VIRUS EVOLUTION

Virus Evolution, 2023, **9(2)**, 1–9
DOI: https://doi.org/10.1093/ve/vead046
Advance access publication date: 27 July 2023
Research Article

Dating reservoir formation in virologically suppressed people living with HIV-1 in Rakai, Uganda

Edward Nelson Kankaka,[1,2,†,*] Andrew D. Redd,[2,3,4] Amjad Khan,[5] Steven J. Reynolds,[1,2,3] Sharada Saraf,[3] Charles Kirby,[2] Briana Lynch,[3] Jada Hackman,[3,‡] Stephen Tomusange,[1] Taddeo Kityamuweesi,[1] Samiri Jamiru,[1] Aggrey Anok,[1] Paul Buule,[1] Daniel Bruno,[6] Craig Martens,[6] Larry W. Chang,[2] Thomas C. Quinn,[2,3] Jessica L. Prodger,[7] and Art Poon[5,§]

[1]Research Department, Rakai Health Sciences Program, 4-6 Sanitary Lane, Old Bukoba Road, Kalisizo 256, Uganda, [2]Division of Infectious Diseases, Johns Hopkins University School of Medicine, 615 N. Wolfe Street, Baltimore, MD 21211, USA, [3]Laboratory of Immunoregulation, Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 5601 Fishers Lane, MSC, Bethesda, MD 9806, USA, [4]Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Faculty of Health Sciences, Anzio Rd, Observatory, Cape Town 7925, South Africa, [5]Department of Pathology, Schulich School of Medicine and Dentistry, Western University, 1151 Richmond Street, London, Ontario N6A 5K8, Canada, [6]Genomic Unit, Rocky Mountain Laboratories, NIAID, NIH, 904 South Fourth Street, Hamilton, MT 59840, USA and [7]Department of Microbiology and Immunology, Schulich School of Medicine and Dentistry, Western University, 1151 Richmond Street, London, Ontario N6A 5K8, Canada
[†]https://orcid.org/0000-0001-6977-3405
[‡]https://orcid.org/0000-0001-8412-8674
[§]https://orcid.org/0000-0003-3779-154×
*Corresponding author: E-mail: ken1edd@gmail.com

Abstract

The timing of the establishment of the HIV latent viral reservoir (LVR) is of particular interest, as there is evidence that proviruses are preferentially archived at the time of antiretroviral therapy (ART) initiation. Quantitative viral outgrowth assays (QVOAs) were performed using Peripheral Blood Mononuclear Cells (PBMC) collected from Ugandans living with HIV who were virally suppressed on ART for >1 year, had known seroconversion windows, and at least two archived ART-naïve plasma samples. QVOA outgrowth populations and pre-ART plasma samples were deep sequenced for the *pol* and *gp41* genes. The *bayroot* program was used to estimate the date that each outgrowth virus was incorporated into the reservoir. *Bayroot* was also applied to previously published data from a South African cohort. In the Ugandan cohort ($n = 11$), 87.9 per cent pre-ART and 56.3 per cent viral outgrowth sequences were unique. Integration dates were estimated to be relatively evenly distributed throughout viremia in 9/11 participants. In contrast, sequences from the South African cohort ($n = 9$) were more commonly estimated to have entered the LVR close to ART initiation, as previously reported. Timing of LVR establishment is variable between populations and potentially viral subtypes, which could limit the effectiveness of interventions that target the LVR only at ART initiation.

Keywords: HIV; latent; reservoir; establishment; formation; Bayesian.

1. Introduction

Antiretroviral therapy (ART) halts HIV-1 replication and disease progression and has markedly reduced morbidity and mortality. However, ART does not eradicate HIV from the body, and in the majority of cases, viremia rebounds relatively quickly upon interruption of ART, requiring lifelong treatment. People living with HIV must endure drug side effects, pill burden, stigma and other psychosocial distress, and increased risk of comorbidities. The primary obstacle to an HIV cure is the presence of stable, latently infected cells that allow persistence of replication-competent provirus despite optimal ART. These cells are collectively called the latent viral reservoir (LVR) (Evelyn and Siliciano 2012).

The LVR predominantly consists of resting CD4+ (rCD4) T cells, which are believed to become infected when transitioning from an activated to a resting state, allowing them to persist without immune detection with HIV proviral DNA stably integrated within the host genome (Shan et al. 2017). In the setting of suppressive ART, these latently infected cells generally decay over time but at a rate that is too slow to be cleared in a lifetime (Finzi et al. 1999; Siliciano et al. 2003). While the LVR is relatively small—about 10–100 cells harboring proviruses per 10^6 rCD4 T cells—only a fraction of these proviruses are intact, and an even smaller fraction are replication-competent (~1/10^6 rCD4 T cells) (Prodger et al. 2017).

There is general agreement that LVR establishment starts almost immediately after infection and is augmented throughout periods of viremia (Archin et al. 2012; Persaud et al. 2014). However, it is not yet known if cells become latently infected consistently throughout viremia (so that the LVR is accrued evenly throughout pre-ART infection) or if there are some events that promote latency. One event of interest is the impact of ART initiation on LVR establishment. Multiple studies have sought to

estimate the integration dates of latently infected cells based on comparing proviral sequences to evolutionary phylogenies created from sequences of virus circulating at different timepoints pre-ART. Several of these studies have reported that most proviruses in the LVR integrated in the time period preceding ART initiation (Johanna et al. 2016; Abrahams et al. 2019; Brooks et al. 2020; Pankau et al. 2020), which has led some to propose that ART initiation may promote the transition of short-lived effector CD4 T cells to long-lived rCD4. If this is found to be true, interventions to limit LVR establishment at ART initiation might result in smaller and less heterogenous LVRs that would be easier to clear by potential cure therapies. However, observations from a Canadian cohort suggest that the LVR is accrued evenly throughout pre-ART infection, with no one period favoring the establishment of latently infected cells (Jones et al. 2018). These variations in patterns of LVR establishment warrant further investigations in other populations of people living with HIV.

A potential explanation for the discordance between previous studies is that the methods used to estimate the timing of LVR establishment do not account for the uncertainty in estimating the rate of pre-ART evolution, which is used to generate a molecular clock to estimate the integration dates of sequences from the LVR. Additionally, previous analysis methods do not directly use additional information, such as dates of seroconversion or ART initiation, to refine integration date estimates. In this study, a novel dating method that addresses these limitations, *bayroot* (a Bayesian extension of root-to-tip regression) (Ferreira, Wong, and Art 2023), was used to determine the timing of LVR establishment in ART-suppressed individuals with known dates of seroconversion living in Rakai, Uganda. In a secondary analysis, we applied this method to data from one of the previously published studies reporting the preponderance of LVR establishment close to the time of ART initiation (Abrahams et al. 2019).

2. Materials and methods

2.1 Study participants

This study was nested in two established cohorts in Rakai, Uganda: the Rakai LVR cohort (Prodger et al. 2017) and the Rakai Community Cohort Study (RCCS) (Kankaka et al. 2022). The RCCS is a longstanding (>30 years) population-based cohort study through which longitudinal HIV testing and archived serum samples were available. The Rakai LVR cohort is a longitudinal study of HIV persistence, nested within the RCCS, following adults (≥18 years) living with HIV-1, all of whom were virally suppressed on ART for >1 year at the time of enrollment.

The current study population consisted of participants with known seroconversion windows (at least one HIV-negative test result prior to confirmed HIV-1 seroconversion) and at least two historical ART-naïve serum samples available through the RCCS.

2.2 Recovery of replication-competent provirus from PBMC

As a part of Rakai LVR cohort study activities, participants had previously provided one or more samples of whole blood (180 ml) for quantitative viral outgrowth assays (QVOA). For participants with multiple blood samples, samples were collected at least 1 year apart. Methods used for QVOA assays have been described in detail elsewhere (Chun et al. 1997; Laird et al. 2016; Prodger et al. 2017). Briefly, rCD4 T cells were isolated using negative selection, stimulated, and plated in a limiting dilution format with MOLT-4 cells stably expressing CD4, CCR5, and CXCR-4. After 14 or 21 days, wells were examined for viral outgrowth using p24 ELISA (Perkin-Elmer, Waltham MA). Supernatants containing outgrowth virus were harvested and stored at −80°C for sequencing.

2.3 Viral sequencing

Prominent QVOA outgrowth viruses (LVR sequences) and total circulating viral populations from pre-ART serum samples (pre-ART sequences) were sequenced using site-directed next-generation sequencing for the reverse transcriptase region of the *pol* gene (POL_RT; HXB2 bases = 2,723–3,225) and the gp41 region of the *env* gene (ENV_GP41; HXB2 bases = 7,938–8,256), as described previously (Illumina Inc, San Diego, CA) (Courtney et al. 2017; Poon et al. 2018).

2.4 Phylogenetic trees

For each participant, unique pre-ART and LVR sequences, as well as the HXB2 reference sequence (van Beveren, Coffin, and Hughes 2002), were aligned using MAFFT (Kazutaka and Standley 2014). Unique sequences were obtained by keeping only the earliest sampled sequence in case of identical sequences (done separately for pre-ART and outgrowth viruses before they were combined). Sequences were screened for hypermutations using the Hypermut tool from Los Alamos (Alamos 2022). Maximum likelihood phylogenies were then generated using RAxML (Stamatakis 2014) under a general time reversible model of evolution with HXB2 as the outgroup to obtain an initial rooted topology. The HXB2 outgroup was then dropped, and the topology of the maximum likelihood phylogeny further improved using known sampling times of pre-ART tips for re-rooting using root-to-tip regression (sampling times of LVR tips were censored). While using HXB2 as an outgroup provided a rough estimate of the location of the root, it was subsequently dropped as it represents a subtype B infection that is distantly related to the virus populations in our study. Root-to-tip regression provides a refined estimate of the root location using the within-host genetic variation (root-to-tip distances) and known pre-ART sampling times for each study participant.

2.5 Dating of LVR variants

The re-rooted tree was used as the input tree for the *bayroot* package to estimate LVR integration dates (Ferreira, Wong, and Art 2023). Briefly, *bayroot* uses the Metropolis-Hastings algorithm to randomly sample the regression parameters from the posterior distribution, in this case being the molecular clock (slope), location of the root in the tree, and the time associated with the root (x–intercept). We used a lognormal prior for the molecular clock (initial rate = 0.01), a uniform prior on the root location, and a uniform distribution on the root time based on the known seroconversion window. After several runs (∼100 burnin), replicate chain samples visibly converged to the same posterior distribution (Supplementary Fig. 1). The resulting posterior sample of regression parameters was used to generate a sample of integration dates for each query LVR sequence using rejection sampling from a probability distribution constrained to prior information (i.e. no expected LVR integration on suppressive ART). The resulting integration date distributions were extracted for further statistical analyses and visualization.

2.6 Modeling of expected distributions

To generate the expected distribution of integration times, we used the deterministic model of continuous reservoir seeding and decay as implemented in Python by Pankau et al. (Pankau et al. 2020).

All model parameters were kept to the same values, except that we adjusted the per-replication probability of producing intact virus from $\tau = 0.05$ to $\tau = 0.5$ to be consistent with empirical estimates (Bruner et al. 2016). In this model, susceptible target cells S are replenished at a rate $\alpha_s = 70$ cells μl^{-1} day^{-1}, die at a rate $\delta_s = 0.2$ day^{-1}, and are infected by virions at a rate $\beta = 10^{-4}$ μL cells^{-1} day^{-1}. Some of these cells become productively infected with intact virions (A_p) while others become unproductively infected with defective virions (A_u), with the per-replication probability of producing intact provirus τ. Upon infection, active infected cells A_p and A_u may enter latency with a probability $\lambda = 10^{-4}$. All active infected cells A_p and A_u die at a rate $\delta_1 = 0.8$ day^{-1} and are killed by adaptive immunity E at a rate $\kappa = 0.3$ μL cells^{-1} day^{-1}. The adaptive immunity has an initial adaptive precursor frequency $\alpha_E = 10$–4 cells μL^{-1} day^{-1}, is additionally recruited by cell infection at a rate $\omega = 1.6$ mL cells^{-1} day^{-1} with a 50 per cent saturation constant $E_{50} = 250$ cells μL^{-1}; and is cleared at a rate $\delta_E = 0.002$ day^{-1}. Intact infectious virions V are produced by the productively infected cells A_p with a burst size $\pi = 5 \times 10^4$ virions cells^{-1}, decay at a rate $\gamma =$ 23 day^{-1}, and infect new susceptible cells at the rate β. Following Pankau et al., we simulated this system with no decay (i.e. reservoir seeding only), with slow decay at a half-life $t_{1/2} = 44$ months (Siliciano et al. 2003), and with a slower decay at a half-life of $t_{1/2} = 139$ months (Golob et al. 2018). The modeled distributions were plotted and compared to empirical distributions in yearly bins using the two-sample Kolmogorov-Smirnov test.

2.7 Categorization of integration date estimates

For each participant and for each sequenced region (POL_RT and ENV_GP41), we categorized the integration dates into relative time periods. Relative time periods allowed controlling for inter-patient differences in length of the viremic window (the time period between seroconversion and ART initiation) and were defined as Oldest, Middle, and Most recent (the first, middle, and last tertiles of the viremic window, respectively). Similar categorization was done for the modeled distributions, and this was visually compared to the empirical distributions.

2.8 Comparison with the CAPRISA cohort

In a previous report, nine female participants in the CAPRISA cohort were found to have predominant LVR seeding in the time period immediately preceding ART initiation (Abrahams et al. 2019). In the original analysis, three methods were used to estimate integration dates: patristic distance, clade support, and phylogenetic placement. These methods were applied to each of the five regions of the genome sequenced (the p17 region of *gag* ('GAG_P17'); three regions from *env*: the C1-C2 region of gp120 with hypervariable loops V1 and V2 trimmed ('ENV_C1C2'), the C2-C3 region of gp120 ('ENV_C2C3'), and the C4-C5 region of gp120 with hypervariable loop V5 trimmed ('ENV_C4C5'); and the 5′ region of *nef* ('NEF_1'), and the weighted median of these estimates was taken as the integration date when each provirus had entered the reservoir. As this approach differs from *bayroot*, we re-analyzed the sequence data from the CAPRISA cohort to ensure that any findings in the Rakai LVR cohort were not due to different analysis methods. Relative time distributions were used for the comparisons.

2.9 Ethical considerations

The study was approved by the Institutional Review Boards of the National Institutes of Health, the Uganda Virus Research Institute, the Uganda National Council for Science and Technology, and Western University. All participants provided written informed consent at each study visit. All were treated according to the existing national guidelines at the time, primarily based on CD4 count before the test and start era.

3. Results

3.1 Participants, samples, and viral load

Five females and six males, aged between 31.6 and 51.1 years at the time of the first LVR sampling, were included in the study. Per the two genetic regions analyzed, ten individuals were infected with HIV-1 subtype D and one with subtype C. The duration of infection time pre-ART initiation ranged from 2.6 to 10.9 years (median 6.3 years), while the time between ART initiation and blood draws for QVOA ranged from 2.0 to 15.3 years (median 10.6 years; Table 1). Between 2 and 4 pre-ART plasma samples were available for circulating HIV sequencing, and between 1 and 3 post-ART PBMC samples were analyzed by QVOA to obtain LVR sequences. The distribution of both pre- and post-ART sampling dates varied between participants (Fig. 1). There were no cases of ART treatment failure by the time of PBMC sampling (defined as viral load >400 copies/ml, Supplementary Fig. 2).

3.2 Phylogenetic trees

For POL_RT, a median of 372 pre-ART (range 216–740) and 21 LVR (range 3–46) sequences were obtained for each participant, and 89.2 per cent of pre-ART and 65.2 per cent of LVR sequences were unique (confidence interval, CI: 80.1–98.3 and 49.4–80.9 per cent, respectively). For ENV_GP41, a median of 141 pre-ART (range 51–191) and 27 LVR (range 3–80) sequences were obtained, and 96.9 per cent of pre-ART and 48.6 per cent of LVR sequences were unique (CI: 92.5–100 and 33.8–63.3 per cent, respectively, Table 1).

Only unique pre-ART and LVR sequences were included in the phylogenetic analysis, using the earliest sampled sequence to represent subsequent identical sequences. In the trees from POL_RT, pre-ART sequences diverged from the root with time, except for Donor_88 where no divergence was observed beyond the first pre-ART sample. There was variation in the distribution of LVR sequences within the POL_RT phylogenetic trees with clustering in some individuals (e.g. Donor_73, who had no LVR sequences mapping to the earliest pre-ART sequences) and near-even distribution across sampled timepoints in others (e.g. Donor_15). LVR sequences were generally less or equally divergent from the root compared to the most divergent pre-ART sequences (Supplementary Fig. 3). In the trees from ENV_GP41, some LVR sequences from Donor_73 and Donor_85, and to a lesser extent Donor_82, were more divergent from the root compared to pre-ART sequences (Supplementary Fig. 3B). Despite differences in the temporal patterns of divergence in phylogenies created from POL_RT and ENV_GP41 (Fig. 2A), the estimates for integration dates of LVR sequences generated using the two gene regions were similar for most sequences from most individuals (Supplementary Fig. 4).

3.3 Estimated LVR integration dates

To obtain distributions of the estimated integration dates throughout the viremic period (i.e. pre-ART initiation), we took 100 estimates for each LVR sequence and combined them for each participant and for each gene region. For the seven participants with

TABLE I. (A) Participant Characteristics and Sequences from the Reverse Transcriptase Region of the Pol Gene. (B) Participant Characteristics and Sequences from the Gp41 Region of the Env Gene

ID	sex	Age (years)	HIV subtype	Duration of untreated infection (years)	Pre-ART samples (count)	Time between ART inititation and last QVOA sample (years)	LVR samples (count)	Unique pre-ART sequences (count)	Total pre-ART sequences (count)	Proportion of pre-ART sequences that are unique (%)	Unique LVR sequences (count)	Total LVR sequences (count)	Proportion of LVR sequences that are unique (%)
A										POL_RT			
Donor_15	F	51.1	D	8.4	2	15	3	359	359	100.0%	13	28	46.4%
Donor_19	M	40.2	C	5.6	2	4.5	1	329	332	99.1%	9	11	81.8%
Donor_40	F	46.6	D	10.9	2	12.7	2	369	385	95.8%	10	15	66.7%
Donor_46	M	46	D	5.8	2	15	2	317	323	98.1%	38	46	82.6%
Donor_73	F	54.1	D	8.5	3	15.5	2	567	571	99.3%	8	31	25.8%
Donor_77	F	37.4	D	2.7	3	11.1	1	284	516	55.0%	3	7	42.9%
Donor_80	M	45.2	D	10.8	4	4.8	1	740	740	100.0%	8	17	47.1%
Donor_82	M	40.4	D	6.7	2	5.1	1	329	389	84.6%	14	24	58.3%
Donor_85	M	46.7	D	6.3	2	13.9	1	286	346	82.7%	29	29	100.0%
Donor_88	M	50.3	D	5.3	2	6.3	1	167	216	77.3%	3	3	100.0%
Donor_27	F	31.6	D	1.9	2	15.3	3	–	–	–	–	–	–
B										ENV_GP41			
Donor_15	F	51.1	D	8.4	2	15	3	–	–	–	–	–	–
Donor_19	M	40.2	C	5.6	2	4.5	1	–	–	–	–	–	–
Donor_40	F	46.6	D	10.9	2	12.7	2	–	–	–	–	–	–
Donor_46	M	46	D	5.8	2	15	2	152	156	97.4%	24	40	60.0%
Donor_73	F	54.1	D	8.5	3	15.5	2	138	138	100.0%	9	19	47.0%
Donor_77	F	37.4	D	2.7	3	11.1	1	75	92	81.5%	2	8	25.0%
Donor_80	M	45.2	D	10.8	4	4.8	1	51	51	100.0%	7	14	50.0%
Donor_82	M	40.4	D	6.7	2	5.1	1	162	162	100.0%	21	27	78.0%
Donor_85	M	46.7	D	6.3	2	13.9	1	184	191	96.3%	17	27	63.0%
Donor_88	M	50.3	D	5.3	2	6.3	1	80	80	100.0%	2	3	67.0%
Donor_27	F	31.6	D	1.9	2	15.3	3	143	143	100.0%	16	32	50.0%

Footnote. QVOA- Quantitavive Viral Outgrowth Assay, LVR—replication-competent latent viral reservoir; ENV_GP41—gp41 region of the env gene; POL_RT—Reverse Transcriptase region of the pol gene.

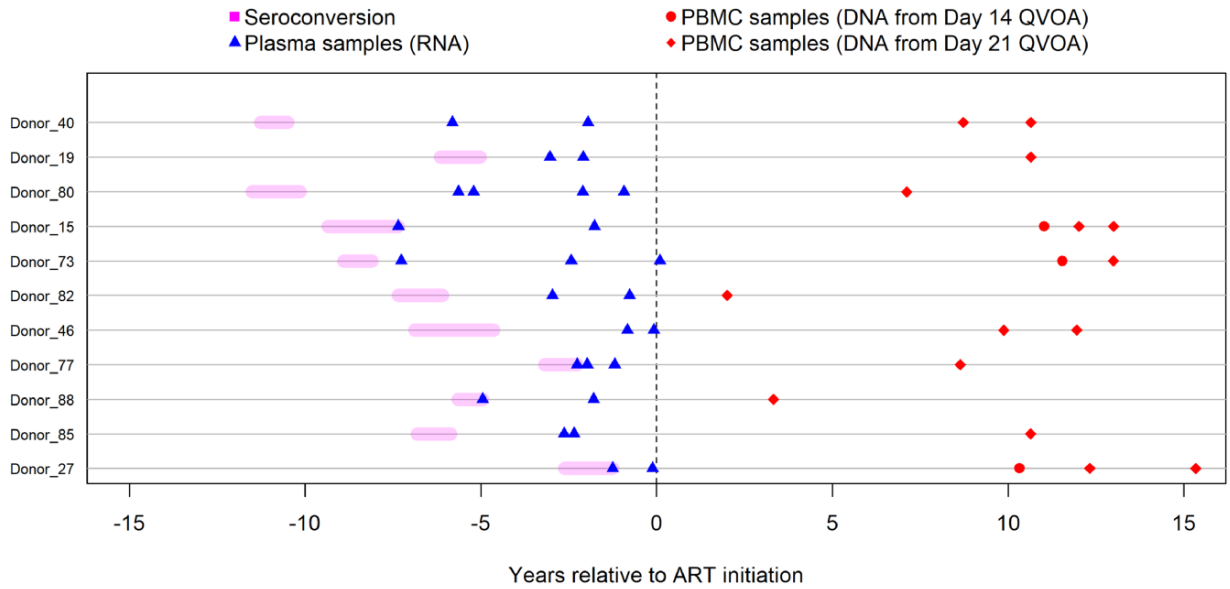

Figure 1. Pre- and post-ART sampling of participants in the Rakai LVR cohort. The seroconversion window—shown in pink—is the time between the last HIV negative test and the first positive HIV test. The blue triangles represent collection dates of serum prior to ART initiation, while the red circles represent dates of collection of peripheral blood for analysis of proviruses archived in the latent viral reservoir. Time is depicted relative to the date of ART initiation for each participant. Gene regions recovered for each participant are noted on the far right. DNA sequencing was done on Day 14 viral outgrowth where Day 21 outgrowth was not available.

phylogenies available for both regions, ENV_GP41 and POL_RT distributions were similar for 6/7 participants (large difference observed for Donor_85). In most cases, the shape of the distribution had no clear peak (Supplementary Fig. 4).

To further examine these distributions, we generated relative time categories by dividing each participant's viremic time period into tertiles. Based on POL_RT phylogenies, integration dates were approximately evenly distributed across the three tertiles for 8/10 participants, while for two participants (Donor_46 and Donor_73) a predominance of integration dates (>70 per cent) fell within the most recent tertile, closest to ART initiation (Fig. 3A). Compared to the models, only the two participants with distributions skewed toward ART initiation were consistent with rapid decay of the reservoir at half-life >44 months (Kolmogorov-Smirnov test p-values 0.147 and 0.056, respectively) and inconsistent with slower decay at half-life of 139 months (p-values 0.038 and 0.011, respectively, Supplementary Fig. 5). Similarly, when examining ENV_GP41 phylogenies, integration dates were approximately evenly distributed across the three tertiles for 6/8 participants, while for two participants (Donor_46 and Donor_85) a predominance of dates (>70 per cent) fell within the tertile closest to ART initiation (Fig. 3A) although they were not statistically significant from the slower decay model at half-life of 139 months (Kolmogorov-Smirnov test p-values 0.474 and 0.053, respectively, figure not shown). The integration date point estimates and confidence intervals for individual participant sequences are shown in Supplementary Fig. 6.

3.4 Sex differences in estimated LVR integration dates

In the Rakai LVR cohort, males appeared to have a higher proportion of proviruses with earlier integration dates when examining the ENV_GP41 region in pooled analyses [median of 1.2 years prior to ART initiation in males (IQR = 3.0 to 0.4) versus a median of 0.5 years prior to ART initiation in females (IQR = 1.1 to 0.2), $p < 0.001$]. However, this difference was not observed when using the POL_RT region [median of 1.8 years prior to ART initiation in males (IQR = 3.3 to 0.7) versus a median of 1.8 years prior to ART initiation in females (IQR = 4.2 to 0.5), $p = 0.077$].

3.5 Comparison with the CAPRISA cohort

When applying *bayroot* to the five gene regions sequenced in the CAPRISA cohort, the temporal signal (i.e. the amount of divergence between root and tip, normalized by evolution time) was strongest in the ENV_C2C3 followed by the ENV_C4C5 and NEF_1 regions (Fig. 2B). These regions are located near the ENV_GP41 region that provided the stronger temporal signal in the Rakai LVR cohort. The trees for each CAPRISA participant and for each gene region are shown in Supplementary Fig. 7.

Similar to the Rakai LVR cohort, *bayroot* found that the different gene regions provided similar estimates for a given LVR sequence's integration date (Fig. 3B and Supplementary Fig. 8). One exception was CAP316, for whom results from ENV_C2C3 showed preponderance to integration closer to ART initiation, but results from GAG_P17 were quite different, with an even distribution of integration throughout viremia.

Similar to the original analysis (Abrahams et al. 2019), the *bayroot* method found that, generally, integration dates of proviral sequences were not evenly distributed throughout viremia, but were more likely to fall in the tertile closest to ART initiation. Interestingly, this uneven distribution appears to be driven by an absence of very early sequences in CAPRISA participants, as for most participants and gene regions the proportion of sequences with integration dates falling in the middle tertile was approximately 1/3, which is what would be expected if there is an even distribution of LVR accrual throughout viremia. Compared to the models, the distributions of CAPRISA participants were generally more consistent with rapid decay of the reservoir at half-life >44 months as reported in the original analysis. The integration date point estimates and confidence intervals for individual sequences are shown in Supplementary Fig. 9.

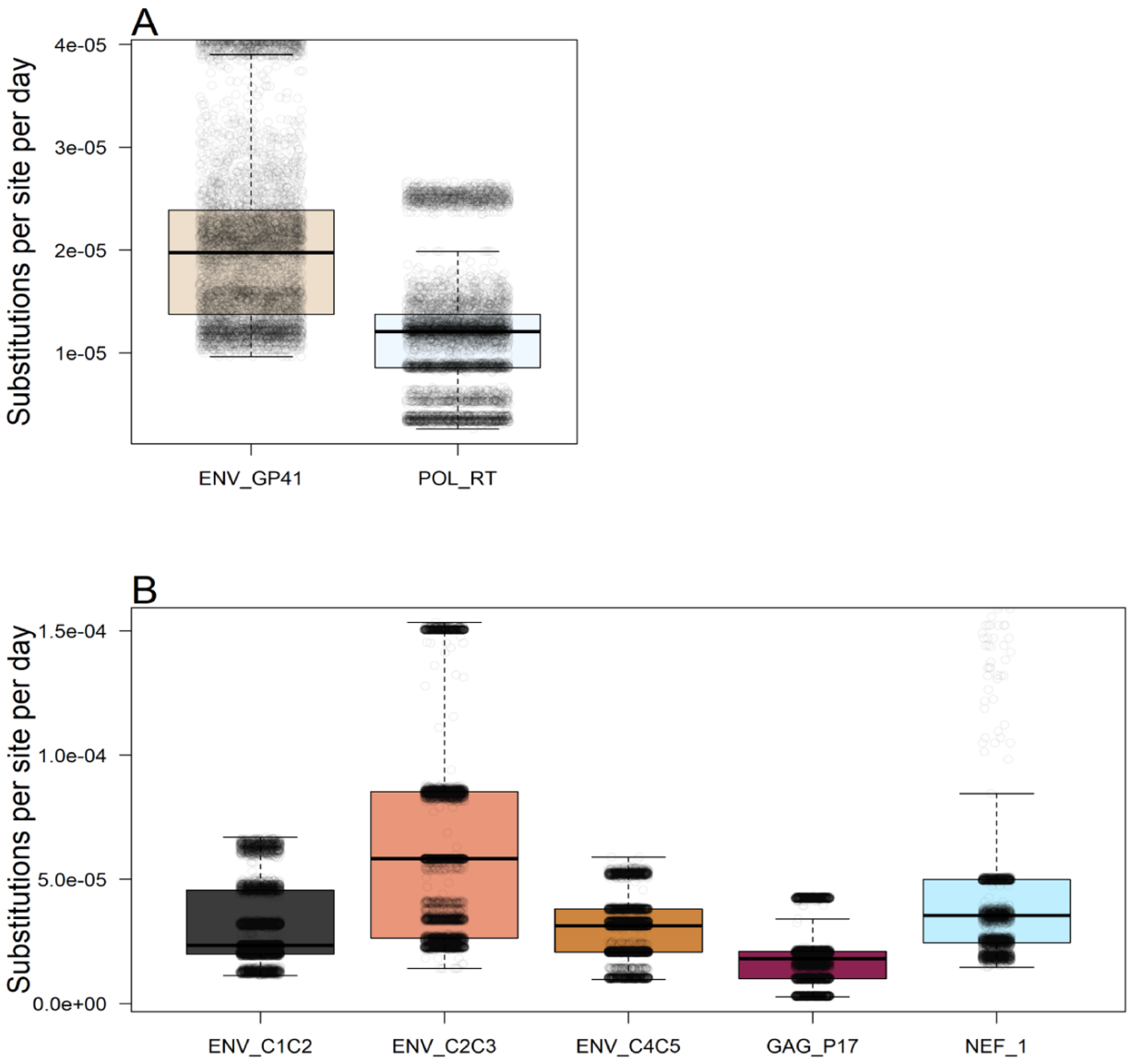

Figure 2. Comparison of the strength of the temporal signal between gene regions sequenced in (A) the Rakai LVR and (B) the CAPRISA cohort. The temporal signal refers the amount of divergence from the ancestral sequence, observed over a given period of time. The jittered points on top of the boxplots show the actual distribution of the data. Only participants who had all regions of interest sequenced were included, in each cohort.

4. Discussion

This study explored the timing of LVR formation in a cohort of male and female Ugandans. Using pre-ART serum samples, we used the novel *bayroot* analysis tool to generate a molecular clock for each participant, and then mapped proviral sequences from their LVR to this clock to estimate when the provirus originally integrated. Similar to previous studies, we observed proviral sequences from throughout untreated infection, from the time immediately after seroconversion to the time immediately preceding ART initiation (Brooks et al. 2020), suggesting that the LVR contains viral sequences archived throughout untreated infection. For most Rakai LVR cohort participants, the estimated integration dates of proviruses in the LVR were evenly distributed throughout the viremic period, consistent with slow or little decay of the reservoir. Only 2/11 participants showed evidence of a skewed distribution consistent with rapid decay of the reservoir, with a higher abundance of proviruses that integrated in the period immediately preceding ART initiation.

Previous studies examining when proviruses are archived into the LVR have suggested that most sequences in the LVR were integrated late in the viremic period, closer to the time prior to ART initiation, with fewer 'old' sequences originating from virus circulating immediately after seroconversion (Johanna et al. 2016; Abrahams et al. 2019; Brooks et al. 2020; Pankau et al. 2020). To ensure that our findings, which were in contrast to these previous reports, were not due to differences in analysis methods, we reanalyzed published data available from the CAPRISA cohort using the *bayroot* method. Previous published findings from the CAPRISA cohort were upheld with *bayroot*, suggesting that analysis methods cannot explain the difference in distribution of proviral integration dates between the two cohorts. In addition, both studies examined nearby portions of the HIV-1 genome (parts of *env*) and have a similar number of participants. Notably, the distributions in CAPRISA participants were generally consistent with rapid decay of the reservoir. In a few cases (e.g. Donor_85 in the Rakai cohort and CAP316 in the CAPRISA cohort), the distribution of proviral

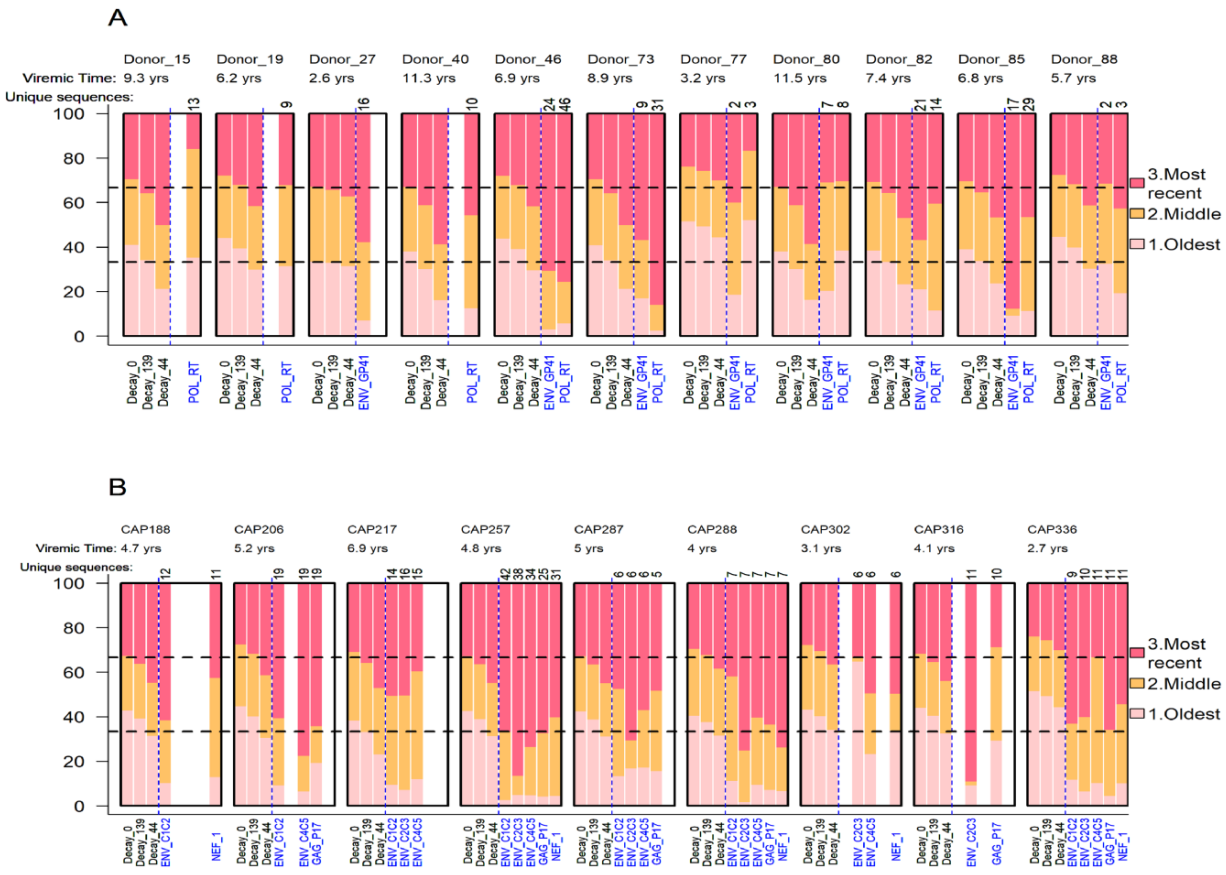

Figure 3. Distribution of integration dates throughout untreated infection in (A) the Rakai LVR cohort and in (B) the CAPRISA cohort. Distribution is depicted across relative time, where the duration of untreated infection was divided equally into three periods for each participant: the oldest tertile (pink) is immediately after seroconversion; the most recent tertile (red) immediately precedes ART initiation; and the middle tertile (orange) is between the two. For each participant, every unique LVR sequence had 100 date estimates from *bayroot* contributing to the overall distribution. Results are presented stratified by gene region (blue x-axis labels) when multiple gene regions were recovered for that participant. The black x-axis labels Decay_0, Decay_139, and Decay_44 indicate the expected distributions from mathematical models with no decay, decay with half-life of 139 months based on total HIV DNA decay (Golob et al. 2018), and decay with half-life of 44 months based on replication-competent provirus decay (Finzi et al. 1999), respectively. The blue vertical dashed line separates distributions from the models and those from patient data. The black horizontal dashed lines represent boundaries of an even distribution of integration dates across the tertiles. Supplementary Table 1 also shows the count and percentage of integration date estimates within each tertile, for each participant and gene region in the Rakai LVR cohort, corresponding to Fig. 3A.

integration dates differed markedly between gene regions. In these individuals, it appears that the gene regions experienced independent evolution, emphasizing the importance of analyzing multiple gene regions where whole genome sequencing is not available. In the Rakai LVR cohort, amplification of both POL_RT and ENV_GP41 was attempted for all participants. Unfortunately, in some cases only one region was amplified from pre-ART samples, which limited our analysis.

However, there are several key differences between the groups, which may influence these results. First, the CAPRISA cohort analyzed more regions of the genome with a stronger temporal signal (i.e. more evolution/divergence during the pre-ART period), and the lack of a strong temporal signal in the two gene regions analyzed in the Rakai LVR cohort may influence our findings. In addition, the CAPRISA cohort was exclusively females infected with subtype C, whereas the Rakai LVR cohort included males and females, predominately infected with subtype D. We could not explore if HIV subtype contributed to differences between the CAPRISA and Rakai cohorts, as there was only one subtype C participant in the Rakai LVR cohort. However, biological differences may exist between subtypes. For example, subtype C has been reported to have reduced replicative fitness, slower disease progression, increased coreceptor fidelity, and probably increased transcriptional activity compared to B-subtypes (Gartner et al. 2020). Sex-based differences in HIV pathogenesis have been reported (Das et al. 2018; Scully 2018; Scully et al. 2019; Gianella et al. 2021) and recently also differences in latent proviral reactivation (Prodger et al. 2020). In the Rakai LVR cohort, the distribution of integration dates appeared to be skewed toward earlier in viremic infection among males, but only when examining the ENV_GP41 region, and not when examining the POL_RT region, precluding any conclusions from these data. There has also been reported differences between cohorts in the estimated rates of HIV-1 superinfection, which may interfere with molecular clock analysis, with higher rates in Rakai (Redd et al. 2012) compared to CAPRISA (Redd et al. 2014). Lastly, participants in the Rakai LVR cohort were infected and viremic for a longer period of time prior to ART initiation (median 6.3 years, IQR 5.5–81) compared to the CAPRISA cohort (median 4.8 years, IQR 4.0–5.0), and Rakai participants had been on ART longer when reservoir sampling was performed (median 10.6 years, IQR 7.9–12.5) compared to the CAPRISA cohort (median 5.0 years, IQR 4.6–5.3). There is a paucity of data on the effects of sex and HIV subtype on HIV persistence in general, but especially with respect to LVR dynamics in

long-term infection. The discordant results from these two cohorts emphasize the need for more collaborative studies of the proviral landscape in diverse, global cohorts.

In addition to differences between cohorts, the Rakai LVR cohort participants had less frequent pre-ART samples available for generating phylogenies compared to the CAPRISA study and most other previous studies (Johanna et al. 2016; Abrahams et al. 2019; Brooks et al. 2020; Pankau et al. 2020). However, the *bayroot* method enabled recovery of integration date estimates with less intense pre-ART sampling by restricting estimates to the time between seroconversion and ART initiation. Also, multiple date estimates per unique sequence ($n = 100$) accounts for uncertainty surrounding estimated integration dates, compared to point estimates reported in previous studies, which can be misleading (Ferreira, Wong, and Art 2023). One strength of the Rakai LVR cohort is that multiple LVR samples (post-ART initiation) were available for many participants, with LVR quantification samples collected up to 15.3 years after ART initiation. This testing well after the initial rapid decay period following ART initiation should allow for a more accurate picture of the long-lived replication-competent LVR (Strain et al. 2005; Besson et al. 2014).

Finally, it should be noted that the relatively limited number of participants examined in both the Rakai LVR and CAPRISA cohorts, as well as in the other previous cohorts from Europe, Kenya, and the Americas, limits conclusions that can be made about LVR formation patterns (Johanna et al. 2016; Jones et al. 2018; Brooks et al. 2020; Pankau et al. 2020). However, current treatment guidelines preclude larger, purpose-designed prospective studies to address this question using molecular clocks of pre-ART viral evolution. This study suggests that there is variability in the timing of reservoir formation between populations, and that reservoir formation is not weighted to the time preceding ART initiation in all people living with HIV. These findings highlight the need for additional investigation in more cohorts with longitudinal pre-ART serum samples if available.

Data availability

Sequences are available in Genbank, accession numbers pending. Supplementary data are available at Virus Evolution online.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Acknowledgements

We thank the participants for their repeat voluntary participation, and the staff of the Rakai Health Sciences Program who collected these data. We would also like to thank CAPRISA investigators, especially Dr Melissa-Rose Abrahams, for providing CAPRISA cohort sequence data and helpful discussions.

Funding

This study was supported in part by the Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health (grant # UM1 AI164565) and by Gilead Sciences (grant # 327799.1). EK received training and salary support from the Foundation for AIDS research (amfAR, grant #109844) and the Fogarty Internal Center (D43TW010557). JP received salary support through a Canada Research Chair, Tier 2, from the Canadian Institutes for Health Sciences (grant # 950 – 233211).

Conflict of interest: All authors have no commercial or other association that might pose a conflict of interest.

References

Abrahams, M.-R. et al. (2019) 'The Replication-Competent HIV-1 Latent Reservoir Is Primarily Established near the Time of Therapy Initiation', *Science Translational Medicine*, 11: eaaw5589.

Alamos, L. (2022) *Hypermut Hypermutation Analysis Page* <https://www.hiv.lanl.gov/content/sequence/HYPERMUT/hypermut.html> accessed 25 Sep 2022.

Archin, N. M. et al. (2012) 'Immediate Antiviral Therapy Appears to Restrict Resting CD4+ Cell HIV-1 Infection without Accelerating the Decay of Latent Infection', *Proceedings of the National Academy of Sciences of the United States of America*, 109: 9523–8.

Besson, G. J. et al. (2014) 'HIV-1 DNA Decay Dynamics in Blood during More than a Decade of Suppressive Antiretroviral Therapy', *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 59: 1312–21.

Brooks, K. et al. (2020) 'HIV-1 Variants are Archived Throughout Infection and Persist in the Reservoir', *PLOS Pathogens*, 16: e1008378.

Bruner, K. M. et al. (2016) 'Defective Proviruses Rapidly Accumulate during Acute HIV-1 Infection', *Nature Medicine*, 22: 1043–9.

Chun, T. W. et al. (1997) 'Quantification of Latent Tissue Reservoirs and Total Body Viral Load in HIV-1 Infection', *Nature*, 387: 183–8.

Courtney, C. R. et al. (2017) 'Contrasting Antibody Responses to Intrasubtype Superinfection with CRF02_AG', *PLoS One*, 12: e0173705.

Das, B. et al. (2018) 'Estrogen Receptor-1 Is a Key Regulator of HIV-1 Latency that Imparts Gender-Specific Restrictions on the Latent Reservoir', *Proceedings of the National Academy of Sciences of the United States of America*, 115: E7795–7804.

Evelyn, E., and Siliciano, R. F. (2012) 'Redefining the Viral Reservoirs that Prevent HIV-1 Eradication', *Immunity*, 37: 377–88.

Ferreira, R.-C., Wong, E., and Art, F. Y. P. (2023) 'Bayroot: Bayesian Sampling of HIV-1 Integration Dates by Root-to-Tip Regression', *Virus Evolution*, 9: veac120.

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', *Nature Medicine*, 5: 512–7.

Gartner, M. J. et al. (2020) 'Understanding the Mechanisms Driving the Spread of Subtype C HIV-1', *EBioMedicine*, 53: 102682.

Gianella, S. et al. (2021) 'Sex Differences in Human Immunodeficiency Virus Persistence and Reservoir Size during Aging', *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 75: 73–80.

Golob, J. L. et al. (2018) 'HIV DNA Levels and Decay in a Cohort of 111 Long-Term Virally Suppressed Patients', *AIDS*, 32: 2113.

Johanna, B. et al. (2016) 'Establishment and Stability of the Latent HIV-1 DNA Reservoir', *ELife*, 5: e18889.

Jones, B. R. et al. (2018) 'Phylogenetic Approach to Recover Integration Dates of Latent HIV Sequences Within-Host', *Proceedings of the National Academy of Sciences of the United States of America*, 115: E8958–67.

Kankaka, E. N. et al. (2022) 'Makerere's Contribution to the Development of a High Impact HIV Research Population-Based Cohort in the Rakai Region, Uganda', *African Health Sciences*, 22: 42–50.

Kazutaka, K., and Standley, D. M. (2014) 'MAFFT: Iterative Refinement and Additional Methods', *Methods in Molecular Biology (Clifton, N.J.)*, 1079: 131–46.

Laird, G. M. et al. (2016) 'Measuring the Frequency of Latent HIV-1 in Resting CD4$^+$ T Cells Using a Limiting Dilution Coculture Assay', *Methods in Molecular Biology (Clifton, N.J.)*, 1354: 239–53.

Pankau, M. D. et al. (2020) 'Dynamics of HIV DNA Reservoir Seeding in a Cohort of Superinfected Kenyan Women', *PLoS Pathogens*, 16: e1008286.

Persaud, D. et al. (2014) 'Age at Virologic Control Influences Peripheral Blood HIV Reservoir Size and Serostatus in Perinatally-Infected Adolescents', *JAMA Pediatrics*, 168: 1138–46.

Poon, A. F. Y. et al. (2018) 'Quantitation of the Latent HIV-1 Reservoir from the Sequence Diversity in Viral Outgrowth Assays', *Retrovirology*, 15: 47.

Prodger, J. L. et al. (2020) 'Reduced HIV-1 Latent Reservoir Outgrowth and Distinct Immune Correlates among Women in Rakai, Uganda', *JCI Insight*, 5: e139287.

——— et al. (2017) 'Reduced Frequency of Cells Latently Infected with Replication-Competent Human Immunodeficiency Virus-1 in Virally Suppressed Individuals Living in Rakai, Uganda', *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 65: 1308–15.

Redd, A. D. et al. (2012) 'The Rates of HIV Superinfection and Primary HIV Incidence in a General Population in Rakai, Uganda', *The Journal of Infectious Diseases*, 206: 267–74.

——— et al. (2014) 'Limited HIV-1 Superinfection in Seroconverters from the CAPRISA 004 Microbicide Trial', *Journal of Clinical Microbiology*, 52: 844–8.

Scully, E. P. (2018) 'Sex Differences in HIV Infection', *Current HIV/AIDS Reports*, 15: 136–46.

——— et al. (2019) 'Sex-Based Differences in Human Immunodeficiency Virus Type 1 Reservoir Activity and Residual Immune Activation', *The Journal of Infectious Diseases*, 219: 1084–94.

Shan, L. et al. (2017) 'Transcriptional Reprogramming during Effector-to-Memory Transition Renders CD4+ T Cells Permissive for Latent HIV-1 Infection', *Immunity*, 47: 766–775.e3.

Siliciano, J. D. et al. (2003) 'Long-Term Follow-up Studies Confirm the Stability of the Latent Reservoir for HIV-1 in Resting CD4+ T Cells', *Nature Medicine*, 9: 727–8.

Stamatakis, A. (2014) 'RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies', *Bioinformatics*, 30: 1312–3.

Strain, M. C. et al. (2005) 'Effect of Treatment, during Primary Infection, on Establishment and Clearance of Cellular Reservoirs of HIV-1', *The Journal of Infectious Diseases*, 191: 1410–8.

van Beveren, C. P., Coffin, J., and Hughes, S. (2002) *Human Immunodeficiency Virus Type 1 (HXB2), Complete Genome; HIV1/HTLV-III/LAV Reference Genome* (1906382, NCBI Nucleotide Database) <http://www.ncbi.nlm.nih.gov/nuccore/K03455.1> accessed 31 Jul 2022.