Observational Study

Medicine®

OPEN

Pairwise diversity and tMRCA as potential markers for HIV infection recency

Sikhulile Moyo, MSc, MPH, PhD[a,b,*], Eduan Wilkinson, MSc, PhD[g], Alain Vandormael, MSc, PhD[c], Rui Wang, PhD[d], Jia Weng, PhD[h], Kenanao P. Kotokwe, BSc[b], Simani Gaseitsiwe, PhD[b,d], Rosemary Musonda, MSc, PhD[b,d], Joseph Makhema, MBc, HB, FRCP[b,d], Max Essex, DVM, PhD[b,d], Susan Engelbrecht, PhD[a,e], Tulio de Oliveira, PhD[c,f,g], Vladimir Novitsky, MD, PhD[b,d]

Abstract

Intrahost human immunodeficiency virus (HIV)-1 diversity increases linearly over time. We assessed the extent to which mean pairwise distances and the time to the most recent common ancestor (tMRCA) inferred from intrahost HIV-1C *env* sequences were associated with the estimated time of HIV infection. Data from a primary HIV-1C infection study in Botswana were used for this analysis (N = 42). A total of 2540 HIV-1C *env* gp120 variable loop region 1 to conserved region 5 (V1C5) of the HIV-1 envelope gp120 viral sequences were generated by single genome amplification and sequencing, with an average of 61 viral sequences per participant and 11 sequences per time point per participant. Raw pairwise distances were calculated for each time point and participant using the ape package in R software. The tMRCA was estimated using phylogenetic inference implemented in Bayesian Evolutionary Analysis by Sampling Trees v1.8.2. Pairwise distances and tMRCA were significantly associated with the estimated time since HIV infection (both $P < 0.001$). Taking into account multiplicity of HIV infection strengthened these associations. HIV-1C *env*-based pairwise distances and tMRCA can be used as potential markers for HIV recency. However, the tMRCA estimates demonstrated no advantage over the pairwise distances estimates.

Abbreviations: HIV = human immunodeficiency virus, MCMC = Markov Chain Monte Carlo, tMRCA = the time to the most recent common ancestor, V1C5 = variable loop region 1 to conserved region 5 of the HIV-1 envelope gp120.

Keywords: HIV incidence, HIV recency, HIV-1C, pairwise distances, tMRCA

1. Introduction

Human immunodeficiency virus (HIV) incidence is critical for monitoring the state of the epidemic, as well as for the design and evaluation of prevention interventions. Recency of HIV infection (time since virus transmission) can be estimated by repeat testing a cohort of HIV-negative participants, although this approach is associated with substantial cost and logistical challenges. The cross-sectional methods based on serological markers have been proposed as a reasonable alternative.[1–14] However, the variability of immune response, diversity of HIV-1 subtypes, and unknown use of antiviral therapy present challenges for serological testing.

The diversity of viral quasispecies can be assessed through the distribution of pairwise distances, or by computing the frequency of ambiguous (mixed) calls. During the transition period from early to chronic HIV-1 infection, the diversity of viral quasispecies within the host increases almost linearly.[15–17] This observation provides an opportunity to apply pairwise distances inferred from the intrahost virus sequences to estimate the time since HIV infection (HIV recency).[1,18–20] Approximately 80% of HIV-1 infections are seeded by a single founder strain as a result of the severe bottleneck upon virus transmission.[21–23] Virus sequences representing HIV-1 quasispecies can then be used to reconstruct viral phylogenies and to infer the time to the most recent common ancestor (tMRCA), which may also be useful for estimating HIV recency.[24,25]

Recent developments in sequencing technology have enabled the high throughput of intrahost sequences representing virus quasispecies. These developments have also facilitated the application of coalescent analysis and the estimation of the tMRCA for HIV infection recency.[26,27] For example, Giorgi

Editor: Ping Zhong.

Funding/support: The primary HIV-1C infection study in Botswana (Tshedimoso Study) was supported and funded by the NIH/NIAID (R01 AI057027). SM was supported by the Oak Foundation Fellowship (Grant # OUSA-12-025). SM and SE were supported by Stellenbosch University Division Medical Virology. AV and TdO were supported by a South African MRC Flagship grant (MRC-RFA-UFSP-01–2013/UKZN HIVEPI). TdO is partially supported by the Royal Society-Newton Advanced Fellowship. RW was supported by the NIH/NIAID (R37 AI51164). SM and SG were partially supported by the Wellcome Trust DELTAS Initiatives/Sub-Saharan Africa Network for TB/HIV Research Excellence (SANTHE) (Grant # 07752/Z/15/Z).

The funders had no role in the study design, data collection, decision to publish, or preparation of the manuscript. We thank Lendsey Melton for excellent editorial assistance.

The remaining authors have no conflicts of interest to disclose.

[a] Division of Medical Virology, Stellenbosch University, Tygerberg, South Africa, [b] Botswana-Harvard AIDS Institute Partnership, Gaborone, Botswana, [c] Africa Health Research Institute, School of Nursing and Public Health, University of KwaZulu-Natal, Durban, Republic of South Africa, [d] Harvard T.H. Chan School of Public Health, Boston, MA, USA, [e] National Health Laboratory Services (NHLS), Tygerberg Coastal, South Africa, [f] Research Department of Infection, University College London, London, United Kingdom, [g] College of Health Sciences, University of KwaZulu-Natal, Durban, Republic of South Africa, [h] Division of Sleep Medicine, Brigham and Women's Hospital, Boston, Massachusetts.

[*] Correspondence: Sikhulile Moyo, Botswana-Harvard AIDS Institute Partnership, Gaborone, Botswana (e-mail: sikhulilemoyo@gmail.com).

Medicine (2017) 96:6(e6041)

Received: 20 October 2016 / Received in final form: 9 January 2017 / Accepted: 11 January 2017

http://dx.doi.org/10.1097/MD.0000000000006041

et al[24] have developed a Poisson–Fitter tool that uses a set of homogeneous sequences and performs statistical tests on the Hamming Distance frequency distributions. The tool computes the best fitting Poisson distribution through Maximum Likelihood, performs a Goodness of Fit test, and tests for Star-Phylogeny (a Star-Phylogeny assumes that all the species radiated simultaneously from 1 ancestor). However, this tool requires samples from the very early stage of HIV infection (2–5 weeks), under the assumption of homogeneous viral sequences prior to selection pressure and fast exponential growth. Poon et al[25] demonstrated that tMRCA reconstructed by coalescent analysis of longitudinal virus sequences (generated by next-generation sequencing) can be used for the accurate estimation of HIV recency. Park et al[19,20] have suggested an algorithm for identifying signatures of incident, chronic, and multiple infections to overcome a limitation on pairwise distance-based assays, which may potentially misclassify early infections with multiple distinct founder strains as chronic infections.

However, it remains unclear whether HIV-1 pairwise distances and/or tMRCA can be used as either independent or complementary markers of HIV recency in cross-sectional sample, particularly in the context of heterogeneity of individual immune responses, levels of virus replication, variation of HIV-1 subtypes, nonuniform diversity across HIV-1 genes, and different modes of virus transmission. In this study, we addressed whether estimated time of HIV infection could be determined from cross-sectional sampling by utilizing a cohort with known time of seroconversion and prospective sampling. Normally, only a single-per-person sample is available in many population-based surveys. To assess the utility of pairwise distances and tMRCA in estimating the time since infection in cross-sectional sampling within the early stage of a predominantly heterosexual HIV-1C epidemic in Botswana, we took advantage of the availability of a unique set of samples with known time since infection. We used longitudinal samples as a source of virus sequences representing HIV quasispecies over time as a reference set, and used appropriate statistical techniques to account for multiple sampling from the same individuals by applying a mixed-effects model.

2. Materials and methods

2.1. Study participants

A total of 42 participants recruited into a primary HIV infection cohort in Botswana (Tshedimoso Study) with longitudinal sampling and estimated time of seroconversion were included.[28] The time of HIV infection was considered to be about 14 days prior to seroconversion.[29,30] HIV-1C *env* gp120 variable loop region 1 to conserved region 5 (V1C5) of the HIV-1 envelope gp120 viral sequences were generated by single genome amplification and sequencing, as described elsewhere.[31,32] A total of 2540 sequences represented an average of 61 viral sequences per participant and 11 sequences per time point per participant. The analyzed region of *env* V1C5 corresponds to nucleotide positions 6615 to 7757 relative to the HXB2 reference strain. The study design and participant characteristics were described elsewhere.[31,33,34] Briefly, study participants were predominantly female (76.2%), with a median age of 27 (interquartile range 25–32.5) years at enrollment.

All participants were infected with HIV-1C. Both viral RNA and proviral DNA were used as templates for amplification and sequencing. The accession numbers of the viral sequences used in this study are KC628761–KC630726. The viral mutations within the targeted V1C5 region of HIV-1C *env* gp120 using serial samples have been described elsewhere.[35,36]

The initial set included 223 time points. In the preliminary analysis, we used phylogenetic inference to identify time points with evidence of HIV-1C super-infection based on branching patterns and presence of phylogenetically distinct clusters separated by other participants' sequences. Based on this analysis, 4 time points from 2 participants with evidence of super-infection were excluded from analysis. In addition, 15 time points were excluded due to the initiation of ART. Because virologically suppressed individuals are unlikely to represent recent HIV infection, and are associated with false-recent infections,[9,37] 35 time points with HIV-1 RNA load below 1000 copies/mL were also excluded from analysis. The final reference set of viral sequences comprised 164 time points from 42 participants spanning over 2 years after estimated time of HIV infection/seroconversion. For a subanalysis, we excluded an additional 19 time points with evidence of subclusters (see section "Test for clustering" below). These cases were assumed to represent transmissions of multiple viral variants from the same (or closely related) source(s) of established (chronic) HIV infection. The sample set in the subanalysis therefore included 145 time points from 42 participants.

This study was conducted according to the principles expressed in the Declaration of Helsinki. The study was approved by the Health Research and Development Committee (HRDC) in Botswana (Protocol number PPME-13/18/1) and the Office of Human Research Administration (OHRA) of the Harvard School of Public Health (Protocol number 10491). All participants provided written informed consent.

2.2. Pairwise distances

Uncorrected (raw) pairwise distances were calculated using the ape package in R.[38]

2.3. Recombination analysis and phylogenetic inference

Each set of viral sequences per time point per participant was checked for potential recombination by the RDP v.4 package.[39] Sequences with evidence of recombination signal based on at least 2 out of 7 methods in the RDP package were excluded before the test for clustering was applied. Phylogenetic tree reconstruction was performed using a maximum likelihood (ML) framework, with PhyML[40] as implemented in SeaView v2.4 software[41] using the GTR + G + I model.

2.4. Test for clustering

For each pool of viral sequences per participant per time point, the pairwise distance matrix was generated by dist.dna (ape package in R) using multiple sequence nucleotide alignment. To identify potential clusters within the pool of viral sequences, kmeans (stat package in R) was utilized and partitioning of the pairwise distance matrix into 2 groups (k = 2) was performed. The ratio of withinss (vector of within-cluster sum of squares, 1 per cluster) to betweenss (the between-cluster sum of squares) was used to determine the validity of partitioning. The clustering was considered valid if the ratio values (withinss to betweenss) for both tested clusters were greater than zero and less than 0.2. This range corresponds to the monophyletic lineage of viral sequences with subclustering, which is evident from a combination of the relatively long branches separating clusters of viral sequences and short branches within each cluster. This branching topology was

assumed to be associated with transmission of multiple viral variants from the same (or closely related) source(s) of established (chronic) HIV infection.

2.5. Estimating tMRCA

The tMRCA was estimated for each time point per participant using Bayesian Markov Chain Monte Carlo (MCMC) phylogenetic inference implemented in Bayesian Evolutionary Analysis by Sampling Trees v.1.8.2 package.[42] Each independent run had a chain length of 100,000,000 with a sample frequency of 10,000. The previously estimated intrahost rate of nucleotide substitution per site within the HIV-1 gp120 V1C5 region (1.58×10^{-2}) was used.[32] To avoid over-parameterization, all runs were performed using the HKY substitution model with a gamma distributed rate variation, a strict molecular clock model, and a constant population size tree prior. The MCMC log output of each run was examined in Tracer v1.6[43] to verify convergence and effective sample sizes (ESS) greater than 200. Mean tree model root height was used to determine the time since infection along with the 95% highest posterior density (HPD) intervals.

2.6. Statistical analysis

Our study benefits from a high frequency of HIV testing and longitudinal sampling[28] which enabled us to estimate the time since infection, measured in days since seroconversion date plus 14 days.[29] Because the time of infection was estimated with high precision, we could then compare the accuracy of the pairwise distances and tMRCA.

We used a linear mixed-effects model to assess the association between estimated time since infection and tMRCA or raw pairwise distances, taking into account the intrahost dependency of repeated measurements. We calculated the accuracy (sensitivity and specificity) in predicting time from infection using either tMRCA or raw pairwise distances in categorizing participants within X (X = 130, 160, 360) days from infection. The sensitivity and specificity for tMRCA are defined as P(predicted time from infection using tMRCA < X days | estimated time from infection < X days) and P(predicted time from infection using tMRCA ≥ X days | estimated time from infection ≥ X days). These quantities for raw pairwise distance are defined in the same way. Confidence intervals were obtained using the bootstrap method, treating each participant as a sampling unit and basing on 100 bootstrap samples. Analyses were performed using data within 2 years of infection and based on measurements from 164 time points from 42 participants. Analyses were repeated in a subset of measurements excluding 19 time points with evidence for subclusters (see section "Test for clustering" above). Mixed-effects models were fit using R version 3.2.3. All other analyses were performed using SAS 9.4 (Cary, NC). All *P*-values were 2-sided. *P*-values < 0.05 were considered statistically significant.

3. Results

We investigated whether pairwise distances of viral intrahost sequences and reconstructed tMRCA could be used as markers of time since infection in a predominantly heterosexual HIV-1C epidemic. The estimated time since infection was significantly associated with tMRCA ($\beta = 0.237$; $P < 0.001$) and pairwise distances ($\beta = 27.6$; $P < 0.001$; Table 1). Excluding subclustering resulted in a stronger association with the estimated time since infection for both tMRCA ($\beta = 0.435$; $P < 0.001$) and pairwise distances ($\beta = 43.8$; $P < 0.001$; Table 1). Pairwise distances and tMRCA tend to overestimate time since infection at the early time points and overestimate at the later time points (Fig. 1A and B). Excluding time points with evidence of subclusters seemed to mitigate this trend (Fig. 1C and D).

We calculated the sensitivity and specificity of using pairwise distances or tMCRA to predict time since infection within 130, 180, and 360 days (Table 2). Consistent with the trends seen in Fig. 1 that predicted times using either tMRCA or raw pairwise distance tend to overestimate times from infection in earlier time points but underestimate those in later time points, the sensitivity was lower for earlier time points and higher for later time points, whereas the specificity was higher for earlier points and lower for earlier time points. Both tMRCA and raw pairwise distances had high sensitivity (but lower specificity) at the 360-day threshold.

4. Discussion

Estimation of HIV incidence is crucial for understanding the trends and dynamics of the HIV epidemic, monitoring of HIV incidence, and designing and evaluating preventions and interventions.[44–48] In this study, we assessed whether pairwise diversity of intrahost HIV sequences and inferred tMRCA could be used as potential markers for the estimation of HIV recency. We found that both pairwise distances and tMRCA inferred from intrahost HIV-1C *env* gp120 V1-C5 sequences (representing single time points) correlated directly with the estimated time of HIV infection. We used only a single time point for each pool of viral sequences, as this is the typical case for the vast majority of epidemiological studies interested in estimating HIV recency from cross-sectional sampling.

The diversity of the HIV-1 *env* gene increases linearly at approximately 1% per year during the early stage of HIV-1 subtype B infection,[16] particularly when the transmitted/founder virus is represented by a single viral variant. Therefore, pairwise distances can be used as a marker of HIV recency for infections with a single transmitted virus. However, HIV infections with multiple transmitted viruses, even in the early stage of HIV infection, could have elevated levels of viral diversity which are associated with established/chronic HIV infection.[19] Thus, recent HIV infections with multiple transmitted viruses could be misclassified as chronic HIV infections. In this study, we

Table 1

Associations between analyzed parameters using linear mixed-effect model.

Subsets	Predictor	Beta (β)	Lower 95% CL	Upper 95% CL	*P*
All time points included, n = 164	Pairwise distances	27.6	18.6	36.6	1.48e-08
	tMRCA, years	0.237	0.125	0.350	5.79e-05
Time points with subclusters excluded, n = 145	Pairwise distances	43.8	33.7	53.9	2.35e-14
	tMRCA, years	0.435	0.283	0.587	1.11e-07

CL = confidence limit, tMRCA = time to the most common recent ancestor.

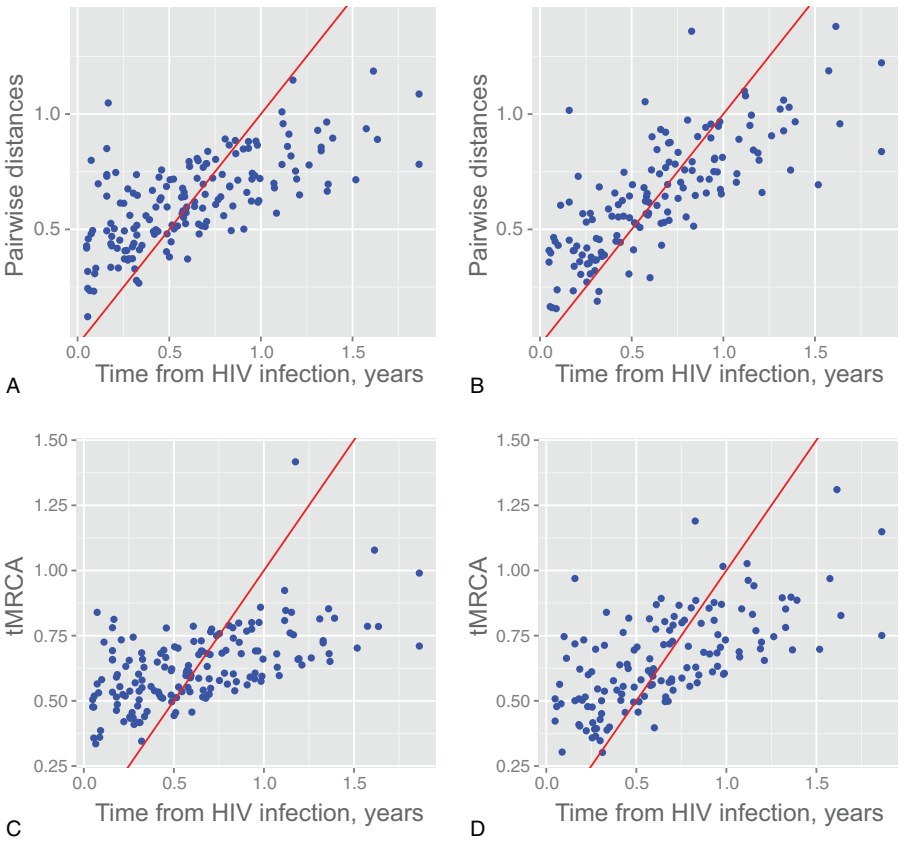

Figure 1. Predicted times from infection using pairwise distance (top panels) or the time to the most recent common ancestor (tMRCA, bottom panels) versus estimated times from infection. The panels on the left represent predictions using all 164 measurements and those on the right represent predictions using a subset of 145 measurements excluding those with evidence for subclustering. The red lines are 45° lines passing origin, reflecting perfect prediction.

demonstrated that controlling for multiplicity of HIV infection improves association between pairwise sequence diversity and the estimated time of HIV infection.

We found no advantage of tMRCA as a marker of HIV recency (in comparison with pairwise distances). It is possible that advanced evolutionary models applied selectively to each pool of viral sequences could improve accuracy in the estimation of tMRCA. Both markers, tMRCA and pairwise distances, are significantly associated with time from HIV infection, suggesting the potential utility of intrahost virus sequences for HIV recency estimation in cross-sectional sampling. However, using either marker alone may not be adequate to predict time from infection. Further research investigating combining information from multiple sources may help improve the prediction accuracy.

The study has limitations. We used the available set of intrahost HIV-1C *env* sequences, which does not necessarily adequately represent the distribution of HIV-1C viruses on a population level. All specimens were collected in 2004 to 2010, and therefore might not reflect the current HIV epidemic in Southern Africa, as circulating viruses could differ over time. The sample set was enriched with early time points, with a relatively small number of later time points. To avoid over-parameterization, we used a relatively simple evolutionary model and simplistic parameters for the estimation of tMRCA, which is another study limitation. To optimize parameters in the Bayesian Evolutionary Analysis by Sampling Trees analysis for a more accurate estimation of tMRCA, further studies are warranted. Our preliminary data suggest that each pool of viral quasispecies

Table 2

Sensitivity and specificity of pairwise diversity and tMRCA estimate for thresholds 130, 180, and 360 days since infection.

	Pairwise distances			tMRCA		
Window, days	**Sensitivity**	**Sensitivity LCL**	**Sensitivity UCL**	**Sensitivity**	**Sensitivity LCL**	**Sensitivity UCL**
≤130	33.3	17.2	49.5	17.8	4.0	31.5
≤180	67.8	53.9	81.7	44.1	28.0	60.1
≤360	97.5	94.8	100.0	98.3	96.1	100.0
	Specificity	**Specificity LCL**	**Specificity UCL**	**Specificity**	**Specificity LCL**	**Specificity UCL**
≤130	98.0	94.9	100.0	100.0	100.0	100.0
≤180	96.5	92.1	100.0	97.7	94.0	100.0
≤360	34.6	5.2	64.1	11.5	0.0	30.4

LCL = lower confidence limit, tMRCA = time to the most common recent ancestor, UCL = upper confidence limit.

might require individual optimization of parameters and model selection. In a limited set of preliminary runs, we found that applying more complex models to a pool of viral quasispecies with low level of viral diversity results in poor convergence and unstable behavior of the MCMC run, which is likely to represent a negative effect of over-parameterization.[49–51] It is also possible that using virus sequences representing multiple time points of sampling (if available) could improve the estimation of tMRCA, as described by Poon et al.[25]

Acknowledgments

The authors thank study participants in the Tshedimoso Study in Botswana; Art Poon and Lesego Gabaitiri for useful discussions; and Botswana Harvard AIDS Institute Partnership and the Tshedimoso Study team. The authors also thank NIH/NIAID (R01 AI057027); Oak Foundation Fellowship (Grant # OUSA-12-025) (to SM); Stellenbosch University Division Medical Virology (to SM and SE); South African MRC Flagship grant (MRC-RFA-UFSP-01–2013/UKZN HIVEPI) (to AV and TdO); Royal Society-Newton Advanced Fellowship (to TdO); NIH/NIAID (R37 AI51164) (to RW); and Wellcome Trust DELTAS Initiatives/Sub-Saharan Africa Network for TB/HIV Research Excellence (SANTHE) (Grant # 07752/Z/15/Z) (to SM and SG) for the support.

References

[1] Moyo S, Wilkinson E, Novitsky V, et al. Identifying recent HIV infections: from serological assays to genomics. Viruses 2015;7: 5508–24.
[2] Hallett TB. Estimating the HIV incidence rate: recent and future developments. Curr Opin HIV AIDS 2011;6:102–7.
[3] Duong YT, Qiu M, De AK, et al. Detection of recent HIV-1 infection using a new limiting-antigen avidity assay: potential for HIV-1 incidence estimates and avidity maturation studies. PLoS One 2012;7:e33328.
[4] Parekh BS, McDougal JS. Application of laboratory methods for estimation of HIV-1 incidence. Indian J Med Res 2005;121:510–8.
[5] Kassanjee R, Pilcher CD, Keating SM, et al. Independent assessment of candidate HIV incidence assays on specimens in the CEPHIA repository. AIDS 2014;28:2439–49.
[6] Brookmeyer R, Laeyendecker O, Donnell D, et al. Cross-sectional HIV incidence estimation in HIV prevention research. J Acquir Immune Defic Syndr 2013;63(Suppl 2):S233–9.
[7] Cousins MM, Konikoff J, Laeyendecker O, et al. HIV diversity as a biomarker for HIV incidence estimation: including a high resolution melting diversity assay in a multi-assay algorithm. J Clin Microbiol 2013;52:115–21.
[8] Cousins MM, Konikoff J, Sabin D, et al. A comparison of two measures of HIV diversity in multi-assay algorithms for HIV incidence estimation. PLoS One 2014;9:e101043.
[9] Konikoff J, Brookmeyer R, Longosz AF, et al. Performance of a limiting-antigen avidity enzyme immunoassay for cross-sectional estimation of HIV incidence in the United States. PLoS One 2013;8:e82772.
[10] Laeyendecker O, Brookmeyer R, Mullis C, et al. Specificity of four laboratory approaches for cross-sectional HIV incidence determination: analysis of samples from adults with known non-recent HIV infection from five African countries. AIDS Res Hum Retroviruses 2012;28: 1177–83.
[11] Laeyendecker O, Kulich M, Donnell D, et al. Development of methods for cross-sectional HIV incidence estimation in a large, community randomized trial. PLoS One 2013;8:e78818.
[12] Laeyendecker O, Piwowar-Manning E, Fiamma A, et al. Estimation of HIV incidence in a large, community-based, randomized clinical trial: NIMH project accept (HIV Prevention Trials Network 043). PLoS One 2013;8:e68349.
[13] Moyo S, LeCuyer T, Wang R, et al. Evaluation of the false recent classification rates of multiassay algorithms in estimating HIV type 1 subtype C incidence. AIDS Res Hum Retroviruses 2014;30:29–36.
[14] Wang R, Weng J, Moyo S, et al. Short communication: effect of short-course antenatal zidovudine and single-dose nevirapine on the BED capture enzyme immunoassay levels in HIV type 1 subtype C infection. AIDS Res Hum Retroviruses 2013;29:901–6.
[15] Ragonnet-Cronin M, Aris-Brosou S, Joanisse I, et al. Genetic diversity as a marker for timing infection in HIV-infected patients: evaluation of a 6-month window and comparison with BED. J Infect Dis 2012;206: 756–64.
[16] Shankarappa R, Margolick JB, Gange SJ, et al. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol 1999;73:10489–502.
[17] Kearney M, Maldarelli F, Shao W, et al. Human immunodeficiency virus type 1 population genetics and adaptation in newly infected individuals. J Virol 2009;83:2715–27.
[18] Xia X-Y, Ge M, Hsi JH, et al. High-accuracy identification of incident HIV-1 infections using a sequence clustering based diversity measure. PLoS One 2014;9:e100081.
[19] Park SY, Love TM, Nelson J, et al. Designing a genome-based HIV incidence assay with high sensitivity and specificity. AIDS 2011;25: F13–9.
[20] Park SY, Goeken N, Lee HJ, et al. Developing high-throughput HIV incidence assay with pyrosequencing platform. J Virol 2014;88: 2977–90.
[21] Keele BF, Giorgi EE, Salazar-Gonzalez JF, 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;105:7552–7.
[22] Abrahams MR, Anderson JA, Giorgi EE, et al. Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-poisson distribution of transmitted variants. J Virol 2009;83:3556–67.
[23] Novitsky V, Wang R, Margolin L, et al. Transmission of single and multiple viral variants in primary HIV-1 subtype C infection. PLoS One 2011;6:e16714.
[24] Giorgi EE, Funkhouser B, Athreya G, et al. Estimating time since infection in early homogeneous HIV-1 samples using a poisson model. BMC Bioinformatics 2010;11:532.
[25] Poon AF, McGovern RA, Mo T, et al. Dates of HIV infection can be estimated for seroprevalent patients by coalescent analysis of serial next-generation sequencing data. AIDS 2011;25:2019–26.
[26] Poon AF, Swenson LC, Bunnik EM, et al. Reconstructing the dynamics of HIV evolution within hosts from serial deep sequence data. PLoS Comput Biol 2012;8:e1002753.
[27] Beerenwinkel N, Gunthard HF, Roth V, et al. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front Microbiol 2012;3:329.
[28] Novitsky V, Wang R, Kebaabetswe L, et al. Better control of early viral replication is associated with slower rate of elicited antiviral antibodies in the detuned enzyme immunoassay during primary HIV-1C infection. J Acquir Immune Defic Syndr 2009;52:265–72.
[29] Fiebig EW, Wright DJ, Rawal BD, et al. Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection. AIDS 2003;17:1871–9.
[30] Cohen MS, Shaw GM, McMichael AJ, et al. Acute HIV-1 Infection. N Engl J Med 2011;364:1943–54.
[31] Novitsky V, Wang R, Margolin L, et al. Timing constraints of in vivo gag mutations during primary HIV-1 subtype C infection. PLoS One 2009;4: e7727.
[32] Novitsky V, Wang R, Rossenkhan R, et al. Intra-host evolutionary rates in HIV-1C env and gag during primary infection. Infect Genet Evol 2013;19:361–8.
[33] Novitsky V, Woldegabriel E, Wester C, et al. Identification of primary HIV-1C infection in Botswana. AIDS Care 2008;20:806–11.
[34] Novitsky V, Wang R, Margolin L, et al. Dynamics and timing of in vivo mutations at Gag residue 242 during primary HIV-1 subtype C infection. Virology 2010;403:37–46.
[35] Novitsky V, Lagakos S, Herzig M, et al. Evolution of proviral gp120 over the first year of HIV-1 subtype C infection. Virology 2009;383:47–59.
[36] Novitsky V, Wang R, Rossenkhan R, et al. Intra-host evolutionary rates in HIV-1C env and gag during primary infection. Infect Genet Evol 2013;19:361–8.
[37] Yu L, Laeyendecker O, Wendel SK, et al. Short communication: low false recent rate of limiting-antigen avidity assay among long-term infected subjects from Guangxi, China. AIDS Res Hum Retroviruses 2015;31: 1247–9.
[38] R Core Team (2016) R: A language and environment for statistical computing. URL https://www.R-project.org/. R Foundation for Statistical Computing, Vienna, Austria.
[39] Martin DP, Lemey P, Lott M, et al. RDP3: A flexible and fast computer program for analysing recombination. Bioinformatics 2010;26:2462–3.
[40] Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 2003;52:696–704.

[41] Gouy M, Guindon S, Gascuel O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 2010;27:221–4.
[42] Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 2007;7:214.
[43] Rambaut A, Suchard MA (2014) Tracer v1.6. Available from http://beast.bio.ed.ac.uk/Tracer.
[44] UNAIDS/WHOTechnical Update on HIV Incidence Assays for Surveillance and Monitoring Purposes. UNAIDS, Geneva:2015.
[45] Guy R, Gold J, Calleja JM, et al. Accuracy of serological assays for detection of recent infection with HIV and estimation of population incidence: a systematic review. Lancet Infect Dis 2009;9:747–59.
[46] Incidence Assay Critical Path Working GroupMore and better information to tackle HIV epidemics: towards improved HIV incidence assays. PLoS Med 2011;8:e1001045.
[47] Kim AA, Hallett T, Stover J, et al. Estimating HIV Incidence among adults in Kenya and Uganda: a systematic comparison of multiple methods. PLoS One 2011;6:e17535.
[48] Sharma UK, Schito M, Welte A, et al. Workshop summary: novel biomarkers for HIV incidence assay development. AIDS Res Hum Retroviruses 2012;28:532–9.
[49] Duchene S, Duchene DA, Di Giallonardo F, et al. Cross-validation to select Bayesian hierarchical models in phylogenetics. BMC Evol Biol 2016;16:115.
[50] Lartillot N, Brinkmann H, Philippe H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol 2007;7(Suppl 1):S4.
[51] Hedge J, Wilson DJ. Practical approaches for detecting selection in microbial genomes. PLoS Comput Biol 2016;12:e1004739.