Phylogenetic studies of transmission dynamics in generalized HIV epidemics: An essential tool where the burden is greatest?
Ann M. Dennis*<sup>a</sup>, Joshua T. Herbeck*<sup>b</sup>, Andrew Leigh Brown<sup>c</sup>, Paul Kellam<sup>d,e</sup>, Tulio de Oliveira<sup>f</sup>, Deenan Pillay<sup>g</sup>, Christophe Fraser<sup>d</sup>, Myron S. Cohen<sup>a</sup>.

<sup>a</sup>Division of Infectious Diseases, University of North Carolina at Chapel Hill, Chapel Hill, NC
<sup>b</sup>Department of Microbiology, University of Washington, Seattle, WA
<sup>c</sup>Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK
<sup>d</sup>Wellcome Trust Sanger Institute, Cambridge, UK
<sup>e</sup>Division of Infection and Immunity, University College London, London, UK
<sup>f</sup>Wellcome Trust-Africa Centre for Health and Population Studies, University of Kwazula-Natal, ZA
<sup>g</sup>Department of Infectious Disease Epidemiology, Imperial College London, London, UK

*Authors contributed equally

Correspondence:
Ann M. Dennis, MD
130 Mason Farm Drive, Suite 2115, CB 7030
Chapel Hill, NC 27999-7030
adennis@med.unc.edu
T 919.966.2536  F 919.966.6714

Running Head: Phylogenetics of generalized HIV epidemics

Word count: 5570; Tables: 3; Figures: 2

Conflicts of Interest and Source of Funding: We declare that we have no conflicts of interest. AD reports receiving support from the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant KL2TR000084. JH reports receiving support from the University of Washington Center for AIDS Research, an NIH funded program under award number P30AI027757 which is supported by the following NIH Institutes and Centers (NIAID, NCI, NIMH, NIDA, NICHD, NHLBI, NIA, NIGMS, NIDDK).
ABSTRACT

Efficient and effective HIV prevention measures for generalized epidemics in sub-Saharan Africa have not yet been validated at the population-level. Design and impact evaluation of such measures requires fine-scale understanding of local HIV transmission dynamics. The novel tools of HIV phylogenetics and molecular epidemiology may elucidate these transmission dynamics. Such methods have been incorporated into studies of concentrated HIV epidemics to identify proximate and determinant traits associated with ongoing transmission. However, applying similar phylogenetic analyses to generalized epidemics, including the design and evaluation of prevention trials, presents additional challenges. Here we review the scope of these methods and present examples of their use in concentrated epidemics in the context of prevention. Next, we describe the current uses for phylogenetics in generalized epidemics, and discuss their promise for elucidating transmission patterns and informing prevention trials. Finally, we review logistic and technical challenges inherent to large-scale molecular epidemiological studies of generalized epidemics, and suggest potential solutions.

Key words: HIV-1, molecular epidemiology, phylogenetic, transmission networks, Sub-Saharan Africa
INTRODUCTION

Despite advances in HIV prevention over the past 30 years, an estimated 2.5 million persons were newly infected in 2011, bringing the number of people living with HIV worldwide to 34 million.\(^1\) The HIV burden continues to be greatest in sub-Saharan Africa; this region accounts for 75% of all HIV infections and the highest adult prevalence at nearly 5% overall,\(^1\) though this average mask extremes reported in subpopulations, some exceeding 50% infected.\(^2\) The use of antiretroviral treatment (ART) to reduce viral loads and associated transmission risk among sero-discordant couples (Treatment-as-Prevention, TasP) has garnered excitement as a means to curb the spread of the virus.\(^3\) A recent population-based cohort study conducted in a high prevalence region in KwaZulu-Natal found that the risk for HIV acquisition was lowest in areas with the highest ART coverage—providing ecological evidence for real-world effectiveness of TasP.\(^4\) As ART coverage has increased in Africa in the last decade, more HIV-infected individuals have been treated, and life-expectancy of infected individuals has increased.\(^5\) Meanwhile, new transmissions continue, and therefore overall prevalence of HIV can be expected to increase.\(^6\)

To sustain an ongoing ART scale-up, and the potential widespread implementation of TasP in the future, expanded financial and public health resources will be required.\(^7,8\) However, the most biologically effective and financially efficient way to implement and evaluate prevention measures at the population level is unclear.\(^9,10\) Comprehensive knowledge about local epidemics will be required for successful prevention campaigns, including basic data about population demographics, transmission risk groups and viral subtypes, and complex estimates about transmission dynamics, social and sexual mixing networks, and patterns of geographic spread. Importantly, prospective information about success or failure of interventions is essential.

The tools of HIV phylogenetics and molecular epidemiology can be used to understand local transmission dynamics and assist in the design and evaluation of prevention trials. Since early in the epidemic these approaches have been employed to track HIV origin and geographic spread\(^11-14\) and in forensic studies evaluating small transmission chains.\(^15-19\) New developments, primarily led by the increased availability of viral
sequences over the past 20 years, have allowed fine-scale transmission dynamics at the community, regional, and country level to be uncovered.\textsuperscript{20} The opportunity for such approaches has been facilitated by: 1) the increasing availability of large HIV sequence databases, driven by the routine provision of gene sequence based drug resistance testing, with sequences linked to epidemiological data (temporal, clinical, demographic, behavioral or geographic); 2) rapid advances in high-throughput sequencing technology and decreases in sequencing costs; and 3) theoretical and methodological advances in studies of viral transmission, using phylogenetics or genetic network analysis together with linking epidemiological and population genetic models. These advances provide a framework to identify individual traits or stages of infection that are associated with high relative infectiousness; the results of such studies can answer questions not easily resolved with standard epidemiological approaches. In effect, molecular epidemiology tools can help identify traits associated with ongoing HIV transmission ("who" is transmitting the virus?), rather than behaviors or demographic characteristics associated with high rates of infection (who is currently infected?). These analyses are increasing deployed for concentrated epidemics where HIV sequences are routinely available. However, the essential next step is to apply these tools to HIV prevention in generalized epidemics and resource-limited settings where the impact would be greatest. Further, the development of such approaches for HIV can provide a model for new approaches to the transmission dynamics of other pathogens.

**SCOPE OF PHYLOGENETIC ANALYSIS TO INFORM MOLECULAR EPIDEMIOLOGY**

Molecular epidemiology is the use of genetic data to inform disease etiology and distribution. In HIV, the term can encompass disparate approaches. For this review, we divide the use of genetic analysis to inform HIV epidemiology into three general categories: Molecular Epidemiology, Phylodynamics and Phylogeography (Table 1). While there is generally overlap across these categories within studies, the main questions associated with each category are distinct. Molecular Epidemiology allows understanding of the risk factors for HIV transmission and epidemic spread. Phylodynamics reconstructs epidemic history, and quantifies epidemic growth or decline, using viral genealogies and explicit population genetic models. Phylogeography describes the distribution of subtype diversity, estimates the impact of human migration on viral spread, and places historical and risk factor data into geographic context, in order to identify hubs of transmission.
As for any approach to elucidate HIV transmission patterns, gene sequences can only be generated for those already diagnosed and sampled, and therefore the contribution to transmission from those unsampled, undiagnosed individuals must be considered. Nevertheless, with a well-designed population sampling and sequencing strategy, with linkage to some key demographic, epidemiological and clinical data, questions from all three categories can be addressed for any given population (using the same sequence dataset).

**PHYLOGENETIC STUDIES OF TRANSMISSION IN CONCENTRATED EPIDEMICS**

Molecular epidemiological tools have been used to address a diverse array of research questions over the last two decades, predominantly in settings with concentrated epidemics. These approaches have advanced to describe transmission dynamics at the local, regional, and national level through the *post-hoc* utilization of large sequence datasets linked to epidemiological data. These sequence datasets are largely a consequence of routine antiretroviral drug resistance testing that accompanies initial HIV diagnosis, and contains partial HIV *pol* sequences (typically full protease and partial reverse transcriptase). The *pol* region has been shown to have sufficient variability to allow for phylogenetic reconstruction to the same extent as the more variable *gag* and *env* regions. Studies focused on transmission networks in concentrated epidemics have most commonly been conducted in Europe or North America, where large *pol* sequence datasets linked with epidemiological data exist (Table 2).

**Basic methodological approach**

The basic method to identify HIV transmission dynamics (traits associated with ongoing transmission) is straightforward (Figure 1): 1) HIV gene sequences are used to reconstruct phylogenetic trees or viral transmission networks; 2) transmission clusters are identified from the networks or phylogenies using *ad-hoc* thresholds for inclusion (e.g. minimum pairwise genetic distance between viral gene sequences, or by statistical robustness of the node that defines a phylogenetic clade); and 3) individual traits are evaluated for the strength of their association with cluster membership and used to infer underlying determinants of transmission. Cluster inclusion thresholds are *ad-hoc* in the sense that there is no widely accepted
threshold definition or determining convention. Variation in threshold definitions and approaches can strongly impact the size and number of inferred clusters in a sequence data set, yet many studies do perform sensitivity analyses to quantitatively assess the impact of this variation on cluster identification in their respective data sets. A recent study of transmission dynamics in Brighton, UK, took this basic methodological approach one step further: clinical data were used to determine the most likely transmitter within transmission pairs or clusters, which allowed for likely onward transmission events (and not simply cluster membership based on pairwise genetic distance) to be evaluated for association with individual traits.  

**What is the contribution of acute and early infection to ongoing transmissions?**

Acute HIV infection is generally defined as the time after HIV acquisition but before seroconversion. Early HIV infection represents the first few months of infection, sometimes using a staging system reported by Feibig et al. Most recent results suggest HIV transmission is increased during acute and early HIV infection, for a period of uncertain duration that may range from weeks to months. Multiple phylogenetic studies have focused on acute and early (primary) HIV infection to assess the contribution of acute infection transmissions to overall HIV incidence in a population. The majority of these studies analyzed only sequences from patients diagnosed during acute or early HIV infection, and found variable rates of cluster formation, ranging from 13% to greater than 50% of the study population. These results suggest that acute and early infections are responsible for a disproportionate number of onward transmissions, relative to chronic infections. This might be expected, as: 1) individuals with newly acquired infections remain sexually active; 2) such individuals are unlikely to know their HIV status; 3) HIV viral load is exceptionally high for weeks to months after infection; and 4) the transmitted/founder viruses that establish successful infections may have transmission advantages. However, the proportion of transmission events attributed to acute or primary infection is dependent on definitions used for these stages of infection; more rigorous definitions of early-stage infection, and focusing on individuals with estimated infection dates or known dates of seroconversion, can help to resolve this issue. It is essential to understand the contribution of early-stage infection to onward transmission, as a high frequency may compromise TasP strategies.

**Other factors linked to transmissions**
Trends or patterns associated with onward transmission have been described in greater detail by incorporating sequences from chronically infected patients thereby increasing the study population sampling density. Using molecular epidemiology tools, transmission network characteristics have been described by various factors such as race/ancestry and ethnicity, transmission risk group, HIV subtype, transmitted drug resistance mutations, and transmission cluster growth. A detailed longitudinal study of MSM sought to find the most likely transmitter for new infections and assess risk factors at the time of transmission. Even with 75% of all diagnosed HIV-infected individuals from the local clinic providing pol sequences, the most likely transmitter could be identified in only 25% of those recently infected. This implies that even high sampling fractions of local epidemics may not completely reveal underlying transmission linkage due to, for instance, extra-community or undiagnosed infections as major sources of new transmissions.

**Large scale analyses can delineate sub-epidemics**

Large scale analyses of pol sequences have been used to assess the potential influence of viral subtype and regional scale on HIV transmission dynamics. The United Kingdom HIV Drug Resistance Database (UKRDB) (containing >85,000 pol sequences, http://www.hivrdb.org.uk/) has provided data for several transmission studies. Using phylodynamics and a relaxed molecular clock approach, transmission dynamics were reconstructed among MSM and heterosexuals. Among MSM, six large transmission clusters were identified, representing separate introductions of subtype B into the U.K. in the 1980s. Additional analysis within the large MSM clusters (reconstructed with ~2,000 individual sequences) indicated that 25% of transmissions likely occurred within six months of infection, with most clusters arising over periods of 3-4 years—an episodic epidemic with multiple clusters of transmission. However, in heterosexuals, where the epidemic is dominated by non-B subtypes, much slower transmission dynamics were found. This demonstrates that phylogenetic approaches actually reveal the differing dynamics between different risk groups.

The Swiss HIV Cohort includes large repositories of pol sequences that have been used to assess changes in subtype B and non-B transmission clusters over time. Largely independent epidemics of MSM and heterosexuals/injection drug users were noted in the subtype B clusters, but heterosexuals alone did not
Over time, the contribution of injection drug use to the heterosexual epidemic notably decreased. The effect of migration was investigated by analyzing non-B subtypes sampled both in and outside Switzerland. Less than 25% of the non-B subtypes sampled in Switzerland were found in clusters with other Swiss sequences, suggesting that most non-B infections in the country could not be prevented through national prevention measures targeting individuals of only Swiss origin.

Transmission network analysis
The framework of social network analysis provides an alternative method to understand HIV transmission dynamics, which can be supplemental to the explicitly evolutionary approach of phylogenetic analysis. Transmission clusters are reconstructed similar to contact tracing, but based on genetic distance metrics. Transmission network parameters among MSM were estimated using phylodynamics of over 14,000 UKRDB sequences (representing ~60% of UK MSM). Using an inferred network distribution and associated parameter values, the HIV epidemic was characterized by preferential association, which predicted that the epidemic would persist even under conditions of poor overall transmission following a randomly distributed intervention. Cluster growth also differed across the MSM and heterosexual transmission risk groups. However, the relationship between phylogenetic clusters and sexual networks is not direct (they are not one and the same), and the use of phylogenies to understand the underlying transmission network is complex and requires further elucidation.

PHYLOGENETICS AND TRANSMISSION DYNAMICS IN GENERALIZED EPIDEMICS OF SUB-SAHARAN AFRICA: WHAT QUESTIONS CAN BE ADDRESSED?

HIV transmission dynamic studies in concentrated epidemics largely involve the post-hoc use of HIV drug resistance screening datasets whose coverage can reach a substantial proportion of the HIV-infected population. Large datasets of this type are rarely found in Africa—despite sub-Saharan Africa accounting for two-thirds of the global HIV infections, only 24% of sequences deposited in the LANL HIV Sequence Database (http://www.hiv.lanl.gov, which receives all HIV sequences deposited in GenBank) are derived from the region (Figure 2). Notably, the Southern African Treatment and Resistance Network (SATuRN) has a growing
database currently with >7,000 HIV sequences, albeit sampled from a very large HIV-infected population. Nonetheless, informative phylogenetic analyses of generalized epidemics will still require significant de novo sequencing effort. To enable detailed molecular epidemiological studies to inform prevention, clinical, demographic and behavioral data must be linked to each sequence. This represents a major logistical challenge, given that current sequence databases are generally based on opportunistic approaches to sequence acquisition. Several important questions surrounding HIV epidemiology and prevention can be addressed in the region with available molecular epidemiology tools, particularly when linked to traditional epidemiologic data.

**Historical pattern of growth or decline for a given HIV epidemic**

Molecular epidemiology and phylogeography can be used to understand historical patterns in epidemic origin, spread, and growth over time and space. The high genetic diversity of HIV-1 Group M has given rise to 9 genetically divergent subtypes (A-D, F-H and J-K), intersubtype recombinants and circulating recombinant forms (CRFs). Sub-Saharan Africa has the highest HIV genetic diversity, and is also characterized by distinct geographical subtype distributions that have remained relatively stable over the past decade. Surveillance of genetic diversity has historically supported tracking global epidemiology and public health strategies to slow further viral spread.

The integration of time-stamped sequence (those with known dates of sampling) data with phylogenetics, coalescent models and molecular clock models allows inferences to be made on the timing of epidemic origin and spread in Africa. These analyses can contribute to the design of intervention strategies through better understanding of epidemic growth potential, or factors contributing to historical geographic spread of subtypes or recombinants. Phylogeographic approaches showed that HIV-1C in Zimbabwe expanded through multiple introductions originating in southern Africa and localized exponential growth in the 1980s corresponding to demographic and political change. Travel accessibility and infrastructure is felt to be critical in epidemic spread, particularly the rapid growth of HIV-A and D into east Africa, and the expansion of HIV-1C into east Africa and Angola.
**Geographic source of local epidemics or outbreaks**

HIV prevalence in Africa is heterogeneously distributed within countries and often communities; epidemics can be overlapping sub-epidemics defined by geography, time, and a complex interplay of local epidemic drivers. This variation in HIV spread through populations requires that prevention efforts be tailored to characteristics of local epidemics.\(^1\)\(^,\)\(^75\) For example, the geographic clustering of HIV along roadways in KwaZulu-Natal indicated the need for more intensified interventions within these communities; this study also indicated that many infections were imported from outside local communities.\(^76\) High transport connectivity and mobile populations may explain the hyper-endemic outbreaks experienced in eastern and southern Africa.\(^77\) Understanding the degree of transmission that occurs from outside communities may have a substantial impact on the design and success of targeted prevention efforts such as TasP.

Communities with high transport connectivity may have local HIV epidemics supported by a significant proportion of transmissions from outside communities. Among 153 HIV *pol* sequences from patients in one community in rural coastal Kenya, multiple subtypes and significant recombination were documented.\(^78\) In a phylogeographic analysis, many of these sequences were related to different regions in Africa, suggesting multiple introductions into this community, likely reflecting its extensive transport links. In the Rakai district in Uganda, 14,595 individuals in 46 communities underwent extensive HIV surveillance including spatial and phylogenetic transmission linkage analysis.\(^79\) Of 189 HIV-incident cases, an estimated 39% of new cases were infected by household partners, and many new cases that were infected by an extra-household partner were from outside the community. The high degree of external HIV introductions into these communities suggests that the ability of test-and-treat strategies to reduce HIV transmissions may be difficult to measure in small populations unless external introductions are reduced, or unless geographic TasP coverage is sufficiently widespread that migration no longer becomes a relevant problem.

Results from phylogeographic studies must be interpreted in the context of the viral sequences analyzed. Studies based on convenience samples with limited temporal or spatial scales could lead to erroneous inferences about transmission rates between populations (e.g. between the studied community and the extra-community). While this potential limitation can be addressed to some extent with extensive sampling of the
community of interest, the use of simulation to validate inferences can be useful.\textsuperscript{80}

**Tracking the transmission and evolution of drug resistance mutations**

The decreases in morbidity and mortality following ART provision\textsuperscript{81,82} have led to widespread ART rollout over the past decade.\textsuperscript{83} With increased ART access in Africa, the prevalence of transmitted drug resistance (TDR) has duly increased.\textsuperscript{84,85} High levels of acquired drug resistance, often unrecognized when laboratory monitoring for virologic failure is unavailable, and the failure of ART to prevent secondary transmission, contribute to increasing TDR in the population.\textsuperscript{86} Simulations suggest that TDR could have a significant impact on mortality,\textsuperscript{87} thus highlighting the need for ongoing attention for tracking and preventing TDR. Further, with TasP programmes expanding, and WHO guidelines\textsuperscript{88} recommending treatment initiation at CD4 lymphocyte counts<500 cells/mL, then it is likely that an increasing proportion of new infections will be with resistant viruses, despite an overall drop in incidence.\textsuperscript{89}

Although TDR prevalence in sub-Saharan Africa is moderate at an estimated 5.7%,\textsuperscript{86} it is projected to have increased by 14-29% per year in southern and east Africa following ART rollout.\textsuperscript{84} Phylogenetic analyses have been employed in TDR prevalence studies primarily to characterize the extensive genetic diversity in specific regions\textsuperscript{90-94} and the dominance of HIV-1C in southern Africa.\textsuperscript{95-99} Few studies have evaluated transmission lineages of drug resistant strains through transmission cluster analysis to support or refute epidemiological linkages.\textsuperscript{90,96,100} Transmission clusters combined with participant life histories revealed a high degree of sexual partner mixing in Ugandan fishing communities and uncovered clusters sharing similar TDR mutations.\textsuperscript{94} Only one study incorporated antiretroviral history in the probable transmitting partners to further characterize linkages among individuals with TDR.\textsuperscript{100}

In contrast to resource-rich settings, drug resistance testing is rarely performed at entry to clinical care, thereby limiting the number of sequences available for phylogenetics. However, sequences will become increasingly available as drug resistance monitoring strategies continue to expand and through ongoing or pending TasP protocols. Phylogenetics could be used to evaluate trends in genetic diversity, further the understanding of
sexual networks or transmission clusters harboring resistant strains, track the evolution of drug resistance on a population-level, and to help assess the effect of TDR on transmissibility by subtype.

**Understanding transmission patterns to help design targeted prevention measures**

Targeting epidemic drivers, or core groups, for enhanced prevention may increase the effectiveness of an intervention, but requires detailed epidemiological understanding of viral transmission in the community. This is challenging because HIV transmission dynamics of generalized African epidemics are largely unknown. Molecular epidemiological approaches can help uncover local HIV epidemic drivers by contributing the links between overlapping sub-epidemics that are characterized by geography, time and social/sexual interaction.

Characterizing HIV subtypes and including linkage analysis within epidemiological studies can shed light on transmission patterns between high risk groups and the general population. Among MSM in Senegal, HIV phylogenetics revealed different subtype distributions compared to the general population. Most of these men reported sex with women, thus likely contributing to bridging between these groups that may modify the subtype distribution. In coastal Kenya, local subtypes among MSM predominated, including those frequent among female sex workers, indicating an epidemic of local origin and confirmation of observed behavioral links between MSM and the general population. By combining partnership histories and phylogenetic analysis among Ugandan female sex workers, partial sexual networks and multiple infections were observed, confirming high-risk networks. Among HIV concordant heterosexual couples in Senegal, most couples had phylogenetically linked sequences. When combined with interview data, the male partner was often the most likely index, implying concurrency associated transmission among stable partners.

In the design of prevention measures, these phylogenetic tools could be expanded to other questions; currently unknown is whether epidemic growth in local epidemics of sub-Saharan Africa is driven primarily by those with high viremia or consistent low-level transmission from chronically infected individuals in concurrent partnerships. Determining epidemic drivers allows for targeted prevention in a more cost effective manner. For example, Avahan, the India AIDS Initiative, focused prevention on groups at high risk for transmission and acquisition in India (sex workers, their clients, and injection drug users; identified without phylogenetic
analyses) leading to over 100,000 estimated HIV infections prevented in the general population between 2003 to 2008. The success of these types of approaches, however, depends on having an in-depth understanding of local epidemic drivers in the HIV epidemics of sub-Saharan Africa.

Assessing the impact of an intervention

Molecular epidemiology approaches also hold promise in evaluating HIV prevention interventions. At the individual level phylogenetics has been used to confirm or refute transmission linkages among seroconverters and their partners in heterosexual serodiscordant partnership trials. By determining genetic linkages between enrolled partners, the primary efficacy of the intervention can be better assessed (if transmission occurred between partners despite the intervention, or did the transmission arise from an outside partnership). The Partners in Prevention HSV/HIV Transmission study (PiP) assessed the efficacy of genital herpes suppression in reducing HIV transmission among serodiscordant couples in east and southern Africa. Nearly 27% of couples were found to be unlinked through phylogenetic analyses, showing that a substantial number of transmissions occurred through outside partnerships. In HPTN 052, pol and env sequences among index-partner pairs and controls were also evaluated with phylogenetic methods. Similar to PiP, 24% of the index cases were not linked to their partner. There was a strong association between linked transmission and the delayed ART initiation study arm: 28 of 29 linked transmission events were in the delayed arm. The association between early ART initiation and transmission reduction became stronger when only the linked events are included in the analysis; this emphasizes the importance of genetic linkage analysis to assess seroconversion events in prevention studies.

Several combination prevention trials that will assess the impact of TasP on a population level are either planned or are currently in progress in sub-Saharan Africa. Most of these trials have planned or are considering integration of molecular epidemiology analyses (Table 3).

At the population level, analysis of genetic data can potentially supplement standard approaches to evaluate the impact of an intervention. Comparative phylogenetic analyses of a baseline trial population and the population over the course of a trial can reveal the emergence or disappearance of clusters associated with
particular traits. This approach was used to assess the impact of targeted hepatitis B vaccination in The Netherlands, and showed that resulting decreases in HBV incidence were due largely to declines in intravenous drug or heterosexual (but not MSM) risk groups.\textsuperscript{112} However, a follow-up study with increased sample size (n=894, versus n=85) suggested that reduced HBV transmission was in fact due to reduced incidence in MSM, highlighting the importance of sample size and extended sampling periods for studies of this type.\textsuperscript{113} Alternatively, gene sequence data can be used to reconstruct HIV transmission networks rather than phylogenies, and cluster size distributions (CSD) can be compared over the course of a prevention trial or intervention.\textsuperscript{59,114,115} In theory, CSDs will be dominated by larger clusters in populations where epidemic drivers (individuals with relatively high infectiousness) persist. Changes in CSD can reflect an intervention’s impact on particular subgroups or the overall transmission patterns. These population level approaches may be most suitable for trials in which clear incidence outcomes are equivocal, or in which there are clear decreases in incidence but the underlying cause is unknown. While the methods can be applied to a trial with any targeting strategy, clinical and demographic data from sequenced individuals are required.

Phylodynamics is the use of pathogen sequences to reconstruct epidemic history using viral genealogies and explicit population genetic models.\textsuperscript{114,116,117} Despite great methodological potential and scientific interest, phylodynamics has to date been rarely used for impact evaluation. This is likely related to a lack of consensus about the interpretation of the estimated parameter \( N_e \), (nominally the effective population size, a quantity proportional to the number of infected individuals, and estimated by coalescent approaches within the product \( N_e \times \tau \), where \( \tau \) is the mean (viral) generation time). Estimates of both \( N_e \) and \( \tau \) can be strongly affected by epidemic stage, transmission dynamics, or population sampling,\textsuperscript{11,118,119} and come with large variances. There is poor resolution of population size changes in the recent past (~5 years).\textsuperscript{119} Additionally, simulation studies have shown that it may be difficult to disentangle the effects of changing incidence and changing transmission networks on phylodynamic parameters;\textsuperscript{119,120} information on transmission network structure might be required for accurate parameter estimation. On a positive note, there is a growing base of modeling approaches to understand the relationships between epidemic models, phylogenetics, and transmission networks, which could be utilized to better understand how transmission network structure affects phylogenetic trees, and to model outcomes of specific prevention trial designs.\textsuperscript{60,114,121-125} Of particular interest is the
incorporation of stochastic birth-death processes into phylodynamic estimation of epidemic parameters in lieu of standard coalescent models, allowing for more realistic assumptions about changes in epidemic size through time.\textsuperscript{126-130}

**CHALLENGES AND SOLUTIONS FOR IMPLEMENTING PHYLOGENETIC STUDIES IN GENERALIZED EPIDEMICS**

Phylogenetic analyses of HIV transmission and other epidemiological questions hold great promise to further our understanding of generalized epidemics and inform prevention efforts. However, we must consider how differences between concentrated and generalized epidemics will affect the design and implementation of such studies. Below we note several key challenges that must be met, followed by potential solutions.

*HIV transmission networks will be difficult to detect with limited population sampling*

As seen in phylogenetic studies of concentrated HIV epidemics, a large sampling fraction (e.g. >25% of infected individuals in a community) is needed to identify transmission pairs or clusters. This result is seen empirically\textsuperscript{24} and in simulations.\textsuperscript{121, 122} For phylogenetic studies in generalized HIV epidemics, especially in regions where prevalence can exceed 10%, a substantial number of individuals will need to be sampled and sequenced.

**Potential solution 1:** Population sampling that is biased toward sequencing of incident cases can both decrease the required sample fraction and increase the probability of identifying transmission linkages. This “targeted” sampling strategy contrasts with the opportunistic sampling strategy generally found in standard phylogenetic studies. The approach is suited for studies that seek to identify HIV transmissions, rather than reconstruct viral evolutionary history. As such, phylogenetic studies of transmission will be most informative and efficient when epidemiological questions are not simply overlaid onto *ad-hoc* phylogenies reconstructed from randomly sampled individuals in a population (the standard approach for phylogenetic studies in *e.g.* systematics, biogeography and phylodynamics).
Potential solution 2: Population sampling conducted over multiple time periods, with the initial sample completed prior to the intervention, increases the probability of identifying transmission pairs or clusters with phylogenetic analysis. Transmission studies in concentrated epidemics generally involve post-hoc use of sequence datasets which allows for retrospective analyses of epidemic history or transmission dynamics. The utility of HIV phylogenetics to inform prevention trials cannot be fully realized, however, based solely on retrospective analyses. The sampling frequency required to improve transmission cluster detection is unclear, and will likely vary according to local incidence rates.

HIV sequencing requirements may be extensive
As large sample fractions will be required for informative phylogenetic analyses of generalized epidemics, significant de novo sequencing effort will be necessary, even with “targeted” sampling of incident infections. This will require extensive technical capability and ability for large volumes of sequences to be generated relatively rapidly. Except for the SATuRN database, there are no standing HIV sequence databases in the region that can be readily used for molecular epidemiological studies, in contrast to resource-rich settings that have a larger proportion of sequences deposited compared to their epidemic size (Figure 2).

Potential solution: Developing a sequencing pipeline is necessary for large scale HIV phylogenetic and molecular epidemiology projects. Five features of this pipeline are required: 1) the pipeline must work on clinical samples; 2) it must scale across multiple diverse HIV genomes (different subtypes), based on a universal primer set, or alternative methods of genome enrichment that produce equivalent amplification rates across diverse samples; 3) it must scale from 100s to 1000s of samples with ~75% sequencing success rate across a wide range of genome copy number (viral loads of the individual samples); 4) it must produce accurate consensus sequences with no manual editing; 5) it must detect and accurately quantify minority variants; and 6) it must maximize the informative phylogenetic signal, by extending sequencing from one gene (typically the partial pol gene sequenced to around 1kbase length), to the whole viral genome (9.8kbase in length). A component of this solution may be the development in Africa of the capacity for high-throughput full genome HIV sequencing.
The development of a pan-African sequence database (analogous to the LANL HIV Sequence Database), will provide an important resource for future studies. The few phylogenetic studies of HIV transmission in Africa to date\textsuperscript{79,132} suggest that extra-community HIV transmission sources are common. The potential role of a large pan-African database would be to provide sets of African outgroup sequences, in order to identify extra-community sources and to clarify their impact vs. undiagnosed local sources. The database would also be useful for studying the spread of inter-subtype recombinants and for characterizing the diversity of regional epidemics in Africa, \textit{i.e.} for phylodynamics and phylogeography of the African HIV epidemic.

\textbf{Methodological challenges: Integrating molecular epidemiology into phylodynamics}

Molecular epidemiology and phylodynamics have both been areas of active methodological development, as evidenced by the articles referenced above. Nonetheless, current studies in molecular epidemiology that define risk factors for transmission do not make full use of the data available, and do not adequately account for the uncertainty and arbitrariness inherent in clustering.

\textbf{Potential solution:} A preferable approach would be to integrate the estimation of transmission risk factors (and other statistics of molecular epidemiology) directly into a phylodynamic inference framework, so that all of the data available could be used and arbitrary clustering would not be a prerequisite step. Some authors have begun this process,\textsuperscript{133} but further methodological development and validation is needed. Additionally, the development of consistent quantitative definitions of transmission clusters, \textit{e.g.} based on tree shape characteristics such as average branch lengths or nodal support, can make the identification of clusters more rigorous;\textsuperscript{134} this includes assessing the statistical significance of trait clusters via simulation procedures similar to those used to examine gene flow among populations.\textsuperscript{135}

\textbf{Ethical challenges: risks to individual privacy and stigmatization}

Ethical challenges in studies involving transmission dynamics in HIV epidemics extend beyond those faced by randomized HIV control trials\textsuperscript{136} and apply to both concentrated and generalized epidemics. Phylogenetic studies conducted at the village or small community-level may involve collection of HIV sequences and individual clinical and demographic data, and in some studies may include geographic location data. The risk
of individual identification could result in the loss of privacy and, in locations where HIV transmission or reckless exposure is a criminal offence, prosecution. Additionally, an important goal in phylogenetic studies of transmission is to identify traits associated with individuals and groups responsible for onward HIV transmission. Thus there is the potential for stigmatization of individuals linked with or have common features of transmission network members, either underlying (e.g. socioeconomic or demographic group) or proximate (e.g. injection-drug use or sexual practice).

In contrast to the stigma related to HIV-infection, this challenge will include HIV-negative individuals as well.

**Potential solution:** Although sampling from generalized epidemics in regions of high prevalence might make such identifications unlikely, principles and governance on patient identifiable data will be necessary. Data from phylogenetic studies of transmission should be reported in ways where individuals cannot be identified. For example, the UKRDB and the Swiss HIV Cohort Study have adopted the strategy where only a minority (e.g. 10%) of the sequences collected will be released, with these sequences chosen at random. This includes location data; one approach is to reduce the resolution of location data by including a set number of individuals (e.g. 200) in the location set, such that individual identification by location is not possible. These data security strategies will also be addressed in the patient consent process and the tight restriction needed on data release must be recognized by funding agencies and scientific journals.

**CONCLUSIONS**

Opportunities to implement phylogenetic methods at the inception of HIV prevention studies should not be lost. Progress in computational and analytic techniques for reconstructing HIV phylogenies is ongoing. The costs associated with HIV gene or genome sequencing will continue to decrease; rapid, high-throughput sequencing will produce more sequences and larger databases. These advances make it all but certain that sequences or specimens collected in broad reaching studies will eventually be sequenced and used for phylogenetic analyses. However, planning for such analyses from the beginning will maximize their usefulness and the likelihood that phylogenetic analyses can be used in impact evaluations. Additionally, implementing these analyses prospectively will help in identifying hidden sub-populations or core-transmitter groups, as well as monitoring the spread of transmitted drug resistance, especially among rapidly transmitting networks or
ACKNOWLEDGEMENTS

We thank Ward Cates and Nancy Padian for their careful review of earlier versions of the manuscript and Vladimir Novitsky and Mary Kate Grabowski for their helpful discussions. This work stemmed from discussions held during meetings sponsored by the NIH HIV Prevention Trials Network (Feb 21-22, 2012) and the Bill and Melinda Gates Foundation (BMGF) [Nov 1-2, 2012]. We thank participants of these meetings and specifically David Burns at the NIH and Gina Dallabeta at the BMGF. We would also like to acknowledge members of the Phylogenetic and Networks for Generalised HIV Epidemics in Africa (PANGEA), a consortium sponsored by the BMGF.
REFERENCES


137. *FigTree v 1.4 [computer program]2009.*


FIGURE CAPTIONS

**Figure 1.** Example population-level HIV phylogeny reconstructed from HIV pol sequences, to illustrate the basic approach to identifying traits associated with transmission using a phylogeny. Putative clusters of linked transmissions are identified using (ad-hoc) criteria such as pairwise genetic distance and/or nodal support (yellow boxes). Individual clinical or demographic traits are then examined for significant association with linked or unlinked individuals in the phylogeny (red lineages designate individuals with a certain trait, e.g. a particular transmission risk group). Note that not all individuals are included in transmission clusters, and that transmission clusters do not include all individuals with the “red” trait.

**Figure 2.** The regional distribution of HIV sequences deposited in the LANL HIV Sequence Database scaled to the estimated number of persons living with HIV, by WHO region and sub-Saharan Africa (inset). Map generated through query of the LANL database for number of sequences sampled by geographic region and country (http://www.hiv.lanl.gov; queried on June 26, 2013). Numbers were scaled to the 2011 WHO estimates on numbers of persons living with HIV by region and country (http://apps.who.int/gho/data/).
### TABLES

**Table 1.** Categories of genetic analysis related to HIV epidemiology, with examples provided from recent studies conducted in areas with generalized epidemics.

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phylogenetics</td>
<td>Quantification of epidemic growth or decline; and the cross-validation or enhancement of epidemiological model fitting.</td>
<td>Epidemic growth in east Africa, Angola, and Cameroon; Comparison of HIV-2 and HIV-1 population dynamics in west Africa.</td>
</tr>
<tr>
<td>Phylogeography</td>
<td>Evaluation of the relationship between human migration (both temporary and permanent) and disease spread; identification of geographical hubs of transmission; and identification of local and imported cases</td>
<td>Subtype diversity and recombination, Spatial distribution in east Africa, Evolution of subtype C in Zimbabwe, Evolution of HIV-1 in Guinea-Bissau, Historical diversity in Kinshasa, DRC, Global spread.</td>
</tr>
<tr>
<td>Region</td>
<td>Study</td>
<td>Setting / Risk</td>
</tr>
<tr>
<td>--------</td>
<td>-------</td>
<td>----------------</td>
</tr>
<tr>
<td>Europe</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Belgium</td>
<td>Chalmel et al. 2010</td>
<td>Ghent / Mixed</td>
</tr>
<tr>
<td>Spain</td>
<td>Gonzalez-Alba et al. 2011</td>
<td>Madrid / Mixed</td>
</tr>
<tr>
<td>Switzerland</td>
<td>Ambrosioni et al. 2012</td>
<td>Geneva / Mixed</td>
</tr>
<tr>
<td></td>
<td>Kouyos et al. 2010</td>
<td>Switzerland / Mixed</td>
</tr>
<tr>
<td></td>
<td>Yerly et al. 2009</td>
<td>Geneva</td>
</tr>
<tr>
<td>UK</td>
<td>Fisher et al. 2010</td>
<td>Brighton / MSM</td>
</tr>
<tr>
<td></td>
<td>Gifford et al. 2007</td>
<td>UK</td>
</tr>
<tr>
<td></td>
<td>Hué et al. 2005</td>
<td>UK / MSM</td>
</tr>
<tr>
<td></td>
<td>Hughes et al. 2009</td>
<td>UK / HET</td>
</tr>
<tr>
<td></td>
<td>Lewis et al. 2008</td>
<td>London / MSM</td>
</tr>
<tr>
<td></td>
<td>Pao et al. 2005</td>
<td>Brighton / MSM</td>
</tr>
<tr>
<td>North America</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Aldous et al. 2012</td>
<td>5 US CFAR sites/Mixed</td>
</tr>
<tr>
<td></td>
<td>Brenner et al. 2007</td>
<td>Montreal/MSM</td>
</tr>
<tr>
<td></td>
<td>Dennis et al. 2012</td>
<td>North Carolina/Mixed</td>
</tr>
</tbody>
</table>

AHI, acute HIV infection; ART, antiretroviral therapy; CFAR, Center for AIDS Research; HET, heterosexual; IDU, injection drug use; MSM, men who have sex with men; TDR, Transmitted drug resistance
Table 3. Randomized controlled trials for combination HIV prevention and HIV cohort studies with ongoing or planned integration of molecular epidemiology in Africa.

<table>
<thead>
<tr>
<th>Study</th>
<th>Location</th>
<th>Population</th>
<th>Est. HIV prevalence</th>
<th>Study arms/Description</th>
<th>Status</th>
<th>Phylogenetics component</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCPP: Botswana Combination Prevention Project*</td>
<td>Botswana</td>
<td>30 rural communities in 3 geographic areas in Botswana (~6,027 individuals per community). The estimated number of targeted adults aged 16-64 years is ~105,000</td>
<td>25%</td>
<td>A: Combination Prevention Communities, n=15: Rapid scale-up: (1) HTC services (2) ART for adults with CD4≤350 cells/mL, or WHO Stage III/IV; (3) ART for adults with a viral load ≥10,000 copies/mL, CD4&gt;350 cells/mL, and WHO Stage I/II; (4) MC service; (5) PMTCT services, with a target of &gt;90% of pregnant women with HIV receive PMTCT antiretrovirals before 28 weeks gestation, and &gt;90% continue ART indefinitely after starting ART for PMTCT if otherwise ineligible for treatment. B: Enhanced Care Communities, n=15.</td>
<td>Starting</td>
<td>Env gp120 V1C5 sequencing to address: (1) the proportion of new infections in 20% of households that cannot be phylogenetically linked to HIV-infected persons in the same study arm; (2) an adjusted analysis of Combination Prevention package efficacy that includes only linked incident HIV infections per study arm; and (3) the proportion of recent HIV-1 infections that can be linked to HIV-infected adults in intervention communities.</td>
</tr>
<tr>
<td>PopART: Population Effects of Antiretroviral Therapy to Reduce HIV Transmission (HPTN 071)**</td>
<td>Zambia &amp; South Africa</td>
<td>21 clusters (60,000 individuals per cluster, with 2,500 enrolled in observation cohort)</td>
<td>15%</td>
<td>A: Universal community home-based testing; active linkage of HIV-positive individuals to care and immediate ART and/or MC. B: Same as A but ART according to national guidelines C: Enhanced standard of care</td>
<td>Starting</td>
<td>Planned in 12/21 clusters (4 triplets). Samples to be collected from ~30,000 untreated individuals. Sequencing method to be determined.</td>
</tr>
<tr>
<td>TasP study (ANRS 1249)**</td>
<td>KwaZulu Natal, South Africa</td>
<td>34 clusters (2,000 adults/cluster)</td>
<td>20%</td>
<td>A: Expanded testing, MC, immediate ART, IEC, STI treatment B: Same as A but ART at CD4&lt;350 cells/ml</td>
<td>Ongoing</td>
<td>Pol sequences for all individuals with virologic failure or seroconversion within the trial; phylodynamic analysis</td>
</tr>
<tr>
<td>Africa Centre Demographic Surveillance Site*</td>
<td>KwaZulu Natal, South Africa</td>
<td>90,000 people from 11,000 households</td>
<td>24%</td>
<td>Observational study; longitudinal prospective cohort. Annual HIV-testing of all resident adults and a sample of non-resident adults</td>
<td>Ongoing</td>
<td>Pol sequences for all seroconverters in the surveillance area (N&gt;700) between 2004 and 2009</td>
</tr>
<tr>
<td>MP3: An HIV Prevention Program for Mochudi, Botswana*</td>
<td>Botswana</td>
<td>Northeastern sector of Mochudi, total population ~15,000</td>
<td>21%</td>
<td>ART at CD4 ≥250/350 cells/mL and HIV-1 RNA ≥50,000 copies/mL</td>
<td>Completed</td>
<td>Env gp120 V1C5 sequences from 785 individuals to assess general clustering patterns and proportions of clustered vs. non-clustered sequences among</td>
</tr>
<tr>
<td>Rakai Community Cohort Study&lt;sup&gt;7&lt;/sup&gt;</td>
<td>Uganda</td>
<td>12-15,000 total individuals from 46 different communities (general Rakai cohort)</td>
<td>12%</td>
<td>Observational study; longitudinal prospective cohort. 14 near-annual surveys, since 1994. Recent expansion into 4 hotspot communities along Lake Victoria</td>
<td>Ongoing</td>
<td>Gag and env sequenced for &gt;1,000 ART-naïve individuals from 3 survey rounds (1994, 2002, 2009) in general cohort communities</td>
</tr>
</tbody>
</table>

ART, Antiretroviral therapy; HCT, HIV testing and counseling; PMTCT, Prevention of mother to child transmission; MC, medical circumcision; IEC, Information Education Communication campaign; STI, sexually transmitted infection

*Partially adapted from Boily (2012)<sup>142</sup>

2. Johns Hopkins/USAID
3. NIH/Harvard/NIH/LSHTM/Imperial College London, ZAMBART, DTCC of Public Health
4. ANRS
5. Wellcome Trust
6. NIH/Harvard, R01 AI083036; PI: M. Essex;
7. NIH
Transmission cluster (defined by e.g. pairwise genetic distance and node support)

Individual trait (e.g. stage of infection, transmission risk group, locality)
HIV sequences per 100,000 persons living with HIV, 2011
- <500
- 500 - 1499
- 1500 - 3999
- 4000 - 5999
- >6000