PANGEA-HIV 2: Phylogenetics And Networks for Generalised Epidemics in Africa

Lucie Abeler-Dörnera, Mary K. Grabowskib, Andrew Rambautc, Deenan Pillayd,e, Christophe Frasera, on behalf of the PANGEA consortium

Purpose of review
The HIV epidemic in sub-Saharan Africa is far from being under control and the ambitious UNAIDS targets are unlikely to be met by 2020 as declines in per-capita incidence being largely offset by demographic trends. There is an increasing number of proven and specific HIV prevention tools, but little consensus on how best to deploy them.

Recent findings
Traditionally, phylogenetics has been used in HIV research to reconstruct the history of the epidemic and date zoonotic infections, whereas more recent publications focus on HIV diversity and drug resistance. However, it is also the most powerful method of source attribution available for the study of HIV transmission. The PANGEA (Phylogenetics And Networks for Generalized Epidemics in Africa) consortium has generated over 18 000 NGS HIV sequences from five countries in sub-Saharan Africa. Using phylogenetic methods, we will identify characteristics of individuals or groups, which are most likely to be at risk of infection or at risk of infecting others.

Summary
Combining phylogenetics, phylodynamics and epidemiology will allow PANGEA to highlight where prevention efforts should be focussed to reduce the HIV epidemic most effectively. To maximise the public health benefit of the data, PANGEA offers accreditation to external researchers, allowing them to access the data and join the consortium. We also welcome submissions of other HIV sequences from sub-Saharan Africa to the database.

Keywords
HIV, Phylogenetics And Networks for Generalized Epidemics in Africa, phylogenetics, sub-Saharan Africa, transmission dynamics

INTRODUCTION
The current review will give a brief overview of how the PANGEA consortium (Phylogenetics And Networks for Generalized Epidemics in Africa) will contribute to understanding the HIV epidemics in sub-Saharan Africa.

In recent years, most phylogenetic studies of HIV sequences sampled in sub-Saharan Africa have focussed on drug resistance mutations and their transmission [1–4], or classification of subtypes and evolutionary questions of HIV diversity [5–8]. However, sequence data is increasingly being used to construct transmission networks to inform prevention efforts [9], and to inform epidemiological studies [10*,11*]. Studies which combine phylogenetics and epidemiology to characterize sources of transmission in key populations and to detect sources of transmission and outbreaks have previously been conducted in other regions [12,13,14*,15*,16**]. However, such studies remain rare in sub-Saharan Africa as sequences are not available in as large numbers as in North American or European settings [17].

*Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, UK, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Rakai Health Sciences Program, Baltimore, USA, Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories, Edinburgh, UK, *Africa Health Research Institute, KwaZulu-Natal, South Africa and *Division of Infection and Immunity, University College London, London, UK

Correspondence to Mary K. Grabowski, PhD, ScM, Johns Hopkins University, Baltimore, USA. E-mail: lucie.abeler-dorner@bdil.ox.ac.uk; mgrabows@jhu.edu

DOI:10.1097/COH.0000000000000542
To date, PANGEA has generated over 18,000 NGS HIV sequences from five countries in sub-Saharan Africa, from diverse settings including cohorts of the population-based cohorts from surveillance sites (Rakai Community Cohort Study) [18], the Mochudi Prevention Project [19,20], the MRC/UVRI Uganda population-based cohorts, and fisherfolk cohorts [21–23] an MRC/UVRI Uganda cohort of female sex-workers [24], a cohort of HIV-1 drug-resistant individuals from northern KwaZulu-Natal in South Africa (Africa Health Research Institute Resistance Cohort) [25], participants from the Partners in Prevention HSV/HIV Transmission Study [26], and participants from the TasP trial [27]. Data from the HPTN 071-2 (PopART) trial Phylogenetics ancillary study [28] will become available in spring 2020. Below we will outline some of the key questions that PANGEA is planning to address with these data.

**AIMS OF THE PHYLOGENETICS AND NETWORKS FOR GENERALIZED EPIDEMICS IN AFRICA CONSORTIUM**

The overarching goal of the consortium is to use phylogenetic methods to identify the characteristics of individuals and groups that make them more likely to be at higher risk of infection (sinks), at higher risk of infecting others (sources), or both (hubs), and to translate these results into information that is directly actionable in HIV prevention [29]. Potentially relevant characteristics include age, sex, geography, occupation, cultural preferences and norms, migrational behaviour, riskiness of sexual behaviour, and use of preexposure prophylaxis (PreP). Figure 1 shows a schematic overview of sources, sinks, and hubs.

**ROLE OF PREVALENCE ‘HOTSPOTS’**

The HIV epidemic in sub-Saharan Africa is markedly heterogeneous with foci of high HIV incidence and prevalence communities (i.e. hotspots) among larger, relatively low-risk general populations [30]. In East Africa, hotspots include Lake Victoria fishing communities, which contain disproportionate numbers of key populations at high risk of infection, such as sex workers [21,31]. In Southern Africa, similar high prevalence foci have been observed along major trading routes and close to mines [32,33]. Using viral phylogenetics, we are measuring

---

**FIGURE 1.** Source–sink model. Different groups in a population may be at high risk of infecting others (source), at high risk of getting infected (sink), or both (hub). Groups can be characterized, for example, by age, sex, geography, occupation, cultural preferences and norms, migrational behaviour, or riskiness of sexual behaviour.
viral flow between high-prevalence and low-prevalence communities to quantify the attributable fraction of cases arising from hotspots. Bbosa et al. [34], for example, discovered that the HIV epidemic in fishing communities around Lake Victoria acts as a sink in comparison to inland communities, contrary to what was previously thought. Ratmann et al. [35] demonstrated how source-sink analyses including resolving the direction of transmission can be carried out using NGS data.

MOBILE POPULATIONS AND THEIR ROLE IN SUSTAINING LOCAL EPIDEMICS

Migration and mobility are important components in understanding the spread of HIV in many African communities [36,37]. There is growing evidence that mobile population are at higher risk for HIV and are less likely to be tested and linked to HIV care and treatment [32,37,38]. Consequently, these populations most likely have a key role in ongoing epidemic dynamics. For example, country-wide HIV incidence rates in Botswana have not declined despite achieving 90-90-90, which may be because Botswanan policy prohibits provision of HIV treatment to migrant populations [39]. Mobility is also a key factor in the interpretation of community randomized trials of HIV prevention [40]. If HIV transmissions are common between control and intervention arms, the impact of Universal Test and Treat (UTT) inventions will likely be underestimated. With phylogenetic approaches, these events can be quantified and adjusted for when analysing treatment effects [41]. Using phylogenetics, we are characterizing the role of mobile groups and non-mobile groups to ongoing HIV spread in the geographic regions represented in the PANGEA-HIV consortium, including Botswana and the areas of the two included UTT trials – HPTN 071 (PopART) and ANRS 12249 (TasP). Furthermore, we will analyse the geographical movement of the virus to quantify movement between the cohorts and more distant geographical locations in sub-Saharan Africa.

CHARACTERISTICS OF TRANSMITTERS AND TRANSMISSIONS

Another factor that can sustain local epidemics, as well as posing a problem for clinical trials and prevention efforts in general, is the heterogeneity of transmitters. Achieving 90-90-90 and having 72% of individuals virally suppressed might not reduce transmission by 72%, if individuals at high risk of infecting others are more likely to be missed. Targeted HIV prevention, therefore, requires an understanding of the risk factors linked to acquiring HIV and those linked to transmitting HIV. The most powerful tool of source attribution currently available to study HIV transmission is the construction of phylogenetic transmission networks [15*,42,43,44**,45].

TRANSMISSION NETWORKS

A viral phylogeny reveals who is close to whom in the transmission network [46]. With additional epidemiological information [15*] or modelling [42] one can identify which of these close individuals are likely sources of new infections. Following theoretical work on incorporating within-host diversity into phylogenetic inference [47*], we have developed a tool that can be used for source attribution in HIV NGS data [48*]. Once sources have been identified with quantified uncertainty, epidemiological questions about the characteristics of transmitters can be addressed. For example, transmission networks will enable us to detect individuals that are at high risk of infecting others and quantify their contribution to transmission. More generally, PANGEA aims to identify demographic, clinical, and virological correlates of being a transmitter compared with the background HIV-infected population. Transmission networks combined with mathematical modelling will also allow us to estimate the proportion of transmissions that arise from individuals in early or acute infection [49]. It has long been suspected that the majority of HIV transmissions derive from undiagnosed and untreated individuals in acute infection [50]. There is a worry that acutely infected individuals in hard-to-reach groups will continue to drive the epidemic even as increases in ART coverage lead to a sharp overall fall in incidence. We will construct matrices of who infects whom, stratified, for example, by age and sex. This may reveal motifs of transmission, which have been hypothesised as important drivers of the epidemic [11*,51*]. We will document transmissions of both drug-sensitive and drug-resistant viruses, and thus quantify important unknown parameters that influence long-term projections of the risk of drug resistance in generalized epidemics.

EMERGENCE AND TRANSMISSION OF DRUG RESISTANCE

The rapidly expanding use of ART, both for treatment and prevention, increases the risk of HIV drug resistance. Indeed, prevalence of transmitted resistance is increasing throughout the generalized epidemics in Africa [52]. This supports modelling work suggesting that 15% of new infections could be associated with drug resistance mutations by 2020 [53]. The source of resistance transmission can
either be from those with virological failure on treatment, or from those themselves untreated but having recently been infected with drug-resistant virus. The implications of this are profound. Should evidence emerge of spread of transmitted drug resistance, there is a more urgent need to change first line therapy. With dolutegravir soon to be made available, our findings will help to determine how individual countries prioritize treatment guidelines. Phylogenetic approaches can help to distinguish if a patient was likely infected with a drug-resistant virus or acquired drug-resistance mutations while on treatment. NGS data also yields information on whether different drug resistance mutations are found in different viruses within the same patient or if a multiresistant lineage is emerging. Combining both approaches with information on antiretroviral treatment will allow us to better understand how resistance spreads through a population.

COMBINING PHYLOGENETICS WITH MATHEMATICAL MODELLING OF HIV PREVENTION

Mathematical models of HIV transmission are key tools for evidence synthesis, impact assessment, prioritization of interventions, and prediction. Yet, despite many developments in methods, statistics, and computation, many uncertainties remain in key parameters that affect long-term projections \[54\]. Uncertainties include the infectiousness of acute infection, estimates of mixing matrices by age and risk group, estimates of between-individual heterogeneity in transmission rates, the existence, size, magnitude and contribution of hidden high-risk subpopulations, and the degree and rate of spatial mixing. These parameters are, at least in principle, in reach of being estimated using phylogenetic analyses. We will use parameters estimated from phylogenetic analysis to inform our mathematical models of the epidemic.

METHODS DEVELOPMENT

Some of the planned analyses will require the development of new methods. For example, PANGEA samples are not representative samples of the population. We will develop methods to adjust for sampling biases based on local knowledge, using Hidden Markov models, and repeat down-sampling. Some of our populations live in areas with a high share of A/D recombinant viruses. We are developing practical approaches to accommodating recombination in phylogenetic analyses. As part of the first phase of the PANGEA project, we held an open-community exercise to evaluate methods of inferring changes in incidence \[55\]. Genetic data were simulated under a range of scenarios at the regional and local community level and research groups were invited to apply their methods blinded to the true dynamics. The results of this exercise will be used to guide analysis of the actual PANGEA sequence data but common to all the approaches was the construction of phylogenetic trees from the genome sequence data. State-of-the-art phylogenetics methods are amongst the most computationally intense used in the study of infectious diseases today. We will, therefore, further develop the BEAST software packages \[56,57\] to deal with the challenge of large numbers of sequences and resulting phylogenetic trees. We will use web-based interactive graphics like NextStrain \[58\] to visualize the resulting complex phylogenies and integrate them with geographical maps and epidemiological data.

ETHICS OF PHYLOGENETICS

In our experience, consent rates for phylogenetics work are very high among study participants. However, transmission data needs to be handled very carefully to avoid the possibility that individual pairs can be identified, especially as some sexual practices are illegal across African countries. In light of these concerns, the PANGEA consortium convened a think tank meeting in 2017, including ethicists, legal experts, and social scientists, to propose guidelines for phylogenetic analysis \[59\]. The guidelines highlight the importance of appropriate communication of phylogenetic results, and ensuring public health benefit is balanced against risks to individuals and communities.

HOW TO GET INVOLVED

There are different ways in which external researchers can be involved with PANGEA (Fig. 2). Details are available on the consortium website (www.pangea-hiv.org).

1. **Sequence your samples with PANGEA:** We offer subsidized state-of-the-art NGS sequencing \[60\] to collaborators who are willing to contribute their sequence data and associated epidemiological data to the PANGEA database 12 months after they have received the data.

2. **Contribute sequence data:** If your project is at a stage where you are ready to reach out to a wider range of collaborators or if you are looking for a secure storing place for your data, consider contributing your sequence data to the PANGEA database.
(3) Join the community of PANGEA researchers: The aim of PANGEA is to make the best possible use of the data for public health and scientific discovery, and give credit to the people who generated the data. PANGEA, therefore, operates a data sharing policy that is as open as possible given the sensitivity of the data. External researchers can request access to the data via a concept sheet proposal. After 6 months and two successful updates on progress to the consortium, researchers named on the concept sheet proposal can apply to become an accredited PANGEA researcher. Accredited PANGEA researchers have access to all sequence data and less sensitive metadata. They are required to agree to a code of conduct and update the consortium regularly on their research. Collaborations with researchers from the institutes that generated the data are encouraged.

**PHYLOGENETICS AND NETWORKS FOR GENERALIZED EPIDEMICS IN AFRICA DATA**

Currently the PANGEA database holds over 18,000 NGS sequence files from sub-Saharan Africa (Table 1) with basic epidemiological metadata associated with them. Details are available on the consortium website (www.pangea-hiv.org). For some cohorts, in-depth metadata is available; contact us for more information. Currently we hold 6766 sequences from Uganda, 1749 from South Africa, 2530 from Botswana, and 29 from Kenya. Sequencing from Kenya is in progress and further sequencing is planned from Uganda and South Africa. Around 7500 sequences from Zambia will become available in early 2020. Overall, 65% of the sequences span the full genome or nearly the full genome (>7000 base pairs), 18% of sequences are partial genomes with at least 3000 base pairs and 17% of sequences contain at least 1000 base pairs. Using an improved
protocol [60], we currently obtain full genomes from close to 90% of high-quality samples, with many of the less-complete sequences originating from participants who are likely already virally suppressed.

CONCLUSION
Studies of transmission precede the use of phylogenetic methods in epidemiology [61–67]. However, phylogenetics can be in many ways more informative and/or more feasible than classical epidemiology or longitudinal studies of serodiscordant couples. Viral phylogenetic data cannot only complement epidemiological data and fill the gaps, it can also provide clinical information, for example, drug resistance information, more cheaply. Viral data can be compared more easily across studies than questionnaires and phylogenetics is much more scalable than partner studies or contact tracing. Crucially, phylogenetics also provides information on transmissions on a different level. Classical epidemiology will tell us that circumcision modifies the transmission rate per sexual contact, but phylogenetics will tell us in addition that uncircumcised individuals are an epidemic driver in a certain population. Phylogenetics can, therefore, help determine how the HIV virus moves through a population and can be translated into directly actionable information for public health, for example, suggesting, which groups within a population need more support and should be prioritized for interventions.

Having generated over 18,000 full or partial NGS genomes from five countries in Eastern and Southern Africa, PANGEA will focus on analysing source–sink dynamics, addressing the impact of mobility and migration, identifying and interpreting patterns of drug resistance, and describing the wider phylodynamic context of HIV spread. Results will be used to guide recommendations for HIV treatment and prevention policy in sub-Saharan Africa. PANGEA is committed to maximize the public health benefit of the data and welcomes project proposals and data contributions from researchers who share our aims.

Acknowledgements
We thank all members of the PANGEA consortium for making this project possible, Dr William Probert, Dr Chris Wymant, and Dr Tanya Golubchik, and the members of the PANGEA 2 Steering Committee for critical reading of the manuscript, and Larisa Gavrila for help with the references.

PANGEA is funded by the Bill & Melinda Gates Foundation (consecutive grants OPP1084362 and OPP1175094).

Members of the PANGEA consortium are listed below: Helen Ayles, Clinical Research Department, Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, UK, and Zambart, University of Zambia, Lusaka, Zambia; David Bonsall, Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, UK; Rory Bowden, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK; Vincent Calvez, Institut Camille Jordan, Université de Lyon, Université Claude Bernard, Université Jean Monnet, France; William Probert, Institute of Infection and Global Health, University of Liverpool, Liverpool, UK; Chris Wymant, Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, UK; Tanya Golubchik, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK; and Jon Abeler, Norwegian Institute of Public Health, Oslo, Norway.

Table 1. HIV sequences generated by the Phylogenetics And Networks for Generalized Epidemics in Africa consortium as of December 2018

<table>
<thead>
<tr>
<th>Country</th>
<th>Research centre</th>
<th>Sampling period</th>
<th>Partial genomes &lt;3000 nt</th>
<th>Partial genomes 3000–7000 nt</th>
<th>Near full genomes &gt;7000 nt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Botswana</td>
<td>Botswana Combination Prevention Project, Partners in Prevention</td>
<td>2008–2013</td>
<td>222</td>
<td>452</td>
<td>1856</td>
</tr>
<tr>
<td></td>
<td>Partners in Prevention</td>
<td>2005–2008</td>
<td>4</td>
<td>0</td>
<td>34</td>
</tr>
<tr>
<td>Kenya</td>
<td>Partners in Prevention</td>
<td>2006–2013</td>
<td>61</td>
<td>31</td>
<td>399</td>
</tr>
<tr>
<td>South Africa</td>
<td>AHRI</td>
<td>2012–2013</td>
<td>87</td>
<td>427</td>
<td>1235</td>
</tr>
<tr>
<td></td>
<td>Partners in Prevention</td>
<td>2004–2010</td>
<td>6</td>
<td>4</td>
<td>128</td>
</tr>
<tr>
<td>Tanzania</td>
<td>Partners in Prevention</td>
<td>2006–2008</td>
<td>1</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>Uganda</td>
<td>MRC Uganda</td>
<td>2009–2013</td>
<td>314</td>
<td>601</td>
<td>1102</td>
</tr>
<tr>
<td></td>
<td>Partners in Prevention</td>
<td>2005–2013</td>
<td>88</td>
<td>45</td>
<td>702</td>
</tr>
<tr>
<td></td>
<td>Rakai Health Sciences Program</td>
<td>2012–2014</td>
<td>1832</td>
<td>1319</td>
<td>1299</td>
</tr>
<tr>
<td>Zambia†</td>
<td>HPTN 071 PopART Phylogenetics†</td>
<td>2016–2018</td>
<td>409</td>
<td>238</td>
<td>5196</td>
</tr>
</tbody>
</table>

†Sequences from HPTN 071 PopART Phylogenetics will become available to accredited researchers in spring 2020.
PANGEA-HIV 2 Abeler-Dörner et al.

REFERENCES AND RECOMMENDED READING

Papers of particular interest, published within the annual period of review, have been highlighted as:
- of special interest
- of outstanding interest

15. The authors inferred that, amongst MSM in Detroit, transmission occurring during acute and early infection accounted for 50–90% of transmission. This study demonstrated the power of structured coalescent methods.
- The authors quantify the drivers of the Dutch HIV epidemic and MSM by identifying ~600 probable transmission pairs, with directionality inferred using dates of seroconversion. The article inferred more than half of transmission (65–75%) came from undiagnosed men; that these men tested less often than average; and that transmission as mechanism that renews and so drives the epidemic. This study has been influential in the question it addresses. As the methods have been questioned [51*], the question requires further analysis.
20. The authors inferred that, amongst MSM in Detroit, transmission occurring during acute and early infection accounted for 40–50% of transmission. This study demonstrated the power of structured coalescent methods.
- The authors quantify the drivers of the Dutch HIV epidemic and MSM by identifying ~600 probable transmission pairs, with directionality inferred using dates of seroconversion. The article inferred more than half of transmission (65–75%) came from undiagnosed men; that these men tested less often than average; and that transmission as mechanism that renews and so drives the epidemic. This study has been influential in the question it addresses. As the methods have been questioned [51*], the question requires further analysis.
25. The authors inferred that, amongst MSM in Detroit, transmission occurring during acute and early infection accounted for 40–50% of transmission. This study demonstrated the power of structured coalescent methods.
- The authors quantify the drivers of the Dutch HIV epidemic and MSM by identifying ~600 probable transmission pairs, with directionality inferred using dates of seroconversion. The article inferred more than half of transmission (65–75%) came from undiagnosed men; that these men tested less often than average; and that transmission as mechanism that renews and so drives the epidemic. This study has been influential in the question it addresses. As the methods have been questioned [51*], the question requires further analysis.
30. The authors inferred that, amongst MSM in Detroit, transmission occurring during acute and early infection accounted for 40–50% of transmission. This study demonstrated the power of structured coalescent methods.
- The authors quantify the drivers of the Dutch HIV epidemic and MSM by identifying ~600 probable transmission pairs, with directionality inferred using dates of seroconversion. The article inferred more than half of transmission (65–75%) came from undiagnosed men; that these men tested less often than average; and that transmission as mechanism that renews and so drives the epidemic. This study has been influential in the question it addresses. As the methods have been questioned [51*], the question requires further analysis.
35. The authors inferred that, amongst MSM in Detroit, transmission occurring during acute and early infection accounted for 40–50% of transmission. This study demonstrated the power of structured coalescent methods.
- The authors quantify the drivers of the Dutch HIV epidemic and MSM by identifying ~600 probable transmission pairs, with directionality inferred using dates of seroconversion. The article inferred more than half of transmission (65–75%) came from undiagnosed men; that these men tested less often than average; and that transmission as mechanism that renews and so drives the epidemic. This study has been influential in the question it addresses. As the methods have been questioned [51*], the question requires further analysis.
Phylogenetics in HIV transmission

35. The authors analyse the dynamics of HIV transmission in very high prevalence and marginalized fishing communities on Lake Victoria, Uganda. Contrary to expectations, the study identifies the fishing communities as sinks rather than sources of the HIV infection.