. expected probability of viral linkage) in the reference category.
- Coefficients, standard errors and confidence bound values are presented on the linear scale and *p*-values are 2-sided.
- Transmission source: *control community*** denotes the effect on the risk of HIV-1 transmission when the source of transmission resides in a control community.
- Transmission type: *same community*** denotes the effect on the risk of HIV-1 transmission when both the source and recipient reside within the same community.
- Drive distance** in the **linear model** denotes the effect on the risk of HIV-1 transmission for a 1 kilometer increase in the drive distance separating a pair of communities.
- Drive distance** in the **log transformed distance model** denotes the effect on the risk of HIV-1 transmission with a unit increase in the Log<sub>e</sub>(drive distance) separating a pair of communities.
- Drive distance** in the **squared distance model** denotes the effect on the risk of HIV-1 transmission with a 1 km increase and a 1 km<sup>2</sup> increase in the drive distance separating a pair of communities.

**Supplementary Table 2:** Negative-binomial regression models describing the expected probability of viral linkage between a pair of individuals randomly sampled from their respective communities in the BCPP trial. Compared with Table 1, the models **exclude potential partner co-visit events to clinics in intervention communities during baseline or post-baseline**.

Variable	Coefficient	Standard Error	95% Conf. Interval	P value
<b>Baseline model: Before the intervention had taken effect</b>				
Intercept	-11.44	0.51	-12.44 to -10.44	< 0.001
Transmission source: <i>control community</i>	0.57	0.31	-0.05 to 1.18	0.07
Transmission type: <i>same community</i>	3.46	0.50	2.48 to 4.43	< 0.001
Drive distance between communities in kilometers	-0.0026	0.0011	-0.0049 to -0.0004	0.02
	<b>AIC</b>	214.21		
	<b>N</b>	870		
<b>Post baseline model: After the intervention had taken effect</b>				
Intercept	-11.50	0.52	-12.51 to -10.48	< 0.001
Transmission source: <i>control community</i>	1.30	0.48	0.35 to 2.25	0.01
Transmission type: <i>same community</i>	2.09	0.63	0.86 to 3.32	0.001
Drive distance between communities in kilometers	-0.0038	0.0014	-0.0065 to -0.0011	0.01
	<b>AIC</b>	151.65		
	<b>N</b>	870		

**Notes:**

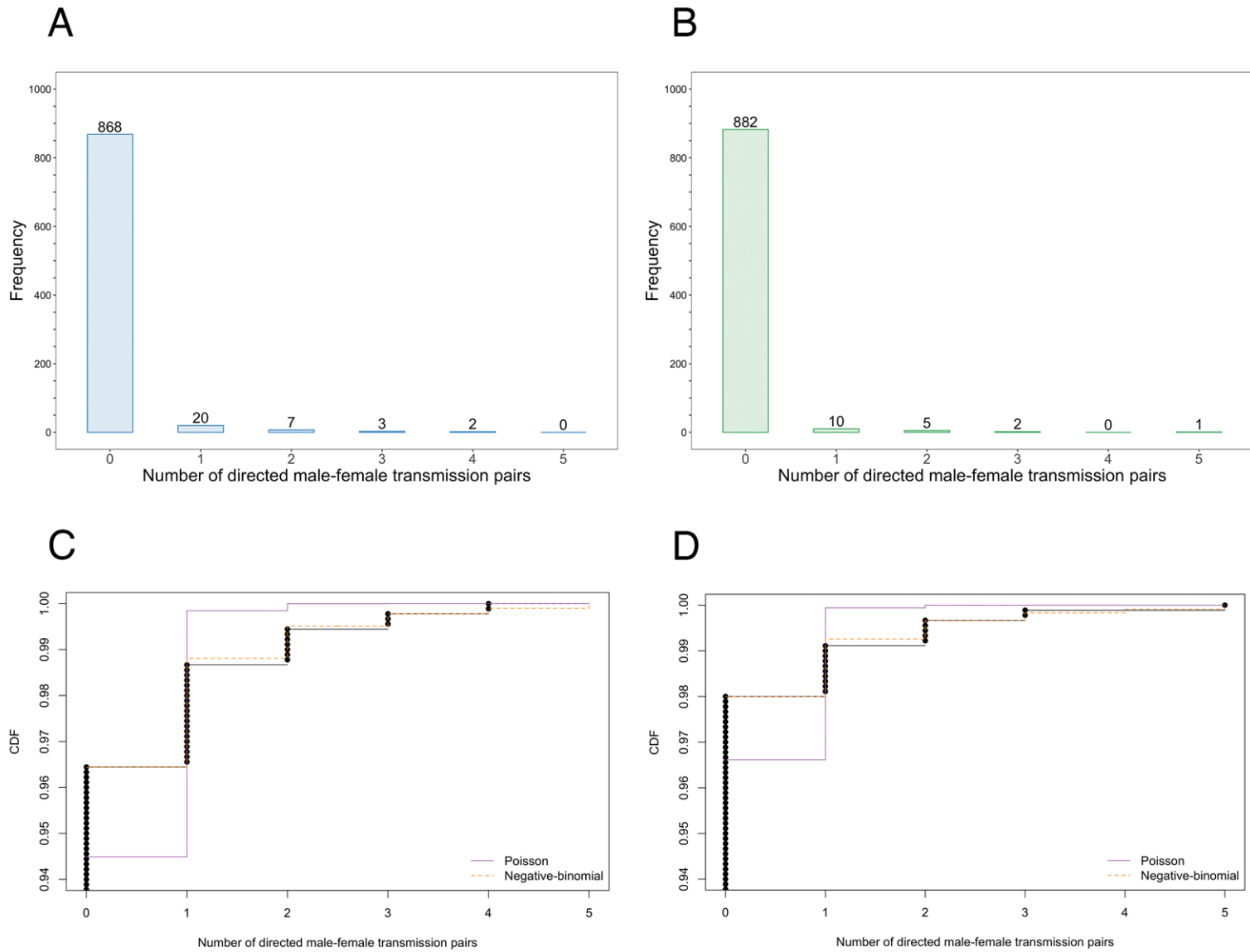
1. Negative binomial regression models were fit to directed opposite-sex HIV-1 transmission pairs identified between ordered pairs of communities in the BCPP trial at baseline (n = 49 / 51) and post-baseline (n = 27 / 31), excluding transmission pairs where both individuals visited a clinic within the same intervention community during the same period of the trial i.e. baseline or post-baseline (see supplementary appendix sections S1.1 and S1.2).
2. Post-baseline is described as at least one year after baseline household survey activities had concluded in a community.
3. The reference category is directed opposite-sex HIV-1 transmission pairs where individuals were randomly sampled from different communities in the BCPP trial and the source of transmission resides in an intervention community.
4. The intercept denotes the risk of HIV-1 transmission (i.e. expected probability of viral linkage) in the reference category.
5. Coefficients, standard errors and confidence bound values are presented on the linear scale and *p*-values are 2-sided.
6. **Transmission source: control community** denotes the effect on the risk of HIV-1 transmission when the source of transmission resides in a control community.
7. **Transmission type: same community** denotes the effect on the risk of HIV-1 transmission when both the source and recipient reside within the same community.
8. **Drive distance** between communities in kilometers denotes the effect on the risk of HIV-1 transmission for a 1 kilometer increase in the drive distance separating a pair of communities.

**Supplementary Table 3:** A comparison of the number of transmission pairs identified in the BCPP trial during the post-baseline period ( $n = 31$ ) with those expected under the post-baseline model in Table 1. Table 1 describes the risk of HIV-1 transmission between communities.

Pairwise drive distance bin in kilometers	Median of bin	Number of distinct possible transmission pairs in bin	Number of identified transmission pairs	Number of expected transmission pairs
[0, 50]	25	226,500	19	26
(50, 120]	85	205,621	3	3
(120, 220]	170	209,215	5	3
(220, 320]	270	232,686	0	2
(320, 410]	365	256,466	2	2
(410, 475]	442.5	220,073	0	1
(475, 710]	592.5	197,104	1	1
(710, 915]	812.5	210,298	1	0
(915, 1045]	980	196,782	0	0
(1045, 1200]	1122.5	235,449	0	0

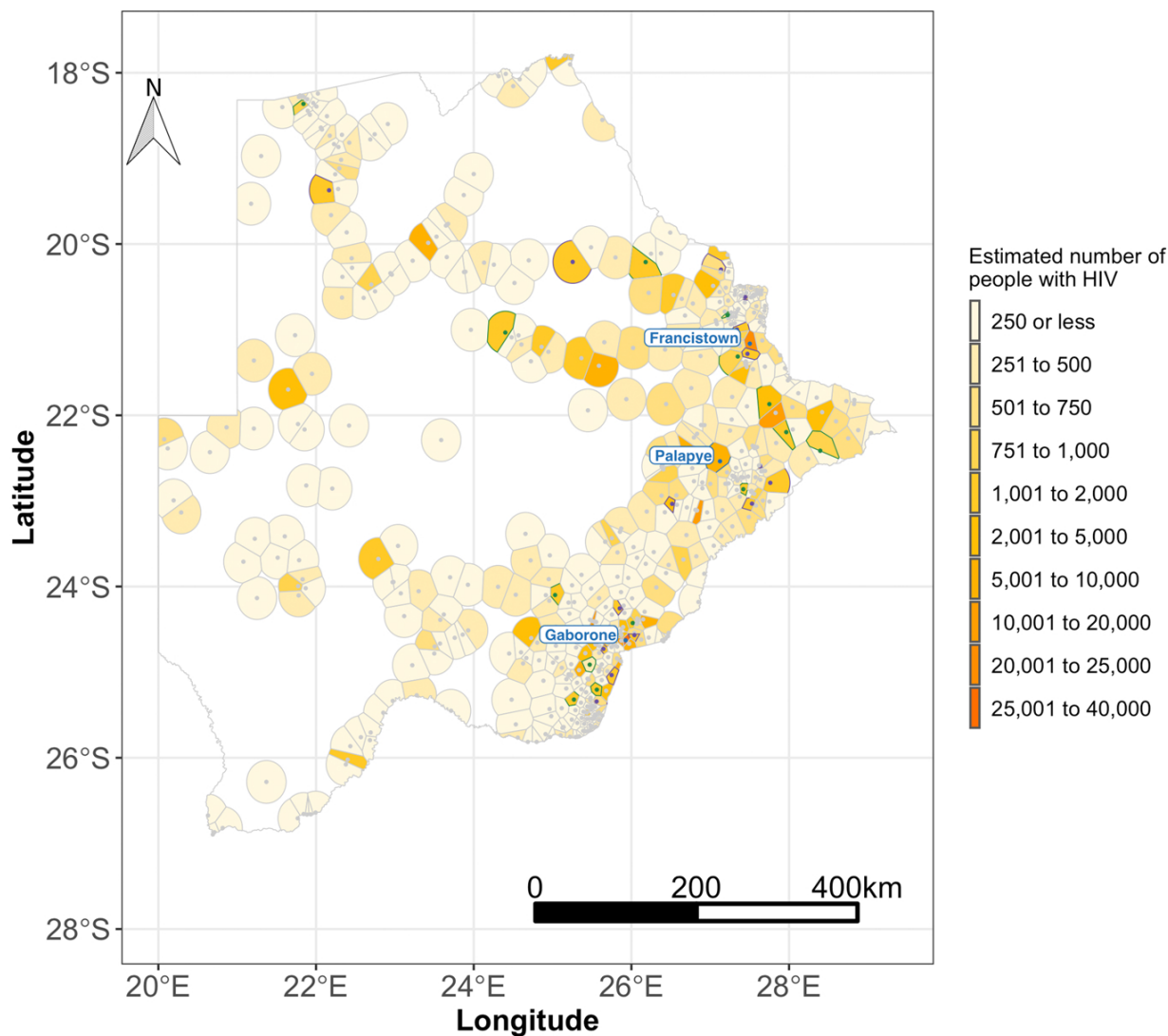
**Notes:**

1. The pairwise drive distances separating communities in the BCPP trial were grouped into ten bins that each had approximately similar numbers of the maximum number of distinct possible opposite-sex transmission pairs (see Model diagnostics in supplementary appendix section S1.2).
2. The expected number of transmission pairs was estimated with the post-baseline model in Table 1.
3. There was little evidence to suggest that the observed counts differed substantially from those expected under the post-baseline model in Table 1 (Fisher exact  $P = 0.792$ ).

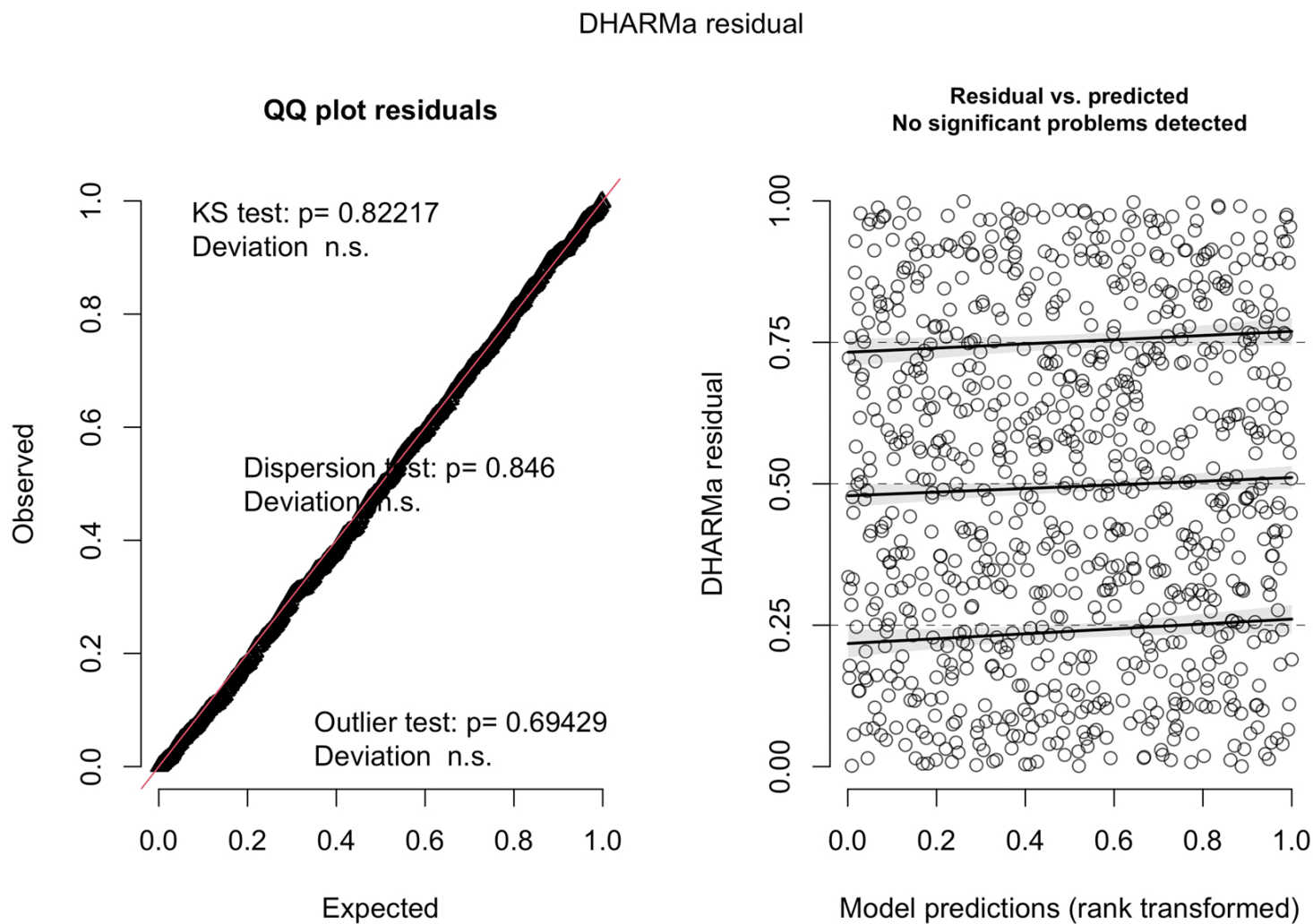


**Supplementary Figure 1. Distribution of directed opposite-sex transmission pairs identified between ordered pairs of the 30 communities in the BCPP trial.** The barplots in panels **A** and **B** show distributions of transmission pairs identified during the baseline (blue bars) and post-baseline (green bars) periods of the BCPP trial, respectively. For example, in panel A there were 20 ordered community pairs that each had a single identified opposite-sex transmission pair and 7 ordered community pairs that each had 2 identified transmission pairs. Panels **C** and **D** show the corresponding empirical and theoretical cumulative distribution functions (cdf) of identified transmission pairs during the baseline (**panel C**) and post-baseline (**panel D**) periods of the BCPP trial. In both panels **C** and **D** the empirical distribution is shown in black, and the theoretical Poisson and negative-binomial distributions are illustrated by solid purple lines and orange broken lines, respectively. The positively skewed distributions of identified transmission pairs in the BCPP trial are better approximated with a negative-binomial distribution compared to a Poisson distribution.

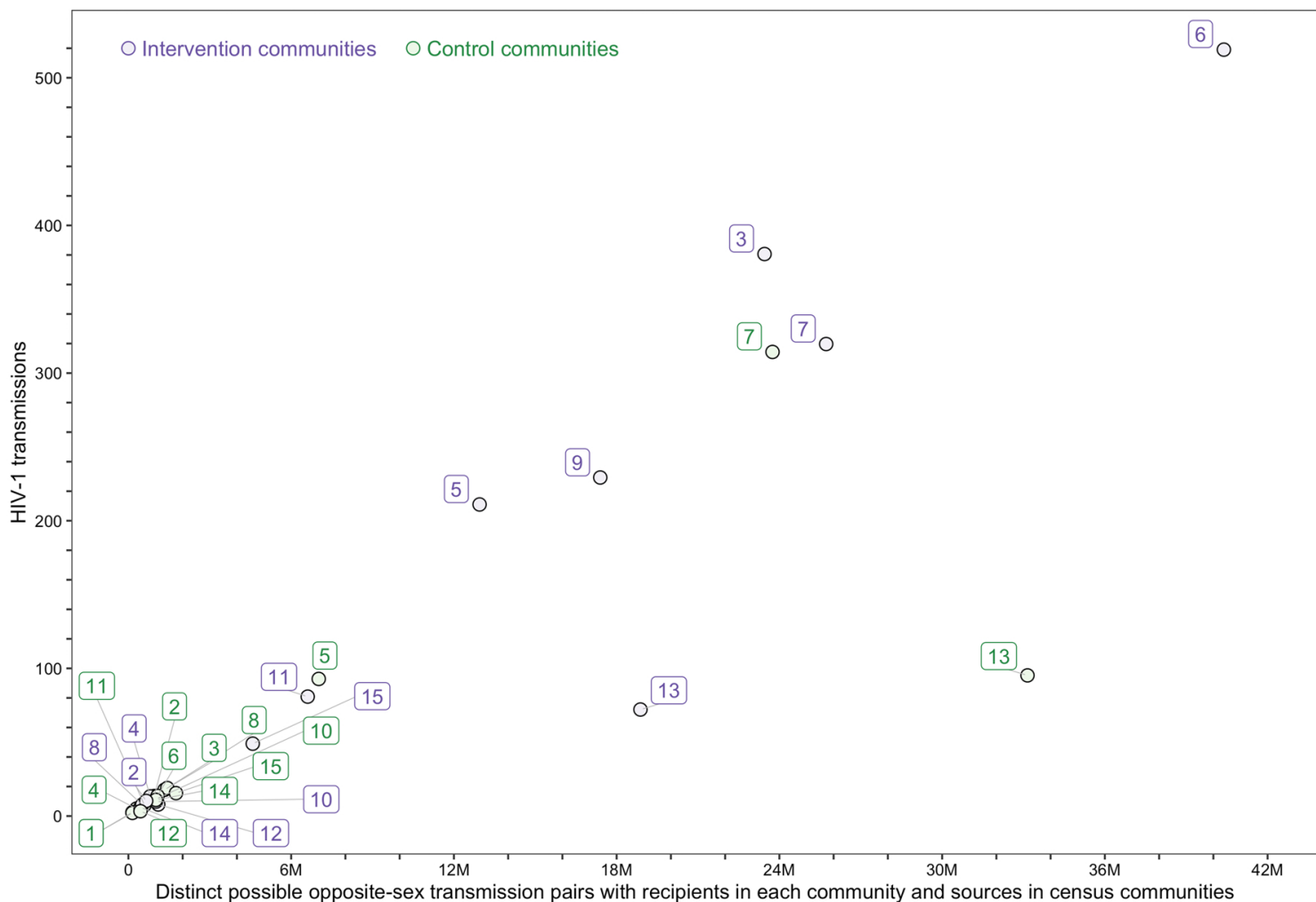




**Supplementary Figure 2. Spatial distribution of the estimated number of people with HIV-1 in Botswana.** Estimates of the number of people with HIV-1 were computed from district HIV-1 prevalence estimates from the 2013 Botswana AIDS Impact Survey (BAIS 2013) and community-size estimates from the 2011 Botswana population and housing census. Intervention communities in the BCPP trial are denoted by purple filled circles and boundaries and control communities are represented by green filled circles and boundaries. The communities in the BCPP trial are distributed around three major urban areas: Gaborone city, Palapye and Francistown city represented by blue filled circles and labels.



**Supplementary Figure 3. A quantile-quantile (QQ) residual plot that compares the distribution of residuals of 31 opposite-sex transmission pairs identified between ordered pairs of communities in the BCPP trial during the post-baseline period with those that would be expected under the post-baseline model in Table 1.**



**Supplementary Figure 4. A scatter plot that shows a positive correlation between the estimated number of HIV-1 transmissions to recipients in trial communities that were predicted by the post-baseline model in Table 1 in the absence of the intervention and the total possible distinct opposite-sex transmission pairs that involve recipients in trial communities.** The maximum (or total) distinct possible opposite-sex transmission pairs that involve recipients in trial communities represent the opportunity for transmission to recipients in the the trial communities, and are computed from the community size and HIV-1 prevalence of the source community and the number of people with HIV-1 sampled from the recipient community during the post-baseline period when the intervention could have taken effect. The BCPP trial matched communities into 15 pairs based on geographical proximity to major urban areas, population-size and age structure, and access to health services.