

Impact of pre-adapted HIV transmission

Jonathan M Carlson¹, Victor Y Du^{2,28}, Nico Pfeifer^{1,27,28}, Anju Bansal², Vincent Y F Tan^{1,27}, Karen Power³, Chanson J Brumme⁴, Anat Kreimer^{1,27}, Charles E DeZiel¹, Nicolo Fusi¹, Malinda Schaefer⁵, Mark A Brockman^{4,6}, Jill Gilmour^{7,8}, Matt A Price^{7,9}, William Kilembe¹⁰, Richard Haubrich¹¹, Mina John^{12,13}, Simon Mallal^{12,14}, Roger Shapiro¹⁵, John Frater^{16–18}, P Richard Harrigan^{4,19}, Thumbi Ndung'u^{3,20–22}, Susan Allen^{10,23,24}, David Heckerman¹, John Sidney²⁵, Todd M Allen³, Philip J R Goulder^{20,26}, Zabrina L Brumme^{4,6}, Eric Hunter^{5,10,23} & Paul A Goepfert²

Human leukocyte antigen class I (HLA)-restricted CD8⁺ T lymphocyte (CTL) responses are crucial to HIV-1 control. Although HIV can evade these responses, the longer-term impact of viral escape mutants remains unclear, as these variants can also reduce intrinsic viral fitness. To address this, we here developed a metric to determine the degree of HIV adaptation to an HLA profile. We demonstrate that transmission of viruses that are pre-adapted to the HLA molecules expressed in the recipient is associated with impaired immunogenicity, elevated viral load and accelerated CD4⁺ T cell decline. Furthermore, the extent of pre-adaptation among circulating viruses explains much of the variation in outcomes attributed to the expression of certain HLA alleles. Thus, viral pre-adaptation exploits 'holes' in the immune response. Accounting for these holes may be key for vaccine strategies seeking to elicit functional responses from viral variants, and to HIV cure strategies that require broad CTL responses to achieve successful eradication of HIV reservoirs.

Immune control of HIV is epidemiologically linked to the expression of certain HLA alleles, which mediate control through the presentation of viral peptides to CTLs^{1,2}. The resulting suppression of viral replication induces strong evolutionary pressure that drives the selection of CTL escape mutations. These mutations may fully or partially abrogate viral peptide–HLA binding, disrupt peptide processing or alter peptide–HLA interactions with the T cell receptor³ (TCR). Within-host selection of escape mutations is thought to increase viral fitness by facilitating immune evasion, which should result in increased plasma viral load (VL) and accelerated CD4⁺ T cell decline. However, at least two factors work against the virus in this context. First, some escape mutations impair the ability of the virus to replicate^{4–10}. Second, the CTL response itself adapts to the changing

virus through the emergence of new TCR variants that either recognize the escaped epitope or shift focus to new epitopes^{11–15}. Indeed, although case studies have reported increased VL after escape from highly immunodominant epitopes^{11,16–19}, the overall impact of within-host escape is unknown.

Once selected, escape mutations are frequently transmitted^{7–9}, and they may be accumulating in some populations^{20–22}. Transmission of these escape variants to HLA-mismatched hosts has been linked to improved clinical outcomes owing to reduced intrinsic viral fitness^{7,8,10}, but the clinical consequences of transmission of viruses pre-adapted to the recipient's HLA profile is unknown. Although mutations that abrogate antigen processing and/or HLA binding may confer universal escape consequences in hosts expressing the relevant

¹Microsoft Research, Redmond, Washington, USA. ²Department of Medicine, University of Alabama at Birmingham, Birmingham, Alabama, USA. ³Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts, USA. ⁴British Columbia Centre for Excellence in HIV/AIDS, Vancouver, British Columbia, Canada. ⁵Emory Vaccine Center, Yerkes National Primate Research Center, Emory University, Atlanta, Georgia, USA. ⁶Faculty of Health Sciences, Simon Fraser University, Burnaby, British Columbia, Canada. ⁷International AIDS Vaccine Initiative, New York, New York, USA. ⁸Imperial College of Science, Technology and Medicine, London, UK. ⁹Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California, USA. ¹⁰Rwanda–Zambia HIV Research Group, Zambia–Emory HIV Research Project, Lusaka, Zambia. ¹¹Gilead Sciences, Foster City, California, USA. ¹²Institute for Immunology and Infectious Diseases, Murdoch University, Murdoch, Western Australia, Australia. ¹³Department of Clinical Immunology, Royal Perth Hospital, Perth, Western Australia, Australia. ¹⁴Center for Translational Immunology and Infectious Diseases, Vanderbilt University School of Medicine, Nashville, Tennessee, USA. ¹⁵Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA. ¹⁶Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK. ¹⁷National Institute of Health Research, Oxford Biomedical Research Centre, Oxford, UK. ¹⁸Institute for Emerging Infections, Oxford Martin School, University of Oxford, Oxford, UK. ¹⁹Department of Medicine, University of British Columbia, Vancouver, British Columbia, Canada. ²⁰HIV Pathogenesis Programme, the Doris Duke Medical Research Institute, University of KwaZulu-Natal, Durban, South Africa. ²¹KwaZulu-Natal Research Institute for Tuberculosis and HIV (K-RITH), Nelson R. Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa. ²²Max Planck Institute for Infection Biology, Berlin, Germany. ²³Department of Pathology and Laboratory Medicine, Emory University, Atlanta, Georgia, USA. ²⁴Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, Georgia, USA. ²⁵Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, La Jolla, California, USA. ²⁶Department of Paediatrics, University of Oxford, Oxford, UK. ²⁷Present addresses: Department of Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Saarbrücken, Germany (N.P.); Department of Electrical and Computer Engineering, National University of Singapore, Singapore (V.Y.F.T.); Department of Mathematics, National University of Singapore, Singapore (V.Y.F.T.); Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, California, USA (A.K.). ²⁸These authors contributed equally to this work. Correspondence should be addressed to J.M.C. (carlson@microsoft.com), E.H. (ehunte4@emory.edu) or P.A.G. (paulg@uab.edu).

HLA allele^{19,23}, TCR escape mutations can retain immunogenicity in subsequent hosts^{1,17,24,25}, and the loss of some epitopes in the founder virus may simply result in targeting of other epitopes¹².

Resolving the role of transmitted escape in HIV progression is central to both vaccine design and epidemiology. A leading hypothesis as to why T cell vaccines based on whole-protein immunogens have failed to reduce postinfection VL is that they have not adequately accounted for the role of immune escape and viral diversity²⁶. Alternative vaccine strategies have thus emerged. One aims to focus the immune response on relatively conserved HIV regions ('conserved element vaccines')^{27–30}, whereas another aims to stimulate variant-specific responses by incorporating multiple immunogens that reflect circulating viral diversity ('polyvalent vaccines')³¹. A key assumption of these strategies—and of the polyvalent approach in particular—is that effective immune responses can be elicited against epitope variants, including those representing HLA-specific escape mutations. This assumption, however, conflicts with concerns that the stable transmission and accumulation of CTL escape mutations at the population level will gradually compromise host immunity and result in increased HIV virulence as the pandemic progresses²⁰. Such concerns assume escape variants are universally nonimmunogenic and carry low fitness costs. Furthermore, efforts to quantify the extent to which VL is 'heritable' (i.e., determined by the viral sequence) make critical simplifying assumptions, such as that viral and host genetics act independently on VL and that escaped epitopes are nonimmunogenic^{32,33}. Thus, fundamental working theories on HIV pathogenesis and vaccine design currently operate on strong—and often opposing—assumptions regarding the impact of transmitted immune escape.

RESULTS

Estimating viral adaptation to HLA

The complexity of escape has prevented in-depth study of the clinical consequences of transmitted and within-host escape. Although escape mutations can be predicted remarkably well on the basis of HLA subtype, there is a strong stochastic component to both CTL targeting³⁴ and escape selection³. We therefore sought to reduce the complexity of escape to a single metric, which we call "adaptation". Adaptation to a particular HLA allele h is rooted in a probabilistic model that compares what an HIV sequence would 'look like' were it to evolve indefinitely in a host whose immune system either (1) solely targeted epitopes restricted by h or (2) did not target any HLA-restricted epitopes. We then wrote the adaptation of a particular sequence s to

h as $\text{Adapt}_h(s) = g((\text{Pr}(s|h))/(\text{Pr}(s|\emptyset)))$, where $\text{Pr}(s|h)$ captures scenario 1, $\text{Pr}(s|\emptyset)$ captures scenario 2 and $g(\cdot)$ scales the ratio so that it is symmetric on the interval from -1 to 1 , such that the extremes respectively indicate that the sequence would never be observed or would be observed only in an individual expressing that allele.

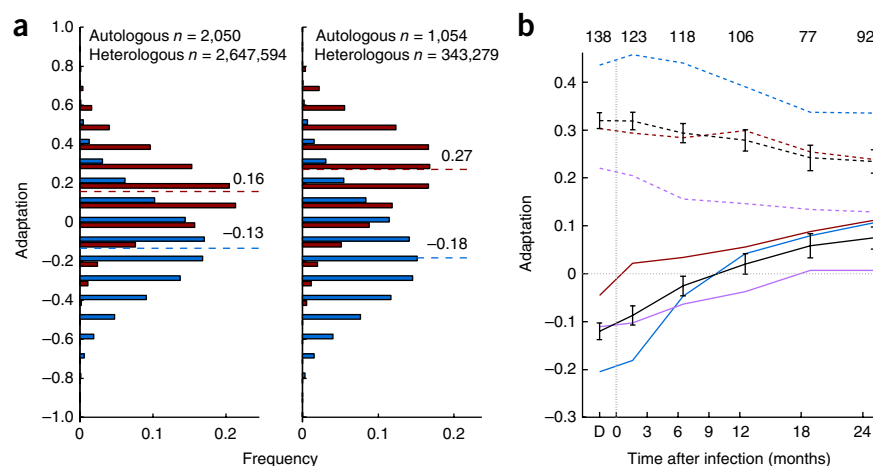
We defined four types of scores: (1) 'autologous adaptation' compares the autologous viral sequence to an individual's alleles; (2) 'heterologous adaptation' compares a nonautologous virus to an individual's alleles; (3) 'circulating adaptation' is the average heterologous adaptation over all viruses in a cohort with respect to an individual's alleles; and (4) 'transmitted adaptation' is the autologous adaptation of an individual's founder virus. These scores can be defined with respect to a single HLA allele ('allele-specific adaptation') or to an individual's HLA repertoire (the average over the individual's allele-specific scores). Further, the adaptation similarity of two alleles (or individuals) is the Pearson correlation coefficient of their respective scores over a panel of heterologous viral sequences. Adaptation can be defined with respect to each viral protein, but it is unclear whether adaptation scores are comparable among proteins (**Supplementary Note 1**).

Estimation of adaptation requires estimation of the conditional probability distribution $\text{Pr}(s|h)$. To this end, we extended the phylogenetic logistic regression framework³⁵ to allow estimation of the probability of observing any amino acid, at any site, conditional on any set of HLA alleles (Online Methods; an implementation is available at <https://phylogd.research.microsoft.com>). We trained two separate models for HIV adaptation on the basis of the availability of linked HLA and sequence data for chronically infected, untreated individuals. The HIV-1 subtype B (HIVB) model was trained on the International HIV Adaptation Collaborative (IHAC) cohort³⁶, which consists of 1,888 individuals from North America and Australia with sequences from all HIV proteins except gp120. The HIV-1 subtype C (HIVC) model was trained on a set of cohorts from southern Africa⁹ consisting of 2,037 individuals with Gag, Pol and Nef sequences. **Supplementary Figure 1** presents a synopsis of the data sets used in this work.

As expected, autologous adaptation was substantially higher than heterologous adaptation (**Fig. 1a** and **Supplementary Fig. 2a**), and mean autologous adaptation increased during the first 2 years of infection and beyond (**Fig. 1b** and **Supplementary Fig. 2b,c**). These results indicate that adaptation is a measure of subject-specific viral variation. Nevertheless, there was substantial overlap between autologous and heterologous adaptation, which indicates that some individuals will by chance be infected by a virus that is pre-adapted to their HLA alleles.

Figure 1 Adaptation of viral sequence to HLA-I alleles.

(a) The distribution of adaptation scores computed for all viruses against all HLA profiles in Southern Africa (left) and British Columbia (right). Adaptation of autologous virus to host HLA (autologous adaptation) is shown in red; adaptation of a virus sequence to a different host's HLA (heterologous adaptation) is shown in blue. Median scores for each distribution are indicated (dashed lines). (b) Autologous adaptation for linked transmission pairs from Zambia. Colors indicate adaptation with respect to recipients' (solid lines) or donors' (dashed lines) HLA-A (red), -B (blue) and -C (purple) alleles or entire repertoire (black). Error bars indicate 95% confidence intervals. The number of samples at each time point is indicated at the top of the graph. Adaptation of linked donor sequence (time point 'D') is set to -50 d for display.



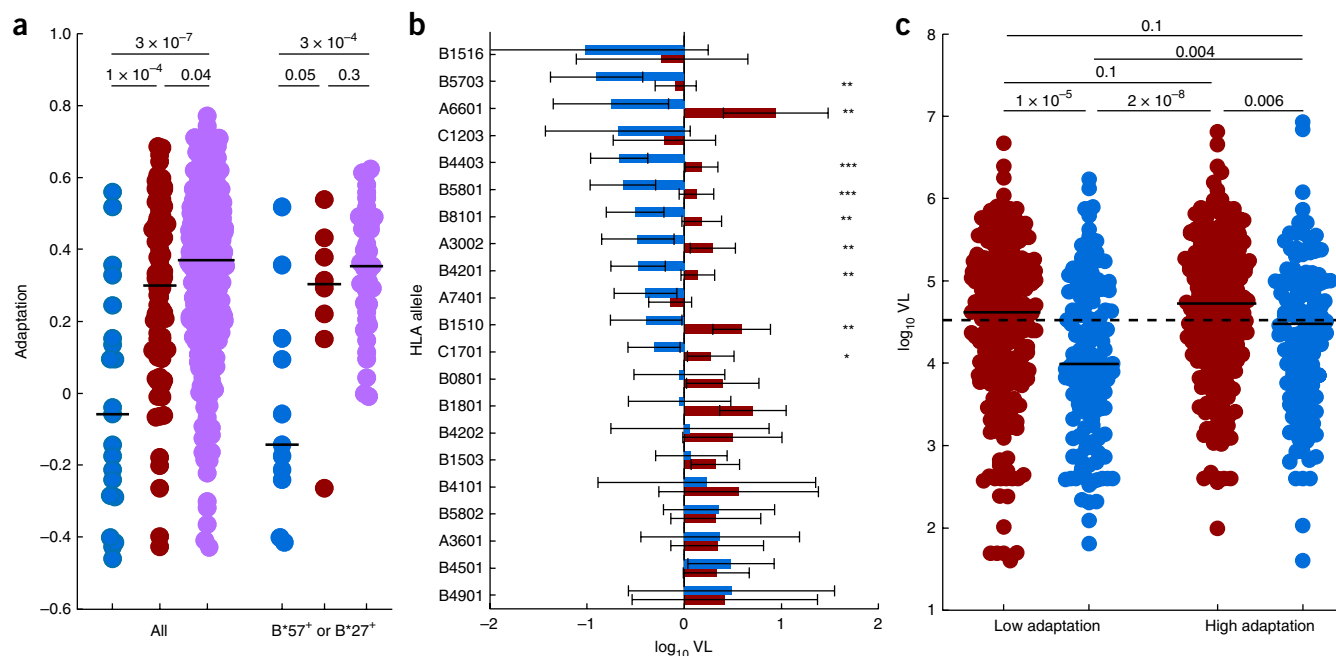


Figure 2 Autologous adaptation predicts faster disease progression. **(a)** Adaptation in controllers (VL < 50 copies per ml; blue, $n = 21$) and noncontrollers (red, $n = 80$, Ragon cohort; purple, $n = 383$, British Columbia cohort with no missing sequence data). Right, data for individuals who express B*57 or B*27 ($n = 11$, 8 and 41 for the three cohorts, respectively). P values (top) determined via two-tailed Mann–Whitney U -test. **(b)** Estimated HLA-specific effects on VL in the Southern Africa cohort. Estimated VL (error bars, 95% CI) relative to cohort average for individuals expressing the allele with no (blue) or with complete (red) allele-specific adaptation. Significant adaptation effects are indicated (*** $P < 0.001$, ** $P < 0.01$ and * $P < 0.05$, likelihood ratio test). **(c)** VL for each of $n = 691$ HIVC-infected subjects from Durban, stratified by Gag-specific adaptation and OLP response breadth (above versus below population averages). Red, below-average (blue, above-average) OLP responses; solid horizontal lines, stratum medians; dashed line, cohort median. P values (top) determined via two-tailed Mann–Whitney U -test (primary and interaction effects remain significant at $P = 0.02$ when treated as continuous variables in a mixed model).

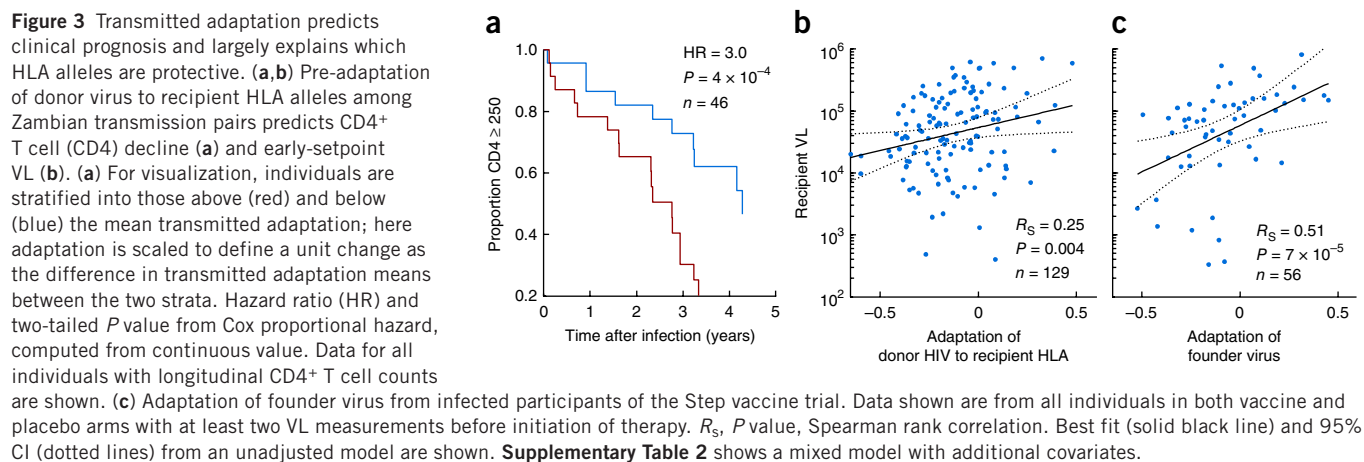
Within-host adaptation accelerates disease progression

If within-host adaptation in the context of a robust CTL response drives pathogenesis, then our measure of autologous adaptation should correlate with clinical markers of disease progression. Consistent with prior reports³⁷, we observed significantly lower levels of autologous adaptation in HIVB-infected controllers than in noncontrollers (Fig. 2a). This pattern held across all HLA loci and proteins, as well as among individuals expressing protective alleles (Supplementary Fig. 3a,b). Similarly, among 2,917 chronically infected noncontrollers, autologous adaptation was the most important predictor of both VL and CD4⁺ T cell counts (Supplementary Table 1). This result was consistent across HIV subtypes and statistical models, and it persisted when host and viral covariates were added to the models.

Critically, allele-specific autologous adaptation completely abrogated the protection attributable to each HLA allele (Fig. 2b and Supplementary Figs. 3c and 4), including alleles for which multiple escape mutations are known to carry substantial *in vitro* fitness costs^{5,6}. This result indicates that the benefit the virus receives from evading the CTL responses outweighs any reduction in intrinsic fitness, and it suggests either that the majority of escape mutations have a negligible impact on intrinsic fitness or that any such reduction is typically compensated for by other mutations. Indeed, there was no clear association between autologous Gag adaptation and *in vitro* viral gag-protease replicative capacity (vRC, measured as the relative growth rate of NL4-3 recombinant viruses encoding subject-derived gag-protease sequences) over all HLA alleles (Supplementary Fig. 5a,b) or among protective alleles (Supplementary Fig. 5c). This is consistent with the disappearance

of any association between protective alleles and vRC over the course of chronic infection^{38,39}.

The conserved element vaccine strategy targets epitopes believed to be relatively resistant to escape, under the assumption that robust CTL responses in the absence of escape are crucial for control^{27–30}. To test this hypothesis using our metric for adaptation, we measured Gag-specific CTL responses among 691 HIVC-infected individuals from Durban using 18-mer overlapping peptides (OLPs) based on the subtype C consensus. We then stratified these individuals by both Gag-specific autologous adaptation and response breadth. This stratification demonstrated that the reduction in VL associated with a broad Gag-specific CTL response is observed primarily among individuals whose virus has not adapted to that response (Fig. 2c). Indeed, high levels of adaptation nearly eliminated the benefit of targeting Gag. In the absence of a CTL response, adaptation was not associated with changes in VL (Fig. 2c). This suggests that escape primarily influences VL by reducing effective immune responses, not by reducing intrinsic viral fitness. Critically, the lowest VLs were observed among individuals who broadly targeted Gag yet harbored low levels of autologous adaptation. These individuals seemed to be mounting a robust CTL response, associated with a substantial reduction of VL, but with limited selection of escape variants. These observations support the idea of protective responses as those that broadly target ‘difficult-to-escape’ epitopes^{2,28}, which in turn directly supports vaccine strategies that aim to elicit such responses^{27–29}. We did not observe an interaction between adaptation and the number of OLP-eliciting responses in Pol or Nef; however, Pol adaptation was positively correlated with VL (Supplementary Fig. 6). In contrast



with responses to Gag, CTL responses against Pol and Nef have not consistently been linked with viral control⁴⁰.

The ability to define a single metric for adaptation allows us to address the question of whether high levels of autologous adaptation are predictive of future disease progression or simply the result of high virus replication. To this end, we applied an autoregression model to longitudinal VL and sequence samples from the Zambian transmission-pair cohort⁹ to test whether autologous adaptation can be used to predict future changes in VL. VL at each time point was modeled as a function of the prior two VL measurements and adaptation at the previous VL measurement, with additional clinical covariates. On average, one s.d. difference in autologous adaptation predicted an additional 0.13 log increase in VL ($P < 0.001$), whereas VL did not significantly predict subsequent changes in adaptation (**Supplementary Fig. 7**). Thus, these longitudinal data are consistent with adaptation (on average) driving subsequent changes in VL, not vice versa.

Transmitted adaptation predicts accelerated disease progression

The majority of amino acid variants present in the donor consensus sequence are transmitted to the recipient⁹. Although some of these variants have been linked to lower VL in HLA-mismatched recipients owing to presumed reduced intrinsic viral fitness^{7,8}, if the variants are adapted with respect to the recipient's HLA alleles, they have the potential to undermine the host immune response⁴¹.

We therefore measured the extent to which the donor HIV Gag, Pol and Nef sequences were by chance pre-adapted to the recipients' HLA alleles in 129 HIVC-infected, epidemiologically linked Zambian transmission pairs⁹. The extent of transmitted adaptation was associated with an increased rate of CD4⁺ T cell decline (**Fig. 3a**) and was correlated with recipient VL (**Fig. 3b**). Overall, transmitted adaptation explained more variation in recipient VL (both early (<12 months after infection) and late) than did HLA alleles, vRC, donor VL, age or (for late VL) sex (**Supplementary Table 2**).

To confirm the role of transmitted adaptation as a predictor of disease progression in newly infected individuals, we evaluated a separate cohort of individuals who were infected during the Step Study HIV vaccine trial⁴². Consistent with our findings in the Zambian cohort, VL was correlated with adaptation of inferred founder viruses to host HLA alleles among seroconverting participants (**Fig. 3c** and **Supplementary Table 2**). The larger correlation coefficient in the Step data compared to that for the Zambian data may be explained by differences among males and females, as the correlation for Zambian males was comparable to that observed in the (all-male) Step data (**Supplementary Table 2**), though this sex difference was not statistically significant within the Zambian cohort.

To determine whether transmitted adaptation affects VL and CD4⁺ T cell counts many years into infection, we used circulating adaptation (over all virus sequences within a cohort) as an estimate for expected transmitted adaptation for each individual in our chronic infection

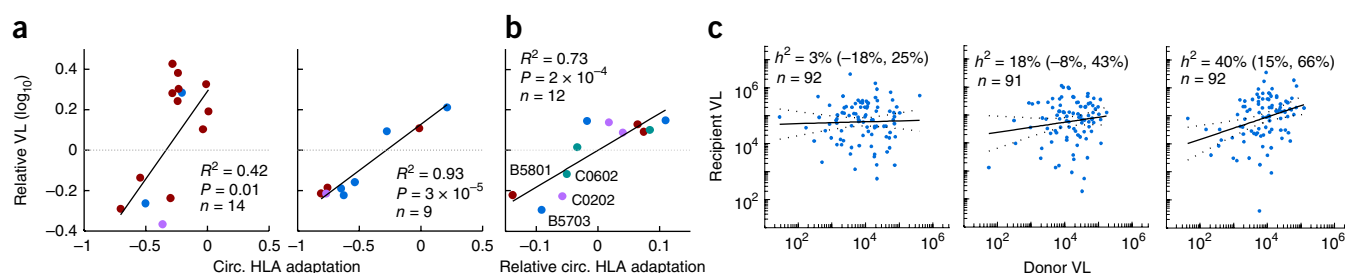
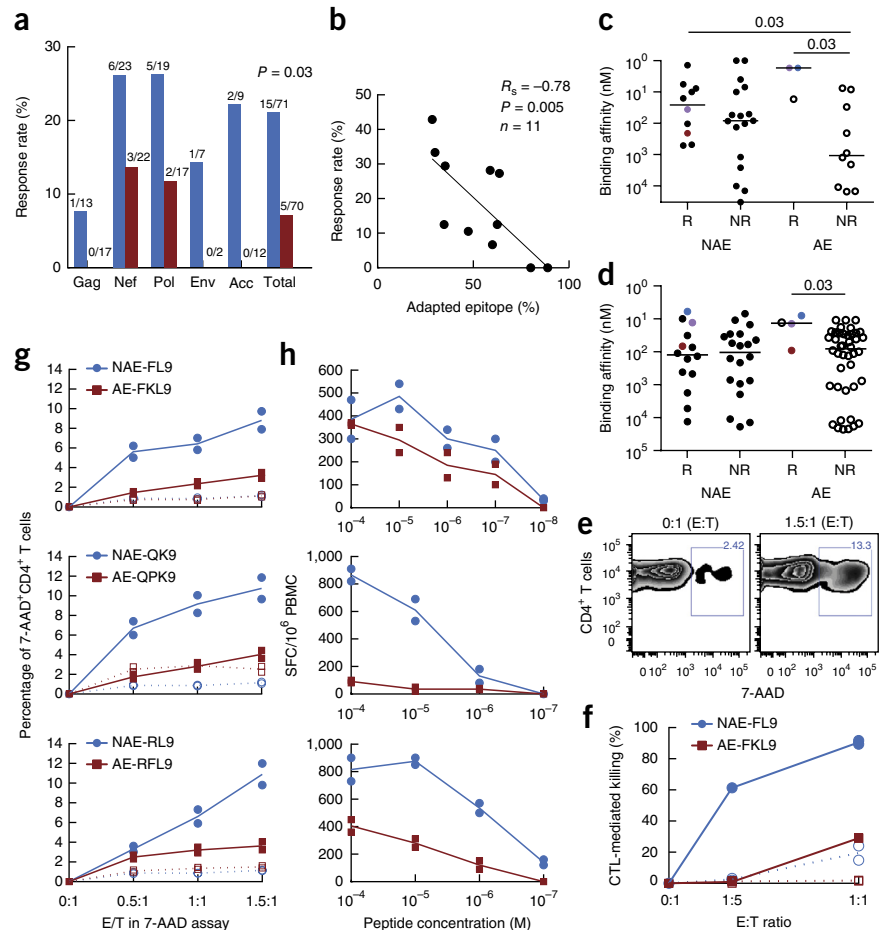


Figure 4 Adaptation affects HLA-VL associations and heritability estimates. (a) Circulating (circ.) allele-specific adaptation compared to allele-specific effects on VL, as estimated from a mixed model fitted to the Southern Africa (left; $n = 2,298$) and British Columbia (right; $n = 1,048$) cohorts. Alleles selected in an independent stepwise regression analysis are shown. *P* values and pseudo-*R*², from a mixed model with random offsets for each locus. Blue, HLA-A; red, HLA-B; purple, HLA-C. (b) Relative VL and circulating adaptation (mean-centered for each allele) for four alleles that showed city-specific VL effects (Durban, Lusaka and Gaborone). Colored points represent matched alleles in different cities. *P* value and pseudo-*R*² from a mixed model with random offset for each city (**Supplementary Fig. 8**). (c) Heritability (*h*²) estimates (95% CI; dotted lines) over all 275 Zambian linked transmission pairs with available VL and HLA types, stratified into tertiles by HLA-B adaptation similarity (from left to right: low, medium and high). Donor and recipient VL adjusted for sex, age and sample year independently for each stratum. Solid black lines indicate best fit.

Figure 5 CTL responses against pre-adapted transmitted epitopes are dysfunctional. (a,b) IFN γ response rates for autologous founder epitopes. (a) Response frequencies for nonadapted (blue) and adapted (red) epitopes. Number of epitopes tested/number of responses for each protein are indicated. P value determined via Fisher's exact test over total. (b) Response rate versus proportion of autologous epitopes that are adapted, on a per-subject level. (c,d) Experimental (c) and predicted (d) HLA-peptide binding affinity for adapted (AE) and nonadapted (NAE) epitopes that elicited a response (R) or not (NR). Colored points represent matched immunogenic NAE-AE pairs. Solid bars, median; P values (top) determined via Mann-Whitney test. (e-h) Epitope-specific CD8 $^{+}$ T cells (effectors (E)) assessed for cytotoxicity of activated CD4 $^{+}$ T cells (targets (T)). All results from duplicate experiments shown as symbols. Solid lines indicate mean values. In f and g, dotted lines and open symbols indicate negative controls. (e) Representative flow cytometry plot for 7-AAD $^{+}$ targets in the absence (0:1 E:T) or presence (1.5:1 E:T) of effectors. Numbers adjacent to outlined areas indicate the percentage of cells in the gate. (f) Relative killing of targets from HLA-matched (solid) and -mismatched (dotted) donors infected with MJ4 virus mutated to contain the NAE or AE FL9 variant and incubated with NAE- or AE-specific effectors. (g,h) Cytotoxicity (g) and antigen sensitivity (h) curves for the three matched NAE-AE epitope pairs. Keys in g apply to h. PBMC, peripheral blood mononuclear cells.



cohorts. Overall, the level of circulating adaptation to an individual's HLA alleles was an independent predictor of both VL and CD4 $^{+}$ T cell counts (Supplementary Table 3), suggesting that transmitted adaptation has long-term effects on natural control.

Adaptation confounds estimation of genetic effects on VL

The impact of autologous and transmitted adaptation on markers of disease progression implies a strong interaction between viral and host genetics with respect to these markers. Such interactions suggest that population-level estimates of epidemiologic parameters will depend on the circulating virus and the predominant host alleles in a particular population.

Indeed, allele-specific circulating adaptation explained much of the variation in HLA-specific VL and CD4 $^{+}$ T cell effects (Fig. 4a and Supplementary Fig. 8a–c), suggesting that protective alleles are those for which the circulating virus is not well adapted. Moreover, among four alleles with evidence of differential impact on VL in three Southern African cities, the relative differences in city-specific VL effects were largely explained by relative differences in city-specific circulating adaptation (Fig. 4b and Supplementary Fig. 8d). Thus, circulating adaptation may explain many of the differences in allele-specific associations with markers of disease progression among diverse cohorts; this supports the hypothesis that accumulation of escape mutations in a population will undermine natural control^{20,22}.

The role of viral genetics in determining VL can be quantified with population-level estimates of VL heritability³². Published estimates vary from 6% to 59%, with a recent meta-analysis of transmission-pair cohorts estimating broad-sense heritability at 33% (95% confidence

interval (CI): 20–46%)³². However, the results presented here suggest that the relationship between donor and recipient VL will be substantially stronger among pairs with 'similar' HLA alleles, as autologous adaptation in such donors (resulting in higher donor VL) will result in increased transmitted adaptation to their recipients (resulting in higher recipient VL).

Indeed, over the set of all 275 HLA-typed Zambian transmission pairs^{43,44}, heritability was estimated at 18% (95% CI: 4–31%). However, when we stratified couples by HLA-B adaptation similarity (Online Methods), heritability ranged from 2% (lower tertile) to 41% (upper tertile; $P = 0.009$; Fig. 4c and Supplementary Table 4). Thus, heritability estimates vary widely as a function of how similar the recipient's alleles are to the donor's, suggesting that discordant heritability estimates in the literature may in part be due to differing levels of HLA heterogeneity in the cohorts.

Dysfunctional responses to pre-adapted transmitted epitopes

The results presented so far suggest that infection by pre-adapted viruses compromises both the initial and the long-term efficacy of CTL responses. To confirm this hypothesis, we tested epitope-specific responses to autologous peptides from 11 individuals recently infected with a single HIVB founder virus (median: 31 d after infection). For each individual, we defined the founder virus sequence and identified all of the HLA-matched, optimally defined epitopes it encoded⁴⁵. Each founder virus epitope that matched the most prevalent circulating HIVB sequence and did not harbor an escape polymorphism was then classified as nonadapted; all other epitopes were classified as adapted. Autologous adapted founder virus epitopes were less likely

than nonadapted epitopes to elicit an interferon- γ (IFN γ) response (Fig. 5a), and the response rate correlated inversely with the proportion of adapted epitopes in the founder virus (Fig. 5b), suggesting that transmitted adapted epitopes are less immunogenic.

In many cases, decreased responses to adapted founder virus epitopes are probably attributable to escape-induced reductions in HLA binding affinity³⁶. However, in the current study the HLA-peptide binding affinity of numerous nonimmunogenic adapted epitopes was similar to that of immunogenic nonadapted epitopes (Fig. 5c,d), suggesting that some adapted mutations confer escape by exploiting holes in the TCR repertoire. Moreover, for all three epitopes that elicited responses in both adapted and nonadapted variants, the adapted epitopes elicited substantially weaker cytotoxic responses than the nonadapted variant did (Fig. 5e–g). These differences could not be explained by HLA-I binding or T cell polyfunctionality (Fig. 5c,d and Supplementary Fig. 9a–c), but they were consistent with reduced antigen sensitivity and magnitude of IFN γ response (Fig. 5h and Supplementary Fig. 9d). Together, these *in vivo* and *in vitro* data indicate that, when present in the founder virus, adapted epitopes are generally poorly immunogenic and, when recognized, elicit suboptimal primary CTL responses.

Vaccination with adapted epitopes

If acquisition of a pre-adapted founder virus at transmission undermines initial host CTL responses, then the quality of vaccine-induced immune responses is likely to depend on the extent to which the vaccine insert is pre-adapted to a recipient's HLA alleles. The Step Study vaccine trial provided us with an opportunity to investigate this hypothesis⁴⁶. Among trial participants, there was evidence of a weak inverse correlation between the extent to which the vaccine insert was adapted to an individual's HLA alleles and preinfection pooled IFN γ response magnitudes in ELISpot assays, as measured by two independent laboratories (Supplementary Fig. 10). These observations suggest that different vaccine insert sequences will result in qualitatively different—yet predictable—immune responses in the same individual. Whether simultaneous immunization by polyvalent vaccines will focus the immune response on functional, nonadapted epitopes, or whether such a strategy risks eliciting suboptimal, nonprotective responses to adapted epitopes, remains an important open question.

DISCUSSION

Although the importance of the CTL response, and of the presence of immune-mediated escape, in the context of HIV infection has been generally recognized for 20 years, the link between transmitted and within-host adaptation and disease progression has been obscured by the complexity of *in vivo* targeting and escape. By directly measuring adaptation using a probabilistic model, we have provided a general framework for estimating host-specific viral adaptation. The results demonstrate the dominant role HLA escape variants have in mediating disease progression, thereby validating some common assumptions while refuting others.

The role of autologous adaptation as a primary correlate of (and predecessor to) key markers of disease progression argues that an effective immune response is one that controls viral replication in the absence of escape (Fig. 2c). This stands in contrast to suggestions that an effective immune response drives the selection of escape mutations that substantially reduce intrinsic fitness. Although it is likely that the cost to intrinsic fitness of potential escape variants is critical to delaying time to escape⁴⁷, and thus will serve as a useful correlate of

protection in identifying potential vaccine candidates^{2,27–29}, escape mutations are selected *in vivo* precisely because they increase overall fitness. In theory, *in vivo* fitness may not be fully restored by escape, but we found that high levels of adaptation were strongly linked to high VL and loss of allele-specific control, indicating that compensatory mutations typically offset any reduced intrinsic fitness within a clinically relevant time frame. It may also be possible to reduce escape by increasing the breadth of the immune response⁴⁷, though such strategies must take care to account for transmitted adaptation.

Indeed, the impact of transmitted adaptation on host immunity and disease progression is critical. From an epidemiologic perspective, the interaction of host and viral genetic effects undermines efforts to predict individual and population outcomes on the basis of host or viral genetics alone. For example, estimates of VL heritability depend on the degree of similarity between donor and recipient HLA alleles, and allele-specific circulating adaptation explains much of the variation in natural control attributed to individual HLA alleles. Moreover, circulating adaptation defines why some alleles are protective only in certain regions^{20,22,48} and predicts clinical outcomes for individuals, suggesting that an individual's prognosis will depend in part on the region in which the infection is acquired. These results further provide explanations for epidemiologic observations that may influence transmitted adaptation, including reduced VL associated with rare alleles⁴⁹ and elevated VL associated with multiple-virus infection⁵⁰ or infection by a partner with a shared B allele⁴³. By undermining HLA-mediated control, accumulation of escape variants in different populations will thus (all else being equal) lead to increasing average viral loads as the pandemic progresses, though other factors may mitigate this process²². Measuring transmitted and within-host adaptation will thus be critical for clinical and observational trials in which reduction in VL or rate of CD4⁺ T cell decline is a primary or secondary endpoint.

From an immunologic perspective, the inability of primary immune responses to effectively target adapted epitopes casts doubt on the efficacy of prophylactic and therapeutic vaccine strategies that seek to elicit responses to such variants and argues instead for conserved element approaches that target a restricted set of epitopes with limited circulating variation. Furthermore, the observation that some adapted epitopes may competently bind their cognate HLA and elicit detectable, yet dysfunctional, responses suggests that the virus is exploiting biases in the circulating naive TCR pool^{51,52}. Such dysfunctional responses argue for in-depth screening for virus-inhibiting responses throughout the vaccine-development cycle and raise the disturbing possibility that dysfunctional responses are causally worse than the complete absence of a response⁵³.

By combining the largest, most feature-rich data sets available with a novel statistical method for summarizing the extent of cellular immune adaptation, we have demonstrated the ability of HIV to exploit universal 'holes' in the adaptive immune response. These holes explain much of the HLA- and region-specific heterogeneity in clinical outcomes and suggest that vaccine-induced 'sieving'⁵⁴ may not simply result from insufficient vaccine coverage, but is in part inherent in the limitations of the naive immune system. Accounting for these holes will be imperative in ongoing efforts to design strategies that leverage the CTL response to prevent infection^{27–31} or clear the latent reservoir⁵⁵.

METHODS

Methods and any associated references are available in the [online version of the paper](#).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank M. Carrington for comments on the manuscript, S. Riddler (University of Pittsburgh, Pittsburgh, Pennsylvania, USA) for access to HLA and sequence data from the ACTG trials, D. Claiborne (Emory University, Atlanta, Georgia, USA) for providing MJ4 proviruses with nonadapted and adapted epitopes, R.A. Kaslow and J. Tang (University of Alabama, Birmingham, Alabama, USA) for access to Zambian HLA data, and D. Goedhals and C. van Vuuren (University of Free State, Bloemfontein, South Africa) for curating additional clinical data from the Bloemfontein cohort. We thank Merck, the NIH National Institute of Allergy and Infectious Diseases (NIAID) and the NIAID-funded HIV Vaccine Trials Network for providing the clinical data set, viral sequences, HLA types and CTL response data from the Step Study (HVTN 502). We also thank the Step and ACTG 5142 and 5128 staff and trial participants, as well as the staff and volunteers of the HOMER, WAHCS, ZEHRP, Durban, Gaborone, Kimberley and Bloemfontein cohorts, for their contributions. This study was funded by NIAID (grants R01 AI112566 (P.A.G.), R56 AI098551 (P.A.G.), R01 AI64060 (E.H.), R37 AI51231 (E.H.), P01 AI074415 (T.M.A.), U01 AI 66454 (R.S.), R01 AI46995 (P.J.R.G.) and R01 AI071906 (R.A. Kaslow and J. Tang)), the Canadian Institutes of Health Research (grants MOP-93536 and HOP-115700 (both to M.A.B. and Z.L.B.)) and the Wellcome Trust (grant WT104748MA (P.J.R.G.)). HLA typing and viral sequencing of the ACTG cohorts were supported by the NIH (grant U01 AI 068636 to R.H.), the National Institute of Mental Health (NIMH) and the National Institute of Dental and Craniofacial Research (NIDCR). Support for the ZEHRP cohort was also provided by the International AIDS Vaccine Initiative (S.A.) and made possible in part by the support of the American people through the US Agency for International Development (USAID). A full list of IAVI donors is available at <http://www.iavi.org>. This work was also supported in part by the Virology Core at the Emory Center for AIDS Research (grant P30 AI050409 (E.H.)), the Flow Cytometry Core at the University of Alabama at Birmingham Center for AIDS Research (grant P30 AI027767 (P.A.G.)), the Tennessee Center for AIDS Research (P30 AI110527 (S.M.)) and the Yerkes National Primate Research Center base (grant P51OD11132 (E.H.)) through the NIH Office of the Director. M.S. was supported in part by an Action Cycling Fellowship. T.N. was supported by the International AIDS Vaccine Initiative, the South African Department of Science and Technology and the National Research Foundation through the South Africa Research Chairs Initiative, by an International Early Career Scientist award from the Howard Hughes Medical Institute, and by the Victor Daitz Foundation. P.R.H. is supported by a CIHR/GSK Professorship in Clinical Virology. Z.L.B. is supported by a Scholar Award from the Michael Smith Foundation for Health Research.

AUTHOR CONTRIBUTIONS

J.M.C. designed and implemented the statistical analyses and adaptation model and wrote the paper. N.P. designed and implemented the adaptation model, with help from V.Y.F.T., A.K., C.E.D. and D.H. N.F. and C.J.B. helped with the design and/or implementation of the statistical analyses. P.A.G. designed the functional studies on primary immune responses, which were performed by V.Y.D., A.B. and J.S. K.P. and T.M.A. provided controller sequences. M.S., S.A. and E.H. provided transmission pair and longitudinal sequence and clinical data. M.A.B., J.G., M.A.P., W.K., R.H., M.J., S.M., R.S., J.F., P.R.H., T.N., S.A., P.J.R.G., Z.L.B. and E.H. provided chronic infection data. J.M.C., E.H., P.A.G., Z.L.B. and P.J.R.G. advised the project and helped write the paper, with input from all other authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

- Goonetilleke, N. *et al.* The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection. *J. Exp. Med.* **206**, 1253–1272 (2009).
- Pereyra, F. *et al.* HIV control is mediated in part by CD8⁺ T-cell targeting of specific epitopes. *J. Virol.* **88**, 12937–12948 (2014).
- Carlson, J.M., Le, A.Q., Shahid, A. & Brumme, Z.L. HIV-1 adaptation to HLA: a window into virus–host immune interactions. *Trends Microbiol.* **23**, 212–224 (2015).
- Martinez-Picado, J. *et al.* Fitness cost of escape mutations in p24 Gag in association with control of human immunodeficiency virus type 1. *J. Virol.* **80**, 3617–3623 (2006).
- Boutwell, C.L., Rowley, C.F. & Essex, M. Reduced viral replication capacity of human immunodeficiency virus type 1 subtype C caused by cytotoxic-T-lymphocyte escape mutations in HLA-B57 epitopes of capsid protein. *J. Virol.* **83**, 2460–2468 (2009).
- Wright, J.K. *et al.* Impact of HLA-B*81-associated mutations in HIV-1 Gag on viral replication capacity. *J. Virol.* **86**, 3193–3199 (2012).
- Goepfert, P.A. *et al.* Transmission of HIV-1 Gag immune escape mutations is associated with reduced viral load in linked recipients. *J. Exp. Med.* **205**, 1009–1017 (2008).
- Chopra, D.R. *et al.* Transmission of HIV-1 CTL escape variants provides HLA-mismatched recipients with a survival advantage. *PLoS Pathog.* **4**, e1000033 (2008).
- Carlson, J.M. *et al.* Selection bias at the heterosexual HIV-1 transmission bottleneck. *Science* **345**, 1254031 (2014).
- Prince, J.L. *et al.* Role of transmitted Gag CTL polymorphisms in defining replicative capacity and early HIV-1 pathogenesis. *PLoS Pathog.* **8**, e1003041 (2012).
- Feeney, M.E. *et al.* Immune escape precedes breakthrough human immunodeficiency virus type 1 viremia and broadening of the cytotoxic T-lymphocyte response in an HLA-B27-positive long-term-nonprogressing child. *J. Virol.* **78**, 8927–8930 (2004).
- Keane, N.M. *et al.* High-avidity, high-IFN γ -producing CD8 T-cell responses following immune selection during HIV-1 infection. *Immunol. Cell Biol.* **90**, 224–234 (2012).
- Almeida, C.-A.M. *et al.* Translation of HLA-HIV associations to the cellular level: HIV adapts to inflate CD8 T cell responses against Nef and HLA-adapted variant epitopes. *J. Immunol.* **187**, 2502–2513 (2011).
- Allen, T.M. *et al.* De novo generation of escape variant-specific CD8⁺ T-cell responses following cytotoxic T-lymphocyte escape in chronic human immunodeficiency virus type 1 infection. *J. Virol.* **79**, 12952–12960 (2005).
- Iglesias, M.C. *et al.* Escape from highly effective public CD8⁺ T-cell clonotypes by HIV. *Blood* **118**, 2138–2149 (2011).
- Ntale, R.S. *et al.* Temporal association of HLA-B*81:01- and B*39:10-mediated HIV-1 p24 sequence evolution with disease progression. *J. Virol.* **86**, 12013–12024 (2012).
- Oxenius, A. *et al.* Loss of viral control in early HIV-1 infection is temporally associated with sequential escape from CD8⁺ T cell responses and decrease in HIV-1-specific CD4⁺ and CD8⁺ T cell frequencies. *J. Infect. Dis.* **190**, 713–721 (2004).
- Goulder, P.J.R. *et al.* Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS. *Nat. Med.* **3**, 212–217 (1997).
- Crawford, H. *et al.* Evolution of HLA-B*5703 HIV-1 escape mutations in HLA-B*5703-positive individuals and their transmission recipients. *J. Exp. Med.* **206**, 909–921 (2009).
- Kawashima, Y. *et al.* Adaptation of HIV-1 to human leukocyte antigen class I. *Nature* **458**, 641–645 (2009).
- Cotton, L.A. *et al.* Genotypic and functional impact of HIV-1 adaptation to its host population during the North American epidemic. *PLoS Genet.* **10**, e1004295 (2014).
- Payne, R. *et al.* Impact of HLA-driven HIV adaptation on virulence in populations of high HIV seroprevalence. *Proc. Natl. Acad. Sci. USA* **111**, E5393–E5400 (2014).
- Goulder, P.J. *et al.* Evolution and transmission of stable CTL escape mutations in HIV infection. *Nature* **412**, 334–338 (2001).
- Asquith, B., Edwards, C.T.T., Lipsitch, M. & McLean, A.R. Inefficient cytotoxic T lymphocyte-mediated killing of HIV-1-infected cells in vivo. *PLoS Biol.* **4**, e90 (2006).
- Iversen, A.K.N. *et al.* Conflicting selective forces affect T cell receptor contacts in an immunodominant human immunodeficiency virus epitope. *Nat. Immunol.* **7**, 179–189 (2006).
- Korber, B.T., Letvin, N.L. & Haynes, B.F. T-cell vaccine strategies for human immunodeficiency virus, the virus with a thousand faces. *J. Virol.* **83**, 8300–8314 (2009).
- Rolland, M., Nickle, D.C. & Mullins, J.I. HIV-1 group M conserved elements vaccine. *PLoS Pathog.* **3**, e157 (2007).
- Mothe, B. *et al.* Definition of the viral targets of protective HIV-1-specific T cell responses. *J. Transl. Med.* **9**, 208 (2011).
- Létourneau, S. *et al.* Design and pre-clinical evaluation of a universal HIV-1 vaccine. *PLoS One* **2**, e984 (2007).
- Borthwick, N. *et al.* Vaccine-elicited human T cells recognizing conserved protein regions inhibit HIV-1. *Mol. Ther.* **22**, 464–475 (2014).
- Fischer, W. *et al.* Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. *Nat. Med.* **13**, 100–106 (2007).
- Fraser, C. *et al.* Virulence and pathogenesis of HIV-1 infection: an evolutionary perspective. *Science* **343**, 1243727 (2014).
- van Dorp, C.H., van Boven, M. & de Boer, R.J. Immuno-epidemiological modeling of HIV-1 predicts high heritability of the set-point virus load, while selection for CTL escape dominates virulence evolution. *PLoS Comput. Biol.* **10**, e1003899 (2014).
- Yewdell, J.W. Confronting complexity: real-world immunodominance in antiviral CD8⁺ T cell responses. *Immunity* **25**, 533–543 (2006).
- Carlson, J.M. *et al.* Widespread impact of HLA restriction on immune control and escape pathways of HIV-1. *J. Virol.* **86**, 5230–5243 (2012).
- Carlson, J.M. *et al.* Correlates of protective cellular immunity revealed by analysis of population-level immune escape pathways in HIV-1. *J. Virol.* **86**, 13202–13216 (2012).

37. Miura, T. *et al.* HLA-associated viral mutations are common in human immunodeficiency virus type 1 elite controllers. *J. Virol.* **83**, 3407–3412 (2009).
38. Brockman, M.A. *et al.* Early selection in Gag by protective HLA alleles contributes to reduced HIV-1 replication capacity that may be largely compensated for in chronic infection. *J. Virol.* **84**, 11937–11949 (2010).
39. Huang, K.-H.G. *et al.* Progression to AIDS in South Africa is associated with both reverting and compensatory viral mutations. *PLoS One* **6**, e19018 (2011).
40. Kiepiela, P. *et al.* CD8⁺ T-cell responses to different HIV proteins have discordant associations with viral load. *Nat. Med.* **13**, 46–53 (2007).
41. Wright, J.K. *et al.* Influence of Gag-protease-mediated replication capacity on disease progression in individuals recently infected with HIV-1 subtype C. *J. Virol.* **85**, 3996–4006 (2011).
42. Buchbinder, S.P. *et al.* Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial. *Lancet* **372**, 1881–1893 (2008).
43. Tang, J. *et al.* HLA allele sharing and HIV type 1 viremia in seroconverting Zambians with known transmitting partners. *AIDS Res. Hum. Retroviruses* **20**, 19–25 (2004).
44. Song, W. *et al.* Disparate associations of HLA class I markers with HIV-1 acquisition and control of viremia in an African population. *PLoS One* **6**, e23469 (2011).
45. Llano, A., Frahm, N. & Brander, C. in *HIV Molecular Immunology* (eds. Yusim, K. *et al.*) 3–24 (Los Alamos National Laboratory, 2009).
46. McElrath, M.J. *et al.* HIV-1 vaccine-induced immunity in the test-of-concept Step Study: a case-cohort analysis. *Lancet* **372**, 1894–1905 (2008).
47. Liu, M.K.P. *et al.* Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. *J. Clin. Invest.* **123**, 380–393 (2013).
48. Matthews, P.C. *et al.* Differential clade-specific HLA-B*3501 association with HIV-1 disease outcome is linked to immunogenicity of a single Gag epitope. *J. Virol.* **86**, 12643–12654 (2012).
49. Trachtenberg, E. *et al.* Advantage of rare HLA supertype in HIV disease progression. *Nat. Med.* **9**, 928–935 (2003).
50. Janes, H. *et al.* HIV-1 infections with multiple founders are associated with higher viral loads than infections with single founders. *Nat. Med.* **21**, 1139–1141 (2015).
51. Miles, J.J., Douek, D.C. & Price, D.A. Bias in the $\alpha\beta$ T-cell repertoire: implications for disease pathogenesis and vaccination. *Immunol. Cell Biol.* **89**, 375–387 (2011).
52. Kløverpris, H.N. *et al.* CD8⁺ TCR bias and immunodominance in HIV-1 infection. *J. Immunol.* **194**, 5329–5345 (2015).
53. Mailliard, R.B. *et al.* Selective induction of CTL helper rather than killer activity by natural epitope variants promotes dendritic cell-mediated HIV-1 dissemination. *J. Immunol.* **191**, 2570–2580 (2013).
54. Rolland, M. *et al.* Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. *Nat. Med.* **17**, 366–371 (2011).
55. Deng, K. *et al.* Broad CTL response is required to clear latent HIV-1 due to dominance of escape mutations. *Nature* **517**, 381–385 (2015).

ONLINE METHODS

Data sets. This study used previously described data sets^{9,10,13,36–38,42–44,46,54,56–79}, as outlined in **Supplementary Figure 1** and detailed in **Supplementary Note 2**.

Functional data for acute infection cohort. Functional experiments, performed as previously described^{36,45,53,78–88}, are detailed in **Supplementary Note 3**.

Statistical methods. For each cohort, VL was transformed using \log_{10} , and CD4⁺ T cell counts were transformed using the Box–Cox procedure⁸⁹, applied separately to all British Columbia ($\lambda = 0.45$) and all Southern Africa ($\lambda = 0.58$) data, to make the counts as close to normally distributed as possible. The resulting distributions in the chronic infection cohorts remained modestly right-skewed (CD4⁺ T cells) and left-skewed (VL). This effect was most extreme for the British Columbia data, where VL was right-censored at 10^6 copies per ml. Such censoring may affect model estimates; therefore we focused cross-sectional analyses on the Southern Africa data and used nonparametric tests where possible (see below). The associations between adaptation and VL or CD4⁺ T cell count were highly significant whether estimated with the mixed model (to account for confounders) or Spearman rank correlation (to account for non-normality).

Where HLA alleles were typed to low or medium resolution, we estimated a probability distribution over HLA haplotypes as previously described⁹⁰. The distributions were used in training and application of the adaptation model as described below. When used as independent variables in standard generalized linear fixed and mixed-effects models, we imputed the HLA alleles by calculating the marginal probability that the individual expressed each HLA supertype, type and subtype. The marginal probabilities for each HLA were then treated as a fractional observation of that HLA. HLA alleles were treated as (possibly fractionally observed) binary variables, such that homozygosity was not encoded in the models.

The adaptation model training was based on HLA and viral genetic data alone; clinical parameters were not considered in the model training. Autologous adaptation for the Southern Africa, British Columbia and Zambian transmission-pair cohorts were estimated out of sample using tenfold cross-validation. Individuals lacking both VL and CD4⁺ T cell data were used in the training sets for all ten folds. For the IHAC cohort, clinical parameters were available for only the British Columbia cohort. Thus, all individuals from the ACTG and Western Australia cohorts were included in all cross-validation partitions. All other adaptation scores were computed from models trained on the entire Southern Africa (HIVC) or IHAC (HIVB) data set.

All reported *P* values are from two-sided tests and are unadjusted for multiple comparisons. Where a large number of comparisons was used, false discovery rates in the form of *q*-values are reported^{91,92}. Wherever possible, analyses were adjusted for subject age, sex, cohort of origin and HLA alleles. Age was dichotomized at ≥ 40 (ref. 93). HLA alleles were treated as random effects (see below). For the British Columbia data, sampling year was treated as a random effect. For the Zambian transmission pairs, no significant effect was observed for sampling year; we therefore used a dichotomous fixed effect (around median sampling year) as a covariate. For the Southern Africa data, sampling dates were not available for a substantial number of samples; however, each data set (with the exception of that for the Zambia cohort) included samples collected within a small time frame. Therefore, any variation due to sampling time is approximately captured by the cohort indicator variables. For all analyses, missing demographic data (age, sex, sampling year) were imputed using linear imputation; individuals with missing clinical, functional or sequence data were excluded. Unless otherwise noted, individuals with partially missing sequence data were included in the Southern Africa and British Columbia autologous adaptation results but were excluded from all other analyses.

Stepwise regression. Stepwise regression for generalized linear models was performed using $P < 0.05$ and $P > 0.05$ as entry and exit criteria, respectively. For the controller data (**Supplementary Fig. 3b**), logistic regression was used, and all HLA alleles, at both type and subtype resolution, observed in at least five individuals were included as potential features. For these analyses, we excluded all HLA supertypes. The *q*-values were estimated from the *P* values of all possible features, conditioned on the final model.

Stepwise regression was also applied to the Southern Africa and British Columbia cohorts to identify the dominant alleles that contribute to clinical

parameters. For these applications, only HLA-subtypes (four-digit) observed in at least 20 individuals were considered, and age, sex and cohort indicator variables were included in the model as covariates.

Cox proportional hazards model. We assessed the association between transmitted adaptation and CD4⁺ T cell decline using the 46 Zambian transmission pairs for whom we had longitudinal CD4⁺ T cell counts from the period immediately after transmission. We evaluated the relationship using the Cox proportional hazards model, treating the adaptation of the donor virus to the recipient alleles as a continuous variable and treating sex, age (≥ 40) and sample date ($>$ median) as covariates (none was significant). For **Figure 3a**, we stratified individuals on the basis of the mean transmitted adaptation over those 46 subjects. For that figure only, the adaptation score was scaled so that a one-unit change corresponded to the difference in mean adaptation within the two strata. Thus, the reported hazard ratio (3.0) indicates that an individual with above-average transmitted adaptation progresses to a CD4⁺ T cell count < 250 cells per mm^3 at a threefold-higher rate than an individual with below-average transmitted adaptation.

Our pre-specified endpoint was a CD4⁺ T cell count < 250 cells per mm^3 . This endpoint was based on two factors: (1) the national therapy guideline at the time of sampling was a CD4⁺ T cell count < 200 cells per mm^3 , and no individuals initiated with a CD4⁺ T cell count > 250 cells per mm^3 ; and (2) 27 of 46 (59%) individuals reached a CD4⁺ T cell count < 250 cells per mm^3 within the timeframe of the study. As a *post hoc* analysis, we repeated the analysis for CD4⁺ T cell count thresholds, incrementing by 50 cells per mm^3 , and including therapy initiation as an additional endpoint; results were significant ($P < 0.05$) for all endpoints from 150 to 350 cells per mm^3 , with hazard ratios ranging from 2.2 to 3.2.

Linear mixed models. With the chronic data, our primary goal was to estimate the effect of adaptation on clinical parameters. However, HLA class I alleles are known to represent the primary host genetic factor that influences VL and CD4⁺ T cell counts^{94,95}. A number of different HLA alleles within the HLA-I loci have been reported as significantly associated with VL, with differences observed among HIV subtype, cohort and disease stage. Furthermore, some HLA alleles appear to have HLA-subtype-specific effects on VL and CD4⁺ T cell count (most prominently, B*58:01 compared to B*58:02, but others as well), whereas other alleles act at the type or even the supertype level. It is therefore clear that all HLA alleles, at all resolutions, need to be accounted for in assessments of the effects of a new independent variable.

To this end, we used linear mixed models (LMMs). In this setup, we conceptually build a linear model with a separate weight for every HLA allele (we provide one each at the supertype, type and subtype levels for each HLA allele). Because such a model is overparameterized, we place a Gaussian prior distribution on the parameter for each HLA and integrate out the HLA effects on the basis of those priors. The parameter-specific Gaussian priors are specified by $\mathcal{N}(0, \sigma_A)$, $\mathcal{N}(0, \sigma_B)$, and $\mathcal{N}(0, \sigma_C)$, for the HLA A, B and C alleles, respectively. In this way, we are able to condition on all HLA alleles while allowing the variance of effect sizes to differ among loci. In addition, we treat the subcohorts (Southern Africa) and sampling year (British Columbia) as random effects, drawn from a separate Gaussian distribution with its own variance, similar to the treatment for other random effects noted in the text. When displayed in tables, all features in italics are treated as random effects with their own effect-size variance. We used the LMM implementation from the Matlab statistics toolbox. The model can be expressed as

$$Y = X\beta + \sum_{i=1}^R Z_i B_i + E$$

where Y is the $N \times 1$ response vector, X is the $N \times P$ fixed-effects design matrix, Z_i ($i = 1, \dots, R$) are the $N \times Q_i$ random effects design matrices, β is a $P \times 1$ fixed-effects vector, $B_i \sim \mathcal{N}(0, \sigma_i^2 I)$, $E \sim \mathcal{N}(0, \sigma_E^2 I)$, I is the $N \times N$ identity matrix, σ_E^2 is the variance of the elements of E , and σ_i^2 is the variance of the estimates of B_i . This model thus groups random effects by categories (for example, each HLA-B allele is grouped with all HLA-B alleles), then estimates different variance components for each effect category.

Models were fit using both restricted maximum likelihood (REML) and maximum likelihood (ML). The fraction of variation explained (pseudo- R^2)

was computed according to the likelihood ratio test method⁹⁶, trained using ML. *P* values for the fixed effects were derived from the standard error of the estimated effects (trained via REML); *P* values for random effects came from the likelihood ratio test (trained via ML), corrected for boundary effects⁹⁷.

Allele-specific adaptation. To estimate the effects of allele-specific adaptation, we constructed a series of independent linear models, one for each HLA subtype *h*. Each linear model was defined as follows:

$$y_i = \beta x_i + \beta_1 h_i + \beta_2 \text{Adapt}_h(s_i) + x_i \beta + \epsilon_i$$

where y_i is the value of the dependent variable (transformed VL or CD4⁺ T cell count) for individual *i*; h_i is a binary variable indicating whether the individual *i* expresses *h*; $\text{Adapt}_h(s_i)$ is the autologous adaptation to *h* for each individual *i*, defined as 0 if $h_i = 0$; x_i is a vector of covariates; β is the corresponding vector of weights; and ϵ_i is independently sampled from a Gaussian distribution. In this context, the covariates are indicator variables for cohort of origin, sex, age (≥ 40) and HLA subtypes identified via an independent stepwise regression analysis. Because $\text{Adapt}_h(s_i)$ ranges from -1 to 1 , $\beta_1 - \beta_2$ defines the expected relative change in VL attributable to *h* in the complete absence of allele-specific adaptation, and $\beta_1 + \beta_2$ defines the expected relative change in VL attributable to *h* in the presence of complete allele-specific adaptation. The 95% confidence intervals and *P* values are readily obtained from the variance–covariance matrix associated with the parameter estimates. **Figure 2b** and **Supplementary Figure 3c** show all alleles for which a likelihood ratio test against a null model with $\beta_1 = \beta_2 = 0$ was significant at $P < 0.05$. This threshold corresponded to a 10% false discovery rate for both VL and CD4⁺ T cell count. In **Figure 2b**, we show the results of testing the null hypothesis $\beta_2 = 0$, indicating the significance of allele-specific adaptation on VL (and similarly for CD4⁺ T cell count; **Supplementary Fig. 3c**).

Longitudinal data analysis. To test whether changes in adaptation predict future changes in VL, we used longitudinal VL and sequence samples from the Zambian transmission-pair data set and analyzed them in an autoregression mixed model with second-order lag and random intercept varying by subject and HLA allele. Adapting the above notation for mixed models, for each subject *i*, we modeled VL at time point *t* (described below) using

$$\text{VL}_{i,t} = \beta_1 \text{VL}_{i,t-1} + \beta_2 \text{VL}_{i,t-2} + \beta_3 \text{Adapt}_{i,t-1} + x_i \beta + \sum_{j=1}^R z_{ij} b_j + \epsilon$$

As above, $x_i \beta$ captures fixed-effect covariates (**Supplementary Fig. 7**), and z_{ij} is a vector of random effects, with $b_j \sim \mathcal{N}(0, \sigma_j)$ denoting the random effect weights. Here we used HLA alleles (as above) and subject identifiers as random effects to account for subject- and HLA-specific effects on the change of VL over time. The model thus assumes that VL at any time point can be predicted on the basis of the 'before' VL measurements, the absolute level of adaptation at the prior time point, and a set of covariates. We chose a lag of two on the basis of exploratory analysis in the absence of adaptation that showed that a third-order lag did not significantly improve model fit.

Our primary endpoint was to determine whether autologous adaptation at the prior time point predicted changes in VL at the next time point. Secondary analyses showed estimated effect sizes of alternative definitions of adaptation (allele-specific and protein-specific; **Supplementary Fig. 7b**). We also explored a model in which adaptation is the dependent variable and VL is the independent variable, and we varied the definition of adaptation.

To specify the time points, we started with available sequence samples, limiting to samples with complete *gag*, *pol* and *nef* sequencing. These were sampled approximately 0–3, 3–9, 9–15, 15–21 and 21–30 months after infection, though precise sampling times varied among individuals. Each sequence sample was discretized to one of the above time points; if multiple samples discretized to the same time point, the latest sample in the time point was used. VL measurements were discretized to the same time points. If multiple VL measurements mapped to the same time point, we used that closest to the sequencing time point. Only VL measurements made within 90 d of the sequencing samples were used. VL and sampling dates were identical for 384 of 422 matched time points.

Circulating adaptation and HLA alleles. **Figure 4a,b** and **Supplementary Figure 8a–c** display the correlation between allele-specific circulating adaptation

and the relative protection attributable to each allele. For each allele, we computed adaptation between that allele and the HIV sequences isolated from all individuals in the cohort (limiting to sequences without missing data). Allele-specific circulating adaptation was then defined as the mean adaptation for each allele over these sequences. We computed this mean separately over all British Columbia and all Southern Africa sequences. City-specific circulating adaptation for the Southern Africa cohort was computed for the three cities with the largest samples sizes (see below). Circulating adaptation thus estimates the expected transmitted allele-specific adaptation were an individual with that allele infected randomly by an HIV sequence selected from that cohort (or city).

To estimate the allele-specific effect on VL or CD4⁺ T cell counts, we used LMMs as described above, but using a single variance parameter for HLA-A, -B and -C to make cross-locus comparison possible. Age, sex and cohort were used as covariates. The allele-specific effect on VL or CD4⁺ T cell count was taken to be the best linear unbiased estimate for each allele, which was then regressed against allele-specific circulating adaptation. *R*² and *P* values were computed as described above for an LMM that uses the best linear unbiased estimate as the dependent variable, HLA locus as a random effect, and allele-specific circulating adaptation as the fixed effect of interest. In all analyses, we limited to HLA subtypes that were observed in at least 20 individuals.

We performed three analyses. In the first (**Fig. 4a** and **Supplementary Fig. 8a**), we limited the analysis to alleles selected in a stepwise regression procedure, as described above. The second analysis (**Supplementary Fig. 8b,c**) included all alleles. When all alleles are included, the effect size estimated for each allele shrinks, because of information sharing across allele pairs that are in linkage disequilibrium, as well as the increased regularization effect of the LMM.

In the third analysis (**Fig. 4c** and **Supplementary Fig. 8d**), we compared city-specific VL effects against city-specific circulating adaptation. To this end, we limited our samples to individuals from the three cities with the most subjects (Durban, South Africa; Gaborone, Botswana; and Lusaka, Zambia), then performed stepwise regression on those individuals to identify alleles that should be used as covariates. We then tested all alleles for significant interaction effects with the set of city indicator variables using a likelihood ratio test, and we identified four alleles with some evidence of differential effects by city ($P < 0.1$). We then estimated the city-specific effects on VL in a mixed model, including as covariates the other alleles and demographic variables; all effects were estimated jointly. The city-specific effects on VL compared to city-specific circulating adaptation are shown in **Supplementary Figure 8d**, which shows a clear trend in which cities with higher circulating adaptation for a given allele are also associated with higher relative VL for that allele. To form an omnibus statistical test, we mean-centered the city-specific VL effects and the circulating adaptation for each of the four alleles (**Fig. 4b**). We then estimated pseudo-*R*² and *P* values by modeling mean-centered VL effects in a mixed model with mean-centered circulating adaptation as a fixed effect and city as a random effect.

Adaptation score. When a CTL response is directed against a particular epitope, there is a fitness advantage for viruses containing genetic mutations that reduce or eliminate that response, provided those mutations do not reduce viral protein function such that the loss of fitness from disrupted protein function is greater than the gain in fitness from reducing the efficacy of the immune response. Although such escape mutations may act by disrupting TCR recognition, HLA binding or epitope processing, the escape mutations are remarkably consistent across individuals with a particular HLA allele. Typically, escape mutations are specific to HLA subtypes^{35,98,99}, though the same mutation may be selected across HLA types and even supertypes³⁵. These observations allow us to conceive of HLA-specific adaptation roughly in terms of the proportion of known HLA-associated sites that have escaped, as we have done previously^{21,22,100}. The problem with this definition is that it ignores the apparent hierarchy of escape: although escape is largely consistent, there are large variations across individuals in the timing of escape¹⁰¹. Some of this variation is due to variation in the frequency with which an epitope is targeted (immunodominance), though even when an epitope is targeted, alternative escape routes may be taken, with some typically preferred over others. In addition, variations in the population-wide prevalence of escape mutations make observation of certain variants more surprising than that of others in any given individual²⁰. As a result, simple counting-based metrics of escape will underemphasize the presence of rare escapes and

overemphasize the presence of common escapes (some of which are consensus in the circulating viral populations).

We consider that a probabilistic approach to estimating HLA-specific adaptation will yield a more intuitive metric that implicitly accounts for the frequency of within-host escape, as well as the baseline frequency of escape polymorphisms in the population. Conceptually, our approach is to estimate the probability distribution over all possible HIV sequences, conditional on all possible HLA repertoires. In practice, these distributions are estimated from observed data, as described below. We then define adaptation of a particular sequence to a particular HLA allele as a function of the likelihood ratio that compares a model in which the immune system is restricted by that single HLA allele against a hypothetical null model in which there is no immune response. This ratio is transformed so as to be in the range from -1 to 1 . The computation of the adaptation score is thus the result of several steps:

1. Training the model
 - a. A feature-selection step in which the HLA alleles that drive selection at each site are identified
 - b. The estimation of the multinomial probability distribution over all amino acids (AAs) at each site, conditional on all HLAs and the transmitted sequence
2. Defining adaptation of a sequence s with respect to an HLA allele h
 - a. Estimating the probabilities of observing a particular HIV sequence in (1) the presence of a specific HLA and (2) the absence of all immune pressure
 - b. A transformation of the likelihood ratio from step 2a

What follows is a detailed description of each of the steps, beginning with a preliminary introduction that defines the notation and outlines the approach, followed by one section for each of these steps.

Parenthetically, we note that the model described in step 1 does not describe the rate of change in the viral population. Rather, it estimates the distribution of AAs among chronically infected individuals. Because adaptation increases during chronic infection (Supplementary Fig. 2c), parameter learning is dependent on the average duration of infection in the training set. Thus models trained on individuals with advanced disease (our HIVB model) will encode different AA distributions than those trained on individuals who are at an earlier stage of infection (our HIVC model). Qualitatively, not observing expected escape mutations will thus be less 'surprising' in the HIVC model.

Preliminaries. Let $S = \{S_i\}$ be a random variable with a state space that covers all possible HIV sequences over $i = 1, \dots, N$ sites (which may span multiple proteins). Our aim is to estimate the probability distribution over S conditional on an individual's HLA alleles. An individual's HLA class I repertoire consists of three to six HLA alleles (two each from the HLA-A, -B and -C loci, with a possibility of homozygosity at each locus). HLA alleles are specified hierarchically¹⁰². For our purposes, we consider three levels: supertype, type and subtype. Because superotypes are defined on the basis of binding profiles, some alleles do not fit within a supertype, whereas others are classified as two different superotypes¹⁰³. We represent the space of possible HLA combinations using a binary vector H , with one entry for each HLA supertype, subtype and type observed in our training data sets. We refer to a binary vector realization of H as \tilde{h} , but for ease of notation we sometimes write $H = \{h_1, \dots, h_k\}$ to represent the binary vector \tilde{h} that consists of all zeros except those superotypes, types and subtypes corresponding with the k specified HLA alleles. For example, $H = \{B^*57:01, B^*58:01\}$, corresponds to the binary vector with five entries set to 1: those for B*57:01, B*57, B*58, B*58:01 and the B58 supertype. Our aim is to estimate

$$\Pr(S = s | H = h)$$

for any sequence s and any set of HLA alleles h . Under the assumption of independence among sites, we factor the distribution as

$$\Pr(S = s | H = h) = \prod_{i=1}^N \Pr(S_i = s_i | H = h)$$

It should be noted that the state of a sequence in chronic infection is strongly dependent on the transmitted sequence, which in turn will be related to other

transmitted sequences on the basis of phylogenetic relatedness. Thus, we write the per-site probability distribution using the law of total probability as

$$\Pr(S_i = s_i | H = h) = \sum_{t_i} \Pr(S_i = s_i | H = h, T_i = t_i) \Pr(T_i = t_i)$$

where $T = \{T_i\}$ represents the space of possible transmitted HIV sequences over $i = 1, \dots, N$ sites. See ref. 104 for a full motivation and explanation of this factorization that is used to create the phylogenetically corrected distribution. As described below, in the present application, $\Pr(S_i = s_i | H = h, T_i = t_i)$ is defined according to a modified logistic regression model. The prior distribution over the transmitted sequence, $\Pr(T_i = t_i)$, is specified differently for model training and estimation of adaptation.

For simplicity of notation, here we often use the shorthand $\Pr(s | h)$ to mean $\Pr(S = s | H = h)$, and similarly for other random variables (capital letters) and their realizations (lowercase letters).

Step 1: training the model. The model is trained from large cross-sectional observational cohorts of chronically infected, therapy-naïve individuals for whom HIV sequence and linked HLA types are available. The process begins with feature selection (step 1a). By assuming independence among sites, we are able to estimate independent per-site models. Importantly, any given site is unlikely to be under selection pressure from more than a couple of HLA alleles. Indeed, in a recent large-scale HLA association study using our subtype B training data, there were an average of 1.3 HLA alleles associated per site that was associated with at least one HLA allele³⁶. Thus, our first step is to identify site-specific HLA alleles, so that estimation of the probability distribution is parameterized only by those alleles for which there is statistical evidence of selection. To this end, we use previously published approaches to identify HLA associations^{35,36,104,105}. Of note, these methods treat each individual amino acid at each site independently, as this simplifies the model and increases statistical power. The result is a list of HLA-AA pairs with a corresponding q -value that is an estimate of the proportion of associations that are false positives among those associations that are deemed significant at the corresponding threshold⁹². We chose $q < 0.2$ as our threshold. The full list of HLA-AA associations is available in the **Supplementary Data Set**.

These methods are based on the phylogenetically corrected logistic regression model, which models $\Pr(S_i | h, t_i)$ using logistic regression, with 0/1 binary features for each HLA allele and a $-1/1$ binary feature for the transmitted state t_i . Thus, we can specify the model for amino acid a at site i as

$$\Pr(S_{ia} = 1 | h, t_{ia}) = \frac{1}{1 + e^{-z}}$$

$$z = h\beta + \beta_0 t_{ia}$$

for the $1 \times M$ binary feature vector h encoding the HLA alleles expressed by the individual, the $M \times 1$ parameter vector β and the scalar offset parameter β_0 . Under this model, in the absence of HLA-mediated selection pressure, the log-odds that $S_{ij} = 1$ is β_0 if the individual was transmitted amino acid a and $-\beta_0$ if the individual was transmitted any other amino acid. To perform feature selection, we start with the null model of no selection pressure ($\beta = 0$), then systematically test each HLA j using maximum likelihood to optimize β_j and β_0 . The HLA that results in the maximum likelihood is allowed to stay in the model, and the process is repeated until no HLA yields a significant addition at $P < 0.05$ by the likelihood ratio test. These P values are used to estimate false discovery rates, which are estimated over all amino acids at all sites within a single protein. All tests at $q < 0.2$ were treated as significant, and the remaining tests had their weights set to $\beta_j = 0$.

Because the transmitted AA was not observed, we average over all possible states ($t_{ia} = 1$ and $t_{ia} = -1$). The probability distribution $\Pr(T_{ia} = 1)$ is estimated from the phylogeny, which is parameterized as a continuous-time Markov process with a reversible substitution-rate matrix. To this end, we begin with a phylogeny, the structure of which is estimated using PhyML 3.0 (ref. 106). For each site, we estimate a general time-reversible (GTR) substitution-rate matrix R (under two states) and a stationary binary AA probability distribution π using the expectation maximization (EM) algorithm^{107,108}. For each expectation step, we fix the β parameters from the logistic regression portion of the model, as well as phylogenetic parameters, then estimate the marginal and pairwise-marginal

distributions of all hidden nodes in the tree, including the hidden nodes that represent the transmitted virus. Using these marginal distributions, we then maximize the likelihood with respect to both the phylogenetic parameters and the logistic regression parameters (β). The process is repeated until convergence. The details of the EM algorithm for phylogenies are given in ref. 108. Maximization of the logistic regression parameters, conditional on $\Pr(T_{ia} = 1)$, is achieved through the creation of fractional observations for each individual, where $\Pr(T_{ia} = 1)$, estimated for each individual in the tree, defines the weight for each fractional observation. For binary models, the matrix exponentials required for the continuous-time Markov process can be computed analytically, leading to substantial simplification and speedup. Additional steps can be taken to deal with ambiguous HLA data, which involve treating the high-resolution HLA variables as missing data, conditional on the low-resolution types and estimated haplotype frequencies, as previously described³⁶.

The next step (step 1b) is multinomial logistic regression. The methods described for step 1a have been used previously to great effect in identifying HIV residues that are likely to serve as adapted or nonadapted residues in the context of HLA-mediated immune escape³. However, because they treat amino acids at the same position as independent, they do not yield consistent estimates of the probability distribution over all amino acids at a site. To address this, we describe here a modification of the phylogenetically corrected logistic regression algorithm that uses a multinomial GTR substitution model in the phylogeny and multinomial logistic regression model to estimate $\Pr(s_i | h, t_i)$.

For a multinomial logistic regression model with A states and M predictors, we can define the probabilities for each of the A states as

$$\Pr(S_i) = AA_a | h, t_i = \frac{1}{Z} e^{\beta_a \cdot h + \beta_0 (2\mathbf{1}(t_i = AA_a) - 1)}$$

$$Z = \sum_{b=1}^A e^{\beta_a \cdot h + \beta_0 (2\mathbf{1}(t_i = AA_b) - 1)}$$

where AA_a denotes the a th amino acid, β_a is the $M \times 1$ parameter vector for AA_a , and $\mathbf{1}(\cdot)$ is evaluated as 1 if the contents are true and as 0 otherwise. Thus the transmitted amino acid places β_0 weight in favor of the same amino acid and $-\beta_0$ weight against all other amino acids.

To set up this model, we begin with all amino acids observed in at least three of our training sequences, then add 'X' to represent all other amino acids. The space of HLA variables that are allowed to have nonzero weights is taken as the union of all HLA variables that were associated with any amino acid at site i in step 1a. The weights are then chosen to maximize the likelihood, subject to an L1-norm regularization penalty term that subtracts $\lambda_i (\sum_{a,j} |\beta_{a,j}| + |\beta_0|)$, $j = 1, \dots, M$, $a = 1, \dots, A$, with λ_i chosen independently for each site using sevenfold cross-validation.

As in step 1a, the distribution over the transmitted amino acid, $\Pr(t_i)$, is taken from the marginal distribution for the hidden node that represents the transmitted sequence in the phylogeny, which in turn is affected by the parameters of the multinomial logistic regression model, as well as the observed amino acids at the tips of the tree and the phylogenetic parameters, which consist of the substitution rate matrix R and stationary distribution π . Optimization of these last parameters is carried out using an EM algorithm, with the M-step including maximization of both phylogenetic and logistic regression parameters, conditional on the inferred marginal and pairwise marginal distributions on the internal nodes. We parameterize the phylogenetic model as a site-specific GTR model with A states. The model is trained with a modification of the algorithm described by Holmes and Rubin¹⁰⁸, as described in the next section. Of note, the trained parameters (R, π) represent HLA-corrected, site-specific estimates of the standard phylogenetic parameters, with π representing the steady-state amino acid frequency distribution, and R representing the transition rate matrix, each corrected for the effect of HLA-mediated selection and the phylogenetic structure.

EM algorithm for multistate phylogenies. Holmes and Rubin¹⁰⁸ describe an EM algorithm for maximizing the likelihood of a continuous time Markov process over A states with respect to the substitution-rate matrix R and the stationary-state probabilities π . Their model is defined for general (nonreversible) R and π , with reversibility heuristically imposed and the constraints that all substitution rates be positive and that π defines a probability distribution imposed by

Lagrange multipliers. In practice, we found that these approaches were numerically unstable and frequently resulted in invalid R (e.g., nonreversible or containing negative entries) and π (containing negative entries or failing to sum to one). We therefore modified the procedure as follows.

We start by following Holmes and Rubin in calculating the expected complete log likelihood with respect to given parameters (R, π) and updated parameters (R', π') as

$$Q(R, \pi, R', \pi') = \sum_a \hat{s}_a \log \pi'_a + \sum_a \hat{w}_a R'_{aa} + \sum_a \sum_{b \neq a} \hat{u}_{ab} \log R'_{ab}$$

where a and b index into the A possible amino acid states observable at site i , and \hat{s}_j , \hat{w}_j and \hat{u}_{jk} are sufficient statistics computed from (R, π) and described by Holmes and Rubin¹⁰⁸. These correspond, respectively, to the expected number of substitution paths that start in state a , the expected time spent in state a , and the expected number of $a \rightarrow b$ transitions. Whereas Holmes and Rubin introduce Lagrange multipliers to Q to enforce boundary constraints, maximize the resulting expression with respect to (R', π') and then make a heuristic correction to ensure that R' is reversible, we begin by parameterizing R with respect to its stationary probability distribution π to enforce reversibility. Specifically, we define the substitution rate matrix as

$$[R]_{ab} = \begin{cases} \pi_b \lambda_{ab} & \text{if } a \neq b \\ -\sum_{b \neq a} [R]_{ab} & \text{if } a = b \end{cases}$$

We then numerically maximize Q with respect to (R', π') by computing the gradient of Q , which yields

$$\frac{\partial Q}{\partial \pi'_a} = \frac{\hat{s}_a}{\pi'_a} - \sum_{b \neq a} \hat{w}_b \lambda'_{ab} + \frac{1}{\pi'_a} \sum_{b \neq a} \hat{u}_{ba}$$

$$\frac{\partial Q}{\partial \lambda'_{ab}} = -\hat{w}_a \pi'_b - \hat{w}_b \pi'_a + \frac{1}{\lambda'_{ab}} (\hat{u}_{ab} + \hat{u}_{ba})$$

To further enforce that $\lambda_{ab} \geq 0$, $a \geq 0$ and $\sum_{b=1}^A \pi_a = 1$, we reparameterize π_a as

$$\pi_a = \frac{\exp(\alpha_a)}{\sum_{b=1}^A \exp(\alpha_b)}$$

and λ_{ab} as

$$\lambda_{ab} = \exp(\beta_{ab})$$

For $a \neq b$, these functions have gradients

$$\frac{\partial \pi'_a}{\partial \alpha_a} = \pi'_a (1 - \pi'_a)$$

$$\frac{\partial \pi'_a}{\partial \alpha_b} = -\pi'_a \pi'_b$$

$$\frac{\partial \lambda'_{ab}}{\partial \beta_{ab}} = \lambda'_{ab}$$

Thus, we maximize Q with respect to α and β using

$$\frac{\partial Q}{\partial \alpha_a} = \frac{\partial Q}{\partial \pi'_a} \frac{\partial \pi'_a}{\partial \alpha_a} = \frac{\partial Q_a}{\partial \pi'_a} \pi'_a - \pi'_a \sum_b \frac{\partial Q}{\partial \pi'_b} \pi'_b$$

and

$$\frac{\partial Q}{\partial \beta_{ab}} = \frac{\partial Q}{\partial \lambda'_{ab}} \frac{\partial \lambda'_{ab}}{\partial \beta_{ab}} = -\lambda'_{ab} (\hat{w}_a \pi'_b + \hat{w}_b \pi'_a) + \hat{u}_{ab} + \hat{u}_{ba}$$

This parameterization allows us to use gradient-based optimizers that expect unbounded parameters, while assuring reversibility of R and appropriate bounds

on the rates and probabilities. For optimization, we use an efficient implementation of L-BFGS quasi-Newton optimization¹⁰⁹.

Step 2: defining the adaptation of a sequence with respect to an HLA allele h . First we estimate the probability of a sequence in an HLA context (step 2a). Step 1 yields a model that allows the estimation of the probability of observing any given AA at site i in the context of any set of HLA alleles and given any transmitted amino acid. As our ultimate goal is to measure adaptation, we consider that a useful distribution over the transmitted AA, $\Pr(t_i)$, is our estimate of the ancestral (equivalently, steady-state) distribution, provided by π , which represents an HLA-corrected estimate of the 'ideal' distribution in the absence of HLA pressure and corrected for the phylogenetic structure of the observed sequences. We therefore estimate

$$\Pr(S_i = s_i | H = h) = \sum_a \Pr(S_i = s_i | H = h, T_i = AA_a) \pi_a$$

where π_a is the stationary probability of amino acid a . Assuming independence of sites yields

$$\Pr(S = s | H = h) = \prod_i \Pr(S_i = s_i | H = h)$$

which is an estimate of the probability of observing any particular sequence in a chronically infected individual who expresses HLA alleles h .

The next step (step 2b) is to define the adaptation score. Step 2a yields a probability distribution over HIV sequences conditional on a set of HLA alleles. In practice, we find it most helpful to consider a single HLA allele at a time. Thus, if $|h| = 1$, then $\Pr(S = s | H = h)$ is the probability of observing s in an individual whose CTLs are targeting only epitopes presented by h . We then define the adaptation of s to h as

$$\text{Adapt}_h(s) \equiv g \left(\frac{\Pr(S = s | H = h)}{\Pr(S = s | H = \emptyset)} \right)$$

$$g(x) = \frac{2}{\pi} \arctan[\ln(x)]$$

where $H = \emptyset$ is a vector with all HLA variables set to zero, representing an HIV sequence evolving in the absence of immune pressure. The transformation $g(x)$ maps the ratio to a heavy-tailed sigmoidal function on the range $(-1, 1)$, with 0, 0.75, 0.85 and 0.90 respectively corresponding to the cases where the HIV sequence is equally, ~ 10 -fold more, ~ 100 -fold more or $\sim 1,000$ -fold more likely in the context of the HLA allele than in the absence of any selection pressure. Because we define $\Pr(t) = \pi$, the adaptation score has the intuitive interpretation as a measure of the extent to which the deviation of s from the idealized ancestral sequence is due to selection pressure mediated by h . Further, by defining the adaptation score in terms of a null immune response, we naturally normalize for variations in sequence coverage due to incomplete or ambiguous sequencing.

Notably, our independence-of-sites assumption allows the straightforward combination of adaptation scores computed from two different genomic regions. For example, if we have computed adaptation of Gag and Nef with respect to an HLA allele h , then we have

$$\begin{aligned} \text{Adapt}_h(\text{Gag} + \text{Nef}) &= g(g^{-1}(\text{Adapt}_h(\text{Gag})) \times g^{-1}(\text{Adapt}_h(\text{Pol}))) \\ &\approx \text{Adapt}_h(\text{Gag}) + \text{Adapt}_h(\text{Pol}) \end{aligned}$$

The shape of the inverse tangent function is such that the approximation is close to equality when the adaptation scores of the regions are both between -0.5 and $+0.5$.

Because the models are trained on high-resolution HLA types, $\text{Adapt}_s(h)$ must be extended to cover low- and medium-resolution data sets. If h_j represents a low- or medium-resolution HLA type, with corresponding subtypes h_{jk} , $k = 1, \dots, K$, then the adaptation of s with respect to h_j is defined as the weighted average of the adaptation of s to the possible subtypes,

$$\text{Adapt}_{h_j}(s; \theta) = \sum_k \text{Adapt}_{h_{jk}}(s) \cdot \Pr(h_{jk} | h_j, \theta)$$

with θ parameterizing the ethnicity-specific distribution of HLA subtypes. In our experiments, $\Pr(h_{jk} | h_j, \theta)$ was taken from a modification of a published statistical

HLA haplotype completion tool⁹⁰. Our modification allowed for averaging over uncertain ethnicities when the ethnicity of individuals was unknown but the distribution over a population could be obtained from external sources.

Finally, when h represents a set of alleles (such as $h = \{h_{a1}, h_{a2}, h_{b1}, h_{b2}, h_{c1}, h_{c2}\}$, representing an individual's full class I repertoire), then adaptation is defined as the average adaptation score over the set of alleles:

$$\text{Adapt}_h(s) = \frac{1}{|h|} \sum_{h_k \in h} \text{Adapt}_{h_k}(s)$$

Thus, we compute the adaptation score for each of an individual's HLA alleles separately, then use those numbers to compute adaptation scores for each locus and for the entire repertoire. Although our model could instead be used to estimate the distribution of s conditional on a set of alleles, we found it more intuitive to think of adaptation of s to a particular allele as independent of the other alleles expressed by an individual. Moreover, because most sites are not under selection by multiple alleles, and because when multiple selection does occur, most individuals do not express both alleles, the fully conditional adaptation scores were highly correlated ($R > 0.97$) to the definition we used.

The distribution of autologous sequences to an individual's HLA alleles in our HIVB cohort is shown in **Figure 1**. The mean was 0.26, which indicates that $\Pr(S = s | H = h)$ is approximately 1.5-fold more likely than $\Pr(S = s | H = \emptyset)$, with a minimum of -0.44 (2.3-fold less likely) and a maximum of 0.8 (22-fold more likely). For HIVC these numbers were, respectively, 0.18 (1.3-fold), -0.49 (-2.6 -fold) and 0.99 (10^{27} -fold). When the numbers are transformed in this way, large differences in fold (say, $1,000$ – 10^{27}) yield a small difference in adaptation score (0.9–0.99), and thus a small difference when used as a linear predictor of clinical outcomes. This property increases robustness against (for example) model overfitting or errors in HLA typing and yields an approximately normal distribution of adaptation scores in any given population (**Fig. 1** and **Supplementary Fig. 2a**).

Adaptation similarity. For the purpose of adaptation, we consider that two HLA alleles h_1 and h_2 are similar if they drive similar escape mutations. In the context of the adaptation score, this suggests that the similarity of h_1 and h_2 can be defined as the Pearson correlation coefficient, ρ_{h_1, h_2} , between the two alleles, over the entire population of HIV sequences. In practice, we must estimate this correlation over a set of observed sequences. Here we will use all Southern Africa sequences that are not missing entire protein sequences. We thus estimate the sample Pearson correlation coefficient, η_{h_1, h_2} , over these sequences. To account for sequence features such as gaps, missing regions and AA mixtures, we perform an additional normalizing step. Specifically, we first compute the matrix $X = \{x_{ij}\}$, where $x_{ij} = \text{Adapt}_{s_i}(h_j)$ is the adaptation of the i th sequence to the j th allele, and then mean-center each row. The sample 'adaptation similarity' of h_1 and h_2 is then defined as the Pearson correlation coefficient between columns i and j of the resulting matrix. The resulting HLA-specific similarity largely recreates supertype definitions (**Supplementary Data Set**). At the subject level, we extend the above definition such that h_i, h_j refers to the sets of alleles (all alleles, or all alleles at one of the loci). Here we focused on HLA-B adaptation similarity between donor and recipient pairs, as adaptation to HLA-B consistently had the largest effect size in all the previous analyses (**Supplementary Tables 1–3**).

Accession codes. The following accessions (all from GenBank) were used in the current study. Durban, South Africa (Southern Africa): FJ198407–FJ199088, EU698132–EU698633, AY838569–AY838639, HM593106–HM593510 (gag); FJ199532–FJ199992, EU698737–EU698888 (pol); FJ199089–FJ199531, EU698634–EU698736, AY838640–AY838756 (nef); AY463217–AY772701, AY838639, AY838567, AY878054–AY878072, AY901965–AY901981, DQ011165–DQ011180, DQ056404–DQ093607, DQ164104–DQ164129, DQ275642–DQ275665, DQ351216–DQ351238, DQ369976–DQ396400, DQ445631–DQ445637 (full length). Bloemfontein, South Africa (Southern Africa): KT736510–KT736715 (gag); KT736966–KT737213 (pol); KT736716–KT736965 (nef). Kimberley, South Africa (Southern Africa): KT860066–KT860091 (gag). Gabarone, Botswana (Southern Africa): FJ497801–FJ497951, KT860175–KT860351 (gag); FJ498244–FJ498543, KT860352–KT860415 (pol); FJ498544–FJ498778, KT860120–KT860174 (nef). Thames Valley Cohort (Southern Africa): FJ645274–FJ645344, FJ645350–FJ645360, FJ645409–FJ645410 (gag); FJ645411–FJ645478, FJ645483–FJ645488, FJ645534–FJ645538 (pol);

KT860092–KT860119 (nef). British Columbia, Canada (IHAC): EU241938–EU242504 (gag); GQ303719–GQ304249, EF368373–EF368603, EF368604–EF369427 (Pr-RT); FJ812899–FJ813480 (integrase); JX147785–JX148365 (tat/rev exon 1); JX147023–JX147784 (gp41; tat/rev exon 2); JX148366–JX148914 (vpu); JX148915–JX149509 (vif); DQ203856–DQ204405, EF567317–EF567389 (vpr); DQ484067–DQ485128 (nef). Western Australian HIV Cohort Study (IHAC): AY856956–AY857186 (full length). US AIDS Clinical Trials Group protocols 5142 and 5128 (IHAC): GQ371216–GQ371763 (gag); GQ371764–GQ372317 (pol); GQ372318–GQ372824, GQ398382–GQ398387 (nef); GU727870–GU731062 (env and accessory). Ragon Elite Controller: EU517772–517812 (gag); EU517898–EU517938 (protease); EU517972–EU518012 (reverse transcriptase); EU517859–EU517897 (integrase); EU518046–EU518086 (vif); EU518088–EU518127 (vpr); EU517721–EU517760 (vpu); EU518013–EU518044 (tat); EU517815–EU517970 (rev); GU046566–GU046603 (nef). Ragon Non-Controllers: DQ886031, DQ886038, FJ469682–FJ469772, JQ403024–JQ403086, JQ403091 (full length). Zambian Transmission Pairs: KM048382–KM049006 (gag); KM049900–KM050767 (pol); KM049007–KM049899 (nef). Step Study: JF320002–JF320643 (full length).

Code availability. Implementation of the adaptation score and adaptation similarity are available as a web service and downloadable software at <https://phylod.research.microsoft.com>.

56. Wright, J.K. *et al.* Gag-protease-mediated replication capacity in HIV-1 subtype C chronic infection: associations with HLA type and clinical parameters. *J. Virol.* **84**, 10820–10831 (2010).
57. Kiepiela, P. *et al.* Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* **432**, 769–775 (2004).
58. Huang, K.-H.G. *et al.* Prevalence of HIV type-1 drug-associated mutations in pre-therapy patients in the Free State, South Africa. *Antivir. Ther.* **14**, 975–984 (2009).
59. Matthews, P.C. *et al.* HLA-A*7401-mediated control of HIV viremia is independent of its linkage disequilibrium with HLA-B*5703. *J. Immunol.* **186**, 5675–5686 (2011).
60. Shapiro, R.L. *et al.* Antiretroviral regimens in pregnancy and breast-feeding in Botswana. *N. Engl. J. Med.* **362**, 2282–2294 (2010).
61. Brumme, Z.L. *et al.* Evidence of differential HLA class I-mediated viral evolution in functional and accessory/regulatory genes of HIV-1. *PLoS Pathog.* **3**, e94 (2007).
62. Brumme, Z.L. *et al.* HLA-associated immune escape pathways in HIV-1 subtype B Gag, Pol and Nef proteins. *PLoS One* **4**, e6687 (2009).
63. Bhattacharya, T. *et al.* Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science* **315**, 1583–1586 (2007).
64. Moore, C.B. *et al.* Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science* **296**, 1439–1443 (2002).
65. Mallal, S.A. The Western Australian HIV Cohort Study, Perth, Australia. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.* **17**, S23–S27 (1998).
66. John, M. *et al.* Adaptive interactions between HLA and HIV-1: highly divergent selection imposed by HLA class I molecules with common supertype motifs. *J. Immunol.* **184**, 4368–4377 (2010).
67. Haas, D.W. *et al.* A multi-investigator/institutional DNA bank for AIDS-related human genetic studies: AACTG Protocol A5128. *HIV Clin. Trials* **4**, 287–300 (2003).
68. Poon, A.F.Y. *et al.* The impact of clinical, demographic and risk factors on rates of HIV transmission: a population-based phylogenetic analysis in British Columbia, Canada. *J. Infect. Dis.* **211**, 926–935 (2015).
69. Miura, T. *et al.* Genetic characterization of human immunodeficiency virus type 1 in elite controllers: lack of gross genetic defects or common amino acid changes. *J. Virol.* **82**, 8422–8430 (2008).
70. Wang, Y.E. *et al.* Protective HLA class I alleles that restrict acute-phase CD8+ T-cell responses are associated with viral escape mutations located in highly conserved regions of human immunodeficiency virus type 1. *J. Virol.* **83**, 1845–1855 (2009).
71. Henn, M.R. *et al.* Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog.* **8**, e1002529 (2012).
72. Frahm, N. *et al.* Increased sequence diversity coverage improves detection of HIV-specific T cell responses. *J. Immunol.* **179**, 6638–6650 (2007).
73. McKenna, S.L. *et al.* Rapid HIV testing and counseling for voluntary testing centers in Africa. *AIDS* **11** (suppl. 1), S103–S110 (1997).
74. Kempf, M.-C. *et al.* Enrollment and retention of HIV discordant couples in Lusaka, Zambia. *J. Acquir. Immune Defic. Syndr.* **47**, 116–125 (2008).
75. Allen, S. *et al.* Promotion of couples' voluntary counselling and testing for HIV through influential networks in two African capital cities. *BMC Public Health* **7**, 349 (2007).
76. Trask, S.A. *et al.* Molecular epidemiology of human immunodeficiency virus type 1 transmission in a heterosexual cohort of discordant couples in Zambia. *J. Virol.* **76**, 397–405 (2002).
77. Yue, L. *et al.* Cumulative impact of host and viral factors on HIV-1 viral-load control during early infection. *J. Virol.* **87**, 708–715 (2013).
78. McMichael, A.J., Borrow, P., Tomaras, G.D., Goonetilleke, N. & Haynes, B.F. The immune response during acute HIV-1 infection: clues for vaccine development. *Nat. Rev. Immunol.* **10**, 11–23 (2010).
79. Salazar-Gonzalez, J.F. *et al.* Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. *J. Exp. Med.* **206**, 1273–1289 (2009).
80. Sidney, J. *et al.* Measurement of MHC/peptide interactions by gel filtration or monoclonal antibody capture. *Curr. Protoc. Immunol.* Chapter 18, Unit 18.3 (2013).
81. Lundegaard, C. *et al.* NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. *Nucleic Acids Res.* **36**, W509–W512 (2008).
82. Lundegaard, C., Lund, O. & Nielsen, M. Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. *Bioinformatics* **24**, 1397–1398 (2008).
83. Bansal, A. *et al.* CD8 T cell response and evolutionary pressure to HIV-1 cryptic epitopes derived from antisense transcription. *J. Exp. Med.* **207**, 51–59 (2010).
84. Bansal, A. *et al.* Immunological control of chronic HIV-1 infection: HLA-mediated immune function and viral evolution in adolescents. *AIDS* **21**, 2387–2397 (2007).
85. Bansal, A. *et al.* Multifunctional T-cell characteristics induced by a polyvalent DNA prime/protein boost human immunodeficiency virus type 1 vaccine regimen given to healthy adults are dependent on the route and dose of administration. *J. Virol.* **82**, 6458–6469 (2008).
86. Roederer, M., Nozzi, J.L. & Nason, M.C. SPICE: exploration and analysis of post-cytometric complex multivariate datasets. *Cytometry A* **79**, 167–174 (2011).
87. Akinsiku, O.T., Bansal, A., Sabbaj, S., Heath, S.L. & Goepfert, P.A. Interleukin-2 production by polyfunctional HIV-1-specific CD8 T cells is associated with enhanced viral suppression. *J. Acquir. Immune Defic. Syndr.* **58**, 132–140 (2011).
88. Ndung'u, T., Renjifo, B. & Essex, M. Construction and analysis of an infectious human immunodeficiency virus type 1 subtype C molecular clone. *J. Virol.* **75**, 4964–4972 (2001).
89. Box, G.E.P. & Cox, D.R. An analysis of transformations. *J. R. Stat. Soc. Series B. Stat. Methodol.* **26**, 211–252 (1964).
90. Listgarten, J. *et al.* Statistical resolution of ambiguous HLA typing data. *PLoS Comput. Biol.* **4**, e1000016 (2008).
91. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
92. Storey, J.D. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Stat.* **31**, 2013–2035 (2003).
93. Vance, D.E., Mugavero, M., Willig, J., Raper, J.L. & Saag, M.S. Aging with HIV: a cross-sectional study of comorbidity prevalence and clinical characteristics across decades of life. *J. Assoc. Nurses AIDS Care* **22**, 17–25 (2011).
94. Carrington, M. & O'Brien, S.J. The influence of HLA genotype on AIDS. *Annu. Rev. Med.* **54**, 535–551 (2003).
95. Pereyra, F. *et al.* The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* **330**, 1551–1557 (2010).
96. Nagelkerke, N.J.D. A note on a general definition of the coefficient of determination. *Biometrika* **78**, 691–692 (1991).
97. Self, S.G. & Liang, K.-Y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.* **82**, 605 (1987).
98. Leslie, A. *et al.* Differential selection pressure exerted on HIV by CTL targeting identical epitopes but restricted by distinct HLA alleles from the same HLA supertype. *J. Immunol.* **177**, 4699–4708 (2006).
99. Payne, R.P. *et al.* Differential escape patterns within the dominant HLA-B*57:03-restricted HIV Gag epitope reflect distinct clade-specific functional constraints. *J. Virol.* **88**, 4668–4678 (2014).
100. Brumme, Z.L. *et al.* Human leukocyte antigen-specific polymorphisms in HIV-1 Gag and their association with viral load in chronic untreated infection. *AIDS* **22**, 1277–1286 (2008).
101. Brumme, Z.L. *et al.* Marked epitope- and allele-specific differences in rates of mutation in human immunodeficiency type 1 (HIV-1) Gag, Pol, and Nef cytotoxic T-lymphocyte epitopes in acute/early HIV-1 infection. *J. Virol.* **82**, 9216–9227 (2008).
102. Marsh, S.G.E. *et al.* Nomenclature for factors of the HLA system, 2010. *Tissue Antigens* **75**, 291–455 (2010).
103. Sidney, J., Peters, B., Frahm, N., Brander, C. & Sette, A. HLA class I supertypes: a revised and updated classification. *BMC Immunol.* **9**, 1 (2008).
104. Carlson, J.M. *et al.* Phylogenetic dependency networks: inferring patterns of CTL escape and codon covariation in HIV-1 Gag. *PLoS Comput. Biol.* **4**, e1000225 (2008).
105. Carlson, J., Kadie, C., Mallal, S.A. & Heckerman, D. Leveraging hierarchical population structure in discrete association studies. *PLoS One* **2**, e591 (2007).
106. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
107. Dempster, A.P., Laird, N.M. & Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series B. Stat. Methodol.* **39**, 1–38 (1977).
108. Holmes, I. & Rubin, G.M. An expectation maximization algorithm for training hidden substitution models. *J. Mol. Biol.* **317**, 753–764 (2002).
109. Andrew, G. & Gao, J. Scalable training of L1-regularized log-linear models. in *Proc. 24th International Conference on Machine Learning* 33–40 (ACM, 2007).