References

[1] Los Alamos National Laboratory (2015) HIV sequence database. (http://www.hiv.lanl.gov/).

[2] Simmonds P, et al. (1991) Discontinuous sequence change of human immunodeficiency virus (HIV) type 1 env sequences in plasma viral and lymphocyte-associated proviral populations in vivo: implications for models of HIV pathogenesis. *Journal of Virology* 65(11):6266–6276.

[3] Shankarappa R, et al. (1999) Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* 73(12):10489–10502.

[4] Edwards CTT, et al. (2006) Population genetic estimation of the loss of genetic diversity during horizontal transmission of HIV-1. *BMC Evolutionary Biology* 6(28).

[5] Fischer M, et al. (2004) Attenuated and nonproductive viral transcription in the lymphatic tissue of HIV-1-infected patients receiving potent antiretroviral therapy. *J. Infect. Dis.* 189(2):273–285.

[6] Novitsky V, et al. (2009) Timing constraints of in vivo gag mutations during primary HIV-1 subtype C infection. *PLoS ONE* 4(11):e7727.

[7] Akaike H (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6):716–723.

Table S1: Summary of the patient data collected from the LANL HIV sequence database [1] in the data sets from public sources.

Reference	ID	Sequences			Time points			Time span	Linear model		MAE		MAD		p-Latency
		Plasma	PBMC	Total	Plasma	PBMC	Total	Plasma	ΔAIC	Root	Years	Scaled	Years	Scaled	
[2]	2658	69		69	5		5	5.2	64	-1.2 (-2.0, -0.46)	0.42	0.081	0.64	0.12	0.23
[3]	825	49		49	6		6	8.2	60	0.11 (-0.53, 0.74)	0.73	0.089	0.89	0.11	1.0
[4]	**7259**	**28**		**28**	**4**		**4**	**2.4**	**9.8**	**-0.97 (-2.2, 0.30)**	**0.69**	**0.28**	**0.61**	**0.25**	**0.79**
	7265	**20**		**20**	**3**		**3**	**0.75**	**8.4**	**-0.73 (-1.4, -0.024)**	**0.18**	**0.25**	**0.33**	**0.44**	**1.0**
	13333	38		38	4		4	1.4	17	-0.49 (-1.1, 0.13)	0.31	0.22	0.30	0.22	1.0
	13334	36		36	5		5	2.0	11	-1.2 (-2.4, 0.071)	0.47	0.23	0.44	0.22	0.24
	13336	42		42	4		4	2.0	28	-0.31 (-0.70, 0.079)	0.34	0.17	0.34	0.17	1.0
[3]	821	69	178	247	7	17	17	6.5	180	-0.20 (-0.46, 0.064)	0.46	0.070	0.78	0.12	$< 10^{-5}$*
	822	29	90	119	3	10	10	5.8	70	-1.6 (-2.2, 0.95)	0.50	0.086	0.89	0.15	0.0080
	824	52	102	154	7	9	13	8.6	130	-0.65 (-1.2, -0.057)	0.63	0.073	0.96	0.11	0.0072
	13889	77	65	142	13	14	18	13	130	-3.5 (-4.7, -2.3)	1.5	0.11	1.8	0.13	$< 10^{-5}$*
[5]	**10769**	**108**	**56**	**164**	**10**	**4**	**11**	**5.4**	**5.8**	**-17 (-30, -3.3)**	**5.6**	**1.0**	**7.6**	**1.4**	**0.02**
[6]	34391	12	35	47	3	5	6	0.91	21	-0.99 (-1.4, -0.57)	0.12	0.14	0.24	0.27	0.017
	34411	14	19	33	3	3	6	1.3	22	-0.027 (-0.35, 0.30)	0.14	0.11	0.25	0.20	0.019

ID corresponds to the anonymized patient identifiers in the LANL database. Time span is in years. ΔAIC is the Akaike Information Criterion (AIC) [7] of the null model minus the AIC of the linear model. Root is the estimate of the root time in years by the linear model with respect to the time of the first sample. MAE is Mean Absolute Error between collection date and estimated date of the training data. MAD is Mean Absolute Difference is between collection date and estimated date) of the censored data. Scaled MAE/D is the mean absolute error/difference divided by the time span of the training data. p-Latency is the p-value of the nonparameteric binomial test (* marks significant results after a Bonferroni correction). The model failed to be calibrated for bold rows.