Supplementary Note: Protein- and allele-specific adaptation

The aim of the adaptation framework introduced here is to reduce the high-dimensional space of immune escape to a single dimension: the adaptation score. While this reduction in complexity enables an intuitive interpretation, and statistical exploration, of HLA-mediated escape, it necessarily loses information. It is thus natural to restore some complexity by focusing on different components of the adaptation score in an effort to address specific scientific questions. In this work, we consider three such components: HLA allele specific adaptation, HLA locus-specific adaptation, and HIV protein-specific adaptation.

Because adaptation with respect to an individual's HLA profile is simply the arithmetic mean of allele-specific adaptation, the first two are straightforward extensions. In analyses where allele-specific effects are the primary outcome, there are thus no direct statistical concerns (though see below). With respect to locus-specific adaptation, we approached all analyses from the standpoint that overall adaptation is the primary measure of interest. Although there is ample evidence in the literature that HLA-B is the primary locus linked with control (and progression), the underlying mechanism is unknown, and there is evidence that HLA-A and -C restricted responses play a role as well. We thus consider locus-specific effects to be secondary analyses. We report these results in supplementary figures and tables, but do not discuss them at length in the text and note that any effects are subject to the usual caveats of post hoc analyses. The one exception is the heritability analysis, which was the final analysis undertaken, and thus had the benefit of the observation that HLA-B adaptation was consistently the most important locus in all of the previous analyses. This observation, combined with prior reports on the Zambian dataset that HLA-B sharing among linked partners was associated with elevated VL in the recipient, led us to treat HLA B-similarity as the primary analysis. However, we note that overall similarity showed the same trend (Heritability = 7% (95% CI: -15–28%), 15% (-6–37%), 25% (-5–55%) when stratified by tertile), though the interaction effect was not statistically significant ($P = 0.2$).

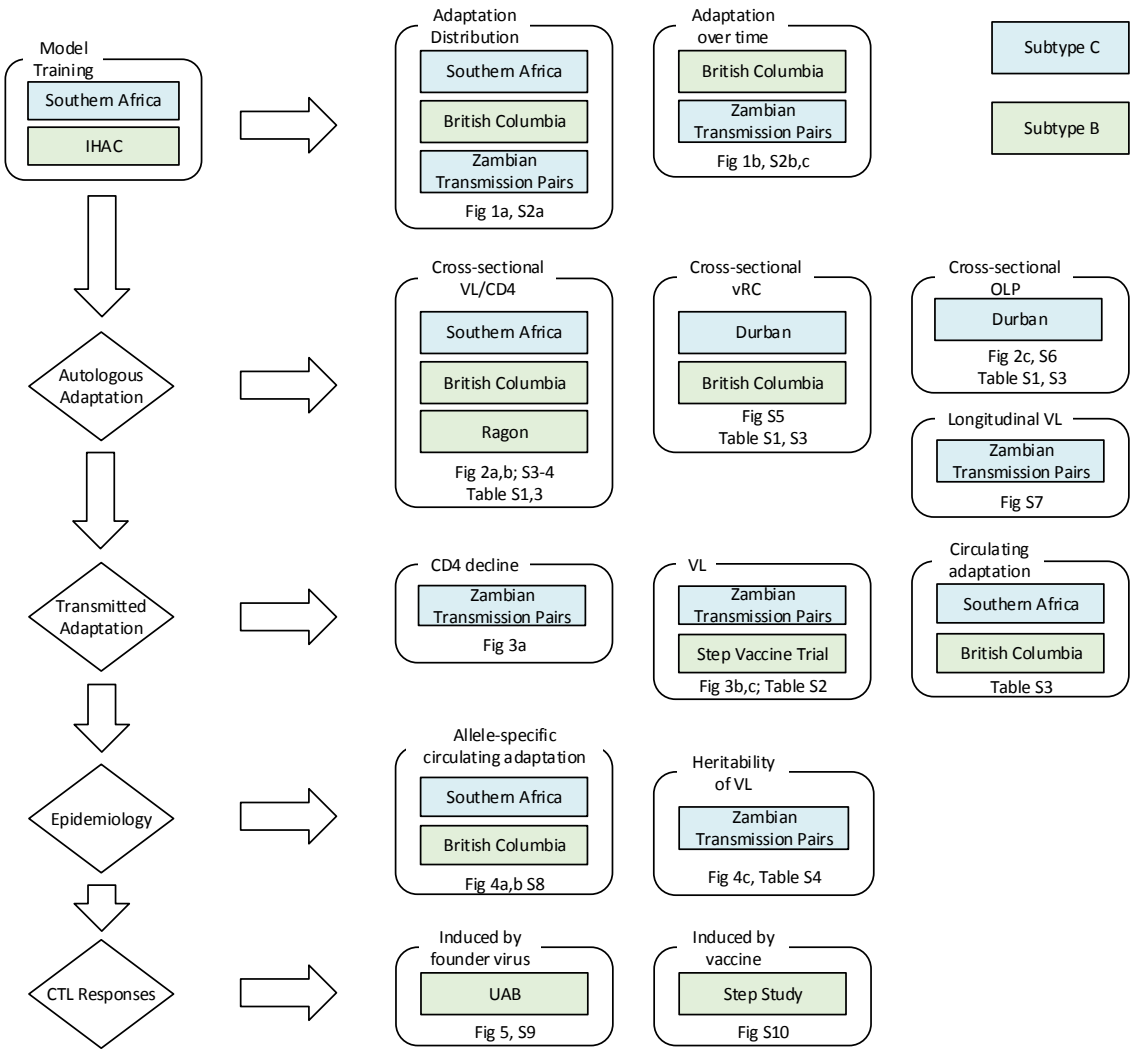
Protein specific adaptation is more problematic. While we present these results in the Supplementary material for the interested reader, we caution against direct interpretation without additional follow up experiments. At issue is the way in which the model is trained. The effect size that the model estimates for each escape substitution is roughly the comparison of the frequency of the amino acid in individuals with the HLA compared to those without. If escape or reversion dynamics differ among proteins (we previously reported substantially slower reversion rates in Pol over all non-consensus variants; Table S2 in Ref. 9), then the model-estimated effect sizes will be systematically different from protein to protein. Indeed, hints of this may be observed in Supplementary Figure 2b, where Pol has substantially higher autologous adaptation (to the donor) and lower heterologous adaptation (here, shown as transmitted adaptation to the new host), and shows very different trends over the first two years of infection. Whether these differences in model-estimated effect sizes are immunologically meaningful is unclear. Given prior work that has highlighted the relative importance of Gag-specific CTL responses (e.g., Ref. 40), we are inclined to dismiss as model artifact the observation that Pol-specific adaptation is consistently the strongest predictor of all outcomes studied here. However, we remain open to the possibility that this is a biologically relevant observation, and have thus elected to present these results in the Supplementary Material and intend to continue to explore this observation in future work.

Finally, it is worth noting that some HLA alleles are associated with a higher number of escape mutations in some proteins than in others (see association lists in the Supplementary Data file). While this difference may be due to chance, any protein-specific biases induced on the model will thus manifest themselves within allele-specific adaptation. While this is of potential concern, we do not see evidence that this biases our results. For example, in Figure 2b, autologous adaptation eliminated control associated with all alleles tested, which included alleles such as B*81:01 (24/41, 59%, of associations in Gag) and B*44:03 (51/79, 65%, of associations in Pol). Similarly, in Fig. 4a, because overall circulating adaptation was lowest in Pol, protein-specific biases would cause us to overestimate the reported effect if alleles preferentially targeting Pol were most likely to be protective. In fact, the majority of protective alleles preferentially target Gag (or have no clear preference), and the allele with the lowest circulating adaptation is B*57:03, with 29/53 (55%) of associations in Gag.

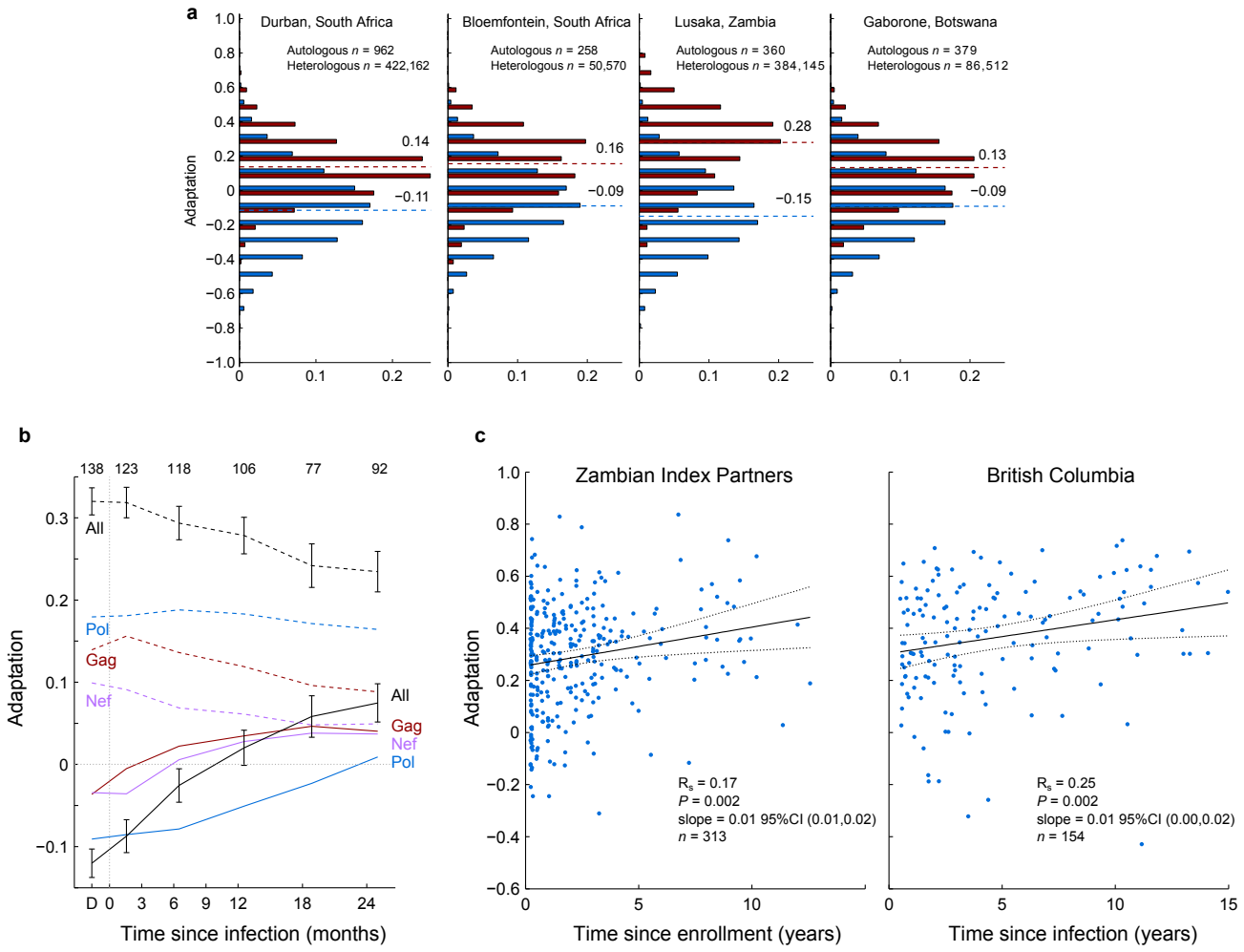
	Name	Location	n	Sequence Coverage* (n)	Sample Dates (range)	Age (IQR)	Male (%)	VL mean† (n)	CD4 mean (n)	Functional Data (n)
Chronic Infection	Southern Africa	Durban, South Africa	1246	Gag/Pol/Nef (949)	1999-2005	23-30	7	29400 (1237)	395 (1158)	vRC (403), OLP (691)
		Bloemfontein, South Africa	261	Gag/Pol/Nef (258)	2006	31-42	27	45700 (163)	266 (261)	
		Kimberley, South Africa	31	Gag (26)	2009	24-38	0	22300 (14)	419 (29)	
		Lusaka, Zambia‡	379	Gag/Pol/Nef (360)	1998-2009	26-36	53	47700 (351)	319 (56)	
		Gaborone, Botswana	514	Gag/Pol/Nef (379)	2007-2008	23-32	0	16100 (471)	367 (415)	
		Thames Valley	102	Gag/Pol/Nef (65)	2007-2009	n.a.	n.a.	4100 (100)	473 (98)	
	Internation HIV Adaptation Collaborative (IHAC)	British Columbia, Canada	1103	All but gp120 (1103)	1996-1999	32-49	86	120226 (1048)	290 (1048)	vRC (734), Duration of Inf (325)
		ACTG Clinical Trials Sites	538	All but gp120 (538)	2003-2004	32-45	81	n.a.	n.a.	
		Western Australia	247	All but gp120 (247)	1997-2002	30-45	79	n.a.	n.a.	
	Ragon Elite Controllers	North America	21	All but Env (21)	2001-2009	50-59	85	<50 (21)	n.a.	
	Ragon Non-Controllers	North America	80	All but Env (80)	2001-20010	47-54	76	n.a.	n.a.	
Early Infection	Zambian Transmission Pairs§	Lusaka, Zambia	129	Gag/Pol/Nef (129)	1998-2009	24-35	43	43600 (129)	longitudinal (46)	vRC (113), Donor VL (129)
			275	n.a.	1995-2011	23-34	43	38100 (275)	n.a.	Donor VL (275)
	Step Study	North America	56	Gag/Pol/Nef (56)	2005-2007	25-36	100	112200 (56)	n.a.	Pre-infection Elispot (1143)
	UAB Clinic	Birmingham, AL	11	Full Proteome (11)	2000-2012	23-46	91	569400 (11)	455 (11)	Autologous CTL response (11)

Blue: Subtype C, Green: Subtype B, Bold box: Adaptation model training data.

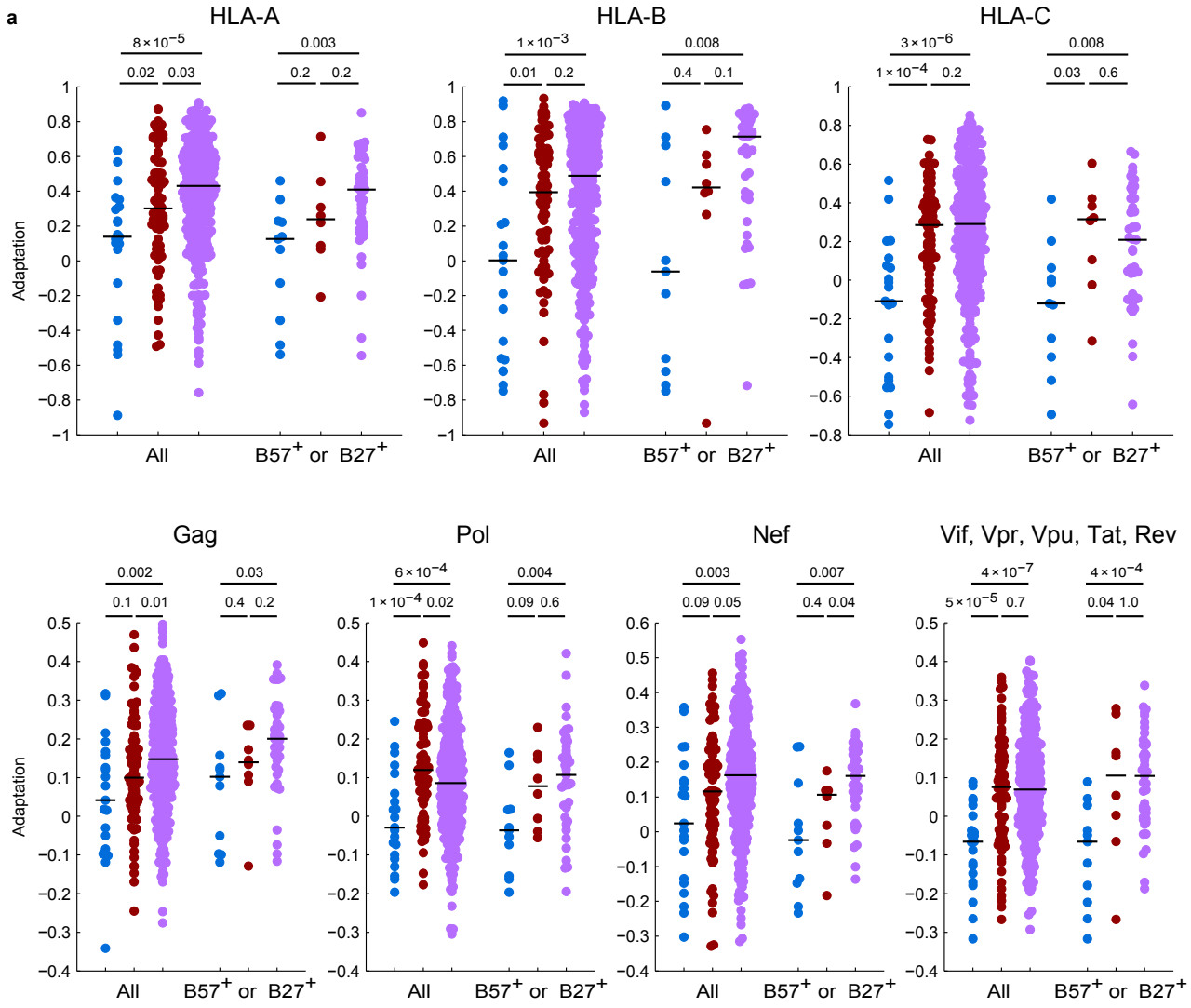
*Counts indicate number of individuals with any sequence data. Early infection and Ragon cohorts had no missing sequence data. †Geometric mean was used for VL. ‡These individuals were taken from the seropositive partners from the Zambian Transmission Pairs cohort. §Two datasets were previously defined for the Zambian transmission pairs. See methods for details.



Supplementary Figure 1. Schematic of datasets used for each analysis. Throughout the work, we refer to each dataset by the name or location, whichever is most specific. See methods for further details. Unless otherwise specified, all data are from cross-sectional samples.



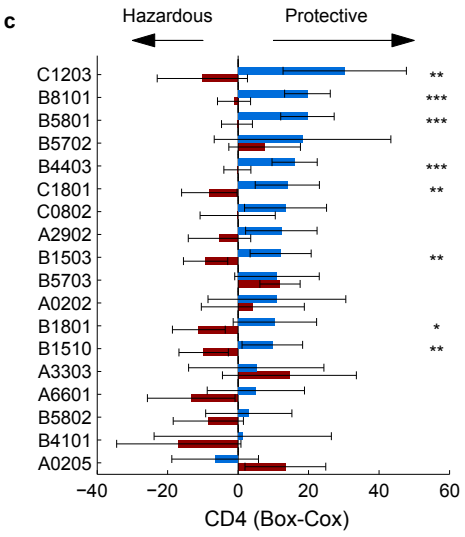
Supplementary Figure 2. Distribution of autologous and heterologous adaptation by cohort and over time. (a) Autologous (red) and heterologous (blue) adaptation distributions for the Southern Africa (HIVC) cohorts. See **Figure 1a** in the main text for British Columbia (HIVB) and combined Southern Africa distributions. Median adaptation for each distribution is indicated. (**b**) Autologous adaptation is shown for linked transmission pairs from Zambia, as in **Figure 1b** in the main text. Adaptation to recipients' (solid) or donors' (dashed) alleles is shown for Gag (red), Pol (blue), Nef (purple) or all proteins (black). Note that adaptation to each protein provides stronger evidence for adaptation overall in a manner that is approximately additive (see methods). Error bars, 95% confidence intervals. Number of samples in each time point indicated at top. Adaptation of linked donor sequence (time point 'D') is set to -50 d for display. (**c**) Autologous adaptation within chronically infected individuals as a function of approximate time since infection. For Zambian Index Partners, a lower bound of time since infection, as defined by time since enrollment, is used. For British Columbia, estimated time since infection for a subset of patients was provided based on patient-doctor surveys. Both plots are limited to individuals with no missing sequence data. R_S, *P*-value, from Spearman ρ.



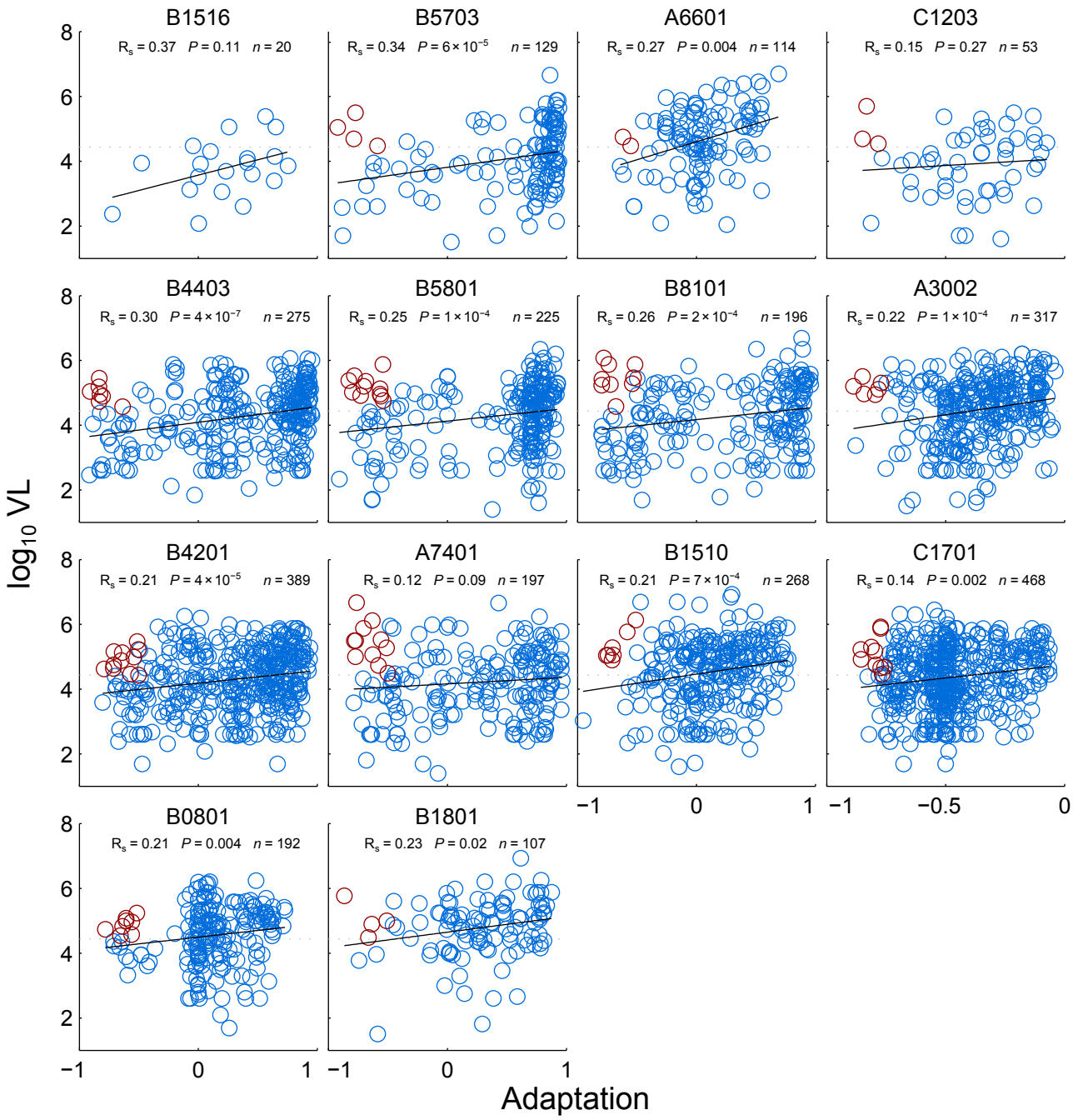
b

Covariate*	log-odds ratio†	SE	p‡	q§
(Intercept)	-1.90	0.50	na	na
Adapt	-5.38	1.36	1E-06	1E-04
B57	4.25	1.19	3E-05	0.001
A68	-Inf	na	0.01	0.36
A03	1.86	0.83	0.02	0.38
C03	1.99	1.00	0.04	0.62

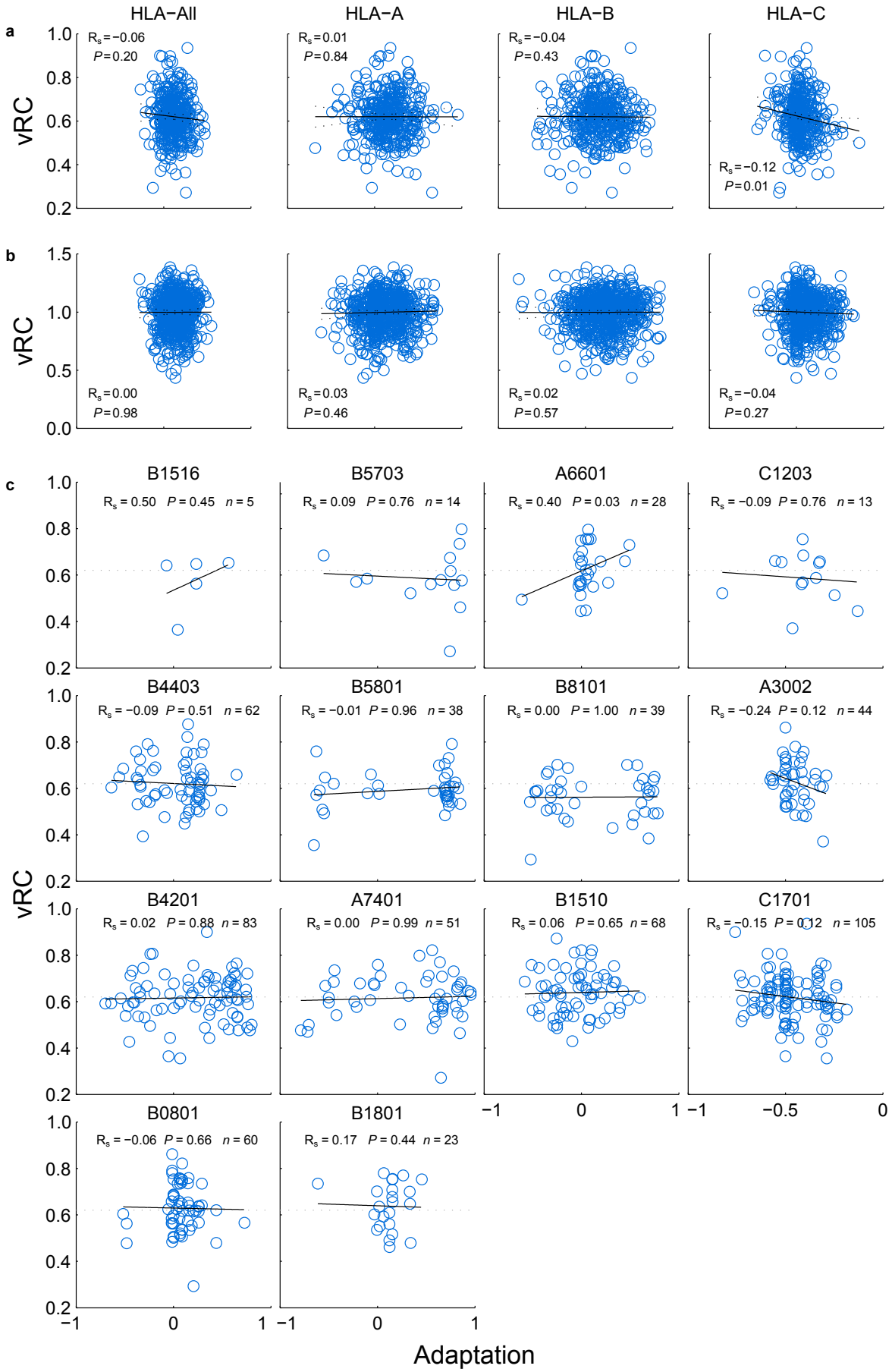
*Possible features include autologous adaptation, and all HLA alleles at type and subtype resolution that are observed in at least 5% (6/101) of individuals. †Model parameters from stepwise logistic regression with $P < 0.05$ as the entry criterion. ‡P-values from a likelihood ratio test, using all other selected features as covariates. §q-value computed over all features considered for the model, using the selected features as covariates.



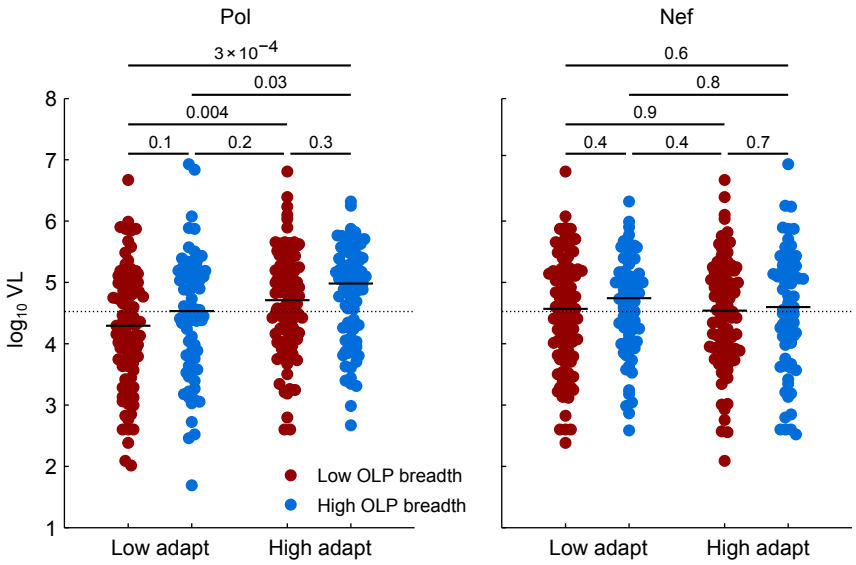
Supplementary Figure 3. Autologous adaptation is associated with markers of disease progression. (**a**, **b**) Autologous adaptation is lower for elite controllers. (**a**) Data are as in **Figure 2a** in the main text, but with adaptation computed separately by HLA locus or protein. *P*-values, Mann-Whitney U test. Blue, Ragon controllers ($n = 21$); Red, Ragon non-controllers ($n = 80$); Green, British Columbia ($n = 383$, limiting to those with complete viral genome sequences). (**b**) Stepwise logistic regression over the Ragon cohorts was performed with controller status as the dependent variable. Adaptation is with respect to all loci and proteins. (**c**) Estimated HLA-specific effects on CD4 count in the HIVC cohort. Estimated CD4 count (Error bars, 95% CI) relative to cohort average for individuals expressing the allele with no (blue) or with complete (red) allele-specific adaptation. Significant (likelihood ratio test) adaptation effects are denoted for $P < 0.001$ (***), $P < 0.01$ (**), and $P < 0.05$ (*). Compare to **Figure 2b** in the main text.



Supplementary Figure 4. Allele-specific adaptation eliminates allele-specific control. Viral load versus allele-specific autologous adaptation among protective alleles. Scatter plots are shown for each allele in **Figure 2b** for which expression of the allele in the absence of adaptation was associated with below average VL (dashed line). All data are for subtype C (Southern Africa cohort). Individuals with allele-specific adaptation less than −0.5 and above average VL are highlighted in red. These individuals may not be mounting effective allele-specific CTL responses. R_s, *P*-value, from Spearman ρ.



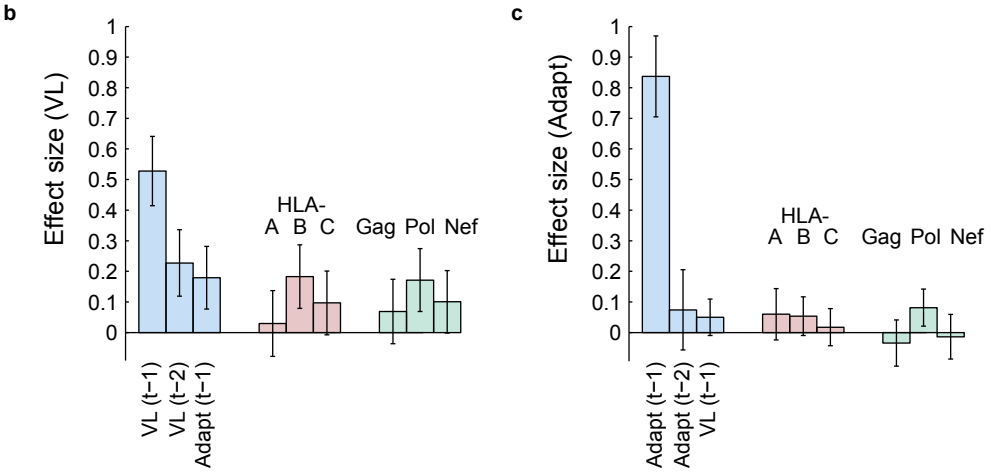
Supplementary Figure 5. Increasing adaptation is not associated with reduced vRC. (a, b) *In vitro gag-protease* viral replicative capacity (vRC; the relative growth rate of NL4-3 recombinant viruses encoding subject-derived *gag-protease* sequences) versus autologous Gag adaptation against all alleles, or a single locus. (**a**) Durban ($n = 403$; HIVC); (**b**) British Columbia ($n = 734$; HIVB). The nominally significant negative correlation for HLA-C in Durban was not significant when accounting for multiple testing. (**c**) Durban vRC versus allele-specific autologous Gag adaptation among protective alleles. Scatter plots are shown for each allele in **Figure 2b** for which expression of the allele in the absence of adaptation was associated with below average VL in Southern Africa. Dashed line, cohort-wide mean vRC. R_s, *P*-value, from Spearman ρ.



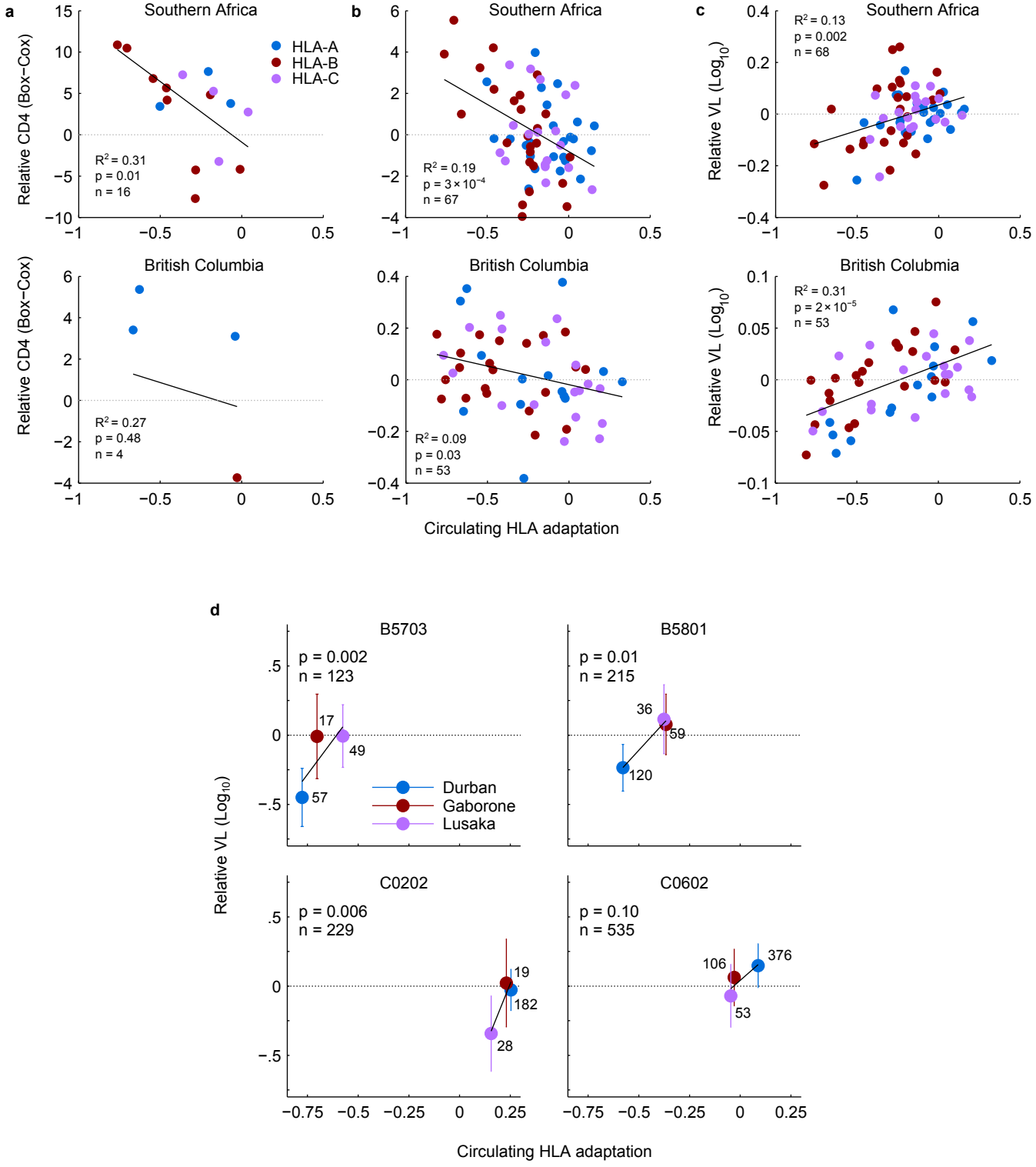
Supplementary Figure 6. OLP breadth and autologous adaptation in Pol and Nef. VL for each of $n = 319$ HIVC-infected patients from Durban are shown, stratified by protein-specific adaptation and OLP response breadth. Solid bars, stratum median. Dashed line, cohort median. *P*-values, two-tailed Mann-Whitney U-test. Compare to **Figure 2c** in the main text. Note that neither Pol nor Nef responses have been consistently linked to lower viral load. In addition, increased targeting may be a result of loss of viral control, if increased virus load increases antigen stimulation or is indicative of escape elsewhere. Such a "reverse" causation is consistent with the pattern observed in Pol, though other explanations are possible as well.

a

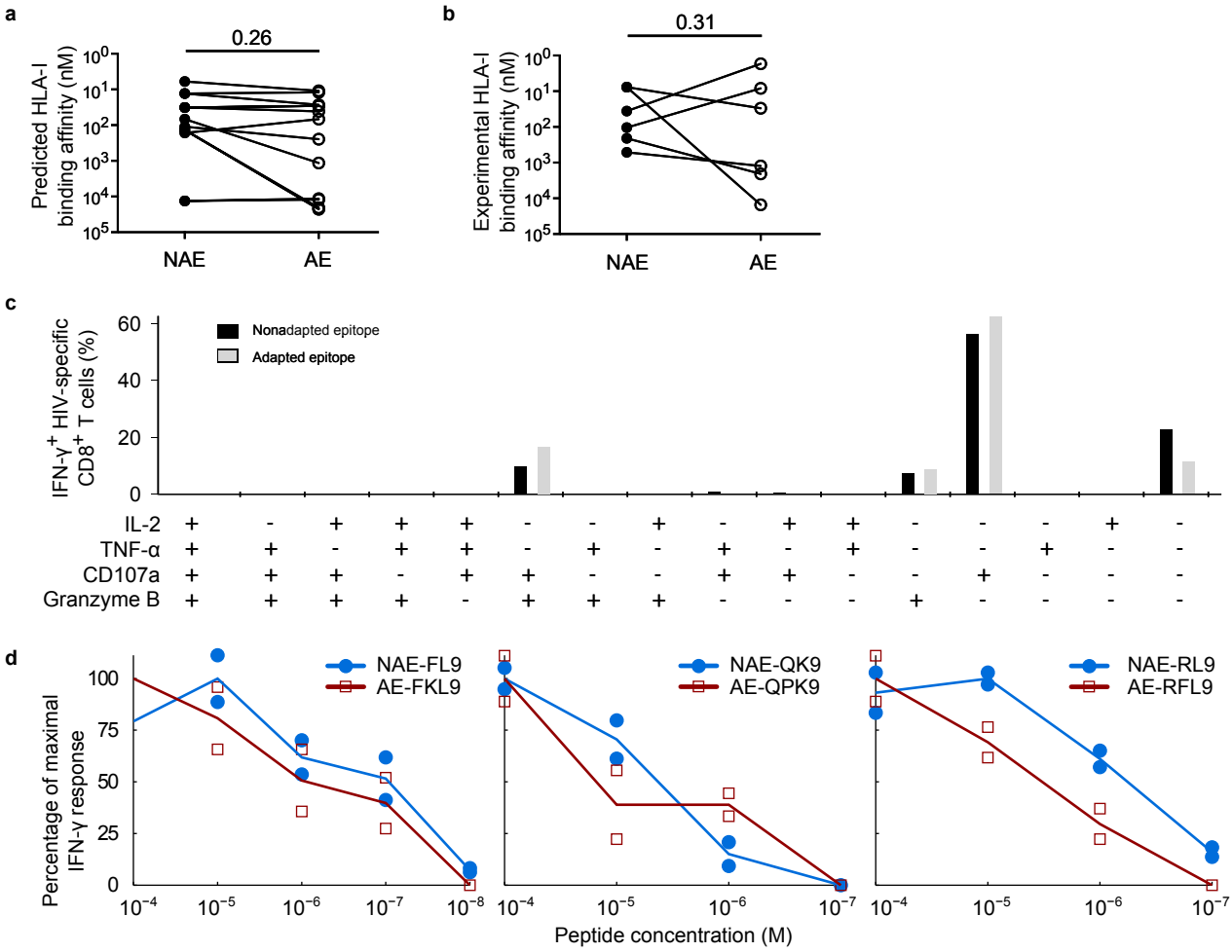
	Viral Load				Adaptation		
	Estimate	**p**	**R^2**		**Estimate**	**p**	**R^2**
VL(t-1)	0.53	<1E-16	22.4%	Adapt(t-1)	0.84	<1E-16	37.5%
VL(t-2)	0.23	6E-05	7.8%	Adapt(t-2)	0.07	0.27	0.7%
Adapt(t-1)	**0.18**	**7E-04**	**6.2%**	**VL(t-1)**	**0.05**	**0.10**	**1.4%**
Age (≥40)	0.01	0.85	0.0%	Age (≥40)	-0.02	0.48	0.3%
Is Female	-0.06	0.31	0.6%	Is Female	-0.01	0.76	0.0%
Infection Year	-0.08	0.12	1.3%	Infection Year	-0.07	0.03	2.5%
SubjectID	*0.16*	*0.40*	*0.0%*	*SubjectID*	*0.00*	*0.50*	*0.0%*
HLA-A	*0.00*	*0.50*	*0.0%*	*HLA-A*	*0.00*	*0.50*	*0.0%*
HLA-B	*0.02*	*0.50*	*0.0%*	*HLA-B*	*0.00*	*0.50*	*0.0%*
HLA-C	*0.00*	*0.50*	*0.0%*	*HLA-C*	*0.00*	*0.50*	*0.0%*
N	190		60.8%	N	195		85.5%



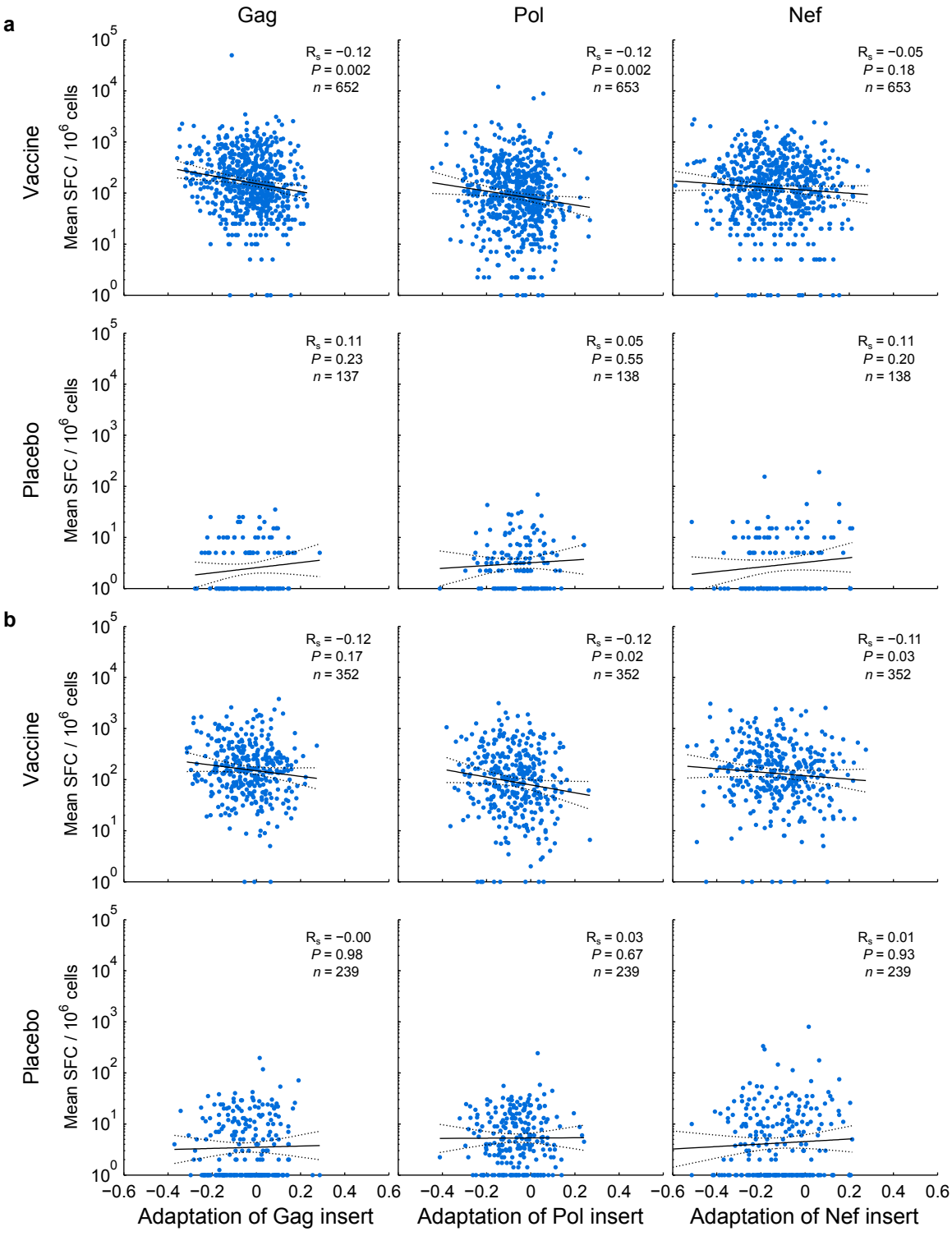
Supplementary Figure S7. Autologous adaptation predicts subsequent changes in VL. (**a**) LMM estimates and effect sizes for autoregression models in which the end point is VL (left) or autologous adaptation (right). In each model, the target variable from the previous two time points is included as independent variables. Bold, the primary end point of the analysis: autologous adaptation as a predictor of VL (left) or VL as a predictor of autologous adaptation (right), at the previous time point. Italics, random effects. Infection year, binary fixed effect around the median. N, total number of observed end points, over 77 subjects. All continuous variables were standardized to make effect sizes comparable. Total variance explained is substantially more than the sum of the components due to high covariance among the lagged VL and adaptation variables. (**b,c**) The effect sizes and 95% CI are shown for the VL (**b**) and Adaptation (**c**) dependent variables. Blue (left cluster), primary analysis as shown in (**a**). Red (middle cluster), substitution of Adaptation variables for locus-specific adaptation. Green (right cluster), substitution of Adaptation variables for protein-specific adaptation.



Supplementary Figure S8. Circulating adaptation and allele-specific effects on VL and CD4 counts. Circulating adaptation explains relative protection associated with HLA alleles. (**a-c**) Circulating allele-specific adaptation is compared against the allele-specific effects on CD4 counts (Box-Cox transformed) (**a,b**) or VL (**c**), as estimated from a random effects model. *P*-values, pseudo-R^2, from LMM with random offsets for each locus. (**a**) Alleles significantly associated with CD4 counts. (**b,c**) All alleles observed in at least 20 individuals. See **Figure 4** in the main text for association among alleles significantly associated with VL. (**d**) There was evidence ($P < 0.1$) of a city-specific interaction effect for four HLA alleles on VL. For each of these alleles, the city-specific effect on VL is compared with the city-specific circulating adaptation. *P*-value, likelihood ratio test for interaction variable. Total number of individuals expressing each allele in cohort representing each city is shown. See **Figure 4** in the main text for an omnibus statistical test.



Supplementary Figure 9. Functional characterization of primary CD8$^+$ T cell responses to adapted and non-adapted epitopes. (**a, b**) Predicted (**a**) and experimental (**b**) peptide binding affinity for matched adapted (AE) and nonadapted (NAE) epitope pairs. See **Figure 4c,d** in the main text for binding affinity of all peptides tested. (**c**) The overall polyfunctionality of responses (1−4 functions in addition to interferon-γ production) was assessed for the three pairs of responding NAE/AE epitopes, using SPICE and PESTLE software. All combinations of effector and degranulation functions were derived from interferon-γ$^+$ CD8$^+$ T cells. (**d**) Antigen sensitivity is shown for the three matched immunogenic NAE/AE epitope-specific CD8$^+$ T cells. All data for duplicate experiments are shown; lines trace the mean response at each concentration. Data are normalized to maximal mean response. See **Figure 4h** for unnormalized curves.



Supplementary Figure 10. Pooled ELISpot responses to Step vaccine inserts are negatively correlated with protein-specific adaptation of the inserts to each participant's HLA alleles. (**a, b**) The mean number of spot forming cells (SFC) per million cells were estimated for each insert using interferon-γ ELISpot pooled peptides that spanned each insert. For polymerase, we use the geometric mean of the SFC/million measured on each of two peptide pools. R_s, *P*-value, from partial rank correlation, correcting for participant Age (≥40), Sex, Race (white), and AD5 serostatus. ELISpot assays were previously performed in duplicate by the HVTN (**a**) and Merck (**b**) laboratories.

Supplementary Table 1. Adaptation predicts cross sectional VL and CD4 counts in a linear mixed model (LMM).

		Viral Load*						CD4†					
		British Columbia (HIVB)			Southern Africa (HIVC)			British Columbia (HIVB)			Southern Africa (HIVC)		
	Feature‡	estimate	p	R^2	estimate	p	R^2	estimate	p	R^2	estimate	p	R^2
HLA	Female	-0.16	4E-03	0.8%	-0.28	2E-05	0.8%	1.05	0.26	0.1%	6.01	4E-04	0.6%
	Age40§	0.04	0.32	0.1%	0.06	0.54	0.0%	-1.00	0.13	0.2%	-4.83	0.03	0.2%
	HLA-A	0.06	9E-03	0.5%	0.05	0.02	0.2%	0.95	0.02	0.4%	0.88	0.07	0.1%
	HLA-B	0.06	0.02	0.4%	0.15	<1E-16	3.0%	0.00	0.50	0.0%	3.19	7E-13	2.5%
	HLA-C	0.00	0.50	0.0%	0.02	0.40	0.0%	0.00	0.07	0.2%	1.26	0.05	0.1%
	Cohort ||	0.18	3E-08	2.8%	0.36	<1E-16	6.9%	0.00	0.07	0.2%	5.98	<1E-16	4.7%
	N¶	1048		5.5%	2298		14.5%	1048		0.8%	1983		14.7%
Adaptation	**Adaptation**	**0.27**	**2E-03**	**0.9%**	**0.99**	**<1E-16**	**3.7%**	**-10.15**	**2E-12**	**4.5%**	**-22.94**	**1E-16**	**4.0%**
	Female	-0.16	4E-03	0.8%	-0.29	8E-06	1.1%	1.11	0.22	0.1%	6.50	1E-04	0.9%
	Age40	0.03	0.42	0.1%	0.07	0.42	0.0%	-0.73	0.26	0.1%	-5.56	0.01	0.4%
	HLA-A	0.06	0.01	0.5%	0.04	0.08	0.1%	0.69	0.34	0.0%	1.30	8E-03	0.3%
	HLA-B	0.06	0.03	0.3%	0.15	6E-13	2.7%	0.00	0.50	0.0%	3.12	6E-11	2.5%
	HLA-C	0.00	0.50	0.0%	0.05	0.22	0.0%	0.00	0.50	0.0%	0.97	0.16	0.1%
	Cohort	0.17	7E-08	2.6%	0.31	4E-16	3.4%	0.00	0.50	0.0%	6.38	<1E-16	5.8%
	N	1048		6.4%	1869		14.0%	1048		5.2%	1659		19.0%
Adaptation by Locus	**Adaptation-A**	**0.06**	**0.38**	**0.1%**	**0.27**	**9E-04**	**0.6%**	**-1.88**	**0.08**	**0.3%**	**-2.00**	**0.29**	**0.1%**
	Adaptation-B	**0.16**	**1E-03**	**1.0%**	**0.47**	**5E-13**	**2.8%**	**-5.61**	**2E-11**	**4.2%**	**-10.83**	**2E-13**	**3.2%**
	Adaptation-C	**0.01**	**0.93**	**0.0%**	**0.16**	**0.08**	**0.2%**	**-1.43**	**0.17**	**0.2%**	**-7.94**	**2E-04**	**0.8%**
	Female	-0.16	3E-03	0.9%	-0.29	8E-06	1.1%	1.24	0.17	0.2%	6.44	1E-04	0.9%
	Age40	0.03	0.43	0.1%	0.08	0.37	0.0%	-0.70	0.28	0.1%	-5.33	0.01	0.4%
	HLA-A	0.06	0.02	0.4%	0.04	0.08	0.1%	0.79	0.16	0.1%	1.40	3E-03	0.5%
	HLA-B	0.06	0.02	0.4%	0.15	9E-14	2.9%	0.00	0.50	0.0%	3.20	6E-12	2.7%
	HLA-C	0.00	0.50	0.0%	0.05	0.25	0.0%	0.14	0.50	0.0%	0.96	0.16	0.1%
	Cohort	0.17	9E-08	2.6%	0.31	1E-16	3.6%	0.00	0.50	0.0%	6.33	<1E-16	5.5%
	N	1048		6.7%	1869		14.4%	1048		6.2%	1659		19.6%
Adaptation by Protein	**Adaptation-Gag**	**0.13**	**0.51**	**0.1%**	**0.48**	**0.03**	**0.5%**	**-10.34**	**1E-03**	**1.5%**	**-2.52**	**0.69**	**0.0%**
	Adaptation-Pol	**0.00**	**0.99**	**0.0%**	**0.69**	**1E-04**	**1.5%**	**-11.42**	**3E-04**	**2.0%**	**-27.89**	**5E-08**	**3.8%**
	Adaptation-Nef	**0.00**	**0.98**	**0.0%**	**0.23**	**0.35**	**0.1%**	**-3.71**	**0.15**	**0.3%**	**-12.20**	**0.06**	**0.5%**
	Adaptation-Acc	**0.25**	**0.24**	**0.2%**				**-1.69**	**0.65**	**0.0%**			
	Adaptation-Gp41	**0.06**	**0.88**	**0.0%**				**-18.57**	**4E-03**	**1.4%**			
	Female	-0.15	0.03	0.8%	-0.36	9E-07	2.5%	1.23	0.29	0.2%	6.68	7E-03	1.0%
	Age40	0.04	0.44	0.1%	0.09	0.34	0.1%	0.20	0.81	0.0%	-3.24	0.22	0.2%
	HLA-A	0.04	0.09	0.3%	0.01	0.48	0.0%	0.00	0.23	0.1%	1.03	0.16	0.1%
	HLA-B	0.05	0.05	0.4%	0.12	3E-05	1.7%	0.31	0.23	0.1%	3.19	3E-05	2.0%
	HLA-C	0.00	0.50	0.0%	0.00	0.50	0.0%	0.00	0.50	0.0%	0.63	0.42	0.0%
	Cohort	0.13	6E-04	1.7%	0.12	3E-03	0.8%	0.00	0.23	0.1%	8.62	1E-13	6.8%
	N	609		4.4%	962		10.0%	609		7.8%	764		18.7%
Conditional on VL or CD4	CD4/LogVL**	-0.01	6E-14	5.3%	-0.02	<1E-16	16.9%	-3.82	2E-14	5.4%	-8.47	<1E-16	17.0%
	Adaptation	**0.14**	**0.11**	**0.2%**	**0.58**	**5E-06**	**1.4%**	**-9.04**	**1E-10**	**3.9%**	**-12.08**	**2E-06**	**1.5%**
	Female	-0.14	7E-03	0.7%	-0.05	0.52	0.0%	0.35	0.69	0.0%	5.55	6E-04	0.8%
	Age40	0.02	0.58	0.0%	0.00	1.00	0.0%	-0.59	0.35	0.1%	-2.73	0.22	0.1%
	HLA-A	0.05	0.07	0.2%	0.04	0.09	0.1%	0.38	0.50	0.0%	0.58	0.19	0.1%
	HLA-B	0.05	0.04	0.3%	0.11	2E-06	1.4%	0.00	0.50	0.0%	1.95	1E-04	0.9%
	HLA-C	0.01	0.50	0.0%	0.02	0.45	0.0%	0.25	0.50	0.0%	0.66	0.27	0.0%
	Cohort	0.16	1E-07	2.5%	0.30	<1E-16	5.2%	0.00	0.50	0.0%	6.92	<1E-16	8.4%
	N	1048		11.3%	1504		29.4%	1048		10.4%	1504		34.1%
All Measured Features	**Adaptation**	**0.40**	**0.01**	**1.9%**	**0.83**	**1E-03**	**2.7%**	**-7.44**	**5E-03**	**2.4%**	**-15.98**	**4E-03**	**2.4%**
	Female	-0.17	0.08	1.0%	-0.15	0.14	0.6%	-0.23	0.89	0.0%	2.83	0.18	0.5%
	Age40	-0.04	0.56	0.1%				1.43	0.20	0.5%			
	Time††	0.00	0.99	0.0%				-2.90	0.03	1.5%			
	vRC§§	0.30	0.21	0.5%	1.77	3E-05	4.6%	-13.29	9E-04	3.4%	-28.20	2E-03	2.7%
	Protein responses ‡‡				0.09	6E-07	6.1%				1.39	8E-04	2.8%
	OLP ¶¶				0.08	0.04	0.8%				0.67	0.39	0.0%
	HLA-A	0.03	0.35	0.0%	0.06	0.24	0.1%	0.29	0.50	0.0%	1.70	0.09	0.5%
	HLA-B	0.00	0.50	0.0%	0.10	0.06	0.7%	0.00	0.50	0.0%	1.77	0.05	0.8%
	HLA-C	0.02	0.41	0.0%	0.00	0.50	0.0%	0.00	0.50	0.0%	0.01	0.50	0.0%
	Cohort	0.20	1E-04	4.1%				1.55	0.08	0.6%			
	N	325		8.7%	372		20.8%	325		9.7%	356		11.2%

*Log10 VL as dependent variable. †Box-Cox tranformed CD4 counts as dependent variable. ‡Random effects in italics; estimate is the fixed effect size or the standard deviation for random effects; pseudo-R2 is estimated using the likelihood ratio method. §Indicator variable for age ≥ 40. ||Year of sampling for British Columbia, treated as a categorical variable; the source cohort for Southern Africa. ¶Total number of observations for this model. The R2 for this row is that for the entire model, which may be greater than the sum of the individual terms, as the likelihood ratio approach tends to under-estimate the fraction of variation explained when features are not independent. **Box-Cox transformed CD4 counts were used as covariates for the VL models; log10 VL was used as a covariate for the CD4 models. ††Log10 transformed estimated days since infection, as estimated by clinician-patient interviews. ‡‡in vitro Gag-Protease viral replicative capacity, measured for a subset of the Durban cohort. §§Total number of OLP responses to each protein. OLP measured as whole proteome 18mers overlapping by 10, for a subset of the Durban cohort. ¶¶Random effects design matrix where each column represents a response to a particular OLP.

Supplementary Table 2. Transmitted adaptation predicts VL.

	Feature§	Mean VL (30-365d)* estimate	p	R^2	Mean VL (>365d)† estimate	p	R^2
Zambian Txn pairs	**Donor-to-recipient Adaption‖**	**0.96**	**4E-04**	**10.7%**	**1.28**	**9E-04**	**14.6%**
	Among male recipients (n=52, 36)¶	1.21	5E-04	20.1%	0.81	0.04	13.0%
	Among female recipients (n=61, 40)	0.67	0.11	4.5%	1.63	6E-03	20.5%
	HLA-A	0.35	0.03	4.3%	0.47	0.03	6.7%
	HLA-B	0.32	0.02	5.5%	0.42	0.02	7.8%
	HLA-C	0.26	0.22	1.5%	0.36	0.20	2.1%
	Gag	0.99	0.03	3.9%	0.25	0.70	0.2%
	Pol	0.65	0.05	3.9%	1.62	9E-04	15.4%
	Nef	0.67	0.18	2.0%	0.79	0.23	1.9%
	Recipient Is Female	-0.57	2E-06	19.4%	-0.35	0.03	6.6%
	Recipient Age (≥40)	-0.10	0.62	0.4%	-0.23	0.31	1.4%
	Transmission year (≥median)	-0.12	0.27	0.9%	-0.23	0.13	3.4%
	Founder virus *gag* vRC	0.06	0.38	0.7%	0.06	0.48	0.8%
	Donor VL**	0.19	0.02	3.9%	-0.09	0.42	1.0%
	HLA-A	0.00	0.50	0.0%	0.05	0.48	0.0%
	HLA-B	0.00	0.50	0.0%	0.10	0.15	1.4%
	HLA-C	0.07	0.50	0.0%	0.00	0.50	0.0%
	N††	113		27.4%	76		22.1%
Step trial	**Founder adaptation§§**	**1.46**	**1E-03**	**18.1%**	**1.08**	**0.09**	**14.2%**
	HLA-A	0.75	8E-03	13.4%	0.22	0.62	1.5%
	HLA-B	0.64	7E-03	13.7%	0.45	0.20	9.6%
	HLA-C	0.19	0.53	1.0%	0.82	0.11	14.6%
	Gag	0.56	0.49	1.0%	-0.67	0.61	1.5%
	Pol	3.05	4E-04	22.0%	1.93	0.15	12.0%
	Nef	0.29	0.65	0.5%	1.22	0.27	7.0%
	Age (≥40)	-0.24	0.46	1.1%	0.14	0.76	0.5%
	Race (white)¶¶	0.25	0.22	1.8%	-0.09	0.75	0.5%
	HLA-A	0.00	0.50	0.0%	0.00	0.50	0.0%
	HLA-B	0.15	0.26	0.7%	0.00	0.50	0.0%
	HLA-C	0.02	0.50	0.0%	0.00	0.50	0.0%
	N	56		20.5%	23		16.6%

*Dependent variable was mean Log_{10} VL from 30-365d post infection. †Dependent variable was mean Log_{10} VL over samples taken after 365d post infection. §Random effects in italics; estimate is the fixed effect size or the standard deviation for random effects; pseudo-R^2 is estimated using the likelihood ratio method. ‖Adaptation of the donor's virus to the recipient's HLA alleles. ¶Features in inset represent independent runs in which autologous adaptation was replaced by allele-specific or protein-specific adaptation, or the subjects were stratified by sex. **VL of the donor, measured at the same time point as that used for sequencing (median 46d post transmission). ††Features in italics were treated as random effects. ††Total number of observations for this model. The R^2 for this row is that for the entire model. §§Mean adaptation over all amplicons sequenced from the first available sample. ¶¶Indicator variable, as specificed in the Step study trial design.

Supplementary Table 3. Circulating adaptation predicts chronic VL and CD4 counts.

		Viral Load*						CD4†					
		British Columbia (HIVB)			Southern Africa (HIVC)			British Columbia (HIVB)			Southern Africa (HIVC)		
	Feature‡	estimate	p	R^2	estimate	p	R^2	estimate	p	R^2	estimate	p	R^2
Autologous and Circulating Adaptation	**Autologous**	**0.25**	**4E-03**	**0.8%**	**1.00**	**<1E-16**	**3.9%**	**-9.96**	**5E-12**	**4.3%**	**-23.32**	**2E-16**	**4.2%**
	Circulating	**0.79**	**1E-04**	**1.2%**	**1.22**	**8E-04**	**0.5%**	**-4.49**	**0.13**	**0.2%**	**-24.34**	**7E-03**	**0.4%**
	Female	-0.15	5E-03	0.8%	-0.33	4E-07	1.4%	1.09	0.23	0.1%	6.79	7E-05	1.0%
	Age40§	0.03	0.45	0.1%	0.06	0.51	0.0%	-0.72	0.27	0.1%	-5.36	0.01	0.4%
	HLA Frequency ‖	0.18	0.46	0.0%	0.00	0.50	0.0%	0.00	0.35	0.0%	0.02	0.50	0.0%
	HLA-A	0.03	0.10	0.2%	0.04	0.22	0.0%	0.67	0.24	0.0%	1.35	0.01	0.3%
	HLA-B	0.03	0.17	0.1%	0.12	1E-08	1.7%	0.00	0.35	0.0%	2.59	2E-05	1.0%
	HLA-C	0.02	0.34	0.0%	0.00	0.50	0.0%	0.00	0.50	0.0%	0.64	0.32	0.0%
	Cohort ¶	0.17	1E-07	2.5%	0.16	3E-07	1.4%	0.00	0.35	0.0%	5.99	<1E-16	4.8%
	N**	1048		7.7%	1792		13.0%	1048		5.5%	1572		19.6%
Adaptation by Locus	**Adaptation-A**	**0.06**	**0.33**	**0.1%**	**0.26**	**1E-03**	**0.6%**	**-1.81**	**0.10**	**0.3%**	**-2.16**	**0.26**	**0.1%**
	Adaptation-B	**0.16**	**9E-04**	**1.0%**	**0.49**	**8E-14**	**3.1%**	**-5.68**	**1E-11**	**4.3%**	**-11.28**	**7E-14**	**3.5%**
	Adaptation-C	**-0.01**	**0.83**	**0.0%**	**0.14**	**0.12**	**0.1%**	**-1.13**	**0.28**	**0.1%**	**-7.53**	**5E-04**	**0.8%**
	Circulating-A	**0.36**	**4E-03**	**0.8%**	**0.22**	**0.23**	**0.1%**	**-1.84**	**0.34**	**0.1%**	**-4.20**	**0.39**	**0.1%**
	Circulating-B	**0.36**	**2E-03**	**0.7%**	**0.77**	**1E-03**	**0.5%**	**-0.82**	**0.63**	**0.0%**	**-20.87**	**4E-05**	**0.7%**
	Circulating-C	**0.08**	**0.45**	**0.1%**	**0.44**	**0.06**	**0.2%**	**-2.74**	**0.09**	**0.3%**	**-2.95**	**0.64**	**0.0%**
	Female	-0.15	6E-03	0.7%	-0.33	4E-07	1.4%	1.25	0.17	0.2%	6.86	6E-05	1.0%
	Age40	0.03	0.46	0.1%	0.06	0.48	0.0%	-0.68	0.29	0.1%	-5.04	0.02	0.3%
	HLA Frequency	0.40	0.06	0.2%	0.00	0.50	0.0%	0.00	0.34	0.0%	0.04	0.50	0.0%
	HLA-A	0.04	0.10	0.2%	0.04	0.26	0.0%	0.74	0.28	0.0%	1.46	4E-03	0.4%
	HLA-B	0.01	0.50	0.0%	0.11	1E-07	1.5%	0.17	0.50	0.0%	1.97	2E-03	0.5%
	HLA-C	0.00	0.50	0.0%	0.00	0.50	0.0%	0.00	0.34	0.0%	0.84	0.27	0.0%
	Cohort	0.16	2E-07	2.4%	0.17	3E-08	1.6%	0.00	0.34	0.0%	5.86	<1E-16	4.3%
	N	1048		8.5%	1792		13.6%	1048		6.6%	1572		20.6%
Adaptation by Protein	**Adaptation-Gag**	**0.09**	**0.65**	**0.0%**	**0.52**	**0.02**	**0.6%**	**-9.53**	**4E-03**	**1.3%**	**-3.54**	**0.57**	**0.0%**
	Adaptation-Pol	**-0.05**	**0.79**	**0.0%**	**0.69**	**1E-04**	**1.5%**	**-11.56**	**5E-04**	**1.8%**	**-28.11**	**3E-08**	**4.0%**
	Adaptation-Nef	**0.00**	**0.98**	**0.0%**	**0.21**	**0.39**	**0.1%**	**-3.74**	**0.16**	**0.3%**	**-11.16**	**0.08**	**0.4%**
	Adaptation-Acc	**0.20**	**0.34**	**0.1%**				**-1.36**	**0.72**	**0.0%**			
	Adaptation-Gp41	**0.11**	**0.76**	**0.0%**				**-18.96**	**3E-03**	**1.3%**			
	Circulating-Gag	**1.19**	**3E-04**	**2.1%**	**1.27**	**0.02**	**0.5%**	**-7.52**	**0.19**	**0.3%**	**-46.79**	**3E-03**	**1.1%**
	Circulating-Pol	**0.66**	**0.18**	**0.4%**	**0.31**	**0.60**	**0.0%**	**0.33**	**0.97**	**0.0%**	**-19.93**	**0.28**	**0.2%**
	Circulating-Nef	**-0.18**	**0.67**	**0.0%**	**0.33**	**0.68**	**0.0%**	**1.51**	**0.84**	**0.0%**	**-25.50**	**0.21**	**0.2%**
	Circulating-Acc	**0.00**	**0.99**	**0.0%**				**-2.65**	**0.69**	**0.0%**			
	Circulating-Gp41	**0.02**	**0.97**	**0.0%**				**-10.56**	**0.43**	**0.1%**			
	Female	-0.15	0.02	0.9%	-0.37	5E-07	2.6%	1.22	0.30	0.2%	6.96	5E-03	1.1%
	Age40	0.03	0.47	0.1%	0.10	0.31	0.1%	0.22	0.78	0.0%	-3.03	0.25	0.2%
	HLA Frequency	0.00	0.50	0.0%	0.00	0.50	0.0%	0.00	0.50	-0.1%	0.05	0.50	0.0%
	HLA-A	0.00	0.50	0.0%	0.00	0.50	0.0%	0.00	0.50	-0.1%	0.00	0.50	0.0%
	HLA-B	0.03	0.48	0.0%	0.09	1E-03	1.0%	0.33	0.50	0.0%	1.52	0.11	0.2%
	HLA-C	0.00	0.50	0.0%	0.00	0.50	0.0%	0.46	0.50	-0.1%	0.97	0.26	0.1%
	Cohort	0.12	1E-03	1.5%	0.13	1E-03	1.0%	0.00	0.50	0.0%	8.31	2E-12	6.1%
	N	609		7.5%	962		10.6%	609		8.3%	764		20.5%
All Measured Features	**Autologous**	**0.11**	**0.25**	**0.2%**	**0.86**	**1E-03**	**2.9%**	**-8.14**	**7E-07**	**3.3%**	**-16.02**	**4E-03**	**2.5%**
	Circulating	**0.54**	**0.01**	**0.8%**	**1.76**	**2E-03**	**2.3%**	**-7.56**	**0.02**	**0.5%**	**-12.67**	**0.36**	**0.4%**
	Female	-0.20	2E-03	1.3%	-0.15	0.15	0.6%	1.04	0.33	0.1%	2.79	0.19	0.5%
	Age40	0.04	0.40	0.1%				-0.38	0.60	0.0%			
	vRC††	0.47	1E-03	1.4%	1.78	2E-05	4.7%	-14.89	9E-10	4.8%	-29.60	1E-03	2.8%
	Protein responses ‡‡				0.08	5E-07	6.2%				1.38	8E-04	2.8%
	OLP §§				0.08	0.04	0.8%				0.01	0.47	0.0%
	HLA Frequency	0.27	0.23	0.1%	0.00	0.50	0.0%	0.00	0.50	0.0%	11.04	0.50	0.0%
	HLA-A	0.04	0.08	0.3%	0.03	0.50	0.0%	0.71	0.28	0.0%	1.67	0.11	0.4%
	HLA-B	0.00	0.50	0.0%	0.05	0.36	0.0%	0.28	0.50	-0.1%	1.27	0.17	0.3%
	HLA-C	0.00	0.50	0.0%	0.00	0.50	0.0%	0.00	0.50	-0.1%	0.01	0.50	0.0%
	Cohort	0.16	9E-06	2.4%				0.00	0.50	0.0%			
	N	749		8.2%	372		22.6%	749		9.4%	356		11.5%
All Measured Features Conditional on VL or CD4	CD4/LogVL¶¶	-0.02	3E-14	7.4%	-0.02	2E-11	11.9%	-4.45	4E-14	7.5%	-7.89	3E-14	12.9%
	Autologous	**-0.01**	**0.90**	**0.0%**	**0.48**	**0.06**	**1.0%**	**-7.53**	**2E-06**	**2.9%**	**-9.64**	**0.06**	**0.9%**
	Circulating	**0.44**	**0.04**	**0.5%**	**1.47**	**7E-03**	**2.1%**	**-4.36**	**0.19**	**0.3%**	**4.16**	**0.75**	**0.0%**
	Female	-0.18	4E-03	1.1%	-0.05	0.58	0.1%	-0.04	0.97	0.0%	2.09	0.30	0.3%
	Age40	0.03	0.46	0.1%				-0.18	0.80	0.0%			
	vRC	0.23	0.11	0.3%	1.38	8E-04	3.2%	-12.96	4E-08	4.0%	-16.19	0.06	0.8%
	Protein responses				0.06	5E-05	4.2%				0.28	0.41	0.0%
	OLP				0.07	0.08	0.6%				0.00	0.50	0.0%
	HLA Frequency	0.28	0.23	0.1%	0.01	0.50	0.0%	0.00	0.15	0.1%	16.39	0.41	0.0%
	HLA-A	0.03	0.13	0.2%	0.00	0.50	0.0%	0.58	0.36	0.0%	1.20	0.24	0.1%
	HLA-B	0.00	0.50	0.0%	0.06	0.35	0.0%	0.31	0.50	0.0%	0.74	0.42	0.0%
	HLA-C	0.02	0.41	0.0%	0.00	0.50	0.0%	0.32	0.35	0.0%	0.50	0.44	0.0%
	Cohort	0.15	9E-06	2.4%				0.00	0.15	0.1%			
	N	749		15.0%	356		31.2%	749		16.2%	356		22.9%

*Log10 VL as dependent variable. †Box-Cox tranformed CD4 counts as dependent variable. ‡Random effects in italics; estimate is the fixed effect size or the standard deviation for random effects; pseudo-R2 is estimated using the likelihood ratio method. §Indicator variable for age ≥ 40. ‖The mean frequency of the 2-digit HLA was computed for each locus. The three resulting features were then treated as random effects. ¶Year of sampling for British Columbia, treated as a categorical variable; the source cohort for Southern Africa. **Total number of observations for this model. The R2 for this row is that for the entire model, which may be greater than the sum of the individual terms, as the likelihood ratio approach tends to under-estimate the fraction of variation explained when features are not independent. ††in vitro Gag-Protease viral replicative capacity, measured for a subset of the Durban cohort. ‡‡Total number of OLP responses to each protein. OLP measured as whole proteome 18mers overlapping by 10, for a subset of the Durban cohort. §§Random effects design matrix where each column represents a response to a particular OLP. ¶¶Box-Cox transformed CD4 counts were used as covariates for the VL models; log10 VL was used as a covariate for the CD4 models.

Supplementary Table 4. Heritability estimate is influenced by HLA-B similarity.

Covariate*	Estimate	SE	tStat	p
(Intercept)	4.84	0.09	54.31	<1E-16
Recipient is Female	-0.49	0.10	-5.08	7E-07
Recipient Age (≥40)	0.06	0.18	0.35	0.73
Transmission Year (≥ Median)	0.03	0.09	0.30	0.77
Donor VL	0.13	0.05	2.73	7E-03
HLA-B Similarity†	**0.15**	**0.05**	**3.17**	**2E-03**
Donor VL : HLA-B Similarity§	**0.13**	**0.05**	**2.63**	**9E-03**

* A linear model was fitted to recipient VL (mean over all observations 30-365d post infection) for N=275 linked transmission pairs from the Zambian cohort. All independent variables were standardized (zero mean, unit variance). †The adaptation similarity between donor and recipient HLA-B alleles. Adaptation similarity is defined to be the Pearson correlation coefficient of two sets of alleles over the reference panel of all Southern Africa HIV sequences. §An interaction term between Donor VL and HLA-B similarity.