Published in final edited form as:

Curr Opin HIV AIDS. 2014 March ; 9(2): 126–133. doi:10.1097/COH.0000000000000040.

Molecular tools for studying HIV transmission in sexual networks

Mary K. Grabowski¹ and Andrew D. Redd²

¹Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD

²Laboratory of Immunoregulation, NIAID, NIH, Baltimore, MD

Abstract

Purpose of review—Phylogenetics is frequently used for studies of population-based HIV transmission. The purpose of this review is to highlight current utilities and limitations of phylogenetics in HIV epidemiological research from sample collection through data analysis.

Recent findings—Studies of HIV phylogenies can provide critical information about HIV epidemics that are otherwise difficult to obtain through traditional study design such as transmission of drug resistant virus, mixing between demographic groups, and rapidity of viral spread within populations. However, recent results from empirical and theoretical studies of HIV phylogenies challenge some of the key assumptions and interpretations from phylogenetic studies. Recent findings include lack of transmission bottlenecks in men who have sex with men and injection drug, evidence for preferential transmission of HIV virus in heterosexual epidemics, and limited evidence that tree topologies correlate with underlying network structures. Other challenges include a lack of a standardized definition for a phylogenetic transmission cluster and biased or sparse sampling of HIV transmission networks.

Summary—Phylogenetics is an important tool for HIV research, and offers opportunities to understand critical aspects of the HIV epidemic. Like all epidemiological research, the methods used and interpretation of results from phylogenetic studies should be made cautiously with careful consideration.

Keywords

HIV; Phylogenetics; Linkage

Introduction

One of the defining features of HIV is its ability to rapidly evolve and persist within individuals despite continued pressure from the cellular and humoral host immune responses[1]. This struggle between virus and human has resulted in one of the most genetically diverse pandemics in recorded history[2]. The diversity of HIV has been undoubtedly critical to its persistence and spread throughout the world; however, recent developments in genetic sequencing technologies, computational methodologies, and

statistics are providing researchers with new tools to utilize viral diversity to combat the global HIV pandemic. Guided by population genetics and epidemiological principles[3], scientists are using viral phylogenetics to improve our understanding of HIV diversity within individuals and populations, generating an unprecedented knowledge of viral dynamics to improve strategies of HIV prevention and treatment of HIV-infected persons[4]. This review examines the application of viral phylogenetics to HIV epidemiological research, with a focus on selected theoretical and practical considerations in phylogenetic studies of HIV transmission.

HIV sequence data for phylogenetic analyses

HIV phylogenetics requires the generation of viral sequences, which in most cases sequences are derived from virions isolated from the peripheral blood of infected persons with unknown duration of disease. However, sexual transmission of HIV, the most common route of infection, involves the transmission of virus within the anogenital mucosa, with the highest probability of viral transmission occurring in the earlier and later stages of disease[5]. In an analysis of HIV envelope sequences obtained from the serum and vaginal secretions of twelve women in a Kenyan cohort, Kemal et al. found phylogenetically related but distinct viral sub-populations within the blood and genital tract (GT)[6]. Similarly, Boeras et al. found distinct, though genetically diverse, viral populations in the GT in chronically infected individuals in HIV serodiscordant relationships, but also showed that only a subset of these GT viruses are transmitted to sexual partners[7]. This study and others of heterosexual (HET) HIV transmission further suggested certain GT viruses are preferentially transmitted to sexual partners[8,9], and that these newly acquired “founder” viruses may be selectively sequestered in the GT of HIV-infected individuals for subsequent transmission[7,10,11].

The extent to which intra-host HIV evolutionary processes affect inference from viral phylogenies is unknown. Volz et al. argued that intra-host selection process have limited impact on inferences from viral phylogenies for three reasons: (1) there is a viral bottleneck at the time of transmission (2) most infections are transmitted during the early stages of infection and (3) transmission of virus is independent of viral factors[12]. While there is strong evidence to support an transmission bottleneck during heterosexual intercourse, studies suggest high multiplicity of HIV infection in MSM and IDU-associated transmissions[13,14]. The role of early stage transmission in HIV epidemics is widely debated[15], but intra-host HIV evolution occurs within only weeks following acquisition and the onset of the adaptive immunologic responses[16]. Phylogenetic studies of HIV epidemics in MSM have suggested early infection sustains MSM epidemics[17-19]; however, this early period may not have as substantial of a role in IDU or HET epidemics [20]. Lastly, given data supporting the preferential HIV transmission and association of viral factors with established viral infection[9,11,21], it seems unlikely that the last of the three arguments is correct.

Viral sequencing technologies

Once HIV has been isolated from the blood or other body fluids and then amplified through a single or series of polymerase chain reactions, a viral sequence(s) can be obtained from the resulting viral amplicon. HIV sequencing methodologies include direct Sanger sequencing[22], single genome amplification or cloning (SGA/cloning)[23], and next-generation sequencing (NGS)[24]. The preponderance of historical sequences currently available in large HIV sequence databases (i.e. Los Alamos HIV database) were generated using Sanger sequencing of a single gene region. Sanger sequencing results is a single consensus sequence, or bulk sequence, from the post-PCR viral amplicon in which each base in the sequence is the most frequent base at that position across all unique sequences within the amplicon[22]. In contrast, SGA/cloning and NGS yield multiple distinct viral sequences, which allows one to quantify and examine viral diversity within samples. The level of viral diversity that can be measured is directly related to the genetic length and depth of sequencing that is obtained[25]. In the case of SGA/cloning this is usually done at a level of a few up to multiple dozens of longer sequences from one sample. NGS sequencing strategies are significantly more robust compared to SGA and can generate tens of thousands of shorter sequences from a single sample. However, certain NGS technologies can be error-prone within homopolymer stretches of DNA[26], and one of the major challenges in using NGS has been to distinguish systematic and random technical error from true viral diversity[27-29]. Due to these inherent limitations, as well as differences in cost and ease of data analysis, these strategies have different applications (Table 1).

One aspect of HIV virology that is particularly important for consideration in sequencing studies of concentrated HIV epidemics, where partner turnover is more frequent and hence lifetime partner numbers higher, is viral coinfection or superinfection with multiple genetically distinct HIV viruses (reviewed by Redd and Tobian[30]). When individuals are infected with multiple viruses either at the time of initial infection or later, a bulk sequence will not adequately represent all viruses that an individual has acquired or possibly transmitted[31,32]. In these instances SGA and NGS methods can provide a more detailed depiction of intra-host HIV diversity[33-35]. Phylogenies reconstructed from bulk sequences may be subject to greater bias with increasing prevalence of HIV coinfection within a population, though this remains an understudied area in phylogenetics.

Reconstructing a phylogenetic tree

HIV phylogenies are evolutionary trees in which the leaves of the tree are the sampled sequences or taxa, branches are the genetic distance between taxa, and the nodes denote estimated speciation events[36]. Reconstruction of HIV phylogenies can be accomplished through a variety of methodological approaches including neighbor joining (NJ), maximum likelihood, and Bayesian methods. These approaches and their strengths and limitations are reviewed in detail by Yang[37]. Briefly, NJ methods employ a hierarchical cluster algorithm wherein aligned sequences are grouped together based on their genetic similarity[38,39]. The NJ algorithm is the most computationally efficient of tree reconstruction approaches in part because the clustering algorithm does not need to search the full tree space, which can be extremely large for datasets with many taxa. Maximum likelihood and Bayesian

approaches are relatively more complex because the searched tree-space is often more extensive and users are required a priori to specify model assumptions, including a model of nucleotide substitution, or a model of HIV evolution[36]. Despite more limited computational efficiency, maximum likelihood and Bayesian approaches remain popular because user-defined evolutionary models can result in more accurate phylogenetic trees when correctly specified, assumptions are explicit, and model specification is flexible[37]. Software has been developed to help analysts determine which evolutionary model is the best for their dataset[40,41].

In a maximum likelihood analysis, the phylogenetic tree that maximizes the likelihood among all possible trees given the sequence data is the maximum likelihood, or best phylogenetic tree[42]. In comparison, posterior probabilities are estimated in Bayesian analyses, where the posterior is a function of the prior probability of the tree and the tree likelihood[36]. Interpretation of Bayesian posterior probabilities are more straightforward than bootstrap values in maximum likelihood analyses; however, node confidence can be higher than expected using Bayesian approaches[43,44]. Formal comparison between Bayesian posterior probabilities and ML bootstrap values reveals that posterior probabilities are almost always higher than corresponding bootstrap values[43,44]. Thus, comparisons of clade support or phylogenetic criteria incorporating clade support values (i.e definitions of HIV transmission clusters) across maximum likelihood and Bayesian analyses should be made with caution.

Bayesian trees were first constructed using a molecular clock assuming a constant rate of evolution across tree branches[37]. This assumption was then relaxed to allow for rate variation across viral lineages and for clock-free analyses[45]. MrBayes and BEAST are two software packages frequently used in Bayesian phylogenetics[46,47], though the latter of these packages is preferred for the estimation of time-scaled HIV phylogenies. Both packages utilize Markov Chain Monte Carlo (MCMC) algorithms[48,49]. The BEAST software is often used for HIV phylodynamic studies, including the estimation of temporal and geographical epidemic dynamics[47]. Critical to the validity of phylodynamic studies is representative sampling over the space and time for which inference is to be made[50,51]. A detailed review of viral phylodynamics and its applications can be found elsewhere[18].

Phylogenetic analysis of HIV transmission networks

Viral linkage is the process of using viral sequence data and phylogenetics to determine connectivity of HIV transmission in groups of individuals that have an underlying epidemiological linkage and one of the members is presumed to be the source of HIV-infection to the other member(s) such as concordant HIV-positive couples, known sexual partners, or needle-sharing groups[31,52-54]. The objective of linkage analysis is to confirm or rule out a specific sexual partner or contact as the infection source; therefore, the criteria used to establish phylogenetic associations should be conservative, with an emphasis on excluding (specificity) rather than confirming (sensitivity) infection sources.

Viral linkage analysis was recently applied in two large HIV clinical trials of acyclovir (Partners in Prevention) and HIV antiretroviral therapy (HPTN-052) for HIV prevention

among HIV serodiscordant couples[31,54]. Criteria for determining whether transmission occurred from the index HIV-positive to the initially HIV-negative partner differed but were highly conservative in both studies. Strict criteria were used so that viral sources of infection were not erroneously attributed to the index HIV infected partner or the benefits of the interventions overestimated. In the Partners in Prevention study (PIP) a newly HIV-seroconcordant couple (n=151) was considered virologically linked if (1) *gag* and/or *env* sequences formed a monophyletic cluster in a maximum likelihood tree and (2) the genetic distance between partner viruses exceeded a probability threshold (50%) using a Bayesian algorithm that was developed by the study investigators. Using these methods the PIP investigators estimated that 26.5% of newly acquired infections were obtained from sources other than the known index HIV-infected partner.

To determine viral linkage of 38 newly HIV seroconcordant couples in HPTN-052 clinical trial, Eshelman et al. also used a similar approach to that applied in the PIP study though they analyzed the *pol* gene and performed 454 pyrosequencing of the *gp41* on all couples that could not be initially linked using consensus sequences. In comparison, the investigators found 18.4% of newly HIV-concordant couples in HPTN-052 were virologically unlinked[31]. The difference between these two rates of linkage is may in part be due to the fact that the PIP study had four times as many transmitting couples.

HIV transmission patterns from phylogenetics

One of the fastest growing applications of HIV phylogenetics is to the study of population-based patterns of HIV transmission. In comparison to viral linkage analyses, the unit of analysis in these studies is the transmission network wherein direct and indirect phylogenetic links between individuals are both of primary interest. In the absence of a fully sampled transmission network and detailed contact histories, direct linkage between two individuals in a phylogenetic tree is not possible[55]. Instead, phylogenetics is used to identify groups of persons who share an HIV transmission chain with unknown directionality and contact structure. Identified chains may then be characterized to inform HIV prevention efforts. Epidemic trajectories and key transmission parameters (reproductive number, generation times[56]) also inform HIV control and these too can be estimated using phylogenetic methods.

Broadly speaking, individuals whose viral sequences are closely related because they have low genetic distance and share a most recent common ancestor constitute a phylogenetic cluster, or putative, recent HIV transmission chain. Definitions of phylogenetic clusters vary widely in the HIV literature, though many studies use both clade support and genetic distance cutoffs (Table 2)[57-76]. Rationale for cutoffs is rarely provided; however, these decisions may affect study inferences. For example, Robinson et al. simulated HIV spread and pathogen phylogenies on two different network topologies and showed that cutoffs choices affect the size and distribution of phylogenetic clusters obtained[77].

The relationship between phylogenetic tree topologies and the structure of the underlying contact network on which those topologies are generated is an ongoing and critical area in HIV phylogenetic research. Interpretation of phylogenetic cluster distribution as indicators

of contact network phenomena is complicated, in part because of the absence of meaningful statistical comparisons between empirical and null cluster distributions. More recently, tree balance statistics have been developed and used to compare phylogenetic clustering patterns against null contact models of random mixing[78]. However, Robinson et al. found that cluster distributions and tree balance statistics revealed little of the global network structure from simulation, particularly when networks were dynamic[77]. Frost and Volz reached similar conclusions, showing tree shape rarely corresponded to population structure[79]. While determining global network structure from HIV phylogenies may be challenging, identifying individuals with related HIV viruses can be done readily from phylogenies and is often very informative. When combined with detailed epidemiological and clinical data, phylogenetic analyses can reveal information of public health relevance including transmission of drug resistant virus and mixing across transmission, demographic and behavioral subgroups[4].

The majority of population-based HIV phylogenetic studies have been conducted outside of Africa, where viral transmission is typically concentrated within high risk populations such as commercial sex workers, MSM and IDUs[80]. Lewis et al. analyzed over 14,000 viral *pol* sequences collected from approximately 60% of HIV-infected men who have sex with men (MSM) in Great Britain[17]. Using dated HIV phylogenies, the investigators found that a substantial fraction of MSM who phylogenetically clustered did so within large groups (10 persons) and over short time periods. In a more recent analysis of this same data but using different methodological approaches, Volz et al. confirmed the importance of early HIV-infection in the UK MSM epidemic[18], and investigations of other MSM epidemics have yielded similar results[19]. However, varying definitions of what constitutes early infection makes conclusions regarding its role in concentrated epidemics difficult from phylogenetics alone[5, 81].

The study of population-based HIV transmission patterns using phylogenetics is also complicated by often biased and or incomplete sampling of transmission networks. Sequences are often obtained through convenience or clinic-based recruitment of participants and sampling fractions of the target network are rarely known or specified (Table 2). Moreover, HIV transmission networks are usually under sampled: the proportion of sequences that phylogenetically cluster rarely exceed 30%[4]. While limited clustering may reflect inadequate sampling of target networks or some loss of tree resolution with intra-host viral evolution, high numbers of singleton lineages may also indicate important epidemiologic phenomenon driving HIV transmission in populations, such as viral introductions via migration or shorter term travel[82].

Conclusion

Phylogenetics is a growing and exciting field in the HIV prevention sciences. Like all sciences, phylogenetics has limitations and when used for epidemiological purposes, it is subject to many of the same measurement and sampling biases of traditional epidemiological study designs. From a public health standpoint, HIV phylogenetics is most powerful when combined with detailed clinical and epidemiological data, in which case HIV phylogenies can reveal critical information relevant to disease control, including

transmission of drug resistant virus, associations between socio-demographic characteristics and viral spread within populations, and the time scales over which HIV epidemics occur. Additional theoretical studies relating HIV phylogenies to network structure and transmission processes are needed.

Acknowledgments

Ms. Grabowski is supported by the Johns Hopkins Bloomberg School of Public Health Sommer Scholarship. Dr. Redd is supported by the Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health.

The authors would like to thank Dr. Ronald Gray for helpful discussions and critical reading of this manuscript.

References

1. Lemey P, Rambaut A, Pybus OG. HIV evolutionary dynamics within and among hosts. *AIDS Rev.* 2006; 8:125–140. [PubMed: 17078483]
2. Hemelaar J, Gouws E, Ghys PD, Osmanov S. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *AIDS.* 2006; 20:W13–23. [PubMed: 17053344]
3. Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, et al. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science.* 2004; 303:327–332. [PubMed: 14726583]
- 4*. Brenner B, Wainberg MA, Roger M. Phylogenetic inferences on HIV-1 transmission: Implications for the design of prevention and treatment interventions. *AIDS.* 2013; 27:1045–1057. Review of the role of phylogenetics in HIV prevention moving forward. Uses the Montreal MSM epidemic as a case study in the power and pitfalls of these approaches in regards to identifying clusters and linking them to new transmissions. [PubMed: 23902920]
5. Wawer MJ, Gray RH, Sewankambo NK, Serwadda D, Li X, et al. Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in rakai, uganda. *J Infect Dis.* 2005; 191:1403–1409. [PubMed: 15809897]
6. Kemal KS, Foley B, Burger H, Anastos K, Minkoff H, et al. HIV-1 in genital tract and plasma of women: Compartmentalization of viral sequences, coreceptor usage, and glycosylation. *Proc Natl Acad Sci U S A.* 2003; 100:12972–12977. [PubMed: 14557540]
7. Boeras DI, Hraber PT, Hurlston M, Evans-Strickfaden T, Bhattacharya T, et al. Role of donor genital tract HIV-1 diversity in the transmission bottleneck. *Proc Natl Acad Sci U S A.* 2011; 108:E1156–63. [PubMed: 22065783]
8. Derdeyn CA, Decker JM, Bibollet-Ruche F, Mokili JL, Muldoon M, et al. Envelope-constrained neutralization-sensitive HIV-1 after heterosexual transmission. *Science.* 2004; 303:2019–2022. [PubMed: 15044802]
9. Hollingsworth TD, Laeyendecker O, Shirreff G, Donnelly CA, Serwadda D, et al. HIV-1 transmitting couples have similar viral load set-points in rakai, uganda. *PLoS Pathog.* 2010; 6:e1000876. [PubMed: 20463808]
10. Lythgoe KA, Fraser C. New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels. *Proc Biol Sci.* 2012; 279:3367–3375. [PubMed: 22593106]
11. Redd AD, Collinson-Streng AN, Chatziandreou N, Mullis CE, Laeyendecker O, et al. Previously transmitted HIV-1 strains are preferentially selected during subsequent sexual transmissions. *J Infect Dis.* 2012; 206:1433–1442. [PubMed: 22997233]
12. Volz EM, Koelle K, Bedford T. Viral phylodynamics. *PLoS Comput Biol.* 2013; 9:e1002947. [PubMed: 23555203]
13. Li H, Bar KJ, Wang S, Decker JM, Chen Y, et al. High multiplicity infection by HIV-1 in men who have sex with men. *PLoS Pathog.* 2010; 6:e1000890. [PubMed: 20485520]
14. Bar KJ, Li H, Chamberland A, Tremblay C, Routy JP, et al. Wide variation in the multiplicity of HIV-1 infection among injection drug users. *J Virol.* 2010; 84:6241–6247. [PubMed: 20375173]

15. Cohen MS, Dye C, Fraser C, Miller WC, Powers KA, et al. HIV treatment as prevention: Debate and commentary--will early infection compromise treatment-as-prevention strategies? *PLoS Med.* 2012; 9:e1001232. [PubMed: 22802728]
16. Wood N, Bhattacharya T, Keele BF, Giorgi E, Liu M, et al. HIV evolution in early infection: Selection pressures, patterns of insertion and deletion, and the impact of APOBEC. *PLoS Pathog.* 2009; 5:e1000414. [PubMed: 19424423]
17. Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ. Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med.* 2008; 5:e50. [PubMed: 18351795]
- 18**. Volz EM, Koopman JS, Ward MJ, Brown AL, Frost SD. Simple epidemiological dynamics explain phylogenetic clustering of HIV from patients with recent infection. *PLoS Comput Biol.* 2012; 8:e1002552. Analysis that argues that the intra-host selection process may have limited impact on inferences from viral phylogenies for three reasons: (1) there is a viral bottleneck at the time of transmission (2) most infections are transmitted during the early stages of infection and (3) transmission of virus is independent of viral factors. [PubMed: 22761556]
19. Brenner BG, Roger M, Routy JP, Moisi D, Ntemgwa M, et al. High rates of forward transmission events after acute/early HIV-1 infection. *J Infect Dis.* 2007; 195:951–959. [PubMed: 17330784]
20. Hughes GJ, Fearnhill E, Dunn D, Lycett SJ, Rambaut A, et al. Molecular phylodynamics of the heterosexual HIV epidemic in the united kingdom. *PLoS Pathog.* 2009; 5:e1000590. [PubMed: 19779560]
21. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, et al. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci U S A.* 2008; 105:7552–7557. [PubMed: 18490657]
22. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A.* 1977; 74:5463–5467. [PubMed: 271968]
23. Salazar-Gonzalez JF, Bailes E, Pham KT, Salazar MG, Guffey MB, et al. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J Virol.* 2008; 82:3952–3970. [PubMed: 18256145]
24. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet.* 2010; 11:31–46. [PubMed: 19997069]
25. Redd AD, Mullis CE, Serwadda D, Kong X, Martens C, et al. The rates of HIV superinfection and primary HIV incidence in a general population in rakai, uganda. *J Infect Dis.* 2012; 206:267–274. [PubMed: 22675216]
26. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature.* 2005; 437:376–80. [PubMed: 16056220]
27. Zagordi O, Klein R, Daumer M, Beerenwinkel N. Error correction of next-generation sequencing data and reliable estimation of HIV quasiespecies. *Nucleic Acids Res.* 2010; 38:7400–7409. [PubMed: 20671025]
28. Beerenwinkel N, Zagordi O. Ultra-deep sequencing for the analysis of viral populations. *Curr Opin Virol.* 2011; 1:413–8. [PubMed: 22440844]
29. Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N. Deep sequencing of a genetically heterogeneous sample: Local haplotype reconstruction and read error correction. *J Comput Biol.* 2010; 17:417–428. [PubMed: 20377454]
30. Redd AD, Quinn TC, Tobian AA. Frequency and implications of HIV superinfection. *Lancet Infect Dis.* 2013; 13:622–628. [PubMed: 23726798]
- 31*. Eshleman SH, Hudelson SE, Redd AD, Wang L, Debes R, et al. Analysis of genetic linkage of HIV from couples enrolled in the HIV prevention trials network 052 trial. *J Infect Dis.* 2011; 204:1918–26. Example of how HIV phylogenetics can be used to help determine the outcome of an HIV prevention trial. [PubMed: 21990420]
32. Redd AD, Collinson-Streng A, Martens C, Ricklefs S, Mullis CE, et al. Identification of HIV superinfection in seroconcordant couples in rakai, uganda, by use of next-generation deep sequencing. *J Clin Microbiol.* 2011; 49:2859–67. [PubMed: 21697329]
33. Kraft CS, Basu D, Hawkins PA, Hraber PT, Chomba E, et al. Timing and source of subtype-C HIV-1 superinfection in the newly infected partner of zambian couples with disparate viruses. *Retrovirology.* 2012; 9:22. 4690-9-22. [PubMed: 22433432]

34. Redd AD, Mullis CE, Serwadda D, Kong X, Martens C, et al. The rates of HIV superinfection and primary HIV incidence in a general population in Rakai, Uganda. *J Infect Dis.* 2012; 206:267–274. [PubMed: 22675216]
35. Barlow KL, Green J, Clewley JP. Viral genome characterisation by the heteroduplex mobility and heteroduplex tracking assays. *Reviews in medical virology.* 2000; 10:321–35. [PubMed: 11015743]
36. Lemey P, Salemi M, VanDamme A, editors. *The phylogenetic handbook: A practical approach to phylogenetic analysis and hypothesis testing.* Cambridge: 2009.
37. Yang Z, Rannala B. Molecular phylogenetics: Principles and practice. *Nat Rev Genet.* 2012; 13:303–314. [PubMed: 22456349]
38. Gascuel O, Steel M. Neighbor-joining revealed. *Mol Biol Evol.* 2006; 23:1997–2000. [PubMed: 16877499]
39. Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 1987; 4:406–425. [PubMed: 3447015]
40. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: More models, new heuristics and parallel computing. *Nat Methods.* 2012; 9:772. [PubMed: 22847109]
41. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 2011; 28:2731–2739. [PubMed: 21546353]
42. Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol.* 1981; 17:368–376. [PubMed: 7288891]
43. Alfaro ME, Zoller S, Lutzoni F. Bayes or bootstrap? A simulation study comparing the performance of bayesian markov chain monte carlo sampling and bootstrapping in assessing phylogenetic confidence. *Mol Biol Evol.* 2003; 20:255–266. [PubMed: 12598693]
44. Erixon P, Svennblad B, Britton T, Oxelman B. Reliability of bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Syst Biol.* 2003; 52:665–672. [PubMed: 14530133]
45. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 2006; 4:e88. [PubMed: 16683862]
46. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, et al. MrBayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst Biol.* 2012; 61:539–542. [PubMed: 22357727]
47. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol.* 2012; 29:1969–1973. [PubMed: 22367748]
48. Yang Z, Rannala B. Bayesian phylogenetic inference using DNA sequences: A markov chain monte carlo method. *Mol Biol Evol.* 1997; 14:717–724. [PubMed: 9214744]
49. Mau B, Newton MA, Larget B. Bayesian phylogenetic inference via markov chain monte carlo methods. *Biometrics.* 1999; 55:1–12. [PubMed: 11318142]
50. Stack JC, Welch JD, Ferrari MJ, Shapiro BU, Grenfell BT. Protocols for sampling viral sequences to study epidemic dynamics. *J R Soc Interface.* 2010; 7:1119–1127. [PubMed: 20147314]
51. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. *PLoS Comput Biol.* 2009; 5:e1000520. [PubMed: 19779555]
52. Nett RJ, Bartschi JL, Ellis GM, Hachey DM, Frenkel LM, et al. Two clusters of HIV-1 infection, rural Idaho, USA, 2008. *Emerg Infect Dis.* 2010; 16:1807–1809. [PubMed: 21029556]
53. Leitner T, Escanilla D, Franzen C, Uhlen M, Albert J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc Natl Acad Sci U S A.* 1996; 93:10864–10869. [PubMed: 8855273]
54. Campbell MS, Mullins JI, Hughes JP, Celum C, Wong KG, et al. Viral linkage in HIV-1 seroconverters and their partners in an HIV-1 prevention clinical trial. *PLoS One.* 2011; 6:e16986. [PubMed: 21399681]
55. Resik S, Lemey P, Ping LH, Kouri V, Joanes J, et al. Limitations to contact tracing and phylogenetic analysis in establishing HIV type 1 transmission networks in Cuba. *AIDS Res Hum Retroviruses.* 2007; 23:347–356. [PubMed: 17411367]

56. Stadler T, Kouyos R, von Wyl V, Yerly S, Boni J, et al. Estimating the basic reproductive number from viral sequence data. *Mol Biol Evol.* 2012; 29:347–357. [PubMed: 21890480]
57. Antoniadou ZA, Kousiappa I, Skoura L, Pilalas D, Metallidis S, et al. Molecular epidemiology of HIV-1 infection in northern greece (2009-2010): Evidence of a transmission cluster of HIV-1 subtype A1 drug resistant strains among men having sex with men. *AIDS Res Hum Retroviruses.* 2013
58. Bezemer D, Faria NR, Hassan A, Hamers RL, Mutua G, et al. HIV type 1 transmission networks among men having sex with men and heterosexuals in kenya. *AIDS Res Hum Retroviruses.* 2013
59. Ruelle J, Ingels MG, Jnaoui K, Ausselet N, Vincent A, et al. Transmission network of an HIV type 1 strain with K103N in young belgian patients from different risk groups. *AIDS Res Hum Retroviruses.* 2013; 29:1306–1309. [PubMed: 23731270]
60. Frentz D, Wensing AM, Albert J, Paraskevis D, Abecasis AB, et al. Limited cross-border infections in patients newly diagnosed with HIV in europe. *Retrovirology.* 2013; 10:36. 4690-10-36. [PubMed: 23551870]
61. Dennis AM, Murillo W, de Maria Hernandez F, Guardado ME, Nieto AI, et al. Social network-based recruitment successfully reveals HIV-1 transmission networks among high-risk individuals in el salvador. *J Acquir Immune Defic Syndr.* 2013; 63:135–141. [PubMed: 23364512]
62. Li L, Chen L, Liang S, Liu W, Li T, et al. Subtype CRF01_AE dominate the sexually transmitted human immunodeficiency virus type 1 epidemic in guangxi, china. *J Med Virol.* 2013; 85:388–395. [PubMed: 23341369]
63. Yebra G, Holguin A, Pillay D, Hue S. Phylogenetic and demographic characterization of HIV-1 transmission in madrid, spain. *Infect Genet Evol.* 2013; 14:232–239. [PubMed: 23291408]
64. Ng KT, Ong LY, Lim SH, Takebe Y, Kamarulzaman A, et al. Evolutionary history of HIV-1 subtype B and CRF01_AE transmission clusters among men who have sex with men (MSM) in kuala lumpur, malaysia. *PLoS One.* 2013; 8:e67286. [PubMed: 23840653]
65. Feng Y, He X, Hsi JH, Li F, Li X, et al. The rapidly expanding CRF01_AE epidemic in china is driven by multiple lineages of HIV-1 viruses introduced in the 1990s. *AIDS.* 2013; 27:1793–1802. [PubMed: 23807275]
66. Siljic M, Salemovic D, Jevtovic D, Pesic-Pavlovic I, Zerjav S, et al. Molecular typing of the local HIV-1 epidemic in serbia. *Infect Genet Evol.* 2013
67. Yebra G, de Mulder M, Holguin A. Description of HIV-1 group M molecular epidemiology and drug resistance prevalence in equatorial guinea from migrants in spain. *PLoS One.* 2013; 8:e64293. [PubMed: 23717585]
68. Murillo W, Veras N, Prosperi M, de Rivera IL, Paz-Bailey G, et al. A single early introduction of HIV-1 subtype B into central america accounts for most current cases. *J Virol.* 2013; 87:7463–7470. [PubMed: 23616665]
69. Temereanca A, Ene L, Mehta S, Manolescu L, Duiculescu D, et al. Transmitted HIV drug resistance in treatment-naïve romanian patients. *J Med Virol.* 2013; 85:1139–1147. [PubMed: 23592112]
70. Audelin AM, Cowan SA, Obel N, Nielsen C, Jorgensen LB, et al. Phylogenetics of the danish HIV epidemic: The role of very late presenters in sustaining the epidemic. *J Acquir Immune Defic Syndr.* 2013; 62:102–108. [PubMed: 23075917]
71. Chen M, Yang L, Ma Y, Su Y, Yang C, et al. Emerging variability in HIV-1 genetics among recently infected individuals in yunnan, china. *PLoS One.* 2013; 8:e60101. [PubMed: 23555898]
72. Han X, An M, Zhang M, Zhao B, Wu H, et al. Identification of 3 distinct HIV-1 founding strains responsible for expanding epidemic among men who have sex with men in 9 chinese cities. *J Acquir Immune Defic Syndr.* 2013; 64:16–24. [PubMed: 23542640]
73. Ivanov IA, Beshkov D, Shankar A, Hanson DL, Paraskevis D, et al. Detailed molecular epidemiologic characterization of HIV-1 infection in bulgaria reveals broad diversity and evolving phylodynamics. *PLoS One.* 2013; 8:e59666. [PubMed: 23527245]
74. Avidor B, Turner D, Mor Z, Chalom S, Riesenber K, et al. Transmission patterns of HIV-subtypes A/AE versus B: Inferring risk-behavior trends and treatment-efficacy limitations from viral genotypic data obtained prior to and during antiretroviral therapy. *PLoS One.* 2013; 8:e57789. [PubMed: 23469241]

75. Ndiaye HD, Tchiakpe E, Vidal N, Ndiaye O, Diop AK, et al. HIV type 1 subtype C remains the predominant subtype in men having sex with men in senegal. *AIDS Res Hum Retroviruses*. 2013; 29:1265–1272. [PubMed: 23742654]
76. Tramuto F, Maida CM, Bonura F, Perna AM, Vitale F. Dynamics and molecular evolution of HIV-1 strains in sicily among antiretroviral naive patients. *Infect Genet Evol*. 2013; 16:290–297. [PubMed: 23470353]
77. Robinson K, Fyson N, Cohen T, Fraser C, Colijn C. How the dynamics and structure of sexual contact networks shape pathogen phylogenies. *PLoS Comput Biol*. 2013; 9:e1003105. [PubMed: 23818840]
78. Leventhal GE, Kouyos R, Stadler T, Wyl V, Yerly S, et al. Inferring epidemic contact structure from phylogenetic trees. *PLoS Comput Biol*. 2012; 8:e1002413. [PubMed: 22412361]
79. Frost SD, Volz EM. Modelling tree shape and structure in viral phylodynamics. *Philos Trans R Soc Lond B Biol Sci*. 2013; 368:20120208. [PubMed: 23382430]
80. UNAIDS. Global HIV/AIDS response: Epidemic update and health sector progress towards universal access. 2011
81. Brown AE, Gifford RJ, Clewley JP, Kucherer C, Masquelier B, et al. Phylogenetic reconstruction of transmission events from individuals with acute HIV infection: Toward more-rigorous epidemiological definitions. *J Infect Dis*. 2009; 199:427–431. [PubMed: 19133810]
82. Holmes EC, Ghedin E, Halpin RA, Stockwell TB, Zhang XQ, et al. Extensive geographical mixing of 2009 human H1N1 influenza A virus in a single university community. *J Virol*. 2011; 85:6923–6929. [PubMed: 21593168]
83. Dennis AM, Hue S, Hurt CB, Napravnik S, Sebastian J, et al. Phylogenetic insights into regional HIV transmission. *AIDS*. 2012; 26:1813–1822. [PubMed: 22739398]

Keynotes

- When using phylogenetics to analyze HIV epidemiological phenomenon, including transmission network structure or transmission dynamics, it is essential that potential measurement and sampling biases are considered in the interpretation of study results.
- HIV sequencing technology is rapidly changing, and it is critical to understand the advantages and limitations of the different sequencing technologies available prior to performing phylogenetic studies.
- There is no consensus on phylogenetic criteria that should be used to establish a putative transmission cluster in population-based studies of HIV transmission networks. Appropriate phylogenetic criteria may depend on the underlying epidemiological and evolutionary dynamics in a given research setting.

Table 1
Overview of common sequencing technologies for HIV research

Technology	No. of sequences	Advantages	Limitations	Utility
Sanger 'bulk'	Single	<ul style="list-style-type: none"> • Low labor • Rapid turnaround • Easily analyzed • Large historical record • Lowest overall cost per sample 	<ul style="list-style-type: none"> • Cannot accurately detect multiple infections • Medium sequence length • Limited ability to capture intra-host viral diversity 	<ul style="list-style-type: none"> • Clinical resistance testing • Large epidemiological studies utilizing phylogenetics • Subtype screening • Initial linkage analysis between individuals with known or suspected linkage
Single genome sequencing (SGS)/Cloning	2-50+	<ul style="list-style-type: none"> • Technically simple • More accurate base calls • Long sequence length • Moderate cost at low copy or sample number 	<ul style="list-style-type: none"> • Viral diversity/ability to detect multiple infections is dependent on number of clones analyzed • Labor intensive and higher costs at high copy or sample numbers 	<ul style="list-style-type: none"> • Viral diversity studies of large genetic regions • Full viral genome sequencing • Recombination studies
Next generation sequencing (NGS)	50-25000+ [§]	<ul style="list-style-type: none"> • More accurately captures intra-host viral diversity including multiple infections • High-throughput • Lowest cost per nucleotide base sequenced 	<ul style="list-style-type: none"> • Highly sensitive to contamination • Requires specialized data cleaning and analysis • Expensive upfront equipment and reagent costs • Limited access globally 	<ul style="list-style-type: none"> • HIV superinfection and dual infection studies • Epidemiological studies of populations at high-risk for multiple infections (IDU, high-incidence populations) • Detection of minor variants

[§] NGS produces sequence reads that are consolidated to form consensus sequences which can total 1-500+ depending on sequencing efficiency

Table 2

Summary of participant recruitment and phylogenetic methods for a convenient sample of 20 phylogenetic studies of HIV-1 transmission clustering published in PubMed Central in 2013

Authors [Ref]	Location	Predominate Risk group	Patient recruitment	No. of Participants	Method of tree reconstruction	Phylogenetic HIV transmission cluster definition
Antoniadou et al. [57]	Greece	HET, IDU, MSM	Clinic	98	NJ	85% bootstrap, 0.015 mean intra-cluster genetic distance
Bezemer et al. [58]	Kenya	HET, MSM	Convenience	674	ML, Bayesian	70% bootstrap, 0.015 mean intra-cluster genetic distance
Ruelle et al. [59]	Belgium	HET, MSM	Clinic	55	ML, Bayesian	0.89 posterior probability
Frentz et al. [60]	Europe, Israel	HET, IDU, MSM	Clinic	4260	ML	98% bootstrap, 0.030 mean intra-cluster genetic distance
Dennis et al. [61]	El Salvador	FSW, MSM	Respondent driven	119	Bayesian	posterior probability=1, 0.015 mean intra-cluster genetic distance
Li et al. [62]	China	HET	Clinic	253	ML	70% bootstrap
Yebra et al. [63]	Spain	HET, IDU, MSM	Clinic	1293	Bayesian	90 posterior probability, cluster depth cutoff ^a
Ng et al. [64]	Malaysia	MSM	Clinic	496	ML, Bayesian	2 individuals from same geographic location, >90% bootstrap, posterior probability=1
Feng et al. [65]	China	HET, IDU, MSM	Clinic	75	ML	90% bootstrap
Siljic et al. [66]	Serbia	HET, MSM	Clinic	221	ML, Bayesian	90% bootstrap, 0.015 mean intra-cluster genetic distance, 0.9 posterior probability
Yebra et al. [67]	Equatorial Guinea ^b	HET	Clinic	278	ML, Bayesian	95% bootstrap, 0.95 posterior probability
Murillo et al. [68]	Central America	HET, FSW, MSM	Clinic	625	ML	S-H test (p-value <0.01) ^c , patristic distance threshold (25th percentile) ^d
Temereanca et al. [69]	Romania	HET, IDU	Clinic	61	ML	0.01 maximum intra-cluster genetic distance
Audelin et al. [70]	Denmark	HET, IDU, MSM	Clinic	1515	NJ, Bayesian	90 % bootstrap, 0.025/ 0.050 mean/ maximum intra-cluster genetic distance, posterior probability=1
Chen et al. [71]	China	HET, IDU	Clinic	308	Neighbor-joining	70% bootstrap
Han et al. [72]	China	MSM	Snowball, Clinic, Cross-sectional survey	583	Neighbor-joining	70% bootstrap
Ivanov et al. [73]	Bulgaria	HET, IDU, MSM	Clinic	125	ML, Bayesian	96 % bootstrap, 0.10 maximum intra-cluster genetic distance, 0.97 posterior probability
Avidor et al. [74]	Israel	IDU, MSM	Clinic	318	Bayesian	0.95 posterior probability
Ndiaye et al. [75]	Senegal	MSM	Snowball	109	ML	98% bootstrap

Authors [Ref]	Location	Predominate Risk group	Patient recruitment	No. of Participants	Method of tree reconstruction	Phylogenetic HIV transmission cluster definition
---------------	----------	------------------------	---------------------	---------------------	-------------------------------	--

Tramuto et al. [76]	Sicily	HET, MSM	Clinic	155	ML	75% bootstrap
---------------------	--------	----------	--------	-----	----	---------------

HET=heterosexuals, IDU= Injection drug users; MSM=Men who have sex with Men, FSW=Female Sex Workers, NJ=Neighbor-joining, ML=Maximum likelihood.

^aDefined as maximum length of time separating the ancestral node and the most recent tip in the cluster; 6 years for non-subtype B viruses and 8 years for subtype B viruses.

^bStudy conducted among Equatorial Guineans attending Spanish clinics

^cShimodaira-Hasegawa test

^dInferred using PhyloPart software