Validation of molecular clock inferred HIV infection ages: evidence for accurate estimation of infection dates

E. G. Kostaki¹, S. Limnaios¹, S. Roussos¹, M. Psychogiou², G.K. Nikolopoulos³, S-R. Friedman⁴, A. Antoniadou⁵, M. Chini⁶, A. Hatzakis¹, V. Sypsa¹, G. Magiorkinis¹, C. Seguin-Devaux⁷, D. Paraskevis¹

¹*Department of Hygiene, Epidemiology and Medical Statistics, Medical School, National and Kapodistrian University of Athens, Athens, Greece*

²*1st Department of Medicine, Laikon General Hospital, Medical School, National and Kapodistrian University of Athens, Athens, Greece*

³*Medical School, University of Cyprus, Nicosia, Cyprus*

⁴*Department of Population Health, NYU Grossman School of Medicine, New York, NY USA*

⁵*4th Department of Medicine, Attikon General Hospital, Medical School, National and Kapodistrian University of Athens, Athens, Greece*

⁶*3rd Department of Internal Medicine-Infectious Diseases Unit, "Korgialeneio-Benakeio" Red Cross General Hospital, Athens, Greece*

⁷*Department of Infection and Immunity, Luxembourg Institute of Health, Esch sur Alzette, Luxembourg*

Correspondence to

Dimitrios Paraskevis, PhD

Associate Professor of Epidemiology and Preventive Medicine,

Department of Hygiene, Epidemiology and Medical Statistics,

Medical School, National and Kapodistrian University of Athens,

75 Mikras Asias Street, GR-11527, Athens, Greece

Phone : +30 2107462119, e-mail : dparask@med.uoa.gr

Abstract

Improving HIV diagnosis, access to care and effective antiretroviral treatment provides our global strategy to reduce HIV incidence. To reach this goal we need to increase our knowledge about local epidemics. HIV infection dates would be an important information towards this goal, but they are largely unknown. To date, methods to estimate the dates of HIV infection are based mainly on laboratory or molecular methods. Our aim was to validate molecular clock inferred infection dates that were estimated by analysing sequences from 145 people living with HIV (PLHIV) with known transmission dates (clinically estimated infection dates). All HIV sequences were obtained by Sanger sequencing and were previously found to belong to well-established molecular transmission clusters (MTCs). Our analysis showed that the molecular clock inferred infection dates were correlated with the clinically estimated ones (Spearman's Correlation coefficient = 0.93, $p < 0.001$) and that there was an agreement between them (Lin's concordance correlation coefficient = 0.92, $p < 0.001$). For most cases (61.4%), the molecular clock inferred preceded the clinically estimated infection dates. The median difference between clinically and molecularly estimated dates of infection was of 0.18 (IQR: -0.21, 0.89) years. The lowest differences were identified in people who inject drugs of our study population. Our study shows that the estimated time to more recent common ancestor (t_{MRCA}) of nodes within clusters provides a reliable approximation of HIV infections for PLHIV infected within MTCs. Next-generation sequencing data and molecular clock estimates based on heterochronous sequences provide, probably, more reliable methods for inferring infection dates. However, since these data are not available in most of the HIV clinical laboratories, our approach, under specific conditions, can provide a reliable estimation of HIV infection dates and can be used for HIV public health interventions.

Introduction

HIV infection is still expanding around the world. According to WHO/UNAIDS, the number of people living with HIV (PLHIV) globally was 38 million in 2019. In the absence of an effective vaccine, the goal to reduce HIV incidence is based on the 90-90-90 target for diagnosis, access to care and effective antiretroviral (ARV) treatment. Knowledge about local HIV epidemics would facilitate the implementation and success of this ambitious global target (1). One of the critical epidemiological parameters is HIV incidence, but HIV infection dates are largely unknown. This is due to the lack of seronegative tests for the majority of the PLHIV, as well as the rare diagnosis during the acute phase of infection.

To date, the available methods to estimate HIV infection are based on serological or genomic-based approaches. Laboratory (serological) methods, such as the Limiting Antigen Avidity Assay (LAG) or the Calypte Incidence Assay (BED), can classify recent from non-recent infections (2-4). Genomic-based or molecular methods can estimate the age of HIV infection based on previous observations that HIV genetic divergence increases over time. Several methods have been described so far including high resolution melting assays (HMA), the degree of sequence ambiguities in Sanger population sequencing (5-7), molecular clock-based methods (8, 9) and the Hamming distances between different quasispecies (7, 10). More recently, the utilization of next-generation sequencing (NGS), and specifically the deep NGS, provided a new advance for the estimation of the time since HIV infection (11-13).

Our aim was to investigate if time to the most recent common ancestor (t_{MRCA}) of nodes within well-established molecular transmission clusters (MTCs) can be used as a proxy for the HIV infection date, using Sanger sequencing data from densely sampled populations with known transmission dates (clinically estimated infection dates).

Materials and Methods

Our study population consisted of PLHIV with known transmission dates from the following datasets: i) a "seek-test-treat" intervention programme named "ARISTOTLE" implemented in Athens, Greece, during 2012-2013 (14), ii) a network-based intervention that targeted persons with recent HIV-infection ("TRIP") implemented in Athens, Greece, between 2013 and 2015 (15), iii) PLHIV infected within a subtype A1 MTC in Greece (16, 17), iv) People who inject drugs (PWID) infected locally during an HIV outbreak in Luxembourg since 2014 and v) PLHIV infected within a CRF42_BF MTC in Luxembourg (18) (Table 1). HIV sequences were selected to belong to PWID-specific MTCs [datasets (i), (ii) and (iv)] and local MTCs [datasets (iii) and (v)] (19). Local MTCs were defined as phylogenetic clusters including ≥ 2 sequences from the same geographic area (Greece or Luxembourg) at a proportion greater than 70% compared to the total number of sequences within the cluster, as previously described in detail (16, 20). All sequences were generated before the ARV treatment initiation. Each patient was specified by a unique serial number to avoid double-counting. Duplicate patients were detected and excluded from the study population.

Known transmission dates were available from: a) the midpoint between the last available seronegative and the first seropositive tests for HIV, the possible date of HIV acquisition based on the diagnosis of acute HIV infection and patient self-report [datasets (i), (iii), (iv) and (v)] (17), or b) the putative date of HIV infection which was determined using the LAg method [dataset (ii)] (2, 15). From now on we will refer at these dates as "clinically estimated infection dates".

Molecular clock calculation using phylodynamic analysis was used for the re-estimation of the infection dates of our study population (PLHIV with an available clinically estimated infection date). Phylodynamic analyses were applied on PR and partial RT sequences

generated by Sanger sequencing and implemented in BEAST (versions 1.8.0 and 2.1.3), using different nucleotide substitution models (GTR+G, HKY+G), uncorrelated lognormal relaxed clock models and the birth-death models (21, 22). The full description of these analyses has been provided in detail elsewhere (17, 19, 23). The age of infection (infection dates) was approximated to the median estimate of the t_{MRC} of the internal nodes within each MTC. We will refer at these dates as “molecular clock inferred infection dates”.

Medians, interquartile ranges (IQRs) and 95% confidence intervals (CIs) were used to summarize the data. Spearman’s correlation coefficient was used to measure the strength and direction of association between the clinically estimated and molecular clock inferred infection dates, while Lin’s concordance correlation coefficient was used to measure the agreement between these estimates. The fitted regression line was estimated by a quartile (median) regression model and was compared to the line of equality by testing the two-tailed hypothesis of slope = 1 and intercept = 0. In addition, the Bland-Altman plot was used to quantify and describe the agreement between the two variables and the presence of outliers. Spearman’s correlation coefficient was also used to assess putative correlation between the difference of the clinically and molecularly estimated dates of infection and the time interval from the last seronegative to the first seropositive tests for HIV, for a subset of our study population (26 PLHIV from whom the clinically estimated infection date was represented by the midpoint between these tests). The level of significance was set at 0.05. Analyses were performed in Stata 13-StataCorp LP software.

Results

The total number of HIV sequences used in the context of our study was 145. These sequences were retrieved from five different datasets, as it was described in detail previously in the methods section. The samples from “ARISTOTLE” [dataset (i), n=23] and “TRIP” [dataset (ii), n=13] were merged since they were drawn from PWID participating in intervention programmes implemented in Athens, Greece, during the same time period (Table 1). The data from Luxembourg included both samples from PWID infected during an outbreak [dataset (iv), n=46] and non-PWID found within a CRF42_BF MTC expanding locally [dataset (v), n=39] (Table 1). The sampling periods spanned over a long time between 17 June 2003 and 11 January 2018 for the data from Luxembourg [datasets (iv) and (v)], and 22 May 2008 - 05 December 2014 for the subtype A1 data from Greece [dataset (iii), n=24] (Table 1).

The molecular clock inferred infection dates were positively correlated with the clinically estimated infection dates with the Spearman's Correlation coefficient equal to 0.93 ($p<0.001$) (Figure 1). For most cases, the molecular clock inferred infection dates preceded the clinically estimated ones (89 of 145, 61.4%). Moreover, the fitted regression line was described by the equation: molecular clock inferred = $-53.49 + 1.03 \times$ clinically estimated, with 95% CI for the estimated intercept at (-138.49, 31.51) ($p=0.216$) and with 95% CI for the estimated slope at (0.98, 1.07) ($p=0.0514$) (Figure 1). Lin's concordance correlation coefficient was 0.92 (95% CI: 0.90, 0.94) and showed an agreement between clinically and molecularly estimated infection dates ($p<0.001$).

The median difference between the clinically estimated and molecular clock inferred infection dates was 0.18 (IQR: -0.21, 0.89) years across all sequences (n=145) (Table 2). The largest differences were found for the non-PWID infected with subtype A1 in Greece [dataset (iii), median value: 2.04 (IQR: 0.62, 2.99) years] and for the non-PWID of the CRF42_BF

clade in Luxembourg [dataset (v), median value: 0.52 (IQR: -0.16, 1.28) years] (Table 2). Generally, for the PWID of our study population the differences were very low [datasets (i) & (ii): median value: 0.09 (IQR: -0.04, 0.36) years; dataset (iv): median value: -0.24 (IQR: -0.46, 0.19) years] (Table 2). The differences between the clinically and molecularly estimated infection dates for all PLHIV of our study population were plotted against the mean obtained by the two variables (Figure 2). The Bland–Altman plot revealed an agreement between the two variables, an absence of a consistent bias and the existence of a small number of outliers [6 of 145 (4.1%) PLHIV with a difference greater than 4 years] (Figure 2). Moreover, the difference between the clinically estimated and molecular clock inferred infection dates was not correlated with the time interval from the last seronegative to the first seropositive tests ($p=0.822$).

The median estimate of the time interval between the molecular clock inferred infection dates and the sampling dates of the sequences was 0.59 (IQR: 0.25, 1.45) years (Table 2). As for the difference between the clinically and molecularly estimated infection dates, the largest values of this time interval were found for the non-PWID infected with subtype A1 [dataset (iii), median value: 2.42 (IQR: 1.09, 3.46) years] and CRF42_BF [dataset (v), median value: 1.26 (IQR: 0.38, 2.13) years] (Table 2). For PWID infected in Luxembourg [dataset (iv)] the median value of the interval was 0.34 (IQR: 0.14, 0.59) years, however this value was small and comparable to that for PWID infected in Athens, Greece [datasets (i) & (ii), median value: 0.32 (IQR: 0.18, 0.60)

Discussion

Knowledge of infection dates is valuable for estimating the number of new HIV infections per year, the improvement of linkage to care and also for tailoring efforts to control HIV transmission (24, 25). Our study provides evidence that, under specific conditions, HIV data generated by Sanger sequencing can be used for reliable estimation of HIV infection dates. Specifically, the t_{MRCA} of nodes within local MTCs can provide a proxy of the HIV infection dates. Our validation using sequences from PLHIV with previously known transmission dates (clinically estimated dates of infection) suggested that the difference between clinically estimated and molecular clock inferred infection dates was smaller for PLHIV infected within MTCs with short branches translated as short time intervals between sampling and infection. On the other hand, this difference was not correlated with the time interval from the last seronegative to the first seropositive tests, suggesting that the discordance between the two infection dates was not due to a larger uncertainty of the clinically estimated dates of infection. The molecular clock inferred age of infection preceded the clinically estimated one for most PLHIV due to the characteristics of HIV transmission. Specifically, the t_{MRCA} of an internal node of viral sequences from two persons (i.e., the source and the recipient) corresponds to the coalescent time of the viral lineages in the source (8). Therefore, the molecular clock inferred dates of infection can precede the clinically estimated dates due to the existing genetic divergence at the time of infection (pre-transmission interval) (8, 9). In our case, only a single date was estimated for both the source and the recipient and thus the accuracy of the infection ages depends also on the serial interval (time between two HIV cases in a transmission chain). This probably explains why the accuracy is higher for PWID, since in this case a large number of transmissions occurred within a short time frame.

To date, molecular markers, such as the low proportion of ambiguity sites in partial HIV-1 sequences, have been shown to provide a reliable marker of recent transmission (5-7). Kouyos *et al* suggested that the proportion of ambiguity sites increased by 0.2% within the first 8 years and a proportion >0.5% is a marker of a non-recent infection (5). Earlier studies have shown that molecular clock analyses using multiple sequences per patient can reliably estimate HIV-1 infection dates (8, 9). More recently, Puller *et al* showed that within patient viral diversity estimated using NGS data can be used for the accurate estimation of infection dates (11). The accuracy of the estimations was improved by increasing the sequence depth and particularly viral diversity in the 3rd codon position increased linearly with time (11). NGS data are superior to the fraction of ambiguous sites or Sanger sequencing (11, 12). Moreover, NGS does not require heterochronous sampling and this provides an additional strength of this method.

In the current study, we show that t_{MRCA} estimations of the infection ages based on Sanger sequencing data can be used reliably as an approximation of HIV infection dates. Our approach has a few limitations, such as that it can be utilized only for sequences from PLHIV infected with MTCs. Our approach is probably less reliable than NGS, however NGS data or heterochronous sequences are largely missing from HIV clinical laboratories. Alternatively, and under specific conditions, widely available HIV-1 Sanger sequencing data can be used for the estimation of HIV infection dates. Finally, if MTCs have been characterized, our study suggests that dating can be used for an on-going epidemic to quickly assess whether a given blood sample represents a recent infection. This is a useful adjunct to LAg for HIV public health interventions in accordance with one of the key strategies to respond quickly to potential HIV outbreaks as part of the Ending HIV plan in the USA.

Funding

The study has been supported in part by the following grants: i) “ARISTOTLE” program was implemented under National Strategic Reference Framework 2007-2013 (MIS 365008) and was cofunded by the European Social Fund, national resources and the Hellenic Scientific Society for the Study of AIDS and STDs, ii) “TRIP” program was supported by the United States National Institute on Drug Abuse (NIDA) (DP1 DA034989), iii) data from Luxembourg were collected in the context of studies supported by grants from the Ministry of Health of Luxembourg (HIV-MSAN) and iv) 2018 Asklepios Gilead Hellas Grants Programme.

Competing financial interests

The authors declare no competing financial interests.

Table 1 Number and risk group of people living with HIV of the study population, sampling period, HIV-1 subtype and country of sampling of the sequences under study.

Characteristic	PLHIV (n)	Risk group	Sampling period	HIV-1 subtype	Country of sampling
Overall	145		2003 – 2018		
Dataset					
i & ii	36	PWID	2012 – 2013		Greece
iii	24	Non-PWID	2008 – 2014	A1	Greece
iv	46	PWID	2005 – 2018		Luxembourg
v	39	Non-PWID	2003 – 2017	CRF42_BF	Luxembourg

PLHIV, people living with HIV; PWID, people who inject drugs

Table 2 Difference between clinically estimated and molecular clock inferred infection dates and time interval between molecular clock inferred infection dates and sampling dates for the study population.

Characteristic	Difference between estimated infection dates, in years, median (IQR)	Time interval between infection and sampling, in years, median (IQR)
Overall	0.18 (-0.21, 0.89)	0.59 (0.25, 1.45)
Dataset		
i & ii	0.09 (-0.04, 0.36)	0.32 (0.18, 0.60)
iii	2.04 (0.62, 2.99)	2.42 (1.09, 3.46)
iv	-0.24 (-0.46, 0.19)	0.34 (0.14, 0.59)
v	0.52 (-0.16, 1.28)	1.26 (0.38, 2.13)

IQR, interquartile range

Figures

Figure 1 Scatter plot showing relationship between molecular clock inferred infection dates (shown along the y-axis) and clinically estimated infection dates (shown along the x-axis) for 145 people living with HIV of the study population. Each individual is represented by one circle. The solid line corresponds to the regression fitted line and the dashed line to the line of equality. Spearman's correlation coefficient: $r=0.93$ ($p<0.001$).

Figure 2 Bland-Altman plot showing the differences (in years) between the clinically estimated and molecular clock inferred infection dates (shown along the y-axis) compared with the mean of the dates (x-axis). The dashed lines represent the 5th and 95th percentiles for the differences and the solid line corresponds to the median difference.

References

1. Hall HI, Green TA, Wolitski RJ, Holtgrave DR, Rhodes P, Lehman JS, et al. Estimated future HIV prevalence, incidence, and potential infections averted in the United States: a multiple scenario analysis. *J Acquir Immune Defic Syndr*. 2010;55(2):271-6.
2. Nikolopoulos GK, Katsoulidou A, Kantzanou M, Rokka C, Tsiara C, Sypsa V, et al. Evaluation of the limiting antigen avidity EIA (LAgi) in people who inject drugs in Greece. *Epidemiol Infect*. 2017;145(2):401-12.
3. Busch MP, Pilcher CD, Mastro TD, Kaldor J, Vercauteren G, Rodriguez W, et al. Beyond detuning: 10 years of progress and new challenges in the development and application of assays for HIV incidence estimation. *AIDS*. 2010;24(18):2763-71.
4. Guy R, Gold J, Calleja JM, Kim AA, Parekh B, Busch M, 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(12):747-59.
5. Kouyos RD, von Wyl V, Yerly S, Boni J, Rieder P, Joos B, et al. Ambiguous nucleotide calls from population-based sequencing of HIV-1 are a marker for viral diversity and the age of infection. *Clin Infect Dis*. 2011;52(4):532-9.
6. Lunar MM, Zidovec Lepej S, Poljak M. Sequence ambiguity determined from routine pol sequencing is a reliable tool for real-time surveillance of HIV incidence trends. *Infect Genet Evol*. 2019;69:146-52.
7. Andersson E, Shao W, Bontell I, Cham F, Cuong do D, Wondwossen A, et al. Evaluation of sequence ambiguities of the HIV-1 pol gene as a method to identify recent HIV-1 infection in transmitted drug resistance surveys. *Infect Genet Evol*. 2013;18:125-31.
8. Leitner T, Albert J. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc Natl Acad Sci U S A*. 1999;96(19):10752-7.
9. Paraskevis D, Magiorkinis E, Magiorkinis G, Kiosses VG, Lemey P, Vandamme AM, et al. Phylogenetic reconstruction of a known HIV-1 CRF04_cpx transmission network using maximum likelihood and Bayesian methods. *J Mol Evol*. 2004;59(5):709-17.
10. Allam O, Samarani S, Ahmad A. Hammering out HIV-1 incidence with Hamming distance. *AIDS*. 2011;25(16):2047-8.
11. Puller V, Neher R, Albert J. Estimating time of HIV-1 infection from next-generation sequence diversity. *PLoS Comput Biol*. 2017;13(10):e1005775.
12. Carlisle LA, Turk T, Kusejko K, Metzner KJ, Leemann C, Schenkel CD, et al. Viral Diversity Based on Next-Generation Sequencing of HIV-1 Provides Precise Estimates of Infection Recency and Time Since Infection. *J Infect Dis*. 2019;220(2):254-65.
13. Leitner T. Phylogenetics in HIV transmission: taking within-host diversity into account. *Curr Opin HIV AIDS*. 2019;14(3):181-7.
14. Hatzakis A, Sypsa V, Paraskevis D, Nikolopoulos G, Tsiara C, Micha K, et al. Design and baseline findings of a large-scale rapid response to an HIV outbreak in people who inject drugs in Athens, Greece: the ARISTOTLE programme. *Addiction*. 2015;110(9):1453-67.
15. Nikolopoulos GK, Pavlitina E, Muth SQ, Schneider J, Psychogiou M, Williams LD, et al. A network intervention that locates and intervenes with recently HIV-infected persons: The Transmission Reduction Intervention Project (TRIP). *Sci Rep*. 2016;6:38100.
16. Kostaki EG, Gova M, Adamis G, Xylomenos G, Chini M, Mangafas N, et al. A Nationwide Study about the Dispersal Patterns of the Predominant HIV-1 Subtypes A1 and B in Greece: Inference of the Molecular Transmission Clusters. *Viruses*. 2020;12(10).
17. Kostaki EG, Hodges-Mameletzis I, Magiorkinis G, Adamis G, Xylomenos G, Nikolopoulos G, et al. Earlier treatment initiation is associated with a decreased number of HIV-1 subtype A1 transmissions in Greece. *Sex Transm Infect*. 2020.
18. Struck D, Roman F, De Landtsheer S, Servais JY, Lambert C, Masquelier C, et al. Near Full-Length Characterization and Population Dynamics of the Human Immunodeficiency Virus Type 1

Circulating Recombinant Form 42 (CRF42_BF) in Luxembourg. *AIDS Res Hum Retroviruses*. 2015;31(5):554-8.

19. Kostaki E, Magiorkinis G, Psychogiou M, Flampouris A, Iliopoulos P, Papachristou E, et al. Detailed Molecular Surveillance of the HIV-1 Outbreak Among People who Inject Drugs (PWID) in Athens During a Period of Four Years. *Curr HIV Res*. 2017;15(6):396-404.

20. Paraskevis D, Beloukas A, Stasinou K, Pantazis N, de Mendoza C, Bannert N, et al. HIV-1 molecular transmission clusters in nine European countries and Canada: association with demographic and clinical factors. *BMC Med*. 2019;17(1):4.

21. Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol*. 2014;10(4):e1003537.

22. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol*. 2012;29(8):1969-73.

23. Paraskevis D, Kostaki EG, Arendt V, Staub T, François JH, Guillorit L, et al. Molecular investigation of the HIV-1 outbreak among people who inject drugs (PWID) in Luxembourg: evidence for local origin and increased transmission rate during 2013-2015. In *Proceedings of the 16th European AIDS Conference, Milan, Italy, 25-27 October 2017*; Abstract Number BPD3/2. Available online: <https://www.abstractserver.com/eacsabstractarchive/> (accessed on 24 November 2020).

24. Abrahams MR, Anderson JA, Giorgi EE, Seoighe C, Mlisana K, Ping LH, 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(8):3556-67.

25. Brenner BG, Roger M, Routy JP, Moisi D, Ntemgwa M, Matte C, et al. High rates of forward transmission events after acute/early HIV-1 infection. *J Infect Dis*. 2007;195(7):951-9.