Virus Evolution, 2025, **11(1)**, veae094

DOI: https://doi.org/10.1093/ve/veae094

Advance Access Publication Date: 11 November 2024

Research Article

A simple phylogenetic approach to analyze hypermutated HIV proviruses reveals insights into their dynamics and persistence during antiretroviral therapy

Aniqa Shahid[1,2], Bradley R. Jones[3], Maggie C. Duncan[1,2], Signe MacLennan[1], Michael J. Dapp[4], Mark H. Kuniholm[5], Bradley Aouizerat[6], Nancie M. Archin[7], Stephen Gange[8], Igho Ofotokun[9], Margaret A. Fischl[10], Seble Kassaye[11], Harris Goldstein[12], Kathryn Anastos[13], Jeffrey B. Joy[2,14,15,‡] and Zabrina L. Brumme,[1,2,‡,*] the MACS/WIHS Combined Cohort Study (MWCCS)

[1]Faculty of Health Sciences, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada

[2]British Columbia Centre for Excellence in HIV/AIDS, 1081 Burrard St., Vancouver, BC V6Z 1Y6, Canada

[3]Department of Mathematics, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada

[4]Department of Microbiology, University of Washington, School of Medicine, 1705 NE Pacific St., Seattle, WA 98195, United States

[5]Department of Epidemiology and Biostatistics, University at Albany, State University of New York, 1 University Place, Rensselaer, NY 12144, United States

[6]College of Dentistry, New York University, 345 E. 24th St., New York, NY 10010, United States

[7]UNC HIV Cure Center, Institute of Global Health and Infectious Diseases, University of North Carolina at Chapel Hill, 130 Mason Farm Rd., Chapel Hill, NC 27599, United States

[8]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Baltimore, MD 21205, United States

[9]Division of Infectious Diseases, Department of Medicine, Emory University School of Medicine, 100 Woodruff Circle, Atlanta, GA 30322, United States

[10]Division of Infectious Diseases, Department of Medicine, University of Miami School of Medicine, 1951 NW 7th Ave., Miami, FL 33136, United States

[11]Division of Infectious Diseases and Tropical Medicine, Georgetown University, 3800 Reservoir Road NW, Washington, DC 20007, United States

[12]Departments of Microbiology and Immunology and Pediatrics, Albert Einstein College of Medicine, 1300 Morris Park Ave., Bronx, NY 10461, United States

[13]Department of Medicine, Albert Einstein College of Medicine, 1300 Morris Park Ave., Bronx, NY 10461, United States

[14]Department of Medicine, University of British Columbia, 2775 Laurel St., Vancouver, BC V5Z 1M9, Canada

[15]Bioinformatics Program, University of British Columbia, 100-570 West 7th Ave., Vancouver, BC V5Z 4S6, Canada

[‡]Equal contribution.

[*]Corresponding author. Faculty of Health Sciences, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada. E-mail: zbrumme@sfu.ca

Abstract

Hypermutated proviruses, which arise in a single Human Immunodeficiency Virus (HIV) replication cycle when host antiviral APOBEC3 proteins introduce extensive guanine to adenine mutations throughout the viral genome, persist in all people living with HIV receiving antiretroviral therapy (ART). However, hypermutated sequences are routinely excluded from phylogenetic trees because their extensive mutations complicate phylogenetic inference, and as a result, we know relatively little about their within-host evolutionary origins and dynamics. Using >1400 longitudinal single-genome-amplified HIV *env-gp120* sequences isolated from six women over a median of 18 years of follow-up—including plasma HIV RNA sequences collected over a median of 9 years between seroconversion and ART initiation, and >500 proviruses isolated over a median of 9 years on ART—we evaluated three approaches for masking hypermutation in nucleotide alignments. Our goals were to (i) reconstruct phylogenies that can be used for molecular dating and (ii) phylogenetically infer the integration dates of hypermutated proviruses persisting during ART. Two of the approaches (stripping all positions containing putative APOBEC3 mutations from the alignment or replacing individual putative APOBEC3 mutations in hypermutated sequences with the ambiguous base R) consistently normalized tree topologies, eliminated erroneous clustering of hypermutated proviruses, and brought *env*-intact and hypermutated proviruses into comparable ranges with respect to multiple tree-based metrics. Importantly, these corrected trees produced integration date estimates for *env*-intact proviruses that were highly concordant with those from benchmark trees that excluded hypermutated sequences, supporting the use of these corrected trees for molecular dating. Subsequent molecular dating of hypermutated proviruses revealed that these sequences spanned a wide within-host age range, with the oldest ones dating to shortly after infection. This indicates that hypermutated proviruses, like other provirus types, begin to be seeded into the proviral pool immediately following infection and can persist for decades. In two of the six participants, hypermutated proviruses differed from *env*-intact ones in terms of their age distributions, suggesting that different provirus types decay at heterogeneous rates in some hosts. These simple approaches to reconstruct hypermutated provirus' evolutionary histories reveal insights into their *in vivo* origins and longevity toward a more comprehensive understanding of HIV persistence during ART.

Keywords: HIV; persistence; reservoir; provirus; hypermutation; evolution; phylogenetics

Introduction

Antiretroviral therapy (ART) is not curative because Human Immunodeficiency Virus (HIV) persists as an integrated provirus within infected cell reservoirs (Finzi et al. 1997, 1999). Entry of HIV sequences into these reservoirs begins immediately following infection (Whitney et al. 2014, Gantner et al. 2023) and continues until viral suppression is achieved on ART, yielding a genetically diverse pool of persisting HIV sequences (Brodin et al. 2016, Jones et al. 2018, Brooks et al. 2020, Pankau et al. 2020, Nicolas et al. 2022, Kinloch et al. 2023). Only a minority (∼2%–5%) of integrated proviruses persisting on ART, however, are genetically intact and potentially capable of producing replication-competent HIV (Sanchez et al. 1997, Ho et al. 2013, Bruner et al. 2016, Imamichi et al. 2016). The remainder are genetically defective and cannot produce infectious viruses, although some can produce viral proteins (Imamichi et al. 2016, 2020, Pollack et al. 2017) that contribute to chronic immune activation and associated comorbidities (Deeks et al. 2013). Large deletions, which occur during the minus-strand synthesis step of reverse transcription, are typically the most common proviral defects, followed by hypermutation (Ho et al. 2013, Bruner et al. 2016, Hiener et al. 2017, Lee et al. 2017, Kinloch et al. 2023).

Hypermutated proviruses arise in a single HIV replication cycle when host antiviral APOBEC3 proteins catalyze widespread cytidine-to-uridine deamination within the minus-strand HIV DNA genome that is produced during reverse transcription, yielding extensive guanine to adenine (G-to-A) mutations during plus-strand synthesis (Goodenow et al. 1989, Vartanian et al. 1991, Fitzgibbon et al. 1993). Hypermutation is normally deleterious, yielding nonsense and/or missense mutations that render viral proteins (or regulatory genetic elements) nonfunctional, thereby inhibiting viral replication (Harris and Liddament 2004, Armitage et al. 2012, Waldron 2015). As a result, hypermutated proviruses do not generally yield evolutionary descendants (Sheehy et al. 2002, Kieffer et al. 2005). Nevertheless, hypermutated sequences readily persist, typically representing 15% (though as much as >50%) of all proviruses during long-term ART (Ho et al. 2013, Bruner et al. 2016, Hiener et al. 2017, Lee et al. 2017, Kinloch et al. 2023).

Hypermutated HIV sequences pose challenges for phylogenetic inference. In general, trees inferred directly from sequence alignments containing hypermutated proviruses will inaccurately reflect the ancestor–descendant relationships of these sequences: the terminal branch lengths (TBL) of hypermutated sequences will typically be extended due to their large number of G-to-A mutations, and they will also often cluster together due to a type of phylogenetic error known as long-branch attraction, whereby divergent sequences are grouped together simply because they have undergone a large amount of change, not because they share recent ancestry (Bergsten 2005). These errors occur in part because phylogenetic algorithms assume that mutations gradually accumulate over generations, not all at once in a single round of replication (Gorbalenya 2017), and also because identical and widespread G-to-A mutations occurring at specific APOBEC3 target sites will cause otherwise unrelated genomes to have many mutations in common. Although hypermutated sequences can be included in phylogenies simply as a way to visualize complete datasets (Kearney et al. 2016, Patro et al. 2019, Halvas et al. 2020), such trees should not be used for formal hypothesis testing.

Because of these challenges, hypermutated sequences are typically removed from HIV alignments prior to phylogenetic inference (Brodin et al. 2016, Jones et al. 2018, 2023, Bozzi et al. 2019, Pinzone et al. 2019, Brooks et al. 2020). However, this is a shame because they carry mutations from prior rounds of replication that could aid phylogenetic reconstruction by increasing the overall sample size and mutational depth of the sampled population. Their routine exclusion from phylogenies also means that we understand relatively little about their within-host origins and longevity.

To address this, we used longitudinal within-host HIV *env-gp120* sequence datasets from six participants of the Women's Interagency HIV Study (WIHS) (Shahid et al. 2024) to evaluate the ability of three simple nucleotide alignment modification strategies to normalize the topologies of trees containing hypermutated proviruses. Using these corrected trees, we then estimated the integration dates of *env*-intact and hypermutated proviruses persisting during ART, in order to better understand the within-host evolutionary dynamics of these different proviral types.

Results

Within-host HIV sequence datasets

We analyzed 1408 single-genome-amplified HIV *env-gp120* sequences collected longitudinally from six WIHS participants who experienced HIV seroconversion (a seventh participant from the original study was not included here, as no hypermutated proviruses were isolated from their samples) (Shahid et al. 2024) (Table 1). The data included 865 distinct HIV RNA *env-gp120* sequences (median 157 per participant) isolated from plasma over a median of 9 time points spanning a median of 7 years between seroconversion and ART initiation. The data also included 542 distinct *env-gp120* proviral sequences, including 449 *env*-intact ones (median 62 per participant) and 93 hypermutated ones (median 19 per participant) isolated from peripheral blood at a minimum of 3 time points over a median of 8.7 years during ART (Table 1). All participants had HIV subtype B, with no evidence of dual or super-infection.

Identifying hypermutation, modifying alignments and evaluating tree metrics

Hypermutated HIV sequences were identified using Hypermut 2.0 (Rose and Korber 2000) (see methods for additional details). Between 6% and 30% of participants' proviral sequences were hypermutated (although hypermutation was not observed in any plasma HIV RNA sequences, as expected). In a given within-host alignment, between 9% and 11% of *env-gp120* nucleotide positions had a putative APOBEC3-driven A in at least one sequence (Table 2). Hypermutated proviruses harbored an overall range of 9–83 putative APOBEC3 mutations per *env-gp120* sequence (representing 6%–61% of all possible target sites and 0.6%–5% of all *env-gp120* nucleotides), with a grand median of 45 (representing 31% of all possible target sites and 3% of all *env-gp120* nucleotides).

Table 1. Participant information, HIV sampling, and sequencing details.

ID[a]	Estimated date of infection	Duration of uncontrolled infection (years)	No. of pre-ART plasma HIV RNA time points	Distinct pre-ART plasma HIV *env-gp120* sequences	ART initiation date	Years of ART until last proviral sampling	No. of on-ART proviral time points	Distinct on-ART HIV *env-gp120* proviral sequences (hypermutated *n*; %)
WIHS-P2	January 2003	9	10	227	January 2012	6.8	3	75 (22; 28%)
WIHS-P4[b]	July 1995	10.9	9	182	June 2006	12.3	4	155 (23; 15%)
WIHS-P1	December 1995	12	13	207	January 2008	10.3	4	85 (15; 13%)
WIHS-P3	July 2002	5.5	9	132	January 2008	8.8	3	59 (5; 8%)
WIHS-P5	March 2008	1.9	2	44	February 2010	8.7	3	74 (22; 30%)
WIHS-P6	August 2006	3.9	6	73	July 2010	8.3	4	94 (6; 6%)

[a]Participants are numbered in the same order as the original manuscript (Shahid et al. 2024). That is, WIHS-P2 in the present study is Participant 2 in Shahid et al. 2024.

[b]The MWCCS database indicated that Participant 4 initiated ART in 2003, but no reductions in plasma viral load were observed until June 2006. For this reason, we considered June 2006 as this participant's effective ART start date.

Table 2. Hypermutated sequence details.

ID	Hypermutated proviruses	Aligned HIV *env-gp120* sequence length (bp)	Putative hypermutated nucleotide positions in the alignment[a]	Hypermutated sites identified per sequence, median (range)[b]
WIHS-P2	22	1515	140	43 (20–68)
WIHS-P4	23	1541	176	55 (34–83)
WIHS-P1	15	1483	141	41 (10–75)
WIHS-P3	5	1486	127	40 (36–64)
WIHS-P5	22	1500	152	57 (9–78)
WIHS-P6	6	1523	122	47 (35–75)

[a]The total number of nucleotide positions that harbored an A base at an APOBEC3 target site in at least one hypermutated sequence in the participant's sequence alignment. These positions were stripped out of the alignment in the HM-Stripped approach.

[b]Statistics summarizing the overall number of A bases at APOBEC3 target sites in the participant's hypermutated sequences. These A bases were changed to R or G, respectively, in the HM-Replacedw/R and HM-Replacedw/G approaches.

For context, the grand median of putative APOBEC3 mutations in *env*-intact (non-hypermutated) proviruses was 5. As described in the methods, we prepared five within-host *env-gp120* sequence alignments for each participant. The first alignment, which we called "env-intact only", contained all pre-ART *env-gp120* plasma HIV RNA sequences plus the *env*-intact proviruses sampled during ART (i.e., hypermutated (HM) proviruses were excluded). The second alignment, which we called "HM-Unaltered", contained all pre-ART plasma HIV RNA sequences plus all proviruses (i.e., both *env*-intact and hypermutated) sampled during ART. The next three alignments used different strategies to mask hypermutation: "HM-Stripped" removed all positions in the alignment that harbored an A at an APOBEC3 target site in at least one hypermutated sequence, "HM-Replacedw/R" replaced all individual A bases at APOBEC3 target sites within hypermutated sequences with R (denoting A or G), while the "HM-Replacedw/G" strategy replaced these with G. Visualizations of the HM-Unaltered, HM-Stripped and HM-Replacedw/R alignments are provided in Supplementary Figure S1.

After inferring phylogenies from these alignments, we then applied a variety of metrics to these trees, as described in the methods and in Fig. 1. These metrics allowed us to evaluate the extent to which the alignment modification strategies normalized the position of hypermutated proviruses in the tree, and the overall tree topology.

Assessing how alignment modification strategies normalized tree topology and metrics

Participant WIHS-P2's dataset included 227 plasma HIV RNA *env-gp120* sequences sampled over 9 years during untreated infection and 75 proviruses (53 *env*-intact, 22 hypermutated) sampled over ∼7 years during ART (Fig. 2a). WIHS-P2's unmodified nucleotide alignment yielded a phylogeny that placed nearly all hypermutated proviruses into a single clade (Fig. 2b), consistent with long-branch attraction. Notably, this erroneous clustering was falsely well supported, with >50% of nodes in this clade having high (≥90%) bootstrap support (see larger version of this tree in Supplementary Fig. S2), illustrating the pitfalls of inferring phylogenies directly from such alignments. Moreover, hypermutated provirus terminal branch lengths (TBL) in this tree were on average four times longer than *env*-intact ones ($P < .0001$; Fig. 2c), although their root-to-tip (RTT) distances were not significantly inflated ($P = .2$, Supplementary Fig. S3a). Hypermutated proviruses also exhibited significantly higher evolutionary distinctiveness (ED) than *env*-intact ones in this tree ($P < .0001$ for both fair proportion ED and equal splits ED); Fig. 2d and Supplementary Fig. S4A). Also reflecting the erroneous clustering of hypermutated sequences in this tree, the median number of nodes separating hypermutated sequences from one another [i.e. topological distance (TD)] was on average only half of that separating *env*-intact proviruses ($P < .0001$; Fig. 2e). A Slatkin–Maddison (SM) test also returned significant evidence of genetic population structure (i.e. "compartmentalization") between hypermutated and *env*-intact proviruses in this tree (three inferred migrations; estimated $P=0$; Fig. 2b, inset), as did the Simmonds Association Index (AI; estimated $P=0$; Supplementary Table S1).

By contrast, the tree inferred from WIHS-P2's HM-Stripped alignment, in which 140 (of 1515) *env-gp120* positions harboring putative APOBEC3 mutations had been removed, exhibited a substantially normalized topology (Fig. 2f). The same was true for the tree inferred from the HM-Replacedw/R alignment, where a median of 43 putative APOBEC3-driven A bases in hypermutated sequences had been replaced with R (Fig. 2g; larger trees in Supplementary Fig. S2). In both trees, hypermutated proviruses were now comparable to *env*-intact ones in terms of TBLs (both $P>.1$; Fig. 2h and i), ED (all $P>.1$; Fig. 2j and k; Supplementary Fig. S4b and c), and TD (both $P>.1$, Fig. 2l and m). Genetic compartmentalization between *env*-intact and hypermutated proviruses was also markedly reduced (15 inferred migrations compared to the original 3 using the SM test), although the *P*-values computed using the SM and Simmonds AI tests remained statistically significant (all $P\leq.01$; Fig. 2f and g, insets, and Supplementary Fig. S2). Of note, RTT distances of hypermutated proviruses in these two trees were now shorter than those of *env*-intact ones (both $P<.001$; Supplementary Fig. S3b and c). In contrast, while the tree inferred from participant WIHS-P2's HM-Replacedw/G alignment (where putative APOBEC3-driven A bases in hypermutated sequences were replaced with G) appeared broadly normalized, *env*-intact and hypermutated sequences remained highly significantly compartmentalized in this tree (estimated $P=0$ using the SM test; Supplementary Fig. S5).

As our second example, participant WIHS-P4's dataset included 182 plasma HIV RNA *env-gp120* sequences sampled over ∼11 years pre-ART, and 155 proviruses (132 *env*-intact; 23 hypermutated) sampled during 12 years of ART (Fig. 3a). The unaltered alignment produced a phylogeny (Fig. 3b; larger tree in Supplementary Fig. S6) where hypermutated sequences exhibited significantly inflated branch lengths, RTT distances and ED (all $P<.0001$; Fig. 3c and d, Supplementary Figs 3d and 4d), erroneous clustering ($P<.0001$ Fig. 3e), and significant compartmentalization (estimated $P=0$ using the SM test; Fig. 3b, inset). By contrast, the HM-Stripped and HM-Replacedw/R alignments produced substantially normalized trees (Fig. 3f and g, respectively; larger trees in Supplementary Fig. S6) with no genetic compartmentalization between *env*-intact and hypermutated sequences (22 migrations compared to the original 6; both $P>.1$; Fig. 3f and g, insets). The ranges of TBLs, RTT distance measurements, ED measures, and TDs were now also comparable between *env*-intact and hypermutated proviruses, although the latter remained modestly yet statistically significantly different from *env*-intact sequences by most measures (*P*-values from .001 to .039, Fig. 3h–m; Supplementary Fig. S3e and f; Supplementary Fig. S4e and f). In contrast, hypermutated sequences remained highly compartmentalized in the phylogeny inferred from WIHS-P4's HM-Replacedw/G alignment (Supplementary Fig. S7).

The same analyses were applied to participants WIHS-P1, WIHS-P3, WIHS-P5, and WIHS-P6 (small trees and select metrics in Supplementary Figs S8–S11; large trees in Supplementary Figs S12–S15; and remaining metrics in Supplementary Figs S3 and S4). Broadly, the trees inferred from the HM-Stripped and HM-Replacedw/R alignments were markedly normalized and yielded metric values for *env*-intact and hypermutated proviruses that spanned comparable ranges. For some participants, these metrics normalized such that *env*-intact and hypermutated viruses became statistically comparable (e.g. WIHS-P5; Supplementary Fig. S10). For others, hypermutated sequences remained somewhat distinctive (e.g. hypermutated provirus TBLs and ED remained slightly elevated for WIHS-P6; Supplementary Figs S4 and S11), but in all cases, these differences were far smaller in magnitude than those from the trees inferred from unaltered alignments. Indeed, the *P*-values derived from comparing *env*-intact and hypermutated proviruses in the HM-Stripped and HM-Replacedw/R trees were an average >3 logs higher than those from the HM-Unaltered trees, with 56% of comparisons yielding *P*-values >.05 (Fig. 4).

By contrast, the HM-Replacedw/G approach did not reliably normalize all trees. In particular, WIHS-P5's HM-Replacedw/G phylogeny maintained obvious clustering of hypermutated sequences and very strong compartmentalization, while TBL, FP-ED, and TD also remained highly skewed for one or more participants (Fig. 4, and data not shown). As such, only the HM-Stripped and HM-Replacedw/R trees were advanced for further evaluation.

Of note, for all participants, maximum likelihood (ML) scores for tree topologies inferred under the HM-Stripped and HM-Replacedw/R strategies were substantially better than the topologies inferred using unaltered alignments as judged by Shimodaira–Hasegawa tests (Supplementary Table S2).

Inferring proviral integration dates from corrected trees: a validation

We next investigated whether accurate evolutionary information can be extracted from these corrected trees, by phylogenetically inferring the integration dates of proviruses sampled during ART. Figure 5 illustrates how this is done. Briefly, we first root the phylogeny at the location that maximizes the correlation between the RTT distances of the pre-ART plasma HIV RNA sequences and their sampling dates (proviruses sampled during ART, though included in the tree, are not considered in this correlation; Fig. 5b). This root represents the most recent common ancestor (MRCA) of the dataset (i.e. the estimated founder virus). We then fit a linear model relating the RTT genetic distances of the pre-ART plasma sequences to their sampling dates (Fig. 5c). This model is then used to convert the RTT distance of each on-ART provirus into its inferred integration date (plus 95% confidence interval; Fig. 5d).

Application of this approach to WIHS-P2's unaltered and corrected trees yielded estimated root dates that were consistent with the clinically estimated infection date (Table 1) and comparable to the root date inferred from the benchmark (*env*-intact only) tree (Supplementary Table S3; the likely reason that the unaltered tree produced reasonable root dates and evolutionary rate estimates is that these metrics are computed from pre-ART plasma HIV RNA sequences only). We next verified the extent to which the integration dates of *env*-intact proviruses inferred from the corrected trees matched those inferred from the benchmark tree (which, per current field standards, excluded hypermutated sequences entirely). Reassuringly, *env*-intact proviral integration dates inferred from the HM-Stripped tree were highly concordant with those inferred from the benchmark tree [Spearman's rho (ρ) = 0.95, $P<.0001$; Lin's concordance correlation coefficient ($\rho_c=0.96$)], as were those inferred from the HM-replacedw/R tree ($\rho=0.98$, $P<.0001$; $\rho_c=0.97$) (Fig. 6a). These results indicate that WIHS-P2's corrected trees can be used for molecular dating, and produce valid proviral integration dates.

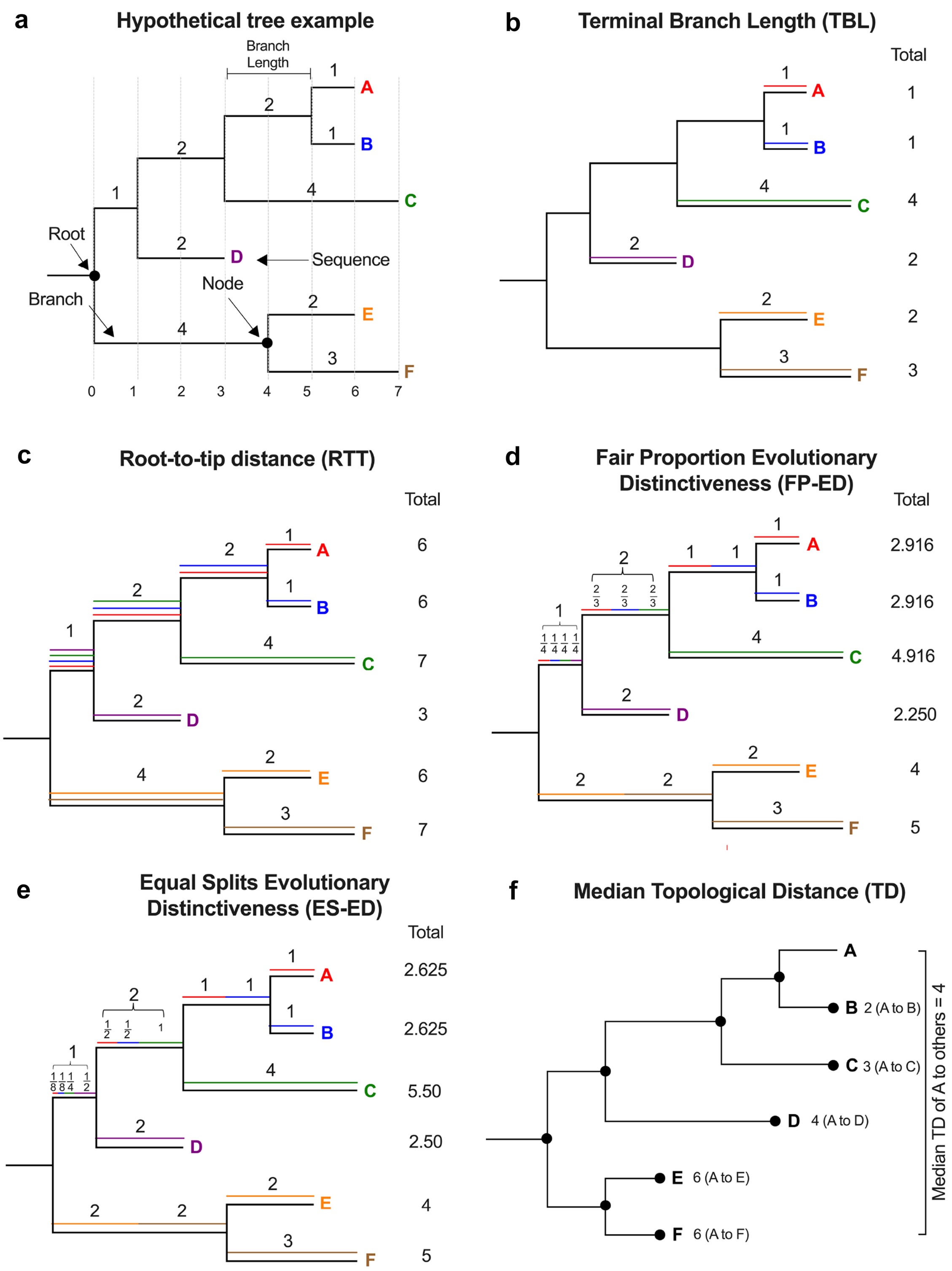

Figure 1. Tree-based metrics. (a) Hypothetical tree containing six sequences, labeled A–F, each in a unique color. Vertical dotted lines depict the distance scale in hypothetical units numbered below the tree. All other panels depict this same tree. (b) Colored horizontal lines trace the TBLs of sequences A–F, with the values also shown at the right of the tree. (c) Each sequence's path from root to tip is traced with a unique color, where the sum of these lengths (representing the RTT distance) is shown at the right of the tree. (d, e) FP-ED divides the shared evolutionary history represented by an internal branch equally among all its descendant sequences at the tips. Here, colored lines and associated fractional branch lengths show how internal branch lengths are apportioned to each sequence. The sum of each sequence's branch measurements, the FP-ED, is shown at the right of the tree. (e) In contrast, ES-ED assigns 50% of each internal branch length to each immediate descendant. As such, branches leading to a single descendant assign 50% of that branch to this descendant, whereas branches leading to multiple descendants further split the remaining 50% among them using this same scheme. The sum of these measurements, the ES-ED, is shown at the right of the tree. (f) The TD separating sequence A from all others is shown to the right of each tip, where TD is computed as the total number of nodes separating A from all others in the tree. Here, the median TD separating A from all others in the tree is 4.

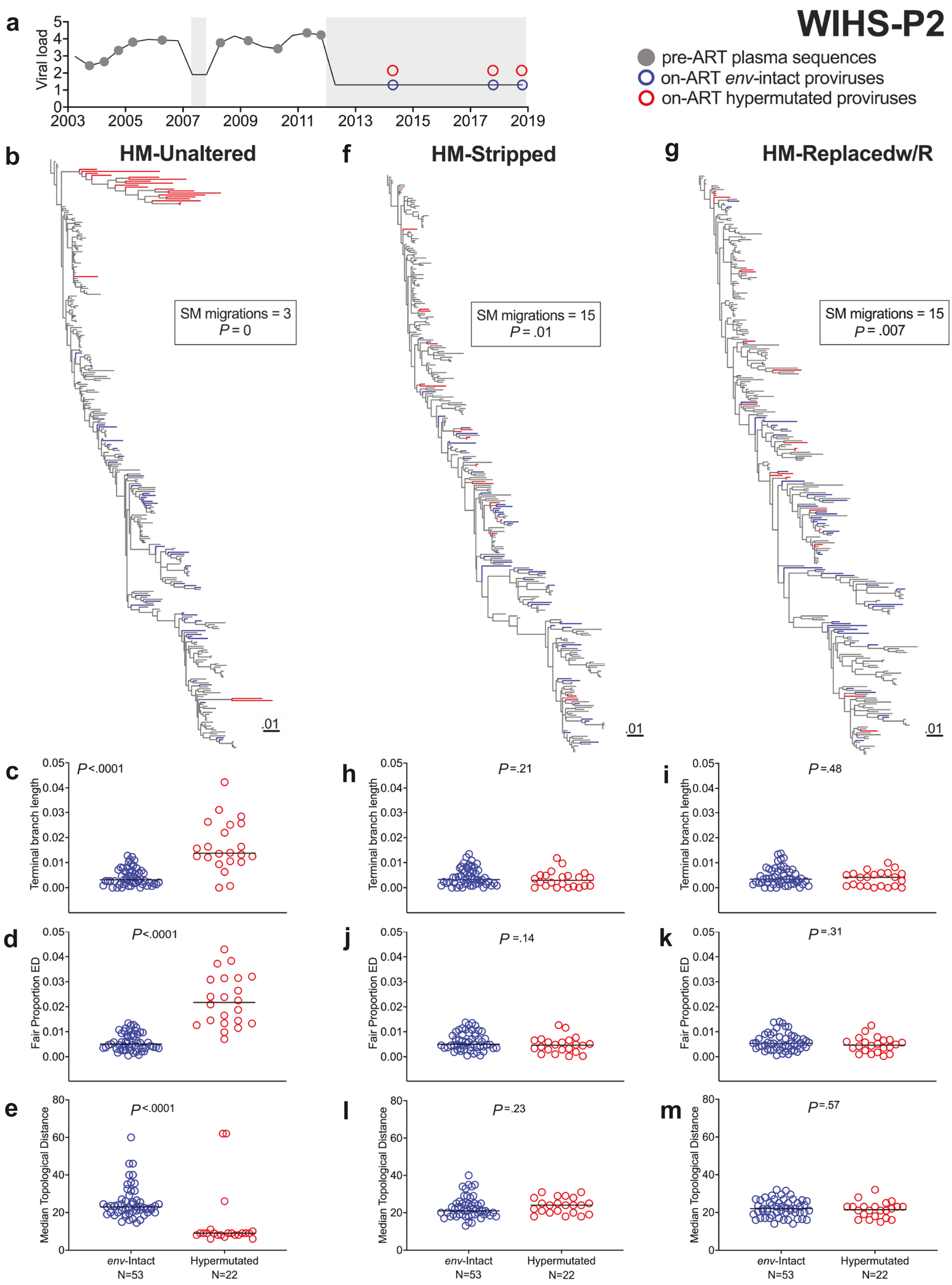

Figure 2. WIHS-P2 clinical history, within-host phylogenies, and tree metrics. (a) Participant WIHS-P2's plasma viral load history and sampling timeline. Closed gray circles denote pre-ART plasma HIV RNA sampling. Open circles denote proviral sampling on ART (blue for *env*-intact proviruses and red for hypermutated proviruses). Gray shading denotes ART. (b) Participant WIHS-P2's rooted ML phylogeny, inferred from all within-host *env-gp120* sequences including hypermutated proviruses. Branches are colored by sequence type (pre-ART HIV RNA = gray; on-ART *env*-intact provirus = blue; on-ART hypermutated provirus = red). Inset shows the number of inferred migrations between *env*-intact and hypermutated sequence groups computed using the SM test, along with the estimated *P*-value. Here, *P* = 0 can be interpreted as *P* < .001, as 1000 permutations were performed. (c) TBLs of *env*-intact and hypermutated sequences in this tree. Horizontal black lines denote the median values. *P*-value computed using the Mann–Whitney *U* test. (d) FP-ED values for *env*-intact and hypermutated sequences in this tree. (e) Median TDs separating *env*-intact and hypermutated proviruses from others of the same type. (f–l) Same as (b–e), but for the phylogeny inferred from an alignment where positions containing hypermutation were stripped out. (g–m) Same as (b–e), but for the phylogeny inferred from an alignment where hypermutated sites were replaced with R.

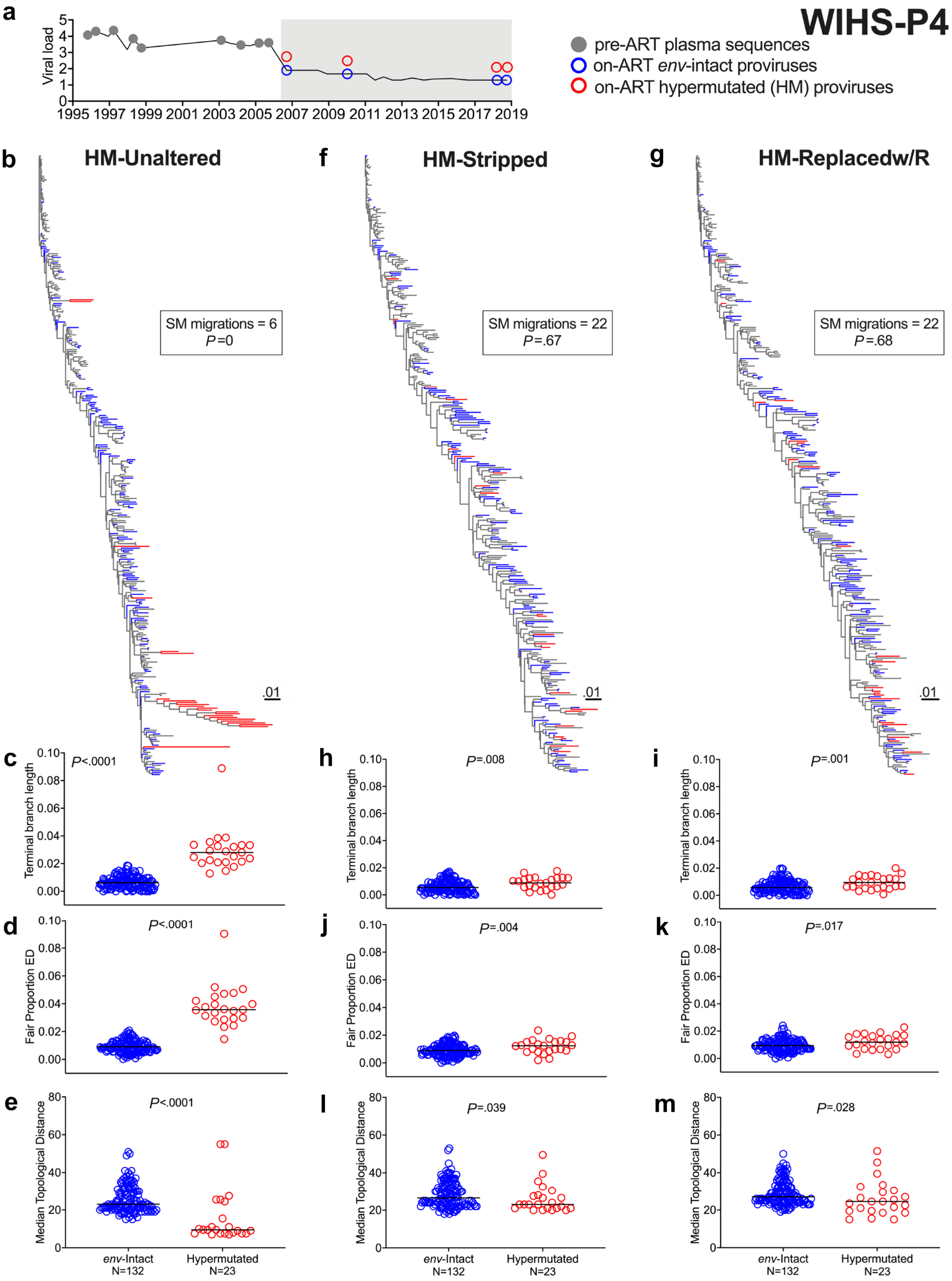

Figure 3. WIHS-P4 clinical history, within-host phylogenies, and tree metrics. Legend as in Fig. 2, except for participant WIHS-P4.

We next inferred the integration dates of all proviruses including the hypermutated ones, from the corrected trees. Inferred integration dates were highly concordant between the two approaches, yielding ρ_c between 0.93 and 0.97 depending on whether we compared *env*-intact, hypermutated, or all proviruses (Fig. 6b). Moreover, there was no bias between the two methods

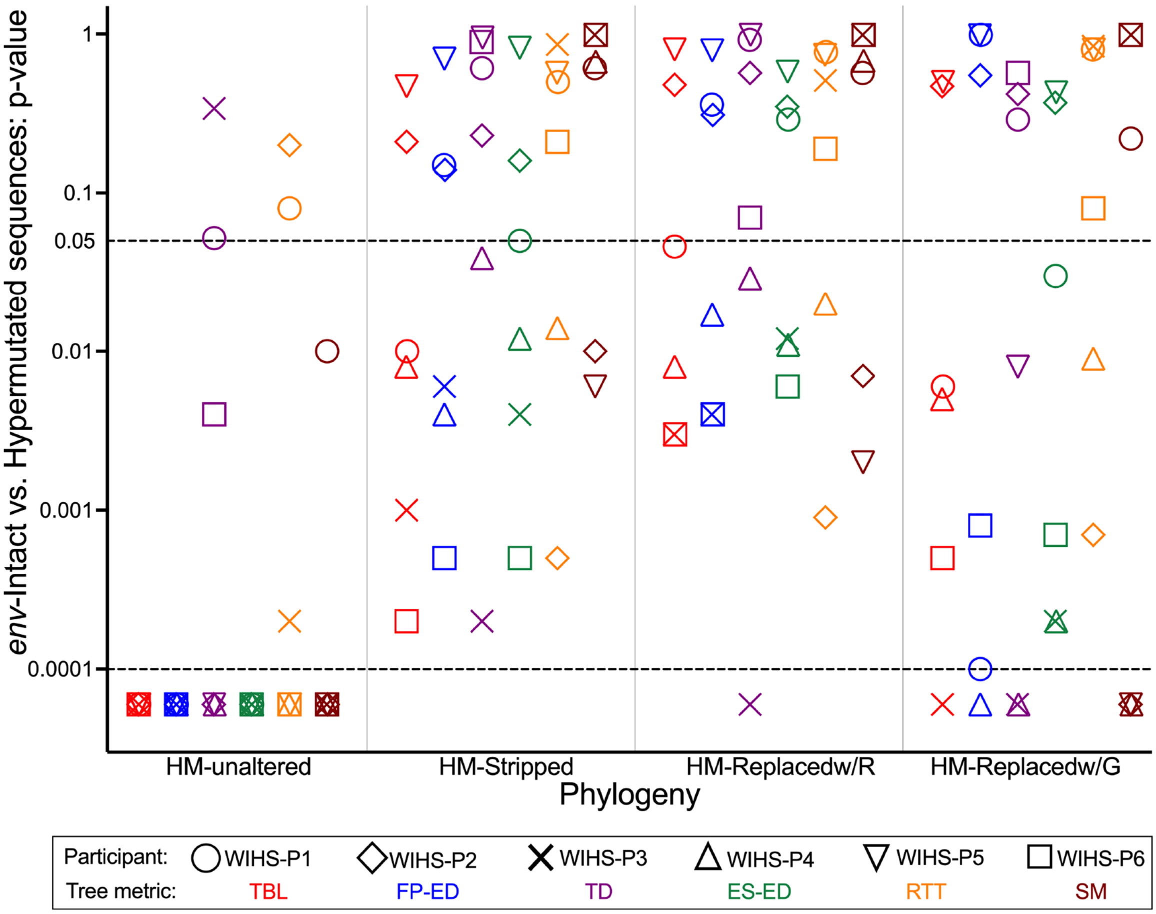

Figure 4. Summary of tree metrics across all participants. For each participant (each shown with a distinct symbol), the *P*-value derived from comparing *env*-intact and hypermutated proviruses in the tree for each of the phylogenetic metrics (each shown in a distinct color) is plotted for each tree type. For consistency with the other metrics, the SM estimated *P*-values of 0 are shown here as $P < .0001$. The horizontal dashed line at $P = .05$ denotes the standard threshold for statistical significance.

($P = .65$) (Fig. 6c). Thus, for participant WIHS-P2, both methods recovered proviral ages equally well. In contrast, yet not surprisingly, the phylogeny inferred from the unaltered alignment produced hypermutated provirus integration dates that were poorly concordant with those from the corrected trees (HM-Stripped $\rho_c = 0.46$; HM-Replacedw/R $\rho_c = 0.45$; Fig. 6d). This illustrates the pitfalls of inferring evolutionary information from the former tree type.

We obtained similar results for WIHS-P4. Again, the integration dates of *env*-intact proviruses inferred from both corrected trees were highly concordant with those inferred from the benchmark tree (both $\rho_c = 0.98$; Fig. 7a), indicating that the corrected trees are appropriate for molecular dating. Moreover, proviral integration dates inferred from the corrected trees were highly concordant with one another ($\rho_c = 0.97$–0.98) (Fig. 7b) and showed no bias between methods ($P = .25$) (Fig. 7c). By contrast, the phylogeny inferred from the unaltered alignment produced hypermutated provirus integration dates that were highly discordant with those inferred from the corrected trees (both $\rho_c = 0.08$; Fig. 7d), again illustrating the pitfalls of inferring evolutionary information from the former tree type.

WIHS-P1, WIHS-P3, WIHS-P5, and WIHS-P6's corrected trees similarly produced *env*-intact proviral integration dates that were strongly concordant with those inferred from their benchmark trees (ρ_c: 0.81–0.93), and overall proviral integration dates that were generally highly concordant with one another, with no bias between methods (Supplementary Figs S16–S19). Again, the phylogenies inferred from their unaltered alignments produced hypermutated provirus integration dates that were generally poorly concordant with those inferred from the corrected trees.

Together, these observations demonstrate that masking hypermutation within alignments is possible and yields phylogenies that can be used to infer the integration dates of both hypermutated and *env*-intact proviruses.

Longevity and dynamics of hypermutated proviruses persisting on ART

Having demonstrated that proviral integration dates can be inferred from the corrected trees, we compared the integration dates of *env*-intact and hypermutated proviruses persisting on ART. Again, we begin with participant WIHS-P2. Both of this participant's corrected trees indicated that the hypermutated proviruses, like the *env*-intact ones, spanned essentially the entire duration of untreated infection, with the earliest dating to early 2004, approximately 1 year after seroconversion, (Fig. 8a and b). On average, however, hypermutated proviruses were older than *env*-intact ones in this participant (both trees $P = .001$; Fig. 8a and b). Longitudinal analysis further revealed that, while integration date distributions of *env*-intact proviruses remained stable during the first 7 years of ART (both trees $P \geq .1$; Fig. 8c and d), hypermutated

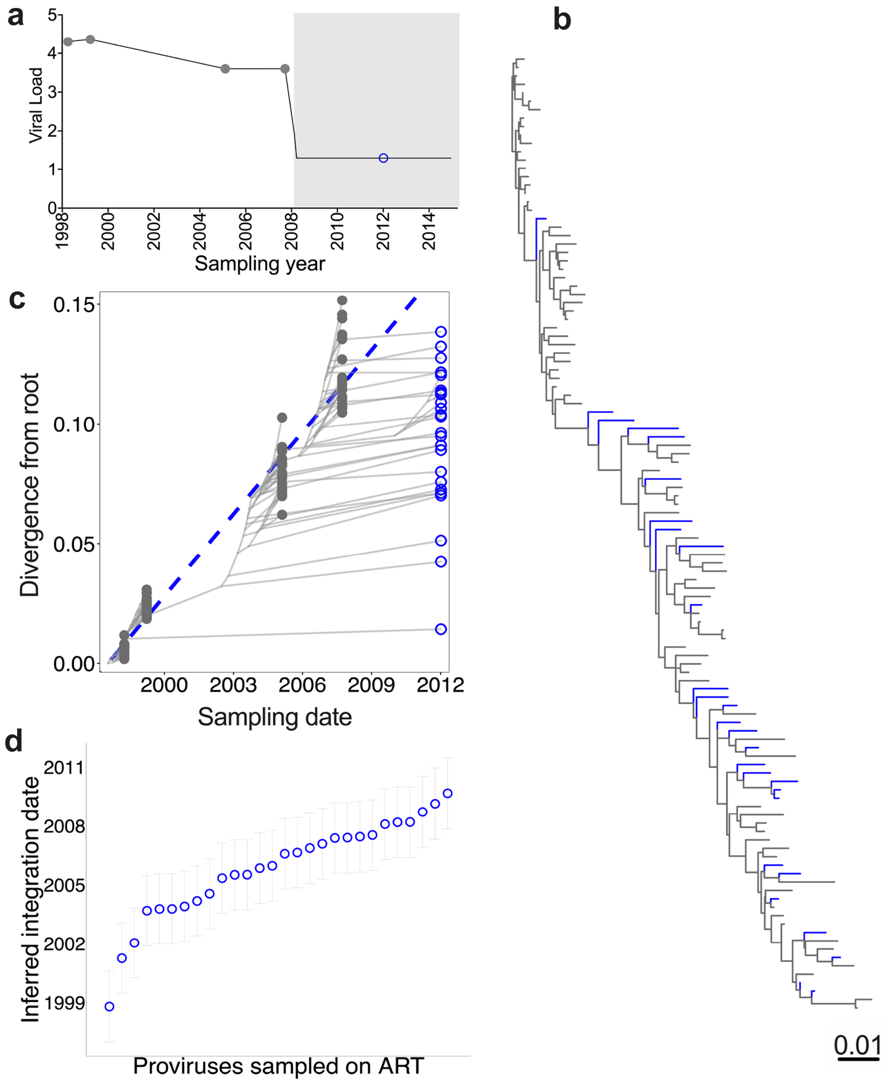

Figure 5. Within-host phylogenetic approach to infer proviral integration dates. (a) Viral load and sampling timeline for a hypothetical participant. Closed gray circles denote plasma HIV RNA sampling prior to ART, while the open blue circle denotes proviral HIV DNA sampling during ART. Light gray shading represents ART. (b) Rooted, ML within-host phylogeny, with branches colored by sequence type (gray = pre-ART plasma HIV RNA; blue = on-ART proviruses). (c) HIV sequence divergence from the root over time. The blue dashed diagonal represents the linear model relating the RTT distances of distinct pre-ART plasma HIV RNA sequences (closed gray circles) to their sampling dates. This model is used to convert the RTT distances of proviral sequences sampled during ART (open blue circles) to their integration dates. Faint gray lines trace the ancestral relationships between HIV sequences. (d) Integration date point estimates (and 95% confidence intervals) for each distinct provirus sequence sampled during ART, sorted from oldest to youngest. The provirus shown at the bottom right of (c), for example, was estimated to have integrated in October 1998 and is shown at the bottom left of (d).

proviruses gradually shifted toward earlier integration dates over time (both trees $P < .02$; Fig. 8e and f). This was presumably because proviruses with more recent integration dates were preferentially eliminated during long-term ART.

WIHS-P4's proviruses also spanned essentially the entire duration of untreated infection (Fig. 8g and h). In contrast to WIHS-P2, however, the integration dates of WIHS-P4's hypermutated proviruses were on average more recent than their *env*-intact ones (both trees $P \leq .02$; Fig. 8g and h). As previously reported (Shahid et al. 2024), WIHS P4's *env*-intact proviruses gradually shifted toward earlier integration dates over time on ART (both trees $P \leq .003$; Fig. 8i and j), likely because those with more recent integration dates decayed more rapidly following ART initiation. In contrast, hypermutated provirus integration date distributions remained stable during ART (both trees $P > .1$; Fig. 8k and l).

WIHS-P1, WIHS-P3, WIHS-P5, and WIHS-P6's hypermutated proviruses also spanned broad age ranges, but in contrast to WIHS-P2 and WIHS-P4, they did not differ from *env*-intact ones in terms of their overall integration date distributions (Supplementary Figs S20 and S21). As reported previously, their *env*-intact proviral integration date distributions remained stable except for participant WIHS-P5 in whom the proviral pool shifted slightly toward later integration dates over time (Supplementary Figs S21c and d) (Shahid et al. 2024). Hypermutated proviral integration date distributions were also stable over time except in WIHS-P1, whose proviral date distributions differed markedly by visit (Supplementary Figs S20e and f). Although this could suggest dynamic changes over time, limited sampling must be acknowledged. Notably, the HM-Stripped and HM-Replacedw/R approaches produced comparable results except in the temporal analysis of *env*-intact

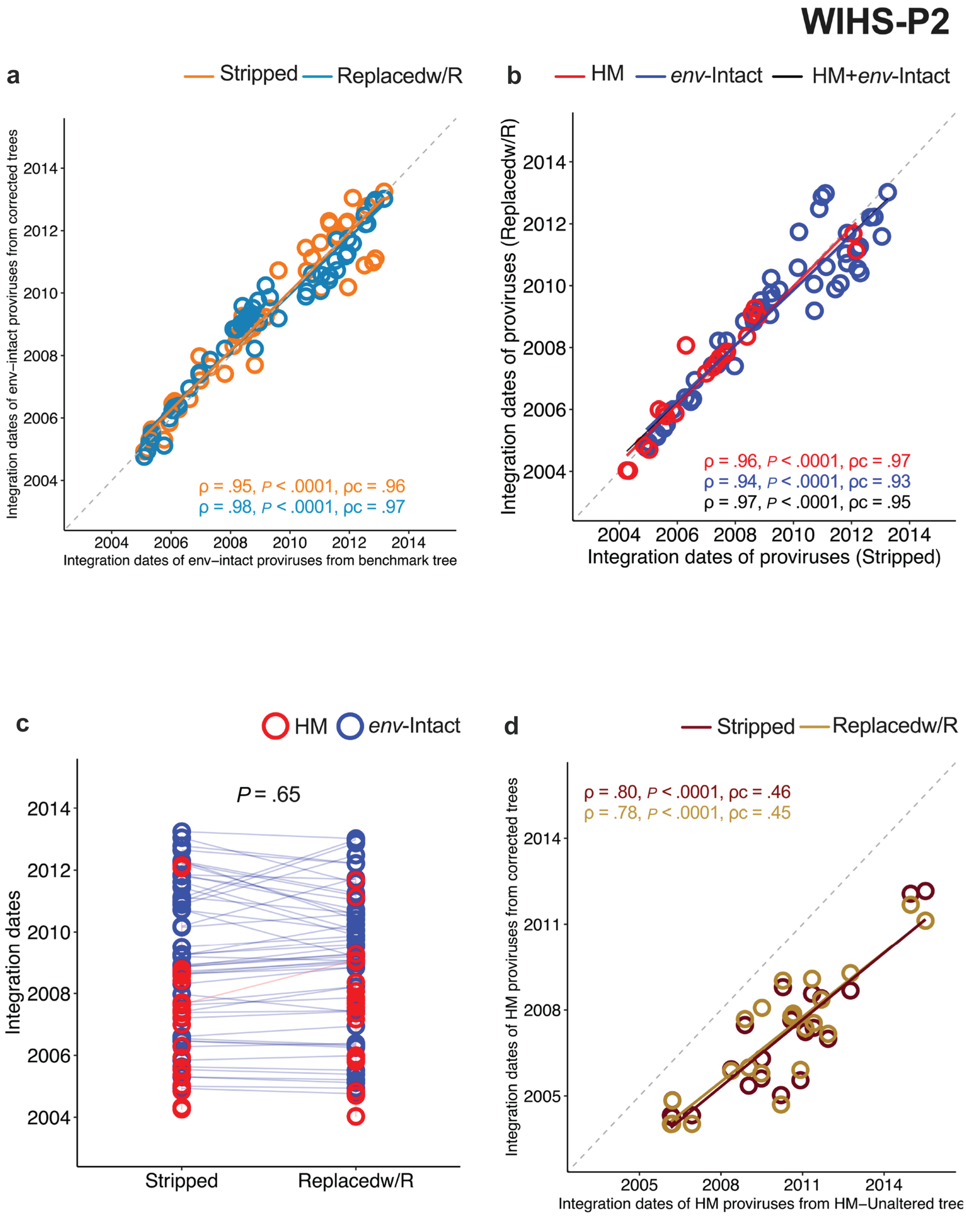

Figure 6. Inferring proviral integration dates from corrected trees: validation using WIHS-P2's data. (a) Correlation between inferred integration dates of *env*-intact proviruses from the benchmark versus corrected trees, where the dates inferred from the HM-Stripped tree are in orange and those inferred from the HM-Replacedw/R tree are in teal. Spearman's ρ, associated P-value, and Lin's concordance correlation coefficient (ρ_c) are shown for each comparison. Regression lines in matching colors are also provided to help visualize these relationships. The dotted diagonal denotes a hypothetical perfect concordance. (b) Correlation between inferred integration dates of all proviruses from HM-Stripped versus HM-Replacedw/R trees, with hypermutated proviruses in red and *env*-intact proviruses in blue. Statistics are computed for all proviruses (black), hypermutated proviruses only (red), and *env*-intact proviruses only (blue). (c) Inferred integration dates of *env*-intact and hypermutated proviruses from HM-Stripped and HM-Replacedw/R trees, presented as paired measurements connected with matching-colored lines. P-value computed using the Wilcoxon matched-pairs signed-rank test. (d) Correlation between inferred integration dates of hypermutated proviruses from the HM-Unaltered and corrected trees (HM-Stripped tree = maroon; HM-Replacedw/R tree = gold).

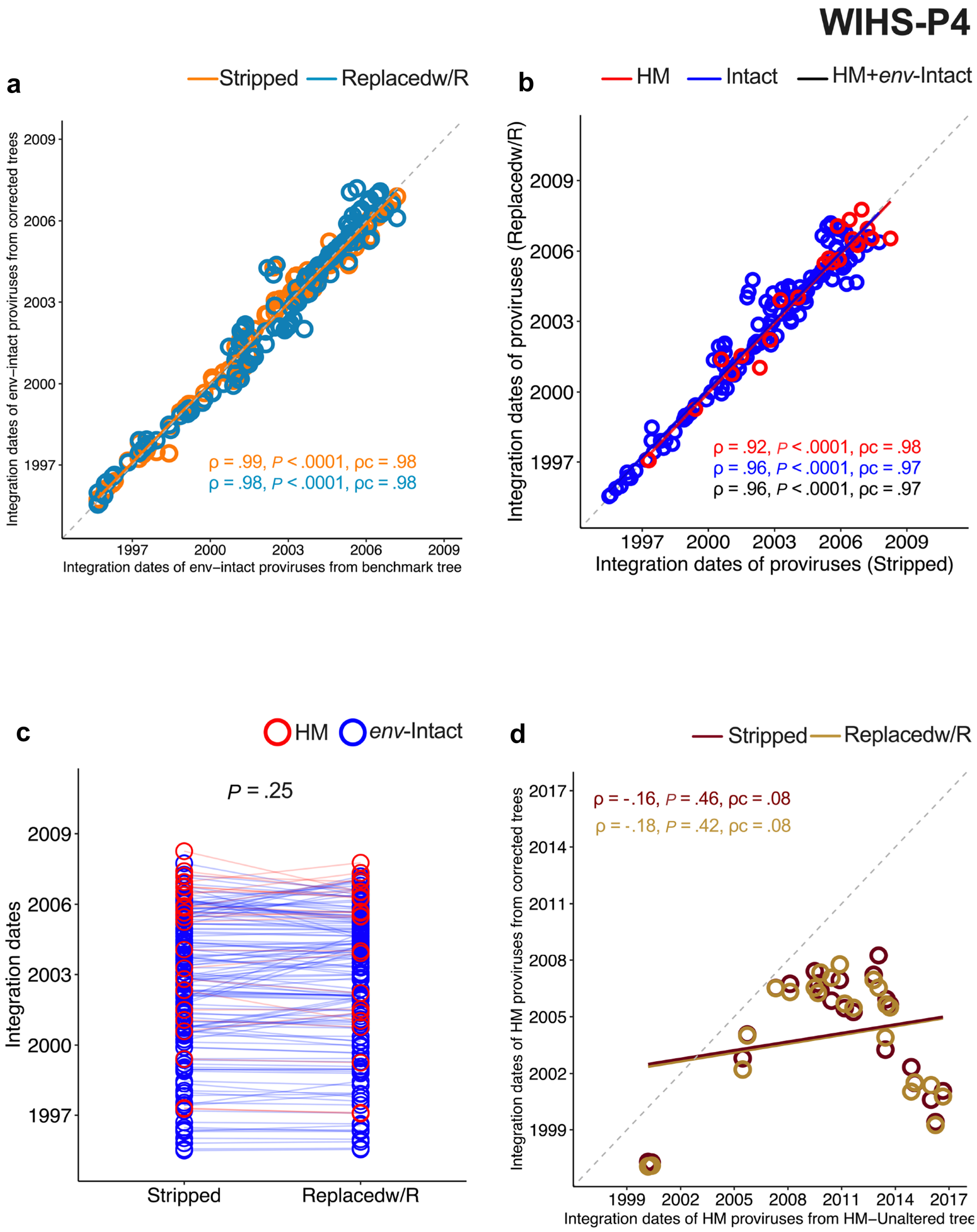

Figure 7. Inferring proviral integration dates from corrected trees: validation using WIHS-P4's data. Legend as in Fig. 6, except for participant WIHS-P4.

proviruses for WIHS-P3, where HM-Stripped suggested a modest shift toward more recent integration dates over time, whereas HM-Replacedw/R indicated no change (Supplementary Figs S20i and j).

Discussion

Although hypermutated proviruses persist in all people living with HIV (Ho et al. 2013, Bruner et al. 2016, Kinloch et al. 2023), we know relatively little about their within-host origins because their extensive mutations complicate phylogenetic inference. We explored three simple approaches to mask hypermutation in nucleotide alignments, with the dual goals of (i) reconstructing phylogenies that accurately reconstruct the within-host evolutionary histories of hypermutated sequences and (ii) applying molecular dating approaches to these trees to gain insights into hypermutated provirus within-host origins and dynamics.

Of the approaches we evaluated, stripping nucleotide positions containing putative APOBEC3 mutations from the alignment, or replacing individual APOBEC3 mutations in hypermutated sequences with R, consistently normalized tree topologies and metrics. By contrast, replacing APOBEC3 mutations in hypermutated sequences with G failed to consistently resolve their erroneous clustering in the tree. We speculate that this is because G replacement is an overcorrection, as not all A bases at target sites are necessarily due to recent APOBEC3 activity. The HIV genome is naturally high in A bases (Kypr and Mrazek 1987, Kypr et al. 1989), so some of the A bases at APOBEC target sites are likely inherited, not due to recent APOBEC3 activity. Given this, G replacement likely obscures some legitimate ancestral information, while also making hypermutated sequences appear to share more G bases than they really do, leaving these sequences at continued risk of long-branch attraction in some cases. By contrast, replacing putative APOBEC3 mutations with R mitigates this risk by acknowledging this ambiguity. We therefore advise against the replacement of APOBEC3 mutations in hypermutated sequences with G.

Importantly, the integration dates of *env*-intact proviruses inferred from the HM-Stripped and HM-Replacedw/R approaches were highly concordant with those inferred from benchmark trees that excluded hypermutated sequences entirely, as is the current practice. The demonstration that these corrected trees provide valid molecular dating results is important because it provides a way to study the within-host evolutionary origins and dynamics of the genetically diverse population of hypermutated proviruses that persist during ART.

Proviral integration date estimates produced by the two approaches were highly concordant, and there was no clear difference in their performance. While the *P*-values derived from comparing the tree-based metrics of *env*-intact and hypermutated sequences, shown in Fig. 4, are overall slightly higher for the HM-Replacedw/R compared to the HM-Stripped approach, we caution against interpreting this to mean that the former is superior. Although we applied statistical tests to guide interpretation, the main goal was to produce tree metric values for hypermutated and *env*-intact sequences that were comparable in range. Both HM-Stripped and HM-Replacedw/R approaches achieved this. We did not necessarily expect that *env*-intact and hypermutated sequence metrics would normalize completely (i.e. produce nonsignificant *P*-values) in all cases, because some evolutionary attributes of hypermutated sequences might plausibly differ from *env*-intact ones. As hypermutated sequences do not normally yield descendants for example, their closest neighbors in the tree might be more distant than those for *env*-intact proviruses, simply because of the lower likelihood of sampling a close relative (which, for a hypermutated sequence, could only be an ancestor). Differential evolutionary dynamics between hypermutated and *env*-intact proviruses could also produce differential RTT measurements (and by extension integration date estimates) between groups, a phenomenon that was indeed observed in WIHS-P2 and WIHS-P4.

We therefore offer the following considerations when choosing an approach. Since the HM-Replacedw/R approach retains the full alignment, it should also preserve more phylogenetic signal than the HM-Stripped approach, where an average of 9% of each *env-gp120* alignment was removed. This could be advantageous for HIV regions that are relatively conserved, yet hotspots for APOBEC3 mutation, for example parts of *pol* (Kieffer et al. 2005, Kijak et al. 2008). However, before implementing the Replacedw/R approach, it is essential to verify that the chosen phylogenetic inference package supports ambiguous characters. IQ-TREE 2 v2.1.3, used in the present study, assigns equal likelihood to each component character (Minh et al. 2020), but other packages, such as the approximate ML algorithm FastTree, treat all non-A/C/T/G characters as missing data (Price et al. 2010). We also wish to acknowledge that, while the approaches described herein involve modification of sequence alignments for general downstream phylogenetic analyses, probabilistic inference frameworks have been developed in the Bayesian Evolutionary Analysis for Sampling Trees software (Drummond and Rambaut 2007) to account for different types of DNA damage, including APOBEC3-mediated hypermutation in HIV (Drummond et al. 2012) and post-mortem damage in ancient DNA (Rambaut et al. 2009) for downstream Bayesian time-calibrated phylogenetic analyses.

It is also important to recognize when sequence alignment modifications are warranted. Hypermutated sequences can be incorporated directly into phylogenies if the goal is simply to visualize a complete dataset. Such trees might even be adequate for some limited tree-based inferences, as suggested by our finding that uncorrected trees produced reasonable root dates and evolutionary rates, likely because these calculations only use information from pre-ART plasma HIV RNA sequences. Nevertheless, our finding that these uncorrected trees produce falsely highly supported clades, erroneously reconstruct the ancestry of hypermutated proviruses, and produce inaccurate (and often nonsensical) proviral integration dates underscores why they should not be used to answer questions about the evolutionary history of hypermutated proviruses. For such questions, strategies to mask hypermutation should be used.

Our results also reveal insights into the evolutionary dynamics of defective proviruses. Even though these cannot reseed infection, studying their dynamics is still important because many can still produce viral proteins (Pollack et al. 2017, Imamichi et al. 2020) that cause chronic immune activation (Deeks et al. 2013) and likely contribute to T-cell exhaustion during ART (Hatano et al. 2013), which in turn could reduce the efficacy of immune-based reservoir elimination strategies (Pollack et al. 2017). Indeed, defective proviruses decay much more slowly than intact ones (Pinzone et al. 2019, Peluso et al. 2020, Gandhi et al. 2021), and differentially with respect to one another, depending on their defect type (Pinzone et al. 2019), presumably because proviruses capable of HIV protein expression have a higher cumulative risk of elimination over time (Imamichi et al. 2020). Our study reveals that, like *env*-intact ones, hypermutated proviruses persisting during ART spanned a very wide age range that largely recapitulates the within-host evolution of HIV prior to ART. From WIHS-P2, for example, we isolated hypermutated proviruses that had integrated as early as a year following seroconversion. This indicates that hypermutated proviruses, like other provirus types, begin to be seeded into the proviral pool essentially immediately following transmission, and can persist for decades thereafter.

Our results also revealed evidence of differential evolutionary dynamics of hypermutated and *env*-intact proviruses in two of the six participants studied, namely WIHS-P2, whose hypermutated proviruses were on average older than *env*-intact ones, and WIHS-P4, in whom the opposite was observed. This suggests that the decay rates of different types of proviruses can be heterogeneous within a given host, as well as heterogeneous between hosts, adding further complexity to the challenge of HIV reservoir elimination.

Our study has some limitations. We analyzed the present dataset (Shahid et al. 2024) because it is among the most comprehensive of its type (in terms of sequence N, follow-up time, and sampling near seroconversion) and because *env-gp120* is commonly used for within-host HIV evolutionary studies (Dapp et al. 2017, Brooks et al. 2020). That said, participants WIHS-P3 and

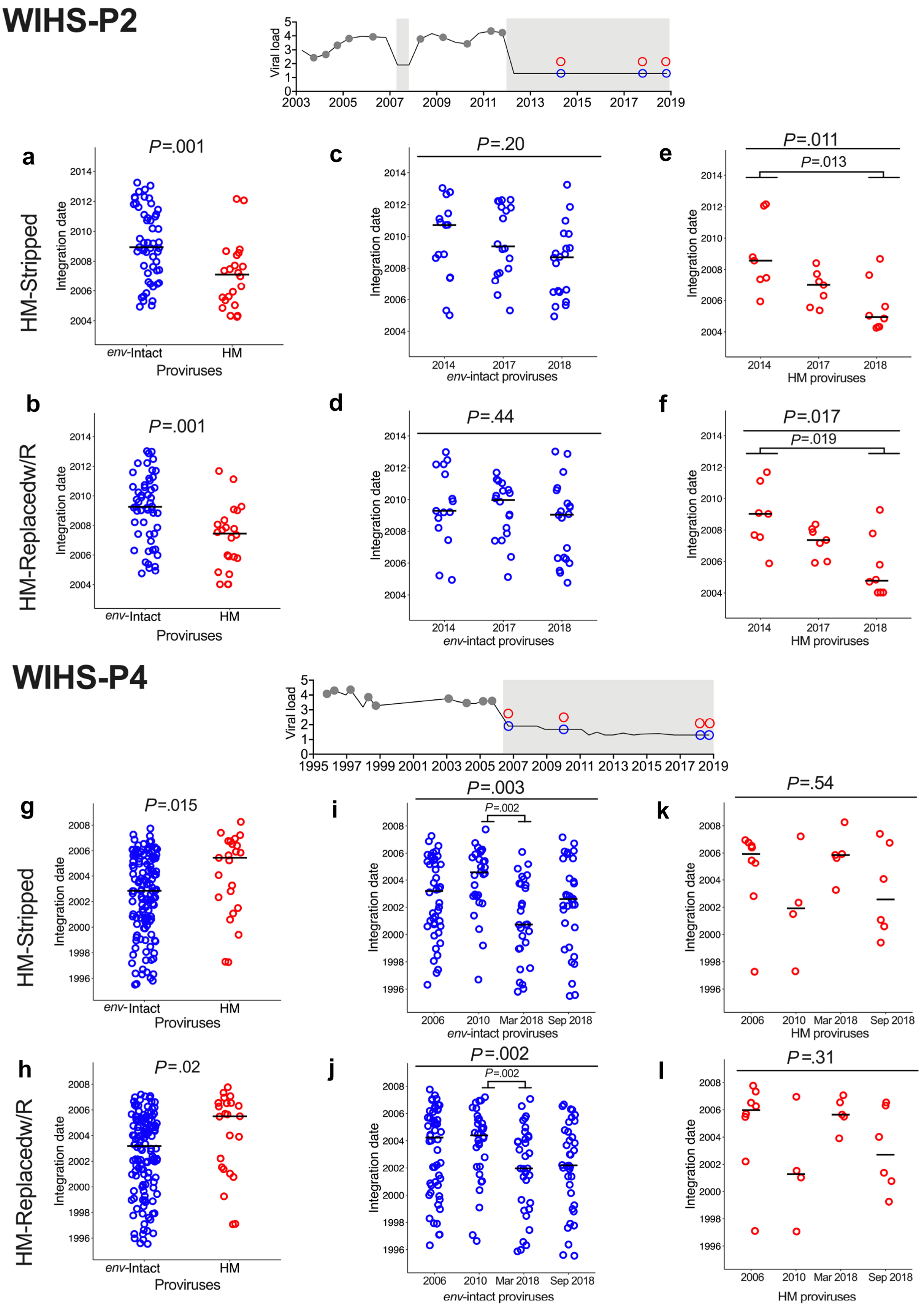

Figure 8. WIHS-P2 and P4: integration dates of *env*-intact and hypermutated proviruses persisting during ART. Top: HIV plasma viral load and sampling history for participant WIHS-P2. (a, b) Integration dates of *env*-intact (blue) and hypermutated proviruses (HM; red) inferred from the HM-Stripped (a) and HM-Replacedw/R (b) trees. Here, all proviruses of the same type are grouped together regardless of their sampling date on ART. *P*-value derived from the Mann–Whitney *U* test. Horizontal black lines represent the median values. (c, d) These are the same *env*-intact provirus integration dates as shown in (a, b), but now stratified by their sampling date on ART. The *P*-value is from a Kruskal–Wallis test comparing all groups. (e–f) These are the same hypermutated provirus integration dates as shown in (a, b), but now stratified by their sampling date on ART. The large *P*-value at the top is from a Kruskal–Wallis test comparing all groups. The smaller *P*-values below represent the significant pairwise post-tests after correction for multiple comparisons. (g–l) Same as for (a–f), except for participant WIHS-P4.

WIHS-P6 had only modest numbers of hypermutated proviruses, which limited our power to detect differences between these and *env*-intact proviruses in their data. Furthermore, while our proposed method should be applicable to any HIV gene, we did not explicitly investigate this. The identification of hypermutated sequences, on which our method depends, is by definition imperfect, as it relies on a statistical cut-off and can be subtly influenced by the choice of reference sequence, particularly if a heterologous sequence (e.g. HXB2 HIV reference strain) is used (Rose and Korber 2000). As recommended, we used the most frequent sequence observed post-seroconversion as the reference (Rose and Korber 2000), although we verified that use of a different sequence impacted the identification of hypermutated sequences minimally or not at all (e.g. using an arbitrarily chosen reference sequence from WIHS-P2's earliest time point yielded 137 out of 1515 nucleotide positions with putative APOBEC3 mutations, versus the original 140). Finally, we cannot assume that intact *env-gp120* sequences come from fully intact HIV genomes. As such, the comparison group for hypermutated sequences in the present study is not the replication-competent HIV reservoir, but rather the pool of proviruses with intact *env-gp120* sequences, many of which will have defects elsewhere.

In summary, the current practice of excluding hypermutated proviruses from phylogenies used for hypothesis testing has limited our understanding of the *in vivo* evolutionary origins and longevity of these sequences. Here, we validated two simple nucleotide alignment modification approaches that allow hypermutated sequences to be correctly incorporated into phylogenies that can be used for molecular dating. Our observations reveal that hypermutated proviruses, like other provirus types, are archived throughout untreated infection and can persist for years on ART. Our results further suggest that the evolutionary dynamics of hypermutated proviruses may differ from those of other proviral types in some individuals. In addition to enriching our understanding of HIV persistence toward the ultimate goal of HIV cure, the approaches developed here could be extended to between-host phylogenies, and testing of other hypotheses related to within-host evolutionary origins of hypermutated sequences.

Materials and methods

Study participants and within-host HIV sequence datasets

We analyzed longitudinal, single-genome-amplified HIV *env-gp120* sequence datasets previously collected from six WIHS participants with documented HIV seroconversion (Shahid et al. 2024). WIHS is a multi-center cohort of women living with (or without) HIV in the United States (Barkan et al. 1998, Bacon et al. 2005, Adimora et al. 2018) that has now merged into the MACS/WIHS Combined Cohort Study (MWCCS) (D'Souza et al. 2021). Each participant's longitudinal dataset comprised plasma HIV RNA *env-gp120* sequences collected between seroconversion and ART initiation, along with *env-gp120* proviral sequences sampled during ART (Shahid et al. 2024) (Table 1). All sequences were collected by single-genome amplification, where those with nucleotide mixtures, defects (e.g. deletions causing frameshifts), or evidence of within-host recombination (identified using RDP4 v4.101; Martin et al. 2015) were excluded (Shahid et al. 2024). Sequences that were 100% identical in *env-gp120* were collapsed to a single representative sequence prior to phylogenetic inference. Within-host datasets comprised a median of 242 (interquartile range 119–337) distinct sequences per participant.

Ethics statement

Institutional review boards at each WIHS clinical research site approved the study protocol. All participants provided written informed consent. This nested substudy was additionally approved by the institutional review boards at Providence Health Care/University of British Columbia and Simon Fraser University.

Identification of hypermutated sequences and sequence alignment modification

Hypermutated HIV sequences were identified using Hypermut 2.0, available at https://www.hiv.lanl.gov/content/sequence/HYPERMUT/hypermut.html (Rose and Korber 2000). This program takes a nucleotide alignment as input, where the first sequence is used as a reference to which all others are compared. As recommended for within-host datasets (Rose and Korber 2000), we chose the most frequently observed *env-gp120* sequence from the first plasma HIV RNA sampling timepoint as the reference wherever possible. Hypermut defines APOBEC3 target sites as **G**RD; that is, a **G** followed by either A or G (denoted by the IUPAC code R; Cornish-Bowden 1985) and then followed by A, G, or T (denoted by D), where the bold and underlined **G** is the APOBEC3 target site. Non-APOBEC3 target sites are defined as GY (where Y denotes C or T) or GRC. Hypermut identifies all target and nontarget sites within each sequence and categorizes each as mutated (i.e. harboring an A) or not (i.e. harboring a C, G, or T). The program then compares the proportion of mutated target and nontarget sites in each sequence using Fisher's exact test. Sequences enriched in G-to-A mutations at target sites with $P < .05$ are identified as hypermutated.

We then prepared five within-host *env-gp120* sequence alignments for each participant, where the first two were controls and the last three used different strategies to mask hypermutation. Sequence alignments were performed in a codon-aware manner using Multiple Alignment Using Fast Fourier Transform (MAFFT) v7.471 (Katoh and Standley 2013) and manually inspected in AliView v1.26 (Larsson 2014). The first alignment contained all pre-ART *env-gp120* plasma HIV RNA sequences plus only the *env*-intact proviruses sampled during ART (i.e. hypermutated proviruses were excluded, as is the current practice in the field; Jones et al. 2018, 2020, Brooks et al. 2020, Kinloch et al. 2023). We called this the "*env*-intact only" alignment, where the resulting phylogeny was used as the benchmark for provirus molecular dating. The second alignment contained all pre-ART plasma HIV RNA sequences plus all (i.e. both *env*-intact and hypermutated) proviruses sampled during ART. The phylogeny inferred from this "HM-Unaltered" alignment served to illustrate the skewed topologies of resulting trees. The next three alignments were modifications of this second one, where we tested different strategies to mask hypermutation and thereby normalize topology. The first strategy, HM-Stripped, removed all nucleotide positions that harbored an A at an APOBEC3 target site in at least one hypermutated sequence, yielding a shorter overall alignment. The second strategy, HM-Replacedw/R, individually replaced all A bases at APOBEC3 target sites within hypermutated sequences with R (denoting A or G). The third strategy, HM-Replacedw/G, individually replaced all A bases at APOBEC3 target sites within hypermutated sequences with G. Both these strategies preserved the alignment length. Here, replacing with G assumes that all A bases at target sites are the result of APOBEC3 effects, whereas replacing with R recognizes the possibility that some of these A bases may be inherited. Visualizations of the HM-Unaltered, HM-Stripped, and HM-Replacedw/R alignments are provided in Supplementary Fig.

S1. Phylogenies inferred from these alignments were evaluated as described in the next section.

Within-host phylogenetic inference, rooting, and tree metrics

ML phylogenies were inferred from sequence alignments following automated model selection using an Akaike information criterion (AIC) in IQ-TREE 2. Best-fit models are reported in Supplementary Table S3. Branch support values were derived using the ultrafast bootstrap option (1000 bootstraps) (Hoang et al. 2018, Minh et al. 2020). Phylogenies were visualized using the R package *ggtree* (Yu 2020).

Most of our downstream analyses required rooting the tree at the inferred MRCA of the dataset. As previously described, we used a modified RTT regression approach where we explored all positions in the tree to identify the location that maximized the (Pearson's) correlation between the RTT distances of all plasma HIV RNA sequences collected prior to ART initiation and their sampling dates (Jones et al. 2018). This location was set as the tree root, which represents the estimated transmitted/founder virus, or a close descendant thereof, in these datasets.

To evaluate the extent to which the three alignment modification strategies normalized the position of hypermutated proviruses in the tree, we compared *env*-intact and hypermutated proviruses with respect to various tree-based metrics, explained in Fig. 1. We quantified each sequence's terminal branch length (TBL), which is the length of the branch connecting each sequence to the tree, in estimated substitutions per nucleotide site (Fig. 1b). We computed each sequence's root-to-tip (RTT) distance, defined as the total distance between each tip and the tree root (Fig. 1c). We computed two measures of evolutionary distinctiveness: fair proportion evolutionary distinctiveness (FP-ED) and equal splits evolutionary distinctiveness (ES-ED), both of which distribute the RTT distances in a tree among the descendant sequences at the tips (Pavoine et al. 2017). FP-ED does this by dividing the shared evolutionary history represented by an internal branch equally among all its descendant tips, regardless of branching order (Isaac et al. 2007, Redding et al. 2014) (Fig. 1d), whereas ES-ED assigns a longer portion of shared internal branches to immediate descendants (Redding and Mooers 2006) (Fig. 1e). FP-ED and ES-ED were computed using a custom R script with package picante (v1.8.2) (Kembel et al. 2010). We computed each proviral sequence's median topological distance (TD) from all other sequences of the same type (i.e. *env*-intact or hypermutated), where distance was defined as the number of nodes separating each pair (Fig. 1f). We used the Slatkin Maddison (SM) test (Slatkin and Maddison 1989), implemented using the R package "slatkin.maddison" (v0.1.0; https://github.com/prmac/slatkin.maddison) to assess the extent to which *env*-intact and hypermutated proviruses displayed population structure in the tree. This test determines the minimum number of migrations between groups to explain the distribution of groups at the tree tips: the smaller the number, the stronger the support for population structure. Statistical support is based on the number of migrations that would be expected in a randomly structured population, simulated by permuting group labels between tips. Note that SM returns an estimated *P*-value, where a value of 0 can be interpreted as $P < .001$, as 1000 permutations were performed. Finally, as the SM test can sometimes produce statistically significant *P*-values for larger datasets that have only minimal levels of compartmentalization (Council et al. 2020, Sarkar et al. 2023), we also applied the Simmonds Association Index (AI) (Wang et al. 2001), implemented in BaTSv0.9 (Parker et al. 2008). The Simmonds AI assesses the degree of population structure by calculating the composition of descendant sequences in each successive node in the tree and summing these in a weighted manner (where nodes closer to the root receive less weight) to generate an overall association value (Wang et al. 2001). The AI represents the ratio of the mean association value calculated from 100 bootstrap replicates of the data, and the mean of 10 control trees with randomly permuted tip labels, where smaller ratios indicate a higher degree of compartmentalization. Finally, we used Shimodaira–Hasegawa tests, as implemented in PAUP* (*Phylogenetic Analysis Using PAUP) (Swofford 2002), to compare the likelihood of tree topologies inferred under the different hypermutation repair strategies with those inferred using unaltered alignments.

Within-host phylogenetic inference and proviral dating

We inferred the integration dates of *env*-intact and hypermutated proviruses persisting during ART using a published phylogenetic approach (Jones et al. 2018). Using the rooted trees, we fit a linear model relating the RTT distances of pre-ART plasma HIV sequences to their collection dates (i.e. proviral sequences were not considered when determining the root). The slope of this line represents the average within-host *env-gp120* evolutionary rate during untreated HIV infection, and the *x*-intercept represents the inferred root date. Model quality was assessed by comparing the model's AIC to that of a null model with zero slope. To pass quality control, the linear model needed to have an AIC value of at least 10 units lower than the null model ($\Delta AIC \geq 10$) and a root date prior to the first plasma sampling. All phylogenies met these criteria (Supplementary Table S3). We then used the linear model to convert proviral RTT distances to their integration dates. The custom R script for this method is available at https://github.com/cfe-lab/phylodating.

Statistical analysis

Spearman's correlation (ρ) and Lin's concordance correlation coefficient (ρ_c) were calculated in R. All other statistical analyses were performed in Prism, v10.0.2 (GraphPad Software). A threshold of $P < .05$ was used to denote statistical significance.

Acknowledgements

We thank Drs Art F.Y. Poon and Natalie N. Kinloch for helpful discussions.

The authors gratefully acknowledge the contributions of the study participants and the dedication of the staff at the MWCCS sites.

Supplementary data

Supplementary data is available at *VEVOLU Journal* online.

Conflict of interest: None declared.

Funding

Data in this manuscript were collected by the WIHS, now the MWCCS.

The contents of this publication are solely the responsibility of the authors and do not represent the official views of the National Institutes of Health (NIH) or other funders.

MWCCS (Principal Investigators): Atlanta CRS (Ighovwerha Ofotokun, Anandi Sheth, and Gina Wingood), U01-HL146241; Baltimore clinical research site (CRS) (Todd Brown and Joseph

Margolick), U01-HL146201; Bronx CRS (Kathryn Anastos, David Hanna, and Anjali Sharma), U01-HL146204; Brooklyn CRS (Deborah Gustafson and Tracey Wilson), U01-HL146202; Data Analysis and Coordination Center (Gypsyamber D'Souza, Stephen Gange, and Elizabeth Topper), U01-HL146193; Chicago-Cook County CRS (Mardge Cohen, Audrey French, and Ryan Ross), U01-HL146245; Chicago-Northwestern CRS (Steven Wolinsky, Frank Palella, and Valentina Stosor), U01-HL146240; Northern California CRS (Bradley Aouizerat, Jennifer Price, and Phyllis Tien), U01-HL146242; Los Angeles CRS (Roger Detels and Matthew Mimiaga), U01-HL146333; Metropolitan Washington CRS (Seble Kassaye and Daniel Merenstein), U01-HL146205; Miami CRS (Maria Alcaide, Margaret Fischl, and Deborah Jones), U01-HL146203; Pittsburgh CRS (Jeremy Martinson and Charles Rinaldo), U01-HL146208; UAB-MS stands for University of Alabama at Birmingham - University of Mississippi Medical Center CRS (Mirjam-Colette Kempf, James B. Brock, and Deborah Konkle-Parker), U01-HL146192; and UNC CRS (M. Bradley Drummond and Michelle Floris-Moore), U01-HL146194. The MWCCS is funded primarily by the National Heart, Lung, and Blood Institute (NHLBI), with additional cofunding from the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institute on Aging, National Institute of Dental and Craniofacial Research, National Institute of Allergy And Infectious Diseases (NIAID), National Institute of Neurological Disorders and Stroke (NINDS), National Institute of Mental Health (NIMH), National Institute on Drug Abuse (NIDA), National Institute of Nursing Research, National Cancer Institute, National Institute on Alcohol Abuse and Alcoholism, National Institute on Deafness and Other Communication Disorders, National Institute of Diabetes and Digestive and Kidney Diseases, and National Institute on Minority Health and Health Disparities, and in coordination and alignment with the research priorities of the NIH, Office of AIDS Research. MWCCS data collection is also supported by UL1-TR000004 (UCSF CTSA), UL1-TR003098 (JHU ICTR), UL1-TR001881 (UCLA CTSI), P30-AI-050409 (Atlanta CFAR), P30-AI-073961 (Miami CFAR), P30-AI-050410 (UNC CFAR), P30-AI-027767 (UAB CFAR), P30-MH-116867 (Miami CHARM), UL1-TR001409 (DC CTSA), KL2-TR001432 (DC CTSA), and TL1-TR001431 (DC CTSA).

In addition, this work was supported by the Canadian Institutes of Health Research (CIHR) through a project grant (PJT-159625 to Z.L.B. and J.B.J.) and a focused team grant (HB1-164063 to Z.L.B.). This work was also supported by the Martin Delaney "REACH" Collaboratory (NIH grant no. 1-UM1AI164565-01 to Z.L.B.), which is supported by the following NIH cofunding institutes: NIMH, NIDA, NINDS, NIDDK, NHLBI, and NIAID. This work was also supported by the Einstein-Rockefeller-CUNY Center for AIDS Research (NIH grant no. P30AI124414 to H.G.). A.S., B.R.J., and M.C.D. were supported by CIHR Doctoral Research Awards; M.C.D. was supported by a CIHR Canada Graduate Scholarship—Master's award. S.M. was supported by Simon Fraser University Undergraduate Student Research Award. Z.L.B. was supported by a Scholar Award from Michael Smith Health Research BC. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability

The nucleotide sequences reported in this paper are available in GenBank (proviral DNA accession numbers: OR404056–OR404777 and OR404820–OR404981; HIV RNA accession numbers: OR403057–OR403738).

References

Adimora AA, Ramirez C, Benning L *et al*. Cohort profile: the Women's Interagency HIV Study (WIHS). *Int J Epidemiol* 2018;**47**:393–94i.

Armitage AE, Deforche K, Chang C-H *et al*. APOBEC3G-induced hypermutation of human immunodeficiency virus type-1 is typically a discrete "all or nothing" phenomenon. *PLoS Genet* 2012;**8**:e1002550.

Bacon MC, von Wyl V, Alden C *et al*. The Women's Interagency HIV Study: an observational cohort brings clinical sciences to the bench. *Clin Diagn Lab Immunol* 2005;**12**:1013–19.

Barkan SE, Melnick SL, Preston-Martin S *et al*. The Women's Interagency HIV Study. WIHS Collaborative Study Group. *Epidemiology* 1998;**9**:117–25.

Bergsten J. A review of long-branch attraction. *Cladistics* 2005;**21**:163–93.

Bozzi G, Simonetti FR, Watters SA *et al*. No evidence of ongoing HIV replication or compartmentalization in tissues during combination antiretroviral therapy: implications for HIV eradication. *Sci Adv* 2019;**5**:eaav2045.

Brodin J, Zanini F, Thebo L *et al*. Establishment and stability of the latent HIV-1 DNA reservoir. *eLife* 2016;**5**:e18889.

Brooks K, Jones BR, Dilernia DA *et al*. HIV-1 variants are archived throughout infection and persist in the reservoir. *PLoS Pathog* 2020;**16**:e1008378.

Bruner KM, Murray AJ, Pollack RA *et al*. Defective proviruses rapidly accumulate during acute HIV-1 infection. *Nat Med* 2016;**22**:1043–49.

Cornish-Bowden A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res* 1985;**13**:3021–30.

Council OD, Zhou S, McCann CD *et al*. Deep sequencing reveals compartmentalized HIV-1 in the semen of men with and without sexually transmitted infection-associated urethritis. *J Virol* 2020;**94**:10–128.

Dapp MJ, Kober KM, Chen L *et al*. Patterns and rates of viral evolution in HIV-1 subtype B infected females and males. *PLoS One* 2017;**12**:e0182443.

Deeks SG, Lewin SR, Havlir DV. The end of AIDS: HIV infection as a chronic disease. *Lancet* 2013;**382**:1525–33.

Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 2007;**7**:214.

Drummond AJ, Suchard MA, Xie D *et al*. Bayesian phylogenetics with BEAUti and the BEAST 1.7′. *Mol Biol Evol* 2012;**29**:1969–73.

D'Souza G, Bhondoekhan F, Benning L *et al*. Characteristics of the MACS/WIHS combined cohort study: opportunities for research on aging with HIV in the longest US observational study of HIV. *Am J Epidemiol* 2021;**190**:1457–75.

Finzi D, Blankson J, Siliciano JD *et al*. Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nat Med* 1999;**5**: 512–17.

Finzi D, Hermankova M, Pierson T *et al*. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* 1997;**278**:1295–300.

Fitzgibbon JE, Mazar S, Dubin DT. A new type of G—>A hypermutation affecting human immunodeficiency virus. *AIDS Res Hum Retroviruses* 1993;**9**:833–38.

Gandhi RT, Cyktor JC, Bosch RJ *et al*. Selective decay of Intact HIV-1 proviral DNA on antiretroviral therapy. *J Infect Dis* 2021;**223**:225–33.

Gantner P, Buranapraditkun S, Pagliuzza A *et al.* HIV rapidly targets a diverse pool of CD4(+) T cells to establish productive and latent infections. *Immunity* 2023;**56**:653–68e5.

Goodenow M, Huet T, Saurin W *et al.* HIV-1 isolates are rapidly evolving quasispecies: evidence for viral mixtures and preferred nucleotide substitutions. *J Acquir Immune Defic Syndr (1988)* 1989;**2**:344–52.

Gorbalenya AE. Phylogeny of viruses. *Ref Mod Biomed Sci* 2017, **4**:125–29.

Halvas EK, Joseph KW, Brandt LD *et al.* HIV-1 viremia not suppressible by antiretroviral therapy can originate from large T cell clones producing infectious virus. *J Clin Invest* 2020;**130**:5847–57.

Harris RS, Liddament MT. Retroviral restriction by APOBEC proteins. *Nat Rev Immunol* 2004;**4**:868–77.

Hatano H, Jain V, Hunt PW *et al.* Cell-based measures of viral persistence are associated with immune activation and programmed cell death protein 1 (PD-1)-expressing CD4+ T cells. *J Infect Dis* 2013;**208**:50–56.

Hiener B, Horsburgh BA, Eden J-S *et al.* Identification of genetically intact HIV-1 proviruses in specific CD4(+) T cells from effectively treated participants. *Cell Rep* 2017;**21**:813–22.

Ho YC, Shan L, Hosmane N *et al.* Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. *Cell* 2013;**155**:540–51.

Hoang DT, Chernomor O, von Haeseler A *et al.* UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol* 2018;**35**: 518–22.

Imamichi H, Dewar RL, Adelsberger JW *et al.* 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* 2016;**113**:8783–88.

Imamichi H, Smith M, Adelsberger JW *et al.* Defective HIV-1 proviruses produce viral proteins. *Proc Natl Acad Sci U S A* 2020;**117**:3704–10.

Isaac NJ, Turvey ST, Collen B *et al.* Mammals on the EDGE: conservation priorities based on threat and phylogeny. *PLoS One* 2007;**2**:e296.

Jones BR, Joy JB, Thorne J. Inferring human immunodeficiency virus 1 proviral integration dates with Bayesian inference. *Mol Biol Evol* 2023;**40**:msad15.

Jones BR, Kinloch NN, Horacsek J *et al.* Phylogenetic approach to recover integration dates of latent HIV sequences within-host. *Proc Natl Acad Sci U S A* 2018;**115**:E8958–E67.

Jones BR, Miller RL, Kinloch NN *et al.* Genetic diversity, compartmentalization, and age of HIV proviruses persisting in CD4 + T cell subsets during long-term combination antiretroviral therapy. *J Virol* 2020;**94**.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013;**30**:772–80.

Kearney MF, Wiegand A, Shao W *et al.* Origin of rebound plasma HIV includes cells with identical proviruses that are transcriptionally active before stopping of antiretroviral therapy. *J Virol* 2016;**90**:1369–76.

Kembel SW, Cowan PD, Helmus MR *et al.* Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* 2010;**26**: 1463–64.

Kieffer TL, Kwon P, Nettles RE *et al.* G→A hypermutation in protease and reverse transcriptase regions of human immunodeficiency virus type 1 residing in resting CD4 + T Cells In Vivo. *J Virol* 2005;**79**:1975–80.

Kijak GH, Janini LM, Tovanabutra S *et al.* Variable contexts and levels of hypermutation in HIV-1 proviral genomes recovered from primary peripheral blood mononuclear cells. *Virology* 2008;**376**:101–11.

Kinloch NN, Shahid A, Dong W *et al.* HIV reservoirs are dominated by genetically younger and clonally enriched proviruses. *mBio* 2023;**14**:e0241723.

Kypr J, Mrazek J. Unusual codon usage of HIV. *Nature* 1987;**327**:20.

Kypr J, Mrazek J, Reich J. Nucleotide composition bias and CpG dinucleotide content in the genomes of HIV and HTLV 1/2. *Biochim Biophys Acta* 1989;**1009**:280–82.

Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 2014;**30**:3276–78.

Lee GQ, Orlova-Fink N, Einkauf K *et al.* Clonal expansion of genome-intact HIV-1 in functionally polarized Th1 CD4+ T cells. *J Clin Invest* 2017;**127**:2689–96.

Martin DP, Murrell B, Golden M *et al.* RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evol* 2015;**1**:vev003.

Minh BQ, Schmidt HA, Chernomor O *et al.* IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* 2020;**37**:1530–34.

Nicolas A, Migraine J, Dutrieux J *et al.* Genotypic and phenotypic diversity of the replication-competent HIV reservoir in treated patients. *Microbiol Spectr* 2022;**10**:e0078422.

Pankau MD, Reeves DB, Harkins E *et al.* Dynamics of HIV DNA reservoir seeding in a cohort of superinfected Kenyan women. *PLoS Pathog* 2020;**16**:e1008286.

Parker J, Rambaut A, Pybus OG. Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect Genet Evol* 2008;**8**:239–46.

Patro SC, Brandt LD, Bale MJ *et al.* Combined HIV-1 sequence and integration site analysis informs viral dynamics and allows reconstruction of replicating viral ancestors. *Proc Natl Acad Sci U S A* 2019;**116**:25891–99.

Pavoine S, Bonsall MB, Dupaix A *et al.* From phylogenetic to functional originality: guide through indices and new developments. *Ecol Indic* 2017;**82**:196–205.

Peluso MJ, Bacchetti P, Ritter KD *et al.* Differential decay of intact and defective proviral DNA in HIV-1-infected individuals on suppressive antiretroviral therapy. *JCI Insight* 2020;**5**:e13299.

Pinzone MR, VanBelzen DJ, Weissman S *et al.* Longitudinal HIV sequencing reveals reservoir expression leading to decay which is obscured by clonal expansion. *Nat Commun* 2019;**10**:728.

Pollack RA, Jones RB, Pertea M *et al.* Defective HIV-1 proviruses are expressed and can Be recognized by cytotoxic T lymphocytes, which shape the proviral landscape. *Cell Host Microbe* 2017;**21**:494–506e4.

Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;**5**:e9490.

Rambaut A, Ho SYW, Drummond AJ *et al.* Accommodating the effect of ancient DNA damage on inferences of demographic histories. *Mol Biol Evol* 2009;**26**:245–48.

Redding DW, Mazel F, Mooers A. Measuring evolutionary isolation for conservation. *PLOS ONE* 2014;**9**:e113490.

Redding DW, Mooers AØ. Incorporating evolutionary measures into conservation prioritization. *Conserv Biol* 2006;**20**:1670–78.

Rose PP, Korber BT. Detecting hypermutations in viral sequences with an emphasis on G—> A hypermutation. *Bioinformatics* 2000;**16**:400–01.

Sanchez G, Xu X, Chermann JC *et al.* Accumulation of defective viral genomes in peripheral blood mononuclear cells of human immunodeficiency virus type 1-infected individuals. *J Virol* 1997;**71**:2233–40.

Sarkar S, Romero-Severson E, Leitner T. Migration coupled with recombination explains disparate HIV-1 anatomical compartmentalization signals. *bioRxiv* 2023. 2023.04.22.537949.

Shahid A, MacLennan S, Jones BR *et al.* The replication-competent HIV reservoir is a genetically restricted, younger subset of the overall pool of HIV proviruses persisting during therapy, which is highly genetically stable over time. *J Virol* 2024;**98**:e0165523.

Sheehy AM, Gaddis NC, Choi JD *et al.* Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* 2002;**418**: 646–50.

Slatkin M, Maddison WP. A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics* 1989;**123**:603–13.

Swofford DL. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). *Sinauer Associates* Sunderland, Massachusetts, 2002.

Vartanian JP, Meyerhans A, Asjö B *et al.* Selection, recombination, and G—-A hypermutation of human immunodeficiency virus type 1 genomes. *J Virol* 1991;**65**:1779–88.

Waldron D. Hypermutation of HIV-1 in vivo. *Nat Rev Genet* 2015;**16**:626–26.

Wang TH, Donaldson YK, Brettle RP *et al.* Identification of shared populations of human immunodeficiency virus type 1 infecting microglia and tissue macrophages outside the central nervous system. *J Virol* 2001;**75**:11686–99.

Whitney JB, Hill AL, Sanisetty S *et al.* Rapid seeding of the viral reservoir prior to SIV viraemia in rhesus monkeys. *Nature* 2014;**512**:74–77.

Yu G. Using ggtree to visualize data on tree-like structures. *Curr Protoc Bioinform* 2020;**69**:e96.

Virus Evolution, 2025, 11(1), veae094 , DOI: https://doi.org/10.1093/ve/veae094, Advance Access Publication Date: 11 November 2024, Research Article