The Journal of Infectious Diseases

BRIEF REPORT

Genetic diversity from proviral DNA as a proxy for time since HIV-1 infection

Marius Zeeb[1,2], Paul Frischknecht[1], Michael Huber[2], Corinne D. Schenkel[1], Kathrin Neumann[1], Christine Leeman[1], Julia Notter[3], Andri Rauch[4], Marcel Stöckle[5,6], Matthias Cavassini[7], Enos Bernasconi[8], Dominique L. Braun[1], Huldrych F. Günthard[1,2,*], Karin J. Metzner[1,2,*], Roger D. Kouyos[1,2,*], and the Swiss HIV Cohort study

[1] Department. of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, 8091 Zurich, Switzerland; [2] Institute of Medical Virology, University Zurich, 8057 Zurich, Switzerland; [3] Division of Infectious Diseases and Hospital Epidemiology, Cantonal Hospital St Gallen, 9007 St. Gallen, Switzerland; [4] Department of Infectious Diseases, Inselspital, Bern University Hospital, University of Bern, 3010 Bern, Switzerland; [5] Division of Infectious Diseases & Hospital Epidemiology, University Hospital Basel, 4031 Basel, Switzerland; [6] Medical Faculty, University of Basel, 4031 Basel, Switzerland; [7] Division of Infectious Diseases, University Hospital Lausanne, University of Lausanne, 1011 Lausanne, Switzerland; [8] Division of infectious diseases, Ente Ospedaliero Cantonale, Lugano, University of Geneva and University of Southem Switzerland, 6900 Lugano, Switzerland.

HIV-1 RNA genetic diversity predicts time since infection which is important for clinical care and research. It's unclear, however, whether proviral DNA genetic diversity sampled under suppressive antiretroviral therapy can be used for this purpose. We tested whether proviral genetic diversity from NGS sequences predicts time since infection and recency in 221 people with HIV-1 with known infection time. Proviral diversity was significantly associated with time since infection ($p<5*10^{-07}$, R^2 up to 25%) and predictive of treatment initiation during recent infection (AUC-ROC up to 0.85). This shows the utility of proviral genetic diversity as a proxy for time since infection.

*These authors contributed equally to this work as last author.
Corresponding author information: Marius Zeeb, Universitätsspital Zürich, Division of Infectious Diseases and Hospital Epidemiology, Rämistrasse 100, CH-8091 Zürich, Switzerland, Phone +41 43 253 01 88, marius.zeeb@usz.ch

DOI: 10.1093/infdis/jiae149

Keywords: HIV-1; next-generation sequencing; proviral diversity; infection recency; time since infection

INTRODUCTION

Knowing the time since infection in people with human immunodeficiency virus type I (PWH) is relevant for transmission epidemiology, HIV therapy and for many research questions in general. Since a longer time of infection without therapy means a longer period of ongoing replication and therefore increased viral evolution, it directly impacts the within host viral diversity and proviral reservoir size. This has implications, e.g., when deciding on simplifying antiretroviral therapy [1] or in investigations about immune responses [2]. Yet, its estimation is often challenging due to lack of a previous negative HIV test or recall of unambiguous risk situations leading to an infection.

As HIV diversity increases with infection time, different diversity-based approaches have been developed for estimating time since infection and especially if a PWH is recently, i.e., less than 1 year, infected. For example, [3] used ambiguous nucleotide frequency from Sanger sequences from routine HIV drug resistance testing, and [4,5] showed that an average pairwise diversity score (APD) based on NGS provides an even more accurate measure: In plasma virus derived sequences from ART-naive PWH, APD score correlates well with time since infection and has an ROC area under the curve (AUC) of over 95% to determine if PWH were infected recently [4].

For a large number of PWH, the pre-ART sequences required for these approaches are not available. However, increasing numbers of PWH may have proviral DNA sequences performed for research purposes or to guide treatment simplifications or treatment with long acting antiretrovirals [6]. Such proviral DNA sequences might in principle inform on the time between infection and therapy initiation, as it is expected that the diversity of the viral reservoir increases with the length of this time window, but then stops after ART has suppressed viral replication [7,8]. However, proviral diversity also differs in important ways from pre-ART viral diversity: proviral diversity represents the accumulated diversity over the entire infection, it may be affected by the decay of the reservoir, and by hypermutations in proviral DNA caused by APOBEC3G/F [9].

As these differences may affect the association with prediction of infection time, we evaluate in this study the utility of proviral sequences sampled post ART as a proxy for the time between infection and ART. Given the role of APOBEC3G/F as a source of noise, we combine this approach with a hypermutation filtering on a next generation sequencing (NGS) read level.

METHODS

PWH/Sequence selection criteria

We included PWH with an accurate date of infection enrolled in the Swiss HIV Cohort Study (SHCS), a prospective, multicentric cohort study enrolling PWH in Switzerland [10] and/or in the Zurich Primary Infection Cohort (ZPHI) a multi-centric cohort study enrolling PWH during primary HIV infection [11]. These include PWH with a negative HIV-1 test within one year prior to the date of diagnosis and PWH with a clinical diagnosis of a documented primary HIV infection based on a comprehensive clinical assessment by a highly experienced research team. We determined the date of diagnosis as the earliest date of the following events: SHCS registration, first HIV-1 positive test, or first HIV-1 laboratory measurement. The date of infection was defined as described previously [3]: (i) for PWH in the ZPHI as the estimated date of infection, (ii) for PWH with primary infections as the date of diagnosis minus 30 days (to account for incubation time), and (iii) for all others as the midpoint between diagnosis date and last negative test. We selected proviral NGS sequences from those selected PWH without ART interruption and virological failure until sampling happened. Samples were predominantly sequenced in a study, which systematically sequenced the proviral DNA of all SHCS participants without HIV-RNA genotyping available [12]. We considered the length of two time-windows for the analysis, the number of years from the date of infection until date of ART start ($t_{InfectionToART}$), i.e., time since infection, and the time number of years from ART start until proviral NGS sequence sampling ($t_{ARTtoSampling}$) (Figure 2A).

NGS sequencing

DNA was isolated from on average 5 million PBMCs and proviral DNA was amplified by (i) near full-length PCR and followed by two nested hemi length PCRs [12]. If unsuccessful, (ii) near full-length PCR followed by nested near full-length PCR or (iii) two hemi-length PCRs amplifying a 5' amplicon and a 3' amplicon followed by nested hemi-length PCRs was performed as previously described [7]. NGS sequencing was performed for the near full-length HIV-1 genome using the MiSeq Reagent Kit v2 (300-cycles). Majority consensus alignments were created from the NGS reads using SmaltAlign (https://github.com/medvir/SmaltAlign). From majority consensus sequences, respective genes (*gag*, *pol*, *env*) were extracted with BLAST and codon alignments were made with the HIV-1 reference strain HXB2 using MACSE2 [13].

APOBEC hypermutation filtering

Hypermutation filtering was performed based on a previously published method [9,14]. We adapted this method to the level of single NGS reads, using three different p-value thresholds to determine hypermutation status of a read and subsequent removal, (i) a constant threshold of $p<0.05$, (ii) a liberal dynamic threshold based on the bootstrapped lower 95% confidence interval of the mean from the hypermutation p-value distribution of RNA sequences, randomly selected from the SHCS NGS database at the University Hospital Zurich, for each HIV-1 genome position

(HXB2 as reference), and (iii) a conservative dynamic threshold based on the bootstrapped upper 95% confidence interval of the upper 90% percentile interval of the p-value distribution of RNA sequences for each HIV-1 genome position (HXB2 as reference). Filters and their effect are shown for an example in Supplementary figure 1. After filtering, we generated a new fastq file, reran SmaltAlign, and recalculated the APD.

Average pairwise diversity score

We calculated the APD score as described by [4,5] based on the third codon position of *gag*, *pol*, and *env* individually on the NGS sequence reads and after applying the three different hypermutation filters described above with a coverage threshold of 100 reads for each position.

Time since infection to ART/recent infection analysis

We used linear regression models to determine the fraction (R^2) of the variance of $t_{InfectionToART}$ (time since infection) explained by the APD score calculated on *gag*, *pol*, and *env*. We used ROC curve analysis to determine optimal APD cutoffs for the prediction of recent infection by the APD score calculated on *gag*, *env*, and *pol* separately and in combination for all different hypermutation thresholds. We used two approaches, (i) including all NGS data sets comprising at least 100 codons of the respective gene *env*/*gag*/*pol* (designated as "partial length"), and (ii) full length, including only NGS data sets covering nearly the entire gene, i.e., >95% of codons of the respective gene *gag*/*pol*/*env* (designated as "full length").

RESULTS

We identified 221 PWH with a total of 247 sequences in the SHCS and ZPHI study with an accurate HIV-1 infection date reported and HIV-1 DNA NGS sequences availability. At least one of the three genes had full length in 127 PWH (Figure 1). The median $t_{InfectionToART}$ was 0.41 years (IQR 0.15, 2.27) and the median $t_{ARTtoSampling}$ was 2.29 years (IQR 0.95, 4.46) (Supplementary table 1). We also found an increasing CD4 T cell count from 431 cells/µl (IQR 300, 627) at ART initiation to 636 (IQR 505, 852) at the NGS sample date and a respective decrease for HIV RNA viral load from 18,000 copies/ml (IQR 26, 146,801) to undetectable (IQR 0, 0).

We found significant associations of APD with $t_{InfectionToART}$, but not with $t_{ARTtoSampling}$ (Supplementary table 2, Supplementary figures 2/3). Depending on the gene considered and the hypermutation-filtering threshold used, APD explained between 5% and 25% of the variance in $t_{InfectionToART}$ (quantified as the R^2 in a linear regression model, Figure 2B/C), with the best performance (R^2 = 25%) obtained for *pol* full length and the dynamic conservative threshold. By contrast, APD explained only between 1% and 6% of the variance of $t_{ARTtoSampling}$,(Supplementary figure 4,5,6). Overall, across genes, hypermutation filter increases the R^2 of $t_{InfectionToART}$, in particular for *pol* full length and *env*. For *gag*, however, R^2 is highest without any filtering (Figure 2C). When assessing the ability of APD to predict $t_{InfectionToART}$ in leave-one-out cross validation,

we found the lowest mean absolute error (MAE) in predicting $t_{InfectionToART}$ by *pol* with dynamic conservative threshold and full length (MAE of 1.19 years). Whereas the MAE was highest for *env* (MAE of 2.19 years), with dynamic liberal threshold and full length (Supplementary table 3).

When testing the ability of APD to predict whether ART was initiated in recent infection (<1 year), we obtained areas under the ROC curve (AUC_ROC) ranging from 0.7 (0.62-0.78) for *env* without hypermutation filtering and partial length to 0.85 (0.73-0.96) for *env* 0.05 and dynamic conservative threshold and full length. We found improvements of AUCs with stricter hypermutation filtering thresholds (Figure 2C, Supplementary figure 7/8). For *gag* APD the AUC peak is reached with the 0.05 and dynamic liberal threshold (0.82, 0.72-0.92) whereas for *pol* and *env* APD the AUC peak is reached with the conservative dynamic (and 0.05) threshold, 0.84 (0.75-0.93) and 0.85 (0.73-0.96) respectively (Figure 2D).

DISCUSSION

In this work we showed that a diversity score derived from proviral DNA HIV-1 NGS sequences from individuals on suppressive ART is associated with the time since infection ($t_{InfectionToART}$) and recent infection status. Its predictive accuracy is lower than that of viral diversity derived from plasma HIV-1 RNA [4], in particular when partial sequences were included. However, when restricting the analysis to full-length sequences and hypermutation filtering, predictive performances are in the range of what is achieved with treatment-naive plasma RNA for *pol*/*env* (AUC of 0.84/0.85 for proviral DNA compared to >=0.95 for viral RNA). For *gag*, hypermutation filtering showed no improvements which may be explained by the lower G→A substitution rates in gag [15]. The performance increase comparing partial *pol* to the entire *pol* gene is striking (Figure 2C). This may be explained by absence of the *pol* positions 3000 to 4000 in almost 50% of sequences (Supplementary Figure 9), which previously were shown to have the highest predictability for time since infection [5]. Finally, we show that the APD only has minor associations with $t_{ARTtoSampling}$, confirming our assumption and previous evidence [7] that there is almost no viral evolution under suppressive ART.

The main limitation of this work is the small number of recovered gene sequences, which is most likely due to low reservoir sizes in early treated PWH [1]. It may also be because of the low specificity from the hypermutation filtering and subsequent failure of NGS assembly due to a lack of reads. Another limitation is the between sequence overlap in partial length sequences which may impact comparability of APDs inferred from different regions within a gene. Further, we could not identify an overall optimal hypermutation filtering threshold across all genes. Nevertheless, we show improvements of both the explained variance and AUC with hypermutation filtering compared to not filtering at all.

In summary, this work shows the utility of APDs derived from proviral sequences as a proxy for the time since infection and recent infection prediction. This may be useful for PWH without a

baseline drug resistance test to decide on treatment simplification strategies in clinical practice or to determine infection recency in HIV research, for example, to retrospectively estimate HIV-1 incidence.

Author contributions

Conceptualization: M.Z., H.F.G., K.J.M., and R.D.K; Data curation: M.Z., P.F., M.H., K.N., C.R., J.N., A.R., M.S., M.C., E.B., D.L.B., H.F.G., K.J.M., and R.D.K.; Formal analysis: M.Z.; Funding acquisition: R.D.K. and H.F.G.; Investigation: M.Z., and R.D.K. Methodology: M.Z., C.D.S., and R.D.K.; Project administration: R.D.K.; Resources: M.Z., P.F., M.H., C.D.S., K.N., C.R., J.N., A.R., M.S., M.C., E.B., D.L.B., H.F.G., K.J.M., and R.D.K.; Supervision: H.F.G., K.J.M., and R.D.K.; Validation: R.D.K.; Visualization: M.Z.; Writing – original draft: M.Z., H.F.G., K.J.M., and R.D.K.; Writing - review & editing: M.Z., P.F., M.H., C.D.S., K.N., C.R., J.N., A.R., M.S., M.C., E.B., D.L.B., H.F.G., K.J.M., and R.D.K.

Funding

This work was supported by the Swiss National Science Foundation [grant number 179571] (to H. F. G.), the Yvonne-Jacob Foundation (to H.F.G.), and the University of Zurich's Clinical Research Priority Program Viral Infectious Diseases: Zurich Primary HIV Infection Study (to H. F. G. and D. L. B). R. K. was supported by the Swiss National Science Foundation [grant numbers PZ00P3-14241, BSSGI0_155851].

Acknowledgements

The authors thank the patients who participated in the Swiss HIV Cohort Study; the physicians and study nurses for the excellent patient care provided to participants; Jan Meier, Yves Schäfer, and Océane Follonier from the Swiss HIV Cohort Study data center for data management; and Danièle Perraudin and Marianne Amstad for administration. We also want to thank Alexandra Calmy from the University Hospital Geneva.

Conflicts of interest

A.R. received research grants from Gilead, paid to his institution; travel expenses from Gilead and Pfizer, paid to his institution; and honoraria for data safety monitoring board or advisory board consultations from MSD and Moderna, paid to his institution. D.L.B. received personal consulting fees from Gilead, MSD, and ViiV; personal honoraria for presentations from Gilead, Pfizer, MSD, and ViiV; and received travel expenses from Gilead and ViiV, paid to his institution. E.B. received research grants from MSD, paid to his institution; consulting fees from Moderna, paid to his institution; honoraria for presentations from Pfizer, paid to his institution; travel expenses from ViiV, MSD, Gilead, and Pfizer, paid to his institution; and honoraria for data safety monitoring board or advisory board consultations from ViiV, MSD, Pfizer, Gilead, Moderna, AstraZeneca, AbbVie, and Ely Lilly, paid to his institution. H. F. G. has received research grants from the Swiss

National Science Foundation, Swiss HIV Cohort Study, Yvonne Jacob Foundation, Gilead, ViiV, and Bill and Melinda Gates foundation, paid to his institution; personal honoraria for data safety monitoring board or advisory board consultations from Merck, ViiV healthcare, Gilead Sciences, Janssen, Johnson and Johnson, Novartis, and GSK; and personal travel expenses from Gilead. J.N. received research grants from the Swiss HIV Cohort study and the cantonal hospital St. Gallen, paid to her institution. K.J.M. received unrestricted research grants from Gilead and Novartis, paid to her institution; and personal honoraria for advisory board consultations from ViiV. M.C. received research grants from Gilead, ViiV, and MSD, paid to his institution; payment for expert testimony from Gilead, ViiV, and MSD, paid to his institution; and travel expenses from Gilead, paid to his institution. M.S. received honoraria for data safety monitoring board advisory board consultations from Gilead, ViiV, Moderna, Pfizer, and MSD, paid to his institution; and travel expenses for conferences from Gilead, paid to his institution. P.F. received personal travel expenses from the University Zurich; payment for equipment from the University Zurich; and personal honoraria for presentations from the University Zurich.

R.D.K. received research grants from Gilead and NIH, paid to his institution. All other authors report no potential conflicts.

This work is accepted for a poster presentation at CROI 2024.

Data availability

The individual level datasets generated or analyzed during the current study do not fulfill the requirements for open data access:

1) The SHCS informed consent states that sharing data outside the SHCS network is only permitted for specific studies on HIV infection and its complications, and to researchers who have signed an agreement detailing the use of the data and biological samples; and
2) the data is too dense and comprehensive to preserve patient privacy in persons living with HIV.

According to the Swiss law, data cannot be shared if data subjects have not agreed or data is too sensitive to share. Investigators with a request for selected data should send a proposal to the respective SHCS address (www.shcs.ch/contact). The provision of data will be considered by the Scientific Board of the SHCS and the study team and is subject to Swiss legal and ethical regulations, and is outlined in a material and data transfer agreement.

References

1. West E, Zeeb M, Grube C, Kuster H, Wanner K, Scheier T, et al. Sustained Viral Suppression With Dolutegravir Monotherapy Over 192 Weeks in Patients Starting Combination Antiretroviral Therapy During Primary Human Immunodeficiency Virus Infection (EARLY-SIMPLIFIED): A

Randomized, Controlled, Multi-site, Noninferiority Trial. Clin Infect Dis. 2023 Oct 5;77(7):1012–20.
2. Landais E, Moore PL. Development of broadly neutralizing antibodies in HIV-1 infected elite neutralizers. Retrovirology. 2018 Dec 5;15(1):61.
3. Kouyos RD, von Wyl V, Yerly S, Böni 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 Feb 15;52(4):532–9.
4. 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 Jun;220(2):254–65.
5. Puller V, Neher R, Albert J. Estimating time of HIV-1 infection from next-generation sequence diversity. PLoS Comput Biol. 2017 Oct 2;13(10):e1005775.
6. Ellis KE, Nawas GT, Chan C, York L, Fisher J, Connick E, et al. Clinical Outcomes Following the Use of Archived Proviral HIV-1 DNA Genotype to Guide Antiretroviral Therapy Adjustment. Open Forum Infect Dis. 2020 Jan 1;7(1).
7. Jörimann L, Tschumi J, Zeeb M, Leemann C, Schenkel CD, Neumann K, et al. Absence of proviral HIV-1 evolution in early treated individuals with HIV switching to dolutegravir monotherapy during 48 weeks. J Infect Dis. 2023 Jul 27;
8. van Zyl G, Bale MJ, Kearney MF. HIV evolution and diversity in ART-treated patients. Retrovirology. 2018 Dec 30;15(1):14.
9. Tzou PL, Kosakovsky Pond SL, Avila-Rios S, Holmes SP, Kantor R, Shafer RW. Analysis of unusual and signature APOBEC-mutations in HIV-1 pol next-generation sequences. PLoS One. 2020 Feb 26;15(2):e0225352.
10. Scherrer AU, Traytel A, Braun DL, Calmy A, Battegay M, Cavassini M, et al. Cohort Profile Update: The Swiss HIV Cohort Study (SHCS). Int J Epidemiol. 2022 Feb 1;51(1):33–34j.
11. Rieder P, Joos B, Scherrer AU, Kuster H, Braun D, Grube C, et al. Characterization of Human Immunodeficiency Virus Type 1 (HIV-1) Diversity and Tropism in 145 Patients With Primary HIV-1 Infection. Clin Infect Dis. 2011 Dec 15;53(12):1271–9.
12. Jaha B, Schenkel CD, Jörimann L, Huber M, Zaheri M, Neumann K, et al. Prevalence of HIV-1 drug resistance mutations in proviral DNA in the Swiss HIV Cohort Study, a retrospective study from 1995 to 2018. J Antimicrob Chemother. 2023 Sep 5;78(9):2323–34.
13. Ranwez V, Douzery EJP, Cambon C, Chantret N, Delsuc F. MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. Mol Biol Evol. 2018 Oct 1;35(10):2582–4.
14. Rose PP, Korber BT. Detecting hypermutations in viral sequences with an emphasis on G A hypermutation. Bioinformatics. 2000 Apr 1;16(4):400–1.
15. Kijak GH, Janini LM, Tovanabutra S, Sanders-Buell E, Arroyo MA, Robb ML, et al. Variable contexts and levels of hypermutation in HIV-1 proviral genomes recovered from primary peripheral blood mononuclear cells. Virology. 2008 Jun;376(1):101–11.

FIGURE LEGENDS

Figure 1 Flowchart of PWH selection and availability of HIV-1 genome sequences. APD, average pairwise diversity score; ART, anti-retroviral therapy; PWH, people with HIV-1; SHCS, Swiss HIV Cohort Study; ZPHI, Zurich Primary HIV Infection study.

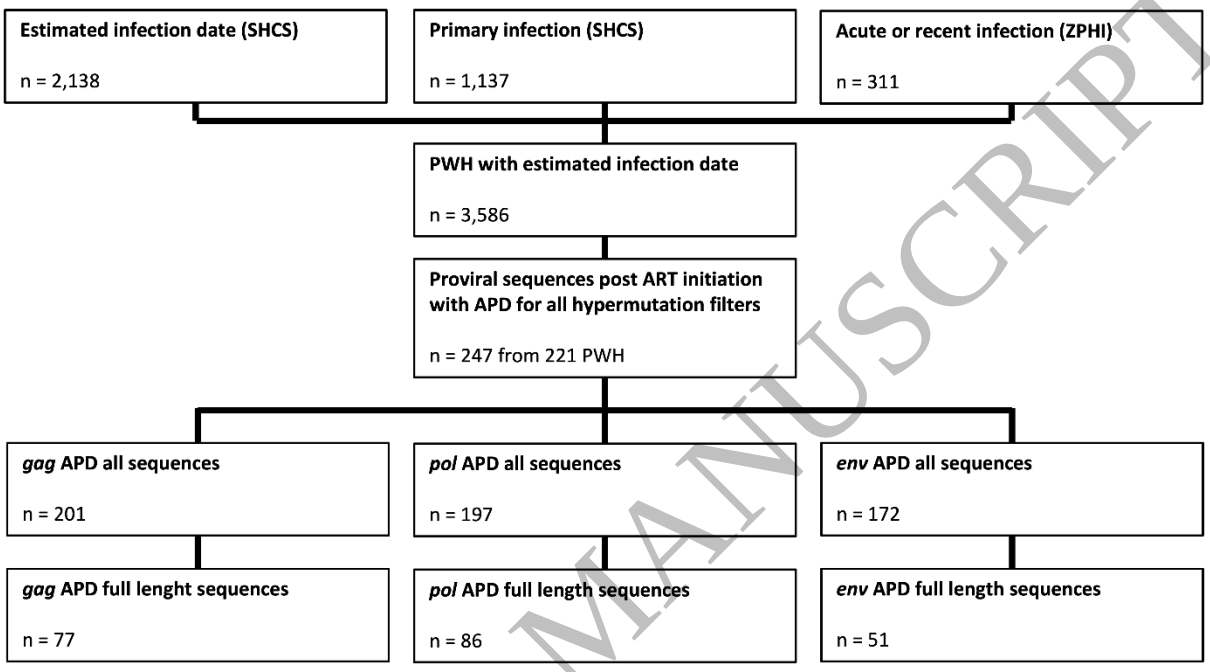
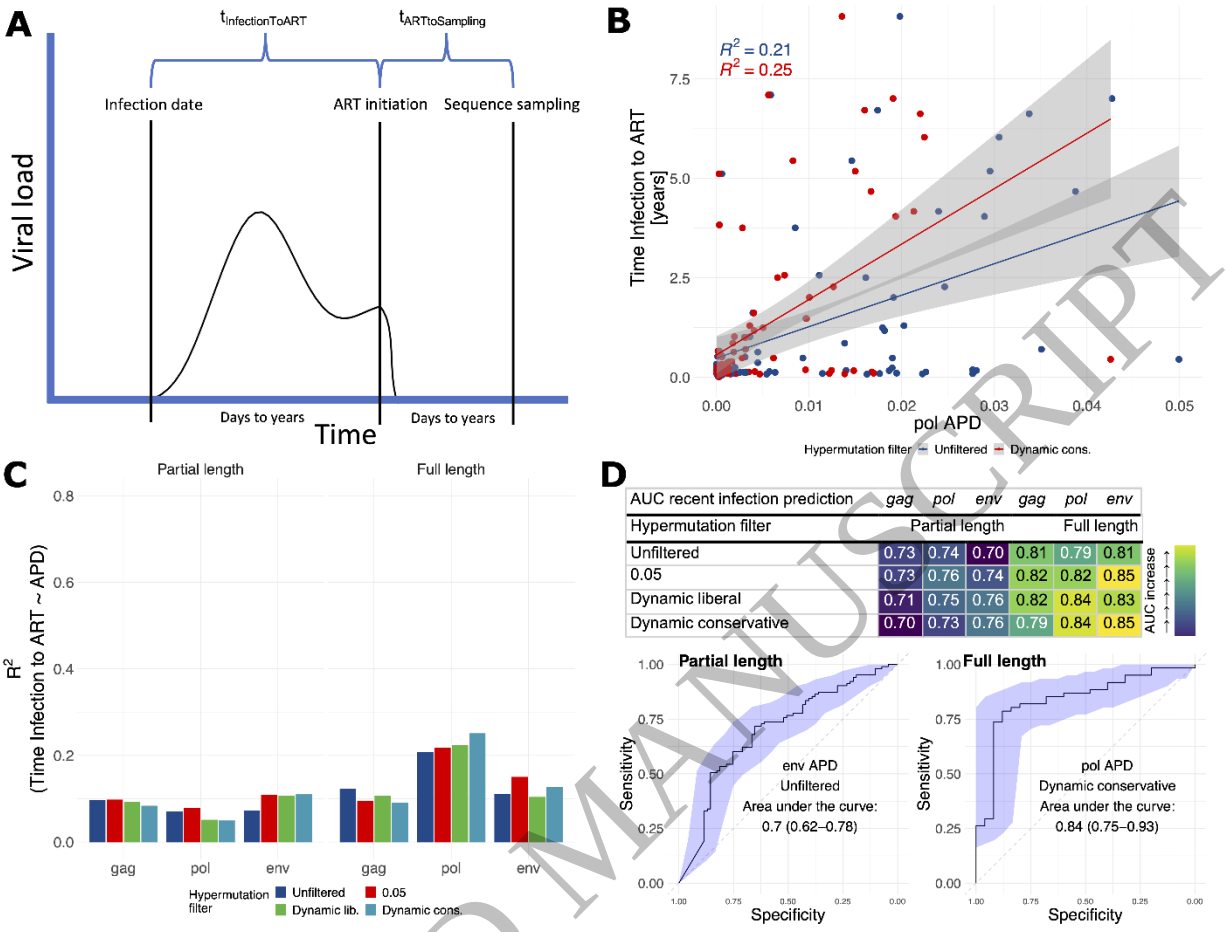


Figure 2 Time since HIV-1 infection to ART initiation prediction with proviral genetic diversity from NGS sequencing A: Illustration of the HIV-infection course and definitions of time since infection to ART initiation and time since ART initiation to proviral NGS sampling B: Time infection to ART in dependence of APD derived from full length pol sequences. C: R^2, the goodness of fit calculated as the explained variation in time infection to ART by APD, of linear regression from time infection to ART in dependence of APD derived from partial length and restricted to full length gag/pol/env sequences D: AUCs and ROC curves for the prediction of time infection to ART <1 year (recent infection status) with APDs derived from partial length and restricted to full length env/pol/gag sequences. AUCs with 95% confidence intervals are shown in Supplementary table 4. All other ROC curves for other hypermutation filters and genes are shown in supplementary figures 7/8. B-D: Analyses were repeated for different levels of hypermutation filtering: (i) hypermutation unfiltered, (ii) 0.05 threshold, (iii) dynamic liberal threshold, and (iv) dynamic conservative threshold (visualized at an example in supplementary figure 1). APD, average pairwise diversity score; ART, anti-retroviral therapy; AUC, area under the curve; ROC, receiver operating characteristic.



DOI: 10.1093/infdis/jiae149