Supporting Information

Lorenzi et al. 10.1073/pnas.1617789113

SI Methods

We applied Bayesian inference methods to estimate jointly the frequency of integrated proviruses and their reactivation probability, an approach enabled by the combined proviral sequencing and viral outgrowth assay data. We model the probability of measuring a certain number k_{DNA} of integrated proviral sequences as binomial,

$$k_{DNA} \sim \text{Binomial}(W_{DNA}, p_{DNA}),$$
$$p_{DNA} = 1-(1-p)^{C_{DNA}} \approx 1-e^{-C_{DNA}p}. \quad \textbf{[S1]}$$

Here, W_{DNA} is the number of wells tested, C_{DNA} is the number of cells per well, and p is the frequency of CD4^{+} T cells with integrated provirus. Similarly, the probability of observing k_{VOA}-positive wells in the viral outgrowth assay is also binomial,

$$k_{RNA} \sim \text{Binomial}(W_{VOA}, p_{VOA}), \quad p_{VOA} \approx 1-e^{-C_{VOA}pr}. \quad \textbf{[S2]}$$

As above, W_{VOA} and C_{VOA} are the number of wells and number of cells tested per well in the viral outgrowth assay, respectively. Here, r is the probability that an integrated provirus reactivates and grows out in the VOA.

We assumed the following prior distributions:

$$\log_{10}(p) \sim \text{Normal}(\mu, 2),$$
$$r \sim \text{Uniform}(0,1), \quad \textbf{[S3]}$$

where $\mu = \max(k_{DNA}/(W_{DNA} \times C_{DNA}), 10^{-7})$ are weakly informative (54), intended simply to prevent pathological inferences. This approach yields conservative parameter estimates. For example, posterior distributions of reactivation probabilities for large clones are nonzero even if no reactivation events are observed. In models that include data from multiple visits, we also included an informative prior to take into account the expected decay of the latent reservoir in individuals on uninterrupted ART,

$$\log_{10}\left(\frac{p_2 r_2}{p_1 r_1}\right) \sim \text{Normal}(-\lambda t, \sigma). \quad \textbf{[S4]}$$

Here, p_i and r_i are the frequency of integrated provirus and reactivation probability at visit i, respectively. Their product is therefore equivalent to the number of infectious units per million cells (IUPM), divided by 10^6. We used $\lambda = -0.007$ mo^{-1} as the IUPM decay rate and $\sigma = 0.38$ as the SD of its measurement, following previous estimates (3). The time between visits (in months) is t. We implemented all models in Stan (52) using the PyStan interface.

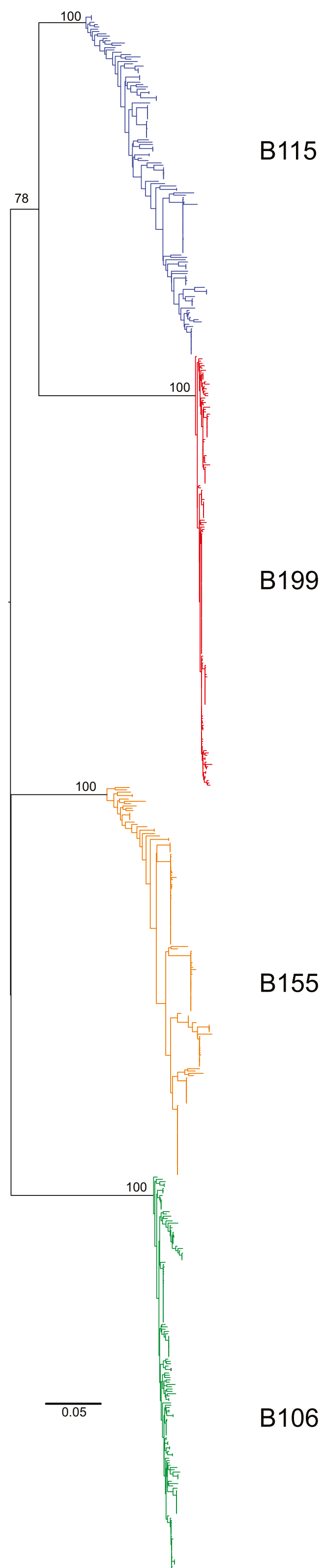
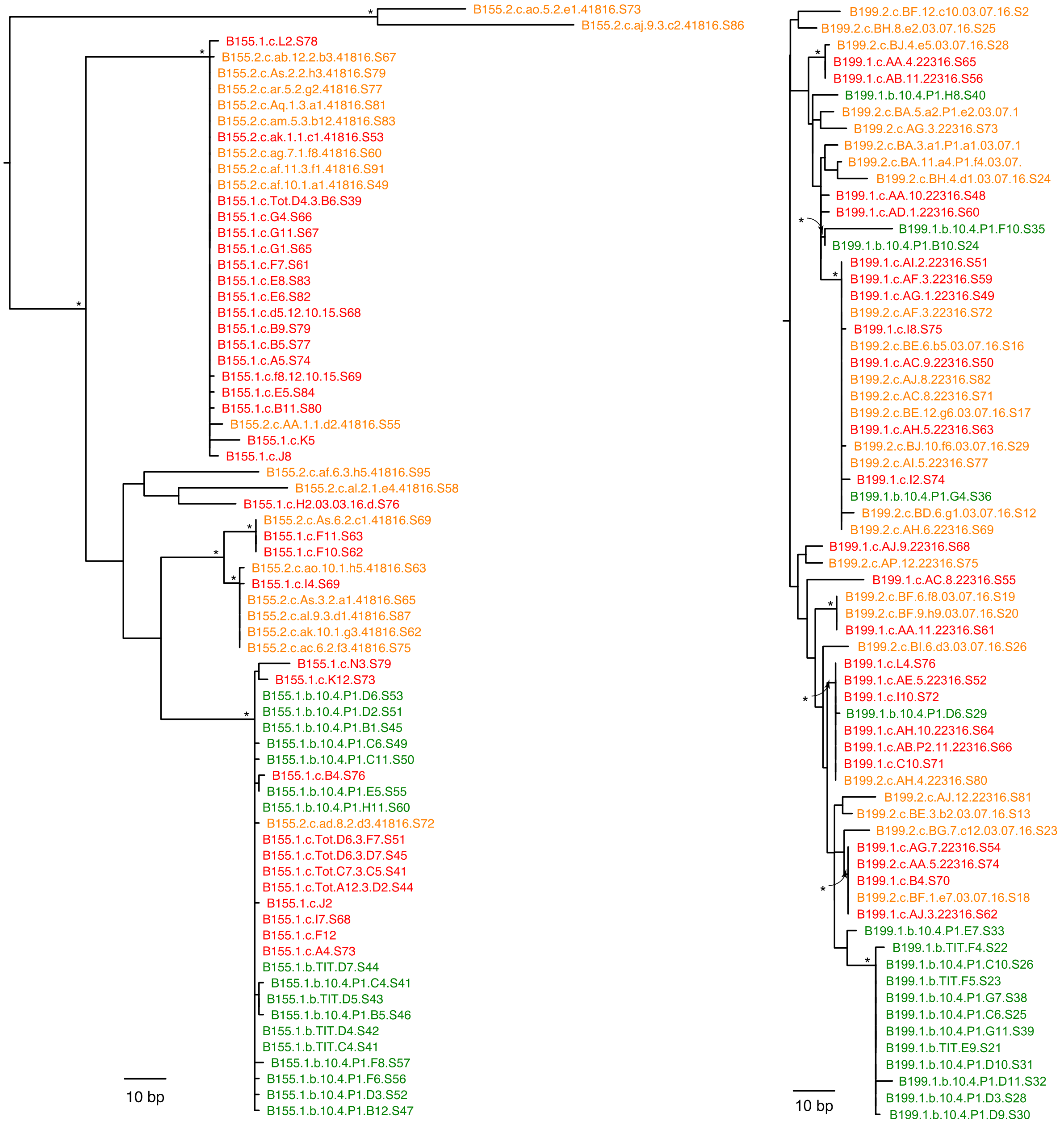


Fig. S1. Maximum likelihood phylogenetic tree was constructed from viral *env* sequences from outgrowth culture supernatants as well as archived proviral DNA from all participants. Hypervariable (as defined in https://www.hiv.lanl.gov/content/sequence/VAR_REG_CHAR/) and other poorly aligned regions were excluded from the analysis. The tree was constructed using RAxML v. 8.0.22 (55) with a GTRGAMMA substitution model, with 1,000 bootstrap replicates and midpoint rooted.

	Virus ID	3BNC117 IC80	VRC01 IC80	10-1074 IC80	PGT121 IC80	PGDM1400 IC80	10E8 IC80
B106	B106.1.c.1.2	1.01	14.1	4.32	1.14	1.42	>50
	B106.1.c.2.9	1.87	>50	0.07	0.27	0.87	>50
	B106.2.c.Q6	1.30	13.1	0.56	0.93	1.38	>50
	B106.1..c.2.2	1.09	13.6	0.04	0.40	0.47	>50
	B106.2.c.P12	0.37	4.88	0.02	0.20	NT	>50
	B106.2.c.P9	0.64	6.46	0.03	0.33	0.15	>50
	B106.1.c.5.10	0.43	8.10	0.03	0.31	0.17	>50
	B106.2.c.P2	0.58	4.24	0.47	0.84	0.07	>50
	B106.1.c.E3	5.40	26.5	0.28	7.32	0.19	>50
	B106.1.c.2.8	1.52	22.8	1.10	0.90	0.22	>50

	Virus ID	3BNC117 IC80	VRC01 IC80	10-1074 IC80	PGT121 IC80	PGDM1400 IC80	10E8 IC80
B115	B115.2.c.h12	18.4	NT	0.34	NT	NT	NT
	B115.2.c.AC4	2.87	NT	0.59	NT	NT	NT
	B115.1.c.G7	2.65	NT	0.49	NT	NT	NT
	B115.1.c.AA4	2.72	NT	0.41	NT	NT	NT
	B115.2.c.P8	7.13	NT	0.42	NT	NT	NT
	B115.1.c.F10	6.85	NT	1.69	NT	NT	NT
	B115.2.c.AE10	10.0	NT	3.70	NT	NT	NT
	B115.2.c.AG2	9.23	NT	2.16	NT	NT	NT
	B115.2.c.AD7	6.85	38.3	1.86	4.69	6.38	>50
	B115.2.AC5	11.4	NT	2.63	NT	NT	NT

Fig. S2. (Continued)



B155

	3BNC117	VRC01	10–1074	PGT121	PGDM1400	10E8
Virus ID	IC80	IC80	IC80	IC80	IC80	IC80
B155.2.c.Aj.9	5.02	4.75	0.10	0.86	>50	>50
B155.2.c.AS2	6.61	35.4	>50	>50	>50	27.5
B155.2.c.AM5	5.77	41.0	>50	>50	>50	35.6
B155.1.c.k5	7.37	NT	>50	NT	NT	NT
B155.2.c.Af.6	4.75	31.0	>50	>50	>50	47.2
B155.2.c.AS6	5.55	38.0	>50	14.5	>50	45.2
B155.1.c.F11	6.20	45.5	>50	32.8	>50	>50
B155.2.c.AK10	5.30	39.2	>50	36.8	>50	43.8
B155.2.c.AD.8	6.07	26.5	>50	>50	>50	36.4
B155.1.c.F12	9.62	NT	>50	NT	NT	NT

B199

	3BNC117	VRC01	10–1074	PGT121	PGDM1400	10E8
Virus ID	IC80	IC80	IC80	IC80	IC80	IC80
B199.1.c.BF12	32.0	>50	0.20	0.77	2.36	47.2
B199.1.c.AA4	2.99	31.0	6.67	>50	21.0	>50
B199.1.c.AG1	20.7	>50	8.63	>50	3.74	>50
B199.1.c.AC8	3.03	>50	0.76	2.75	0.76	>50
B199.1.c.AJ9	6.55	>50	2.84	42.6	2.53	>50
B199.1.c.AA11	4.61	>50	0.59	1.15	0.59	>50
B199.1.c.AG7	1.88	36.0	1.88	12.2	0.56	>50
B199.2.c.BE.3	34.6	>50	1.33	10.6	0.52	>50
B199.1.c.AE5	11.1	>50	1.60	3.34	0.50	>50

Fig. S2. Maximum likelihood phylogenetic tree was constructed from viral *env* sequences from outgrowth culture supernatants for each participant. Viruses from time point 1 are red, viruses from time point 2 are orange, and bulk culture viruses are green. Asterisks indicate nodes with significant bootstrap values (bootstrap support ≥ 70%). Tables beneath each tree show concentration that inhibits response by 80% (IC_{80}) titers for selected outgrowth culture viruses (red, IC_{80} of 0–0.1 μg/mL; orange, IC_{80} of 0.1–1.0 μg/mL; yellow, IC_{80} of 1.0–10 μg/mL; green, IC_{80} of 10–50 μg/mL). NT, not tested.

Table S1. Clinical characteristics of study subjects

Study ID	Age	Sex	Year of HIV diagnosis	CD4 nadir	Years since HIV diagnosis	Years on ART	ART regimen
B106	27	M	2008	390	7	7	TDF/FTC/RPV
B115	44	M	1993	200	22	22	DRV/r, ABC, 3TC
B155	59	M	1993	444	22	15	TDF/FTC/RPV
B199	49	M	2009	200	6	4	TDF/FTC, RAL

ABC, abacavir; DRV/r, darunavir/ritonavir; FTC, emtricitabine; ID, identification; M, male; RAL, raltegravir; RPV, rilpivirine; 3TC, lamivudine; TDF, tenofovoir disoproxil fumarate.