Phylogenetic Tools for Generalized HIV-1 Epidemics: Findings from the PANGEA-HIV Methods Comparison

Oliver Ratmann,*[,1] Emma B. Hodcroft,[2] Michael Pickles,[1] Anne Cori,[1] Matthew Hall,[2,4] Samantha Lycett,[2,3] Caroline Colijn,[5] Bethany Dearlove,[6] Xavier Didelot,[1] Simon Frost,[6] A.S. Md Mukarram Hossain,[6] Jeffrey B. Joy,[7,8] Michelle Kendall,[5] Denise Kühnert,[9,10] Gabriel E. Leventhal,[9,11] Richard Liang,[8] Giacomo Plazzotta,[5] Art F.Y. Poon,[12] David A. Rasmussen,[10] Tanja Stadler,[10] Erik Volz,[1] Caroline Weis,[10] Andrew J. Leigh Brown,[2] and Christophe Fraser[1,4] on behalf of the PANGEA-HIV Consortium

[1]Department of Infectious Disease Epidemiology, MRC Centre for Outbreak Analyses and Modelling, School of Public Health, Imperial College London, London, United Kingdom

[2]School of Biological Sciences, Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom

[3]The Roslin Institute, University of Edinburgh, Edinburgh, United Kingdom

[4]Nuffield Department of Medicine, Li Ka Shing Centre for Health Information and Discovery, Oxford Big Data Institute, University of Oxford, Oxford, United Kingdom

[5]Department of Mathematics, Imperial College London, London, United Kingdom

[6]Department of Veterinary Medicine, Cambridge Veterinary School, Cambridge, United Kingdom

[7]Department of Medicine, University of British Columbia, Vancouver, BC, Canada

[8]British Columbia Centre for Excellence in HIV/AIDS, Vancouver, BC, Canada

[9]Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland

[10]Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland

[11]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology (MIT), Cambridge, MA

[12]Department of Pathology & Laboratory Medicine, Western University, Ontario, Canada

***Corresponding author:** E-mail: oliver.ratmann@imperial.ac.uk.

Associate editor: Jeffrey Townsend

Abstract

Viral phylogenetic methods contribute to understanding how HIV spreads in populations, and thereby help guide the design of prevention interventions. So far, most analyses have been applied to well-sampled concentrated HIV-1 epidemics in wealthy countries. To direct the use of phylogenetic tools to where the impact of HIV-1 is greatest, the Phylogenetics And Networks for Generalized HIV Epidemics in Africa (PANGEA-HIV) consortium generates full-genome viral sequences from across sub-Saharan Africa. Analyzing these data presents new challenges, since epidemics are principally driven by heterosexual transmission and a smaller fraction of cases is sampled. Here, we show that viral phylogenetic tools can be adapted and used to estimate epidemiological quantities of central importance to HIV-1 prevention in sub-Saharan Africa. We used a community-wide methods comparison exercise on simulated data, where participants were blinded to the true dynamics they were inferring. Two distinct simulations captured generalized HIV-1 epidemics, before and after a large community-level intervention that reduced infection levels. Five research groups participated. Structured coalescent modeling approaches were most successful: phylogenetic estimates of HIV-1 incidence, incidence reductions, and the proportion of transmissions from individuals in their first 3 months of infection correlated with the true values (Pearson correlation $>$ 90%), with small bias. However, on some simulations, true values were markedly outside reported confidence or credibility intervals. The blinded comparison revealed current limits and strengths in using HIV phylogenetics in challenging settings, provided benchmarks for future methods' development, and supports using the latest generation of phylogenetic tools to advance HIV surveillance and prevention.

Key words: **HIV transmission and prevention, molecular epidemiology of infectious diseases, viral phylogenetic methods validation.**

Introduction

Recent breakthroughs in human immunodeficiency virus type 1 (HIV-1) prevention and treatment have provided a range of tools to reduce HIV-1 transmission (WHO 2015). Incorporating these strategies into routine care services and delivering on the commitment to end the HIV-1 epidemic by 2030 remains a major challenge (UNAIDS 2014), particularly in sub-Saharan Africa where the burden of HIV-1 is greatest.

Open Access

Article

This region suffers 75% of all new HIV-1 infections worldwide, with adult HIV-1 prevalence exceeding 25% in some regions, and averaging ~5% overall (UNAIDS 2015). To sustain public health interventions at this scale with limited resources, a sufficiently detailed understanding of the local and regional drivers of HIV-1 spread is often indispensable. Universal prevention packages (Iwuji et al. 2013; Hayes et al. 2014) benefit from data that allows monitoring incidence trends and drivers of residual spread, whereas more targeted prevention approaches (Vassall et al. 2014) by definition require a detailed knowledge of at-risk populations.

The Phylogenetics And Networks for Generalized HIV Epidemics in Africa (PANGEA-HIV) consortium aims to provide viral sequence data from across sub-Saharan Africa, and to evaluate their viral phylogenetic relationship as a marker of recent HIV-1 transmission dynamics (Pillay et al. 2015). Previous molecular epidemiological studies indicate that this approach can characterize transmission landscapes across a diverse array of epidemic contexts in order to guide prevention efforts (Fisher et al. 2010; Kouyos et al. 2010; von Wyl et al. 2011; Stadler et al. 2013; Volz et al. 2013; Grabowski et al. 2014; Bezemer et al. 2015; Ratmann et al. 2016). Rather than the partial gene sequences frequently used, the consortium is generating near full-length HIV-1 sequences in order to further increase the resolution and power of viral phylogenomic methods (Dennis et al. 2014). Indeed, such increases in power are needed to disentangle signal from noise in epidemic settings with frequent co-infection and recombination events (Grabowski et al. 2014), and to shift focus to recent transmission dynamics (Dennis et al. 2014).

Available viral phylogenetic techniques can provide estimates of key epidemiological quantities of concentrated HIV-1 epidemics (Brenner et al. 2007; Fisher et al. 2010; Stadler and Bonhoeffer 2013; Volz et al. 2013; Bezemer et al. 2015; Ratmann et al. 2016). But the generalized epidemics in sub-Saharan Africa and sequence availability in these resource-poor settings differ fundamentally from well sampled concentrated epidemics in wealthy countries, where viral phylogenetic tools have been proven to be most effective to date (Dennis et al. 2014). To strengthen the application of viral phylogenetics in sub-Saharan Africa, in October 2014 PANGEA-HIV invited research groups to participate in a blinded methods comparison exercise. Two individual-level HIV epidemic models were used to simulate generalized HIV-1 epidemics. From these, we generated corresponding viral sequence datasets comprising simulated *pol*, *gag* and *env* genes (which we refer to as full genome sequences for brevity), as well as basic individual-level epidemiological data on those infected individuals that were sequenced in the simulations. External research groups then analyzed the blinded data.

Overall, we aimed to evaluate if the most recent generation of viral phylogenetic tools could be adapted and used to estimate epidemiological quantities of central importance to HIV-1 prevention in sub-Saharan Africa. The specific objectives were inspired by current HIV-1 prevention trials in sub-Saharan Africa (Iwuji et al. 2013; Moore et al. 2013; Hayes et al. 2014). The primary goal of these trials is to achieve substantial reductions in HIV-1 incidence over a short period. Viral phylogenetics could be an effective tool to measure similar reductions, especially in contexts where incidence cohorts do not exist, and thereby contribute to monitoring the impact of prevention strategies. First, participants were asked to estimate recent reductions in HIV-1 incidence resulting from a simulated community-based intervention over a 3- to 5-year period. Here, incidence was defined as the proportion of new cases per year among uninfected adults, and reductions in incidence were measured in terms of the incidence ratio before and after the intervention. Second, it has been debated whether frequent transmission during the early acute phase of HIV infection could undermine the impact in reducing incidence of universal test and treat (Cohen et al. 2012). In concentrated epidemics, viral phylogenetics based on partial *pol* sequences have been used to provide estimates of the proportion of transmissions arising from individuals in their first year of infection (Volz et al. 2013; Ratmann et al. 2016). Here, we sought to evaluate whether viral phylogenetics based on full-genome sequences can provide accurate estimates of the proportion of transmissions from individuals in early and acute HIV (defined here as in their first 3 months of infection), because these are likely not preventable in current prevention trials where testing intervals are 1 year or more (Iwuji et al. 2013; Moore et al. 2013; Hayes et al. 2014). Third, as sequence data are now collected as part of HIV-1 prevention trials (HPTN 071 (PopART) Phylogenetics Protocol Team 2015; Novitsky et al. 2015), different approaches to prospective sequence sampling have emerged. Sequences could be collected at high coverage in villages or smaller townships at the risk of missing long-range transmissions, or at lower coverage over geographically much larger areas. We sought to compare the impact of these sampling strategies on viral phylogenetic analyses by simulating epidemics in village and larger regional populations, and sampling sequences at high and low coverage respectively. Other objectives included evaluating the benefit of using concatenated HIV-1 sequences comprising simulated *pol*, *gag* and *env* genes, as compared with using simulated *pol* sequences alone, and the impact of frequent viral introductions into the modeled population as a result of long-distance transmission. Table 1 describes the objectives and reporting variables of the exercise more fully.

Five external research groups participated in the exercise, out of eight teams that initially indicated interest. Table 2 lists the phylogenetic methods that were used: the ABC-kernel method (A. Poon, J. Joy, R. Liang; team Vancouver) (Poon 2015), the birth-death skyline method with sampled ancestors (C. Weis, G.E. Leventhal, D. Kühnert, D.A. Rasmussen, T. Stadler; team Basel-Zürich) (Gavryushkina et al. 2014; Kühnert et al. 2016), a metapopulation coalescent approach (B. Dearlove, M. Hossain, S. Frost; team Cambridge) (Dearlove and Wilson 2013), the structured coalescent (E. Volz, M. Hossain, S. Frost; team Cambridge-London) (Volz et al. 2009), and a Bayesian transmission chain analyser (C. Colijn, M. Kendall, X. Didelot, G. Plazotta; team London) (Didelot et al. 2014). These methods differed in the underlying transmission and intervention models, assumptions to facilitate estimation of the reporting variables, and computational estimation routines. Here, we

Table 1. Aims of the PANGEA Phylodynamic Methods Comparison Exercise.

Objectives		Reporting Variable
Primary objectives		
1	Identify incident trends during the intervention	Consider the year t_s before the intervention started, and the second last year t_e of the simulation. Participants were asked to report HIV-1 incidence trends from t_s to t_e in terms of "declining", "stable", "increasing"
2	Estimate HIV-1 incidence after the intervention	Participants were asked to report %Incidence defined as $\%INC(t_e) = INC(t_e)/S(t_e)$, where $INC(t_e)$ is the number of new cases in year t_e, and $S(t_e)$ is the number of sexually active individuals that were not infected in year t_e
3	Quantify the reduction in HIV-1 incidence at the end of the intervention	Participants were asked to report the incidence ratio $\%INC(t_e)/\%INC(t_s)$
4	Estimate the proportion of transmissions from early and acute HIV, just before the intervention	Participants were asked to report the proportion of new cases in year t_s from individuals in their first 3 months of infection
5	Estimate the proportion of transmissions from early and acute HIV, after the intervention.	Participants were asked to report the proportion of new cases in year t_e from individuals in their first 3 months of infection
Secondary objectives		
To estimate the impact of the following controlled covariates on the reporting variables:		
6	Availability of full genome sequences (HIV-1 *gag*, *pol* and *env* genes) as compared with partial sequences (HIV-1 *pol* gene only)	
7	Sequence sampling frame: Sequence coverage at the end of the simulation; Rapid increases in sequence coverage; Sampling duration after intervention start	
8	Frequency of viral introductions into the modeled study population	
9	Inference of dated viral phylogenies from sequence data	

summarize the findings of the exercise, and discuss their implications for using phylogenetic methods to estimate recent aspects of HIV-1 transmission dynamics in generalized epidemics. Datasets and simulations generated here may be of use for testing other applications of viral phylogenetic methods, and are made available alongside this article.

Results

PANGEA-HIV Reference Datasets for Benchmarking Molecular Epidemiological Transmission Analysis Methods

The simulations capture a variety of transmission and intervention scenarios across two demographic settings in sub-Saharan Africa, and are available from https://dx.doi.org/10.6084/m9.figshare.3103015 (last accessed October 14, 2016).

20 datasets correspond to generalized HIV-1 epidemics in a region of ~80,000 individuals between 1980 and 2020 (table 3). The proportion of infected individuals of whom one sequence was sampled (sequence coverage) was 8–16% by the end of the simulation. These data were simulated under the individual-based HPTN071 (PopART) model, version 1.1, developed at Imperial College London ("Regional" model). The overall simulation pipeline and model components are illustrated in figure 1, and further information is provided in supplementary table S1, Supplementary Material online. The Regional model was calibrated to generate an epidemic with a comparable prevalence at the start of the intervention to that seen currently in HPTN071 (PopART) trial sites in South Africa (Hayes et al. 2014). In the model, standard of care improved according to national guidelines over time, resulting in steady declines in incidence. In 18 of the 20 simulations, a combination prevention intervention was started in 2015 for 3 years at varying degrees of uptake and coverage, resulting in 30% or 60% reductions in incidence relative to the start of the intervention, when incidence was close to 2% per year. In half of the 20 simulations, the proportion of early transmissions in 2015 was respectively calibrated to 10% and 40% (fig. 2). Ranges in incidence reduction reflect modeled, optimistic and pessimistic scenarios in ongoing prevention trials in sub-Saharan Africa (Iwuji et al. 2013; Moore et al. 2013; Hayes et al. 2014). The proportion of transmissions from early and acute HIV has been challenging to estimate without sequence data, and the ranges used here reflect estimates from several settings in sub-Saharan Africa (Cohen et al. 2012). About 5–20% of all transmissions per year occurred from outside the model population, which hindered prevention efforts in the simulations through continual replenishment of the epidemic.

13 simulated datasets capture generalized HIV-1 epidemics over 45 years in a smaller village population of ~8,000 individuals (table 3). Sequence coverage was higher in this smaller population, 25–50% by the end of the simulation. These data were simulated under an individual-based household model using the Discrete Spatial Phylo Simulator for HIV, developed at the University of Edinburgh ("Village" model). Model components are illustrated in figure 1, and further information is provided in supplementary table S2, Supplementary Material online. The Village model was parameterized to simulate an HIV-1 epidemic mostly contained within a small rural African village, with a peak prevalence of 20–25% and peak incidence of 5–7% without treatment (fig. 2). In 12 out of 13 simulations, a community-level intervention providing antiretroviral treatment took place for the last 5 years of the simulation.

Table 2. Phylogenetic Methods Used in the PANGEA Phylodynamic Methods Comparison Exercise.

Team	Team Members	Method	Model-based analysis	Model Overview	Simulated Data Used To Inform Inference	Fitting Process	Availability
Basel-Zürich	C. Weis, G.E. Leventhal, D. Kühnert, D.A. Rasmussen, T. Stadler	Birth–death skyline method with sampled ancestors	Yes	Stochastic birth–death model with sampled ancestors to estimate incidence and incidence reductions, and multi-type birth death model corresponding to two stages of infection to estimate the proportion of early transmissions. Time trends in parameters were modeled with serial time intervals during which parameters were assumed constant. Viral introductions were not modeled	All sequences and full trees to estimate birth–death parameters; cross-sectional survey data	Markov Chain Monte Carlo	http://beast2.org/ (last accessed October 14, 2016) using add-ons bdsky, SA, bdmm
Cambridge	B. Dearlove, M. Hossain, S. Frost	Meta-population coalescent approach	Yes	Standard SI, SIS and SIR models were averaged. Model parameters did not change over time. Viral introductions were not modeled.	All sequences and full trees.	Markov Chain Monte Carlo	http://beast.bio.ed.ac.uk/ (last accessed October 14, 2016) using XML specification described in (Dearlove and Wilson 2013)
Cambridge-London	E. Volz, M. Hossain, S. Frost	Structured coalescent	Yes	Deterministic compartment model stratified by gender, disease progression, diagnosis and treatment status, risk behavior. Time trends in baseline transmission rates were modeled with 4-parameter generalized logistic function. Diagnosis and treatment uptake rates changed at intervention start. Viral introductions were modeled with a source deme.	All sequences and subtrees including all internal nodes 30 before the last sample; cross-sectional survey data; and gender and CD4 count at time of diagnosis for Regional datasets.	Parallel Markov Chain Monte Carlo	http://colgem.r-forge.r-project.org/ (last accessed October 14, 2016)
London	C. Colijn, M. Kendall, G. Plazotta, X. Didelot	Bayesian transmission chain analyzer	Yes	Stochastic generalized branching model with generation time modeled to represent three infection stages. Model parameters did not change over time. Viral introductions were not modeled.	All sequences and full trees on village datasets; sequences in trees with at least 80 tips on regional datasets.	Reversible-jump Markov Chain Monte Carlo	https://github.com/xavierdidelot/TransPhylo (last accessed October 14, 2016) PANGEA release available from authors
Vancouver	A. Poon, J. Joy, R. Liang	ABC kernel method	Yes	Deterministic compartment model stratified by infection status, three stages of infection, and risk behavior. Model parameters did not change over time. Viral introductions were modeled with a source deme.	All sequences and full trees.	Approximate Bayesian Computation	https://github.com/ArtPoon/kamphir (last accessed October 14, 2016) PANGEA release available from authors

Table 3. Simulated Datasets of the Phylodynamic Methods Comparison Exercise.

Model	Dataset	Purpose	%Acute[a,b] (Low=L, High=H)	Intervention Scale Up[a,c] (Fast=F, Slow=S)	Viral Introductions[a,d] (% of All Transmissions per Year)	Sequences (#)	Sequence Coverage in the Last Year of the Simulation[a,e] (% of All Infected and Alive)	Sequences After Intervention Start[f] (% of All Sequences)	Sampling Duration After Intervention Start[g] (Years)
Regional[h]									
	D	Identify 60% reduction in incidence during intervention and 10% early transmissions.	L	F	5	1,600	8	50	5
	C	Identify 30% reduction in incidence during intervention and 10% early transmissions.	L	S	5	1,600	8	50	5
	A	Identify 60% reduction in incidence during intervention and 40% early transmissions.	H	F	5	1,600	8	50	5
	B	Identify 30% reduction in incidence during intervention and 40% early transmissions.	H	S	5	1,600	8	50	5
	O	As D, and evaluate impact of sampling frame: shorter duration of intensive sequencing.	L	F	5	1,280	8	50	3
	T	As D, and evaluate impact of tree reconstruction.	L	F	5	1,600	8	50	5
	S	As D, and evaluate impact of sampling frame: most sequences from after intervention start.	L	F	5	1,600	8	85	5
	I	As D, and evaluate impact of sampling frame: higher sequence coverage.	L	F	5	3,200	16	50	5
	R	As C, and evaluate impact of tree reconstruction.	L	S	5	1,600	8	50	5
	Q	As C, and evaluate impact of sampling frame: most sequences from after intervention start.	L	S	5	1,600	8	85	5
	G	As C, and evaluate impact of sampling frame: higher sequence coverage.	L	S	5	3,200	16	50	5
	N	Control simulation, no intervention.	L	None	5	1,600	8	50	5
	F	As A, and evaluate impact of sampling frame: shorter duration of intensive sequencing.	H	F	5	1,280	8	50	3
	L	As A, and evaluate impact of tree reconstruction.	H	F	5	1,600	8	50	5
	J	As A, and evaluate impact of sampling frame: higher sequence coverage.	H	F	5	3,200	16	50	5
	P	As A, and evaluate impact of higher proportion of viral introductions.	H	F	20	1,600	8	50	5
	H	As B, and evaluate impact of tree reconstruction.	H	S	5	1,600	8	50	5
	K	As B, and evaluate impact of sampling frame: higher sequence coverage.	H	S	5	3,200	16	50	5
	E	As B, and evaluate impact of higher proportion of viral introductions.	H	S	20	1,600	8	50	5
	M	Control simulation, no intervention.	H	None	5	1,600	8	50	5
Village[h]									
	3	Identify 40% reduction in incidence during intervention and 4% early transmissions.	L	F	<2	777	25	>95	5
	2	Identify 15% reduction in incidence during intervention and 4% early transmissions.	L	S	<2	857	25	>95	5

(continued)

Table 3. Continued

Model	Dataset	Purpose	%Acute[a,b] (Low=L, High=H)	Intervention Scale Up[a,c] (Fast=F, Slow=S)	Viral Introductions[a,d] (% of All Transmissions per Year)	Sequences (#)	Sequence Coverage in the Last Year of the Simulation[a,e] (% of All Infected and Alive)	Sequences After Intervention Start[f] (% of All Sequences)	Sampling Duration After Intervention Start[g] (Years)
	1	**Identify 40% reduction in incidence during intervention and 20% early transmissions.**	**H**	**F**	**<2**	**957**	**25**	**>95**	**5**
	4	**Identify 15% reduction in incidence during intervention and 20% early transmissions.**	**H**	**S**	**<2**	**1,040**	**25**	**>95**	**5**
	5	**As 3, and evaluate impact of sampling frame: higher sequence coverage.**	**L**	**F**	**<2**	**1,469**	**50**	**>95**	**5**
	11	**Similar to 3, without imported sequences.**	**L**	**F**	**0**	**638**	**25**	**>95**	**5**
	8	**As 2, and evaluate impact of sampling frame: higher sequence coverage.**	**L**	**S**	**<2**	**1,630**	**50**	**>95**	**5**
	9	**Similar to 2, without imported sequences.**	**L**	**S**	**0**	**686**	**25**	**>95**	**5**
	0	**Control simulation, no intervention.**	**L**	**None**	**<2**	**872**	**25**	**>95**	**5**
	6	**As 1, and evaluate impact of sampling frame: higher sequence coverage.**	**H**	**F**	**<2**	**1,831**	**50**	**>95**	**5**
	12	**Similar to 1, without imported sequences.**	**H**	**F**	**0**	**956**	**25**	**>95**	**5**
	7	**As 4, and evaluate impact of sampling frame: higher sequence coverage.**	**H**	**S**	**<2**	**1,996**	**50**	**>95**	**5**
	10	**Similar to 4, without imported sequences.**	**H**	**S**	**0**	**1,012**	**25**	**>95**	**5**

[a]Variables in shaded columns were unknown to participants at time of analysis.
[b]Values range from 5% to 40%, reflecting recent estimates for endemic-phase epidemics in sub-Saharan Africa (Cohen et al. 2012).
[c]Range reflects optimistic and pessimistic scenarios in prevention trials in sub-Saharan Africa (Iwuji et al. 2013; Moore et al. 2013; Hayes et al. 2014).
[d]Range includes frequent viral introductions as reported in settings with highly mobile populations (Grabowski et al. 2014).
[e]In comparison to the large sequence datasets that are available for concentrated epidemics in Europe or North America, the lower values here reflect challenges in achieving high sequence coverage where large populations are infected. Higher values reflect geographically focused sequencing efforts such as in Mochudi, Botswana (Carnegie et al. 2014).
[f]Values reflect the duration of typical prevention trial settings, and that most sequences are obtained after intervention start (Iwuji et al. 2013; Moore et al. 2013; Hayes et al. 2014).
[g]Out of all individuals that were alive and infected in the last calendar year of the simulation, the proportion that had ever a sequence taken.
[h]For datasets in bold, only viral sequences were disclosed. For all other datasets, only viral phylogenies were provided.

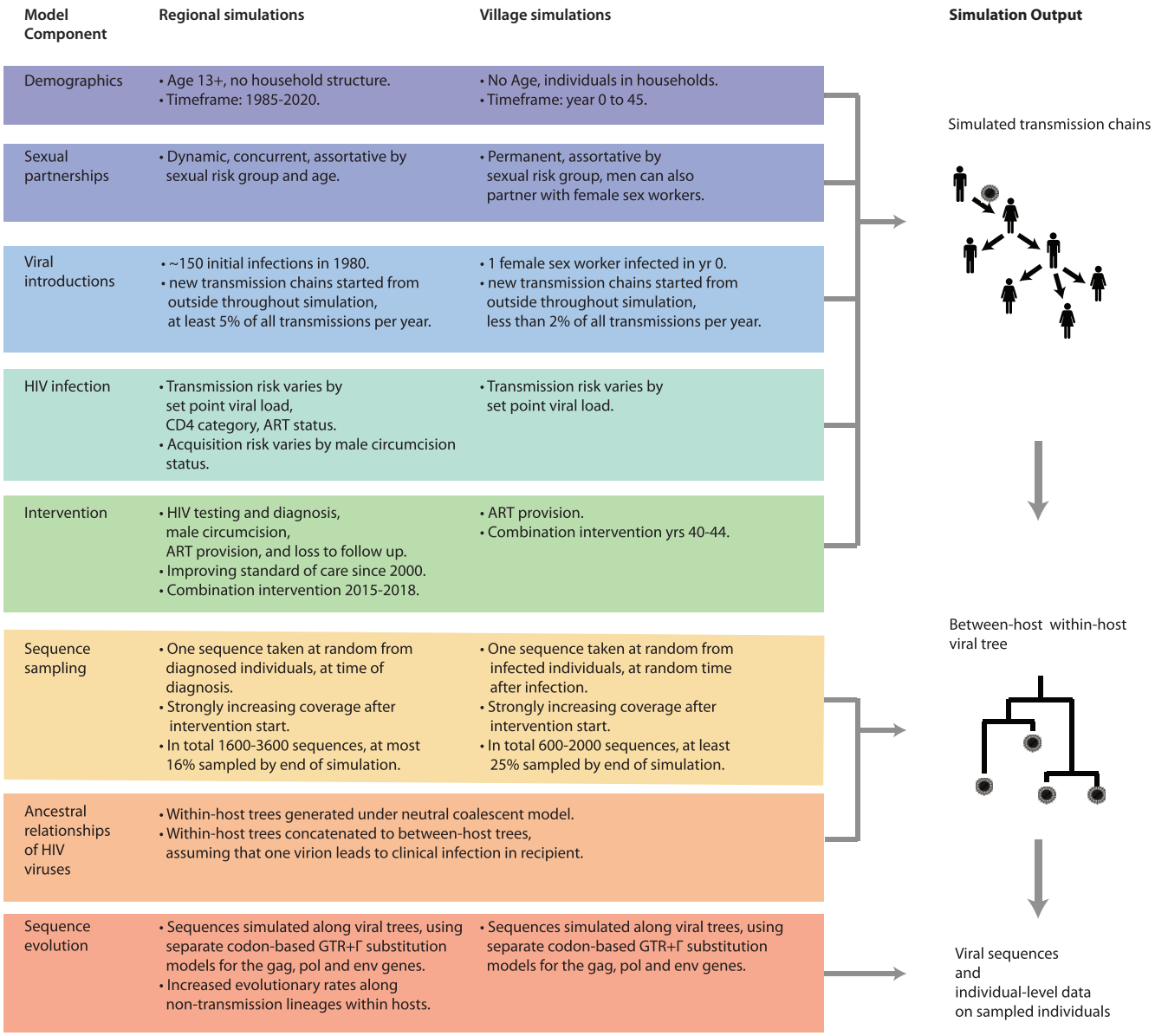

Fig. 1. Simulation pipeline to generate HIV-1 sequence data, viral phylogenies, and accompanying individual-level data. Two simulation models (Regional and Village) were implemented for the methods comparison. The two individual-level epidemic and intervention models generated HIV-1 transmission chains in the model population, and its components are shown in blue to green. Next, individuals were sampled for sequencing, and a viral tree was generated for these individuals. Tree generation accounted for within-host viral evolution under a neutral coalescent model. Finally, viral sequences comprising the *gag*, *pol* and *env* genes were simulated along the viral tree. Sequence generation accounted for known variation in evolutionary rates across genes, codon positions, and along within-host lineages. Further details are provided in supplementary tables S1 and S2, Supplementary Material online.

Treatment uptake was either "fast" or "slow", with reductions in incidence averaging between 10% and 40% relative to before intervention start. Additionally, simulations were configured so that either a small (4%) or large (20%) proportion of transmissions occurred during the first 3 months of infection. Some infections originated from outside the model population in half of the simulations.

Viral sequences were generated from the simulated transmission chains (fig. 1). First, individuals were sampled at random for sequencing. The majority of individuals were only sampled in the last years of the simulations, reflecting that sequences are only beginning to be more routinely collected in sub-Saharan Africa (Iwuji et al. 2013; Moore et al. 2013; Dennis et al. 2014; Grabowski et al. 2014; HPTN 071 (PopART) Phylogenetics Protocol Team 2015; Pillay et al. 2015). Sequence sampling biases can be substantial in real datasets, but were not included in the model (Carnegie et al. 2014; Ratmann et al. 2016). Second, viral trees were generated under a hybrid within- and between-host coalescent model. The viral trees did not always correspond to the transmission trees, because viruses diversified within infected individuals before transmission (Pybus and Rambaut 2009). In 25 of the 33 datasets, these viral trees were made available, in order to reduce the computational burden of molecular epidemiological

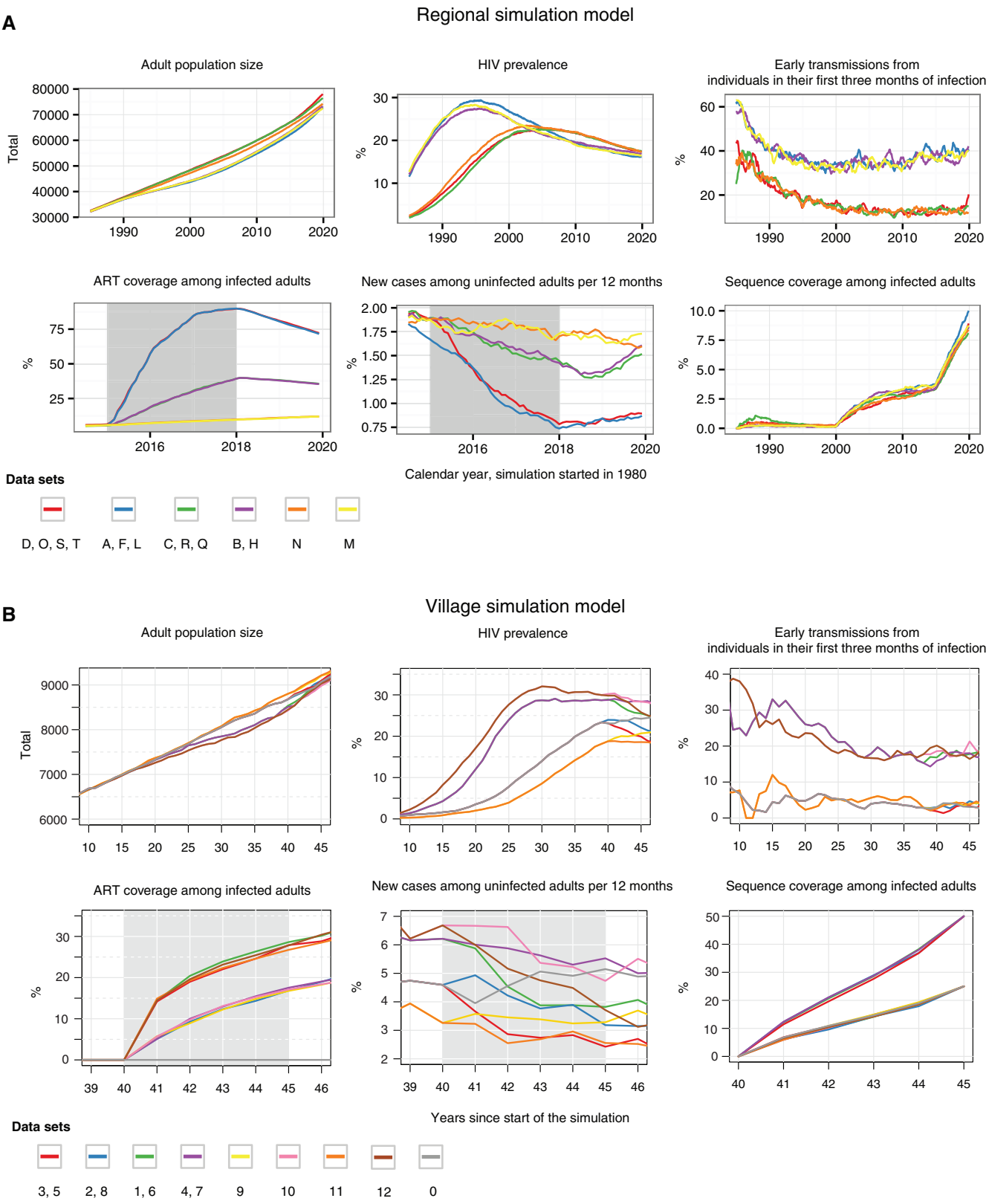

FIG. 2. Simulated epidemic scenarios under the Regional and Village models. (*A*) Six generalized HIV-1 epidemic scenarios were simulated in a region of ~80,000 adult individuals using the Regional model, and (*B*) nine scenarios were simulated in a rural village population with an initial population of ~6,000 individuals using the Village model. The scenarios differ in terms of incidence, the proportion of early transmissions, and scale-up of the combination prevention package during the intervention period (gray-shaded time period). From these, 33 datasets were generated, that included either viral sequences or viral trees. These datasets further varied in the sequence sampling frame and the frequency of viral introductions; see also figure 1 and table 3. Datasets E, G, I, J, K, P had more frequent viral introductions or higher sequence coverage, and are not shown. The proportion of early transmissions under the Village model was smoothed with a 3-year sliding window to better visualize trends in this smaller model population.

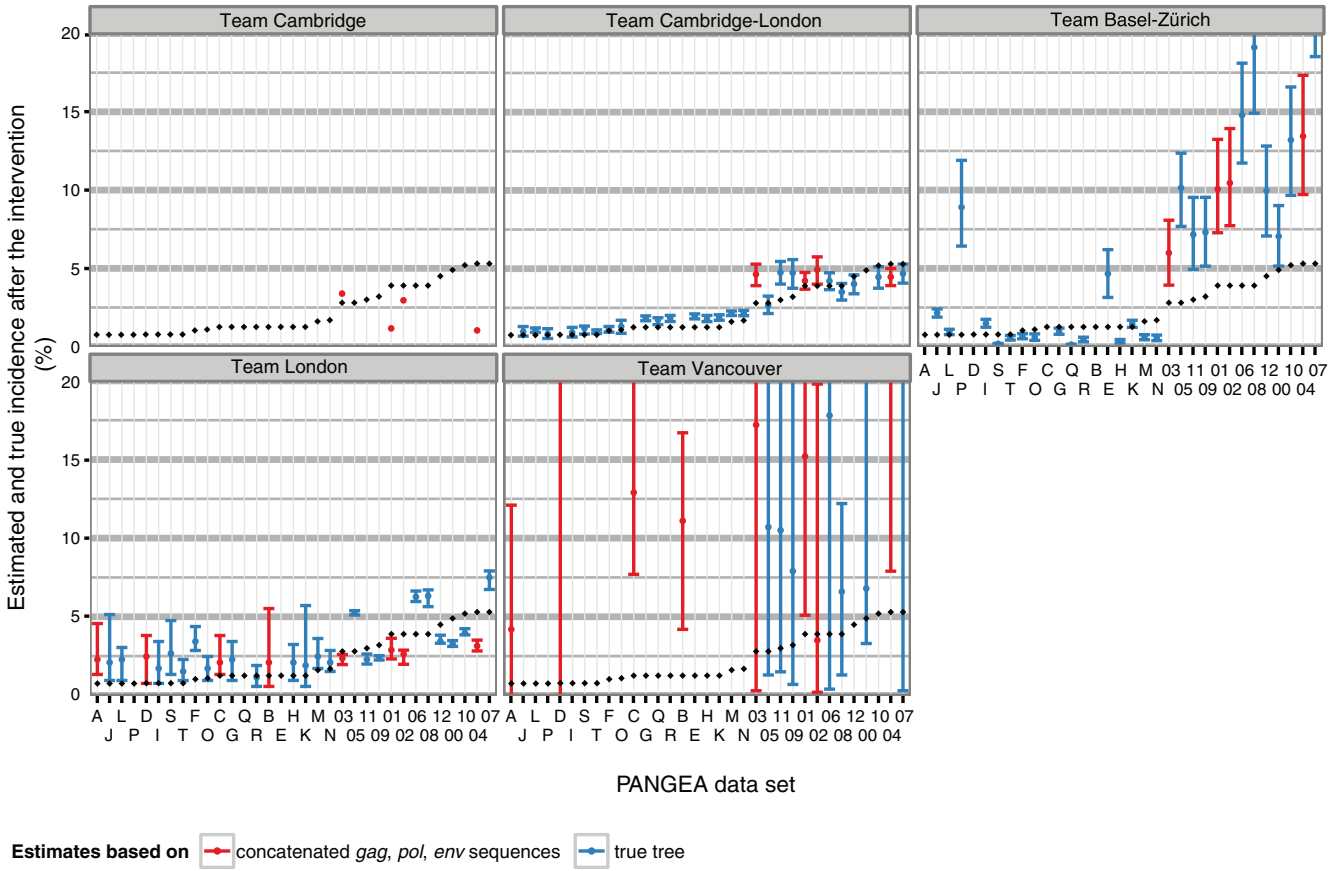

FIG. 3. Estimates of HIV-1 incidence from phylogenetic methods on simulated PANGEA datasets. Submitted estimates are shown for each PANGEA dataset by research team (panel) and type of data provided (either sequences or the viral phylogenetic tree, color). Error bars correspond to 95% credibility or confidence intervals. True values are shown in black.

analyses (table 3 and supplementary figs. S1 and S2, Supplementary Material online). For the remaining 13 datasets, viral sequences of HIV-1 *gag*, *pol* and *env* genes were simulated along the viral trees (~1,500, ~3,000 and ~2,500 nucleotides respectively, for a total of approximately 6,000 nucleotides), from an HIV-1 subtype C starting sequence. The sequences thus represent generalized subtype C epidemics, as in most Southern African countries. The nucleotide sequence evolution model that was used incorporated known differences in evolutionary rates by gene and codon position and relative differences in substitution rates by gene and codon position (Shapiro et al. 2006; Alizon and Fraser 2013). The coalescent and sequence evolution models did not account for recombination, sequencing errors, or selection beyond differential evolutionary rates across genes, codons and within-host lineages (supplementary tables S1 and S2, Supplementary Material online). As a key indicator of the realism of the simulated sequences, we calculated the proportion of the variation in evolutionary diversification among the simulated HIV-1 sequences, that can be explained by a constant molecular clock model. The proportion explained ranged from 25% to 60% (supplementary figs. S3 and S4, Supplementary Material online), broadly in line with estimates on real HIV-1 sequence datasets (Lemey et al. 2006).

The simulations were designed to retain signal for differentiating between the "fast", "slow" and "no" community-level intervention scenarios through the viral sequences provided (supplementary fig. S5, Supplementary Material online). However, we expected that rapid increases in sequence coverage after the intervention would complicate phylogenetic inference. The simulations also retained, on average, information for differentiating between the 10% and 40% early transmission scenarios of the Regional simulations at very low sequence coverage (supplementary fig. S6, Supplementary Material online). More challenges were expected on the Village simulations despite higher sequence coverage, partly because the effect size between the low and high %Acute scenarios was smaller (supplementary fig. S7, Supplementary Material online).

Responses to the Methods Comparison Exercise

Participants were primarily asked to estimate incidence reductions from before the intervention (year 39 or 2014) to just after the intervention (year 43 or 2018), and to estimate the proportion of early transmissions in the year before and after the intervention (table 1). Participating teams developed fast computational strategies for handling full-genome HIV sequence datasets within given timelines (3 months for 13 Village datasets and 6 months for 20 regional datasets). First, where only sequences were provided, viral phylogenies were reconstructed with maximum likelihood methods (Price et al. 2010; Stamatakis 2014). Second, these phylogenies were dated

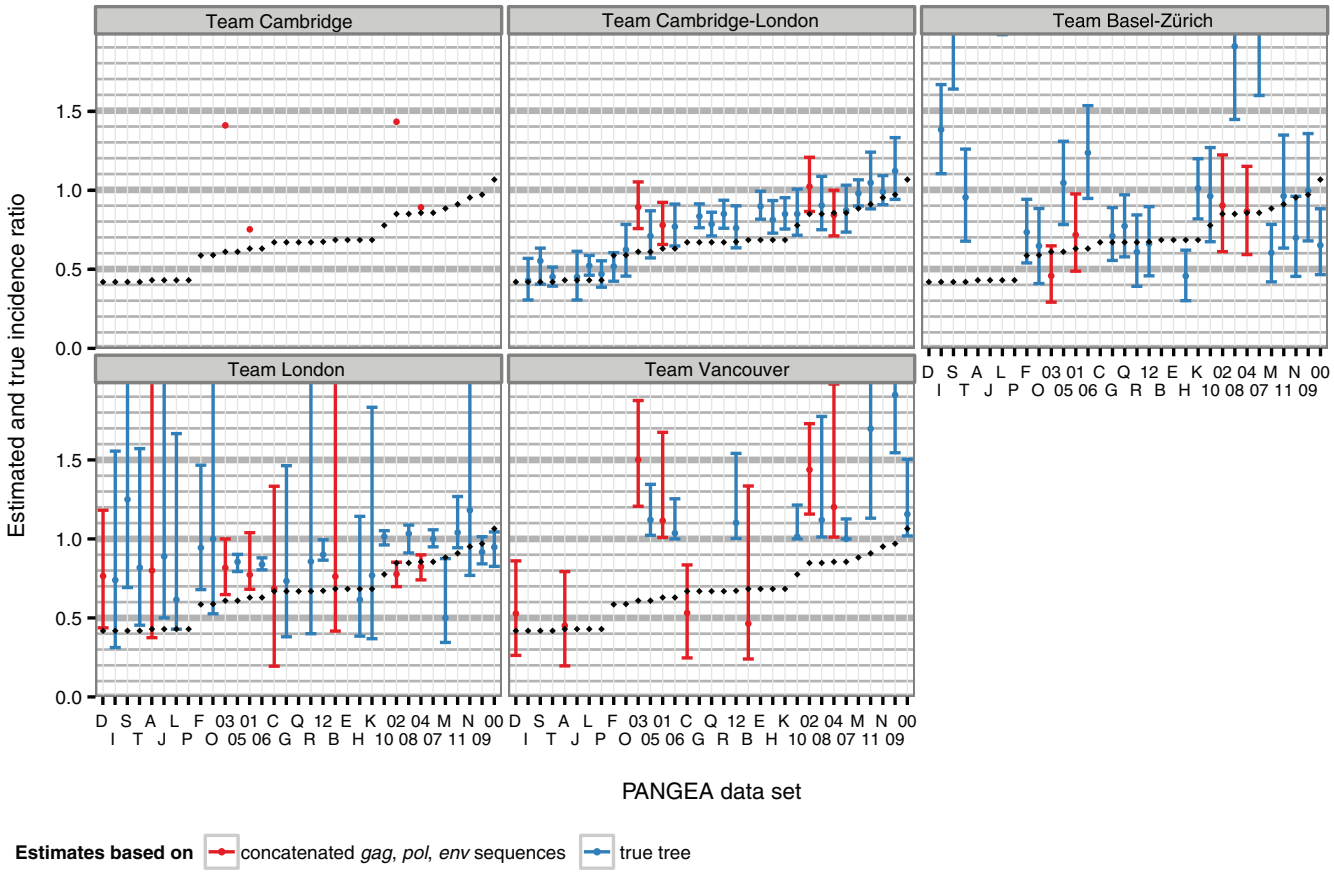

Fig. 4. Estimates of HIV-1 incidence reductions from phylogenetic methods on simulated PANGEA datasets. Submitted estimates are shown for each PANGEA dataset by research team (panel) and type of data provided (either sequences or the viral phylogenetic tree, color). Error bars correspond to 95% credibility or confidence intervals. True values are shown in black.

under least-squares criteria or similar fast approaches (To et al. 2015). Third, dated phylogenies were used as input to the transmission analysis methods described in table 2. This sequential approach allowed the teams to obtain phylogenetic estimates to all reporting variables for the large majority of the datasets (see supplementary table S3, Supplementary Material online). Team Vancouver did not provide estimates to datasets of the Regional model that contained true phylogenetic trees; and teams Cambridge-London and Basel-Zürich did not provide estimates to datasets of the Regional model that contained sequences. The most common reasons for incomplete recall were limited availability of computing resources, tight timelines to evaluate the simulations, and difficulties in tree estimation when viral introductions occurred frequently. Nearly all participants focused on inference from full viral genomes (supplementary table S3, Supplementary Material online), meaning that the impact of full genome sequences (concatenated HIV-1 *gag*, *pol* and *env* genes) as compared with partial sequences (HIV-1 *pol* gene only) could not be evaluated.

Estimating Incidence and Reductions in Incidence

Phylogenetic methods differed in their ability to estimate incidence after the intervention (fig. 3). Under the most successful computational approach, phylogenetic estimates of incidence were correlated with true values by 91% (supplementary table S2, Supplementary Material online, team Cambridge-London who used a structured coalescent model). Bias in these estimates was relatively small for estimates of two teams (on an average 0.35% by team Cambridge-London and 0.57% by team London). Team Basel-Zürich achieved substantially more accurate estimates on the Regional datasets than the Village datasets, whereas the converse was true for team London (supplementary table S2, Supplementary Material online).

The accuracy of phylogenetic estimates of changes in incidence as a result of the intervention largely reflected the accuracy of the underlying incidence estimates (fig. 4). Phylogenetic estimates of incidence ratios correlated with the true values by 93% under the structured coalescent approach of team Cambridge-London, and had only slight upward bias (supplementary table S4, Supplementary Material online). This meant that large reductions in incidence, which are expected from combination prevention interventions, could be correctly detected at relatively low sequence coverage when sequences were sampled for 5 years since intervention start by the most successful method. Epidemic simulations with >25% reductions in incidence were correctly classified as declining in 15/17 (88%) of all simulations with a submission by team Cambridge-London, although the true positive rate was lower with other phylogenetic methods (supplementary table S5, Supplementary Material online).

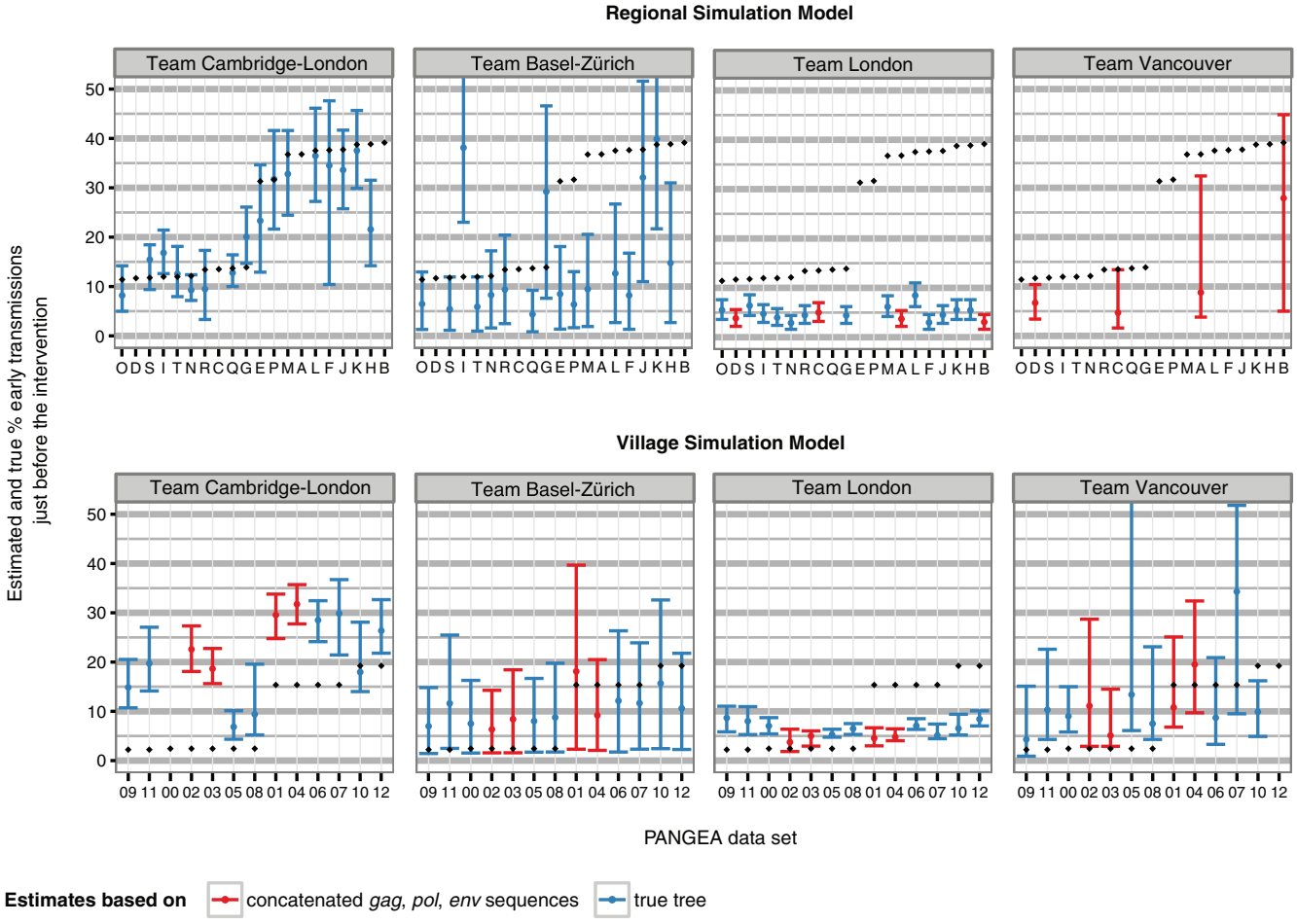

Fig. 5. Estimates of the proportion of transmissions from individuals in their first 3 months of infection (early and acute HIV), before the intervention from phylogenetic methods on simulated PANGEA datasets. Submitted estimates are shown for each PANGEA dataset by research team and model simulation (panels) and type of data provided (either sequences or the viral phylogenetic tree, color). Error bars correspond to 95% credibility or confidence intervals. True values are shown in black.

Estimating the Proportion of Transmissions from Individuals in Their First Three Months of Infection (Early and Acute HIV)

Phylogenetic estimates of the proportion of early transmissions just before and after the intervention were more accurate on the Regional simulations than the Village simulations, potentially reflecting stronger signal as a result of larger effect sizes in the Regional simulations (fig. 5 and supplementary figs. S6–S8, Supplementary Material online). On the regional simulations, estimates by team Cambridge-London had a mean absolute error of 3.9% and correlated with true values by 92%. However, on the Village simulations, the mean absolute error in estimates by team Cambridge-London was 12% (supplementary table S6, Supplementary Material online). Other teams had, overall, difficulties recovering the frequent early transmission scenarios. Team Basel-Zürich achieved the smallest mean absolute error on the Village simulations (supplementary table S6, Supplementary Material online).

Predictors of Large Error in Phylogenetic Estimates

We evaluated to what extent the variation in errors of phylogenetic estimates could be associated to systematic differences in the simulation datasets (referred to as "covariates"), such as sequence coverage and frequency of viral introductions (table 3). Figure 6A illustrates the phylogenetic estimates that deviated largely from the true values (referred to as "outliers"). We focused on quantifying the association of outlier presence with the covariates listed in table 3 using a partial least squares regression approach, which enabled us to handle a relatively large number of co-dependent covariates (see "Materials and Methods" section).

Several covariates could be excluded from this analysis. Estimates obtained from the simulated full genome sequence datasets were not more strongly associated with estimation error than estimates obtained using the phylogenetic trees from which the sequences were simulated (supplementary fig. S9 and supplementary table S7, Supplementary Material online). Shorter, intense sampling periods after intervention start of 3 years compared with a default of 5 years were also not strongly associated with larger estimation error (supplementary table S7, Supplementary Material online).

Figure 6B shows the proportion of variance in outlier presence that is explained by each of the remaining covariates. Signs indicate the impact of a change in predictor values on the number of phylogenetic estimates with

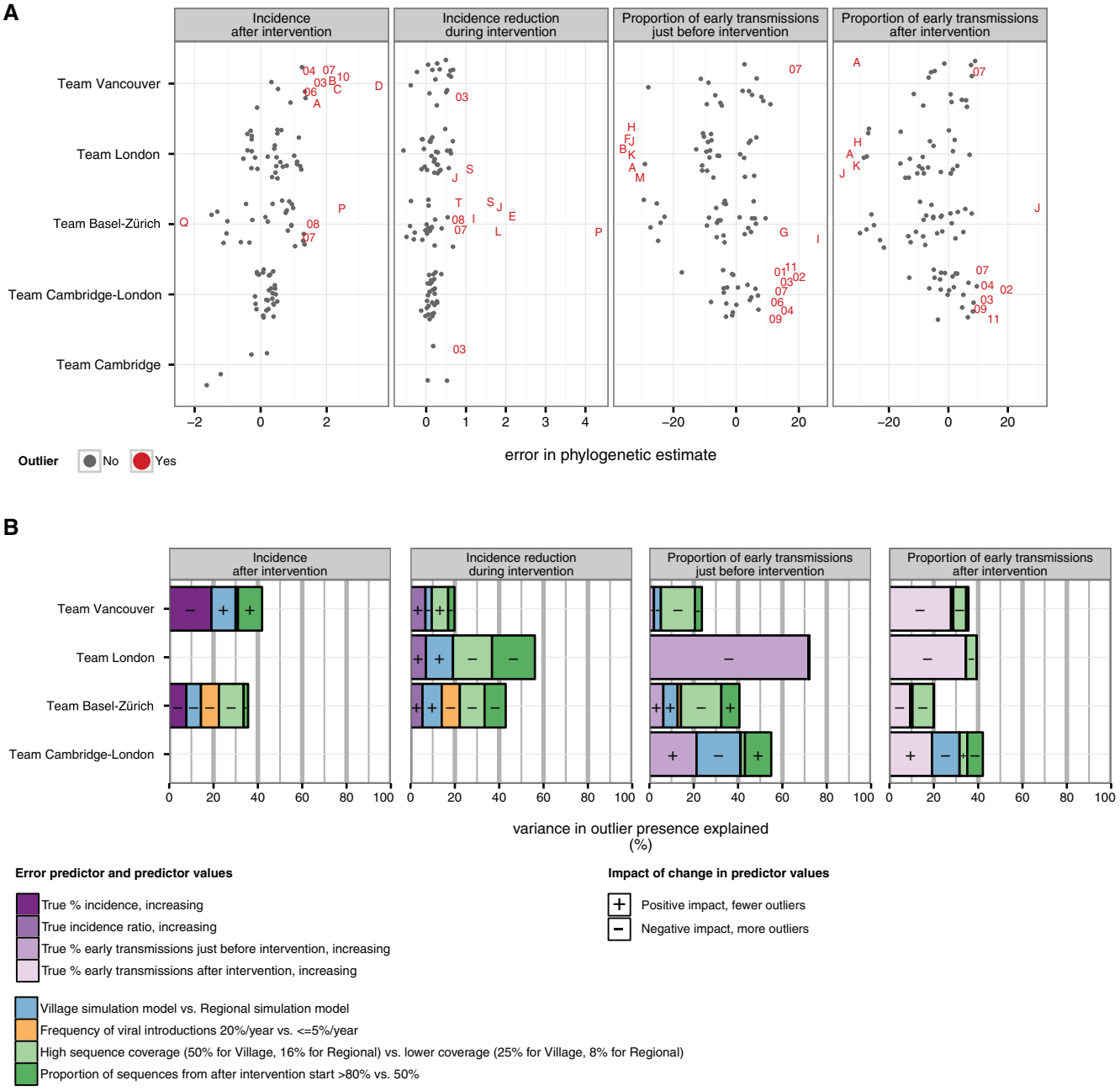

Fig. 6. Predictors of large error in phylogenetic estimates. (*A*) For each response, the error in the phylogenetic estimate was calculated, and statistical outliers were identified. The plot shows error in phylogenetic estimates by team and outcome measure. For large errors, the corresponding PANGEA dataset code in table 1 is indicated. (*B*) The contribution of the systematically varied covariates in table 1 to the presence of outliers was quantified through partial least squares regression (PLS, see "Materials and Methods" section). The plot shows the contribution of each predictor to the variance in outlier presence in colors, and the corresponding signs of the regression coefficients are added. Estimates from team Cambridge could not be characterized due to small sample size. The impact of the error predictors varied across the primary objectives of phylogenetic inference, as well as the phylogenetic methods used. With regard to estimates of incidence and incidence reduction, a subset of phylogenetic methods was particularly sensitive to high sequence coverage, a very large proportion of sequences obtained after intervention start, and a large frequency of viral introductions. With regard to estimates of the proportion of early transmissions, outliers were in several cases best explained by true differences in the proportion of early transmissions.

very large error. Subplots are empty when phylogenetic methods did not produce estimates with large error (indicating a higher degree of success). Overall, with regard to estimates of incidence and incidence reduction, higher sequence coverage (16% vs. 8% in the Regional datasets and 50% vs. 25% in the Village datasets) and a large proportion of sequences obtained after intervention start (>80% vs. 50%) were associated with more outliers for more than one phylogenetic method. Frequent viral introductions (20%/year vs. <=5%/year) were associated with more outliers by team Basel-Zürich. These predictors tended to outweigh the impact that true differences in incidence and incidence reduction had on outlier presence.

In contrast, with regard to estimates of the proportion of early transmissions, outliers were in several cases best explained by true differences in the proportion of early transmissions. Several phylogenetic methods had substantial difficulty estimating frequent early transmissions. Low sampling coverage did not contribute substantially to the presence of outliers. To substantiate this observation further, we compared phylogenetic estimates from just before the intervention to those after the intervention, and found no consistent improvements in accuracy with a doubling of sampling coverage (supplementary fig. S10, Supplementary Material online). Instead, outlier presence could be explained through the simulation model, with more outliers on the Village datasets. These simulations were characterized by smaller sample sizes and smaller effect size (table 3 and supplementary figs. S6 and S7, Supplementary Material online).

Discussion

The PANGEA methods comparison exercise represents a community-wide effort for advancing the use of phylogenetic methods to estimate aspects of recent HIV-1 transmission dynamics of generalized epidemics in sub-Saharan Africa. This region is affected by the largest HIV-1 epidemics worldwide. Viral phylogenetics could be a central tool to guide HIV-1 prevention in these settings (Dennis et al. 2014).

It is not possible for phylogenetic methods to capture all factors that influence the spread of HIV-1, ranging all the way from biological factors determining person-to-person transmission (Cohen et al. 2011) to the structure of sexual networks on the community level (Gregson et al. 2002; Tanser et al. 2011), and the broader impact of prevention and care services (Gardner et al. 2011). Of course, capturing all such features may not be needed: particular aspects of HIV-1 spread in generalized epidemics could be estimable from sequence data under the simplifying assumptions of phylogenetic methods, and at relatively low sequence coverage.

To validate this hypothesis from the outset, the PANGEA-HIV team simulated data under two highly complex HIV transmission and intervention models, whose components are considered essential for understanding long-term HIV transmission dynamics (Eaton et al. 2012). The aspects of HIV-1 spread evaluated here (table 1) were chosen both because molecular epidemiological studies into the sources of transmission and temporal changes in epidemic spread are in principle feasible (von Wyl et al. 2011; Stadler et al. 2013; Volz et al. 2013; Dennis et al. 2014; Ratmann et al. 2016), and because of their relevance to on-going HIV-1 prevention efforts in sub-Saharan Africa. Crucially, the model simulations were constrained to pessimistic and optimistic projections of the likely outcomes of on-going HIV-1 prevention efforts in sub-Saharan Africa (Iwuji et al. 2013; Moore et al. 2013; Hayes et al. 2014), as well as what sequence data could become available in these settings.

The methods comparison exercise was challenging. First, the exercise focused on quantifying recent transmission dynamics, whereas HIV-1 sequence data are more routinely used to characterize the origins and spread of the virus (Faria et al. 2014), or to undertake descriptive analyses of putative transmission chains (Brenner et al. 2007; Dennis et al. 2012). To be precise, the challenge here was in obtaining quantitative estimates of HIV-1 incidence and the sources of transmission in generalized epidemics, and to do so close to the present, when the phylogenetic signal weakens (de Silva et al. 2012). Second, sequence coverage was relatively low in most simulations, as is expected for most endemic-phase settings in sub-Saharan Africa. Furthermore, frequent viral introductions complicated the interpretation of viral trees, timelines were tight (3 months for the Village datasets, and 6 months for the Regional datasets), and phylodynamic models had to represent viral spread in heterogeneous populations (males and females with different risk profiles). We aspired to evaluate the extent to which these challenges can be addressed with full genome HIV-1 sequences, and through customized phylogenetic methods.

The methods comparison exercise demonstrates that viral phylogenetic tools can successfully estimate aspects of recent transmission dynamics of generalized HIV-1 epidemics at limited sequence coverage of the infected population, when full-genome sequences are available. Two methods, the ABC kernel method of team Vancouver and the Bayesian transmission analyzer of team London (table 2), were newly developed in response to the exercise. The birth–death skyline model with sampled ancestors (Gavryushkina et al. 2014) and its multi-type analogue (Kühnert et al. 2016) are readily available through the BEAST2 software package. The structured coalescent (Volz et al. 2009) was customized to reflect available information on the simulated epidemics, and required considerable resources (roughly 1 week of computation time on a 64-core machine of 2.5 Ghz processors per analysis). The methods comparison reflects these different stages in development and customization. In this context, the structured coalescent approach was overall most accurate, producing accurate estimates of incidence and changes in incidence, as well as broadly accurate estimates into the proportion of early transmissions on the Regional simulations from full-genome sequences. Confidence intervals were sufficiently tight for epidemiological interpretation, bearing in mind that uncertainty in tree reconstructions was ignored. This indicates that the latest generation of viral phylogenetic methods can complement standard incidence estimation techniques where full-genome sequences are available from the general population. The use of sequence data for estimating incidence trends in sub-Saharan Africa could be particularly useful where demographic and health survey data are sparse (Pillay et al. 2015), no relevant observational HIV cohorts exist, or where estimates would otherwise be solely reliant on data from particular population groups such as pregnant women (Montana et al. 2008). Further, this study supports using viral phylogenetic methods for identifying sources of HIV-1 transmission from full-genome sequences in certain settings. Broadly accurate estimates into the fraction of transmissions attributable to a population group were obtained when both transmission from that group was not infrequent (at least 10%) and sample size was not too small (thousands of sequences for the HIV-infected populations

considered). Viral phylogenetic methods could thus help to quantify the contribution of several other source populations that are of key interest for prevention in sub-Saharan Africa, including the proportion of individuals infected within localized high prevalence areas (Tanser et al. 2013), or the proportion of young women infected by male peers (Dellar et al. 2015).

We varied aspects of transmission dynamics and the sampling frame in the simulations, to obtain a more systematic understanding of methods' performance (fig. 5). Most phylogenetic methods did not identify significant differences between the high/low early transmission scenarios, and this was also the case when basic genetic distance measures recovered differences between the high/low early transmission scenarios (regional simulations, supplementary fig. S6, Supplementary Material online). The true proportions of early transmissions were also frequently outside 95% confidence or credibility intervals. This indicates that further methods' improvement is needed for estimating the proportion of early transmissions, and potentially for attributing sources of HIV-1 transmission more broadly at the low sequence coverage scenarios considered. Further, nearly all participants reported difficulties in achieving numerical convergence of their methods on full-genome sequence data (unpublished submission reports). This could explain the above observations in part, and in particular why the accuracy of early transmission estimates did not improve when using larger datasets with higher sequence coverage (fig. 5 and supplementary fig. S10, Supplementary Material online). Further investigations are needed. Finally, our error analysis suggests that explicit modeling of unobserved source demes (team Cambridge-London) or identification of spatially localized phylogenetic clusters prior to transmission analyses (team London) could be effective approaches for mitigating the negative impact of viral introductions on phylogenetic analyses on mobile populations (Grabowski et al. 2014). The simulated PANGEA datasets as well as various aspects of the corresponding true epidemics and interventions are available for future benchmarking.

This study has limitations. First, phylogenetic methods were evaluated on simulated HIV-1 epidemics. While the use of two models guards to some extent against over-interpretation, analyses of real datasets may be more complex and could be associated with overall larger error. Of note, the simulated datasets are free of sequence sampling biases, which can substantially distort phylogenetic inferences (Carnegie et al. 2014). Second, the evolutionary components of the two models generated sequences that do not contain gaps or sequencing errors, cannot be translated to amino acids, were correctly aligned, and did not contain recombinant sequences. Viral trees reconstructed from real sequence data are likely less accurate than those used in this analysis, a potential source of error that is not represented in our evaluations. Frequent recombination could imply that full HIV-1 genomes are more appropriately analyzed on a gene-by-gene basis (Hollingsworth et al. 2010; Ward et al. 2013), in contrast to our full-genome analyses of simulated sequences that excluded recombinants. This limitation is particularly relevant to epidemic settings in sub-Saharan Africa where multiple subtypes and recombinant forms circulate at high frequencies. Third, phylogenetic analyses of full-genome sequences were not compared with similar analyses using shorter fragments of the genome such as, e.g., several 250 base pair regions from the *gag*, *pol* or *env* genes. Full-genome sequences may not be required for estimating recent changes in HIV-1 incidence or for quantifying the sources of HIV-1 transmission, and more cost-effective sequencing approaches could provide similar results.

The PANGEA-HIV methods comparison exercise showed viral phylogenetic methods can be adapted to provide quantitative estimates on aspects of recent HIV-1 transmission dynamics in sub-Saharan Africa, where sequence coverage remains limited. On simulations, the structured coalescent approach was overall most accurate for estimating recent changes in incidence and the proportion of early transmissions in modeled populations with generalized, and large HIV-1 epidemics. Future molecular epidemiological analyses would ideally make use of several of the evaluated phylogenetic tools, in order to obtain robust insights into HIV-1 transmission flows and how to disrupt them. Further methods' refinement is required to this end, with our analysis suggesting a focus on estimating the sources of HIV-1 transmission from full-genome HIV-1 sequence data. These findings were obtained through a community-wide, blinded evaluation, and thereby add confidence into the use and interpretation of viral phylogenetic tools for HIV-1 surveillance and prevention in sub-Saharan Arica and beyond.

Materials and Methods

Study Design

The blinded PANGEA-HIV methods comparison exercise was announced in October 2014 at HIV Dynamics & Evolution, and later on the PANGEA-HIV website. In a training round (round 1), participants were asked to identify trends in incidence on simulated sequence datasets that were similar in size to the datasets in table 3, but that had qualitatively different epidemic dynamics. Data included full-genome viral sequences, patient meta-data, and further broad information on the simulated epidemic (supplementary text S1, Supplementary Material online). Participation was unrestricted. In December 2014, the training data were unblinded. All participants shared their findings. PANGEA-HIV and the participants agreed on the objectives and reporting variables listed in table 1; on the timelines for the second final round; and that participation will be retrospectively restricted to teams addressing at least one of the pre-specified reporting variables. Simulation models were updated to include explicit HIV care and intervention components, and re-calibrated to generate the epidemic scenarios shown in figures 1 and 2. Blinded datasets were released on 10 February 2015 (supplementary text S2, Supplementary Material online). The deadline for submissions was 8 May 2015. Questions and clarifications during the exercise were disseminated to all participants. Submissions were checked manually, and teams were given the opportunity to fix conceptual errors. Few

submissions to the Regional simulations were obtained, and the deadline for submission to Regional datasets was extended to 18 August 2015. The Village simulations were un-blinded on 14 May 2015, and a preliminary evaluation was presented and reviewed by all participants at the 22nd HIV Dynamics & Evolution conference. Teams Vancouver and Basel-Zürich informed the evaluation group of a conceptual misunderstanding of the reporting variables, and provided updated incidence estimates after the intervention 1 day after the presentation. These updates on the Village datasets were used in the evaluation reported here. The Regional datasets were un-blinded on 3 September 2015.

Village Simulations

The Village simulations were generated using the Discrete Spatial Phylo Simulator with HIV-specific components (DSPS-HIV, https://github.com/PangeaHIV/DSPS-HIV_PANGEA; last accessed October 14, 2016). The DSPS-HIV is an individual-based stochastic simulator which models HIV-1 transmissions along a specifiable contact network of individuals and produces a line-list of all events (Hodcroft 2015). Viral phylogenies that reflect between- and within-host viral evolution were generated along transmission chains using VirusTreeSimulator (https://github.com/PangeaHIV/VirusTreeSimulator; last accessed October 14, 2016). HIV-1 subtype C sequences were simulated along these viral phylogenies using πBUSS (Bielejec et al. 2014), with substitution rates parameterized from analyses of African subtype C sequences. An overview of the simulation pipeline is shown in figure 1, and details about the parameter values and assumptions used in the DSPS-HIV and to generate phylogenies and sequences are found in supplementary table S2, Supplementary Material online. Notably, assumptions were made in sexual mixing partners, partner duration, interventions, sampling, and between- and within-evolution complexity. Disease progression and transmission within the DSPS-HIV are determined by set-point viral load using previously described relationships (Fraser et al. 2007). Simulations were parameterized to reflect estimates of prevalence and incidence from the peak of the HIV-1 epidemic in the late 1980s and early 1990s (Serwadda et al. 1992; Wawer et al. 1994), before treatment was widely available, with the root of the sequences dating back $\sim$40 years previously, coinciding with the recent subtype C estimates of a common ancestor in the 1940s (Faria et al. 2014). Further information about the DSPS-HIV will be available in a forthcoming publication.

Regional Simulations

The Regional simulation model consists of a stochastic, individual-level epidemic transmission and intervention model, and an evolutionary model that generates viral phylogenies and sequence data to simulated transmission chains. Figure 1 and supplementary table S1, Supplementary Material online, describe the overall simulation pipeline, model components, parameters, and parameter values. Notably, assumptions were made on: sexual risk behavior (proportion of individuals in risk groups, mixing between risk groups, partner change rates); HIV infection (relative transmission rates); interventions (population-level effectiveness of ART); within-host evolution (neutral coalescent model, no co-infection and no recombination); between-host evolution (transmission of one virion, no recombination); and sequence sampling (at time of diagnosis of randomly selected individuals). To obtain the six epidemic scenarios shown in figure 2, we varied the relative transmission rate from early infections as well as parameters relating to uptake of the combination intervention respectively. The simulation algorithm is available from https://github.com/olli0601/PANGEA.HIV.sim (last accessed October 14, 2016), and combines (with further code): the individual-based HPTN071 (PopART) model version 1.1 to generate transmission chains, the VirusTreeSimulator (https://github.com/PangeaHIV/VirusTreeSimulator; last accessed October 14, 2016) to generate viral trees from transmission chains, and SeqGen version 1.3 (Rambaut and Grassly 1997) to simulate viral sequences along viral trees.

Protocols for Phylogenetic Transmission Analyses

All participants adopted overall similar computational strategies that first reconstructed dated maximum-likelihood trees (Price et al. 2010; Stamatakis 2014; To et al. 2015), and then considered the viral trees fixed in one of the following transmission analyses:

ABC Kernel Method

Reporting variables were estimated with an experimental kernel-ABC method that combines a kernel method on tree shapes (Poon et al. 2013) with a framework for approximate Bayesian computation (ABC). The basic premise of ABC is that it is usually easier to simulate data from a model than to calculate its exact likelihood for the observed data. A model can then be fit to the observed data by adjusting its parameters until it yields simulations that resemble these data, bypassing the calculation of likelihoods altogether. We formulated a structured compartmental SI model (Jacquez et al. 1988) that was informed by the descriptions of the agent-based simulations that were distributed to all participants. Specifically, the model comprised three populations: a main local population, a second local high-risk minority population, and an external source population. Each population was further partitioned into susceptible and infected groups, where the latter was stratified into three stages of infection (acute, asymptomatic, and chronic). Mixing rates between the main and minority local populations were controlled by two parameters to allow for asymmetric mixing. Individuals with acute or asymptomatic infections migrated from the external region to the local region at a constant rate m, and replaced with new susceptible individuals in the external region. One infected individual in the external source population started the simulation. Coalescent trees were then simulated based on population trajectories derived from the numerical solution of the ordinary differential equations that represent the model, using the R package *rcolgem*. The subset tree kernel (Poon et al. 2013) was used as a distance measure between the simulated coalescent trees and the

reconstructed viral phylogenies on available sequence data, or the provided phylogenies. A Markov chain Monte Carlo implementation of ABC was used to fit the model. This kernel-ABC approach was validated on simulated data from more conventional compartmental models (Poon 2015).

Birth–Death Skyline Method with Sampled Ancestors

Phylodynamic analyses were performed in BEAST v2.0 (Bouckaert et al. 2014) using the add-ons "bdsky" (Stadler et al. 2013), "SA" (Gavryushkina et al. 2014) and "bdmm" (Kühnert et al. 2016). Under the birth–death skyline model with sampled ancestors ("SA" module), individuals could transmit with some probability after sampling which improved estimation of the reporting variables in preliminary analyses (round 1 of the exercise). To estimate the proportion of early transmissions, the multi-type birth–death model was used with two compartments ("bdmm" module) to consider individuals in their first 3 months of infection separately from those in later stages of infection. In all analyses, time was partitioned into different intervals to obtain estimates of varying transmission rates through time. As further described in supplementary text S3, Supplementary Material online, for both Village and Regional simulations, lognormal priors were used for the effective reproductive number (mu = 0 and sigma = 0.75) and the becoming-non-infectious rate (lognormal with mu = −1 and sigma 0.5). Uniform priors were used for the sampling proportion, and specified based on available meta-data. For the Village datasets 0, 1, 2, 3, 4, 9, 10, 11 and 12, we assumed a priori a sampling proportion between 15% and 40%; for Village datasets 5, 6, 7 and 8 between 40% and 100%; and for the Regional datasets between 5% and 10%. The prior distribution for the removal probability r was chosen based on an estimate of the proportion of sampled infected individuals that are on treatment, and calculated from available survey data before intervention start. Sensitivity analyses on these prior choices were conducted. The reporting variables were estimated from MCMC output of the posterior model parameters using a customized procedure that is fully described in supplementary text S3, Supplementary Material online.

Bayesian Transmission Chain Analyser

The Bayesian approach reported in (Didelot et al. 2014) was adapted to account for incomplete sampling as well as heterogeneity in HIV transmission rates. In place of a susceptible-infectious-recovered (SIR) model (as in Didelot et al. 2014) a generalized branching model was used to describe transmission dynamics. In this model, the (prior) time interval between a case becoming infected and infecting others (t_{gen}) is distributed such that there is a peak after infection, a chronic phase, and increased infectivity with progression to AIDS. Cases were sampled after a random time since becoming infected (t_{samp}). The prior distribution of the numbers of secondary cases was negative binomial ($n = 5$, $P = 0.7$), reflecting a convolution of a Poisson distribution conditioned on a gamma-distributed overall infectivity. To account for infected individuals in transmission chains for whom a sequence was not available, likelihood terms were adjusted by numerically calculating the probability that a case infected at a given time had no sampled descendant cases by the time the study finished, and then conditioning on each case's number of sampled and unsampled descendants. A reversible-jump Bayesian MCMC approach with proposal moves as described in (Didelot et al. 2014) was used to fit the model. This approach produces a posterior collection of transmission trees. From these, we extracted the portion of infections in the acute stage, recent changes in incidence and other outcomes required for the comparison study. The generation time t_{gen} had prior $t_{gen} \sim 0.4\ \text{gamma}(1.3,1) + 0.6\ \text{gamma}(3.5,3.5)$ where the arguments are the shape and scale parameters. The time to sampling had prior $t_{samp} \sim \text{gamma}(0.7, 1.5)$.

Structured Coalescent

Structured coalescent models were implemented in the rcolgem R package and were based on compartmental infectious disease models using the approach described in (Volz 2012). These models were tailored to the Regional and Village scenarios, and included compartments for stage of infection (early HIV infection through AIDS as in Cori et al. 2014), sex, and diagnosis/treatment status. Transmission rates were allowed to vary between compartments, and generalized logistic functions described secular trends in the force of infection through time. Coalescent models also included a deme for the unsampled source deme to capture the effects of lineage importation into the surveyed region. Models were fitted to the dated viral phylogenetic trees and to available epidemiological data under the approximation that the corresponding likelihood terms are independent. For the Regional simulations, the contribution to the likelihood model of the CD4 counts at diagnosis and gender of all sequenced individuals was assumed multinomial; the proportion of diagnoses with a sequence was assumed binomial; and that of survey data (sex, diagnosis, and treatment status) was assumed multinomial. For the Village simulations, fewer meta-data variables were available. The likelihood model assumed that estimated HIV prevalence was within the bounds given by the available survey data. A parallel Bayesian MCMC technique (Calderhead 2014) was used to obtain posterior distributions of model parameters.

Statistical Analysis

Phylogenetic estimates and true values were transformed so that their differences were approximately normally distributed. For incidence and incidence reductions, the error e_i of response i was calculated as $e_i = \log(\hat{x}_i) - \log(x_i)$, where $\hat{x}_i$ is the phylogenetic estimate and x_i the true value on dataset i; for proportions, the error was calculated as $e_i = \hat{x}_i - x_i$. Data points outside the whiskers of Tukey boxplots were considered as outliers.

To identify covariates associated with large error in phylogenetic estimates, stepwise model selection with the *stepGAIV.VR* procedure in the *gamlss* R package was used to reduce the number of covariates at significance level 0.01

(supplementary table S4, Supplementary Material online). The contribution of the remaining covariates to outlier presence (response) was evaluated with partial least squares (PLS) regression (Boulesteix and Strimmer 2007), because of the limited number of datasets and dependencies amongst the covariates. PLS regression is a dimension reduction technique that identifies combinations of covariates (PLS latent factors) that are maximally correlated with the response variable, and then regresses the response variable against the latent factors. The first four latent factors that explained most of the variance in outlier presence were considered in the error analysis. Figure 5B shows, in the notation of (Boulesteix and Strimmer 2007), the sign of the PLS regression coefficients B_{j1} for each covariate j to the univariate response variable across the first $c = 4$ latent factors. The proportion of variance p_j in the response variable attributable to each covariate j is calculated as $p_j = \sum_{k=1}^{c} \left(\frac{w_{jk}}{w}\right)^2 \nu_k$, where w_{jk} is the weight of covariate j to the kth latent factor and ν_k is the variance explained by the kth latent factor. PLS regression was performed with the *plsr* routine in the *pls* R package.

Supplementary Material

Supplementary figures S1–S10, tables S1–S7, and text S1–S4 are available at *Molecular Biology and Evolution* online.

Author Contributions

A.L.B. and C.F. conceived the study. O.R., E.H., A.L.B., and C.F. designed and coordinated the study. E.H., M.H., S.L., and A.L.B. designed and generated the Village simulations. M.H. contributed the virus tree simulator from transmission chains. M.P., A.C., O.R., and C.F. designed and generated the Regional simulations. O.R. checked the submissions received, performed the statistical analysis and wrote the first draft except parts of the "Methods" section. C.C., M.K., X.D., G.P., A.P., J.J., R.L., C.W., G.L., D.R., D.K., T.S., E.V., B.D., M.H., and S.F. evaluated the simulated data and wrote parts of the "Methods" section. All authors reviewed and approved the statistical analysis, and the final version of the article.

Acknowledgments

We thank Andrew Rambaut for his comments on the design of the exercise; the PANGEA-HIV steering committee and participants of the PANGEA-HIV satellite workshop of the 21st and 22nd HIV Dynamics & Evolution conference for their comments during the exercise; and three anonymous reviewers and associate editors for their comments that improved an earlier version of the article. Regional simulations were designed and generated using resources at the Imperial College High Performance Computing Service (http://www3.imperial.ac.uk/ict/services/hpc). Team Cambridge-London thanks the MRC Centre for Outbreak Analysis and Modeling for support. Team Vancouver thanks Rosemary McCloskey for help with tree reconstructions; their contribution was enabled in part by support provided by Westgrid (www.westgrid.ca) and Compute Canada Calcul Canada (www.computecanada.ca). This work was supported by the Bill & Melinda Gates Foundation through the PANGEA-HIV consortium (to O.R., E.H., A.L.B., and C.F.); the NIH through the NIAID cooperative agreement UM1AI068619 for work on the HPTN 071 trial (to A.C., M.P., and C.F.); the Wellcome Trust (WR092311MF to O.R.); the European Research Council (PBDR-339251 to C.F., PhyPD-335529 to T.S.); the National Institutes of Health (NIH MIDAS U01 GM110749 to E.V. and A.L.B., NIH R01 AI087520 to E.V.); the Biotechnology and Biological Sciences Research Council (BB/J004227/1 to S.J.L.); the Canadian Institutes of Health Research (CIHR HOP-111406 to A.F.Y.P., New Investigator Award 175594 to A.F.Y.P.), the Michael Smith Foundation for Health Research/St. Paul's Hospital Foundation/the Providence Health Care Research Institute (Scholar Award 5127 to A.F.Y.P.); the ETH Zürich Postdoctoral Fellowship Program (to D.R. and D.K.); the Marie Curie Actions for People COFUND Program (to D.R. and D.K.); the University of Edinburgh Chancellor's Fellowship scheme (to S.J.L.); the Centre of Expertise in Animal Disease Outbreaks (to S.J.L.); and the Swiss National Science Foundation (162251 to G.E.L.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the article.

References

Alizon S, Fraser C. 2013. Within-host and between-host evolutionary rates across the HIV-1 genome. *Retrovirology* 10. Available from: https://retrovirology.biomedcentral.com/articles/10.1186/1742-4690-10-49.

Bezemer D, Cori A, Ratmann O, van Sighem A, Hermanides HS, Dutilh BE, Gras L, Rodrigues Faria N, van den Hengel R, Duits AJ, et al. 2015. Dispersion of the HIV-1 epidemic in men who have sex with men in the Netherlands: a combined mathematical model and phylogenetic analysis. *PLoS Med.* 12:e1001898.

Bielejec F, Lemey P, Carvalho LM, Baele G, Rambaut A, Suchard MA. 2014. piBUSS: a parallel BEAST/BEAGLE utility for sequence simulation under complex evolutionary scenarios. *BMC Bioinformatics* 15:133.

Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol.* 10:e1003537.

Boulesteix AL, Strimmer K. 2007. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief Bioinform.* 8:32–44.

Brenner BG, Roger M, Routy JP, Moisi D, Ntemgwa M, Matte C, Baril JG, Thomas R, Rouleau D, Bruneau J, et al. 2007. High rates of forward transmission events after acute/early HIV-1 infection. *J Infect Dis.* 195:951–959.

Calderhead B. 2014. A general construction for parallelizing Metropolis-Hastings algorithms. *Proc Natl Acad Sci U S A.* 111:17408–17413.

Carnegie NB, Wang R, Novitsky V, De Gruttola V. 2014. Linkage of viral sequences among HIV-infected village residents in Botswana: estimation of linkage rates in the presence of missing data. *PLoS Comput Biol.* 10:e1003430.

Cohen MS, Dye C, Fraser C, Miller WC, Powers KA, Williams BG. 2012. HIV treatment as prevention: debate and commentary–will early infection compromise treatment-as-prevention strategies?. *PLoS Med.* 9:e1001232.

Cohen MS, Shaw GM, McMichael AJ, Haynes BF. 2011. Acute HIV-1 Infection. *N Engl J Med.* 364:1943–1954.

Cori A, Ayles H, Beyers N, Schaap A, Floyd S, Sabapathy K, Eaton JW, Hauck K, Smith P, Griffith S, et al. 2014. HPTN 071 (PopART): a cluster-randomized trial of the population impact of an HIV combination prevention intervention including universal testing and treatment: mathematical model. *PLoS One* 9:e84511.

de Silva E, Ferguson NM, Fraser C. 2012. Inferring pandemic growth rates from sequence data. *J R Soc Interface* 9:1797–1808.

Dearlove B, Wilson DJ. 2013. Coalescent inference for infectious disease: meta-analysis of hepatitis C. *Philos Trans R Soc Lond B Biol Sci.* 368:20120314.

Dellar RC, Dlamini S, Karim QA. 2015. Adolescent girls and young women: key populations for HIV epidemic control. *J Int AIDS Soc.* 18:19408.

Dennis AM, Herbeck JT, Brown AL, Kellam P, de Oliveira T, Pillay D, Fraser C, Cohen MS. 2014. Phylogenetic studies of transmission dynamics in generalized HIV epidemics: an essential tool where the burden is greatest?. *J Acquir Immune Defic Syndr.* 67:181–195.

Dennis AM, Hue S, Hurt CB, Napravnik S, Sebastian J, Pillay D, Eron JJ. 2012. Phylogenetic insights into regional HIV transmission. *Aids* 26:1813–1822.

Didelot X, Gardy J, Colijn C. 2014. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol Biol Evol.* 31:1869–1879.

Eaton JW, Johnson LF, Salomon JA, Barnighausen T, Bendavid E, Bershteyn A, Bloom DE, Cambiano V, Fraser C, Hontelez JA, et al. 2012. HIV treatment as prevention: systematic comparison of mathematical models of the potential impact of antiretroviral therapy on HIV incidence in South Africa. *PLoS Med.* 9:e1001245.

Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, Tatem AJ, Sousa JD, Arinaminpathy N, Pepin J, et al. 2014. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science* 346:56–61.

Fisher M, Pao D, Brown AE, Sudarshi D, Gill ON, Cane P, Buckton AJ, Parry JV, Johnson AM, Sabin C, et al. 2010. Determinants of HIV-1 transmission in men who have sex with men: a combined clinical, epidemiological and phylogenetic approach. *Aids* 24:1739–1747.

Fraser C, Hollingsworth TD, Chapman R, de Wolf F, Hanage WP. 2007. Variation in HIV-1 set-point viral load: epidemiological analysis and an evolutionary hypothesis. *Proc Natl Acad Sci U S A.* 104:17441–17446.

Gardner EM, McLees MP, Steiner JF, Del Rio C, Burman WJ. 2011. The spectrum of engagement in HIV care and its relevance to test-and-treat strategies for prevention of HIV infection. *Clin Infect Dis.* 52:793–800.

Gavryushkina A, Welch D, Stadler T, Drummond AJ. 2014. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Comput Biol.* 10:e1003919.

Grabowski MK, Lessler J, Redd AD, Kagaayi J, Laeyendecker O, Ndyanabo A, Nelson MI, Cummings DA, Bwanika JB, Mueller AC, et al. 2014. The role of viral introductions in sustaining community-based HIV epidemics in rural Uganda: evidence from spatial clustering, phylogenetics, and egocentric transmission models. *PLoS Med.* 11:e1001610.

Gregson S, Nyamukapa CA, Garnett GP, Mason PR, Zhuwau T, Carael M, Chandiwana SK, Anderson RM. 2002. Sexual mixing patterns and sex-differentials in teenage exposure to HIV infection in rural Zimbabwe. *Lancet* 359:1896–1903.

Hayes R, Ayles H, Beyers N, Sabapathy K, Floyd S, Shanaube K, Bock P, Griffith S, Moore A, Watson-Jones D, et al. 2014. HPTN 071 (PopART): rationale and design of a cluster-randomised trial of the population impact of an HIV combination prevention intervention including universal testing and treatment – a study protocol for a cluster randomised trial. *Trials* 15:57.

Hodcroft E. 2015. Estimating the heritability of virulence in HIV. PhD thesis, University of Edinburgh. Available from: https://www.era.lib.ed.ac.uk/handle/1842/15814.

Hollingsworth TD, Laeyendecker O, Shirreff G, Donnelly CA, Serwadda D, Wawer MJ, Kiwanuka N, Nalugoda F, Collinson-Streng A, Ssempijja V, et al. 2010. HIV-1 transmitting couples have similar viral load set-points in Rakai, Uganda. *PLoS Pathog.* 6:e1000876.

HPTN 071-2 Phylogenetics in HPTN 071: An ancillary study to "Population Effects of Antiretroviral Therapy to Reduce HIV Transmission (PopART): A cluster-randomized trial of the impact of a combination prevention package on population-level HIV incidence in Zambia and South Africa" [Internet]. 2015. HIV Prevention Trials Network. Available from: https://www.hptn.org/sites/default/files/2016-05/HPTN%20071-2_Phylogenetics%20Ancillary%20Protocol_v%201.0_15Jan2015.pdf.

Iwuji CC, Orne-Gliemann J, Tanser F, Boyer S, Lessells RJ, Lert F, Imrie J, Barnighausen T, Rekacewicz C, Bazin B, et al. 2013. Evaluation of the impact of immediate versus WHO recommendations-guided antiretroviral therapy initiation on HIV incidence: the ANRS 12249 TasP (Treatment as Prevention) trial in Hlabisa sub-district, KwaZulu-Natal, South Africa: study protocol for a cluster randomised controlled trial. *Trials* 14:230.

Jacquez JA, Simon CP, Koopman J, Sattenspiel L, Perry T. 1988. Modeling and analyzing HIV transmission – the effect of contact patterns. *Math Biosci.* 92:119–199.

Kouyos RD, von Wyl V, Yerly S, Boni J, Taffe P, Shah C, Burgisser P, Klimkait T, Weber R, Hirschel B, et al. 2010. Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. *J Infect Dis.* 201:1488–1497.

Kühnert D, Stadler T, Vaughan TG, Drummond A. 2016. Phylodynamics with migration: a computational framework to quantify population structure from genomic data. *Mol Biol Evol.* 33:2102–2116.

Lemey P, Rambaut A, Pybus OG. 2006. HIV evolutionary dynamics within and among hosts. *AIDS Rev.* 8:125–140.

Montana LS, Mishra V, Hong R. 2008. Comparison of HIV prevalence estimates from antenatal care surveillance and population-based surveys in sub-Saharan Africa. *Sex Transm Infect.* 84(Suppl 1):i78–i84.

Moore JS, Essex M, Lebelonyane R, El Halabi S, Makhema J, Lockman S, Tchetgen E, Holme MP, Mills L, Bachanas P, Marukutira T, et al. 2013. Botswana Combination Prevention Project (BCPP). *ClincialTrials.gov.* Available from: https://clinicaltrials.gov/ct2/show/NCT01965470.

Novitsky V, Kühnert D, Moyo S, Widenfelt E, Okui L, Essex M. 2015. Phylodynamic analysis of HIV sub-epidemics in Mochudi, Botswana. *Epidemics* 13:44–55.

Pillay D, Herbeck J, Cohen MS, de Oliveira T, Fraser C, Ratmann O, Brown AL, Kellam P, Consortium P-H. 2015. PANGEA-HIV: phylogenetics for generalised epidemics in Africa. *Lancet Infect Dis.* 15:259–261.

Poon AF. 2015. Phylodynamic inference with kernel ABC and its application to HIV epidemiology. *Mol Biol Evol.* 32:2483–2495.

Poon AF, Walker LW, Murray H, McCloskey RM, Harrigan PR, Liang RH. 2013. Mapping the shapes of phylogenetic trees from human and zoonotic RNA viruses. *PLoS One* 8:e78122.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490.

Pybus OG, Rambaut A. 2009. Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet.* 10:540–550.

Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci.* 13:235–238.

Ratmann O, van Sighem A, Bezemer D, Gavryushkina A, Juurrians S, Wensing AM, de Wolf F, Reiss P, Fraser C. 2016. Sources of HIV infection among men having sex with men and implications for prevention. *Sci Transl Med.* 8:320ra322.

Serwadda D, Wawer MJ, Musgrave SD, Sewankambo NK, Kaplan JE, Gray RH. 1992. HIV risk factors in three geographic strata of rural Rakai District, Uganda. *Aids* 6:983–989.

Shapiro B, Rambaut A, Drummond AJ. 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol Biol Evol.* 23:7–9.

Stadler T, Bonhoeffer S. 2013. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philos Trans R Soc Lond B Biol Sci.* 368:20120198.

Stadler T, Kuhnert D, Bonhoeffer S, Drummond AJ. 2013. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc Natl Acad Sci U S A.* 110:228–233.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.
Tanser F, Barnighausen T, Grapsa E, Zaidi J, Newell ML. 2013. High coverage of ART associated with decline in risk of HIV acquisition in rural KwaZulu-Natal, South Africa. *Science* 339:966–971.
Tanser F, Barnighausen T, Hund L, Garnett GP, McGrath N, Newell ML. 2011. Effect of concurrent sexual partnerships on rate of new HIV infections in a high-prevalence, rural South African population: a cohort study. *Lancet* 378:247–255.
To TH, Jung M, Lycett S, Gascuel O. 2015. Fast dating using least-squares criteria and algorithms. *Syst Biol.* 65:82–97.
UNAIDS. 2014. Fast-Track – Ending the AIDS epidemic by 2030. Geneva: UNAIDS. Available from: http://www.unaids.org/en/resources/documents/2014/JC2686_WAD2014report
UNAIDS. 2015. AIDS by the numbers 2015. Geneva: UNAIDS. Available from: http://www.unaids.org/sites/default/files/media_asset/AIDS_by_the_numbers_2015_en.pdf.
Vassall A, Pickles M, Chandrashekar S, Boily MC, Shetty G, Guinness L, Lowndes CM, Bradley J, Moses S, Alary M, et al. 2014. Cost-effectiveness of HIV prevention for high-risk groups at scale: an economic evaluation of the Avahan programme in south India. *Lancet Glob Health* 2:e531–e540.
Volz E, Ionides E, Romero-Severson E, Brandt MG, Mokotoff E, Koopman J. 2013. HIV-1 transmission during early infection in men who have sex with men: a phylodynamic analysis. *PLoS Med.* 10:e1001568.
Volz EM. 2012. Complex population dynamics and the coalescent under neutrality. *Genetics* 190:187–201.
Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SD. 2009. Phylodynamics of infectious disease epidemics. *Genetics* 183:1421–1430.
von Wyl V, Kouyos RD, Yerly S, Boni J, Shah C, Burgisser P, Klimkait T, Weber R, Hirschel B, Cavassini M, et al. 2011. The role of migration and domestic transmission in the spread of HIV-1 non-B subtypes in Switzerland. *J Infect Dis.* 204:1095–1103.
Ward MJ, Lycett SJ, Kalish ML, Rambaut A, Leigh Brown A. 2013. Estimating the rate of intersubtype recombination in early HIV-1 group M strains. *J Virol.* 87:1967–1973.
Wawer MJ, Sewankambo NK, Berkley S, Serwadda D, Musgrave SD, Gray RH, Musagara M, Stallings RY, Konde-Lule JK. 1994. Incidence of HIV-1 infection in a rural region of Uganda. *BMJ* 308:171–173.
WHO. 2015. Guideline on when to start antiretroviral therapy and on pre-exposure prophylaxis for HIV. Geneva: WHO Press. Available from: http://www.who.int/hiv/pub/guidelines/earlyrelease-arv/en/.

Phylogenetic Tools For Generalized HIV-1 Epidemics: Findings from the PANGEA-HIV Methods Comparison

Supplementary Text S1: Supplementary tables and figures

5

Table S1 Model components and assumptions of the Regional Model.

Model parameters	Choice of parameter values	Type of evidence
Demographics Individuals enter the modelled population continuously at age 13, leaving the population when they die, either from disease related mortality or other reasons. The overall population is slowly growing in size at a rate similar to South Africa.		
Population size in 1960	15,000 individuals aged 13 or older	Assumed
Number of new individuals aged 13 per year	The fertility rate was age-dependent, and calibrated to UNPD WPP 2006 South Africa estimates. A truncated Gamma distribution with shape parameter 15, scale parameter 1.6, minimum 13 years and maximum 50 years was used. At any time step in the model simulation, the number of new individuals of age 13 was obtained as the expected number of new arrivals under the truncated Gamma distribution. New individuals had 50% probability of being male.	Empirical
Population mortality rate	The mortality rate was age- and time-dependent, and calibrated to World Bank Global Health Observatory data (http://apps.who.int/gho/data/view.main.1360?lang=en). The final model had the form $\exp(0.001*(a\text{-}13)^{1.7} + 0.002*(a\text{-}13)^{1.7} * 60/(t\text{-}1900) - 4.9)$.	Empirical
Sexual partnerships Once individuals entered the model, they may form and break up sexual partnerships. Individuals form partnerships assortatively by risk group and age. All partnerships are heterosexual.		
Sexual risk behavior	3 risk groups categories (high/medium/low) as in (1).	Assumed
Proportion in high/ medium/ low risk group when entering population	Men: 50% / 40% / 10%; Women: 60% / 30% / 10%.	Assumed
Maximum number of simultaneous partners	High: 10; Medium: 3; Low: 1.	Assumed
Rate of partner acquisition	1.2 partners per year if maximum number of simultaneous partners not reached.	Assumed
Mixing between risk groups	10% of one's partnerships made within a risk group (assortatively), and the remainder made at random in any group.	Assumed
Mixing between age groups	Strongly assortative, determined from Manicaland cohort survey data.	Empirical
Partnership duration	Gamma distribution with shape parameter 10 and scale parameter 2.5.	Assumed
Viral Introductions The epidemic was seeded in 1980 and further viral introductions occurred throughout the simulation.		
Seed cases	The simulation was initially run for 20 years without HIV-1, to allow partnerships to reach a steady state. In 1980, 0.5% of low-risk and 1% of medium and high-risk individuals are seeded HIV-1 positive, with HIV-1 transmission occurring from that point onwards.	Assumed
Proportion of viral introductions among annual	5% or 20%, range includes frequent viral introductions as reported in settings with highly mobile populations (2, 3)	Varied in simulations

new cases		
HIV-1 infection HIV-1 negative individuals are exposed to risk of infection when they are in a serodiscordant partnership. Infection can occur at any time during a serodiscordant partnership, with the risk of infection depending upon the HIV-1 stage of the infected partner, and whether they are on ART or not at that time. For male HIV-1 negative individuals risk of infection also depends on their circumcision status.		
Proportion in SPVL group < 4, 4-4.5, 4.5-5, ≥ 5 log_{10} copies/μL after seroconversion	25% in each group, similar to (4).	Empirical
Duration of early transmission phase (months)	Sampled uniformly between 1 to 5 months, from (5).	Empirical
Duration of CD4 stages >500, 350-500, 200-350 and ≤200 cells/mm^3 when not on ART	Sampled uniformly from ranges in (4) dependent on CD4 stage and SPVL group.	Empirical
Duration CD4 stages >500, 350-500, 200-350 and ≤200 cells/mm^3 when on ART	No progression if virally suppressed; duration of each CD4 stage is doubled if virally unsuppressed as in (1).	Assumed
CD4 of individual after end of early transmission phase	Individuals can start at a lower CD4 stage with a probability from (4) dependent on their SPVL group.	Empirical
Probability of transmission from individual with CD4>500cells/mm^3 not on ART, per time step	Baseline transmission probability	Calibrated to model incidence and prevalence
Relative increase in transmission probability during early transmission phase	6.0 (when ~10% of transmissions early) and 26.0 (when ~40% of transmissions early), values from (5, 6)	Varied in simulations
Relative increase in transmission probability (compared to baseline transmission probability) when CD4 350-500 / 200-350 / ≤200 cells/mm^3	1.0 / 1.9 / 3.0	Assumed
Intervention The intervention model includes HIV-1 testing, male circumcision, ART provision and loss-to-follow up from ART. HIV-1 testing is divided into two separate rates. Firstly, there is a standard of care (background) rate that increases over time, with HIV-1 testing beginning in 2000 and ART becoming available in 2004, reflecting historical scale-up of testing in sub-Saharan Africa. Secondly, starting in 2015, intensive annual testing rounds are modelled, that mimic the HIV-1 testing component of the HPTN-071 combination HIV-1 prevention intervention. Men testing HIV-1 negative, who were not previously circumcised, are offered medical male circumcision in the model. Once circumcised, susceptibility to HIV-1 is reduced. Medical male circumcision rates differ over time, reflecting historical scale-up and strengthened testing as part of the combination prevention intervention. Individuals only start ART after a positive HIV-1 test result, although they may be lost to follow-up before this occurs. After ART start, individuals remain virally unsuppressed during an early ART period. This periods lasts on average 6 months. Thereafter, individuals become either virally suppressed, or not fully suppressed. Infectivity is reduced when an individual is on ART, but it is more substantially reduced if individuals are virally suppressed. Individuals on ART may drop out of treatment at any time after ART start. After drop-out, individuals may re-start therapy. The proportion of HIV+ individuals on ART under the different intervention scenarios is shown in figure 2 of the main text.		
Relative reduction in	0.6, from (7-9)	Empirical

susceptibility when circumcised		
Effectiveness of ART when no virally suppressed, or during early ART	0.45	Assumed
Effectiveness of ART when virally suppressed	0.9	Assumed
Annual intervention coverage	Fast: 90%; Slow: 20%; No intervention: 0%.	Varied in simulations
Uptake (% who successfully start ART)	Background: 30% (CD4>200); 60% (CD4≤200 cells/mm^3) Intervention: 50% (CD4>200); 75% (CD4≤200 cells/mm^3)	Assumed
Sequence Sampling Since 2000, sequences were randomly sampled at time of diagnosis in proportion to the number of annual new diagnoses. The proportions of individuals sampled between 2000-2014 and 2015-2020 are controlled by two parameters. One sequence was sampled per individual. The first parameter is the total number of sequences sampled. The second parameter is the proportion of sampled sequences that are obtained after intervention start in 2015. In addition, the sampling duration was also varied.		
Duration of sampling after intervention start.	3 years or 5 years.	Varied in simulations
Total number of sequences sampled	1600 or 3600, corresponding to 8% and 16% sequence coverage in the last year of the simulation. In comparison to the large sequence data sets that are available for concentrated epidemics in Europe or North America, these lower values reflect challenges in achieving high sequence coverage where large populations are infected.	Varied in simulations
Proportion of sampled sequences that are obtained after intervention start in 2015.	50% or 85%, corresponding to strong increases in sequence coverage after intervention start as expected in trial settings (10-12).	Varied in simulations
Ancestral relationships of HIV-1 viruses The topology of viral phylogenies does not necessarily correspond to the transmission tree, especially when viral infections persist life-long (13). To allow for such disagreement, we used a particular within- and between host coalescent model that is more fully described elsewhere (14, 15). The same model was used in the Village simulations. For each transmission chain, viral phylogenies with branch lengths in calendar time are generated through recursive application of a neutral within-host coalescent model. The infection time of the index case is considered as root of the within-host phylogeny of the index case, and any onward transmission events or sampling events as tips. Under these tip and date constrains, the within-host phylogeny of the index case is simulated assuming an increasing effective population size. For each new infection, the process is repeated and the within-host phylogenies of newly infected individuals are concatenated to the corresponding transmission tips of their transmitter. The model assumes that a single transmitted virion leads to clinical infection of the newly infected individual. For each transmission chain, the simulation produces a dated viral phylogeny that is rooted at the index case and has as tips the sampling times of all individuals in the same transmission chain that are sampled. The sub-trees that correspond to each transmission chain were concatenated to one multi-furcating root in order to obtain a single tree. For each sub-tree, the branch length of each sub-tree to the root reflects the time between the root age and the time of infection of the index case of the corresponding transmission chain in the model population.		
Within-host population size model	The logistic effective population size model is inherited from BEAST, BEAST::LogisticGrowthN0, with parameters N0tau=1, r=2.851904, v.T50=-2. These parameters were chosen so the final effective population size is broadly similar to estimates typically obtained with a BEAST Skyline model (16). See figure S11.	Assumed
Transmission bottleneck size	One virion transmitted.	Assumed

Age of multi-furcating root	Set so that the root age corresponds to estimated dates of origin of subtype C virus in South Africa (17).	Empirical
Sequence evolution		
Viral sequences were simulated along the viral tree from a starting sequence. To this end, branch lengths of the tree were first translated from calendar time to average number of substitutions per site per year. The evolutionary rate model included two components to reflect differences in evolutionary rates along transmission and non-transmission lineages (18, 19). Because one sequence is sampled per individual under the Regional model, non-transmission lineages correspond to the part of tip branches that correspond to viral evolution within sampled individuals. All other branches are part of transmission lineages. Evolutionary rates were drawn from two rate models of transmission and non-transmission lineages, and multiplied with branch lengths in units of calendar time to obtain branch lengths in units of average number of nucleotide substitutions per site per year. The starting sequence, from which all viral sequences were simulated, was obtained through ancestral state reconstruction of full-genome HIV-1 subtype C sequences. Viral sequences were simulated for the *gag*, *pol* and *env* genes from the starting sequence under a codon-based GTR+G sequence evolution model for each gene. The simulated *gag* gene was 1440 nucleotides long, the *pol* gene 2844 nucleotides, and the *env* gene 2523 nucleotides.		
Evolutionary rate of transmission lineages	Sampled from lognormal density with mean evolutionary rate 0.0022 and standard deviation (on the log scale) 0.3. Parameterized from phylogenetic analyses of subtype C sequences from southern Africa. See figure S12.	Empirical
Evolutionary rate of non-transmission lineages	Sampled from lognormal density with mean evolutionary rate 0.0044 and standard deviation (on the log scale) 0.5. Set to twice the rate of transmission lineages (18, 19). See figure S12.	Assumed
Nucleotide substitution rates	Informed from phylogenetic analyses of subtype C sequences from southern Africa. See figure S13.	Empirical

Table S2 Model components and assumptions of the Village Model.

Model parameters	Choice of parameter values	Type of evidence
Demographics Age is not explicitly modelled. Individuals enter the modelled population at 'birth,' already sexually mature, and leave the population when they die, either from disease related mortality or other reasons. The overall population is slowly growing in size.		
Population growth	Population growth was set at 1%/year to achieve incidence/prevalence comparable to a small Ugandan fishing village (20, 21).	Calibrated
Sexual partnerships Individuals are in a sexual partnership with one other individual. Partnerships are formed at 'birth' and last until the death of either partner. No partner switching is modelled. Individuals form partnerships assortatively based on risk group, and the frequency of extra-partner contacts is also determined by risk group. Sex workers do not form partnerships.		
Sexual risk behavior	3 risk group categories	Assumed
Proportion in risk group when entering population	Men: High: 50% / Low: 50%; Women: High: 47% / Low: 47% / Sex Worker: 6%.	Assumed
Mixing between risk groups	All partnerships are within the same risk group. 50-80% of contacts are with partner (if present); remaining contacts are weighted by risk group (ex: high risk more likely to contact other high risk and to contact sex workers)	Assumed

Partnership duration	Partnership lasts until the death of a partner	Assumed
Viral Introductions The epidemic begins with 1 infection in year 0. In simulations where imported sequences were included, this initial infection is also the ancestor of the 'imported' sequences. Viral introductions from outside of the focal population occur stochastically throughout the simulation.		
Seed cases	One female sex worker is infected at year 0, who automatically infects the populations outside the focal population, where the strain can evolve independently. HIV-1 transmission occurs from this point onwards.	Assumed
Proportion of viral introductions	Half of the simulations had no imported sequence migration. In simulations where this was included, 20% of transmissions were descendants of imported sequences by the end of the simulation.	Varied in simulations
HIV-1 infection HIV-1 negative individuals are exposed to risk of infection when they make a serodiscordant contact, either with their partner or with another individual. Transmission risk is dictated by stage of infection and set-point viral load, with acute stage and higher viral loads conferring higher transmission risk. If the HIV-1 infected individual is on ART, transmission does not occur. Individuals also only become infectious 2 weeks after infection.		
Set-Point Viral Load Value	4.5 $\log_{10}$ copies/mL, based on mean of subtype C infected individuals in the UK HIV epidemic	Empirical
Duration of early transmission phase (months)	3 months	Assumed
Probability of transmission	Values based on equations given in (22), based on viral load and whether in acute stage. Divided by 100 to convert to per-act rather than per-year risks.	Empirical
Relative increase in transmission probability during early transmission phase	0 (when ~4% of transmissions early) and 12.5 (when ~20% of transmissions early), value from (22).	Varied in simulations
Intervention No intervention or treatment is available before year 40. After year 40, ART is provided to approximately 20% of the population, including all sex workers. All individuals on ART are immediately fully suppressed, with viral load reduced to 50 copies/mL. ART is permanent (there is no loss to follow-up). As viral load determines disease progression, these individuals live much longer than individuals not on ART.		
Relative reduction in susceptibility when on ART	0.005 from (22).	Empirical
Relative increase in ART uptake in the 'fast' ART simulations	1.49	Assumed
Sequence Sampling Since year 40, sequences were randomly sampled at some point after acute infection. Sampling was done after the simulation was complete, with either 25% or 50% of the total number of individuals HIV+ at any time between years 40 and 45, with approximately the same number of individuals sampled each year. In simulations that were released as sequences only, 42 individuals were also sampled from the pre-intervention time period to replicate limited availability of older samples and to aid in phylogenetic reconstruction.		
Duration of sampling after intervention start.	3 years or 5 years.	Varied in simulations
Total number of sequences sampled	Between 638 and 1996, corresponding to 25% and 50% of the total number of infected individuals in the last 5 years of the simulation. This represented a situation where a small population was intensively sampled for the duration of an intervention, simulating a 'best possible' scenario where high sequence	Varied in simulations

	coverage was available.	
Proportion of sampled sequences that are obtained after intervention start in year 40.	100% in all simulations released as phylogenies. 95% in the four simulations released as sequences, corresponding to strong increases in sequence coverage after intervention start as expected in trial settings (10-12).	Varied in simulations
Ancestral relationships of HIV-1 viruses		
The topology of viral phylogenies does not necessarily correspond to the transmission tree, especially when viral infections persist life-long (13). To allow for such disagreement, we used a particular within- and between host coalescent model that is more fully described elsewhere (14, 15). The same model was used in the Regional simulations. For the transmission chain, viral phylogenies with branch lengths in calendar time are generated through recursive application of a neutral within-host coalescent model. The infection time of the index case is considered as root of the within-host phylogeny of the index case, and any onward transmission events or sampling events as tips. Under these tip and date constrains, the within-host phylogeny of the index case is simulated assuming an increasing effective population size. For each new infection, the process is repeated and the within-host phylogenies of newly infected individuals are concatenated to the corresponding transmission tips of their transmitter. The model assumes that a single transmitted virion leads to clinical infection of the newly infected individual. For the transmission chain, the simulation produces a dated viral phylogeny that is rooted at the index case and has as tips the sampling times of all individuals that are sampled. As all transmissions in the simulation descend from a single ancestral infection, there is only one transmission chain, and all generated sequences naturally coalesce to one ancestral sequence.		
Within-host population size model	The logistic effective population size model is inherited from BEAST, BEAST::LogisticGrowthN0, with parameters N0tau=0.00593, r=2.851904, v.T50=-2. See figure S11.	Assumed
Transmission bottleneck size	One virion transmitted.	Assumed
Age of multi-furcating root	Set so that the root age corresponds to estimated dates of origin of subtype C virus in South Africa (17).	Empirical
Sequence evolution		
Viral sequences were simulated along the viral tree from a starting sequence, which was obtained through ancestral state reconstruction of full-genome HIV-1 subtype C sequences from southern Africa by Gonzalo Yebra. Each viral phylogeny was run through piBUSS three times, once each for *gag*, *pol*, and *env*. All parameters used to simulate the sequences were taken from BEAST analysis of full-genome HIV-1 subtype C sequences from Southern Africa. Viral sequences were simulated for the *gag*, *pol* and *env* genes from the starting sequence under a codon-based GTR+G sequence evolution model for each gene. The simulated *gag* gene was 1479 nucleotides long, the *pol* gene 2999 nucleotides, and the *env* gene 2507 nucleotides.		
Nucleotide substitution rate gamma distribution shape parameter	Codons 1&2: 7.743; codon 3: 11.688	Empirical
Nucleotide substitution rate - *env*	Sampled from gamma distribution with mean evolutionary rate for codons 1&2: 2.98E-3 and codon 3: 5.52E-3, both with standard deviation 9.49E-7	Empirical
Nucleotide substitution rate – *gag/pol*	Sampled from gamma distribution with mean evolutionary rate for codons 1&2: 1.49E-3 and codon 3: 2.76E-3, both with standard deviation 4.75E-7	Empirical
Transition/transversion ratio	Codons 1&2: 0.139; codon 3: 0.765	Empirical

10

Table S3 Responses to the Phylodynamic Methods Comparison Exercise

Simulation model	Data set	Responses				
		Team Cambridge	Team Cambridge-London	Team Basel-Zürich[§]	Team London	Team Vancouver[§]
		(Total responses to the 5 reporting variables for each data set)				
Regional						
	D	0	0	0	5	5 *
	C	0	0	0	5	5 *
	A	0	0	0	5	5 *
	B	0	0	0	5	5 *
	O	0	5	5	5	0
	T	0	5	5	5	0
	S	0	5	5	5	0
	I	0	5	5	5	0
	R	0	4	5	5	0
	Q	0	5	5	0	0
	G	0	5	5	5	0
	N	0	5	5	5	0
	F	0	5	5	5	0
	L	0	5	5	5	0
	J	0	5	5	5	0
	P	0	5	5	0	0
	H	0	5	5	5	0
	K	0	4	5	5	0
	E	0	5	5	0	0
	M	0	5	5	5	0
Village						
	3	3	5	5	5	5
	2	3	5	5	5	5
	1	3	5	5	5	5
	4	3	5	5	5	5
	5	0	5	5	5	5
	11	0	5	5	5	5
	8	0	5	5	5	5
	9	0	5	5	5	5
	0	0	0	5	5	5
	6	0	5	5	5	5
	12	0	5	5	5	2
	7	0	5	5	5	5
	10	0	5	5	5	5

[§] Teams Basel-Zürich and Vancouver updated %Incidence estimates (Primary objective 2) after the data sets were unblinded. * Where sequences were provided, participants used full viral genomes (*gag+pol+env)* for inference. Team Vancouver also provided estimates based on partial *pol* sequences for two reporting variables on the indicated data sets.

Table S4. Estimating incidence and incidence reduction after a community-based intervention with phylogenetic methods on simulated PANGEA data sets.

Statistic	Responses

	Team Vancouver	Team Cambridge	Team Cambridge-London	Team Basel-Zürich	Team London
Correlation between phylogenetic estimates and true values					
%Incidence [1]	0.15	-0.78	0.91	0.83	0.64
Incidence ratio [2]	0.66	0.10	0.92	-0.07	0.15
Bias					
(Overall)					
%Incidence [3]	7.90	-1.83	0.35	3.15	0.57
Incidence ratio [3]	0.36	0.38	0.10	0.17	0.19
(Village Simulation Model)					
%Incidence [3]	7.75	-1.83	0.30	7.20	0.06
Incidence ratio [3]	0.50	0.38	0.11	0.21	0.13
(Regional Simulation Model)					
%Incidence [3]	8.31	-	0.39	0.44	1.06
Incidence ratio [3]	-0.06	-	0.09	0.12	0.23
Mean absolute error on the log scale for cross-comparison					
(Overall)					
%Incidence [4]	1.28	0.83	0.25	0.97	0.56
Incidence ratio [4]	0.43	0.39	0.14	0.32	0.33
(Village Simulation Model)					
%Incidence [4]	1.10	0.83	0.21	1.02	0.37
Incidence ratio [4]	0.50	0.39	0.14	0.26	0.20
(Regional Simulation Model)					
%Incidence [4]	2.07	-	0.29	0.98	0.71
Incidence ratio [4]	0.23	-	0.14	0.37	0.42

[1] Denote true % HIV-1 incidence per year after the intervention in PANGEA data set i by h_i, and estimated incidence by $\hat{h}_i$. Outliers with $\hat{h}_i$>20% were excluded, and the sample Pearson correlation between the remaining $\hat{h}_i$, h_i is reported. [2] Denote true incidence ratios after the intervention in PANGEA data set i by r_i, and estimated incidence ratios by $\hat{r}_i$. Outliers with $\hat{r}_i$>2 were excluded, and the sample Pearson correlation is reported. [3] Bias estimates of incidence and incidence reduction was calculated as $1/n \sum_i \hat{h}_i - h_i$ and $1/n \sum_i \hat{r}_i - r_i$ respectively, after outliers were removed as described above. [4] Mean absolute error in phylogenetic estimates of incidence and incidence reductions was calculated as $1/n \sum_i |log\ \hat{h}_i - log\ h_i|$ and $1/n \sum_i |log\ \hat{r}_i - log\ r_i|$ respectively on the log scale for cross-comparison, after outliers were removed as described above.

Table S5. Identification of HIV-1 incidence trends during a community-based intervention with phylogenetic methods on simulated PANGEA data sets.

True incidence trend	Classified as	Responses				
		Team Vancouver	Team Cambridge	Team Cambridge-London	Team Basel-Zürich	Team London
		Number (Percentage of responses correctly classified)				
Larger than 25% reduction in incidence	Declining	4 (44%)	0 (0%)	15 (88%)	9 (47%)	11 (55%)
	Stable	4	1	2	4	9
	Increasing	1	1	0	6	0
	Scenarios not evaluated	14	21	6	4	3
No or smaller than 25% reduction in incidence	Declining	0	1	1	3	4
	Stable	3	0	5	3	2
	Increasing	2	0	0	1	1
	Scenarios not evaluated	2	6	1	0	0

Table S6. Estimating the proportion of early transmissions before and after a community-based intervention with phylogenetic methods on simulated PANGEA data sets.

Statistic	Responses			
	Team Vancouver	Team Cambridge-London	Team Basel-Zürich	Team London
Correlation between phylogenetic estimates and true values (Village Simulation Model) [1]				
Just before the intervention	0.46	0.69	0.69	0
After the intervention	0.59	0.83	0.28	0
Correlation (Regional Simulation Model) [1]				
Just before the intervention	0.72	0.90	0.20	0.13
After the intervention	0.53	0.92	0.49	0.71
Bias (Village Simulation Model) [2]				
Just before the intervention	3.9	11.8	1.4	-2.7
After the intervention	2.7	10.1	-2.6	-5.3
Bias (Regional Simulation Model) [2]				
Just before the intervention	-13.3	-2.1	-9.4	-19.6
After the intervention	-11.9	-1.4	-13.2	-17.2
Mean absolute error (Village Simulation Model) [3]				
Just before the intervention	7.3	12.0	5.3	7.0
After the intervention	6.2	10.1	6.6	6.8
Mean absolute error (Regional Simulation Model) [3]				
Just before the intervention	13.3	4.1	14.8	20.0
After the intervention	12.0	3.9	13.2	18.0

[1] Denote true % early transmission just before or after the intervention in PANGEA data set i by p_i, and estimated proportions by $\hat{p}_i$. The sample Pearson correlation is reported. [2] Bias was calculated as $1/n \sum_i \hat{p}_i - p_i$. [3] The mean absolute error was calculated as $1/n \sum_i |\hat{p}_i - p_i|$.

Table S7. Significant predictors of error in phylogenetic estimates on simulated PANGEA data sets.

Primary objective	Covariates varied in the simulations (values of covariates) [1]	Significance of association with error			
		Team Cambridge-London	Team Basel-Zürich	Team London	Team Vancouver
		(P-value significance codes: ***: p<1e-3; **: p= 0.001-0.01; - : p>0.05)			
Incidence after intervention	True incidence after intervention (numerical)	***	-	0.02	-
	Simulation model (Village or Regional)	**	***	***	0.01
	Data provided (Sequences or trees)	-	-	-	-
	Frequency of viral introductions (<=5% or 20%)	-	***	-	-
	Sampling coverage at end of simulation (standard or high [2])	0.01	0.02	***	-
	Sampling duration after intervention start (3 years or 5 years)	0.04	-	-	-
	Proportion of sequences from after intervention start (50% or >80%)	-	***	0.04	-
Incidence reduction during intervention	True incidence ratio (numerical)	-	***	***	**
	Simulation model (Village or Regional)	-	***	0.03	0.02
	Data provided (Sequences or trees)	-	0.02	-	-
	Frequency of viral introductions (<=5% or 20%)	-	***	-	-
	Sampling coverage at end of simulation (standard or high [2])	-	-	-	-
	Sampling duration after intervention start (3 years or 5 years)	0.02	-	-	-
	Proportion of sequences from after intervention start (50% or >80%)	-	-	**	-
Proportion of early transmissions just before intervention	True proportion of early transmissions just before intervention (numerical)	0.01	***	***	**
	Simulation model (Village or Regional)	**	-	-	-
	Data provided (Sequences or trees)	0.017	-	0.03	-
	Frequency of viral introductions (<=5% or 20%)	-	0.02	***	-
	Sampling coverage at end of simulation (standard or high [2])	-	-	-	0.04
	Sampling duration after intervention start (3 years or 5 years)	-	0.04	-	-

	Proportion of sequences from after intervention start (50% or >80%)	-	0.02	-	-
Proportion of early transmissions after intervention	True proportion of early transmissions after intervention (numerical)	-	**	***	***
	Simulation model (Village or Regional)	**	-	-	-
	Data provided (Sequences or trees)	0.036	-	-	-
	Frequency of viral introductions (<=5% or 20%)	-	-	-	-
	Sampling coverage at end of simulation (numerical)	-	-	-	-
	Sampling duration after intervention start (3 years or 5 years)	-	-	-	-
	Proportion of sequences from after intervention start (50% or >80%)	-	-	-	-

[1] For each objective, the error e_i in phylogenetic estimates was defined so that errors were approximately normally distributed. Specifically, $e_i = log(\hat{x}_i) - log(x_i)$ for estimates $\hat{x}_i$ and true values x_i of incidence and incidence reduction, and $e_i = \hat{p}_i - p_i$ for estimates $\hat{p}_i$ and true values p_i of proportion of early transmissions. As predictors, we considered all variables v_j along which the PANGEA data sets were systematically varied in table 3. Variables took on either numerical or categorical values as indicated in brackets. We identified those v_j that were significantly associated with e_i. Specifically, we started with the full regression model containing all v_j as explanatory variables and then sequentially dropped predictors according to the generalised likelihood ratio test based on the BIC criterion (function stepGAIV.VR in the gamlss R package with $k = log(n)$).

[2] 16% versus 8% in the Regional data sets and 50% versus 25% in the Village data sets.

35

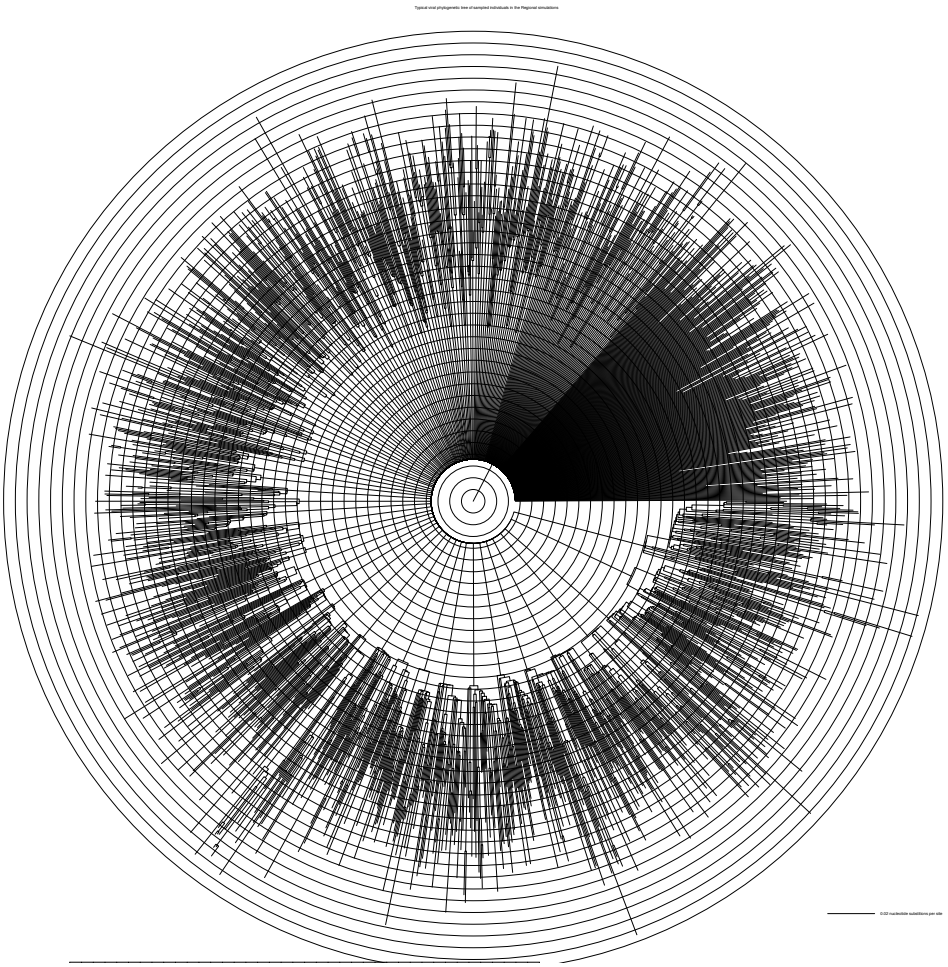

Figure S1. Viral tree of sampled individuals in a typical Regional simulation. 1600 (8%) of infected individuals from transmission chains in the Regional population were usually sampled. Due to frequent viral introductions, a large number of separate transmission chains were present in the modelled population. Each transmission chain was collapsed to sampled individuals. Corresponding viral trees with branch lengths in units of average nucleotide substitutions per site were generated under a coalescent model that also accounted for within-host viral evolution. Viral trees were connected to a single root sequence, with the root branch lengths reflecting time of viral introduction. The resulting tree of a typical Regional simulation is shown. Viral sequences were simulated along this viral tree.

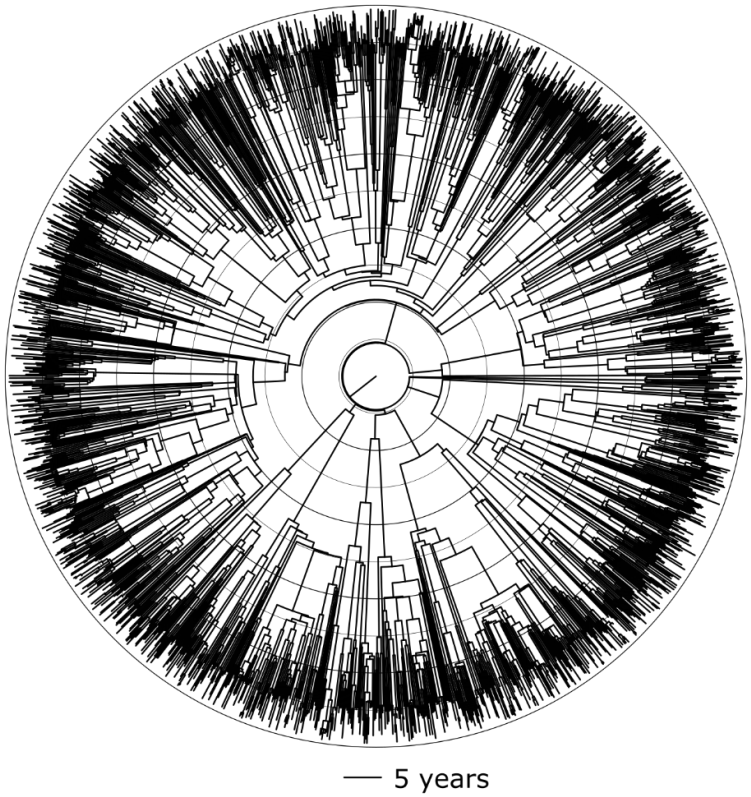

Figure S2. Viral tree of sampled individuals in a typical Village simulation. 638 to 1996 (25%-50%) of infected individuals from the entire transmission chain in the Village population were usually sampled. Parts of the entire transmission chain were modelled to have taken place outside the Village population. The transmission chain was collapsed to sampled individuals. Corresponding viral trees with branch lengths in units of time were generated under a coalescent model that also accounted for within-host viral evolution. The resulting tree of a typical Village simulation is shown. Viral sequences were simulated along this viral tree.

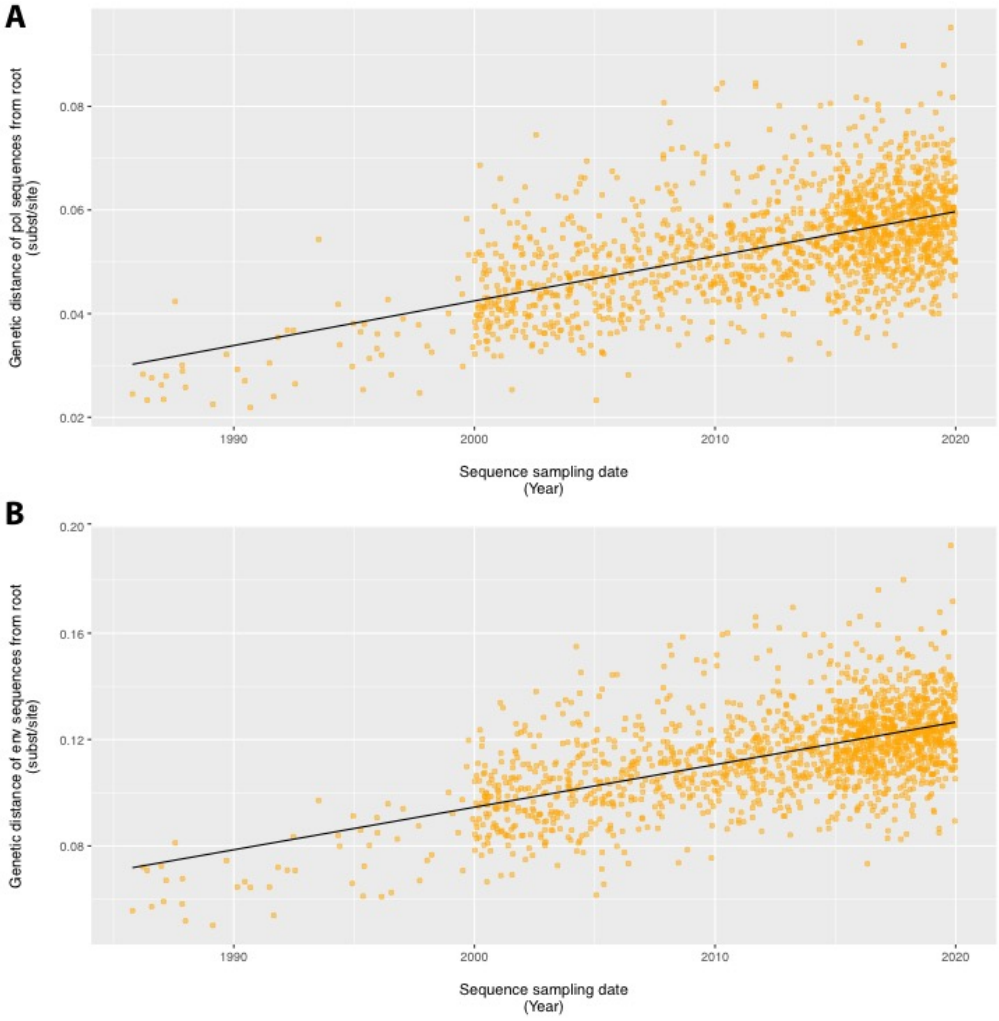

Figure S3. Root to tip divergence of a maximum likelihood trees reconstructed from viral sequences generated under the Regional model. One indicator of realism of simulated HIV-1 sequences is the degree to which viral evolution can be described by a single molecular clock. Viral trees were reconstructed with maximum-likelihood methods from simulated HIV-1 *pol* and *env* genes, and the patristic distances between the root and sampled taxa were plotted against sequence sampling dates. A linear regression model was fitted. (A) For *pol*, the mean evolutionary rate was 0.9e-3 subst/site/year. The variance explained by the constant clock model was $R^2 = 31\%$. (B) For *env*, the mean evolutionary rate was 1.6e-3 subst/site/year. The variance explained by the constant clock model was $R^2 = 35\%$.

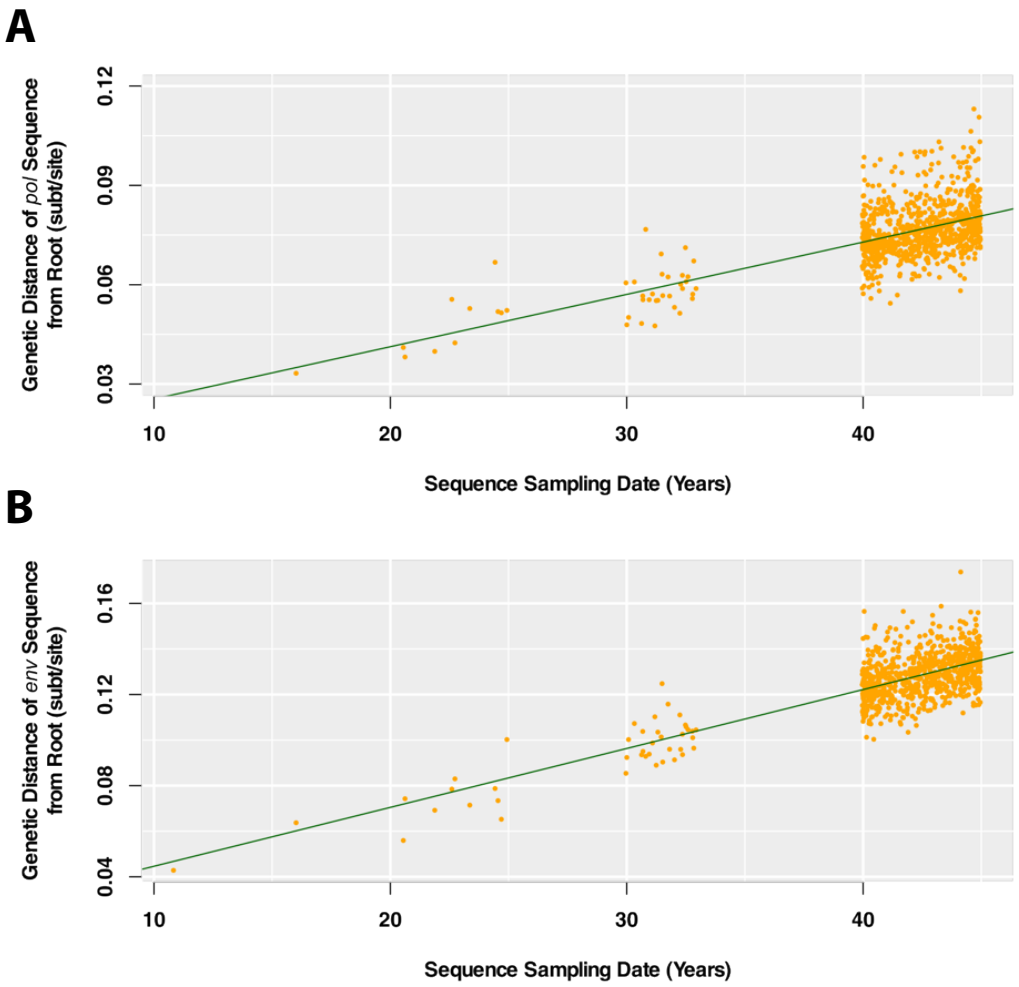

Figure S4. Root to tip divergence of a maximum likelihood trees reconstructed from viral sequences generated under the Village model. One indicator of realism of simulated HIV-1 sequences is the degree to which viral evolution can be described by a single molecular clock. Viral trees were reconstructed with maximum-likelihood methods from simulated HIV-1 *pol* and *env* genes, and the patristic distances between the root and sampled taxa were plotted against sequence sampling dates. A linear regression model was fitted. (A) For *pol*, the mean evolutionary rate was 1.6e-3 subst/site/year. The variance explained by the constant clock model was $R^2 = 34\%$. (B) For *env*, the mean evolutionary rate was 2.6e-3 subst/site/year. The variance explained by the constant clock model was $R^2 = 52\%$.

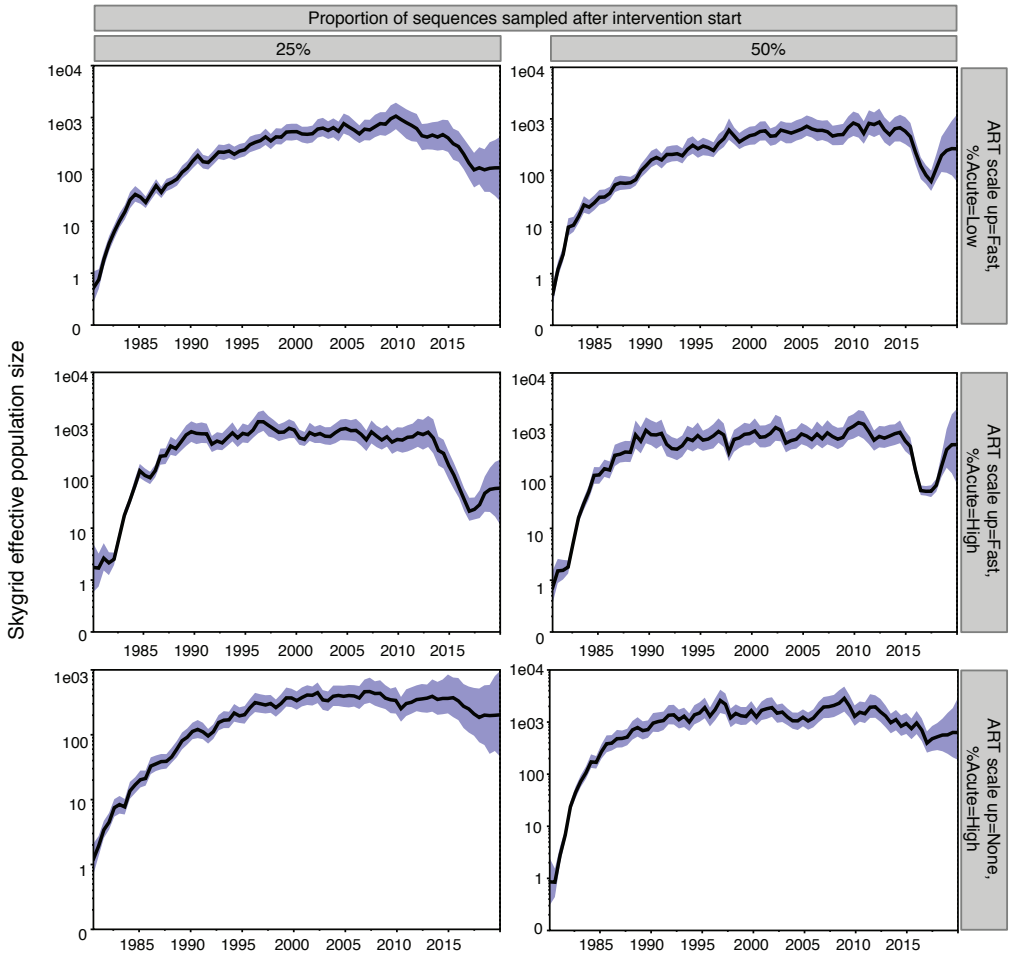

Figure S5. Signal to noise indicators for estimating incidence reductions on the Regional simulations at 8% sequence coverage. Skylines plots were reconstructed under the BEAST 1.8 Skygrid model to indicate if sufficient signal is present in the simulations to identify changes in incidence towards the present. For this analysis, true viral trees as provided for data sets E-T were used, using a previous multi-locus approach (23). Skyline plots are shown for data sets L, M, T (right column), and data sets similar to L, M, T but time homogeneous sequence sampling (~75 sequences sampled per year since 2000) (left column). For data sets L and T, incidence fell by approximately 60% reduction during the intervention. For data set M, incidence declined by approximately 10% as a result of improving standard of care. Qualitatively, these differences are visible in the Skyline plots under time homogeneous sequence sampling (left column). However, under rapidly increasing sampling as in the released data sets, these differences appear confounded (right column).

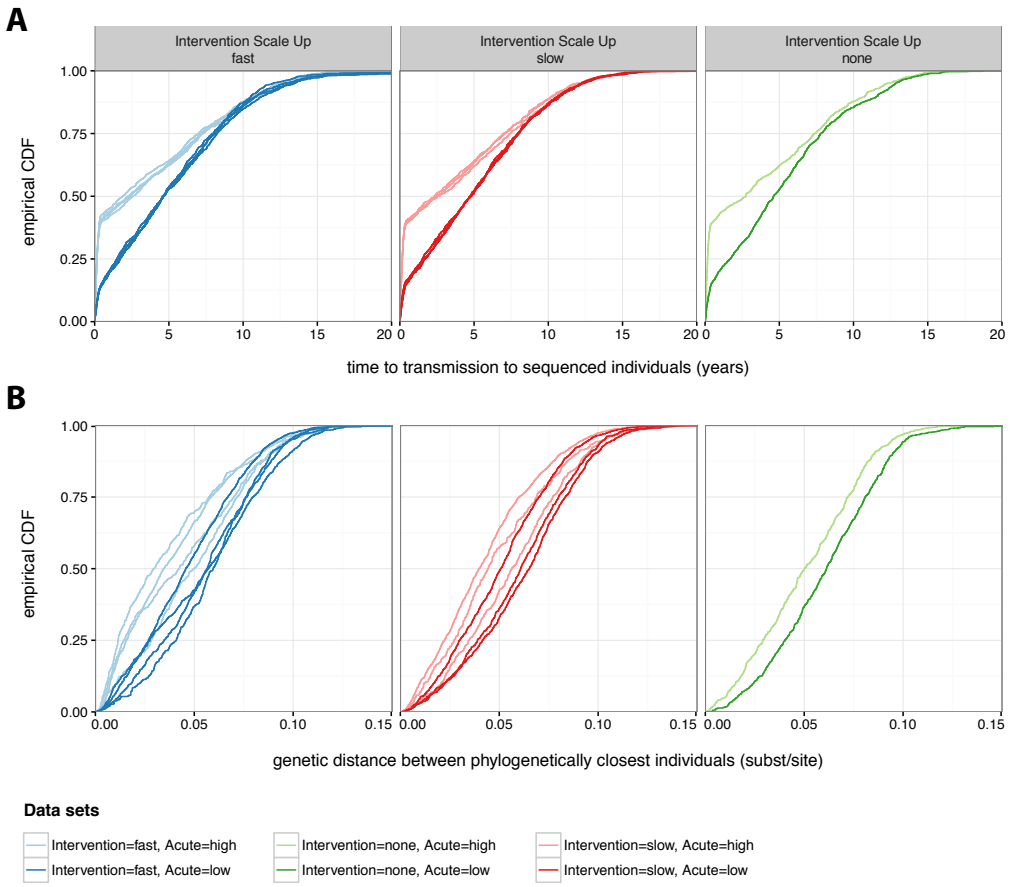

Figure S6. Signal to noise indicators for estimating the proportion of early transmission on the Regional simulations. Phylogenetic methods for estimating the proportion of early transmissions make use of information in branch length distributions, with shorter branches indicating faster transmission from individuals in early stages of infection. We report population-level indicators of differences in branch length distributions for the 10% and 40% Acute scenarios in the Regional simulations. Each line represents empirical cumulative distribution functions of a particular indicator, calculated on one simulation. Simulations are grouped into 10% and 40% Acute scenarios (darker and lighter lines) and intervention scenarios (color), to visualize signal versus noise. (**A**) Empirical CDF of the generation time distribution, among the subset of sequenced individuals. The generation time of sequenced individual was the time from infection of his transmitter to infection of the sequenced individual. High and low %Acute scenarios are clearly distinct from each other. The generation time was not known by participants. (**B**) In the simulations, generation times are reflected in the genetic distance between transmitters and recipients along the tree. However, not all transmitters appeared in the sequence data set. As a proxy, we considered the genetic distance between newly infected individuals and their genetically closest individual in the simulated data set. This proxy reflects information that was available to the participants. On average, high and low %Acute scenarios remained different from each other. This analysis suggests that some, but not strong, information exists in the Regional data sets for differentiating the 10% versus 40% Acute scenarios even at 8% sequence coverage.

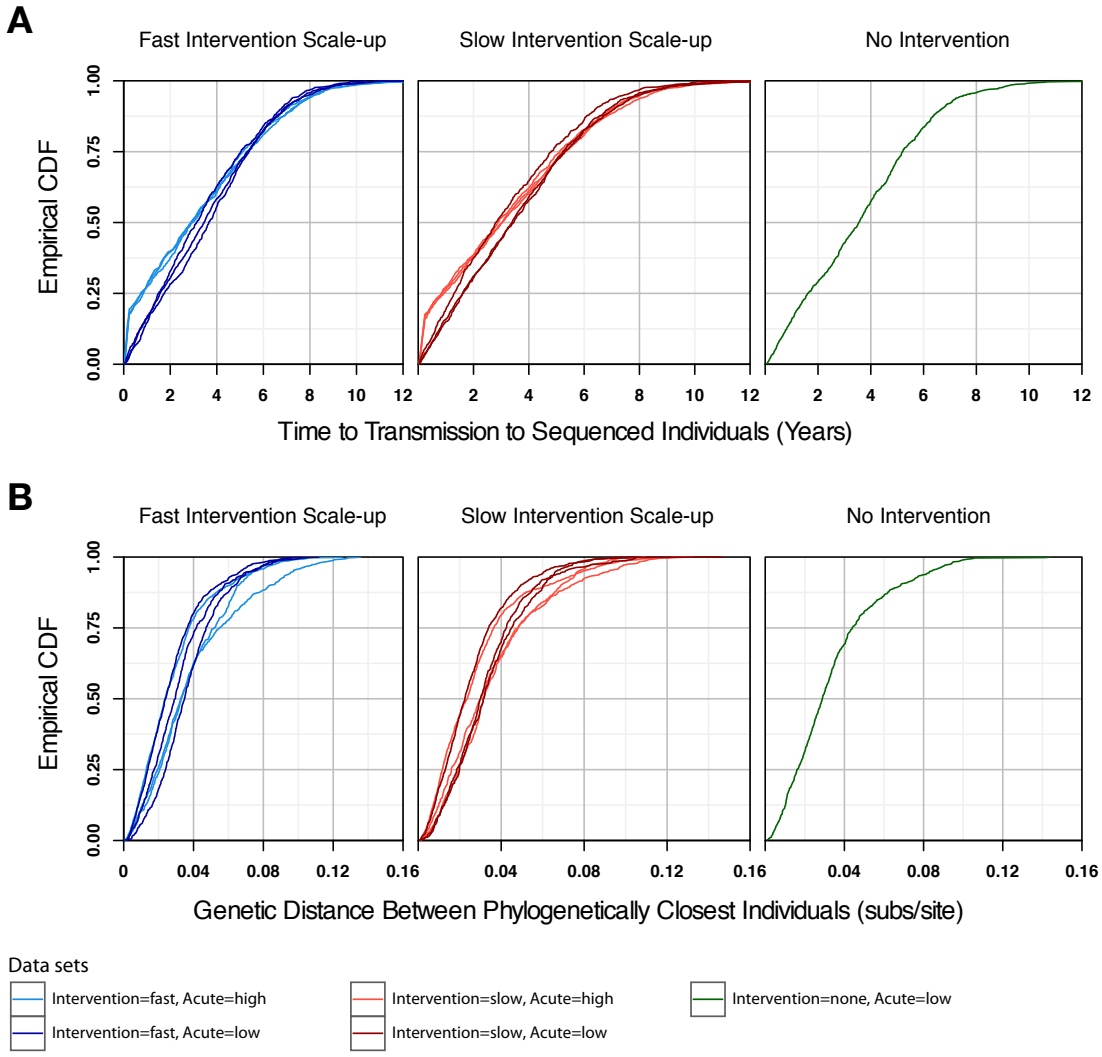

Figure S7. Signal to noise indicators for estimating the proportion of early transmission on the Village simulations. Phylogenetic methods for estimating the proportion of early transmissions make use of information in branch length distributions, with shorter branches indicating faster transmission from individuals in early stages of infection. We report population-level indicators of differences in branch length distributions for the 5% and 20% Acute scenarios in the Village simulations, see also figure S6. Each line represents empirical cumulative distribution functions of a particular indicator, calculated on one simulation. Simulations are grouped into 5% and 20% Acute scenarios (darker and lighter lines) and intervention scenarios (color), to visualize signal versus noise. (**A**) Empirical CDF of the generation time distribution, among the subset of sequenced individuals. The generation time of sequenced individual was the time from infection of his transmitter to infection of the sequenced individual. High and low %Acute scenarios are distinct from each other, although to lesser extent than for the Regional simulations. (**B**) In the simulations, generation times are reflected in the genetic distance between transmitters and recipients along the tree. We evaluated the extent to which the signal in (A) is still present at 25% and 50% sequence coverage, after transmission chains were collapsed to sampled individuals (transmitter may be lost) and then translated into viral trees. The plot shows, for all data sets, empirical CDFs of the genetic distances between phylogenetically closest individuals, as a proxy of transmission pairs. Unlike in (A), there is no apparent difference in high and low %Acute scenarios.

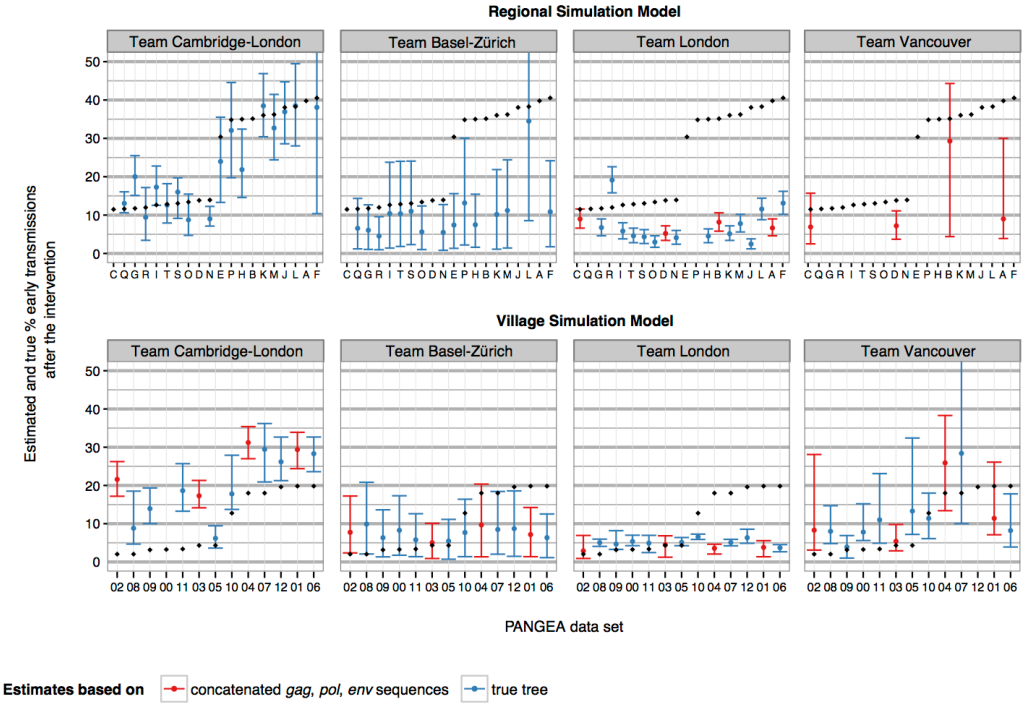

Figure S8 Estimates of the proportion of early transmissions after the intervention from phylogenetic methods on simulated PANGEA data sets. Submitted estimates are shown for each PANGEA data set by research team and model simulation (panels) and type of data provided (either sequences or the viral phylogenetic tree, color). Error bars correspond to 95% credibility or confidence intervals. True values are shown in black.

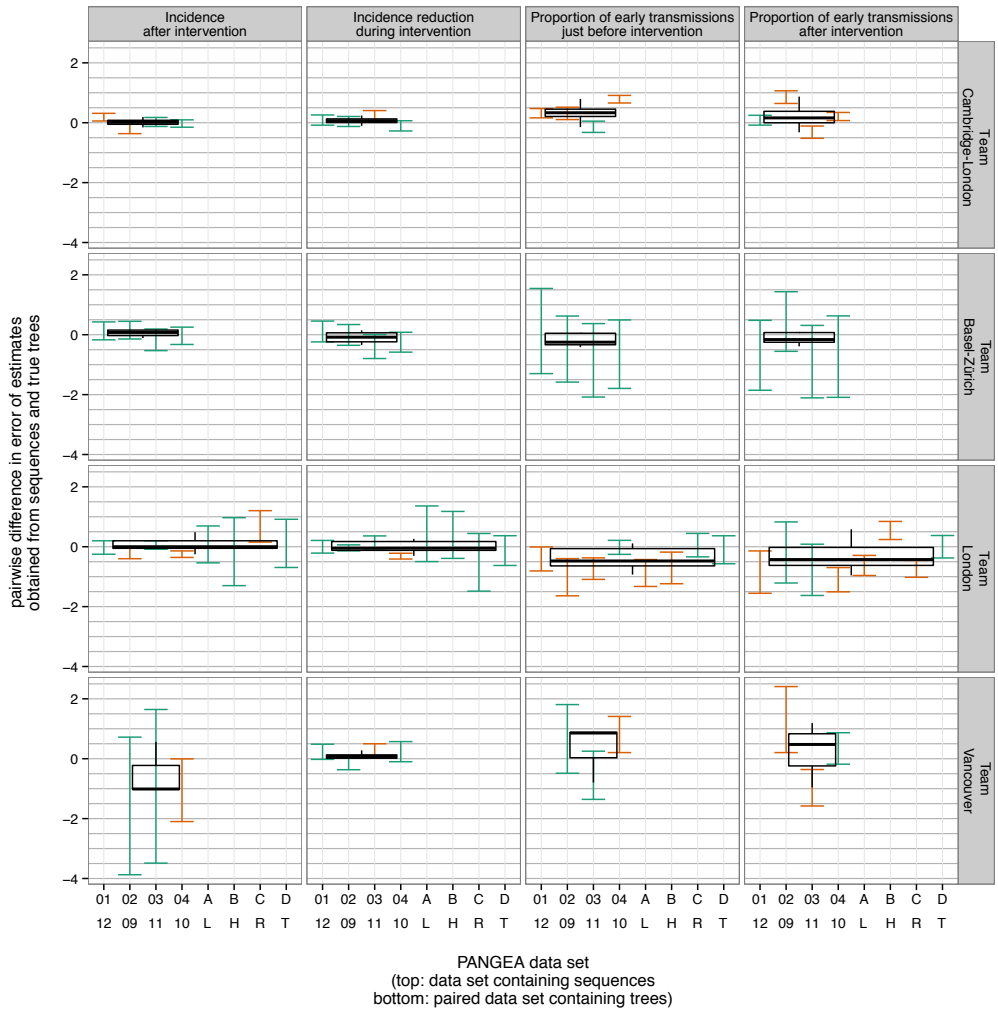

Figure S9. Differences in error of phylogenetic estimates obtained from sequence data, versus estimates obtained from sequence data and true phylogenetic trees known. PANGEA data sets containing sequences or true trees were considered, and paired if the underlying epidemiological scenario was identical (see x-axis and compare to table 3). For each objective, phylogenetic estimates $\hat{x}_i$ and true values x_i to each of these data sets were considered, and the error $e_i = log(\hat{x}_i) - log(x_i)$ was computed. The log scale was chosen so that errors were approximately normally distributed. The difference in errors $e_i - e_j$ between paired data sets i, j is plotted for each objective (columns) and each team (rows). Error bars indicate log transformed 95% confidence intervals; boxplots the distribution of central estimates; and significantly non-zero differences are highlighted in orange. Overall, phylogenetic estimates obtained from full genome sequence data sets were not significantly less accurate compared to estimates obtained with the true phylogenetic trees known (paired t-test: team Cambridge-London n=16, p=0.07; team Basel-Zürich n=16, p=0.79; team London n=32, p=0.033; team Vancouver n=13, p=0.87).

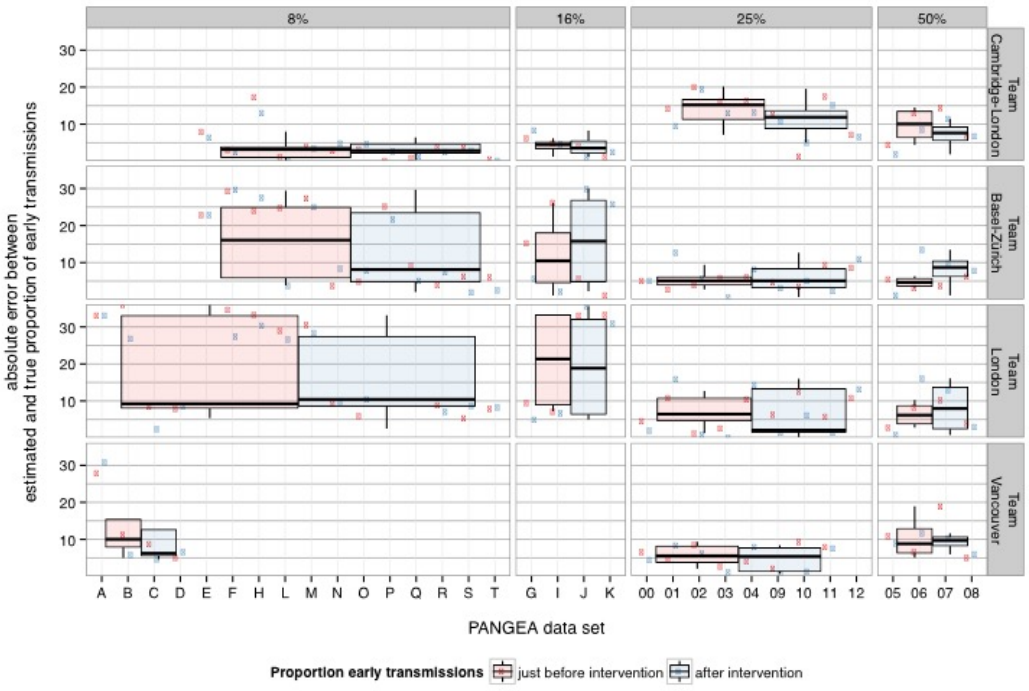

Figure S10. Accuracy of phylogenetic estimates of the proportion of early transmissions on simulated PANGEA data sets as a function of sampling coverage. For each PANGEA data set, the absolute error in the phylogenetic estimates of the proportion of early transmissions from individuals in their first three months of infection is shown by sequence coverage at the end of the simulation (panels). Each panel also compares the absolute error in estimates for the year just before the intervention (red) to that after the intervention (blue). Boxplots highlight the median absolute error and the interquartile range.

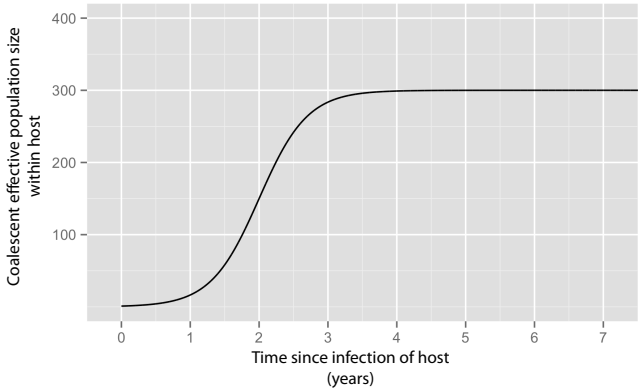

Figure S11. Within-host effective population size model of the Village and Regional simulations. Viral trees were generated under a hybrid within- and between-host coalescent model as described in tables S1 and S2, using the logistic effective population size model shown in this figure.

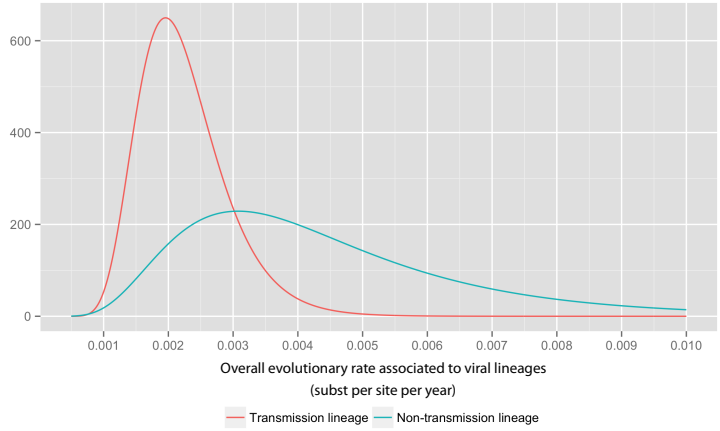

Figure S12. Sampling distribution of evolutionary rates of the Regional simulations. To simulate viral sequences along viral trees model as described in table S1, overall evolutionary rates were sampled from the log-normal distribution models shown in this figure and associated with transmission and non-transmission lineages of the viral tree. Sampled rates were used to translate branches into units of average substitutions per site per year.

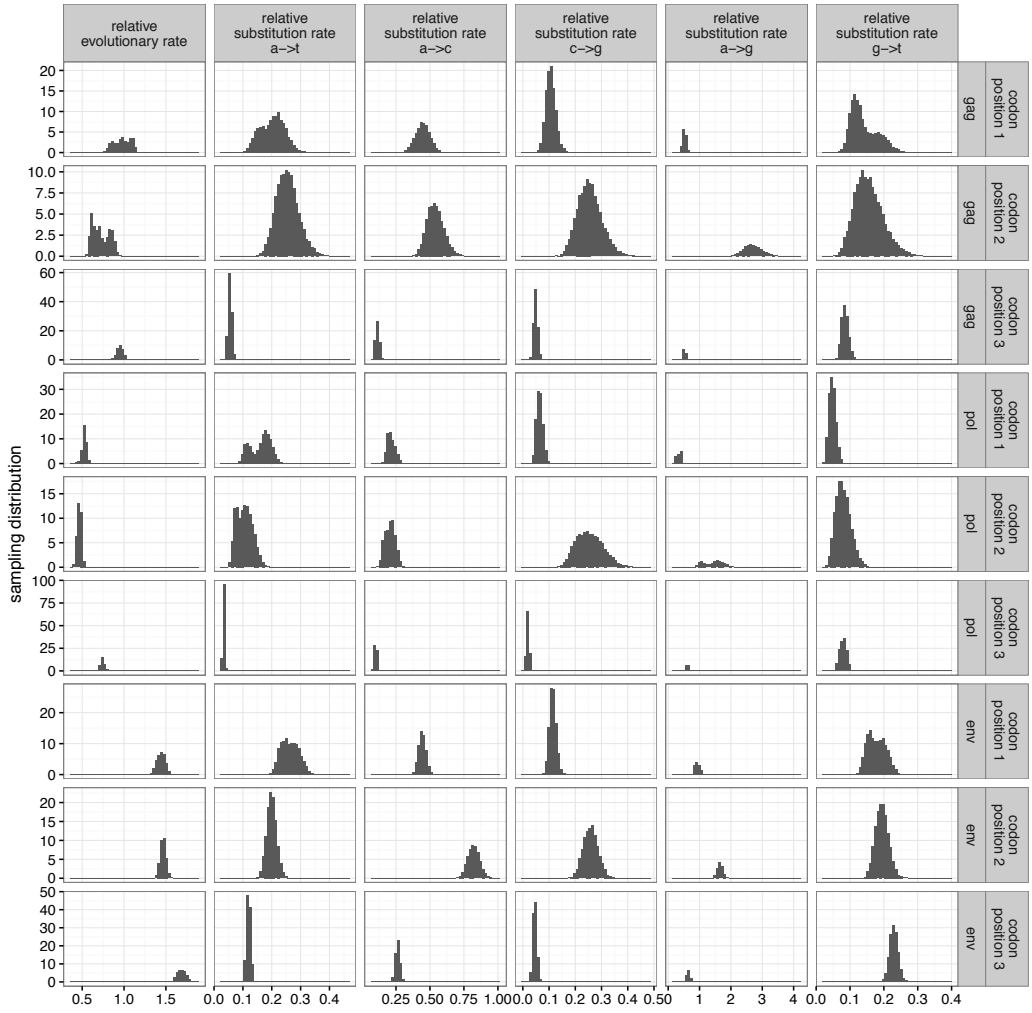

Figure S13. Sampling distribution of relative evolutionary rates and relative substitution rates of the Regional simulations. To simulate viral sequences along viral trees model as described in table S1, relative evolutionary rates by gene and codon position were sampled for each transmission chain as shown in this figure. The sampling distributions were obtained through BEAST phylogenetic analyses of full-genome HIV-1 subtype C sequences. GTR+Γ substitution models were used by gene and codon position, and relative substitution rates were sampled in the same manner. In the Village simulation, relative evolutionary rates and relative substitution rates were similar.

REFERENCES

1. Cori A, Ayles H, Beyers N, Schaap A, Floyd S, Sabapathy K, et al. HPTN 071 (PopART): a cluster-randomized trial of the population impact of an HIV combination prevention intervention including universal testing and treatment: mathematical model. PLoS One. 2014;9(1):e84511.
2. Hue S, Hassan AS, Nabwera H, Sanders EJ, Pillay D, Berkley JA, et al. HIV type 1 in a rural coastal town in Kenya shows multiple introductions with many subtypes and much recombination. AIDS research and human retroviruses. 2012;28(2):220-4.
3. Grabowski MK, Lessler J, Redd AD, Kagaayi J, Laeyendecker O, Ndyanabo A, et al. The role of viral introductions in sustaining community-based HIV epidemics in rural Uganda: evidence from spatial clustering, phylogenetics, and egocentric transmission models. PLoS Med. 2014;11(3):e1001610.
4. Cori A, Pickles M, van Sighem A, Gras L, Bezemer D, Reiss P, et al. CD4+ cell dynamics in untreated HIV-1 infection: overall rates, and effects of age, viral load, sex and calendar time. Aids. 2015;29(18):2435-46.
5. Hollingsworth TD, Anderson RM, Fraser C. HIV-1 transmission, by stage of infection. The Journal of infectious diseases. 2008;198(5):687-93.
6. Boily MC, Baggaley RF, Wang L, Masse B, White RG, Hayes RJ, et al. Heterosexual risk of HIV-1 infection per sexual act: systematic review and meta-analysis of observational studies. The Lancet infectious diseases. 2009;9(2):118-29.
7. Auvert B, Taljaard D, Lagarde E, Sobngwi-Tambekou J, Sitta R, Puren A. Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: the ANRS 1265 Trial. PLoS Med. 2005;2(11):e298.
8. Bailey RC, Moses S, Parker CB, Agot K, Maclean I, Krieger JN, et al. Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. Lancet. 2007;369(9562):643-56.
9. Gray RH, Wawer MJ, Polis CB, Kigozi G, Serwadda D. Male circumcision and prevention of HIV and sexually transmitted infections. Curr Infect Dis Rep. 2008;10(2):121-7.
10. Iwuji CC, Orne-Gliemann J, Tanser F, Boyer S, Lessells RJ, Lert F, et al. Evaluation of the impact of immediate versus WHO recommendations-guided antiretroviral therapy initiation on HIV incidence: the ANRS 12249 TasP (Treatment as Prevention) trial in Hlabisa sub-district, KwaZulu-Natal, South Africa: study protocol for a cluster randomised controlled trial. Trials. 2013;14:230.
11. Moore JS, Essex M, Lebelonyane R, El Halabi S, Makhema J, Lockman S, et al. Botswana Combination Prevention Project (BCPP) 2013 [Available from: https://clinicaltrials.gov/ct2/show/NCT01965470.
12. Hayes R, Ayles H, Beyers N, Sabapathy K, Floyd S, Shanaube K, et al. HPTN 071 (PopART): rationale and design of a cluster-randomised trial of the population impact of an HIV combination prevention intervention including universal testing and treatment - a study protocol for a cluster randomised trial. Trials. 2014;15:57.
13. Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet. 2009;10(8):540-50.
14. Didelot X, Gardy J, Colijn C. Bayesian inference of infectious disease transmission from whole-genome sequence data. Mol Biol Evol. 2014;31(7):1869-79.
15. Hall MD. Phylodynamics of infectious diseases of livestock: preparing for the era of large-scale sequencing: PhD thesis, University of Edinburgh; 2016.

16. Lemey P, Rambaut A, Pybus OG. HIV evolutionary dynamics within and among hosts. AIDS reviews. 2006;8(3):125-40.
17. Walker PR, Pybus OG, Rambaut A, Holmes EC. Comparative population dynamics of HIV-1 subtypes B and C: subtype-specific differences in patterns of epidemic growth. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases. 2005;5(3):199-208.
18. Vrancken B, Rambaut A, Suchard MA, Drummond A, Baele G, Derdelinckx I, et al. The genealogical population dynamics of HIV-1 in a large transmission chain: bridging within and among host evolutionary rates. PLoS Comput Biol. 2014;10(4):e1003505.
19. Alizon S, Fraser C. Within-host and between-host evolutionary rates across the HIV-1 genome. Retrovirology. 2013;10.
20. Opio A, Muyonga M, Mulumba N. HIV infection in fishing communities of Lake Victoria Basin of Uganda--a cross-sectional sero-behavioral survey. PLoS One. 2013;8(8):e70770.
21. Seeley J, Nakiyingi-Miiro J, Kamali A, Mpendo J, Asiki G, Abaasa A, et al. High HIV incidence and socio-behavioral risk patterns in fishing communities on the shores of Lake Victoria, Uganda. Sexually transmitted diseases. 2012;39(6):433-9.
22. Fraser C, Hollingsworth TD, Chapman R, de Wolf F, Hanage WP. Variation in HIV-1 set-point viral load: epidemiological analysis and an evolutionary hypothesis. Proc Natl Acad Sci U S A. 2007;104(44):17441-6.
23. Heled J, Drummond AJ. Bayesian inference of species trees from multilocus data. Mol Biol Evol. 2010;27(3):570-80.

Phylogenetic Tools For Generalized HIV-1 Epidemics: Findings from the PANGEA-HIV Methods Comparison

Supplementary Text S2: Information provided to participants – round 1

PANGEA-HIV (Phylogenetics and Networks for Generalised HIV Epidemics in Africa) is a major new initiative funded by the Bill and Melinda Gates Foundation to generate a large volume of next generation sequence data from African HIV cohorts to facilitate the phylodynamic characterization of generalized HIV epidemics.

The *PANGEA-HIV Methods Milestone 1* aims to evaluate existing phylogenetic methods in their ability to identify recent changes in HIV incidence in order to inform HIV prevention efforts in sub-Saharan Africa. Research groups are invited to participate in a blinded methods comparison exercise on simulated sequence data sets that capture different HIV transmission dynamics in generalized HIV-1 epidemics. Secondary aims of the exercise are to evaluate the merits of full genome sequence data, and the impact of changing sequence coverage.

With this exercise, *PANGEA-HIV* aims to direct further methods development in collaboration with participating research groups. Collaborative research teams will be formed to analyse the approximately 20,000 full genome HIV sequences with matched demographic and clinical patient data that are to be generated by PANGEA-HIV.

The PANGEA methods comparison working group, and *the PANGEA Consortium Executive group*

Table of Contents

PANGEA – HIV

PANGEA-HIV (Phylogenetics and Networks for Generalised HIV Epidemics in Africa) is a major new initiative funded by the Bill and Melinda Gates foundation to

1. deliver ~20,000 full length HIV-1 gene sequences along with associated clinical and demographic patient covariates from several African cohort and study sites: the Botswana Combination Prevention Project (Botswana), the Africa Centre for Health and Population Studies at the University of KwaZulu-Natal (South Africa), the MRC/UVRI Uganda research unit on AIDS (Uganda), the Rakai Health Sciences Programme (Uganda), and HPTN071/ Popart (Zambia and South Africa).
2. direct the further development of phylogenetic and phylodynamic methods to address key challenges in measuring, understanding and controlling HIV transmission dynamics of generalised HIV epidemics

Central questions to be addressed with existing or new phylogenetic/phylodynamic methods in the context of the generalised HIV epidemics in sub-Saharan Africa are

1. *What can be inferred about epidemic dynamics and sexual network characteristics from phylogenetic and self-reported epidemiologic data? What does that imply for control strategies in local or regional settings where HIV prevalence is well in excess of 20% of the adult population?*
2. *What are the transmission dynamics of a generalized epidemic and how do they differ from those of a concentrated epidemic where data are already available?*
3. *What are, at the individual level, the characteristics of infectiousness? Can we identify individuals at greater risk of transmitting the virus, and should these be prioritized for frequent testing and immediate ART?*
4. *How does Next Generation Sequence HIV full genome data improve the inference of transmission dynamics?*

PANGEA Methodology Milestone 1

Introduction

Different phylogenetic and phylodynamic methods have been adopted to characterize concentrated HIV epidemics in Europe and the US, largely from partial HIV-1 pol sequences collected through local, regional or national HIV treatment monitoring studies. There is little consensus on the ability of the various methods to accurately

analyse declining, stable or increasing HIV epidemics at different scales, both in terms of geographical range and epidemic scale. Little is known on the power of the various methods in reliably assessing these HIV epidemics from data that differs in completeness and may be biased. This is particularly so for the application of these methods to HIV next generation sequencing data from generalised epidemics. We expect these methods to have - ultimately - profound implications to our understanding of HIV-1 transmission and our ability to prevent transmission. It is of critical importance to understand - now - the applicability and potential shortcomings of these methods to the kind of data that will be generated by the PANGEA consortium.

Objectives

Research groups are invited to participate in a blinded methods comparison exercise on simulated sequence data sets that capture different HIV transmission dynamics in generalized HIV-1 epidemics.

The primary objective of the *PANGEA-HIV Methods Milestone 1* is to evaluate existing phylogenetic methods in their ability to accurately and reliably identify changes in HIV incidence that might occur over a few years representing a community-based intervention in sub-Saharan Africa in the simulation.

Secondary objectives of the exercise are to evaluate

- improvements in accuracy and power through the use of concatenated HIV-1 *gag*, *pol* and *env* sequence data as compared to HIV-1 *pol* sequence data,
- accuracy and power with respect to different sequence sampling intensities.

Simulation scenarios

Generalised HIV-1 epidemics were simulated for a relatively small "Ugandan" village population of ~8,000 individuals and a larger "South African" regional population of ~40,000 individuals from two structurally different, agent-based epidemiological models. Different incidence scenarios and contamination scenarios (source cases from outside the study population) were simulated. Different proportions of the population were sampled. Each of these scenarios is tagged with a unique identifier (sc[A-Z]). Further details on the simulated data are available below.

Each data set consists of several hundred simulated HIV-1 subtype C viral sequences, comprising *gag*, *pol* and *env* sequences. Several highly variable genome regions were excluded in the simulation. The label of each sequence contains additional information on the individual ID, date of sequence sampling, date of birth if available, and gender.

For the "village" simulation, each scenario contains sequences from a short time period of 3 years. Data sets with the same epi and sample identifiers are sampled from the same epidemic at the same sampling fraction. For the "regional" simulation, each scenario contains sequences sampled for a longer time period spanning 40 years. Scenarios differ in HIV-1 incidence dynamics from a time point after the year 2000. 20 replicate data sets using different random number seeds (rep[1-20]) were generated to evaluate power.

Evaluation criteria

We ask participating research groups to address, where possible, the following questions.

For each scenario of the village simulation,

- Was the epidemic growing, stationary or declining?
- If the epidemic was not stationary, what was the growth / negative growth rate?
- What is the proportion of annual new HIV infections relative to the population at risk of HIV infection?

For each scenario and each replicate of the regional simulation,

- Was the epidemic growing, stationary or declining by the end of the simulation?
- In which calendar year did incidence start to change?
- If the epidemic was not stationary, what was the growth / negative growth rate by the end of the simulation?
- What is the proportion of annual new HIV infections relative to the population at risk of HIV infection by the end of the simulation?

Comparing scenarios of the village simulation with the same epi and sample identifier,

- How large is the relative change in incidence between scenarios?

Comparing scenarios of the regional simulation,

- How large is the relative change in incidence by the end of the simulation between the stationary scenario(s) and the changing incidence scenario(s)?

Please use the PANGEAHIVsim_EvaluationSheet to return your responses (on Dropbox). Where possible, please conduct two analyses, the first using the concatenated *gag+pol+env* genome and the second using only the *pol* gene.

Research timeline and outputs

- **7th November 2014**
  Deadline for early research reports. Using the PANGEAHIVsim_Report document on Dropbox as a template, please describe briefly the methods you are using or have developed and provide a short summary of your preliminary findings on up to 2-3 pages. Feedback to / from participating research groups as needed..
- **2nd December 2014**
  Workshop to compare and consolidate initial results in London, UK. Participating analysis groups to give summaries of progress and feedback regarding additional simulations.
- **End December 2014**
  Deadline for submission of analyses. The PANGEA steering committee will consolidate and communicate the findings jointly with the participating groups to report to the Bill and Melinda Gates foundation, and in a publication.
- **16th May 2015**
  Satellite meeting to bring together final results of simulation based collaboration at HIV Dynamics and Evolution meeting in Budapest, Hungary.

PANGEA-HIV methods comparison working group

In alphabetical order

*Anne Cori ¶, Christophe Fraser ¶, Matthew Hall *, Emma Hodcroft *, Andrew Leigh Brown *, Mike Pickles ¶, Andrew Rambaut *, Manon Ragonnet-Cronin *, Oliver Ratmann ¶*

**University of Edinburgh, United Kingdom*

¶Imperial College London, United Kingdom

The data were generated by

- "African village" simulation: Emma Hodcroft
- "South African"-like regional simulation: Anne Cori, Mike Pickles
- HIV-1 *gag*, *pol* and *env* genome sequences: Matthew Hall, Oliver Ratmann

PANGEA Methodology Milestone 1 – Further details

South African regional simulation

The "South African" regional simulation scenarios were generated under an agent-based epidemiological model that has been developed as part of the HPTN 071 / PopART community randomized trial in South Africa and Zambia.

The epidemiological simulation starts in 1975 and ends in 2020. Individuals are stratified by gender, age, and level of sexual risk. Partnerships form and dissolve, with partner acquisition and concurrency depending on the sexual risk category. HIV transmissibility varies over the natural history of HIV by CD4 stages, acute/chronic HIV infection, circumcision status and condom use. Individuals within the simulated region have sexual partnerships with individuals outside the simulated region.

The viral molecular genetic simulation turns generated transmission chains into multiple phylogenies under a coalescent model that has within-host and between-host evolutionary components. The tips of the phylogeny correspond to sampling events. For each phylogeny, root sequences were generated from real sequences in the Los Alamos sequence database. Tip sequences were generated along the simulated phylogeny under a GTR site substitution model for the following genomic regions

1. gag: p17 start to pol PROT start; length 1440 nucleotides. The simulated gag gene does not include the last 14 amino acids of p6, due to the overlap with pol.
2. pol: PROT start to Integrase end; length 2844 nucleotides.
3. env: CDS signal peptide start to gp41 end; length 2523 nucleotides.

Three epidemiological scenarios A, B, C are generated, which differ in HIV-1 incidence dynamics from a time point after the year 2000. The following patient metavariables are available: Gender, Date of Birth (DOB), Date of Death (DOD), Time of sequence sampling (TIME_SEQ), CD4 count at time of sequence sampling (CD4_SEQ), Infected within one year of sequence sampling (INCIDENT_WITHIN1YEAR_SEQ).

Approximately 1,000 viral sequences are randomly sampled from HIV infected individuals between 1980 and 2020. Over time and across scenarios, the fraction of sampled sequences changes. Evolutionary simulation parameters are held fixed across scenarios.

FAQ

1. **How does the population size change over time?**
   The population size follows South African census estimates. In 1980, the population is a bit smaller than the census estimate, closer to ~ 20 million.
2. **Are there multiple introductions at the start of the simulation in 1975?**
   There are multiple introductions from outside the 'region', including the baseline year 1975. We generated the starting sequences based on phylogenetic estimates of the HIV introduction into South Africa, and expect TMRCA's before 1975.
3. **Are the viral lineages recombinating?**
   No, they are not – phew. ☺
4. **Is there ART in the model?**

Changing ART coverage is implicitly accounted for through changing transmission intensities.

5. **What has changed between the October and November simulations?** We changed the way the sequences were sampled as the specification changed to include CD4 counts. So, different individuals with different population identifiers are now sampled. The epidemic model parameters remained exactly the same.

African Village simulation

For this scenario, an HIV epidemic was simulated in a population of ~8,000 individuals using an individual-based model from first introduction until incidence stabilised. The simulations were run for 70 years. Partnerships and contact rate depend on the gender and risk group of the individuals, with HIV transmissibility varying through acute, chronic, and AIDS stages. Individuals may have contact with individuals from villages outside the focal population.

Three similar HIV epidemic scenarios were simulated. From each simulation, samples were taken during 3 different time periods each lasting 3 years. Each of these time intervals corresponds to a period of increasing, decreasing or stationary incidence dynamics. Two different sampling fractions are represented among the simulations, with one of the 3 scenarios sampled at both fractions leading in total to 9 scenarios A, B, …, I.

The simulation is set to keep the population approximately constant. During the peak years of the epidemic the population declines by about 1% per year.

Participants will notice that the sample times for all scenarios have been blinded. First, meaningless years were used to avoid preconceived bias about what was happening in the HIV epidemic in Africa at any given real date. Second, the sample dates for each time point have been adjusted to avoid bias based on the relative timing of the samples in each. As in real life, participants do not know beforehand the current dynamics of the epidemic.

Because of this, combining the data from any of the separate samples will give erroneous results.

Each sequence is a concatenated sequence of gag, pol, and env. Gag runs from 1-1479bp, pol from 1480-4479, and env from 4480-6987. Each sequence is labelled with the user ID, gender, and sample date (in decimal-year format). User IDs are randomly assigned and meaningless.

Phylogenetic Tools For Generaliz**ed HIV-1 Epidemics: Findings from the PANGEA-HIV Methods Comparison**

Supplementary Text S3: Information for participants – round 2

PANGEA-HIV (Phylogenetics and Networks for Generalised HIV Epidemics in Africa) is a major new initiative funded by the Bill and Melinda Gates Foundation to generate a large amount of next generation sequence data and to provide phylogenetic tools to measure the impact of HIV prevention efforts in generalized epidemics in sub-Saharan Africa.

Research groups are invited to participate in a blinded methods comparison exercise on simulated HIV sequence data sets to test the performance of current phylogenetic methods before their application on real data.

The *PANGEA-HIV Methods Milestone 1* aims to evaluate current phylogenetic methods in their ability to identify recent changes in HIV incidence and the proportion of transmissions that originate from individuals in early HIV infection. Secondary aims of the exercise are to evaluate the merits of full genome sequence data, the impact of sequence coverage, and the impact of the proportion of transmissions originating from outside the study area. The simulation scenarios are challenging and capture detailed aspects of HIV transmission dynamics and intervention efforts that are typical for sub-Saharan Africa.

Based on the outcomes of this exercise, collaborative research teams will be formed to analyse the full genome HIV that are to be generated by PANGEA-HIV.

The PANGEA methods comparison working group, and *the PANGEA Consortium Executive group*

PANGEA – HIV

PANGEA-HIV (Phylogenetics and Networks for Generalised HIV Epidemics in Africa) is a major new initiative funded by the Bill and Melinda Gates foundation to

1. deliver a large volume of full length HIV-1 gene sequences along with associated clinical and demographic patient covariates from several African cohort and study sites: the Botswana Combination Prevention Project (Botswana), the Africa Centre for Health and Population Studies at the University of KwaZulu-Natal (South Africa), the MRC/UVRI Uganda research unit on AIDS (Uganda), the Rakai Health Sciences Programme (Uganda), and HPTN071/ Popart (Zambia and South Africa).
2. direct the further development of phylogenetic and phylodynamic methods to address key challenges in measuring, understanding and controlling HIV transmission dynamics of generalised HIV epidemics

PANGEA-HIV aims to address the questions

1. *What can be inferred about epidemic dynamics and sexual network characteristics from phylogenetic and self-reported epidemiologic data? What does that imply for control strategies in local or regional settings where HIV prevalence is well in excess of 20% of the adult population?*
2. *What are the transmission dynamics of a generalized epidemic and how do they differ from those of a concentrated epidemic where data are already available?*
3. *What are, at the individual level, the characteristics of infectiousness? Can we identify individuals at greater risk of transmitting the virus, and should these be prioritized for frequent testing and immediate ART?*
4. *How does Next Generation Sequence HIV full genome data improve the inference of transmission dynamics?*

PANGEA HIV Methods Milestone 1

Introduction

Phylogenetic methods have been widely applied to characterize concentrated HIV epidemics such as Europe and the US, but not in the context of generalized HIV epidemics in sub-Saharan Africa. These analyses have been based on partial HIV sequence data, but not on full genome sequence data. It is not known how current methods are best scaled to full genome sequence data, and if they can accurately uncover aspects of HIV transmission dynamics from generalized HIV epidemics under typical sequence sampling conditions. We expect these methods to have - ultimately - profound implications to our understanding of HIV-1 transmission and our ability to prevent transmission. It is of critical importance to understand - now - the applicability and potential shortcomings of these methods to the kind of data that will be generated by the PANGEA consortium.

To assess the performance of current phylogenetic methods in a controlled setting, the PANGEA HIV methods comparison working group implemented two highly detailed epidemiological and evolutionary models of generalized HIV epidemics to simulate full genome HIV sequences and phylogenetic trees. The simulation scenarios are designed to capture central aspects of partnering HIV prevention study sites, e.g. the Treatment as Prevention (TasP) trial in South Africa.

Objectives

Research groups are invited to participate in a blinded methods comparison exercise on simulated sequence data sets and simulated phylogenetic trees to address the following specific objectives.

Primary objectives
To evaluate existing phylogenetic methods in their ability to measure

1. changes in HIV incidence that might occur over a few years representing a community-based intervention in sub-Saharan Africa in the simulation. The outcome measure is annual HIV incidence in % of the number of individuals that are at risk of HIV infection.

2. the proportion of HIV transmissions arising from individuals in early HIV infection at the start of the community intervention. The outcome measure is the proportion of new HIV cases from those in early HIV infection in the year before the start of the community intervention.

Secondary objectives
To evaluate

3. improvements through the use of concatenated HIV-1 *gag*, *pol* and *env* sequence data as compared to HIV-1 *pol* sequence data. The outcomes measure is the accuracy in answering the primary objectives.
4. the impact of sequence sampling coverage. The outcomes measure is the accuracy in answering the primary objectives.
5. the impact of the proportion of transmissions that occur from outside the study area. The outcomes measure is the accuracy in answering the primary objectives.

Overview of simulation models

Generalised HIV-1 epidemics were simulated for a relatively small "Ugandan" village population of ~8,000 individuals and a larger "South African" regional population of ~80,000 individuals from two structurally different, agent-based epidemiological models.

The regional simulation captures individual-level HIV transmission dynamics in a larger regional population that is broadly similar to a site (cluster) of the HPTN071/PopART HIV prevention trial in South Africa. Standard of care improved according to national guidelines over time. In a subset of simulations, an additional comprehensive HIV prevention combination package started in 2015 for three years, broadly similar to the HPTN071/PopART intervention. Since 2015, the population is monitored more actively, resulting in a moderate sampling coverage at a large scale (<10% of the infected population). Contamination through transmission from outside the regional area occurs in the range of available estimates for sub-Saharan Africa. More information is available in the appendix 'Regional simulation'.

The village simulation captures individual-level HIV transmission dynamics in a small village population. An intervention campaign started at some point after the epidemic peaked, and was followed for a long period of time. Some time after intervention was started, the simulated campaign was intensely monitored for three years, resulting in a relatively high sampling coverage at a small scale (>10% of the infected population). There is no additional intervention. Contamination through transmission from outside the village area is minimal. More information is available in the appendix 'Village simulation'.

Overview of simulation scenarios

Data sets were simulated from both models. Phylogenetic inference is often computationally expensive. To ease the computational requirements, the simulated phylogeny is provided for a subset of simulations that is of secondary importance. To address the primary objectives, parameters relating to HIV transmission dynamics and the efficacy of the prevention campaigns were varied for each model. Sequence data was generated. To address the

secondary objectives, parameters relating to the sampling frame and the proportion of transmissions from outside the study area were varied. Phylogenies were generated. Additional data for each sequenced individual, and additional population surveys on the course of the HIV epidemic until 2015 are available for each scenario. Please see the appendices for further information.

Primary Evaluation criteria

We aim to evaluate HIV transmission dynamics around an evaluation period that is close to the end of the simulation. For the village simulation, the evaluation period coincides with the intensely sampled period (see appendix). For the regional simulation, the evaluation period starts in January 2015 and ends one year before the end of the simulation, in almost all cases December 2019.

For each data set containing simulated sequences:

i. During the evaluation period, was incidence stable, declining or increasing? Please provide answers in terms of
$$'stable','declining','increasing'$$
ii. What is the annual % incidence in the last year of the evaluation period? Please provide answers in % incidence,
$$\%INC_{t\_e} = \frac{INC_{t\_e}}{S_{t\_e}},$$
where t_e is the last year of the evaluation period, INC_{t_e} is the estimated number of new cases in year t_e, and S_{t_e} is the estimated number of sexually active individuals that have not been infected until t_e.
iii. Comparing the year preceding the evaluation period to the last year of the evaluation period, what is the ratio in annual % incidence? Please provide answers in terms of
$$Ratio = \frac{\%INC_{t\_e}}{\%INC_{t\_s}}$$
where t_s denotes the year preceding the study period.
iv. Was the proportion of transmissions that originated from individuals in early HIV infection in the year preceding the evaluation period below 10%, between 10-30%, or above 30%? Please provide answers in terms of
$$'<10\%','10-30\%','>30\%'$$
Here, early HIV infection is understood as the first 3 months after HIV infection.
v. What is the proportion of transmissions that originated from individuals in early HIV infection in the year preceding the evaluation period? Please provide answers in terms of
$$\%Early_{t\_s} = \frac{INC_{t\_s}(from\ early)}{INC_{t\_s}}$$
vi. What is the proportion of transmissions that originated from individuals in early HIV infection in the last year of the evaluation period? Please provide answers in terms of
$$\%Early_{t\_e} = \frac{INC_{t\_e}(from\ early)}{INC_{t\_e}}$$

Secondary Evaluation criteria

For each of the above data set, please report the outcome measures in (i) to (vi) above for two analyses:

vii. Using only the *pol* sequences
viii. Using the concatenated gag+pol+env sequences.

For each of the data sets containing a simulated phylogeny, please also report the outcome measures in (i) to (vi).

These simulation scenarios may vary from those used to evaluate the primary objectives in terms of

- The sequence sampling coverage
- When sequences are sampled during the course of infection
- The annual proportion of transmissions that originate from outside the study area.

Other Evaluation criteria

We encourage participants to fine-tune their phylogenetic methods to address the above outcome measures. These will be given preference in the methods comparison, presentations and publications. The methods comparison group may consider supplementary outcome measures provided by participants, if these can be directly calculated on the simulated data. Please note that effective population sizes or reproduction numbers cannot be calculated from the agent-based simulations.

Reporting

We will make evaluation sheets available as for the training round of the methods comparison exercise.

Thank you

The PANGEA-HIV methods comparison working group would like to thank all participants for their interest and contributions thus far.

Timelines

February 2015	Release of simulated sequence data sets and simulated phylogenies
27.02.2015	Presentation of interim results based on the training round at CROI
06.05.2015	Deadline for submission of analyses
13.05.2015	HIV Dynamics and Evolution 2015, where we will present an overview of the results of the methods comparison exercise. Individual submissions from participants are encouraged.
16.05.2015	PANGEA-HIV satellite meeting to present and discuss final results of the exercise in detail. All participating teams will have the opportunity to present their work.

PANGEA-HIV methods comparison working group

Leads: Christophe Fraser, Oliver Ratmann (Imperial College London); Andrew Leigh Brown, Emma Hodcroft (Edinburgh) Contributors: *Mike Pickles*, *Anne Cori* (Imperial College London)*; Matthew Hall, Samantha Lycett, Manon Ragonnet-Cronin, Gonzalo Yebra, Andrew Rambaut (University of Edinburgh)*

Phylogenetic Tools For Generalized HIV-1 Epidemics: Findings from the PANGEA-HIV Methods Comparison

Supplementary Text S4: PANGEA-HIV specification of the birth-death skyline method with sampled ancestors

We performed our phylodynamic analyses using the add-ons bdsky [2] and SA [3] in BEAST v2.0 [1]. We estimated the posterior distribution of the epidemiological parameters using fixed trees. In case of sequencing data, we first estimated time trees using RAXML/ExaML [4][6] & LSD [5], and then inferred epidemiological parameters in BEAST.

Parameter inference

For calculation of epidemiological parameters we assumed the birth-death skyline model [2] with sampled ancestors (SA) [3]. The model assumes a transmission rate lambda, removal rate without sampling μ and sampling rate Ψ. Upon sampling, an individual is removed with probability r (called removal probability). It turned out to be crucial to allow for sampled individuals to further transmit with some probability $(1-r)>0$ (the original model in [2] assumed $r=1$).

Under the skyline model, time is partitioned into different intervals. Within each interval, the parameters are assumed to be constant. Across intervals, the rates may change in an arbitrary fashion.

For each interval, we estimate the effective reproductive number $R = \lambda / (\mu + \Psi r)$ and the becoming-non-infectious rate $\delta = \mu + \Psi r$ at which an infected individual becomes non-infectious. The transmission rate λ is therefore given by $R*\delta$, and the sampling proportion is $\Psi /(\mu + \Psi)$.

We partitioned time into three intervals:

- The Village data has one interval over the evaluation period and two in the first 40 years, with their lengths being chosen such that there are an equal number of branching events in each interval. The data was not informative enough to split up the evaluation period into several intervals to gain a higher understanding of what happened during the treatment time. We assumed the HKY model for sequence evolution for the datasets consisting of sequences.
- For the Regional dataset, time was partitioned following the provided information. In particular, for the removal probability r the first interval runs from 1970 until 2004, the second until 2015 and the third over the evaluation period, 2015-2019. For the sampling proportion R and the becoming-non-infectious rate δ, the intervals run from 1970 to 2000, the second until 2015 and the third over the evaluation period.

Sampling proportion

For the sampling proportion $\Psi / (\mu + \Psi)$, we set a prior according to the given sampling densities stated in the information for participants, i.e. for villages 0, 1, 2, 3, 4, 9, 10, 11 and 12 we assumed a uniform prior between 15 - 40% and for villages 5, 6, 7 and 8 between 40 – 100%. For the regional data we used a uniform prior for the sampling proportion between 5%-10%.

Removal probability - Villages

The prior distribution for the removal probability r was chosen based on the probability of a sampled infected individual to be on treatment, and therefore presumed to be non-infectious. Using the provided numbers, for the villages, we calculated the ratio of number of patients on

treatment in year 40 (T_{40}, given in metadata) and the number of infected people at the end of year 39 (P_{39}*8000 with prevalence P_{39} from metadata), $r = T_{40} / P_{39}$*8000. For village 0 the removal probability r was set 0, as no infected individuals are on treatment.

Removal probability - Regionals

For regionals, we defined r in the middle interval as the number of sampled individuals on treatment and sampled before 2015 out of the total number of samples. We defined r in the last interval as the number of sampled individuals on treatment and sampled after 2015 out of all samples.

Prior assumptions for the epidemiological parameters R and δ

We used a lognormal prior for R (lognormal with mu=0 and sigma=0.75) and δ (lognormal with mu=-1 and sigma 0.5) for the villeages and regions.

Sensitivity analyses

To explore sensitivity of our estimates towards prior assumptions, analyses were repeated with a wider prior on R and δ for villages (lognormal with mu=0.0 and sigma=1.5), as well as with removal probability r=0. To see how sensitive the calculation is to changes in the sampling proportion, the true sampling proportion (25% and 50% respectively) was used for an additional analysis. Results did not change noticeably using these stricter priors (not shown).

Assessing criteria i-vi

We used the posterior distributions for R and δ to assess the criteria i-vi of the simulation study.

Criteria i:

When analysing the Village data, we calculated the difference of expected incidence in year 44 (EI_{44}, see Criteria ii) and year 39 (EI_{39}, see Criteria ii). This difference was calculated for all values sampled by the MCMC, i.e. we obtained its posterior distribution. The 95% highest posterior density (HPD) interval was calculated to get the lower and upper bound for the difference.

The same calculation was performed on the Regional dataset, using the last year before the simulation ends (usually year 2018, except for Region F and O, where it is 2016) and year 2014.

If the HPD interval contained 0, we report no significant evidence rejecting stable incidence, if the HPD interval is entirely below 0 we report decreasing incidence, and if the HPD interval is above 0 we report increasing incidence.

Criteria ii:

We calculated incidence and number of susceptible people in year 43 based on our estimates of R and δ at the end of the evaluation period. We assumed that the total population at the end of year 43 is N_{43} = 8000*1.01^4, based on a population size of 8000 at the end of year 39 and a growth rate of 1% per year. The number of infected individuals at end of year 43 is, I_{43} = 8000* P_{39}* exp((λ - δ)*4), using $R = \lambda/\delta$ and δ estimates for the evaluation period and P_{39} being the prevalence at end of year 39. Therefore the susceptible population size at the end of year 43 is $S_{43} = N_{43} - I_{43}$.

The expected incidence in year 44 (EI_{44}) was calculated by subtracting the number of infected individuals at the end of year 43 (I_{43}) from the number of infected individuals at the end of year 44 (I_{44}), and by adding the number of individuals (out of I_{43}) that became non-infectious

during year 44 (BU_{44}), i.e. $EI_{44} = I_{44}-I_{43}+BU_{44}$, with $BU_{44} = I_{43}$ (1- e^(- δ)). The annual percent incidence was calculated via $\%INC_{t\_e} = EI_{44}/S_{43}$.

This value EI_{44}/S_{43} was calculated for each sample of the MCMC, i.e. we obtained its posterior distribution. The final value was calculated by taking the mean of all EI_{44}/S_{43} values. The 95% HPD interval was calculated to get the lower and upper bound EI_{44} values.

Criteria iii:

Now we calculated the ratio *Ratio* = $(EI_{44}/S_{43}) / (EI_{39}/S_{38})$, with $N_{38} = 8000/1.01$, and $I_{38} = 8000*P_{39} * \exp(-(\lambda - \delta))$, where λ and δ result from the middle interval in the skyline model. This value *Ratio* was calculated for all values sampled by the MCMC, i.e. we obtained its posterior distribution. The final value was calculated by taking the mean of all *Ratio* values. The 95% HPD interval was calculated to get the lower and upper bound *Ratio* values.

Criteria iv-vi

We addressed criteria iv-vi (proportion of transmissions that originated from individuals within their first 3 months of HIV infection) by employing the multi-type birth-death model [6]. This model allows us to analyse the data using exactly the same setup as described above, but with two different types of infected individuals: acute (within their first 3 months of HIV infection) and chronic individuals. This yields separate transmission rates λ_a and λ_c for transmissions caused by acute or chronic individuals, respectively, for each of the three intervals. We assume that $\delta_a = \delta_c$. All priors and interval lengths were set to the priors above, and the rate of becoming chronic was set so that individuals remained acutely infected for 3 months on average.The proportion of transmissions (criteria v & vi) that originated from acute individuals is $\%Early = (f_a * \lambda_a)/ (f_a * \lambda_a + f_c * \lambda_c)$, where f_a and f_c are the expected fractions of individuals in the acute and chronic state, respectively. Criteria iv was assessed if the mean of the posterior distribution for P_a was below 10%, between 10-30%, or above 30%".

Remarks

A drawback of our approach is that we cannot directly infer incidence / prevalence, but only the epidemiological parameters (removal and effective reproductive number). From the parameters of the birth-death skyline model (i.e. R and δ) we work out (in a slightly ad hoc way, as described above) the incidence values. Our model could be used to directly infer incidences using particle filtering approaches which remains to be investigated in future work.

References

[1] Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C. H., Xie, D., ... & Drummond, A. J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS computational biology*, 10(4), e1003537.

[2] Stadler, T., Kühnert, D., Bonhoeffer, S., & Drummond, A. J. (2013). Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences*, 110(1), 228-233.

[3] Gavryushkina, A., Welch, D., Stadler, T., & Drummond, A. J. (2014). Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS computational biology*, *10*(12), e1003919.

[4] Stamatakis, A. (2014). RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics*, 1;30(9):1312-3.

[5] T.-H. To, M. Jung, S. Lycett, O. Gascuel. Fast dating using least-squares criteria and algorithms. *Submitted*.

[6] Kühnert D, Stadler T, Vaughan TG, Drummand AJ. Phylodynamics with migration: A computational framework to quantify population structure from genomic data. *Under review*.