Phylogenetic approach to recover integration dates of latent HIV sequences within-host

Bradley R. Jones[a], Natalie N. Kinloch[b], Joshua Horacsek[a], Bruce Ganase[a], Marianne Harris[a], P. Richard Harrigan[c], R. Brad Jones[d], Mark A. Brockman[a,b], Jeffrey B. Joy[a,c,1,2], Art F. Y. Poon[e,1,2], and Zabrina L. Brumme[a,b,1,2]

[a]British Columbia Centre for Excellence in HIV/AIDS, Vancouver, BC, Canada V6Z 1Y6; [b]Faculty of Health Sciences, Simon Fraser University, Burnaby, BC, Canada V5A 1S6; [c]Department of Medicine, University of British Columbia, Vancouver, BC, Canada V5Z 1M9; [d]Department of Microbiology, Immunology and Tropical Medicine, George Washington University, Washington, DC 20037; and [e]Department of Pathology and Laboratory Medicine, University of Western Ontario, London, ON, Canada N6A 5C1

Edited by Robert F. Siliciano, Johns Hopkins University School of Medicine, Baltimore, MD, and approved August 9, 2018 (received for review February 2, 2018)

Given that HIV evolution and latent reservoir establishment occur continually within-host, and that latently infected cells can persist long-term, the HIV reservoir should comprise a genetically heterogeneous archive recapitulating within-host HIV evolution. However, this has yet to be conclusively demonstrated, in part due to the challenges of reconstructing within-host reservoir establishment dynamics over long timescales. We developed a phylogenetic framework to reconstruct the integration dates of individual latent HIV lineages. The framework first involves inference and rooting of a maximum-likelihood phylogeny relating plasma HIV RNA sequences serially sampled before the initiation of suppressive antiretroviral therapy, along with putative latent sequences sampled thereafter. A linear model relating root-to-tip distances of plasma HIV RNA sequences to their sampling dates is used to convert root-to-tip distances of putative latent lineages to their establishment (integration) dates. Reconstruction of the ages of putative latent sequences sampled from chronically HIV-infected individuals up to 10 y following initiation of suppressive therapy revealed a genetically heterogeneous reservoir that recapitulated HIV's within-host evolutionary history. Reservoir sequences were interspersed throughout multiple within-host lineages, with the oldest dating to >20 y before sampling; historic genetic bottleneck events were also recorded therein. Notably, plasma HIV RNA sequences isolated from a viremia blip in an individual receiving otherwise suppressive therapy were highly genetically diverse and spanned a 20-y age range, suggestive of spontaneous in vivo HIV reactivation from a large latently infected cell pool. Our framework for reservoir dating provides a potentially powerful addition to the HIV persistence research toolkit.

HIV | latency | reservoir | phylogenetics | evolution

HIV, like all retroviruses, integrates its genome into that of the infected host cell. Although actively infected cells typically die as a result of viral cytopathic effects or immune-mediated elimination, a minority [broadly estimated as one in every million resting CD4+ T cells (1, 2)] harbor integrated HIV DNA in a state of reduced transcriptional activity (or quiescence) over long periods (3–5). Termed latently infected cells or latent HIV reservoirs, these represent the major barrier to a cure as they can persist for years and can reactivate at any time to produce infectious virions (5–10). It is for this reason that combination antiretroviral therapies (cART), which do not act upon latent HIV reservoirs, need to be maintained for life.

Characterization of latent HIV sequences provides information about reservoir stability, distribution, and dynamics (11), which can in turn inform HIV elimination strategies. It is now established that a majority of sexually acquired HIV infections are initiated by a single transmitted/founder variant, from which descendant viral populations rapidly reaccumulate genetic diversity (12–20). Concomitantly, establishment of latent HIV reservoirs begins within hours or days following infection and continues as long as the virus is actively replicating within the host (21, 22). Given that latent HIV genomes can persist for decades, either in the original cell or clonal descendants thereof (23, 24), the reservoir should constitute a genetically heterogeneous archive of within-host HIV evolution, even after years of cART (25–27). However, while numerous studies have confirmed that the HIV reservoir is genetically diverse (11, 28–35), our knowledge of the within-host ancestor–descendant relationships of these sequences remains limited, because few studies (ref. 36 being a notable exception) have interpreted reservoir diversity in the context of HIV's within-host evolutionary history. In particular, it remains unclear whether HIV sequences sampled from the reservoir during long-term suppressive cART truly recapitulate within-host HIV evolution or whether dynamic processes such as homeostatic proliferation (33), clonal expansion (11, 37–41), and/or preferential elimination of certain latently HIV-infected cells skew reservoir sequence composition over time. These knowledge gaps persist due to the challenges of studying within-host HIV dynamics over long timescales, combined with a lack of methods to infer latent HIV integration dates. While phylogenetic principles have been used to identify viral reservoirs (25), and techniques have been developed to detect latent lineages (42) and to quantify their turnover (43),

Significance

Studies characterizing within-host latent HIV sequence diversity have yielded insight into reservoir dynamics and persistence. Our understanding of these processes, however, can be further enhanced if reservoir diversity is interpreted in context of HIV's within-host evolutionary history. Approaches to infer the original establishment (i.e., integration) dates of individual within-host latent HIV lineages would be particularly useful in this regard. We describe a phylogenetic framework to infer latent HIV ages from viral sequence information and apply it to latent HIV sequences sampled up to 10 y on suppressive therapy to yield insights into HIV reservoir dynamics. The ability to infer within-host latent HIV ages from sequence information has broad potential applications that may advance us toward an HIV cure.

Author contributions: B.R.J., J.B.J., A.F.Y.P., and Z.L.B. designed research; B.R.J., N.N.K., J.H., B.G., M.H., P.R.H., R.B.J., M.A.B., and Z.L.B. performed research; B.R.J., N.N.K., J.H., J.B.J., A.F.Y.P., and Z.L.B. analyzed data; and B.R.J., J.B.J., A.F.Y.P., and Z.L.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. MG822917–MG823179).

[1]J.B.J., A.F.Y.P., and Z.L.B. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: jjoy@cfenet.ubc.ca, apoon42@uwo.ca, or zbrumme@sfu.ca.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1802028115/-/DCSupplemental.

Published online September 5, 2018.

latent HIV deposition times have generally been estimated by assessing genetic similarity of these sequences to pre-cART plasma HIV RNA sequences (e.g., refs. 30 and 36). However, this approach is limited by the ability to previously sample similar sequences within-host. To our knowledge, no studies have directly leveraged information about within-host HIV evolutionary rates to infer latent HIV sequence ages.

As evolution of a given HIV lineage effectively ceases upon integration (and only resumes once a new round of virus replication ensues), latent HIV sequences will display a characteristic discordance between their sampling date and their "true" (older) age based on their genetic divergence from the transmitted/founder virus. Inference of the latter distance, along with a host-specific evolutionary rate, would thus allow a reservoir's "true" age to be inferred from its sequence. Within-host ancestor–descendant relationships can be phylogenetically reconstructed using serially sampled HIV sequences (44, 45); moreover, by "fixing" tree tips at their respective sampling dates, phylogenies can be rescaled chronologically, allowing rates of evolution to be estimated directly (46). As the rate of evolution remains fairly consistent among actively replicating within-host HIV lineages, the divergence of descendant lineages from the transmitted/founder virus can be approximated, at least over the initial years of infection, by a "strict" molecular clock model (36, 47, 48). Based on these principles, we developed a phylogenetic framework that recovers the integration date of latent HIV lineages by reconstructing within-host HIV evolutionary history from virus sequences sampled pre-cART.

Results

Reconstructing Integration Dates Phylogenetically. Phylogenetic inference of HIV evolutionary rates from time-stamped between-host HIV sequence datasets has been used to date key events in the pandemic's history (49, 50), to reconstruct putative transmission histories (44, 51, 52), and to infer unknown sequence ages (53). We adapt these techniques to the within-host context with the goal of estimating latent HIV sequence ages (Fig. 1). We begin by inferring a maximum-likelihood phylogeny (54) from plasma HIV RNA sequences longitudinally sampled pre-cART ("training data") along with sequences whose ages are unknown ("censored data," for example proviral DNA sequences subsequently sampled during suppressive cART) (Fig. 1*A*). We then identify the root location that optimizes the relationship between evolutionary distance and sampling time in the training data (Fig. 1*B*) and calibrate a strict molecular clock by fitting a linear model that relates their root-to-tip evolutionary distances to their sampling times, given by $D = \mu T + a$, where response variable D denotes the genetic distance from the root, predictor variable T denotes the sample collection date, μ denotes the evolutionary rate, and a denotes the y intercept (Fig. 1*C*). We then test for the presence of a molecular clock (i.e., evidence that sequence divergence increases over time) by comparing the model against the null model of zero slope (where D is constant over time). Linear models need to fulfill two criteria to pass quality control: first, their Akaike information criterion (AIC) (55) needs to be at least 10 units lower than that of the null model, and second, the 95% CI of the model-estimated root date needs to contain or precede the first sampling date. For linear models that pass model selection, the establishment (i.e., integration) date of each censored sequence x is estimated by $T(x) = (D(x) - a)/\mu$ (Fig. 1*D*). The uncertainty in $T(x)$ is given by $\varepsilon_x = t_{0.025,n-2}\frac{\varepsilon}{\mu}\sqrt{1+\frac{1}{n}+(D(x)-\bar{D})^2/(\mu^2\sum(T-\bar{T})^2)}$, where $t_{0.025,n-2}$ is derived from the t distribution, ε is the error of the linear model, n is the number of training sequences, and $\bar{D}$ and $\bar{T}$ are the mean genetic distance and collection date of the training sequences, respectively (56).

Framework Proof of Concept. To test the concept, we simulated 1,000 phylogenies with 100 tips each under a strict molecular clock, censored the sampling dates for 50% of sequences at random, and inferred linear models from the remainder (representative dataset in Fig. 2 *A*–*C*). All linear models yielded ΔAIC values above the predefined threshold (mean ΔAIC 187); however, 39/1,000 (4%) returned root date estimates whose 95% CIs were later than the first training time point and were thus rejected, yielding a 96% success rate. Linear models fit the training data very well: the overall scaled mean absolute error (MAE), calculated as the grand mean of the absolute error between predicted and true sampling dates of the training data, scaled by the overall training data timespan, was 0.034 (i.e., 3.4%). That is, for datasets spanning a 1-y period, the linear model recovered training data sampling dates to within an average of 12 d. Dates of censored sequences were similarly recovered with minimal error: The overall scaled mean absolute difference (MAD, calculated as the grand mean of the absolute difference between predicted and true sampling dates of censored sequences, scaled by the total training data timespan) was 0.037, with no evidence of distributional asymmetry using a nonparametric binomial test (57) (all 961 simulations returned *P* values below the significance cutoff; see *Materials and Methods*) (Fig. 2*D*).

To test the framework's performance on real HIV sequence datasets sampled at various depths, frequencies, and timespans, we applied it to 424 longitudinal partial HIV RNA *env* sequences

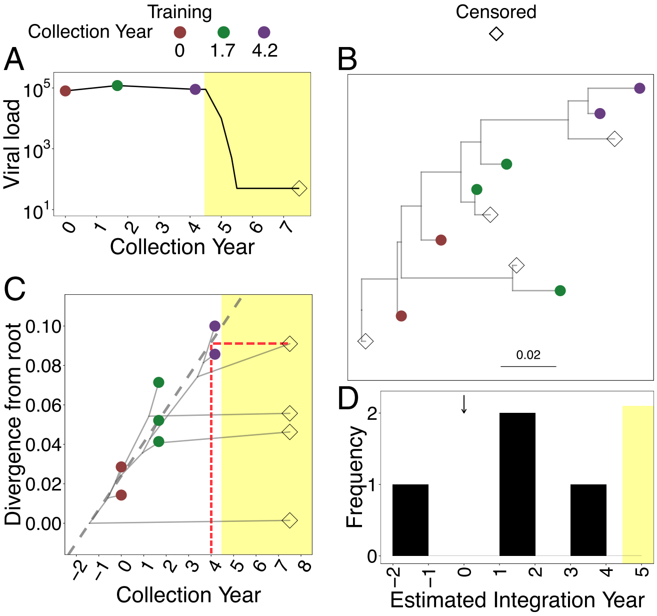

Fig. 1. Framework illustration. (*A*) Hypothetical pVL and sampling history of an HIV-infected individual who initiated cART in chronic infection. Throughout all figures, circles denote plasma HIV RNA, diamonds denote HIV DNA, filled symbols denote training data (plasma HIV RNA sequences used for model calibration), and open symbols denote censored data (sequences destined for molecular dating). Training data are colored based on collection date, while censored data are shown in black. Yellow shading denotes cART. Here, plasma HIV RNA sequences collected at baseline and 1.7 and 4.2 y (training data; filled colored circles) are used to infer integration dates of proviral DNA sequences sampled during suppressive cART in year 7 (censored data; open black diamond). (*B*) Maximum-likelihood within-host phylogeny relating training and censored sequences, where the root represents the inferred MRCA (i.e., the date of the transmitted/founder event). Scale in nucleotide substitutions per site. (*C*) The thick gray dotted diagonal represents the linear model relating root-to-tip distances of the training data to their sampling dates. The *x* intercept (here, 1 y before baseline sampling) represents the inferred root date. The linear model is used to convert root-to-tip distances of censored sequences to their establishment (i.e., integration) dates. For example, the latent sequence at the top right, whose divergence from the root is 0.09, is inferred to have integrated at the beginning of year 4 (dotted red line). Light gray lines trace the ancestor–descendant relationships of HIV lineages. (*D*) Histogram summarizing inferred integration dates of censored sequences. Arrow denotes baseline sampling.

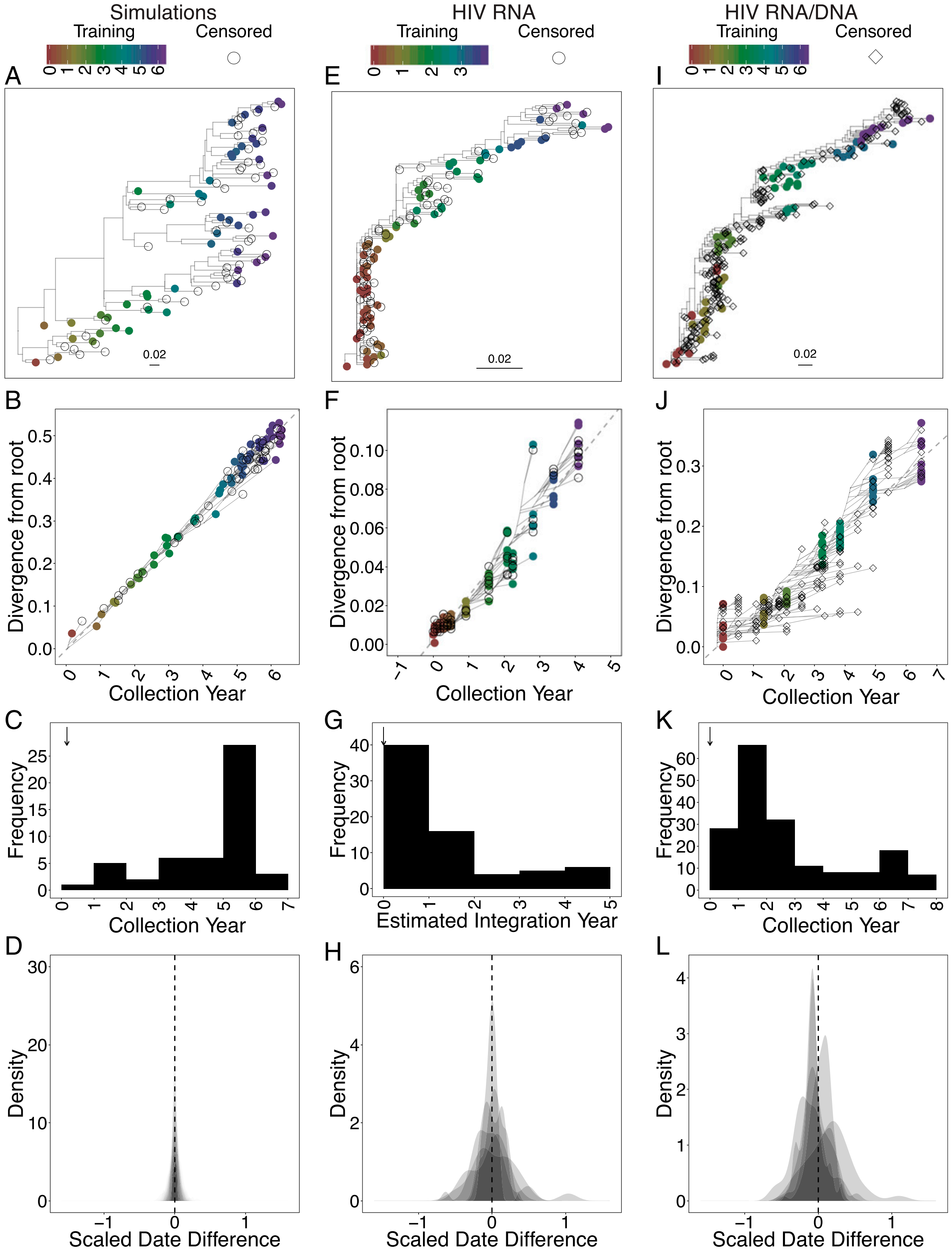

Fig. 2. Framework proof of concept using simulated and published HIV sequences. (*A*) Representative rooted tree relating simulated longitudinal within-host plasma HIV RNA sequences with 50% of tips randomly assigned as training data (circles colored by sampling time point) or censored for molecular dating (open black circles). (*B*) Resulting linear model with ancestor traces overlaid. (*C*) Inferred dates of censored sequences; arrow indicates baseline sampling date. (*D*) Density plots of normalized error distributions (expressed as the absolute difference between predicted and true sampling dates of the censored sequences, scaled by the total dataset timespan, where −1 and 1 represent −100% and 100%, respectively) from 100 (of 961) successful simulations selected at random. (*E*–*G*) Same as *A*–*C*, but for within-host plasma HIV RNA sequences from LANL participant 13654 where 50% of tips were randomly assigned as training data. (*H*) Density plots of normalized error distributions for all six successful LANL RNA datasets. (*I*–*K*) Same as *A*–*C*, but for LANL participant 821 with HIV RNA and DNA sequences treated as training and censored, respectively. (*L*) Density plots of normalized differences between HIV DNA predicted and sampling dates for the six successful LANL RNA/DNA datasets.

retrieved from the Los Alamos National Laboratory HIV sequence database (LANL) (58) from eight individuals with known infection dates (median 40 sequences per individual, collected at a median of 4.5 time points over a median 2.2 y) whose within-host sequence diversity was consistent with a molecular clock (*SI Appendix*, Table S1). After censoring 50% of sequences at random from each dataset, linear models could still be calibrated for six of them (representative data shown in Fig. 2 *E*–*G*; summary in *SI Appendix*, Table S1). In part because real datasets were on average smaller and spanned shorter time periods than simulated data, scaled prediction errors were higher. Nevertheless, the model predicted censored dates to within an average of 15% (SD 6.3%) of the total timespan of each dataset, where prediction errors reflected overall model fit to the training data (Pearson's correlation between model MAE and MAD was $R =$ 0.98, P = 0.007), with no significant distributional asymmetry (Fig. 2*H* and *SI Appendix*, Table S1). These observations demonstrate that the framework can be applied to sparse data (e.g., for individual 13333, a linear model was calibrated from only 19 training sequences sampled over 1.4 y).

We then applied the framework to seven published datasets (58) featuring plasma RNA and proviral DNA sequences sampled at various times pre- and post-treatment (median 142 sequences per individual collected at a median 11 time points over a median 5.8 y), although none featured HIV DNA sampled during long-term suppressive cART (*SI Appendix*, Table S1). Thus, while the majority of HIV DNA sequences should represent contemporary within-host strains, we hypothesized that a minority would represent latent genomes (59). To test this, plasma HIV RNA sequences were used as training data to reconstruct proviral DNA archival dates, with the expectation that datasets containing latent genomes would yield left-skewed date difference distributions due to latent genome archival dates preceding their sampling dates. Of the seven datasets, six (including three with only three training time points) yielded successful linear models. (Representative data are shown in Fig. 2 *I*–*K* and a summary is given in *SI Appendix*, Table S1.) Overall, HIV DNA dates were predicted to within an average of 16% (SD 5.9%) of the total training data timespan where MAD again correlated strongly with overall model MAE (Pearson's $R = 0.98$, $P = 0.00048$; *SI Appendix*, Table S1). Moreover, significantly left-skewed date difference distributions were noted in two individuals, consistent with the presence of latent lineages (*SI Appendix*, Fig. S1 and Table S1). These results demonstrate that HIV sequence ages can be reconstructed from a wide range of datasets from individuals with various clinical histories and that the framework can detect the expected presence of latent lineages.

Dating the Latent HIV Reservoir. As the framework is designed to date HIV sequences within (or released from) HIV reservoirs, we applied it to reconstruct integration dates of HIV sequences isolated from two individuals who had maintained prolonged plasma viral load (pVL) suppression on cART. Participant 1 was diagnosed with HIV in August 1996 and did not initiate a regimen that suppressed pVL to <50 copies HIV RNA per mL until August 2006 (Fig. 3*A*). Viremia rebounded in the fall of 2007 after an unsuccessful regimen change, but suppression was reachieved in April 2008 and sustained to the present day with the exception of one "blip" to 76 copies per mL in September 2015. Using single-genome amplification, we collected 102 HIV RNA *nef* sequences from 14 pre-cART plasma specimens spanning August 1996 to June 2006 as training data. We also collected 42 *nef* sequences sampled at four time points post-cART for molecular dating; these included proviral DNA sequences retrieved from whole blood collected in July 2011 and peripheral blood mononuclear cells (PBMC) in June 2016 (5 and 10 y post-cART, respectively) and HIV RNA sequences sampled during the September 2007 pVL rebound and the September 2015 viremia blip. These data included eight instances where identical HIV RNA sequences were isolated from the same or temporally adjacent plasma samples and eight instances where identical sequences were isolated from putative reservoirs (including two instances where the same sequence was isolated from the plasma viremia event in 2007 and PBMC sampled in 2016) (*SI Appendix*, Fig. S2*A*).

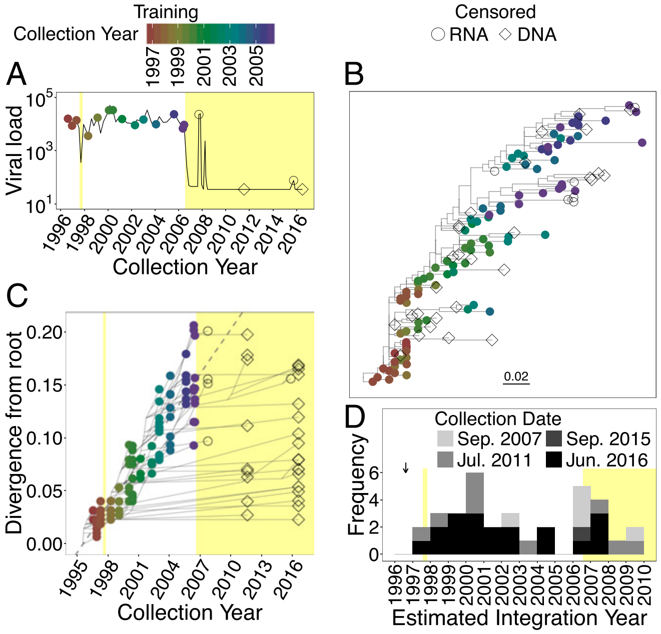

Fig. 3. Reservoir dating: participant 1. (*A*) Plasma HIV RNA sequences from 14 pre-cART time points spanning August 1996 to June 2006 were used as training data (colored circles) to infer the integration dates of censored sequences sampled at four time points between 2007 and 2016, including proviral DNA sequences sampled in 2011 and 2016 (open black diamonds) and plasma HIV RNA sequences from viremic episodes in 2007 and 2015 (open black circles). Yellow shading denotes cART. (*B*) Rooted tree relating training and censored sequences. (*C*) Linear model (gray dotted diagonal) with ancestor traces overlaid. (*D*) Inferred integration dates of censored sequences, colored by sampling date. Arrow denotes baseline sampling date.

As identical HIV sequences are not phylogenetically informative, a rooted maximum-likelihood phylogeny was inferred from 93 unique HIV RNA (training data) and 34 unique putative reservoir sequences (censored data). The characteristic genetic bottlenecks that typify within-host HIV evolution (15, 16, 20) are clearly visible in the "ladder-like" tree shape (48, 60) (Fig. 3*B*) and in the temporally ordered plasma HIV RNA *nef* amino acid alignments (*SI Appendix*, Fig. S2*B*). Training data root-to-tip patristic distances also correlated strongly with sampling times (Pearson $R = 0.92$, $P < 9.8 \times 10^{-38}$), indicative of a molecular clock. Consistent with the reservoir as a genetically diverse archive of within-host HIV evolution [where it would be expected that reservoir sequences would be dispersed throughout a phylogeny of viruses sampled over time from an individual (25)], censored sequences were embedded within multiple within-host lineages and exhibited overall diversity comparable to that of pre-cART plasma HIV RNA sequences sampled over a decade (mean patristic distances of 0.12 vs. 0.095 expected substitutions per base, respectively). The linear model fit the training data well, particularly in the early years (overall ΔAIC = 172; MAE = 1.1 y), yielding an estimated *nef* mutation rate of 3.9×10^{-5} substitutions per base per d and an estimated root date of August 1995 (Fig. 3*C*). Evolutionary rate and root date reconstructions were confirmed using established Bayesian methods (grand mean rate 4.1×10^{-5} [95% highest posterior density (HPD) 3.1×10^{-5}–5.2×10^{-5}] substitutions per base per d; grand mean root date December 1995 [95% HPD July 1995–May 1996]).

Further supporting the reservoir as a within-host HIV evolutionary archive, unique proviral DNA sequences sampled in 2016, after nearly 10 y of suppressive cART, dated to between

1997 and 2007, making them an average of 14 y old (Fig. 3*D*). Moreover, consistent with multiple ancestral HIV lineages contributing to viremia recrudescence upon therapy discontinuation, four of the five HIV RNA sequences isolated from the 2007 viremia rebound dated to distinct pre-cART time points (one to January 2002 and the others to early 2006). The single HIV RNA sequence isolated from the 2015 plasma viremia blip dated to August 2006, around the time of cART initiation, suggesting the blip originated from spontaneous reactivation of a latent HIV lineage seeded just before therapy. While the linear model returned 2009 integration dates for two sequences, in both cases their 95% CIs overlapped with the last period of uncontrolled viremia (*SI Appendix*, Fig. S3). Overall these results suggest that HIV evolution was not ongoing during suppressive cART in this participant, at least not in blood.

Participant 2 was diagnosed with HIV in April 1995 (Fig. 4*A*). In July 2000 he initiated an incompletely suppressive dual ART regimen that was maintained until August 2006, when he initiated cART. Plasma viremia was suppressed to <50 copies per mL by December 2006 and maintained until May 2011, after which intermittent viremia blips occurred, the highest of which was in March 2013 (1,063 copies per mL). We collected 47 plasma HIV RNA *nef* sequences from four pre-ART time points between February 1997 to December 1999 and 100 plasma *nef* sequences from 12 time points between April 2001 and August 2006 while on dual ART, for use as training data. An additional 30 *nef* sequences sampled up to 10 y post-cART, including plasma sequences from the March 2013 viremia blip and PBMC-derived HIV DNA sequences sampled in August 2016, were isolated for molecular dating. We noted 16 cases where identical HIV sequences were isolated from the same or different time points, including one case where a sequence isolated from the 2013 plasma viremia blip exactly matched one isolated from plasma HIV RNA in 2005 (*SI Appendix*, Fig. S4*A*).

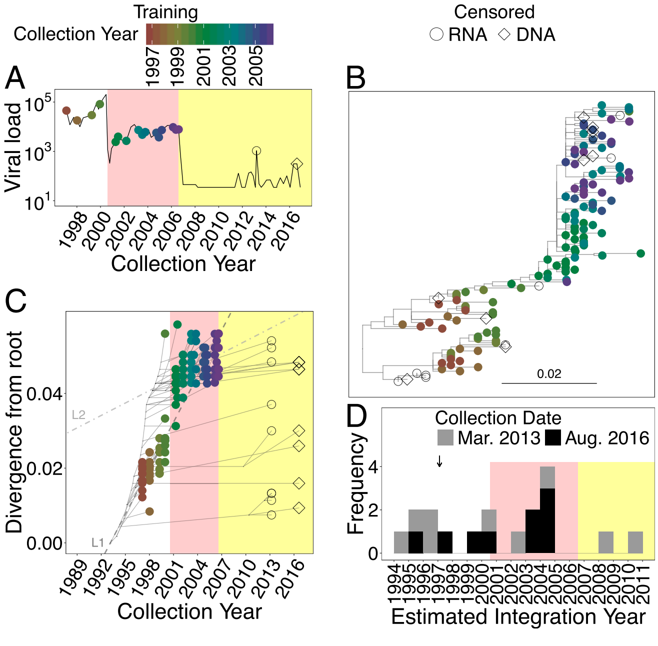

Fig. 4. Reservoir dating: participant 2. (*A*) Plasma HIV RNA sequences from four pre-ART time points between February 1997 and December 1999, and an additional 12 time points between April 2001 and August 2006 during incompletely suppressive dual ART (circles colored by sampling time point) were used as training data to infer the integration dates of censored sequences sampled 7 and 10 y post-cART, including HIV RNA sequences from a viremia episode in 2013 (black open circles) and proviral DNA sampled in 2016 (black open diamonds). Pink and yellow shading denote dual ART and cART, respectively. (*B*) Rooted tree relating training and censored sequences. (*C*) Linear models for the pre-ART period (thick gray diagonal, L1) and dual ART period (hatched gray diagonal, L2), with HIV ancestor traces overlaid. Censored sequences are dated using the model inferred from the lineages in which they reside. (*D*) Inferred integration dates of censored sequences, colored by sampling date. Arrow denotes baseline sampling.

The phylogeny inferred from 119 unique plasma HIV RNA and 18 putative reservoir sequences revealed two initial viral lineages that underwent a severe genetic bottleneck following dual ART (Fig. 4*B* and *SI Appendix*, Fig. S4*B*). Again, training data root-to-tip distances correlated strongly with sampling times (e.g., Pearson $R = 0.61$, $P = 3.4 \times 10^{-5}$ for pre-ART period) and putative reservoir sequences were dispersed throughout all lineages. These included one ancestral subclade branching close to the root that included five unique sequences isolated from both reservoir samplings, whose most recent common ancestor (MRCA) gave rise to the clade that disappeared from circulation after dual ART. Two linear models were trained, one for the ART-naive period (ΔAIC = 16; MAE = 0.85 y) which yielded an estimated mutation rate of 1.1×10^{-5} substitutions per base per d and an estimated root date of February 1993 and the second for the dual-therapy period (ΔAIC = 14; MAE = 3.0 y; estimated mutation rate 3.0×10^{-6} substitutions per base per d) (Fig. 4*C*). Evolutionary rate and root date estimates were again consistent with those estimated using Bayesian methods (grand mean rate 8.8×10^{-6} [95% HPD 6.8×10^{-6}–1.1×10^{-5}] substitutions per base per d; grand mean root date February 1994 [95% HPD January 1993–March 1994]).

Reservoir sequences were dated using the linear model inferred from the lineages in which they resided, yielding integration dates that preceded collection dates by up to 21 y (Fig. 4*D*). Two observations are notable. First, the five sequences in the most ancestral subclade, which included sequences sampled from the 2013 viremia blip as well as from proviral DNA in 2016, dated to between December 1994 and May 1996, before the participant's first plasma sampling. Detection of archival sequences in viremia blips is consistent with the reservoir harboring reactivation-competent viral lineages that can date back to transmission. Second, sequences from the 2013 viremia blip were genetically diverse and dated to between 1994 and 2010, suggesting that spontaneous in vivo reactivation of numerous ancestral HIV lineages led to this viremia episode.

Framework Is Robust to Minimal Training Data. Training linear models from ∼100 HIV sequences sampled over a decade is not feasible in most cases, so we tested model robustness to sampling depth and frequency. We began by evaluating the framework's ability to retrieve known plasma HIV RNA dates using progressively sparser training data ("RNA censoring validation"). To do this, we randomly subsampled training time points from participant 1's data to generate 1,096 new training datasets (featuring all 91 combinations comprising two training time points, 100 datasets each comprising between 3 and 11 training time points, all 91 combinations comprising 12 training time points and all 14 combinations comprising 13 training time points), where the remainder were censored. As all datasets contained the same sequences, a single phylogeny was inferred where the root location (and thus the resulting linear model) differed for each dataset, depending on which sequences were designated "training" vs. "censored." Overall MAE distributions (here, expressed as the difference between known and model-predicted sampling dates for all sequences) as well as their respective concordance coefficients (61) are shown in Fig. 5 *A*–*C*. To provide context on maximal model performance, the full (14 time points) dataset yielded an overall MAE of 1.3 y (Fig. 5*B*) and a concordance coefficient of 0.90 (Fig. 5*C*).

These results demonstrated that the framework could be reliably applied to as few as three training data time points, and sometimes even two. For example, 86% of linear models trained on three time points passed calibration (Fig. 5*A*) and yielded a median MAE only 14% more than the full dataset (1.4 vs. 1.3 y, respectively; Fig. 5*B*) and a median concordance coefficient only 4.4% lower than the full dataset (0.86 vs. 0.90 respectively; Fig.

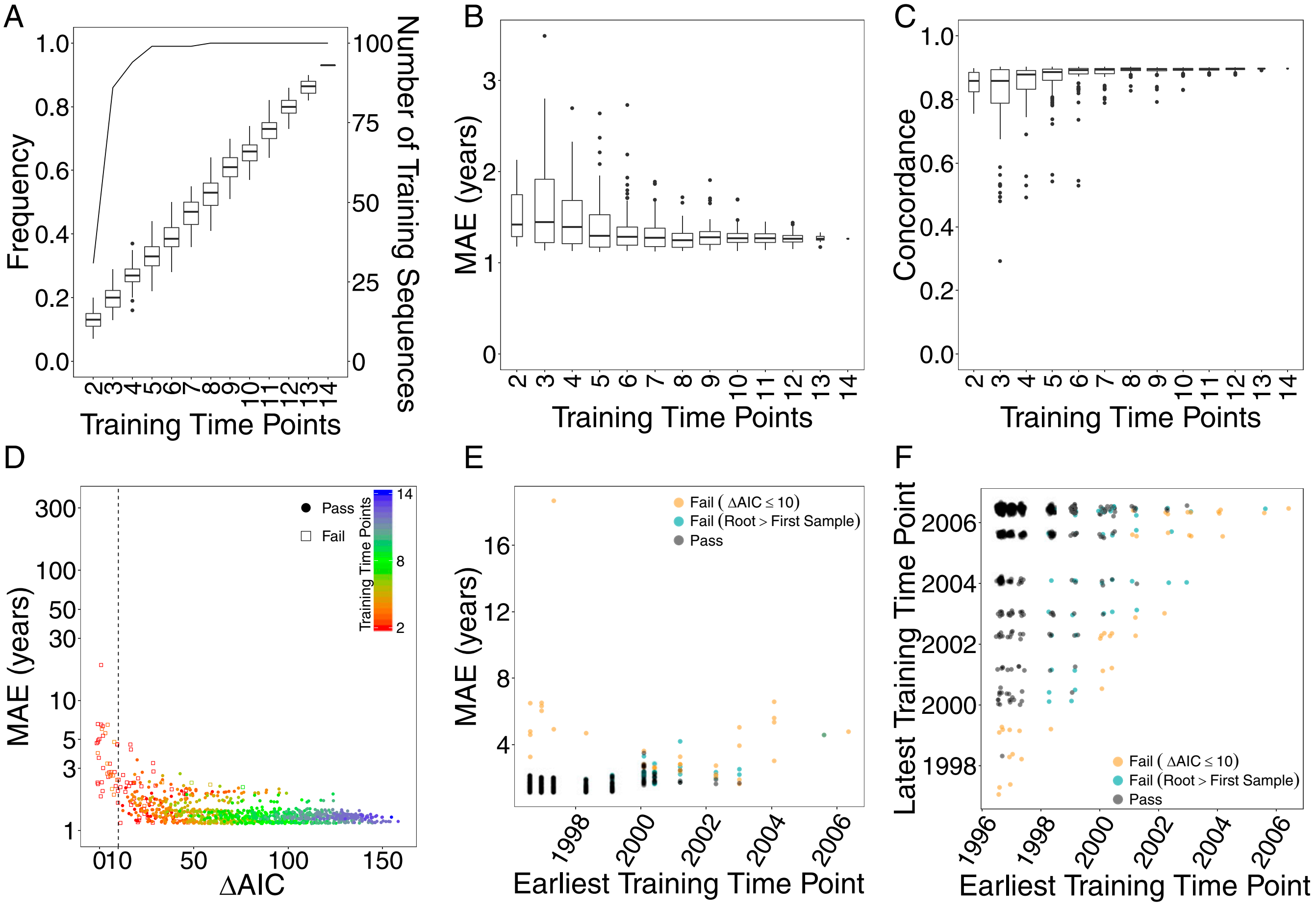

Fig. 5. Framework robustness to training data sampling depth and frequency: RNA censoring validation. This figure summarizes the framework's ability to recover known plasma HIV RNA sampling dates with progressively fewer training data (censoring validation). (*A*) The proportion of linear models passing validation (solid line) and the number of sequences used for model training (floating box plots) are shown for the n = 1,096 subsampled as well as the full (14 time points) dataset. Throughout, box width is scaled to dataset size, box plot horizontal indicates the median, edges indicate interquartile ranges, whiskers denote values within 150% of the quartiles, and circles denote outliers. (*B* and *C*) MAE and concordance coefficient distributions (between recovered and known sampling dates), respectively, stratified by number of training time points. (*D*) Model ΔAIC for all 1,097 datasets, where color denotes the number of training time points and shape denotes whether the model passed or failed. A dotted vertical line denotes ΔAIC = 10. (*E*) Graphic relating model success (black, pass; orange, fail due to ΔAIC <10; teal, fail due to root date criterion), MAE, and date of the earliest training time point for all 1,097 datasets. (*F*) Graphic relating model success with respect to earliest and latest training sampling time points.

5*C*). Remarkably, linear models trained on only two time points (and an average of only 13 training sequences) were still successful in 31% of cases. In general, however, models trained on more time points, where the first was as early as possible and where sampling spanned the widest temporal range, were more likely be successful (Fig. 5 *D–F*). For example, all models where the first training time point was before 1998 (i.e., within ~3 y of the estimated root date) and the last was in 2000 or later were successful (Fig. 5*F*). Our observation that failed models returned a 2.1-fold higher grand mean MAE than passing ones further corroborates our model assessment criteria (*Materials and Methods* and Fig. 5 *D* and *E*).

The framework's ability to reconstruct sequences of unknown age from sparse training data ("RNA subsampling validation") was similarly robust (*SI Appendix*, Fig. S5). The above-described 1,096 temporally subsampled training datasets were used to reconstruct integration dates of putative reservoir sequences; here, however, unused plasma HIV RNA sequences were removed such that phylogenies inferred from fewer training data would be progressively sparser. Again, reservoir establishment dates were reliably retrieved. For example, 84% of linear models trained on only three time points passed calibration (*SI Appendix*, Fig. S5*A*) and yielded reconstructed sequence ages with a median MAD of only 1.5 y and a median concordance coefficient of 0.87 from those estimated using the full training dataset (*SI Appendix*, Fig. S5 *B* and *C*). For context, 1.5 y represents 7.5% of the total (20-y) follow-up period for this participant. Overall, results indicate that our framework is robust to limited sampling.

Framework Is Robust to Rooting Strategy. Rooting phylogenies is inherently difficult. To validate our primary rooting strategy (RTT; see *Materials and Methods*), we outgroup-rooted all datasets on the HIV-1 subtype B reference strain HXB2 (OGR). Outgroup rooting of the eight LANL plasma HIV RNA datasets identified two that did not subsequently exhibit clock-like signal; thus, for the remaining six we censored 50% of the sequences at random and used the remainder as training data. While only three of the resulting linear models passed, these yielded dates highly concordant with RTT-recovered ones (0.98 overall; Fig. 6*A*). Outgroup rooting the seven LANL HIV RNA/DNA datasets, which were generally larger than the HIV RNA datasets, yielded five successful linear models (compared with six of seven by RTT), which returned dates highly concordant with RTT-estimated ones (0.97 overall; Fig. 6*B*). For participants 1 and 2, OGR yielded slightly earlier MRCA dates than RTT (April 1995 and April 1990, respectively; *SI Appendix*, Fig. S6). This is

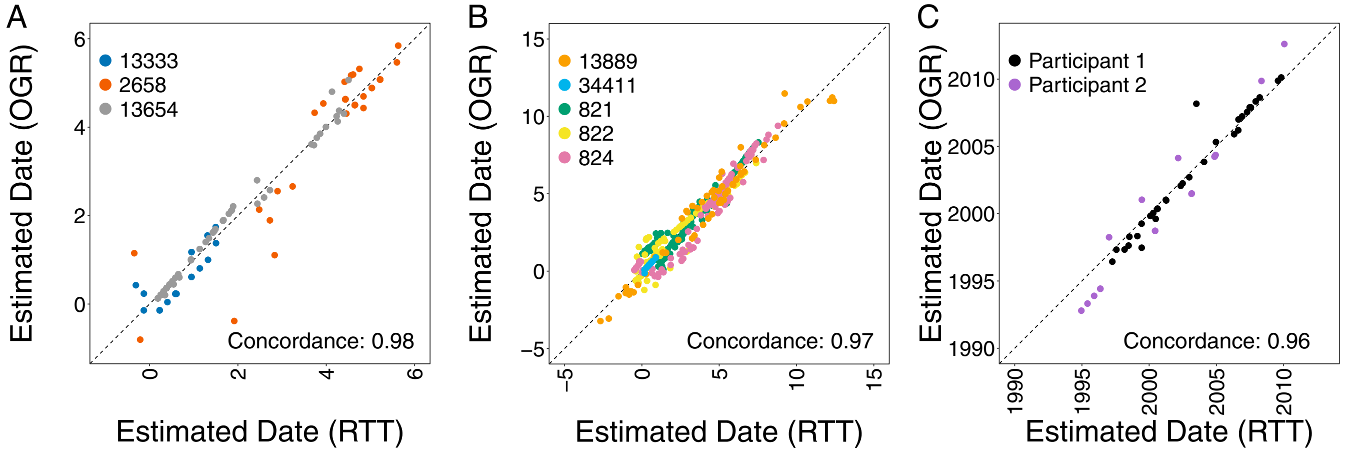

Fig. 6. Model robustness to rooting strategy. Overall concordance of HIV integration dates estimated from RTT vs. OGR for LANL HIV RNA (*A*), LANL HIV RNA/DNA (*B*), and reservoir characterization (*C*) datasets. Results colored by unique individuals. For *A* and *B*, scales are in years before first sampling.

likely because, while RTT is somewhat biased toward placing the root on a branch relating the earliest samples (which is unlikely to be long), OGR root placement has a more uniform probability distribution along the tree (as the outgroup is relatively genetically distant). Folding on a longer branch will thus place the root further back in time relative to the earliest samples on its tips. Nevertheless, OGR-estimated latent HIV dates were highly concordant with RTT-estimated ones (0.96 overall; Fig. 6*C*).

OGR models were also generally robust (*SI Appendix*, Figs. S7 and S8). In the RNA censoring validations, OGR models required slightly more training time points (e.g., whereas 86% of RTT models passed validation, only 73% of OGR ones did so; *SI Appendix*, Fig. S7 *A* and *E*) and returned higher error (e.g., when trained on the full dataset, MAE was 1.9 y for OGR compared with 1.3 y for RTT) (*SI Appendix*, Fig. S7 *B* and *F*), yielding median concordance coefficients of 0.79 vs. 0.90, respectively (*SI Appendix*, Fig. S7 *C* and *G*). OGR models also yielded lower ΔAIC values than RTT ones (*SI Appendix*, Fig. S7 *D* and *H*). Nevertheless, successful OGR models yielded comparable error distributions regardless of training dataset size. For example, the median MAE of OGR models trained on only three time points was only 2.7% higher (1.95 vs. 1.89 y, respectively), and the median concordance coefficient only 4.5% lower, than that of the full dataset (0.76 vs. 0.79) (*SI Appendix*, Fig. S7 *F* and *G*). In the RNA subsampling validations, OGR performed nearly as well as RTT (*SI Appendix*, Fig. S8). For example, compared with dates recovered using the full training dataset, the median MAD of latent HIV ages recovered using only three training time points was 1.8 y for OGR and 1.5 y for RTT (*SI Appendix*, Fig. S8 *B* and *F*), yielding median concordance coefficients of 0.85 and 0.87, respectively (*SI Appendix*, Fig. S8 *C* and *G*). These results indicate that the framework is robust to rooting strategy.

Additional Validations. Since relatively few phylogenetic studies have focused on *nef*, we wished to confirm that evolution in this region is representative of the rest of the HIV genome. To do this we analyzed a published dataset of 16 nearly full genome HIV sequences sampled from an ART-naive individual over ~3.5 y (62). We confirmed that, for each within-host HIV sequence, the number of differences from the baseline consensus within *nef* correlated strongly with that in the remainder of the HIV genome (Pearson $R = 0.96$, $P = 2.0 \times 10^{-9}$). Moreover, a partition-homogeneity test applied to *nef* vs. *env* sequences yielded $P = 0.44$, indicating that we cannot reject the null hypothesis that these regions evolved under the same evolutionary process. These observations support within-host variation in *nef* as being representative of the HIV genome.

Finally it is possible that, even during untreated HIV infection, a minority of plasma HIV RNA sequences may have been recently released from the reservoir. Such sequences would display reduced relative evolutionary rates and could reduce model fit. However, tests for such sequences within participant 1's training data (*Materials and Methods*) identified only one such outlier, and removing it from the analysis yielded latent HIV ages that were 99% concordant with original estimates (*SI Appendix*, Fig. S9). These observations indicate that such sequences are relatively uncommon during untreated HIV infection and that their presence does not substantially affect framework performance.

Discussion

The ability to infer latent HIV integration dates would greatly enhance our understanding of reservoir longevity and dynamics. We developed a simple phylogenetic framework to do this and began by demonstrating its ability to recover known sampling dates on a variety of within-host datasets with acceptable error rates (e.g., in participant 1's HIV RNA training data spanning a 10-y period, the model recovered known sampling dates to within an average of 1.3 y of their actual values; Fig. 5*B*). Subsequent application of the framework to date HIV sequences recovered after up to ~10 y on cART supported several long-held notions. Results confirm proviral DNA in PBMC as a genetically heterogeneous archive recapitulating within-host HIV evolutionary events dating back to transmission (Figs. 3 and 4) (36). Isolation of 18-y-old HIV sequences from a viremia blip on otherwise suppressive cART (Fig. 4*D*) supports the idea that such events can be explained by antigen-driven reactivation of latently infected cells (63) and extends a report of a long-lived (>17 y) defective HIV sequence (64) by demonstrating that replication-competent viral lineages can similarly persist long-term before spontaneously reactivating. Isolation of HIV RNA sequences ranging from 3 to 18 y old from this same viremia blip extends a report that in vivo treatment with latency reversal agents (65) can activate diverse latent HIV lineages by further suggesting that natural stimuli can do the same. This observation also supports a model where latently HIV-infected cells reactivate relatively frequently [although specific rates remain a matter of debate (66, 67)]. Our detection of plasma RNA sequences during participant 1's treatment interruption that exactly matched HIV DNA sequences isolated from their reservoir also corroborates a previous report (68). Of note, the fact that we failed to identify a single HIV sequence whose date estimate's 95% CI unambiguously fell within the period of complete viremia suppression also supports the notion that residual HIV replication on cART, at least as measured in blood, is absent or negligible in this individual (26, 31, 36, 69–71).

While ours is not the first study to infer reservoir temporal origins via analysis of pre-cART HIV sequence variation, previous studies have generally inferred phylogenies solely on the basis of sequence homology (e.g., ref. 30) and have "dated" latent lineages to the period when similar sequences circulated in plasma [a recent elegant study by Brodin et al. (36) significantly advanced the field by inferring host-specific pre-cART evolutionary rates from longitudinal HIV sequences, yet latent reservoir establishment dates were still inferred via genetic similarity to pre-cART sequences]. While reasonable and appropriate, such approaches are limited by their reliance on retrospective sampling of plasma HIV RNA sequences similar to those retrieved from the reservoir, which is not always possible, particularly for individuals who are not diagnosed until later in infection. Recovering latent HIV sequence ages via calculation of a host-specific evolutionary rate overcomes this limitation, allowing the inference of HIV reservoir ages dating back to transmission (72) even if sampling does not extend that far back (e.g., Fig. 4). Our method is also fast. While unknown sequence dates can be recovered using Bayesian approaches (53), these require an additional parameter for each unknown date, significantly increasing computational requirements such that only a small number of sequences can be dated at a time. Our framework can recover all unknown sequence dates simultaneously using a desktop computer and yielded root date and rate estimates concordant with Bayesian estimates. Our framework is also robust to rooting strategy. RTT, which places the tree root at the location that maximizes the relationship between the training data root-to-tip distances and sampling dates, will inherently yield better-fitting linear models than OGR. Nevertheless, OGR performed nearly as well as RTT at reconstructing unknown HIV sequence ages (Fig. 6), even on sparse training data (*SI Appendix*, Figs. S7 and S8). Data requirements are also relatively modest. We dated latent HIV lineages in an individual who did not achieve pVL suppression on cART until a decade after infection, using as few as two training time points (although more were preferable). Framework application to minimal training data should generally be feasible for individuals who initiate cART early, a prediction that is consistent with our previous work on reconstructing HIV infection dates from sequence data (44, 45, 72).

Some limitations merit mention. The framework assumes that within-host HIV evolution can be adequately modeled using a strict molecular clock, which may not be ideal over prolonged timeframes (14). While we addressed variable within-host evolutionary rates in participant 2 by inferring a linear model for each treatment era, adaptation of the framework to work under alternative clock models is warranted. Our framework also assumes that latent lineages remain unchanged until sampling (that is, our model does not allow viral lineages to undergo multiple latent periods). As rates of within-host HIV evolution are gene-specific, training and censored sequences must be derived from the same HIV region: While our previous work suggests that the framework should be adaptable to any HIV gene (72), only *nef* and *env* were validated here. To ensure that HIV sequences represented true within-host lineages we used single-genome amplification; however, only a limited portion of the HIV genome was analyzed and integration sites were not mapped. As such, we cannot confirm whether the studied HIV DNA sequences reside within intact proviruses, nor can we conclusively state that identical sequences derive from latently infected, clonally expanded cell populations. Moreover, as proviral DNA was amplified from PBMC, cellular origin of HIV lineages remains unknown. Finally, as latent HIV ages were only estimated for two participants who initiated cART late, we cannot generalize too broadly. For example, while our observation that HIV sequences sampled during long-term cART recapitulated pretherapy HIV RNA diversity was consistent with Brodin et al. (36), these sequences did not predominantly date to the period shortly before cART as reported in their study. Whether this is attributable to modest participant numbers, methodological differences [e.g., Brodin et al. (36) used next-generation sequencing and inferred reservoir ages via genetic similarity], or substantial interindividual diversity in latent HIV reservoir dynamics and composition merits further investigation.

Despite these limitations, our framework can potentially address key knowledge gaps. For example, framework application to recrudescent plasma HIV RNA sequences following spontaneous or therapeutic latency reversal could advance our understanding of the relationship between reservoir age, lineage origin, and reactivation potential. The framework could be expanded to infer reservoir tissue or cellular origin by adapting phylogeography methods (73–75) to a within-host context. It could be used to investigate the relationship between reservoir age and HIV genomic integrity (provided the viral region used for dating is intact) or genomic integration site (e.g., by using the LTR for model calibration) and should be scalable to next-generation sequence data (44). In conclusion, our framework for reservoir dating corroborates several long-held notions regarding the genetic composition and longevity of the latent reservoir and represents a versatile and potentially powerful addition to the HIV persistence research toolkit.

Materials and Methods

Simulations. Birth–death models have been used to study speciation (76) and epidemics (77) and to model the proliferation of virus lineages within-host (78). We generated 1,000 maximum likelihood phylogenies with 100 tips each under a birth–death model (79) with serial sampling using the sim.bdsky.stt function in the R package TreeSim (80). The model was parameterized by fitting it to data from McCloskey et al. (72). We used BEAST v2.3.2 (81) to estimate the posterior distribution defined by the serial model with a birth–death skyline serial tree prior (77), a strict molecular clock, and an HKY85 model of nucleotide substitution (82). The lineage birth rate λ was estimated as 5.12×10^{-2} per d, the death rate δ as 5.01×10^{-2} per d, and the sampling probability s as 5.24×10^{-3}. We used NELSI (83) to introduce temporal variation in individual branch lengths by assigning rates of evolution to these trees by drawing values from a Gaussian distribution with a mean rate $\mu = 1.96 \times 10^{-4}$ substitutions per generation and SD $\sigma = 1.42 \times 10^{-5}$. To model uncertainty in phylogenetic reconstruction we simulated sequence evolution along each tree with INDELible v1.03 (84) under an HKY85 substitution model where parameters were set to empirical estimates derived from published within-host HIV datasets (stationary distributions 0.42, 0.15, 0.15, and 0.28 for A, C, G, and T, respectively; transition bias parameter 8.5) (72).

Published Datasets. Two datasets from the LANL were assembled for model evaluation (*SI Appendix*, Table S1) (58). The first comprised 424 longitudinal plasma HIV RNA partial *env* sequences from seven individuals from ref. 72 and one from ref. 62, selected based on three predefined criteria: individuals were treatment-naive and had two or more clonal or single-genome sequences per time point where the baseline sample was within 186 d of infection or seroconversion and where at least one subsequent time point was ≥6 mo later. We also required that the overall within-host phylogeny display molecular clock signal (discussed below). The second dataset comprised 361 plasma HIV RNA-derived and 545 PBMC-derived partial *env* sequences, all clonal or single-genome-amplified, from seven individuals for whom at least three RNA time points were available, where the earliest HIV RNA time point was at or before the earliest HIV DNA time point (*SI Appendix*, Table S1). Infection date did not need to be known and the dataset comprised both treatment-naive (14, 85) and treated (86) persons.

Participant Recruitment and Sampling. The framework was applied to two participants recruited from the British Columbia Centre for Excellence in HIV/AIDS for whom longitudinal archived plasma (and in one case, a single archived whole-blood specimen), dating back to 1996, were available. Participants provided blood in summer 2016 from which PBMC were isolated by standard density gradient separation (Histopaque-1077; Sigma), counted (TC20 automated cell counter; Bio-Rad), and cryopreserved at −150 °C in 90% FBS:10% DMSO. This study was approved by the Providence Health Care/University of British Columbia and Simon Fraser University research ethics boards (harmonized protocol H15-03077) and both participants gave written informed consent.

Single-Genome HIV Sequencing. HIV RNA was extracted from 0.5 mL plasma using the NucliSENS EasyMag system (Biomerieux) and genomic DNA was extracted from 200 uL whole blood or 5 million PBMC using the Invitrogen genomic DNA isolation kit. *Nef* was selected for framework application due to its relatively high within-host diversity and richness in phylogenetic signal. Single-genome amplification was performed by limiting dilution using high-fidelity enzymes and HIV sequence-specific primers optimized for amplification of multiple HIV-1 group M subtypes. For HIV RNA extracts, cDNA was generated using Expand Reverse Transcriptase (Roche). Next, cDNA as well as genomic DNA extracts were endpoint-diluted such that $\sim$25–30% of the resulting nested PCR reactions, performed using the Expand High Fidelity PCR system (Roche), would yield an amplicon. Primers used for cDNA generation/first round PCR were Nef8683F_pan (forward; 5-TAGCAGTAGCTGRGKGRACAGATAG-3) and Nef9536R_pan (reverse; 5-TACAGGCAAAAAGCAGCTGCTTATATGYAG-3); primers used for second round PCR were Nef8746F_pan (forward; 5-TCCACATACCTASAAGAATMAGACARG-3) and Nef9474R (reverse; 5-CAGGCCACRCCTCCCTGGAAASKCCC-3). Amplicons were sequenced on an ABI 3130xl or 3730xl automated DNA sequencer using BigDye v3.1 chemistry (Applied Biosystems). Chromatograms were base-called using Sequencher v5.0 (Gene Codes). Sequences exhibiting nucleotide mixtures or gross defects (e.g., large deletions), hypermutated sequences [identified using Hypermut 2.0 (87)], and within-host recombinants [identified using rdp4 Beta 95 (88)] were discarded. Unique sequences have been deposited in GenBank (accession nos. MG822917–MG823179).

Alignments, Phylogenetic Inference, Model Construction, and Evaluation. HIV sequences were aligned using MUSCLE v3.8.31 (89) and manually edited with AliView v1.18 (90). Where identical sequences were observed, the earliest was retained. We reconstructed maximum likelihood phylogenies using RAxML v8.2.10 (54) with 100 bootstraps under the generalized time reversible model and rooted these at the inferred MRCA as follows. Our primary rooting method involved a modification of root-to-tip regression (RTT) (91), where we exhaustively reroot the tree to identify the root location that maximizes the correlation between the root-to-tip distances and collection dates of the training data. We also outgroup rooted (OGR) on the HIV-1 subtype B reference strain HXB2 (GenBank accession no. K03455); here, the node intersecting the branch leading to the outgroup provides the root location. The linear and null model equations (*Results*) were computed using the R function lm (92). Two criteria were required to advance linear models for molecular dating: The null model's AIC (55) needed to be at least 10 units higher than that of the linear model [ΔAIC > 10 criterion, a conservative threshold (93) that in our experiments corresponds to $P = 0.00053$ when using a log-likelihood ratio test] and the 95% CI of the linear model-estimated root date needed to contain or precede the first sampling date. Model error was quantified using two metrics. The discordance between model-predicted and training data sampling dates was expressed as MAE, while the discordance between model-predicted and censored data sampling dates was expressed as MAD. We make this distinction because, for latent HIV sequences, model-predicted dates will legitimately precede sampling. To facilitate comparison between heterogenous validation datasets, MAE and MAD were normalized by scaling values to the total timeframe of the training data. Training sequences that did not conform to a molecular clock (*SI Appendix*, Fig. S9) were identified by computing the relative evolutionary rate (94) of each pair of sequences with unique collection dates and identifying those whose relative rates deviated from the mean by $\geq$2 SD in at least 10% of replicates.

Verification of Root Date Estimates Using Bayesian Methods. Linear model-derived evolutionary rate and root date estimates were validated using BEAST v1.8.4 (95). Posterior distributions for the root date were estimated using a SRD06 substitution model (96), a lognormal uncorrelated relaxed molecular clock (97), and a coalescent Gaussian Markov random field Bayesian skyride tree prior (98) in four parallel Markov chain Monte Carlo chains with 10^8 iterations each. After discarding the first 10% as burn-in, chains were combined with LogCombiner v2.4.2 (81) and analyzed in Tracer v1.6 to ensure convergence and effective sample size values >200 for all parameters.

Comparison of *nef* to the Rest of the Genome. We analyzed a published longitudinal within-host near-whole genome HIV sequence dataset from an untreated individual (GenBank accession no. DQ853436-65) (62), removed 14 sequences that exhibited evidence for recombination (88), and aligned the remaining 16 to HXB2 using MAFFT v7.313 (99). We used Pearson's correlation to evaluate the relationship between the number of nucleotide differences between each sequence and the consensus of the earliest time point within *nef* vs. the rest of the HIV genome (ignoring gaps). We also used the partition-homogeneity test implemented in PAUP* v4.0a (100, 101) with 1,000 replicates to compare evolution in *nef* vs. *env*.

Statistical Analyses. Statistical calculations were performed using R v3.3.3. CIs were calculated using the inverse.predict function of the chemCal package in R. Distributional asymmetry in reconstructed dates was tested using a two-tailed nonparametric binomial test followed by a Bonferroni correction (57). Plots were generated using ggplot2 and ggtree (102) packages in R. Divergence times for ancestor traces were estimated using the estimate.dates function of the ape package in R (103, 104).

ACKNOWLEDGMENTS. We thank Conan Woods for database assistance, Don Kirkby and Walter Scott for cluster assistance, and Drs. Julio Montaner and Rolando Barrios for support (British Columbia Centre for Excellence in HIV/AIDS). We gratefully thank the participants who donated blood for this study. This work was supported by Canadian Institutes for Health Research (CIHR) Grants PJT-148621 (to Z.L.B. and M.A.B.), PJT- 153391 (to A.F.Y.P.), and PJT-155990 (to A.F.Y.P.); National Institute of Allergy and Infectious Diseases of the NIH Grant UM1AI126617, with cofunding support from the National Institute on Drug Abuse, the National Institute of Mental Health, and the National Institute of Neurological Disorders and Stroke (to Z.L.B. and M.A.B.); NIH Grant R21A127029 (to Z.L.B., M.A.B., A.F.Y.P., and J.B.J.); and Canadian HIV Cure Enterprise Team Grant HIG-133050 (to Z.L.B., M.A.B., and A.F.Y.P.) from the CIHR in partnership with the Canadian Foundation for AIDS Research and the International AIDS Society. N.N.K. was supported by a CIHR Frederick Banting and Charles Best Master of Science Award. M.A.B. holds a Canada Research Chair in Viral Pathogenesis and Immunity. A.F.Y.P. is supported by a CIHR New Investigator Award FRN-130609. Z.L.B. is supported by a Michael Smith Foundation for Health Research Scholar Award.

1. Chun TW, et al. (1997) Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *Proc Natl Acad Sci USA* 94:13193–13197.
2. Finzi D, et al. (1997) Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* 278:1295–1300.
3. Archin NM, Sung JM, Garrido C, Soriano-Sarabia N, Margolis DM (2014) Eradicating HIV-1 infection: Seeking to clear a persistent pathogen. *Nat Rev Microbiol* 12:750–764.
4. Pace MJ, Agosto L, Graf EH, O'Doherty U (2011) HIV reservoirs and latency models. *Virology* 411:344–354.
5. Richman DD, et al. (2009) The challenge of finding a cure for HIV infection. *Science* 323:1304–1307.
6. Durand CM, Blankson JN, Siliciano RF (2012) Developing strategies for HIV-1 eradication. *Trends Immunol* 33:554–562.
7. Joos B, et al.; Swiss HIV Cohort Study (2008) HIV rebounds from latently infected cells, rather than from continuing low-level replication. *Proc Natl Acad Sci USA* 105:16725–16730.
8. Katlama C, et al. (2013) Barriers to a cure for HIV: New ways to target and eradicate HIV-1 reservoirs. *Lancet* 381:2109–2117.
9. Pomerantz RJ (2003) Reservoirs, sanctuaries, and residual disease: The hiding spots of HIV-1. *HIV Clin Trials* 4:137–143.
10. Shen L, Siliciano RF (2008) Viral reservoirs, residual viremia, and the potential of highly active antiretroviral therapy to eradicate HIV infection. *J Allergy Clin Immunol* 122:22–28.
11. Hiener B, et al. (2017) Identification of genetically intact HIV-1 proviruses in specific CD4$^+$ T cells from effectively treated participants. *Cell Rep* 21:813–822.
12. Alizon S, Fraser C (2013) Within-host and between-host evolutionary rates across the HIV-1 genome. *Retrovirology* 10:49.
13. Rambaut A, Posada D, Crandall KA, Holmes EC (2004) The causes and consequences of HIV evolution. *Nat Rev Genet* 5:52–61.
14. Shankarappa R, et al. (1999) Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J Virol* 73:10489–10502.
15. Herbeck JT, et al. (2011) Demographic processes affect HIV-1 evolution in primary infection before the onset of selective processes. *J Virol* 85:7523–7534.
16. Henn MR, et al. (2012) Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog* 8:e1002529.
17. Salazar-Gonzalez JF, et al. (2009) Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. *J Exp Med* 206:1273–1289.
18. Fischer W, et al. (2010) Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS One* 5:e12303.
19. Tully DC, et al. (2016) Differences in the selection bottleneck between modes of sexual transmission influence the genetic composition of the HIV-1 founder virus. *PLoS Pathog* 12:e1005619.
20. Keele BF, et al. (2008) Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci USA* 105:7552–7557.
21. Whitney JB, et al. (2014) Rapid seeding of the viral reservoir prior to SIV viraemia in rhesus monkeys. *Nature* 512:74–77.
22. Ledford H (July 10, 2014) HIV rebound dashes hope of 'Mississippi baby' cure. *Nature*, 10.1038/nature.2014.15535.
23. Finzi D, et al. (1999) Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nat Med* 5:512–517.
24. Siliciano JD, et al. (2003) Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+ T cells. *Nat Med* 9:727–728.
25. Nickle DC, et al. (2003) Evolutionary indicators of human immunodeficiency virus type 1 reservoirs and compartments. *J Virol* 77:5540–5546.
26. Kearney MF, et al. (2014) Lack of detectable HIV-1 molecular evolution during suppressive antiretroviral therapy. *PLoS Pathog* 10:e1004010.
27. Lambotte O, et al. (2004) The lymphocyte HIV reservoir in patients on long-term HAART is a memory of virus evolution. *AIDS* 18:1147–1158.
28. Bruner KM, et al. (2016) Defective proviruses rapidly accumulate during acute HIV-1 infection. *Nat Med* 22:1043–1049.

29. Lee GQ, Lichterfeld M (2016) Diversity of HIV-1 reservoirs in CD4+ T-cell subpopulations. *Curr Opin HIV AIDS* 11:383–387.
30. Buzon MJ, et al. (2014) HIV-1 persistence in CD4+ T cells with stem cell-like properties. *Nat Med* 20:139–142.
31. Josefsson L, et al. (2013) The HIV-1 reservoir in eight patients on long-term suppressive antiretroviral therapy is stable with few genetic changes over time. *Proc Natl Acad Sci USA* 110:E4987–E4996.
32. Evering TH, et al. (2012) Absence of HIV-1 evolution in the gut-associated lymphoid tissue from patients on combination antiviral therapy initiated during primary infection. *PLoS Pathog* 8:e1002506.
33. Chomont N, et al. (2009) HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation. *Nat Med* 15:893–900.
34. von Stockenstrom S, et al. (2015) Longitudinal genetic characterization reveals that cell proliferation maintains a persistent HIV type 1 DNA pool during effective HIV therapy. *J Infect Dis* 212:596–607.
35. Rothenberger MK, et al. (2015) Large number of rebounding/founder HIV variants emerge from multifocal infection in lymphatic tissues after treatment interruption. *Proc Natl Acad Sci USA* 112:E1126–E1134.
36. Brodin J, et al. (2016) Establishment and stability of the latent HIV-1 DNA reservoir. *eLife* 5:e18889.
37. Simonetti FR, et al. (2016) Clonally expanded CD4+ T cells can produce infectious HIV-1 in vivo. *Proc Natl Acad Sci USA* 113:1883–1888.
38. Bui JK, et al. (2017) Proviruses with identical sequences comprise a large fraction of the replication-competent HIV reservoir. *PLoS Pathog* 13:e1006283.
39. Lee GQ, et al. (2017) Clonal expansion of genome-intact HIV-1 in functionally polarized Th1 CD4+ T cells. *J Clin Invest* 127:2689–2696.
40. Cohn LB, et al. (2015) HIV-1 integration landscape during latent and active infection. *Cell* 160:420–432.
41. Maldarelli F, et al. (2014) HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* 345:179–183.
42. Immonen TT, Leitner T (2014) Reduced evolutionary rates in HIV-1 reveal extensive latency periods among replicating lineages. *Retrovirology* 11:81.
43. Althaus CL, Joos B, Perelson AS, Günthard HF (2014) Quantifying the turnover of transcriptional subclasses of HIV-1-infected cells. *PLOS Comput Biol* 10:e1003871.
44. Poon AF, et al. (2011) Dates of HIV infection can be estimated for seroprevalent patients by coalescent analysis of serial next-generation sequencing data. *AIDS* 25:2019–2026.
45. Poon AF, et al. (2012) Reconstructing the dynamics of HIV evolution within hosts from serial deep sequence data. *PLOS Comput Biol* 8:e1002753.
46. Rodrigo AG, Felsenstein J (1999) Coalescent approaches to HIV population genetics. *The Evolution of HIV*, ed Crandall KA (Johns Hopkins Univ Press, Baltimore), pp 233–271.
47. Leitner T, Albert J (1999) The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc Natl Acad Sci USA* 96:10752–10757.
48. Lemey P, Rambaut A, Pybus OG (2006) HIV evolutionary dynamics within and among hosts. *AIDS Rev* 8:125–140.
49. Korber B, et al. (2000) Timing the ancestor of the HIV-1 pandemic strains. *Science* 288:1789–1796.
50. Worobey M, et al. (2008) Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* 455:661–664.
51. Le AQ, et al. (2015) Differential evolution of a CXCR4-using HIV-1 strain in CCR5wt/wt and CCR5Δ32/Δ32 hosts revealed by longitudinal deep sequencing and phylogenetic reconstruction. *Sci Rep* 5:17607.
52. Vrancken B, et al. (2014) The genealogical population dynamics of HIV-1 in a large transmission chain: Bridging within and among host evolutionary rates. *PLOS Comput Biol* 10:e1003505.
53. Shapiro B, et al. (2011) A Bayesian phylogenetic method to estimate unknown sequence ages. *Mol Biol Evol* 28:879–887.
54. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.
55. Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Automat Contr* 19:716–723.
56. Massart DL, et al. (1998) *Handbook of Chemometrics and Qualimetrics: Part A* (Elsevier, Amsterdam).
57. Goeman JJ, Solari A (2014) Multiple hypothesis testing in genomics. *Stat Med* 33: 1946–1978.
58. Los Alamos National Laboratory (2018) HIV sequence database. Available at www.hiv.lanl.gov/. Accessed June 24, 2015.
59. Simmonds P, et al. (1991) Discontinuous sequence change of human immunodeficiency virus (HIV) type 1 env sequences in plasma viral and lymphocyte-associated proviral populations in vivo: Implications for models of HIV pathogenesis. *J Virol* 65: 6266–6276.
60. Poon AF, et al. (2013) Mapping the shapes of phylogenetic trees from human and zoonotic RNA viruses. *PLoS One* 8:e78122.
61. Lin LI (1989) A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45:255–268.
62. Liu Y, McNevin JP, Holte S, McElrath MJ, Mullins JI (2011) Dynamics of viral evolution and CTL responses in HIV-1 infection. *PLoS One* 6:e15639.
63. Rong L, Perelson AS (2009) Modeling latently infected cell activation: Viral and latent reservoir persistence, and viral blips in HIV-infected patients on potent therapy. *PLOS Comput Biol* 5:e1000533.
64. Imamichi H, et al. (2014) Lifespan of effector memory CD4+ T cells determined by replication-incompetent integrated HIV-1 provirus. *AIDS* 28:1091–1099.
65. Barton K, et al. (2016) Broad activation of latent HIV-1 in vivo. *Nat Commun* 7:12731.
66. Pinkevych M, et al. (2015) HIV reactivation from latency after treatment interruption occurs on average every 5-8 days–Implications for HIV remission. *PLoS Pathog* 11: e1005000.
67. Hill AL, Rosenbloom DI, Siliciano JD, Siliciano RF (2016) Insufficient evidence for rare activation of latent HIV in the absence of reservoir-reducing interventions. *PLoS Pathog* 12:e1005679.
68. Kearney MF, et al. (2015) Origin of rebound plasma HIV includes cells with identical proviruses that are transcriptionally active before stopping of antiretroviral therapy. *J Virol* 90:1369–1376.
69. Rosenbloom DIS, Hill AL, Laskey SB, Siliciano RF (2017) Re-evaluating evolution in the HIV reservoir. *Nature* 551:E6–E9.
70. Besson GJ, et al. (2014) HIV-1 DNA decay dynamics in blood during more than a decade of suppressive antiretroviral therapy. *Clin Infect Dis* 59:1312–1321.
71. Kieffer TL, et al. (2004) Genotypic analysis of HIV-1 drug resistance at the limit of detection: Virus production without evolution in treated adults with undetectable HIV loads. *J Infect Dis* 189:1452–1465.
72. McCloskey RM, Liang RH, Harrigan PR, Brumme ZL, Poon AF (2014) An evaluation of phylogenetic methods for reconstructing transmitted HIV variants using longitudinal clonal HIV sequence data. *J Virol* 88:6181–6194.
73. Paraskevis D, et al.; SPREAD Programme (2009) Tracing the HIV-1 subtype B mobility in Europe: A phylogeographic approach. *Retrovirology* 6:49.
74. Mehta SR, et al. (2011) Using phylogeography to characterize the origins of the HIV-1 subtype F epidemic in Romania. *Infect Genet Evol* 11:975–979.
75. Holmes EC (2004) The phylogeography of human viruses. *Mol Ecol* 13:745–756.
76. Nee S (2006) Birth-death models in macroevolution. *Annu Rev Ecol Evol Syst* 37:1–17.
77. Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ (2013) Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc Natl Acad Sci USA* 110:228–233.
78. Hartfield M, Alizon S (2015) Within-host stochastic emergence dynamics of immune-escape mutants. *PLOS Comput Biol* 11:e1004149.
79. Boskova V, Bonhoeffer S, Stadler T (2014) Inference of epidemiological dynamics based on simulated phylogenies using birth-death and coalescent models. *PLOS Comput Biol* 10:e1003913.
80. Stadler T (2011) Simulating trees with a fixed number of extant species. *Syst Biol* 60: 676–684.
81. Bouckaert R, et al. (2014) BEAST 2: A software platform for Bayesian evolutionary analysis. *PLOS Comput Biol* 10:e1003537.
82. Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22:160–174.
83. Ho SYW, Duchêne S, Duchêne D (2015) Simulating and detecting autocorrelation of molecular evolutionary rates among lineages. *Mol Ecol Resour* 15:689–696.
84. Fletcher W, Yang Z (2009) INDELible: A flexible simulator of biological sequence evolution. *Mol Biol Evol* 26:1879–1888.
85. Fischer M, et al.; Swiss HIV Cohort Study (2004) Attenuated and nonproductive viral transcription in the lymphatic tissue of HIV-1-infected patients receiving potent antiretroviral therapy. *J Infect Dis* 189:273–285.
86. Novitsky V, et al. (2009) Timing constraints of in vivo gag mutations during primary HIV-1 subtype C infection. *PLoS One* 4:e7727.
87. Rose PP, Korber BT (2000) Detecting hypermutations in viral sequences with an emphasis on G –> A hypermutation. *Bioinformatics* 16:400–401.
88. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B (2015) RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol* 1:vev003.
89. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
90. Larsson A (2014) AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 30:3276–3278.
91. Maljkovic Berry I, et al. (2009) The evolutionary rate dynamically tracks changes in HIV-1 epidemics: Application of a simple method for optimizing the evolutionary rate in phylogenetic trees with longitudinal data. *Epidemics* 1:230–239.
92. R Development Core Team (2008) R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna).
93. Burnham KP, Anderson DR (2002) *Model Selection and Multimodel Inference* (Springer, New York).
94. Rambaut A (2000) Estimating the rate of molecular evolution: Incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* 16:395–399.
95. Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 29:1969–1973.
96. Shapiro B, Rambaut A, Drummond AJ (2006) Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol Biol Evol* 23:7–9.
97. Drummond AJ, Ho SY, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4:e88.
98. Minin VN, Bloomquist EW, Suchard MA (2008) Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol Biol Evol* 25:1459–1471.
99. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* 30:772–780.
100. Swofford DL (2003) PAUP* [Phylogenetic Analysis Using Parsimony (and Other Methods)]. Available at http://paup.phylosolutions.com/. Accessed May 15, 2018.
101. Farris JS, Kallersjo M, Kluge AG, Bult C (1994) Testing significance of incongruence. *Cladistics* 10:315–319.
102. Yu G, Smith D, Zhu H, Guan Y, Lam TTY (2017) GGTREE: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 8:28–36.
103. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.
104. Jones BR, Poon AFY (2017) node.dating: Dating ancestors in phylogenetic trees in R. *Bioinformatics* 33:932–934.