# HIV proviral burden, genetic diversity and dynamics in viremic controllers who subsequently initiated suppressive antiretroviral therapy

F. Harrison Omondi[a,b], Hanwei Sudderuddin[b], Aniqa Shahid[a,b], Natalie N. Kinloch[a,b], Bradley R. Jones[b,c], Rachel L. Miller[b,c], Olivia Tsai[a], Daniel MacMillan[b], Alicja Trocha[d], Mark A. Brockman[a,e], Chanson J. Brumme[b,f], Jeffrey B. Joy[b,c,f], Richard Liang[b], Bruce D. Walker[d], Zabrina L. Brumme[a,b#]

[a] Faculty of Health Sciences, Simon Fraser University, Burnaby, BC, Canada

[b] British Columbia Centre for Excellence in HIV/AIDS, Vancouver, BC, Canada

[c] Bioinformatics Program, University of British Columbia, Vancouver, BC, Canada

[d] Ragon Institute of MGH, MIT and Harvard, Cambridge, MA, USA

[e] Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada

[f] Department of Medicine, University of British Columbia, Vancouver, BC, Canada

Running title: Proviral dynamics and longevity in viremic controllers

#Address correspondence to Zabrina L. Brumme, zbrumme@sfu.ca.

F.H.O. and H.S. contributed equally to this work. F.H.O. is listed first as he wrote the first draft of the manuscript.

Abstract Word Count: 250

Text Word Count: 4382

**Abstract**

Curing HIV will require eliminating the reservoir of integrated, replication-competent proviruses that persist despite antiretroviral therapy. Understanding the burden, genetic diversity and longevity of persisting proviruses in diverse individuals with HIV is critical to this goal, but these characteristics remain understudied in some groups. Viremic controllers, individuals who naturally suppress HIV to low levels but for whom therapy is nevertheless recommended, represent one such group who could yield insights into proviral seeding and turnover in the setting of natural yet incomplete HIV control. We reconstructed within-host HIV evolutionary histories from single-genome amplified viral sequences in four viremic controllers who eventually initiated therapy, and leveraged this information to characterize the diversity and longevity of proviruses persisting on therapy. Despite natural viremic control, all participants displayed significant within-host HIV evolution pre-therapy, where the overall burden and diversity of proviruses persisting on therapy reflected the extent of viral replication, and plasma viral diversity generated, during untreated infection. While the proviral pools of two participants were skewed towards sequences that integrated near ART initiation, proviruses in the other two participants dated to various time-points that were more evenly spread throughout infection and included sequences that integrated close to transmission. Estimated *in vivo* proviral half-lives were <1 year for two participants and >2 years for a third, consistent with dynamic proviral turnover during untreated infection; the fourth was consistent with negligible proviral decay following deposition. HIV cure strategies will need to overcome within-host proviral diversity, even in individuals who naturally controlled HIV replication before therapy.

## Importance

HIV therapy is life-long because integrated, replication-competent viral copies persist within long-lived cells. To cure HIV, we need to understand when these viral reservoirs form, how large and genetically diverse they are, and how long they endure *in vivo*. Elite controllers, individuals who naturally suppress HIV to undetectable levels, have been intensely studied, in part because their small reservoirs may make HIV eradication more achievable. Viremic controllers, individuals who naturally suppress HIV to low levels, may also hold insights but remain understudied. We used a phylogenetic approach to reconstruct reservoir seeding and decay in four viremic controllers from infection to HIV suppression on therapy. In all participants, we recovered diverse HIV sequences persisting during therapy that broadly reflected HIV's evolutionary history during untreated infection. Even in individuals who naturally controlled HIV replication prior to therapy, HIV cure strategies will need to contend with diverse proviral variants archived throughout infection.

## Background

Like all retroviruses, HIV integrates its genome into that of its host cell. Most infected cells die − or are eliminated by the immune system − within a day or two of infection, but a small number persist long-term, even during suppressive antiretroviral therapy (ART) (1–3). While most of these long-lived cells harbor HIV proviruses with large deletions or other genetic defects (4–6), a minority harbor replication-competent proviruses that could reactivate at any time. It is for this reason that ART needs to be maintained for life. Achieving ART-free HIV remission ("functional cure") or a full, sterilizing cure (7) will therefore require permanently inactivating (8, 9) or eliminating (10, 11) these HIV reservoirs, respectively. Characterizing the burden, genetic diversity and longevity of the proviruses that persist within-host can advance us towards these goals.

HIV elite controllers, rare individuals who spontaneously control viremia to undetectable levels without ART, may hold clues to move HIV cure efforts forward. They represent a model of natural functional HIV cure (12) that could help inform ART-free HIV remission strategies, and they may also represent a group in which sterilizing cure might be more easily achieved (13, 14), in part because their HIV reservoirs are smaller and less genetically diverse (8, 12, 15). Proviral burden and diversity however are less well characterized in viremic HIV controllers, individuals who naturally suppress HIV to low levels (normally defined as <2000 copies/ml of plasma (16, 17)), but who are nevertheless recommended to initiate ART under current guidelines (18). In particular, this group could yield insights into persisting proviral composition in the context of natural yet incomplete HIV control. More broadly, our understanding of within-host proviral *dynamics* − that is, the rates of proviral deposition and subsequent decay − remains incomplete for HIV controllers in general. This is in part because longitudinal studies dating

back to infection are rare in this group, and because it is challenging to isolate HIV sequences from samples with low or undetectable viral loads.

In recent years, our understanding of within-host proviral dynamics has been enriched by the development of genetic and phylogenetic approaches to analyze persisting proviral diversity in context of HIV's within-host evolution prior to ART (19–23). A recent study that recovered replication-competent proviruses during therapy revealed that the majority represented HIV variants that circulated close to therapy initiation, though more ancestral sequences, some dating back to transmission, were also recovered (21). Studies employing similar approaches to "date" the overall pool of proviruses persisting on ART, which comprise both genetically intact and defective sequences, have confirmed that these often span the individual's whole pre-ART history, where individuals differ in the extent to which their proviral pool is "skewed" towards sequences that date to around ART initiation (19, 20, 22, 23). Rates of proviral turnover during untreated infection can also be inferred from the age distributions of proviruses sampled shortly after ART (19, 23), with recent data from non-controllers suggesting that proviral half-lives during this period average less than a year (19), though another study estimated an average of 25 months (23). Because viremic controllers naturally limit viral replication to low levels, one could hypothesize that their persisting proviral pools may be smaller and less genetically diverse than those of non-controllers, but proviral diversity and dynamics have not been investigated in this group using phylogenetic approaches.

Here, we combined single-genome sequencing, proviral quantification, phylogenetic analysis, and mathematical modeling to characterize proviral burden, diversity and dynamics in four viremic controllers. Despite limited viral replication, significant within-host HIV evolution nevertheless occurred in all four participants pre-ART, giving rise to within-host proviral pools whose size and genetic diversity reflected this evolution. Moreover, though estimated pre-ART

proviral turnover rates varied between individuals, estimates from three of the four participants were not substantially different from recent estimates from non-controllers (23). Our observations underscore proviral diversity as a barrier that HIV remission and cure strategies will need to overcome, even in individuals who naturally control HIV replication prior to therapy.

## Results

### Participant characteristics and sampling

We studied four viremic controllers: three who broadly maintained plasma viral loads (pVL) <2000 copies HIV RNA/ml during untreated infection (participants 1, 2, 3), and one who maintained pVL <2000 copies HIV RNA/ml for 3 years before losing viremic control (participant 4). Participants initiated ART a mean of 6.1 (range 3.8-8.4) years following their estimated date of infection (EDI). An average of 13 (range 10-17) pre-ART plasma samples per participant, that spanned a mean 4.5 (range 2.5-6.7) year period prior to therapy, were available for HIV RNA sequencing, where the first was sampled a mean of 20.5 (range 15-26) months following the EDI (**Figure 1** and **Table 1**). In addition, PBMC were available at two timepoints on suppressive ART (10 Million cells per timepoint). Due to limited cell numbers, the first PBMC sample, taken an average of 18.7 (range 9-38) months after ART, was used to quantify each individual's total and genomically intact proviral burden in CD4+ T cells (24), while the second, taken an average of 28.8 (range 16-54) months after ART, was used for proviral sequencing and integration date inference (20).

### Quantifying total and genetically intact proviral burden

We quantified total and genomically intact proviral burden in CD4+ T cells using the Intact Proviral DNA Assay (IPDA). This is a duplexed droplet digital PCR (ddPCR) assay that targets two HIV regions, the Packaging Signal (Ψ) near the 5' end of the viral genome and the Rev Responsive Element (RRE) within Envelope (*env*) that, together, have high predictive power to distinguish genomically intact proviruses from those with large deletions and/or extensive hypermutation, which typically dominate *in vivo* (24). Proviruses yielding double (Ψ and *env*) positive signals are inferred to be genetically intact while those yielding only single positive signal are classified as defective.

We observed marked inter-individual differences in total proviral burden. Participant 3 harbored the fewest proviruses (44 HIV DNA copies/$10^6$ CD4+ T cells) while participant 4, the individual who lost their controller status, harbored the most (778 HIV DNA copies/$10^6$ CD4+ T cells) (**Figure 1** and **Table 1**). The percent of genomically intact proviruses also differed markedly, with participant 2 harboring the lowest (10% intact proviruses) and participant 4 the highest (94% intact). The latter is remarkable though not without precedent (25). Participant 4's large overall and intact proviral burden allowed us to isolate one near-full-length intact provirus from this timepoint (**Figure 1**) for integration date inference.

**Single-genome plasma HIV RNA and proviral sequencing**

We next characterized HIV RNA *nef* sequences from longitudinal pre-ART plasma, and proviral sequences from on-ART PBMC, by single-genome amplification. HIV *nef* was sequenced because of its richness in phylogenetic signal despite its relatively short size (a shorter amplicon was critical as viral loads were low for most samples), because *nef* is representative of within-host HIV diversity and evolution relative to the rest of the viral genome (20), and because it is the gene most likely to be intact in proviruses sampled on ART (24, 26). Despite low viral loads we isolated 546 HIV RNA *nef* sequences from pre-ART plasma samples (range 67-245/participant), where an average of 66.2% of these (range 50.0-89.6%) were distinct within that participant (**Table 1**). We further isolated 267 intact proviral *nef* sequences during suppressive ART (range 24-129/participant), where an average of 69.8% (range 50.0-91.5%) were distinct within that participant. All participants had HIV subtype B, and each participant's sequences were monophyletic, with no evidence of co-infection or superinfection (**Figure 2**). As expected, there were multiple instances where we recovered a proviral sequence on ART that was genetically identical to a pre-ART plasma sequence, consistent with the long-term persistence of these sequences *in vivo*.

**Within-host HIV evolutionary reconstruction and proviral dating**

We next characterized the diversity of proviruses persisting on ART, and inferred their within-host ages using a phylogenetic approach (20). For each participant, we inferred a maximum likelihood phylogeny relating their plasma HIV RNA and proviral sequences and identified a root location that represented the inferred transmitted founder viral sequence. We then fit a linear model relating the root-to tip distances of distinct pre-ART plasma HIV RNA sequences to their collection dates, where the slope of this line represents the within-host HIV *nef* evolutionary rate under a strict molecular clock assumption, and the x-intercept represents the phylogenetically-estimated infection date. Overall, we observed statistically significant within-host HIV evolution in plasma in all participants pre-ART, as defined by increasing divergence from the inferred transmitted founder viral sequence, where the 95% confidence interval (CI) of the phylogenetically-estimated infection date captured the participant's clinically estimated one in all cases (**Table 2**). We next describe each participant's results in detail.

Participant 1 (EDI April 2006) maintained viremic control except for "blips" to 8075 and 2685 copies/ml in February and October 2008 respectively, prior to initiating ART in August 2010 (**Figure 3A**). Their proviral load on suppressive ART was 215 HIV copies/$10^6$ CD4+ T cells, with 51% genetically intact (**Figure 1; Table 1**). We isolated 104 intact HIV RNA *nef* sequences from 8 pre-ART plasma timepoints spanning 2.4 years, of which 52 (50%) were distinct, and 48 proviral *nef* sequences sampled approximately two years post-ART, of which 40 (83%) were distinct (**Table 1**). The recovery of identical proviral sequences is consistent with clonal expansion of CD4+ T cells harboring integrated HIV (27–29), though we cannot definitively identify these as clonal since we only performed subgenomic HIV sequencing. Within-host phylogenetic analysis revealed increasing plasma HIV RNA divergence from the inferred root over time, along with non-synonymous substitutions that typify within-host HIV

evolution (**Figure 3B**). Notably, proviruses sampled on ART interspersed throughout the phylogeny except within the clade closest to the root, which comprised the earliest sampled plasma HIV RNA sequences. The linear model relating the root-to-tip distances of distinct pre-ART plasma HIV RNA sequences to their collection dates yielded a pre-ART *nef* evolutionary rate of $3.22 \times 10^{-5}$ substitutions per nucleotide site per day (**Figure 3C** and **Table 2**). Based on this, the oldest sampled provirus was estimated to have integrated in August 2008, 4 years prior to sampling (**Figure 3D**), though the 95% CIs around the proviral date estimates are wide. Nevertheless, these data suggest that proviruses seeded throughout untreated infection persisted during ART.

Participant 2 (EDI December 2005) maintained pVL <2000 copies/ml except for "blips" to 2160 and 3230 copies/ml in June 2010 and September 2013 respectively, and initiated ART in May 2014 (**Figure 4A**). Their total proviral load on ART was 580 HIV copies/$10^6$ CD4+ T cells, with only 10% intact, the smallest intact proportion of all participants (**Figure 1; Table 1**). We recovered 67 HIV RNA *nef* sequences from 9 pre-ART plasma timepoints spanning more than four years, of which 60 (90%) were distinct, and 66 intact proviral *nef* sequences after 1.2 years on ART, of which 36 (55%) were distinct. Of note, the sequences closest to the root of this participant's phylogeny were proviruses sampled on ART, consistent with these having integrated shortly after transmission (**Figure 4B**). Proviruses also interspersed throughout almost all descendant clades. Based on this individual's calculated pre-ART *nef* evolutionary rate of $1.17 \times 10^{-5}$ substitutions per nucleotide site per day (**Figure 4C** and **Table 2**), proviruses persisting on ART were inferred to have integrated on dates that spanned the individual's entire pre-ART history (**Figure 4D**). One particular proviral sequence, recovered 26 times, was estimated to have integrated 6 years prior to sampling.

Participant 3 (EDI Mar 2008) continuously maintained pVL <1000 copies/ml before initiating ART in January 2012. Thereafter, pVL remained suppressed except for a "blip" to 1610 copies/ml in November 2013, but unfortunately no sample was available from that event. This participant had the smallest proviral burden of all (44 HIV copies/$10^6$ CD4+ T cells), with an estimated 31% genetically intact (**Figure 1; Table 1**). We recovered 130 pre-ART plasma HIV *nef* sequences (55% distinct) from 7 pre-ART timepoints, where early plasma timepoints more frequently yielded identical HIV sequences, consistent with limited viral diversity following infection. We recovered only 24 proviral sequences (50% distinct) on ART, consistent with the small proviral burden. Despite maintenance of pVL<1000 pre-ART, the within-host phylogeny displayed significant molecular clock signal ($1.16 \times 10^{-5}$ substitutions per nucleotide site per day) where, similar to participant 2, proviruses sampled on ART interspersed throughout the tree and represented those closest to the root (**Figures 5B, 5C**). The earliest proviruses dated to around transmission while the latest dated to the viremic episode following ART initiation; the narrower 95% confidence intervals around these point estimates allow us to say with more certainty that this individual's proviral pool spans a wide age range (**Figure 5D**). One sequence, recovered 9 times on ART and representing 38% of recovered proviruses in this individual, dated to around ART initiation.

Participant 4 (EDI March 2005) initially controlled pVL to <2000 copies/ml for 3 years, but afterwards lost control, where pVL reached a maximum of 65,500 copies/ml before initiating ART in February 2013 (**Figure 6A**). Two linear models were therefore calibrated, one each for the "control" and "post-control" periods, where the former was used to phylogenetically verify the clinically estimated infection date. We recovered 245 pre-ART plasma HIV *nef* sequences from 12 pre-ART timepoints (71% distinct), where, like participant 3, identical HIV sequences were most frequently recovered from early timepoints, consistent with limited HIV diversity

following infection (**Figure 6B)**. We also recovered 129 proviral sequences during ART, a remarkable 92% of which were unique, supporting substantial proviral diversity in this individual. The linear model inferred from the viremic control period yielded strong molecular clock signal despite low viremia during this period, but no proviral sequences were recovered that interspersed with these very early sequences (**Figures 6B, 6C**). All recovered proviruses were therefore "dated" using the linear model representing the post-control period, yielding integration dates spanning this whole period (**Figure 6C**). Of note, the near-full-length genetically-intact provirus isolated during ART dated to 2011.

**Correlates of proviral burden and diversity during ART**

Early ART limits HIV reservoir size (30–32), and positive correlations have been observed between the duration of uncontrolled viremia, and both HIV reservoir size and diversity in non-controllers (32–36). We therefore investigated the relationship between proviral burden, genetic diversity and cumulative viral load in our cohort. Overall, we observed a strong correlation between HIV genetic diversity ever observed in plasma pre-ART, and proviral diversity on-ART (Spearman's $\rho=1$; $p=0.08$; **Figure 7A**). We further observed that cumulative viral load, measured as $\log_{10}$ viremia copy-days during untreated infection, correlated strongly with both total proviral burden as measured by the IPDA as well as proviral diversity on ART (both Spearman's $\rho=1$; $p=0.08$; **Figures 7B, C**).

**Inferring proviral turnover rates during untreated HIV infection**

At the time of sampling, the majority of proviral turnover would have already occurred prior to ART. As such, we can infer *in vivo* proviral turnover rates from our data. We did this in two ways: using a mathematical modeling approach, and directly from the data. To do the former, we modified an existing mathematical model that assumes that, during untreated infection, HIV sequences enter the proviral pool at a rate proportional to their abundance in

plasma, after which they are cleared at a constant rate (see methods and (23)). While the original approach modeled pre-ART viral load dynamics using typical parameterization appropriate for non-controllers (23), we used a piecewise linear function that merged each participant's pVL data with acute phase dynamics drawn from the (limited) literature on HIV controllers (37–39) (see methods and **Table S1**). **Figures 8A-D** show the participants' viral load histories incorporating these acute-phase dynamics, while **Figures 8E-H** show their proviral age distributions on ART, depicted as the proportion of distinct proviruses that dated to each year prior to ART (or, in the case of participant 3, to each year prior to their viremic episode after ART), as blue bars. We then used the model to predict what each participant's proviral age distributions would have "looked like" under different decay rates, under the assumption that proviral seeding during untreated infection occurred proportionally to viral load, and subsequent decay was constant at that rate. To provide visual context, predicted proviral age distributions assuming no decay**,** as well as those assuming proviral half-lives of 140 (40) and 44 (41) months (values that represent published estimates of DNA and replication-competent reservoir decay *during suppressive ART*), are shown in **Figures 8E-H**. We then identified the half-life that best fit each participants' proviral age distribution, shown as the green line in **Figures 8E-H (**95% confidence intervals shown in **Table S1**). Participants 1 and 4, for whom half of their sampled proviruses dated to the year preceding ART, had the shortest model-estimated proviral half-lives of 0.49 (95% CI 0-1.42) years and 0.82 (95% CI 0-2.84) years respectively (**Figures 8E, 8H**). Participant 3, for whom proviruses were recovered that dated to shortly following infection and whose reservoir was reseeded during their post-ART viremic episode, had a substantially longer estimated proviral half-life of 2.2 (95% CI 0-10.03) years (**Figure 8G**). The best fit to participant 2, whose proviruses dated throughout their 9-year pre-ART history, was the model with no decay, though even this fit was suboptimal (**Figures 8F).** The fact that this participant had the highest defective

proviral burden (90%) may had contributed to this negligible decay rate, but we cannot rule out that the model's assumptions were a poor fit for their proviral dynamics. Sensitivity analyses incorporating different acute-phase dynamics yielded similar results (**Table S1**), suggesting that our reconstructions of this phase did not influence model-estimated half-lives.

Finally, we inferred proviral half-lives directly from the observed proviral age distributions using a Poisson generalized linear model (**Figures 8I-L).** Consistent with the model-derived estimates, participants 1 and 4 had the shortest proviral half-lives of 0.63 (95% CI 0.46-0.97) years and 0.76 (95% CI 0.64-0.94) years respectively, participant 3's was the next longest at 4.79 (95% CI 1.44- $\infty$) years, while participant 2's was the longest at 23.74 (95% CI 4.44- $\infty$) years. The agreement between the model-predicted and directly-calculated proviral half-lives was excellent (Spearman's correlation $\rho=1$, $p=0.08$).

## Discussion

Though modest in size, our longitudinal cohort of ART-treated viremic controllers for whom dates of infection could be reliably estimated allowed us to characterize proviral burden, genetic diversity and dynamics in this understudied group. Overall, our observations indicate that the persisting proviral pools within viremic controllers on ART can share considerable similarities in terms of size, diversity and turnover with those from non-controllers on ART.

Within-host HIV evolutionary studies have revealed that the majority of infections are initiated by a single transmitted founder virus that gives rise to increasingly diverse and divergent descendants (42–48). The extent and rates of within-host HIV evolution in controllers however remain somewhat unclear. While some studies performed on shorter timeframes have reported increases in plasma HIV diversity and/or the selection of immune escape mutations over time in elite (49–51) and viremic (52) controllers, others have found no or limited evidence of evolution within-host (12, 15, 53). The first notable observation from our study was that significant within-host HIV evolution occurred in all participants, as evidenced by ongoing divergence from a phylogenetically-inferred root whose date was consistent with the clinically-estimated infection date in all cases. Also indicative of within-host evolution, non-synonymous substitutions consistent with escape from Human Leukocyte Antigen (HLA) class I-restricted cellular immune responses occurred even in the participant in whom pre-ART plasma viral loads never reached above 1000 HIV RNA copies/ml (participant 3). In this participant for example who expresses both HLA-A*03:01 and HLA-B*57:01, 25/25 sequences sampled in 2009 harbored the predicted HLA-B*57:01-restricted epitope AGNNAACAW at Nef codons 49-57, which is predicted to bind B*57:01 with an affinity of ~1400nM (54). By 2011 however, 8/27 sequences harbored A***E***NNAACAW, a predicted B*57:01 escape variant (affinity ~15,000 nM), where this shift in frequency was statistically significant (Fisher's exact test; p=0.0044).

Similarly, 25/25 sequences sampled in 2009 harbored the predicted A*03:01 epitope KLVPVPEK at Nef codons 144-152 (predicted binding affinity ~170nM), but by late 2011, 6/16 sequences harbored the predicted A*03:01 escape variant KLVPVPE***E*** (affinity ~28,000 nM), a significant frequency shift (p=0.0069). Participants' estimated within-host *nef* evolutionary rates, which ranged from $1.16 \times 10^{-5}$ to $5.35 \times 10^{-5}$ substitutions per nucleotide site per day were also comparable to those reported in non-controllers using similar methods (20). Together, these observations indicate that, despite natural viremia control, significant within-host HIV evolution nevertheless occurs in this group in the absence of therapy.

Our results also revealed that total and genomically intact proviral DNA burdens in ART-treated viremic controllers can be substantial. While few studies have estimated reservoir size in controllers using the IPDA, a recent study reported 20-fold lower frequencies of total and intact proviral DNA in elite controllers (median of 30 and 1.6 copies/$10^6$ $CD4^+$ T cells respectively) compared to non-controllers on ART (median of 603 and 37 copies/$10^6$ $CD4^+$ T cells respectively) (55). In the present study by contrast, the median total and intact proviral burdens were 398 and 84 copies/$10^6$ $CD4^+$ T cells respectively (or 215 and 58 copies/$10^6$ $CD4^+$ T cells respectively if we excluded participant 4 who lost viremic control prior to ART), values that were more in line with those in non-controllers on ART (25, 55, 56). It is important to note however that proviral loads in our study were sampled only ~1.5 years after ART on average. As studies of non-controllers using the IPDA have revealed that, during long-term ART, intact proviruses gradually decay while total proviral DNA remains stable (25, 56, 57), longitudinal studies of treated viremic controllers will be required to elucidate whether on-ART proviral decay dynamics are similar in this group.

Our results also yielded insights into proviral diversity during ART, as well as the length of time these sequences had persisted within-host. Similar to studies of non-controllers (19–23),

on-ART proviral age distributions varied among the participants of our study: whereas participants 1 and 4's proviral pools were skewed towards sequences that integrated in the year or two prior to ART, participants 2 and 3's proviral pools spanned the infection course more evenly, and included proviruses that dated to shortly following infection. Despite these differences, the overall diversity of unique proviruses persisting on ART reflected the overall plasma HIV diversity generated prior to ART (**Figure 7A**). Moreover, the overall level of viral replication during untreated infection correlated positively with both proviral burden (**Figure 7B**) and diversity (**Figure 7C**) on ART. Though our diversity analyses focused on distinct proviruses, our recovery of identical sequences deserves brief mention, as these are consistent with clonal expansion of HIV-infected cells (27–29). As in non-controllers however (26, 58), the frequency of identical sequences varied markedly, consistent with differential clonal expansion across individuals. In participant 3 for example, half of sampled proviruses were identical to at least one other sequence, whereas 83% of participant 1's proviruses were distinct. Taken together with a recent study also reporting high frequencies of distinct proviruses in viremic controllers (52), our observations confirm that diverse proviral pools persist even in individuals who controlled viremia prior to therapy, underscoring the importance of early ART even in this group.

Our inference of pre-ART proviral half-lives also advances our understanding of proviral longevity during active viral replication. In particular, our estimate that pre-ART proviral half-lives were < 1 year in two participants (1 and 4) is consistent with the notion that, in many individuals, proviral turnover during untreated infection is much faster than during long-term suppressive ART (19, 21, 23). Specifically, while historic and recent half-life estimates for replication competent proviruses on ART are on the order of 4 years (25, 41) (with rates slowing further after 7 years on ART (25), while half-life estimates for the overall proviral pool are well over 10 years (25, 40), recent studies of non-controllers estimated pre-ART proviral half-lives of

~8 months (19) and ~25 months (23) using similar methods as the present study. The proviral half-lives calculated from participants 1 and 4 are consistent with proviral turnover during untreated infection being quite dynamic until ART initiation, after which proviral clearance substantially decreases, such that ART effectively "freezes" proviral composition in its immediate pretherapy state (21, 23). Participant 3' pre-ART proviral half-life however was estimated to be between 2.2 and 4.8 years, whereas participant 2's data was consistent with a zero-decay rate. While these participants' data are clearly consistent with ongoing deposition of proviral sequences into the reservoir, they suggest that subsequent clearance is not so rapid in all persons, allowing some individuals to harbor proviral pools that more fully recapitulate HIV's within-host evolutionary history. This phenomenon has previously been observed in non-controllers (20) and further emphasizes the unique nature of each individual's persisting HIV reservoir.

Our study had several limitations. Firstly, our cohort is small due to the challenges of recruiting viremic controllers with known infection dates for longitudinal studies. Secondly, since biological material was limited and viral loads were low in most samples, near-full-genome HIV amplification was not feasible. Thirdly, though *nef* is an appropriate region for within-host HIV evolutionary studies (20, 58, 59), sequencing of a subgenomic region does not allow us to classify recovered sequences as intact or defective, nor definitively classify sequences as clonally expanded. Our use of the IPDA to quantify intact (vs. defective) proviral burden mitigates this only partially. Furthermore, sample availability also prevented us from investigating possible mechanisms underlying participant 4's loss of viremia control. Finally, the peak viremia kinetics that we drew from the literature (12, 37, 50) to use in the mathematical model may not have reflected *in vivo* kinetics, though the high concordance between the results of the primary and sensitivity analyses suggest that the impact of this assumption is minor.

In conclusion, despite their natural ability to control HIV to low levels, viremic controllers nevertheless exhibited significant within-host HIV evolution pre-ART, that gave rise to persisting proviral pools whose size and diversity broadly reflected the extent of this viral replication and evolution. Though the proviral pools of two participants were skewed towards integration dates in the year or two prior to ART initiation, consistent with relatively rapid proviral turnover during untreated infection, those of two others featured sequences that nearly uniformly spanned their whole within-host HIV evolutionary histories. HIV remission and cure strategies will need to overcome within-host proviral diversity, even in individuals who controlled HIV replication prior to ART.

## Methods

### Study participants

Four viremic controllers who naturally maintained plasma viral loads (pVL) <2000 HIV RNA copies/ml for at least three years, and for whom an infection date could be reliably estimated, were studied. All participants were male with a median age of 55 years. Participants 1, 2 and 3 broadly maintained pVL < ~2000 pre-ART whereas participant 4 lost viremic control prior to ART. The estimated dates of infection (EDI) for participants 1, 2, and 4, were calculated as the midpoint between the last seronegative and first seropositive HIV test dates, while for participant 3 the EDI was determined as March 2008, 3 months prior to diagnosis, due to a documented exposure followed by seroconversion-like illness at this time. Participant nadir CD4+ T-cell counts ranged from 320 to 516 cells/mm$^3$, while CD4+ T cells counts at ART initiation ranged from 401-626 cells/mm$^3$. HLA class I types were: A*30:01-A*32:01/B*13:02-B*40:02/C*06:02-C*15:02 (participant 1); A*24:02-A*32:01/B*08:01-B*14:01/C*07:02-C*08:02 (participant 2); A*03:01-A*03:01/B*14:02-B*57:01/C*06:02-C*08:02 (participant 3); and A*03:01-A*32:01/B*14:02-B*39:01/C*08:02-C*12:03 (participant 4). All participants provided written informed consent. This study was approved by the Massachusetts General Hospital, Simon Fraser University and Providence Health Care/University of British Columbia Research Ethics Boards.

### Single genome HIV RNA and proviral amplification

Phylogenetic inference of within-host reservoir ages requires the longitudinal isolation of HIV RNA sequences pre-ART, alongside proviral DNA sequences sampled on ART, ideally using single-genome approaches. Because plasma viral loads were low and biological material was limited (*e.g.* only 10 million PBMCs per timepoint) full-length HIV amplification was not feasible, and HIV *nef* was selected for amplification for the reasons described in the results.

Total nucleic acids were extracted from plasma using the BioMerieux NucliSENS EasyMag system (BioMerieux, Marcy-l'Étoile, France), while genomic DNA was extracted from PBMCs using the Purelink® Genomic DNA kit (Invitrogen). For pre-ART plasma samples destined for HIV RNA *nef* amplification, 1 mL of sample was extracted, eluted in 60 µL and subjected to DNase I digestion (New England Biolabs Ltd, catalogue number: M0303S) to minimize the risk of amplifying HIV DNA in these low pVL samples. For on-ART PBMC samples destined for proviral *nef* amplification, cells were resuspended and split into two aliquots each of 5 million cells before extraction and elution into 60µl.

HIV *nef* was amplified using limiting-dilution nested RT-PCR (for HIV RNA) or nested PCR (for proviral DNA) using high fidelity enzymes and sequence-specific primers optimized for amplification of multiple HIV group M subtypes (58, 59). For HIV RNA extracts, cDNA was generated using NxtScript Reverse Transcriptase (Roche). Next, cDNA as well as genomic DNA extracts were endpoint diluted such that ~25-30% of the resulting nested PCR reactions, performed using the Expand™ High Fidelity PCR system (Roche), would yield an amplicon. Primers (5' -> 3') used for cDNA generation/1st round PCR were Nef8683F_pan (Forward; TAGCAGTAGCTGRGKGRACAGATAG) and Nef9536R_pan (Reverse; TACAGGCAAAAAGCAGCTGCTTATATGYAG). Primers used for 2nd round PCR were Nef8746F_pan (Forward; TCCACATACCTASAAGAATMAGACARG) and Nef9474R (Reverse; CAGGCCACRCCTCCCTGGAAASKCCC). Negative amplification controls were included in every run, and we also confirmed that plasma HIV RNA amplification did not occur in the absence of reverse transcription, indicating that DNAse treatment was effective. Amplicons were sequenced on an ABI 3730xl automated DNA analyzer. Chromatograms were basecalled using Sequencher v5.0 (Gene Codes) or the custom software RECall (60). *Nef* sequences that contained nucleotide mixtures, hypermutations (identified using Hypermut 2.0

(61)), suspected within-host recombinant sequences (identified using rdp4 Beta 95 (62)) or other defects were excluded from phylogenetic analysis, as were duplicate sequences (instead, the latter were added back at the data visualization stage). Within-host plasma and proviral *nef* sequences were codon-aligned using MAFFT v7 (63) implemented in HIV Align (https://www.hiv.lanl.gov/content/sequence/VIRALIGN/viralign.html) and manually edited in AliView v1.18 (64). Maximum likelihood phylogenies were inferred from aligned, gap-stripped within-host sequence datasets using W-IQ-TREE (65) following automated model selection with ModelFinder (66) using an Akaike Information Criterion (AIC) selection criterion (67). The models selected were GTR+F+I+G4 (general time reversible plus invariable sites and FreeRate heterogeneity model) for participant 1, K3Pu+F+I+G4 (Kimura 3 parameter model plus invariable sites and FreeRate heterogeneity) for participants 2 and 3 and TVM+ F+I+G4 (transversion model plus invariable sites and FreeRate heterogeneity) for participant 4. Patristic (tip-to-tip phylogenetic) distances were extracted from maximum likelihood phylogenies using the cophenetic.phylo function from the R package ape (v5.3) (68).

**HIV proviral full genome amplification and sequencing**

Single-template, near-full-length proviral amplification was performed on DNA extracted from CD4+ T cells by nested PCR using Platinum Taq DNA Polymerase High Fidelity (Invitrogen) such that ~25% of the resulting PCR reactions yielded an amplicon. First round primers were: Forward - AAATCTCTAGCAGTGGCGCCCGAACAG, Reverse - TGAGGGATCTCTAGTTACCAGAGTC. Second round primers were: Forward - GCGCCCGAACAGGGACYTGAAARCGAAAG, Reverse- GCACTCAAGGCAAGCTTTATTGAGGCTTA. Reactions were cycled as follows: 92°C for 2 minutes; 10 cycles of (92°C for 10 seconds, 60°C for 30 seconds and 68°C for 10 minutes); 20 cycles of (92°C for 10 seconds, 55°C for 30 seconds and 68°C for 10 minutes); 68°C for 10

minutes (69, 70). Amplicons were sequenced using Illumina MiSeq technology and *de novo* assembled using the custom software MiCall (https://github.com/cfe-lab/MiCall) which features an in-house modification of the Iterative Virus Assembler (IVA) (71).

**Within-host phylogenetic inference and proviral age reconstruction**

We used a published within-host phylogenetic approach to infer proviral sequence ages (20). Briefly, for each participant we inferred a maximum likelihood phylogeny relating within-host longitudinal pre-ART plasma HIV RNA and proviral sequences sampled on ART. We then exhaustively re-root each tree to identify the root location that maximizes the correlation between the root-to-tip distances and collection dates of the pre-ART plasma HIV RNA sequences. Here, the root represents the inferred most recent common ancestor (MRCA) of the within-host dataset (*i.e.* the inferred transmitted founder virus). We then fit a linear model relating the collection dates of the plasma HIV RNA sequences to their divergence from the root, where the slope represents the average host-specific rate of HIV evolution prior to ART and the x-intercept represents the phylogenetically-estimated MRCA date. For participant 4, who lost viremic control, two regression lines were fit - one each for the "viremic control" and "post-control" eras. The integration dates of the proviral sequences sampled during ART, along with their 95% confidence intervals (CI), are then estimated from their divergence from the root using the linear regression. We assessed molecular clock and model fit using a delta (Δ) Akaike Information Criterion (ΔAIC) (67), computed as the difference between the AIC of the null model (no evidence of within-host evolution, *i.e.* a zero slope) and the AIC of the linear regression. A within-host phylogeny was deemed to have a sufficient molecular clock signal if the linear regression produced a $\Delta AIC \geq 10$, (which corresponds to a p-value of 0.00053 when using a log-likelihood ratio test), and the 95% confidence interval of the estimated root date contained or preceded the first sampling point. Since the original description of our proviral

dating framework (20), we have also implemented it as a web-service, available at https://bblab-hivresearchtools.ca/django/tools/phylodating, with open source code available at https://www.github.com/cfe-lab/phylodating.

**Intact Proviral DNA Assay (IPDA)**

Each participant's first PBMC sample on ART was used to estimate total and intact proviral DNA burdens using the Intact Proviral DNA Assay (IPDA) (24). As this method reports proviral burdens in terms of HIV copies per million CD4+ T cells, we first isolated CD4+ T cells from 10M PBMC by negative selection using the EasySep Human CD4+ T cell Enrichment Kit (STEMCELL Technologies, Cat #: 19052). This yielded a mean of 1.6 (range 1.3-2.0) million CD4+ T cells per participant, which were then extracted using the QIAamp DNA Mini Kit (Qiagen) with precautions to minimize shearing. In the IPDA, HIV and human DNA quantification reactions (where the latter target the human RPP30 gene) are conducted in parallel, where copies are normalized to the quantity of input DNA and subsequently corrected for DNA shearing. In each ddPCR reaction, 7ng (RPP30) or a median 588ng (IQR 488-697ng) (HIV) of genomic DNA was combined with ddPCR Supermix for Probes (no dUTPs, BioRad), primers (final concentration 900nM, Integrated DNA Technologies), probe(s) (final concentration 250nM, ThermoFisher Scientific), XhoI restriction enzyme (New England Biolabs) and nuclease free water. Human RPP30 primer and probe sequences (5’ -> 3’) are: RPP30 Forward Primer- GATTTGGACCTGCGAGCG, RPP30 Probe- VIC-CTGACCTGAAGGCTCT- MGBNFQ, RPP30 Reverse Primer- GCGGCTGTCTCCACAAGT; RPP30 Shear Forward Primer- CCAATTTGCTGCTCCTTGGG, RPP30 Shear Probe- FAM-AAGGAGCAAGGTTCTATTGTAG- MGBNFQ, RPP30 Shear Reverse Primer-CATGCAAAGGAGGAAGCCG. The default (24) HIV primers and probes were: Ψ Forward Primer- CAGGACTCGGCTTGCTGAAG, Ψ Probe- FAM- TTTTGGCGTACTCACCAGT-

MGBNFQ, Ψ Reverse Primer- GCACCCATCTCTCTCCTTCTAGC; *env* Forward Primer- AGTGGTGCAGAGAGAAAAAAGAGC, *env* Probe- VIC-CCTTGGGTTCTTGGGA-MGBNFQ, anti-Hypermutant *env* Probe- CCTTAGGTTCTTAGGAGC- MGBNFQ, *env* Reverse Primer- GTCTGGCCTGTACCGTCAGC. The published HIV primers/probes however can sometimes fail to detect autologous HIV sequences due to naturally-occurring polymorphism (72, 73), thus requiring autologous primer/probes. For this reason, targeted HIV sequencing of the IPDA *Ψ* and *env* regions was performed for all participants prior to IPDA measurement, and autologous primers/probes substituted where necessary. The substituted primer and probe sequences (5' -> 3') were: Participant 2 Ψ Probe- FAM- TTTCAGCGTACTCACCAGT, Participant 3 Ψ Probe- FAM- ATATGGCGTACTCACCGGT, Participant 1 Ψ Probe- FAM-AATTGGCGTACTCACCAGT, Participant 1 Ψ Probe- FAM- AATTGGCGTACTCACCAGC, Participant 1 *env* Forward Primer- GGTGGTGCAGAGAGAAAAAAGAGC, Participant 1 *env* Reverse Primer- GCTGACGGCACAGGCCAGGC, Participant 3 *env* Reverse Primer-GCTGACGGTACAGGCCAGAT. Droplets were prepared using the Automated Droplet Generator and cycled as previously described (24). Droplets were analyzed on a QX200 Droplet Reader (BioRad) using QuantaSoft software (BioRad, version 1.7.4), where the results of a minimum of 8 technical replicates, comprising a median of 589,820 (IQR 501,539- 721,871) cells assayed in total, were merged prior to analysis. Intact HIV-1 copies (Ψ and *env* double-positive droplets) were corrected for DNA shearing based on the frequency of RPP30 and RPP30-Shear double positive droplets. The median DNA shearing index, which measures the proportion of sheared DNA in a sample, was 0.40 (IQR 0.39- 0.43), which is in line with acceptable levels in this assay (24, 72).

**Inference of proviral half-lives prior to ART**

We estimated host-specific pre-ART proviral half-lives from each participant's phylogenetically-inferred proviral age distributions two ways: using a mathematical modeling approach (23), and directly from the data. To do the former, we implemented a published "cumulative" mathematical model that describes each individual's untreated HIV infection via a set of ordinary differential equations that model the dynamics of the concentrations of cells and virus in each individual over time, as described in (23). Briefly, the model includes susceptible target cells, actively and latently infected cells that can produce viable virus, actively and latently infected cells that *cannot* produce viable virus, the virus itself, and an immune response. The model assumes that HIV sequences enter the latent proviral pool at a rate proportional to their abundance in plasma at the time, and a subsequent independent step assumes that these proviruses then "decay" out of this pool at a constant, exponential rate. To recreate each participant's *in vivo* plasma viral dynamics as realistically as possible, we merged each participant’s available pVL data to acute phase dynamics observed in viremic controllers (37) (because peak viremia was not captured in our participants' measured viral load data). We also performed sensitivity analyses using peak viremia kinetics observed in elite controllers (38, 39). The acute-phase dynamics we used are shown in **Table S1**. For each participant we then modeled proviral deposition into their latent pool according to their reconstructed viral load histories (shown in **Figures 8A-D**) and grouped latent proviruses by their year of creation. We then allowed each group of latent proviruses to decay exponentially, up until each participant's actual proviral sampling date, under half-lives ranging from 30 to 6000 days, in increments of 30 days to create predicted proviral distributions under these different decay rates. The model was also run under conditions of no decay. Note that this model assumed that, after ART, the overall size of the proviral pool decreased but the relative proportions of sequences in each group remained the same. Finally, we identified the best-fitting proviral decay rate to each participant's

*observed* proviral age distribution by maximum likelihood and used standard theory to identify 95% confidence bounds.

To estimate the reservoir half-life directly from the data, we grouped each participant's observed proviral sequences by their age, in years, relative to ART initiation (or in the case of participant 3, the date of their viremic episode after ART). We will refer to the bin containing proviruses from $t-1$ to $t$ years old as "bin $t$", where we make the simplifying assumptions that all proviruses in bin $t$ are exactly $t$ years old. We then applied a Poisson generalized linear model with the canonical natural logarithm link function to the binned counts, using the age of the bin $t$ as the predictor. This choice can be justified as follows: Let $t_{1/2}$ be the proviral half-life. Assuming the participant's proviral reservoir decays at an exponential rate, we would expect that the size of bin $t$ would be approximately

$$\lfloor C \exp(-\theta t) \rfloor$$

where $C$ represents the initial size of the age bin as if there was no decay, and $\theta = \ln(2)/t_{1/2}$.

Now, we assume that every provirus has the same small independent probability $p$ of being sampled. If so, we would then expect the number of observed proviruses from bin $t$ to be binomially distributed with $\lfloor C \exp(-\theta t) \rfloor$ trials and success probability $p$. Assuming that $p$ is small, we can approximate the distribution with a Poisson distribution with parameter

$$\lambda = Cp \exp(-\theta t) = \exp(D - \theta t)$$

where $D = \ln(Cp)$, or in other words,

$$\ln \lambda = D - \theta t.$$

which is precisely the setup required for a Poisson generalized linear model with natural logarithm link function.

**Statistical Analyses**

Spearman's correlation was used to assess the relationship between continuous parameters. Statistical analysis was performed in Prism (version 8) software.

**Data Availability**

GenBank accession numbers for sequences reported in this study are MW781931 to MW782197 (proviral DNA) and MW782198 to MW782743 (plasma HIV RNA).

**Acknowledgements**

The authors gratefully thank the study participants without whom this research would not be possible.

**Funding Statement**

This work was supported in part by the Canadian Institutes of Health Research (CIHR) through a project grant (PJT-159625 to ZLB and JBJ) and a focused team grant (HB1-164063 to ZLB and MAB). This work was supported in part by the National Institutes of Health (NIH) under award number NIHA127029 (to ZLB) and the Martin Delaney 'BELIEVE' Collaboratory (NIH grant 1UM1AI26617 to ZLB and MAB), which is supported by the following NIH Co-Funding and Participating Institutes and Centers: NIAID, NCI, NICHD, NHLBI, NIDA, NIMH, NIA, FIC, and OAR. FHO is supported by a Ph.D. fellowship from the Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE), a DELTAS Africa Initiative [grant # DEL-15-006]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [grant # 107752/Z/15/Z] and the UK government. The views expressed in this publication are those of the authors and not necessarily those of AAS, NEPAD Agency, Wellcome Trust or the UK government. AS and BRJ are supported by CIHR Doctoral Research Awards. NNK is supported

by a CIHR Vanier Doctoral Award. RLM is supported by a CIHR MSc award. MAB holds a Canada Research Chair, Tier 2 in Viral Pathogenesis and Immunity. ZLB is supported by a Scholar Award from the Michael Smith Foundation for Health Research.

## References


1. Finzi D, Hermankova M, Pierson T, Carruth LM, Buck C, Chaisson RE, Quinn TC, Chadwick K, Margolick J, Brookmeyer R, Gallant J, Markowitz M, Ho DD, Richman DD, Siliciano RF. 1997. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. Science (80- ) 278:1295–1300.
2. Wong JK, Hezareh M, Günthard HF, Havlir D V., Ignacio CC, Spina CA, Richman DD. 1997. Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. Science (80- ) 278:1291–1295.
3. Chun TW, Stuyver L, Mizell SB, Ehler LA, Mican JAM, Baseler M, Lloyd AL, Nowak MA, Fauci AS. 1997. Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. Proc Natl Acad Sci U S A 94:13193–13197.
4. Sanchez G, Xu X, Chermann JC, Hirsch I. 1997. Accumulation of defective viral genomes in peripheral blood mononuclear cells of human immunodeficiency virus type 1-infected individuals. J Virol 71:2233–2240.
5. Bruner KM, Murray AJ, Pollack RA, Soliman MG, Laskey SB, Capoferri AA, Lai J, Strain MC, Lada SM, Hoh R, Ho YC, Richman DD, Deeks SG, Siliciano JD, Siliciano RF. 2016. Defective proviruses rapidly accumulate during acute HIV-1 infection. Nat Med 22:1043–1049.
6. Imamichi H, Dewar RL, Adelsberger JW, Rehm CA, O'Doherty U, Paxinos EE, Fauci AS, Lane HC. 2016. Defective HIV-1 proviruses produce novel protein-coding RNA species in HIV-infected patients on combination antiretroviral therapy. Proc Natl Acad Sci U S A 113:8783–8.
7. Deeks SG, Lewin SR, Ross AL, Ananworanich J, Benkirane M, Cannon P, Chomont N, Douek D, Lifson JD, Lo YR, Kuritzkes D, Margolis D, Mellors J, Persaud D, Tucker JD,

Barre-Sinoussi F, Alter G, Auerbach J, Autran B, Barouch DH, Behrens G, Cavazzana M, Chen Z, Cohen ÉA, Corbelli GM, Eholié S, Eyal N, Fidler S, Garcia L, Grossman C, Henderson G, Henrich TJ, Jefferys R, Kiem HP, McCune J, Moodley K, Newman PA, Nijhuis M, Nsubuga MS, Ott M, Palmer S, Richman D, Saez-Cirion A, Sharp M, Siliciano J, Silvestri G, Singh J, Spire B, Taylor J, Tolstrup M, Valente S, Van Lunzen J, Walensky R, Wilson I, Zack J. 2016. International AIDS Society global scientific strategy: Towards an HIV cure 2016. Nat Med. Nature Publishing Group.

8. Jiang C, Lian X, Gao C, Sun X, Einkauf KB, Chevalier JM, Chen SMY, Hua S, Rhee B, Chang K, Blackmer JE, Osborn M, Peluso MJ, Hoh R, Somsouk M, Milush J, Bertagnolli LN, Sweet SE, Varriale JA, Burbelo PD, Chun TW, Laird GM, Serrao E, Engelman AN, Carrington M, Siliciano RF, Siliciano JM, Deeks SG, Walker BD, Lichterfeld M, Yu XG. 2020. Distinct viral reservoirs in individuals with spontaneous control of HIV-1. Nature 585:261–267.

9. Li C, Mori L, Valente ST. 2021. The Block-and-Lock Strategy for Human Immunodeficiency Virus Cure: Lessons Learned from Didehydro–Cortistatin A. J Infect Dis 223:S46–S53.

10. Allers K, Hütter G, Hofmann J, Loddenkemper C, Rieger K, Thiel E, Schneider T. 2011. Evidence for the cure of HIV infection by CCR5Δ32/Δ32 stem cell transplantation. Blood 117:2791–2799.

11. Gupta RK, Peppa D, Hill AL, Gálvez C, Salgado M, Pace M, McCoy LE, Griffith SA, Thornhill J, Alrubayyi A, Huyveneers LEP, Nastouli E, Grant P, Edwards SG, Innes AJ, Frater J, Nijhuis M, Wensing AMJ, Martinez-Picado J, Olavarria E. 2020. Evidence for HIV-1 cure after CCR5Δ32/Δ32 allogeneic haemopoietic stem-cell transplantation 30 months post analytical treatment interruption: a case report. Lancet HIV 7:e340–e347.

12. Casado C, Galvez C, Pernas M, Tarancon-Diez L, Rodriguez C, Sanchez-Merino V, Vera M, Olivares I, De Pablo-Bernal R, Merino-Mansilla A, Del Romero J, Lorenzo-Redondo R, Ruiz-Mateos E, Salgado M, Martinez-Picado J, Lopez-Galindez C. 2020. Permanent control of HIV-1 pathogenesis in exceptional elite controllers: a model of spontaneous cure. Sci Rep 10.

13. Woldemeskel BA, Kwaa AK, Blankson JN. 2020. Viral reservoirs in elite controllers of HIV-1 infection: Implications for HIV cure strategies. EBioMedicine 62:103118.

14. Promer K, Karris MY. 2018. Current Treatment Options for HIV Elite Controllers: a Review. Curr Treat Options Infect Dis 10:302–309.

15. Bello G, Casado C, García S, Rodríguez C, del Romero J, Carvajal-Rodriguez A, Posada D, López-Galíndez C. 2007. Lack of temporal structure in the short term HIV-1 evolution within asymptomatic naïve patients. Virology 362:294–303.

16. Okulicz JF, Marconi VC, Landrum ML, Wegner S, Weintrob A, Ganesan A, Hale B, Crum-Cianflone N, Delmar J, Barthel V, Quinnan G, Agan BK, Dolan MJ. 2009. Clinical outcomes of elite controllers, viremic controllers, and long-term nonprogressors in the US department of defense HIV natural history study. J Infect Dis 200:1714–1723.

17. Grabar S, Selinger-Leneman H, Abgrall S, Pialoux G, Weiss L, Costagliola D. 2009. Prevalence and comparative characteristics of long-term nonprogressors and HIV controller patients in the French Hospital Database on HIV. AIDS 23:1163–1169.

18. Initiation of Antiretroviral Therapy | Adult and Adolescent ARV | ClinicalInfo.

19. Brodin J, Zanini F, Thebo L, Lanz C, Bratt G, Neher RA, Albert J. 2016. Establishment and stability of the latent HIV-1 DNA reservoir. Elife 5:e18889.

20. Jones BR, Kinloch NN, Horacsek J, Ganase B, Harris M, Harrigan PR, Jones RB, Brockman MA, Joy JB, Poon AFY, Brumme ZL. 2018. Phylogenetic approach to recover

integration dates of latent HIV sequences within-host. Proc Natl Acad Sci U S A 115:E8958–E8967.

21. Abrahams MR, Joseph SB, Garrett N, Tyers L, Moeser M, Archin N, Council OD, Matten D, Zhou S, Doolabh D, Anthony C, Goonetilleke N, Karim SA, Margolis DM, Pond SK, Williamson C, Swanstrom R. 2019. The replication-competent HIV-1 latent reservoir is primarily established near the time of therapy initiation. Sci Transl Med 11:5589.

22. Brooks K, Jones BR, Dilernia DA, Wilkins DJ, Claiborne DT, McInally S, Gilmour J, Kilembe W, Joy JB, Allen SA, Brumme ZL, Hunter E. 2020. HIV-1 variants are archived throughout infection and persist in the reservoir. PLOS Pathog 16:e1008378.

23. Pankau MD, Reeves DB, Harkins E, Ronen K, Jaoko W, Mandaliya K, Graham SM, McClelland RS, Matsen IV FA, Schiffer JT, Overbaugh J, Lehman DA. 2020. Dynamics of HIV DNA reservoir seeding in a cohort of superinfected Kenyan women. PLOS Pathog 16:e1008286.

24. Bruner KM, Wang Z, Simonetti FR, Bender AM, Kwon KJ, Sengupta S, Fray EJ, Beg SA, Antar AAR, Jenike KM, Bertagnolli LN, Capoferri AA, Kufera JT, Timmons A, Nobles C, Gregg J, Wada N, Ho YC, Zhang H, Margolick JB, Blankson JN, Deeks SG, Bushman FD, Siliciano JD, Laird GM, Siliciano RF. 2019. A quantitative approach for measuring the reservoir of latent HIV-1 provirusesNature. Nature Publishing Group.

25. Peluso MJ, Bacchetti P, Ritter KD, Beg S, Lai J, Martin JN, Hunt PW, Henrich TJ, Siliciano JD, Siliciano RF, Laird GM, Deeks SG. 2020. Differential decay of intact and defective proviral DNA in HIV-1–infected individuals on suppressive antiretroviral therapy. JCI Insight 5.

26. Pinzone MR, VanBelzen DJ, Weissman S, Bertuccio MP, Cannon LM, Venanzi-Rullo E, Migueles S, Jones RB, Mota T, Joseph SB, Groen K, Pasternak AO, Hwang WT, Sherman

B, Vourekas A, Nunnari G, O'Doherty U. 2019. Longitudinal HIV sequencing reveals reservoir expression leading to decay which is obscured by clonal expansion. Nat Commun 10:1–12.

27. Lee GQ, Orlova-Fink N, Einkauf K, Chowdhury FZ, Sun X, Harrington S, Kuo HH, Hua S, Chen HR, Ouyang Z, Reddy K, Dong K, Ndung'U T, Walker BD, Rosenberg ES, Yu XG, Lichterfeld M. 2017. Clonal expansion of genome-intact HIV-1 in functionally polarized Th1 CD4+ T cells. J Clin Invest 127:2689–2696.

28. Maldarelli F, Wu X, Su L, Simonetti FR, Shao W, Hill S, Spindler J, Ferris AL, Mellors JW, Kearney MF, Coffin JM, Hughes SH. 2014. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. Science (80- ) 345:179–183.

29. Wagner TA, McLaughlin S, Garg K, Cheung CYK, Larsen BB, Styrchak S, Huang HC, Edlefsen PT, Mullins JI, Frenkel LM. 2014. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. Science (80- ) 345:570–573.

30. Buzon MJ, Martin-Gayo E, Pereyra F, Ouyang Z, Sun H, Li JZ, Piovoso M, Shaw A, Dalmau J, Zangger N, Martinez-Picado J, Zurakowski R, Yu XG, Telenti A, Walker BD, Rosenberg ES, Lichterfeld M. 2014. Long-term antiretroviral treatment initiated at primary HIV-1 infection affects the size, composition, and decay kinetics of the reservoir of HIV-1-infected CD4 T cells. J Virol 88:10056–65.

31. Jain V, Hartogensis W, Bacchetti P, Hunt PW, Hatano H, Sinclair E, Epling L, Lee TH, Busch MP, McCune JM, Pilcher CD, Hecht FM, Deeks SG. 2013. Antiretroviral therapy initiated within 6 months of HIV infection is associated with lower T-cell activation and smaller HIV reservoir size. J Infect Dis 208:1202–1211.

32. Novelli S, Lécuroux C, Avettand-Fenoel V, Seng R, Essat A, Morlat P, Viard J-P, Rouzioux C, Meyer L, Goujard C. 2018. Long-term Therapeutic Impact of the Timing of

Antiretroviral Therapy in Patients Diagnosed With Primary Human Immunodeficiency Virus Type 1 Infection. Clin Infect Dis 66:1519–1527.

33. Brumme ZL, Sudderuddin H, Ziemniak C, Luzuriaga K, Jones BR, Joy JB, Cunningham CK, Greenough T, Persaud D. 2019. Genetic complexity in the replication-competent latent HIV reservoir increases with untreated infection duration in infected youth. AIDS 33:211–218.

34. Kearney M, Maldarelli F, Shao W, Margolick JB, Daar ES, Mellors JW, Rao V, Coffin JM, Palmer S. 2009. Human Immunodeficiency Virus Type 1 Population Genetics and Adaptation in Newly Infected Individuals. J Virol 83:2715–2727.

35. Lukashov V V, Kuiken CL, Goudsmit J. 1995. Intrahost human immunodeficiency virus type 1 evolution is related to length of the immunocompetent period. J Virol 69:6911–6916.

36. Buzon MJ, Martin-Gayo E, Pereyra F, Ouyang Z, Sun H, Li JZ, Piovoso M, Shaw A, Dalmau J, Zangger N, Martinez-Picado J, Zurakowski R, Yu XG, Telenti A, Walker BD, Rosenberg ES, Lichterfeld M. 2014. Long-Term Antiretroviral Treatment Initiated at Primary HIV-1 Infection Affects the Size, Composition, and Decay Kinetics of the Reservoir of HIV-1-Infected CD4 T Cells. J Virol 88:10056–10065.

37. Miura T, Brumme ZL, Brockman MA, Rosato P, Sela J, Brumme CJ, Pereyra F, Kaufmann DE, Trocha A, Block BL, Daar ES, Connick E, Jessen H, Kelleher AD, Rosenberg E, Markowitz M, Schafer K, Vaida F, Iwamoto A, Little S, Walker BD. 2010. Impaired Replication Capacity of Acute/Early Viruses in Persons Who Become HIV Controllers. J Virol 84:7581–7591.

38. Moosa Y, Tanko RF, Ramsuran V, Singh R, Madzivhandila M, Yende-Zuma N, Abrahams MR, Selhorst P, Gounder K, Moore PL, Williamson C, Abdool Karim SS,

Garrett NJ, Burgers WA. 2018. Case report: Mechanisms of HIV elite control in two African women. BMC Infect Dis 18:54.

39. Morley D, Lambert JS, Hogan LE, De Gascun C, Redmond N, Rutishauser RL, Thanh C, Gibson EA, Hobbs K, Bakkour S, Busch MP, Farrell J, McGetrick P, Henrich TJ. 2019. Rapid development of HIV elite control in a patient with acute infection. BMC Infect Dis 19:1–6.

40. Golob JL, Stern J, Holte S, Kitahata MM, Crane HM, Coombs RW, Goecker E, Woolfrey AE, Harrington RD. 2018. HIV DNA levels and decay in a cohort of 111 long-term virally suppressed patients. AIDS 32:2113–2118.

41. Siliciano JD, Kajdas J, Finzi D, Quinn TC, Chadwick K, Margolick JB, Kovacs C, Gange SJ, Siliciano RF. 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.

42. Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, Berlin AM, Malboeuf CM, Ryan EM, Gnerre S, Zody MC, Erlich RL, Green LM, Berical A, Wang Y, Casali M, Streeck H, Bloom AK, Dudek T, Tully D, Newman R, Axten KL, Gladden AD, Battis L, Kemper M, Zeng Q, Shea TP, Gujja S, Zedlack C, Gasser O, Brander C, Hess C, Günthard HF, Brumme ZL, Brumme CJ, Bazner S, Rychert J, Tinsley JP, Mayer KH, Rosenberg E, Pereyra F, Levin JZ, Young SK, Jessen H, Altfeld M, Birren BW, Walker BD, Allen TM. 2012. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog 8:1002529.

43. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, Sun C, Grayson T, Wang S, Li H, Wei X, Jiang C, Kirchherr JL, Gao F, Anderson JA, Ping LH, Swanstrom R, Tomaras GD, Blattner WA, Goepfert PA, Kilby JM, Saag MS, Delwart EL,

Busch MP, Cohen MS, Montefiori DC, Haynes BF, Gaschen B, Athreya GS, Lee HY, Wood N, Seoighe C, Perelson AS, Bhattacharya T, Korber BT, Hahn BH, Shaw GM. 2008. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci U S A 105:7552–7557.

44. Fischer W, Ganusov V V., Giorgi EE, Hraber PT, Keele BF, Leitner T, Han CS, Gleasner CD, Green L, Lo CC, Nag A, Wallstrom TC, Wang S, McMichael AJ, Haynes BF, Hahn BH, Perelson AS, Borrow P, Shaw GM, Bhattacharya T, Korber BT. 2010. Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. PLoS One 5.

45. Salazar-Gonzalez JF, Salazar MG, Keele BF, Learn GH, Giorgi EE, Li H, Decker JM, Wang S, Baalwa J, Kraus MH, Parrish NF, Shaw KS, Guffey MB, Bar KJ, Davis KL, Ochsenbauer-Jambor C, Kappes JC, Saag MS, Cohen MS, Mulenga J, Derdeyn CA, Allen S, Hunter E, Markowitz M, Hraber P, Perelson AS, Bhattacharya T, Haynes BF, Korber BT, Hahn BH, Shaw GM. 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.

46. Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, Gupta P, Rinaldo CR, Learn GH, He X, Huang X-L, Mullins JI. 1999. Consistent Viral Evolutionary Changes Associated with the Progression of Human Immunodeficiency Virus Type 1 Infection. J Virol 73:10489–10502.

47. Rambaut A, Posada D, Crandall KA, Holmes EC. 2004. The causes and consequences of HIV evolution. Nat Rev Genet. Nature Publishing Group.

48. Alizon S, Fraser C. 2013. Within-host and between-host evolutionary rates across the HIV-1 genome. Retrovirology 10:49.

49. O'Connell KA, Brennan TP, Bailey JR, Ray SC, Siliciano RF, Blankson JN. 2010. Control of HIV-1 in Elite Suppressors despite Ongoing Replication and Evolution in Plasma Virus. J Virol 84:7018–7028.

50. Mens H, Kearney M, Wiegand A, Shao W, Schønning K, Gerstoft J, Obel N, Maldarelli F, Mellors JW, Benfield T, Coffin JM. 2010. HIV-1 Continues To Replicate and Evolve in Patients with Natural Control of HIV Infection. J Virol 84:12971–12981.

51. Miura T, Brockman MA, Schneidewind A, Lobritz M, Pereyra F, Rathod A, Block BL, Brumme ZL, Brumme CJ, Baker B, Rothchild AC, Li B, Trocha A, Cutrell E, Frahm N, Brander C, Toth I, Arts EJ, Allen TM, Walker BD. 2009. HLA-B57/B*5801 Human Immunodeficiency Virus Type 1 Elite Controllers Select for Rare Gag Variants Associated with Reduced Viral Replication Capacity and Strong Cytotoxic T-Lymphotye Recognition. J Virol 83:2743–2755.

52. de Azevedo SSD, Caetano DG, Côrtes FH, Teixeira SLM, dos Santos Silva K, Hoagland B, Grinsztejn B, Veloso VG, Morgado MG, Bello G. 2017. Highly divergent patterns of genetic diversity and evolution in proviral quasispecies from HIV controllers. Retrovirology 14:29.

53. Wang B, Mikhail M, Dyer WB, Zaunders JJ, Kelleher AD, Saksena NK. 2003. First demonstration of a lack of viral sequence evolution in a nonprogressor, defining replication-incompetent HIV-1 infection. Virology 312:135–150.

54. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. 2020. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res 48:W449–W454.

55. Kwaa AK, Garliss CC, Ritter KD, Laird GM, Blankson JN. 2020. Elite suppressors have

low frequencies of intact HIV-1 proviral DNA. AIDS 34:641–643.

56. Falcinelli SD, Kilpatrick KW, Read J, Murtagh R, Allard B, Ghofrani S, Kirchherr J, James KS, Stuelke E, Baker C, Kuruc JD, Eron JJ, Hudgens MG, Gay CL, Margolis DM, Archin NM. 2020. Longitudinal Dynamics of Intact HIV Proviral DNA and Outgrowth Virus Frequencies in a Cohort of Individuals Receiving Antiretroviral Therapy. J Infect Dis.

57. Gandhi RT, Cyktor JC, Bosch RJ, Mar H, Laird GM, Martin A, Collier AC, Riddler SA, Macatangay BJ, Rinaldo CR, Eron JJ, Siliciano JD, McMahon DK, Mellors JW, Hogg E, LeBlanc R, Scello C, Palm D, Gandhi M, Fletcher C, Podany A, Aweeka F, Halvas L, Dragavon J, Joseph J, Lagattuta R, Lin L, Pederson S, Robertson K, Rubin L, Smith D, Spudich S, Tsibris A, Hogg E, LeBlanc R, Scello C, Palm D, Gandhi M, Fletcher C, Podany A, Aweeka F, Halvas L, Dragavon J, Joseph J, Lagattuta R, Lin L, Pederson S, Robertson K, Rubin L, Smith D, Spudich S, Tsibris A. 2020. Selective Decay of Intact HIV-1 Proviral DNA on Antiretroviral Therapy. J Infect Dis.

58. Miller RL, Ponte R, Jones BR, Kinloch NN, Omondi FH, Jenabian M-A, Dupuy FP, Fromentin R, Brassard P, Mehraj V, Chomont N, Poon AFY, Joy JB, Brumme ZL, Routy J-P. 2019. HIV Diversity and Genetic Compartmentalization in Blood and Testes during Suppressive Antiretroviral Therapy. J Virol 93:755–774.

59. Jones BR, Miller RL, Kinloch NN, Tsai O, Rigsby H, Sudderuddin H, Shahid A, Ganase B, Brumme CJ, Harris M, Poon AFY, Brockman MA, Fromentin R, Chomont N, Joy JB, Brumme ZL. 2019. Genetic Diversity, Compartmentalization, and Age of HIV Proviruses Persisting in CD4 + T Cell Subsets during Long-Term Combination Antiretroviral Therapy . J Virol 94:1786–1805.

60. Woods CK, Brumme CJ, Liu TF, Chui CKS, Chu AL, Wynhoven B, Hall TA, Trevino C,

Shafer RW, Harrigan PR. 2012. Automating HIV drug resistance genotyping with RECall, a freely accessible sequence analysis tool. J Clin Microbiol 50:1936–1942.

61. Rose PP, Korber BT. 2000. Detecting hypermutations in viral sequences with an emphasis on G→A hypermutation. Bioinformatics 16:400–401.

62. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol 1.

63. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol 30:772–780.

64. Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30:3276–3278.

65. Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. 2016. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res 44:W232–W235.

66. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. 2017. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat Methods 14:587–589.

67. Akaike H. 1974. A New Look at the Statistical Model Identification. IEEE Trans Automat Contr 19:716–723.

68. Paradis E, Schliep K. 2019. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35:526–528.

69. Lee GQ, Orlova-Fink N, Einkauf K, Chowdhury FZ, Sun X, Harrington S, Kuo HH, Hua S, Chen HR, Ouyang Z, Reddy K, Dong K, Ndung'U T, Walker BD, Rosenberg ES, Yu XG, Lichterfeld M. 2017. Clonal expansion of genome-intact HIV-1 in functionally polarized Th1 CD4+ T cells. J Clin Invest 127:2689–2696.

70. Li B, Gladden AD, Altfeld M, Kaldor JM, Cooper DA, Kelleher AD, Allen TM. 2007.

Rapid Reversion of Sequence Polymorphisms Dominates Early Human Immunodeficiency Virus Type 1 Evolution. J Virol 81:193–201.

71. Hunt M, Gall A, Ong SH, Brener J, Ferns B, Goulder P, Nastouli E, Keane JA, Kellam P, Otto TD. 2015. IVA: Accurate de novo assembly of RNA virus genomes. Bioinformatics 31:2374–2376.

72. Kinloch NN, Ren Y, Conce Alberto WD, Dong W, Khadka P, Huang SH, Mota TM, Wilson A, Shahid A, Kirkby D, Harris M, Kovacs C, Benko E, Ostrowski MA, Del Rio Estrada PM, Wimpelberg A, Cannon C, Hardy WD, MacLaren L, Goldstein H, Brumme CJ, Lee GQ, Lynch RM, Brumme ZL, Jones RB. 2021. HIV-1 diversity considerations in the application of the Intact Proviral DNA Assay (IPDA). Nat Commun 12:1–10.

73. Simonetti FR, White JA, Tumiotto C, Ritter KD, Cai M, Gandhi RT, Deeks SG, Howell BJ, Montaner LJ, Blankson JN, Martin A, Laird GM, Siliciano RF, Mellors JW, Siliciano JD. 2020. Intact proviral DNA assay analysis of large cohorts of people with HIV provides a benchmark for the frequency and composition of persistent proviral DNA. Proc Natl Acad Sci 202006816.

## Figure legends

**Fig. 1: Participant sampling timeline and reservoir quantification.** Timeline is depicted as years since ART initiation. Shading represents ART. Inverted black triangles denote the clinically-estimated date of infection. Coloured circles denote pre-ART plasma samples from which HIV RNA sequences were isolated. Red diamonds denote cell samples on ART from which proviral sequences were isolated. Black diamonds denote proviral quantification dates using the Intact Proviral DNA Assay (IPDA). IPDA results are shown as pie charts, where the pie size denotes the total proviral burden and the coloured slices denote intact and defective HIV genome proportions (actual values shown in **Table 1**).

**Fig. 2: Between-host HIV phylogeny.** Maximum likelihood (ML) phylogeny inferred from 546 plasma HIV RNA sequences (circles) and 267 proviral DNA sequences (open diamonds) isolated from study participants. Numbers on internal branches indicate ML bootstrap values supporting within-host monophyletic clades. Scale in estimated substitutions per nucleotide site. The phylogeny is midpoint-rooted. The black dot represents the HIV-1 subtype B reference strain HXB2.

**Fig. 3: Participant 1.** (A) Clinical history and sampling timeline. Throughout all figures, circles denote plasma HIV RNA sampling and diamonds denote HIV DNA sampling. Shading represents ART. (B) Maximum-likelihood within-host phylogeny and corresponding amino acid highlighter plot. In the phylogeny, colored circles denote distinct pre-ART plasma HIV RNA sequences and red diamonds indicate distinct proviral sequences sampled during suppressive ART. Sequences that were recovered repeatedly are shown as open grey circles (for plasma HIV RNA) and diamonds (for proviruses) adjacent to the relevant tip. The root represents the inferred most recent common ancestor of the dataset, representing the phylogenetically-inferred transmitted founder virus event. The highlighter plot is ordered according to the phylogeny and

depicts *amino acid* sequences. The top sequence serves as the reference, where colored ticks in sequences beneath it denote non-synonymous substitutions with respect to the reference. (C) HIV sequence divergence-versus-time plot. The blue dashed line represents the linear model relating the root-to-tip distances of distinct pre-ART plasma HIV RNA sequences (colored circles) to their sampling times. This model is then used to convert the root-to-tip distances of distinct proviral sequences sampled during ART (red diamonds) to their original integration dates. The slope of the regression line, which represents the inferred within-host evolutionary rate (ER) in estimated substitutions per nucleotide site per day, is shown at the bottom right. Faint grey lines denote the ancestral relationships between within-host HIV sequences. The dotted grey lines represent on-ART proviral sequences that were identical to pre-ART plasma sequences. (D) Integration date point estimates (and 95% confidence intervals) for distinct proviral sequences recovered from this participant.

**Fig. 4: Participant 2.**

The panels are as described in the legend of Figure 3.

**Fig. 5: Participant 3.**

The panels are as described in the legend of Figure 3.

**Fig. 6: Participant 4.**

The panels are as described in the legend of Figure 3, with the following additions. Throughout the figure, the red diamond indicated by an asterisk represents the genomically-intact HIV sequence isolated from the sample remnants following the Intact Proviral DNA Assay (IPDA). In panel C, two linear models were fit to the data, encompassing the viremic control (blue dashed line) and the "loss of control" (green dashed line) periods. Corresponding evolutionary rates are shown in the bottom right corner in matching colours. The regression for the control period was

performed using plasma timepoints 2006-06-14 to 2008-09-25, while post-control period regression was performed using plasma timepoints 2009-01-29 to 2012-05-14.

**Fig. 7 Correlates of proviral burden and diversity.**

(A) Relationship between pre-ART plasma HIV RNA diversity and proviral diversity on suppressive ART. (B) Relationship between total area under the plasma viral load curve pre-ART, measured in log10 viremia copy-days, and total proviral *burden* during suppressive ART. (C) Relationship between total area under the plasma viral load curve pre-ART and overall proviral *diversity* during suppressive ART, where the latter is measured as average patristic (tip-to-tip phylogenetic) distance between all distinct proviral sequences recovered from each participant.

**Fig. 8 Model of reservoir composition.**

(A-D) Viral load histories, with engrafted "peak viremia" kinetics from the literature, for participants 1 (panel A), 2 (panel B), 3 (panel C, where the dotted line represents the viremic episode after ART that serves as the reference point in this analysis) and 4 (panel D). (E-H) Phylogenetically-determined proviral ages, and predicted proviral age distributions under different rates of proviral decay, for participants 1-4, respectively. Blue histograms denote the proportion each participant's distinct proviruses that "dated" to each year prior to ART initiation; derived from the integration date point estimates shown in Figures 3D, 4D, 5D and 6D, respectively. The yellow line indicates each participant’s model-predicted proviral age distribution on ART under a scenario of no decay following deposition. The grey and dashed black lines indicate model-predicted proviral age distributions under decay rates of 140 and 44 months respectively (which represent historic estimates of total DNA (40) and replication-competent reservoir (41) half-lives *on ART,* respectively). The green line represents the model-predicted proviral age distributions under the proviral decay rate that best fit each participant's

observed data. (I-L) Best-fit half-lives estimated directly from each participant's observed proviral age distributions using a Poisson generalized linear model along with 95% confidence intervals, for participants 1-4, respectively.

**Table 1**: Participant clinical details, sampling timeline and data summary

| Participant | Clinically estimated infection date | Diagnosis date | First plasma viral load[a] | N Plasma timepoints with HIV sequence data[b] | N Plasma HIV sequences (distinct: N; %) [c] | ART initiation date | Total proviral burden on ART (% genomically intact)[d] | Proviral sequencing date(s) | N Proviral sequences (distinct: N, %) [c] |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Apr 2006 | Feb 2007 | 1615 | 8 | 104 (52; 50%) | 2010-08-01 | 215 (51) | 2012-09-13 | 48 (40, 83%) |
| 2 | Dec 2005 | Jan 2006 | 977 | 9 | 67 (60; 90%) | 2014-05-01 | 580 (10) | 2015-09-08 | 66 (36, 55%) |
| 3 | Mar 2008 | Jun 2008 | 97 | 7 | 130 (71; 55%) | 2012-01-01 | 44 (31) | 2016-07-14 | 24 (12, 50%) |
| 4 | Mar 2005 | Nov 2005 | 383 | 12 | 245 (173; 71%) | 2013-02-01 | 778 (94) | 2013-11-18; 2014-10-20 | 129 (119, 92%) |

[a] First detectable sampled plasma viral load, expressed as HIV RNA copies/ml plasma
[b] Total number of plasma timepoints from which HIV RNA *nef* sequences were successfully amplified.
[c] These values excludes defective, hypermutated and putative within-host recombinant HIV sequences. "Distinct" refers to sequences that were observed only once in that compartment.
[d] as measured by the Intact Proviral DNA Assay and expressed as HIV copies/$10^6$ CD4+ T cells.

**Table 2**: Phylogenetically-inferred root date estimates and within-host evolutionary rates

| Participant | ΔAIC[a] | Clinically estimated infection date | Phylogenetically-estimated Root Date (95% Confidence Interval) | Within-host HIV evolutionary rate[b] |
|---|---|---|---|---|
| 1 | 21.1 | Apr 2006 | Feb 2007 (Dec 2004 to Apr 2009) | $3.22 \times 10^{-5}$ |
| 2 | 37.8 | Dec 2005 | Jan 2003 (Feb 1999 to Dec 2006) | $1.17 \times 10^{-5}$ |
| 3 | 39.2 | Mar 2008 | Oct 2006 (May 2004 to Feb 2009) | $1.16 \times 10^{-5}$ |
| 4 | 30.3<br>167.2 | Mar 2005 | June 2005 (Nov 2003 to Jan 2007) | $1.28 \times 10^{-5}$ (control era)<br>$5.35 \times 10^{-5}$ (post-control era) |

[a] ΔAIC = Delta Akaike Information Criterion. A ΔAIC ≥10 was considered evidence of a molecular clock signal (see methods).
[b] Expressed in estimated substitutions per nucleotide site per day.

Fig 1.

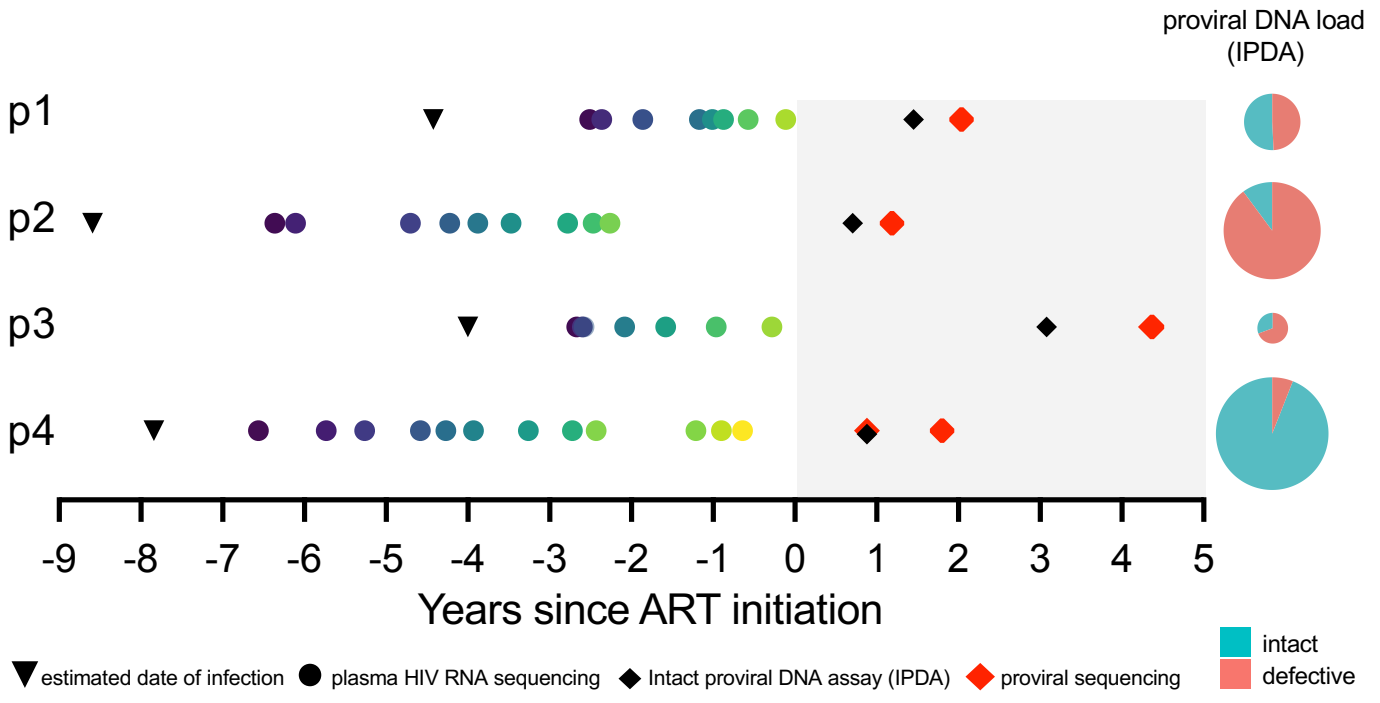


**Fig. 1: Participant sampling timeline and reservoir quantification.** Timeline is depicted as years since ART initiation. Shading represents ART. Inverted black triangles denote the clinically-estimated date of infection. Coloured circles denote pre-ART plasma samples from which HIV RNA sequences were isolated. Red diamonds denote cell samples on ART from which proviral sequences were isolated. Black diamonds denote proviral quantification dates using the Intact Proviral DNA Assay, where results are shown as pie charts where the pie size denotes the total proviral burden and the coloured slices denote intact and defective HIV genome proportions (actual values shown in **Table 1**).

Fig 2.

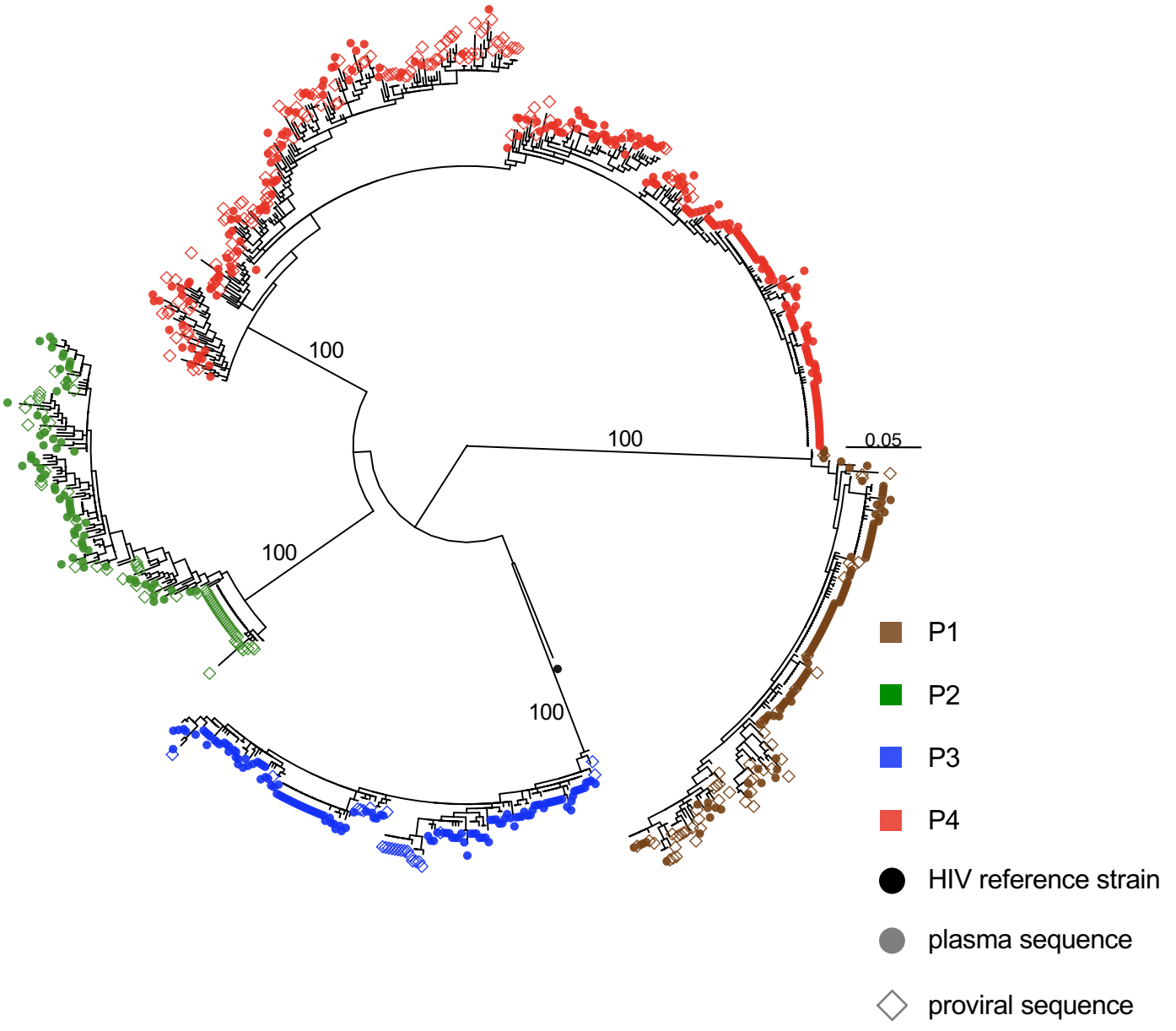


**Fig. 2: Between-host HIV phylogeny.** Maximum likelihood (ML) phylogeny inferred from 546 plasma HIV RNA sequences (circles) and 267 proviral DNA sequences (open diamonds) isolated from study participants. Numbers on internal branches indicate ML bootstrap values supporting within-host monophyletic clades. Scale in estimated substitutions per nucleotide site. The phylogeny is midpoint-rooted. The black dot represents the HIV-1 subtype B reference strain HXB2.

Fig 3.

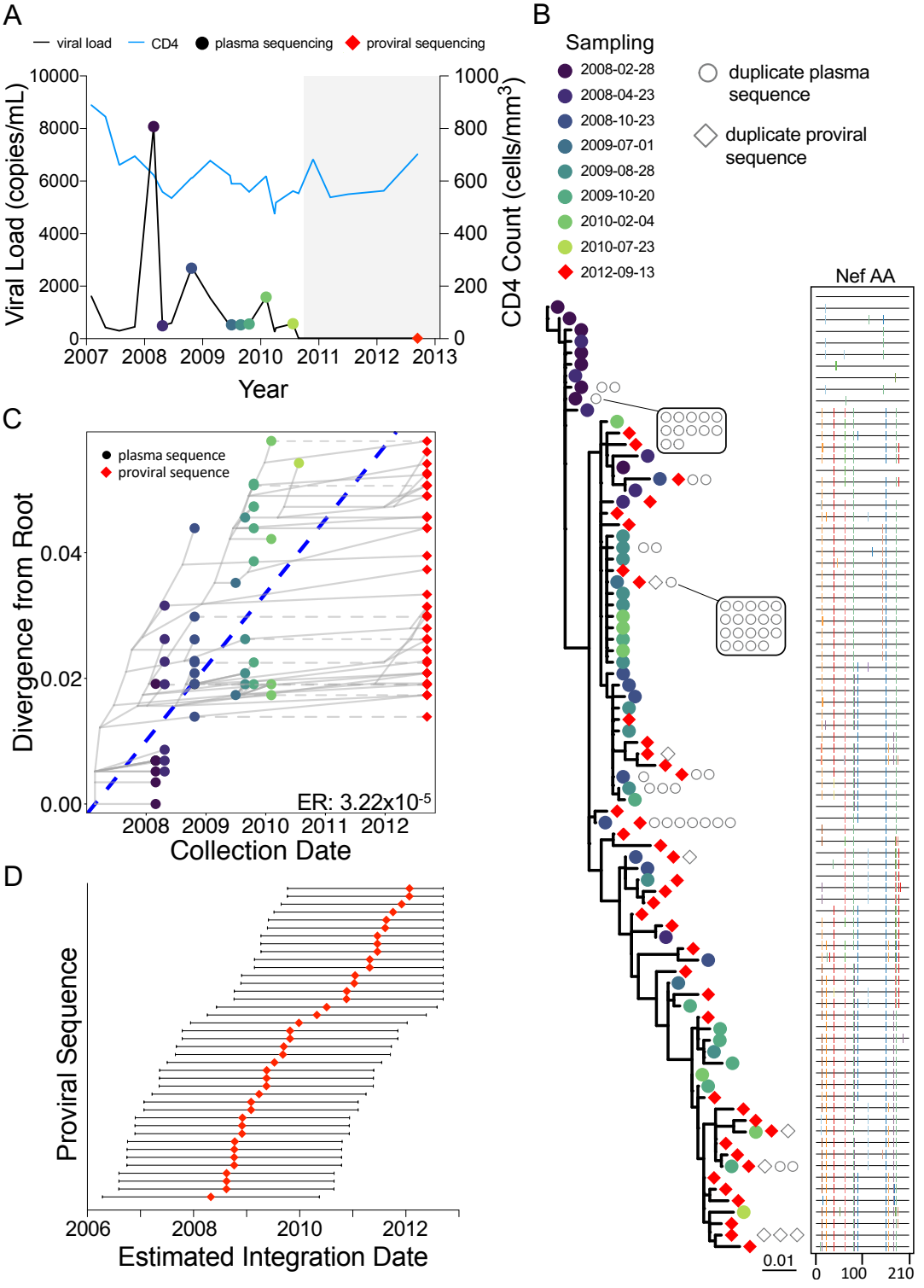


**Fig. 3: Participant 1.** (A) Clinical history and sampling timeline. Throughout all figures, circles denote plasma HIV RNA sampling and diamonds denote HIV DNA sampling. Shading represents ART. (B) Maximum-likelihood within-host phylogeny and corresponding amino acid highlighter plot. In the phylogeny, colored circles denote distinct pre-ART plasma HIV RNA sequences and red diamonds indicate distinct proviral sequences sampled during suppressive ART. Sequences that were recovered repeatedly are shown as open grey circles (for plasma HIV RNA) and diamonds (for proviruses) adjacent to the relevant tip. The root represents the inferred most recent common ancestor of the dataset, representing the phylogenetically-inferred transmitted founder virus event. The highlighter plot is ordered according to the phylogeny, and depicts *amino acid* sequences. The top sequence serves as the reference, where colored ticks in sequences beneath it denote non-synonymous substitutions with respect to the reference. (C) HIV sequence divergence-versus-time plot. The blue dashed line represents the linear model relating the root-to-tip distances of distinct pre-ART plasma HIV RNA sequences (colored circles) to their sampling times. This model is then used to convert the root-to-tip distances of distinct proviral sequences sampled during ART (red diamonds) to their original integration dates. The slope of the regression line, which represents the inferred within-host evolutionary rate (ER) in estimated substitutions per nucleotide site per day, is shown at the bottom right. Faint grey lines denote the ancestral relationships between within-host HIV sequences. The dotted grey lines represent on-ART proviral sequences that were identical to pre-ART plasma sequences. (D) Integration date point estimates (and 95% confidence intervals) for distinct proviral sequences recovered from this participant.

Fig 4.

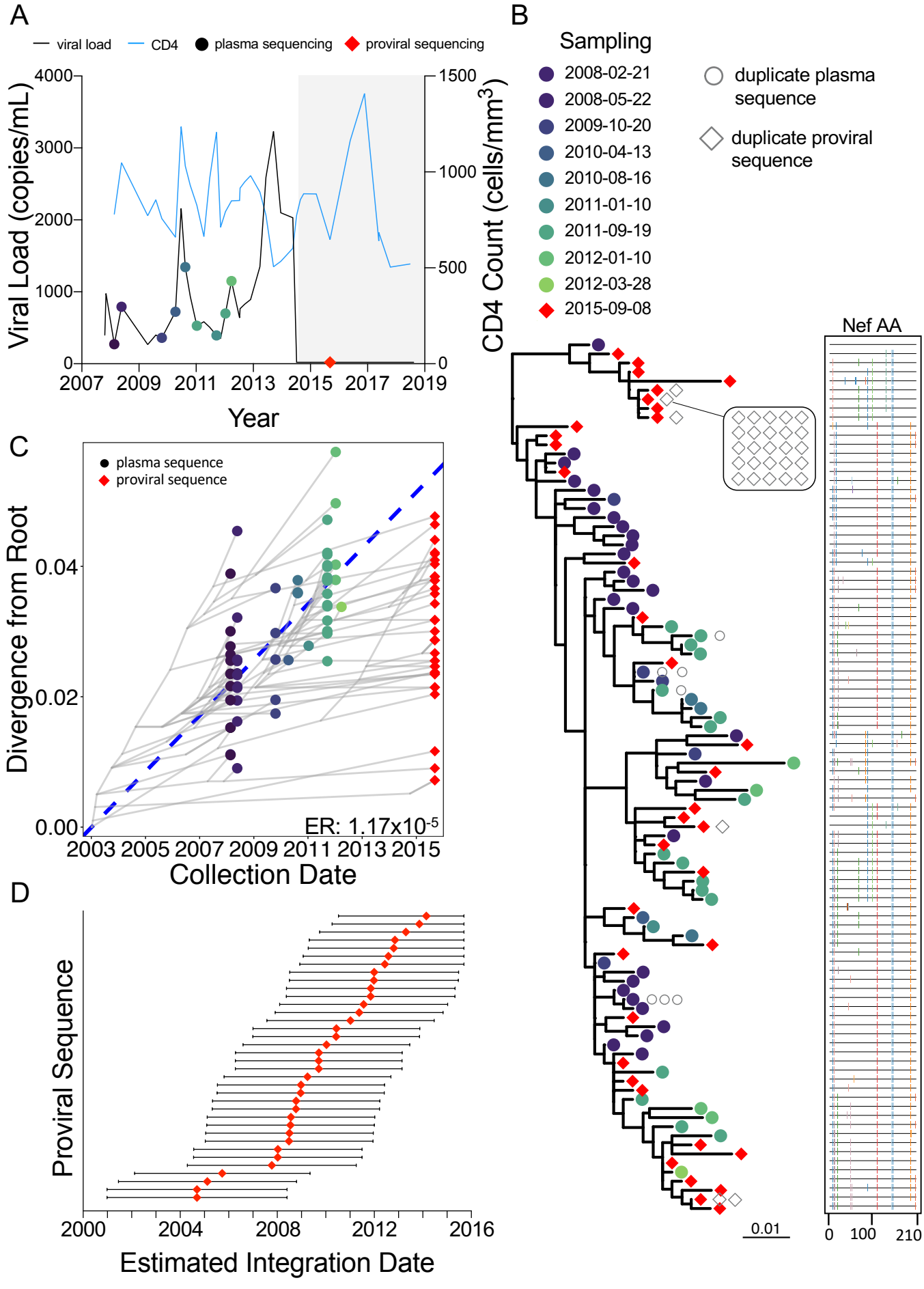


**Fig. 4: Participant 2.** The panels are as described in the legend of Figure 3.

Fig 5.

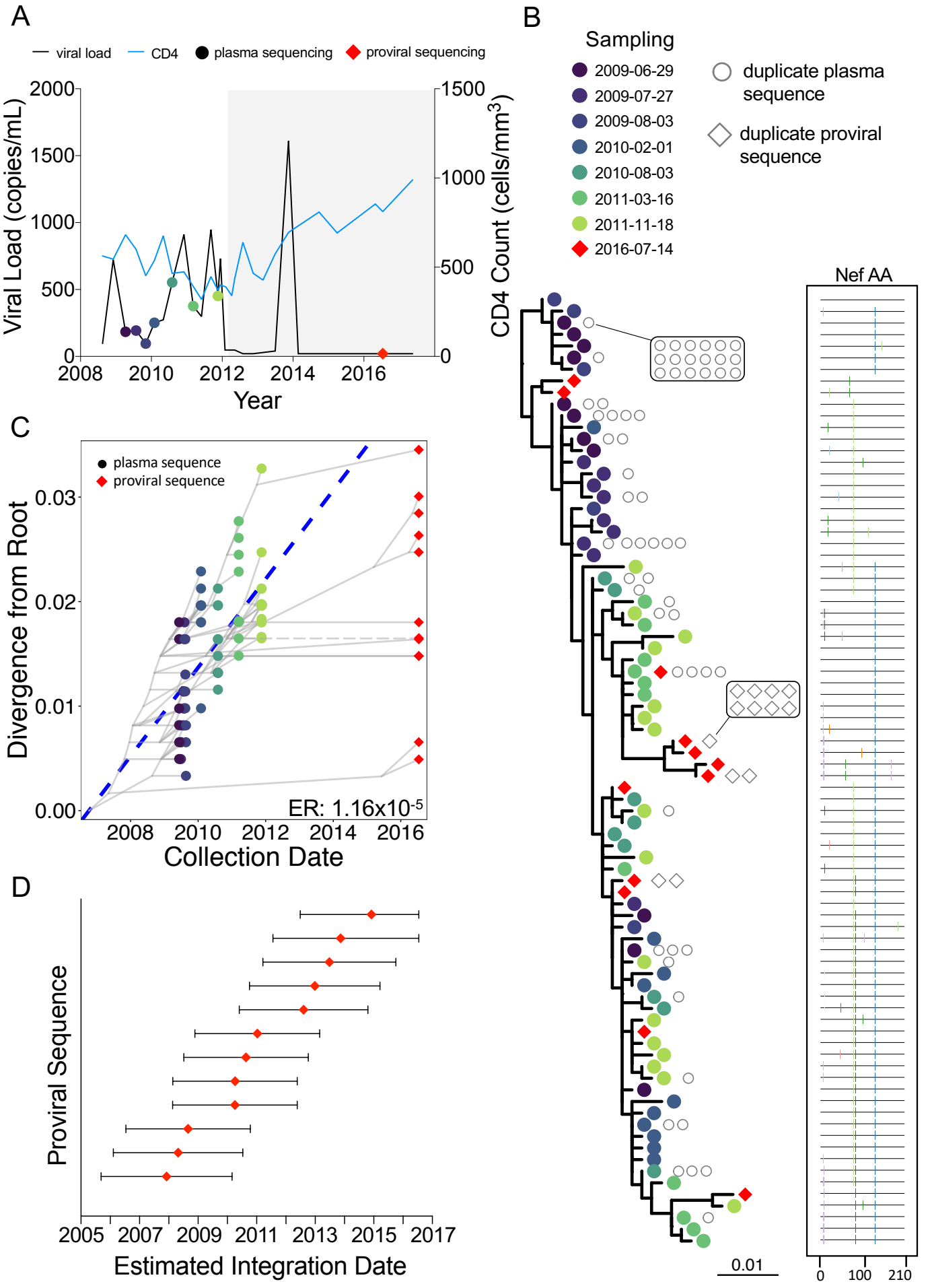


**Fig. 5: Participant 3.** The panels are as described in the legend of Figure 3.

Fig 6.

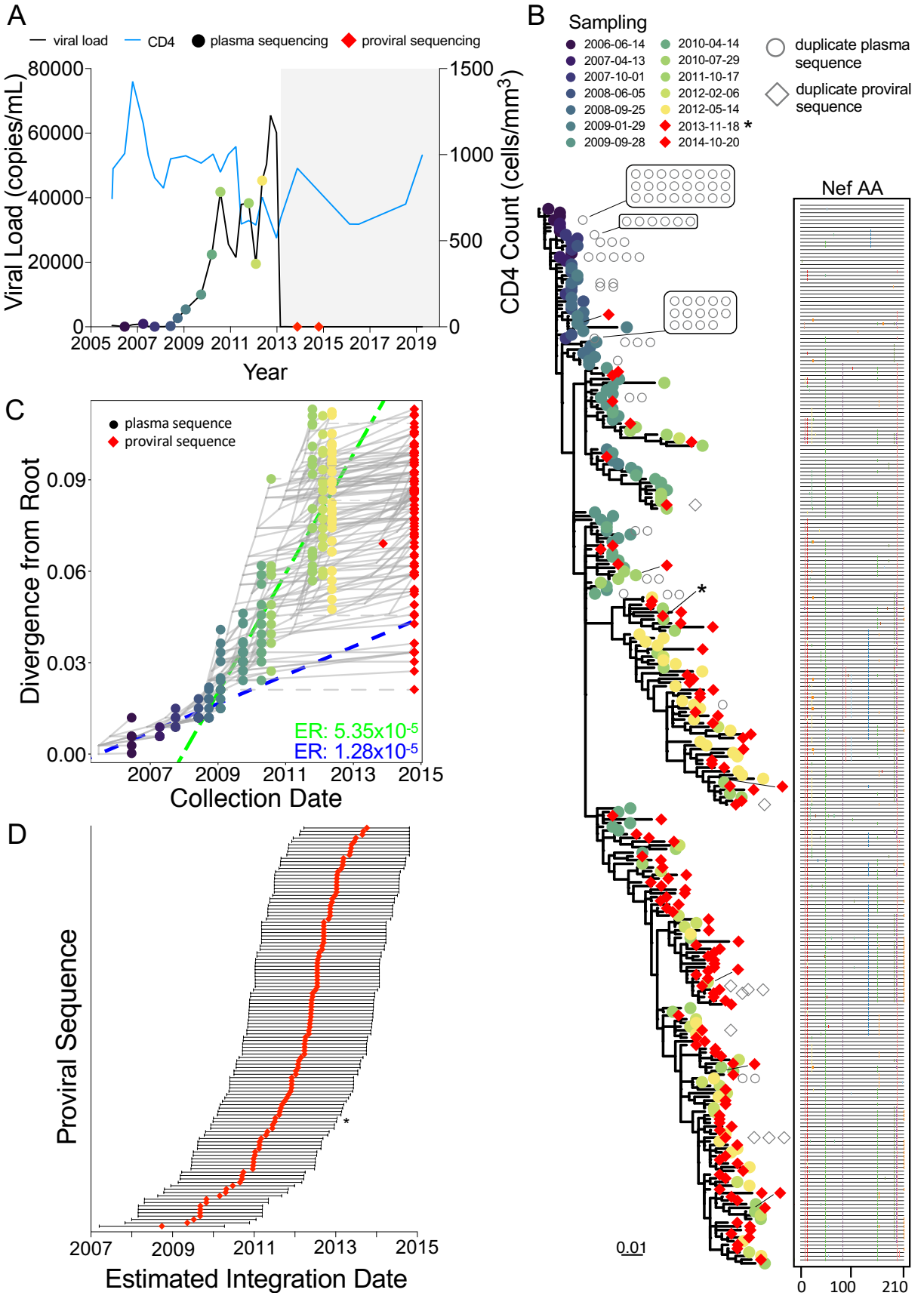


**Fig. 6: Participant 4.** The panels are as described in the legend of Figure 3, with the following additions. Throughout the figure, the red diamond indicated by an asterisk represents the genomically-intact HIV sequence isolated from the sample remnants following the Intact Proviral DNA Assay (IPDA). In panel C, two linear models were fit to the data, encompassing the viremic control (blue dashed line) and the "loss of control" (green dashed line) periods. Corresponding evolutionary rates are shown in the bottom right corner in matching colours. The regression for the control period was performed using plasma timepoints 2006-06-14 to 2008-09-25, while post-control period regression was performed using plasma timepoints 2009-01-29 to 2012-05-14.

Fig 7.

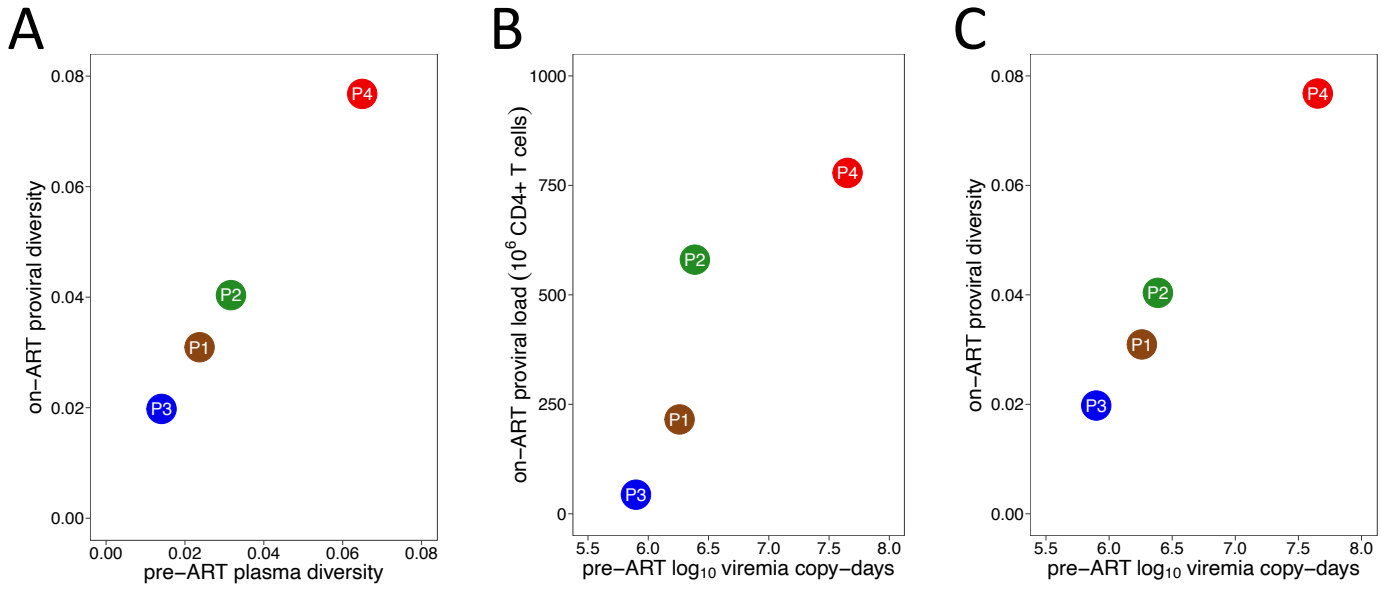


**Fig. 7 Correlates of proviral burden and diversity.** (A) Relationship between pre-ART plasma HIV RNA diversity and proviral diversity on suppressive ART. (B) Relationship between total area under the plasma viral load curve pre-ART, measured in log10 viremia copy-days, and total proviral *burden* during suppressive ART. (C) Relationship between total area under the plasma viral load curve pre-ART and overall proviral *diversity* during suppressive ART, where the latter is measured as average patristic (tip-to-tip phylogenetic) distance between all distinct proviral sequences recovered from each participant.

## Figure 8. Model of reservoir composition

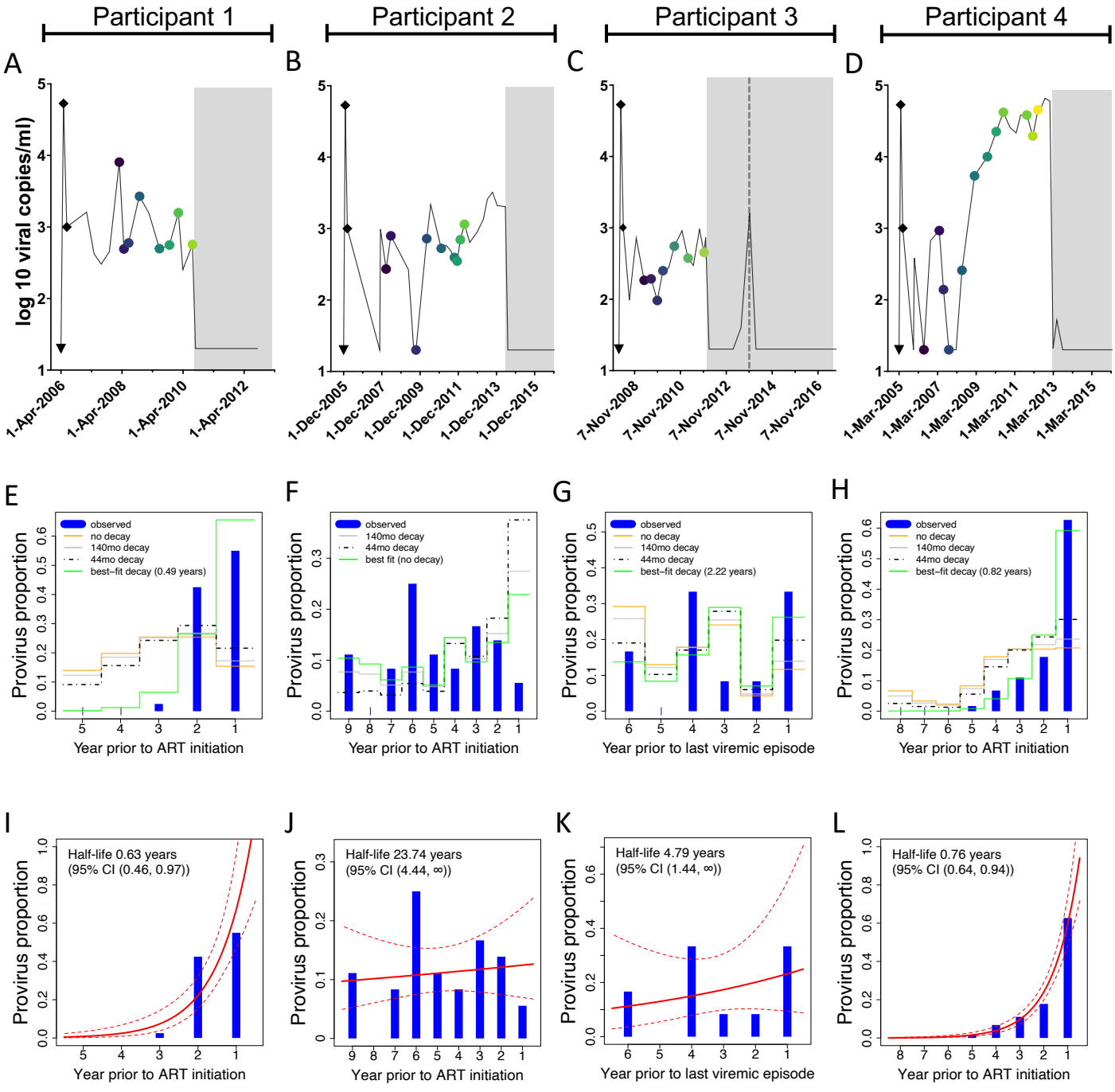


**Fig. 8 Model of reservoir composition.** (A-D) Viral load histories, with engrafted "peak viremia" kinetics from the literature, for participants 1 (panel A), 2 (panel B), 3 (panel C, where the dotted line represents the viremic episode after ART that serves as the reference point in this analysis) and 4 (panel D). (E-H) Phylogenetically-determined proviral ages, and predicted proviral age distributions under different rates of proviral decay, for participants 1-4, respectively. Blue histograms denote the proportion each participant's distinct proviruses that "dated" to each year prior to ART initiation; derived from the integration date point estimates shown in Figures 3D, 4D, 5D and 6D, respectively. The yellow line indicates each participant's model-predicted proviral age distribution on ART under a scenario of no decay following deposition. The grey and dashed black lines indicate model-predicted proviral age distributions under decay rates of 140 and 44 months respectively (which represent historic estimates of total DNA (40) and replication-competent reservoir (41) half-lives *on ART,* respectively). The green line represents the model-predicted proviral age distributions under the proviral decay rate that best fit each participant's observed data. (I-L) Best-fit half-lives estimated directly from each participant's observed proviral age distributions using a Poisson generalized linear model along with 95% confidence intervals, for participants 1-4, respectively.