JVI Journals.ASM.org

# Correlates of Protective Cellular Immunity Revealed by Analysis of Population-Level Immune Escape Pathways in HIV-1

**Jonathan M. Carlson,[a] Chanson J. Brumme,[b] Eric Martin,[c] Jennifer Listgarten,[a] Mark A. Brockman,[b,c] Anh Q. Le,[c] Celia K. S. Chui,[b] Laura A. Cotton,[c] David J. H. F. Knapp,[b] Sharon A. Riddler,[d] Richard Haubrich,[e] George Nelson,[f] Nico Pfeifer,[a] Charles E. DeZiel,[a] David Heckerman,[a] Richard Apps,[g] Mary Carrington,[g] Simon Mallal,[h,i] P. Richard Harrigan,[b] Mina John,[h,i] Zabrina L. Brumme,[b,c] and the International HIV Adaptation Collaborative**

Microsoft Research, Los Angeles, California, USA[a]; British Columbia Centre for Excellence in HIV/AIDS, Vancouver, Canada[b]; Faculty of Health Sciences, Simon Fraser University, Burnaby, British Columbia, Canada[c]; Department of Infectious Diseases and Microbiology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA[d]; Department of Medicine, University of California San Diego, San Diego, California, USA[e]; Basic Research Program, Center for Cancer Research Genetics Core, SAIC-Frederick, Inc., Frederick National Laboratory for Cancer Research, Frederick, Maryland, USA[f]; Cancer and Inflammation Program, Laboratory of Experimental Immunology, SAIC-Frederick, Inc., Frederick National Laboratory for Cancer Research, Frederick, Maryland, USA, and Ragon Institute of Massachusetts General Hospital, MIT, and Harvard, Charlestown, Massachusetts, USA[g]; Centre for Clinical Immunology and Biomedical Statistics, Institute for Immunology and Infectious Diseases, Murdoch University, Murdoch, Western Australia, Australia[h]; and Department of Clinical Immunology, Royal Perth Hospital, Perth, Western Australia, Australia[i]

**HLA class I-associated polymorphisms identified at the population level mark viral sites under immune pressure by individual HLA alleles. As such, analysis of their distribution, frequency, location, statistical strength, sequence conservation, and other properties offers a unique perspective from which to identify correlates of protective cellular immunity. We analyzed HLA-associated HIV-1 subtype B polymorphisms in 1,888 treatment-naïve, chronically infected individuals using phylogenetically informed methods and identified characteristics of HLA-associated immune pressures that differentiate protective and nonprotective alleles. Over 2,100 HLA-associated HIV-1 polymorphisms were identified, approximately one-third of which occurred inside or within 3 residues of an optimally defined cytotoxic T-lymphocyte (CTL) epitope. Differential CTL escape patterns between closely related HLA alleles were common and increased with greater evolutionary distance between allele group members. Among 9-mer epitopes, mutations at HLA-specific anchor residues represented the most frequently detected escape type: these occurred nearly 2-fold more frequently than expected by chance and were computationally predicted to reduce peptide-HLA binding nearly 10-fold on average. Characteristics associated with protective HLA alleles (defined using hazard ratios for progression to AIDS from natural history cohorts) included the potential to mount broad immune selection pressures across all HIV-1 proteins except Nef, the tendency to drive multisite and/or anchor residue escape mutations within known CTL epitopes, and the ability to strongly select mutations in conserved regions within HIV's structural and functional proteins. Thus, the factors defining protective cellular immune responses may be more complex than simply targeting conserved viral regions. The results provide new information to guide vaccine design and immunogenicity studies.**

HIV-1 is notorious for its genetic diversity and its ability to adapt to selection pressures (44, 87, 116). Despite this, within-host HIV-1 evolution in response to antiretroviral (61, 72), host cellular immune (15, 50, 71, 94, 95), antibody (46), and vaccine-induced (103) selection pressures occurs along generally predictable mutational pathways (3, 84). Studying these evolutionary pathways can offer insight into the immunopathogenesis of HIV-1 and may help inform the design of immune-based interventions and vaccines.

Substantial progress has been made in our understanding of HIV-1's ability to evade human leukocyte antigen (HLA) class I-restricted CD8$^+$ cytotoxic T-lymphocytes (CTL). In particular, application of novel statistical methods (13, 25, 84) to large population-based data sets of linked host and viral genetic information has facilitated the systematic identification of HLA-associated immune escape and covarying mutations in HIV-1 (19, 20, 60, 81, 99, 104), revealing important insights into HIV-1 adaptation to its host. We now appreciate that immune selection represents a major force shaping HIV-1 diversity (3, 80, 84) and that HIV-1 escape pathways are generally predictable in the context of host HLA allele expression (3, 19, 84). Immune escape mutations occur within and outside CTL epitopes (13, 20, 60, 99, 104) and can compromise peptide-HLA binding (8, 65), disrupt intracellular antigen processing (35, 122), affect T-cell recognition of the peptide-HLA complex (23, 58, 59, 96, 117), and/or potentially affect killer immunoglobulin-like receptor (KIR) binding (16, 113). While some escape pathways are likely to be universal across HLA alleles and/or HIV-1 subtypes, widespread examples of differential escape between infected populations (9, 60) and between closely related HLA class I alleles (26, 69, 86) have also been demonstrated. Population-level studies have also allowed us to estimate rates (18) and clinical implications (21, 81) of immune escape, infer fitness costs of specific mutations (81), discover novel epitopes in conventional (5, 13) and cryptic (10, 12) reading

Received 1 August 2012 Accepted 2 October 2012
Published ahead of print 10 October 2012
Address correspondence to Zabrina L. Brumme, zbrumme@sfu.ca, or Jonathan M. Carlson, carlson@microsoft.com.
J.M.C., C.J.B., and E.M. contributed equally to this article.
Supplemental material for this article may be found at http://jvi.asm.org/.
Copyright © 2012, American Society for Microbiology. All Rights Reserved.
doi:10.1128/JVI.01998-12
The authors have paid a fee to allow immediate free access to this article.

Downloaded from http://jvi.asm.org/ on November 13, 2018 by guest

frames, and forecast the consequences of continued immune-mediated adaptation for the future of the epidemic (64, 73). These and other findings have led to recommendations that escape information be incorporated into HIV-1 vaccine strategies, for example, via immunogens that incorporate variant sequences (13, 100) and/or that are located in mutationally constrained regions where viral escape mutations would be anticipated to incur substantial fitness costs (7, 30, 102).

Although population-level studies have significantly advanced our understanding of the specific mutational pathways whereby HIV-1 escapes from HLA-restricted immune pressures, much remains to be learned. The key to further exploring these data lies in the recognition that HLA-associated polymorphisms identified at the population level serve as markers of viral sites under strong *in vivo* immune pressure by individual HLA alleles—sufficiently strong that the virus responds by escape. As such, analysis of the distribution, frequency, location, statistical strength, and sequence conservation of HLA-associated polymorphic sites identified at the population level offers a unique perspective from which to identify (albeit indirectly) specific features that render certain immune responses more effective at controlling HIV progression than others. In other words, systematic analysis of HLA-driven evolutionary "imprints" on the HIV proteome can help identify correlates of protective cellular immunity.

In the largest population-level, proteome-wide analysis of immune escape in HIV-1 undertaken to date, we refine existing immune escape data by reporting HLA-associated polymorphisms at HLA supertype-, type-, and subtype-level resolutions across the viral proteome in an HIV-1 subtype B context. We analyze this well-powered data set to investigate the characteristics of HLA-driven immune selection on HIV-1 in order to draw inferences regarding the general mechanisms HIV-1 uses *in vivo* to evade such pressures. Furthermore, by treating escape mutation pathways as evolutionary markers of strong HLA-restricted immune pressure by CTL, we identify specific features of these responses that differentiate protective from nonprotective HLA class I alleles.

## MATERIALS AND METHODS

**IHAC.** The International HIV Adaptation Collaborative (IHAC) is an open multicenter cohort of chronically infected antiretroviral-naïve individuals from Canada, the United States, and Australia for whom HLA class I and nearly full-genome HIV plasma RNA sequences have been characterized. The present study was restricted to 1,888 HIV-1 subtype B-infected individuals (>95% of total cohort), including 1,103 individuals from the British Columbia HOMER cohort (British Columbia, Canada) (19, 20), 247 individuals from the Western Australian HIV Cohort Study (WAHCS; Western Australia, Australia) (13, 77, 84), and 538 U.S. AIDS Clinical Trials Group (ACTG) protocol 5142 participants (4, 60) who also provided human DNA under ACTG protocol 5128 (51). This newly expanded cohort is approximately one-third larger than that studied in 2009 (20) and now features full-proteome HIV coverage. Ethical approval was obtained from Providence Health Care/University of British Columbia (HOMER cohort), Royal Perth Hospital Ethics Committee (WAHCS), and the NIH's National Institute of Allergy and Infectious Diseases (NIAID) Clinical Science Review Committee (CSRC) (ACTG 5142/5128).

**HIV-1 sequencing.** Nearly full-genome plasma HIV-1 RNA sequencing (all regions except gp120) for the HOMER cohort was performed at the BC Centre for Excellence in HIV/AIDS (BCCfE) as described in reference 20. Briefly, HIV RNA was extracted from plasma using standard methods, and regions of interest were amplified by nested reverse transcription-PCR (RT-PCR) using HIV-specific primers. Amplicons were bulk sequenced on an Applied Biosystems 3100, 3700, and/or 3730xl automated DNA sequencer. Data were analyzed using Sequencher software (Genecodes) or the custom software RECall (120). Nucleotide mixtures were called if the secondary peak height exceeded 25% of the dominant peak (Sequencher) or if the secondary peak area exceeded 20% of the dominant peak (RECall). Nearly full-genome HIV-1 sequencing for the WAHCS and ACTG 5142/5128 cohort participants was performed at the Centre for Clinical Immunology and Biological Statistics (CCIBS) laboratory in Perth, Australia, as previously described (20, 60). Plasma HIV RNA was extracted using standard methods, and nearly complete viral genomes were amplified using nested RT-PCR. Amplicons were bulk sequenced using an Applied Biosystems 3730xl automated sequencer. Data were analyzed using semiautomated ASSIGN software with a nucleotide mixture threshold of 15% after consideration of the signal/noise ratio, yielding nearly full-genome sequences. The BCCfE and CCIBS previously undertook blinded interlaboratory genotyping quality control comparisons and observed excellent intersite concordance (20; also unpublished data).

Non-subtype B sequences were identified by comparison to subtype references in the Los Alamos HIV Database using the Recombination Identification Program (RIP [http://www.hiv.lanl.gov/content/sequence/RIP/RIP.html]) and removed from the analysis. HIV-1 sequences were aligned to subtype B reference strain HXB2 (GenBank accession number K03455). Final HLA/HIV sequence data set sizes were as follows: Gag, 1,548 sequences; Pol, 1,799 (protease/reverse transcriptase [PR/RT], 1,786; integrase [INT], 1,566); Nef, 1,685; Vif, 1,325; Vpr, 1,310; Vpu, 1,243; gp120, 655; gp41, 1,425; Tat, 1,734; and Rev, 1,731. HIV-1 sequences from the ACTG, Perth, and HOMER cohorts were previously deposited in GenBank (3, 19–21, 60). Accession numbers for additional HOMER sequences are JX147023 to JX147784 (gp41), JX147785 to JX148365 (Tat/Rev exon 1 region), JX148366 to JX148914 (Vpu), and JX148915 to JX149509 (Vif). Linked HLA/HIV data sets from the BC HOMER cohort are available for sharing with interested researchers in accordance with UBC/Providence Health Care Research Ethics Board protocols (please contact the corresponding author for information).

**HLA class I sequence-based typing and subtype imputations.** HLA class I typing for the HOMER cohort was performed at the BCCfE using an in-house sequence-based typing protocol and interpretation algorithm (19, 20, 32), yielding a mixture of intermediate- and high (subtype-level)-resolution data. High-resolution HLA class I typing for the WAHCS and ACTG 5142/5128 cohorts was performed at the CCIBS as previously described (20, 60). Allele interpretation was performed using ASSIGN (60).

In previous studies we had addressed the issue of mixed-resolution HLA data by truncating all data to the two-digit (type) level (19–21). However, in the present study we wished to identify HLA-associated polymorphisms at the supertype and subtype levels, which requires high-resolution data. Therefore, for all individuals with missing or low- or intermediate-level resolution at one or more loci (711 of 1,888; 38%), we employed a machine learning algorithm trained on a data set of complete high-resolution HLA-A, -B, and -C types from >13,000 individuals with known ethnicity (75) (available at http://research.microsoft.com/en-us/projects/bio/mbt.aspx#HLA-Completion) to complete the data to high resolution at all loci. The resulting output lists the inferred possible HLA-A, -B, and -C haplotypes, along with their respective probabilities, for each individual. Rather than assigning individuals their highest-probability HLA haplotypes, downstream analyses incorporate all HLA haplotypes with probabilities of >1% as weighted averages for each individual. HLA types could not be imputed when data were missing from two (or more) loci. In such cases, an HLA haplotype containing null values for the missing data and a probability of 1.0 were used.

**Identification of HLA-associated polymorphisms.** HLA-associated polymorphisms were identified using phylogenetically informed methods as previously described (20, 24–26), with some modifications. Briefly, a maximum-likelihood phylogenetic tree is constructed for each gene, and

Downloaded from http://jvi.asm.org/ on November 13, 2018 by guest

a model of conditional adaptation is inferred for each observed amino acid at each codon. In this model, the amino acid is assumed to evolve independently along the phylogeny until it reaches the observed hosts (tree tips). In each host, the selection pressure arising from HLA-mediated T-cell responses and amino acid covariation is directly modeled using a weighted logistic regression, in which the individual's HLA repertoire and covarying amino acids are used as predictors, and the bias is determined by the transmitted sequence (26). Because the transmitted sequence is not observed, we average over the possible transmitted sequences, as inferred from the phylogeny.

We extended this approach to incorporate predicted high-resolution HLA data as follows. For each subject in the cohort, a number of "fractional" individuals are generated, representing each possible transmitted polymorphism as well as each possible completed HLA repertoire, with fractional weights representing the probability of observing the given HLA repertoire and transmitted polymorphism, as determined by the HLA completion (see above) and phylogeny, respectively. When the observed polymorphism represents a mixture, we extend the weighting scheme to represent all possible mixture outcomes, with relative weights inversely proportional to the number of amino acids in the mixture. Maximum-likelihood estimation for logistic regression and phylogenetic parameters is jointly estimated using expectation maximization (34).

To identify which factors contribute to the selection pressure, a forward selection procedure is employed, in which the most significant association is iteratively added to the model, with $P$ values computed using the likelihood ratio test. To increase our statistical power, each codon is divided into a set of binary variables, one for each observed amino acid. In addition, we consider only HLA alleles and amino acids that are observed in at least 10 individuals in the population or for which at least 10 individuals do not express the HLA or polymorphism. In the case of imputed HLA alleles, we use the expected number of individuals expressing that allele, where expectation is taken with respect to the HLA completion probability. Statistical significance is reported using $q$ values, the $P$-value analogue of the false-discovery rate (FDR), for each $P$-value threshold (114). The FDR is the expected proportion of false positives among results deemed significant at a given threshold; for example, at a $q$ of $\leq 0.2$, we expect 20% of identified associations to be false positives. We compute $q$ values separately for each protein.

HLA-associated polymorphisms are grouped into two categories: (i) amino acids significantly enriched in the presence of the HLA allele in question ("adapted" forms) and (ii) amino acids significantly enriched in the absence of the HLA allele in question ("nonadapted" forms). To provide an example, the B*27-associated R264K is a well-known escape mutation that commonly occurs at position 2 of the KK10 epitope in $p24^{Gag}$ (65). At this codon, "R" represents the B*27-associated nonadapted form while K represents the B*27-associated adapted form. In general, nonadapted forms correspond to the subtype B consensus sequence while adapted forms correspond to polymorphic variants, but exceptions occur. In addition, we will sometimes differentiate between "direct" and "indirect" HLA-associated polymorphisms: the former represent associations that are detected when HIV sequence covariation is considered, and the latter arise only when covarying amino acids are not considered (and thus may include compensatory mutations). Analyses in this paper employ all observed direct and indirect associations.

**Verifying the impact of HLA imputation on the ability to identify HLA-associated polymorphisms.** As described above, some level of HLA imputation was required for 711 of 1,888 (38%) individuals in the present study. To investigate the impact of HLA imputation on identification of HLA-associated polymorphisms in a "worst-case" scenario, we took the subset of patients for whom full high-resolution data were available (1,177 patients), truncated all HLA types to two-digit resolution, and reimputed them to high resolution. The highest-probability imputed allele matched the original call in 93% of cases. We then identified HLA-associated polymorphisms from the imputed data set and compared to those obtained from analysis of the same data set at high resolution (data not shown). When comparing unique HLA (type)-HIV codon pairs, 82% of associations derived from the imputed data set matched those derived from the high-resolution data set, a result that is expected given the false-discovery rate of 0.2 and the expected loss of statistical power when imputed data are analyzed.

**Testing for differential escape.** The statistical model for differential escape among HLA subtypes belonging to a given allele group (type) was defined as previously described (26). However, whereas the previous study (26) explicitly investigated differential escape within published epitopes, here we tested for differential escape in general across the entire proteome, that is, cases where one subtype selects a given polymorphism and another does not select for that polymorphism or selects for that polymorphism at a lower frequency. As such, differential escape in the present context could be due to less effective (or lack of) responses to specific epitopes. Briefly, the phylogenetically corrected logistic regression model was used to test for evidence that the odds of selection differed between an HLA type and a particular subtype. For example, to test if the probability $P$ of escape to 242N (in $p24^{Gag}$) differed between B*58:01 and the rest of the B*58 subtypes, we used the phylogenetically corrected logistic regression model defined by the following: $\ln[P/(1 - P)] = a(\text{B58}) + b(\text{B5801}) + cT$, where $T$ is a $-1/1$ binary variable representing the transmitted polymorphism, B58 and B5801 are 0/1 binary variables indicating whether the individual has the HLA allele in question, and the parameters ($a$, $b$, and $c$) are chosen to maximize the likelihood of the data. The likelihood ratio was then used to test the null hypothesis that $b = 0$ (i.e., that knowing that an individual expresses B*58:01 confers no additional information beyond knowing that the individual expresses B*58).

To validate the distinction between associations identified at type- versus subtype-level resolution, we applied the above model as follows. For each association originally identified at the subtype level, we tested for differential escape between the identified subtype and the corresponding HLA type. If the resulting $q$ value was less than 0.2, we considered the test an instance of true differential escape (differential escape was confirmed in >80% of cases where escape was originally defined at the subtype level). For each association originally defined at the type level, we tested for differential escape against all subtypes observed in at least 10 individuals in the subset for whom high-resolution HLA data were available across all loci (1,177 of 1,888 subjects, or 62%). Together, a total of 29 HLA types with at least two subtype members were available for analysis. If the $q$ value of the most significant such test was less than 0.2, we considered the HLA-polymorphism pair to represent a true instance of differential escape (a reclassification that occurred in <20% of cases where escape was originally defined at the type level). For both positional analyses and analyses involving the stratification of tests by HLA type, we limited the analysis to the most significant HLA-type/HIV-position pair to avoid double counting nonadapted/adapted pairs.

**Calculation of median genetic distances among HLA alleles of a given type.** To estimate relative protein distances among HLA subtypes included in our differential escape analysis (see above), we retrieved their exon 2 and 3 amino acid sequences from http://hla.alleles.org/data/txt/classi_prot.txt (98) and used the protdist program from the PHYLIP software package (41) to calculate pairwise distances using the Henikoff Tillier probability matrix from blocks (PBM) model of amino acid replacement (119). For each HLA allele group (type), the median pairwise genetic distance between subtype members (intratype distance) was calculated.

**Definition of optimally defined $CD8^+$ epitopes.** The Los Alamos optimal $CD8^+$ epitope list comprises all published epitopes defined according to rigorous *in vitro* epitope fine-mapping and HLA restriction experiments (http://www.hiv.lanl.gov/content/immunology/tables/optimal_ctl_summary.html; 31 August 2009 update) (76). However, optimal epitopes have not necessarily been tested in context of all possible restricting HLA alleles, nor are the alleles defined at the same resolution (i.e., HLA type, subtype, or serotype) throughout. To address these biases (and in recognition that alleles with shared properties are likely to present similar peptides (111),

Downloaded from http://jvi.asm.org/ on November 13, 2018 by guest

we expanded each optimal epitope to include all members of the HLA type, supertype, or serotype to which the published restricting allele belonged, as described in reference 26. Briefly, optimally defined epitopes were retrieved from the Los Alamos Database and hand edited to include recently published epitopes (22, 48, 55, 66, 69, 74, 79, 105; also A. Bansal and P. Goepfert, personal communication). For optimal epitopes restricted by HLA alleles defined at type- or subtype-level resolution, we expanded to include all subtypes belonging to the supertype to which the restricting allele belonged (111). For example, an epitope originally defined as B*57:01 restricted was assigned to all members of the B58 supertype. For epitopes restricted by alleles defined at the serotype level ($n = 38$), we expanded the list to include all HLA alleles belonging to that serotype (54), identified supertypes common to a majority of the resulting expansion, and expanded the list to include all subtypes matching those supertypes. If no supertype was defined (as is the case for HLA-C alleles), we expanded the list to include all HLA subtypes belonging to the HLA type of the original restricting allele. The full list of expanded epitopes (along with their original restrictions) is provided in Table S1 in the supplemental material.

**Analysis of population-level immune escape pathways to identify correlates of protective immunity: data definitions.** Protective effects of specific HLA class I alleles on HIV progression were defined according to published hazard ratios for progression to AIDS (HR-AIDS), determined for 54 individual class I alleles at type-level (2-digit) resolution in a natural history cohort of 600 Caucasian seroconverters (89). To allow analysis at subtype-level resolution using HLA-associated polymorphisms specific to the population in which HR-AIDS were originally derived, we assigned each HR to the most frequently observed HLA subtype in Caucasians (e.g., A*02 was assigned to A*02:01) and analyzed only HLA-associated polymorphisms restricted by these subtypes. In both univariate and multivariate analyses, HR-AIDS values were log transformed to render their distribution more normal.

HLA-associated polymorphism-related variables that we investigated could be classified into two broad categories: features related to the frequency and/or distribution of sites under selection by a given HLA allele (defined as unique HIV codons harboring a polymorphism associated with that HLA allele) and features related to the signal strength of sites under HLA selection and/or to evolutionary constraints on the sites themselves. The first category contained the following variables: (i) number of sites under HLA selection (computed as the total number of HLA-associated polymorphic sites in HIV-1, both overall and within/near published epitopes only); (ii) proportion (percent) of sites under HLA selection, by protein (computed as the proportion of the total HLA-associated sites occurring within the HIV-1 protein of interest); (iii) number of sites under HLA selection occurring at anchor residues (computed as the total number of HLA-associated sites occurring at anchor positions within known epitopes); (iv) proportion (percent) of sites under HLA selection occurring at anchor residues (proportion of total HLA-associated sites inside or within $\pm$ 3 amino acids [aa] of an epitope that occurred at an anchor position); (v) number of active epitopes (defined as the number of published epitopes harboring HLA-associated polymorphisms); and (vi) median number of sites under HLA selection per active epitope. The second category contained the following variables: (vii) median odds ratio (OR) of selection (computed by taking the maximum absolute log odds ratio of all nonadapted and adapted associations per HLA at each unique site, followed by computing the median of all sites per HLA, where the odds ratio for a given HLA-polymorphism pair was computed as the ratio of the odds of observing the polymorphism among individuals expressing the allele to the odds of observing the polymorphism among individuals not expressing the allele); (viii) median conservation of sites under HLA selection (defined as the median sequence conservation of all sites associated with a particular HLA, where conservation was defined to be the proportion of individuals with the consensus sequence among individuals who did not express any HLA alleles associated with that site) (102); and (ix) median number of covarying codons per site (computed by summing the number of unique HIV-1 codons identified as covarying with each site and computing the median for all sites per HLA). Where appropriate, analyses were undertaken at both the proteome-wide and individual-protein levels (defined as Gag, Pol, Env, Nef, and accessory) to identify location-specific correlates of protection.

Multivariate analyses were attempted using a small number of distinct variables significant at the univariate level. To maximize the number of HLA alleles included in the multivariate analysis, alleles with "missing" variables (e.g., those excluded from the original univariate analysis of OR of escape in Gag because they possessed no escape sites in that protein) were assigned the mean observed value in the data set. Selected variables were regressed against log-transformed HR-AIDS values by automated forward selection using a stepwise Akaike information criterion (AIC) procedure (2).

## RESULTS

**Identification of HLA-associated polymorphisms at three levels of HLA resolution.** HLA-associated polymorphisms were identified across the entire HIV-1 proteome in an international cohort of 1,888 treatment-naïve, chronically subtype B-infected individuals using phylogenetically informed methods that incorporated corrections for HIV codon covariation and HLA linkage disequilibrium (20, 25, 26) and used a $q$-value correction for multiple tests (114). Missing or low- or incomplete-resolution HLA data were imputed to subtype-level resolution using a published machine learning algorithm (75), and associations were computed by averaging over all possible HLA resolutions (see Materials and Methods); this represents the first use of HLA completion algorithms in population-level immune escape analyses.

Studies of HLA-associated polymorphisms have generally featured individual pairwise analyses at the HLA type (i.e., 2-digit) and/or subtype (i.e., 4-digit) levels (19, 20, 25, 104). Supertype-level analyses have been uncommon even though statistical power could be enhanced to identify universal escape pathways within epitopes capable of binding related alleles. We therefore report HLA-associated polymorphisms at three different levels of resolution: first, at the combined supertype/type/subtype levels; second; at the combined type/subtype levels; and third; at the subtype level only. In the first two analyses, the level of resolution yielding the lowest $P$ value for each HLA-associated polymorphism is reported (that is, if the statistical signal for the association is stronger at the supertype than at the type or subtype level, then it is reported at the supertype level). A full list of HLA-associated polymorphisms identified at all three levels of resolution is provided in Table S2 in the supplemental material, while a full list of intraprotein codon covariation pathways is provided in Table S3. In addition, protein-specific immune escape maps for all associations with a $q$ of $\leq$0.05 identified at the broadest (supertype/type/subtype) resolution level analysis are provided in Fig. S1 in the supplemental material.

At a $q$ of $<$0.2, over 2,100 unique HLA-associated escape pathways (defined as those unique over HLA restriction, HIV codon coordinate, amino acid, and direction of escape) occurring at over 750 HIV-1 codons were identified (see Table S2). When the inherent sequence conservation of each HIV protein was taken into consideration, the proportion of sequence variation attributable, at least in part, to HLA selection pressures (defined as the proportion of sites harboring at least one HLA-associated polymorphism) differed markedly by viral protein (Fig. 1). For example, with 59% of its residues exhibiting $\geq$99.5% amino acid conservation, $p24^{Gag}$ is the most highly conserved HIV-1 protein, yet HLA pressures influence sequence variation at 40% of its variable

Downloaded from http://jvi.asm.org/ on November 13, 2018 by guest

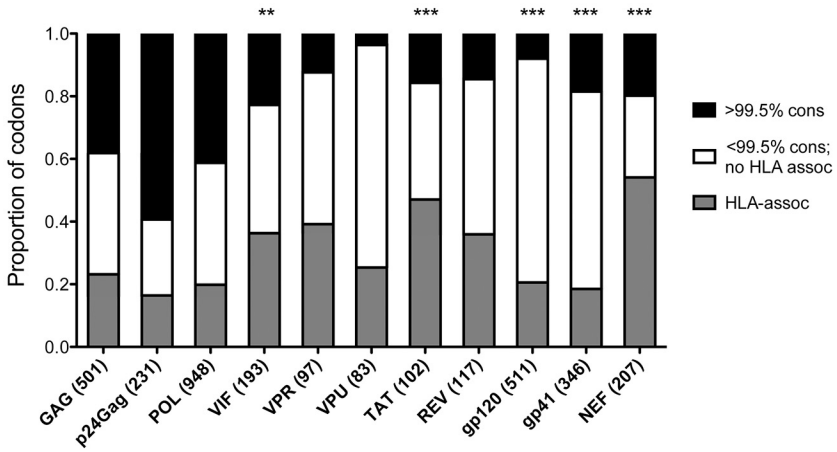


**FIG 1** The proportion of sequence variation attributable to HLA-associated selection pressures differs markedly by HIV-1 protein. The proportion of codons exhibiting ≥99.5% amino acid conservation at the population level in our cohort and the proportion of variable codons (those exhibiting <99.5% amino acid conservation) harboring no known HLA-associated polymorphisms are indicated, along with the proportion of codons harboring at least one HLA-associated polymorphism. Numbers in parentheses in the *x*-axis labels indicate the total length in amino acids of each protein (including the stop codon, if present). **, $P < 0.001$; ***, $P < 0.0001$ (Fisher's exact test compared to all other proteins combined).

codons. HLA-mediated polymorphisms were observed most frequently in the highly immunogenic Nef protein, where 67.4% of variable codons harbor HLA associations ($P = 1.9 \times 10^{-18}$; Fisher's exact test compared to other proteins). Conversely, the highly variable Vpu protein exhibited the least evidence of HLA-driven evolution (26.3% [$P = 0.009$]) (Fig. 1) (19, 52, 60). Across the entire HIV-1 proteome, polymorphisms at 804 of 1,481 (35.2%) variable sites were associated with at least one HLA allele.

**The extent of differential CTL escape between subtype members varies markedly by HLA.** Although the HLA-associated polymorphisms identified at the three different levels of HLA resolution are highly concordant (see Table S2 in the supplemental material), each analysis features relative merits and limitations. Analysis at broader resolution can enhance statistical signal when related alleles bind the same epitope and escape along the same pathway. For example, the Gag-TW10 epitope binds many members of the B58 supertype. At Gag codon 242 (position 3 of this epitope), relative odds of escape are 193:1 for B58 supertype members compared to non-B58 supertype members ($P = 1.6 \times 10^{-117}$). However, when relative odds are calculated at the type (e.g., B*57) or subtype (e.g., B*57:01) level, lower odds are observed (128:1, $P = 1.2 \times 10^{-87}$, and 114:1, $P = 5.4 \times 10^{-72}$, respectively) because persons harboring closely related alleles (e.g., B*58 and/or B*58:01) are classified as lacking the specific allele under investigation even though they also select the same escape mutation. Conversely, when alleles belonging to the same type or supertype bind different peptides or select different escape pathways, analyzing at broader resolution may yield inappropriate results while still achieving high levels of statistical significance and may mask differential escape (where similar HLA alleles bind the same epitope but escape along distinct pathways) (26). For example, despite belonging to the B58 supertype, B*58:02 does not select for the T242N escape mutation (see Table S2).

Notably, even when analyzed at the broadest resolution level, the majority of HLA-associated polymorphisms (62%) are still subtype restricted, while only 31% and 7% are type and supertype restricted, respectively. These distributions were robust to HLA imputation (see Materials and Methods and Fig. S2 in the supplemental material). To characterize the extent of differential escape between and within HLA types/subtypes across the HIV-1 proteome, we applied recently developed techniques (26; see also Materials and Methods) to explicitly test for differential escape (defined here as cases where one subtype selects a polymorphism and another selects a different—or no—polymorphism at any given codon). Overall, we estimate that when polymorphisms are identified at the type level, there is an 81.5% chance that no differential escape occurs at the subtype level. Conversely, when polymorphisms are identified at the subtype level, there is an 81.2% chance that the subtype of interest indeed displays a significant differential escape pattern compared to the type as a whole (data not shown). Given our 20% false-discovery rate significance threshold, results indicate that our method of reporting associations at the level of resolution yielding the lowest *P* value is broadly appropriate. Stratifying analyses by individual HLA alleles revealed that the heterogeneous B*15 type exhibited the most evidence for differential escape (with over 85% of pathways being unique at the subtype level) while B*07 exhibited the least (with only 20% of pathways being unique at the subtype level) (Fig. 2A). Overall, differential CTL escape was widespread at the subtype level, with allele group members exhibiting distinct escape mutations in 58% of the cases on average (HLA-A, 63%; HLA-B, 55%; HLA-C, 60%).

The underlying mechanisms that drive differential CTL escape are unclear and may include differences in epitope binding, T-cell receptor (TCR)-HLA-epitope interactions, TCR repertoire, or may reflect differences that result from linkage of certain HLA subtypes to other factors, such as antigenic exposure, antigenic processing, or linkage with other HLA class I or class II alleles (26). In support of a mechanism that involves properties of the HLA alleles themselves, we observed a striking positive correlation between the overall genetic divergence of HLA exons 2 and 3 (the major genetic determinants of the peptide-HLA binding groove) between allele group members (see Fig. S3 in the supplemental material) and the frequency of differential escape between them (Spearman $R = 0.60$; $P = 0.0043$) (Fig. 2B).

Downloaded from http://jvi.asm.org/ on November 13, 2018 by guest

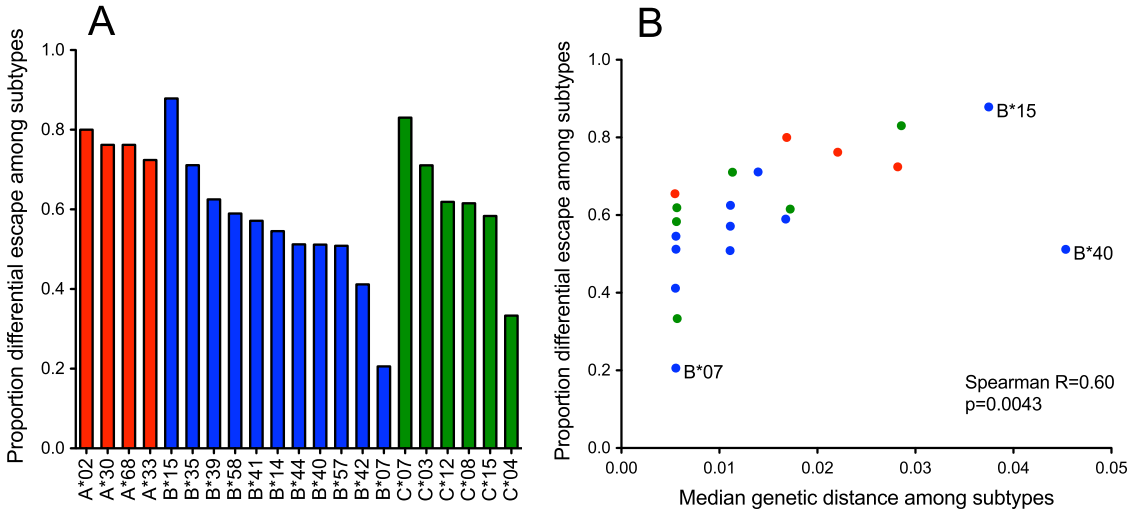


**FIG 2** Widespread differential escape between HLA allele group members correlates with evolutionary distance between HLA genes. Individual HLA-A, -B, and -C types are indicated. The proportion of differential escape (defined as the proportion of HLA-associated polymorphisms for which the odds of escape differed significantly between at least one HLA subtype and the rest of the group) differs markedly for various common types of the HLA-A, -B, and -C loci (A). HLA types exhibiting the highest level of differential escape were those exhibiting the greatest intratype genetic diversity (B). HLA-B*07 and B*15 (exhibiting the highest and lowest proportion of differential escape, respectively) and B*40 are labeled for special interest.

**Enrichment of escape at epitope-HLA anchor residues.** We next examined the distribution and positioning of HLA-associated polymorphisms with respect to CTL epitopes (76). Our initial epitope reference list consisted of all optimally described CTL epitopes restricted by one or more HLA class I alleles. We then expanded this list by assigning the published epitope to all allele group members (i.e., subtypes) belonging to the HLA allele group (i.e., supertype, type, or serotype, depending on the published restriction) to which the original published epitope belonged (see Materials and Methods and Table S1 in the supplemental material). For example, an epitope originally defined as B*57:01 restricted was assigned to all members of the B58 supertype. Since all epitopes on our expanded list were defined at subtype-level resolution, we utilized the subtype-level HLA-associated polymorphism list in the following analysis although sensitivity analyses confirmed that results were consistent regardless of the association list used (data not shown).

Of the 2,161 HLA-restricted viral escape pathways identified at subtype-level resolution, 29% occurred inside or within ±3 residues of an optimally described CTL epitope. This value was 21.5% when epitope expansion was not undertaken. In support of including flanking regions in the epitope definition, a bootstrap analysis (100,000 replicates) indicated that escape frequency at epitope flanking residues was significantly higher than escape outside epitopes and their immediate flanking regions for HLA-B alleles ($P = 0.00012$) though this was not significant for HLA-A or -C alleles. Because the majority of defined epitopes are 9-mers, we first examined these. Of the 262 unique 9-mer HLA-epitope pairs (defined as epitopes restricted by the same HLA supertype and mapping to the same HIV coordinates) for which an HLA-associated polymorphism was identified, positions 2 and 3 and the C terminus represented the most frequently escaping sites, while escape at N-terminal flanking residue −3 and C-terminal flanking residue +2 was less frequent than expected under the null hypothesis of equal escape frequency across all sites within/flanking epitopes (binomial test for departure from expected 1/15 frequency, all $P < 0.01$) (Fig. 3A). Stratification by locus revealed that HLA-A-restricted epitopes tended to select for escape mutations at position 3 and the C terminus, HLA-B-restricted epitopes typically developed escape mutations at position 2, and the limited number of HLA-C epitopes most often encoded C-terminal escape mutations (see Fig. S4 in the supplemental material).

Classification of escape mutations with respect to their position at an anchor residue (defined according to HLA subtype-specific anchor motifs available at http://www.hiv.lanl.gov/content/immunology/motif_scan/motif.html) (37, 78, 91, 110) revealed that escape occurred at anchor residues 1.8-fold more frequently than expected under the null hypothesis of equal escape frequency across all sites within/flanking epitopes (binomial test for departure from expected 2/15 frequency, $P = 1.0 \times 10^{-5}$) (Fig. 3B) and 1.4-fold more frequently than expected across all sites within the epitope only (binomial test for departure from expected 2/9 frequency, $P = 0.003$) (data not shown). The frequency of escape at N- and C-terminal epitope flanking residues was nearly 2-fold lower than that expected under the null hypothesis of equal escape frequency across all sites within/flanking epitopes (binomial test for departure from expected 6/15 frequency, $P < 1.0 \times 10^{-5}$) (Fig. 3B). Moreover, the average statistical strength of escape at anchor sites (measured as the maximum absolute natural-log-transformed odds ratio [lnOR] for each unique HLA-HIV codon pair) was modestly yet significantly increased compared to that of nonanchor or flanking sites (median lnOR of 2.15 [interquartile range (IQR), 1.15 to 2.58] for anchor sites versus 1.62 [IQR, 0.93 to 2.35] for nonanchor sites versus 1.30 [IQR, 0.81 to 1.98] for flanking regions; Kruskal-Wallis, $P = 0.035$) (Fig. 3C).

Epitopes of other lengths (8-, 10-, and 11-mer [8/10/11-mer]; $n = 201$) were also examined. To ensure alignment of anchor residues, positions were analyzed with respect to their distance from the N and C termini. Relatively uniform frequency of escape was observed at epitope positions 1 to 4 and at the penultimate and C-terminal positions, while escape at position C−3 (three positions upstream of the C terminus) was relatively rare (Fig.

Downloaded from http://jvi.asm.org/ on November 13, 2018 by guest

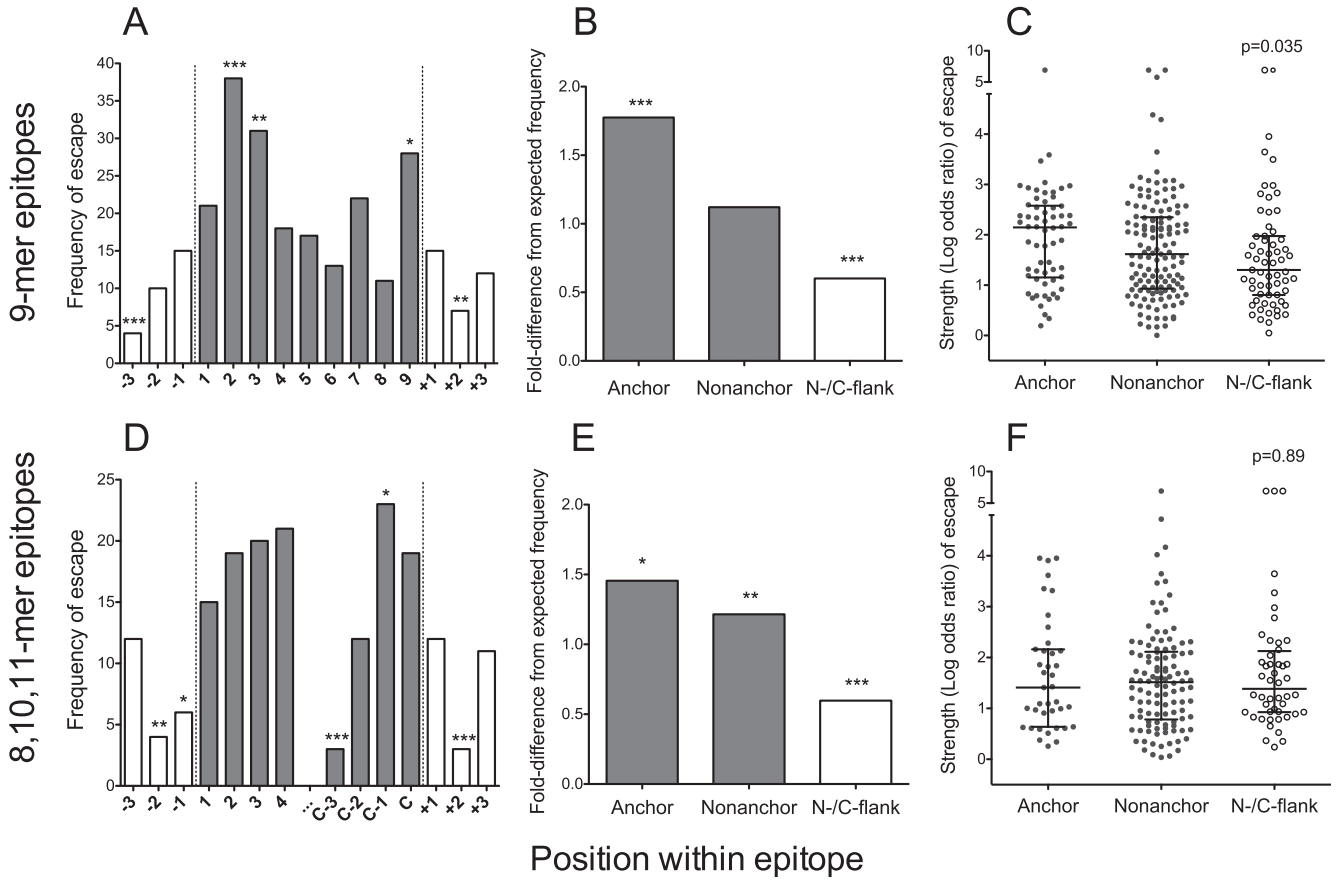


**FIG 3** The distribution and natural log odds ratio of escape within known epitopes is nonuniform and biased toward anchor residues. The distribution and natural log odds ratio of escape mutations within known CD8$^+$ 9-mer epitopes (A, B, and C) and 8-/10-/11-mer epitopes (D, E, and F) are shown. Frequencies of escape by relative position within or occurring at the ±3 amino acids flanking the epitope N and C termini are depicted as histograms in panels A and D. In panel D, notation of the type C−3 on the *x* axis indicates the number of positions upstream of the C terminus (in this case, 3). These same data are depicted as fold differences from expected frequency of escape occurring at HLA-specific anchor, nonanchor, and epitope flanking positions (based on the null hypothesis of equal escape probability across all positions) and are shown in panels B and E. *, $P < 0.05$; **, $P < 0.001$; ***, $P < 0.0001$ (binomial test). The natural log of the odds ratio of escape at HLA-specific anchor, nonanchor, and epitope flanking positions are shown as scatter plots (C and F).

3D). When analyzed separately, 10-mers tended to escape more frequently at position 3 but not the C terminus, while 11-mers tended to escape more frequently at position 4 (see Fig. S4 in the supplemental material), an observation that may be due to the bulging of oversized epitopes and its effect on TCR contacts (83). Overall, a 1.45-fold enrichment of escape mutations at anchor residues was observed compared to other sites within and flanking the epitopes ($P = 0.014$) (Fig. 3E); however, this did not remain significant when analysis was restricted to within-epitope sites only ($P = 0.3$) (data not shown). The average statistical strength of escape (lnOR) did not differ significantly between anchor and nonanchor sites for 8-, 10-, and 11-mer epitopes (median lnOR was between ~1.4 and 1.5 for all positions; $P = 0.89$) (Fig. 3F).

**Estimating the effect of anchor residue escape on HLA binding.** Given the enrichment of CTL escape mutations at HLA anchor residues, we next estimated the extent to which these mutations were predicted to affect peptide-HLA binding using NetMHCpan2.4, an artificial neural network-based peptide-HLA binding prediction tool trained on >37,000 quantitative binding data for more than 42 different HLA molecules (88). For each HLA-associated adapted polymorphism that occurred inside an epitope (147 9-mers and 133 8/10/11-mers), we compared the predicted HLA binding affinity for the published epitope in its HLA-restricted nonadapted (i.e., susceptible) versus its HLA-associated adapted (i.e., escaped) form, using the subtype B consensus for the epitope backbone (in cases where no HLA-restricted nonadapted residue was identified, the subtype B consensus residue was used). After removing epitopes predicted to bind the restricting HLA with a 50% inhibitory concentration ($IC_{50}$) of >1,000 nM in their nonadapted form (representing a 2-fold lower affinity than NetMHCpan2.4-defined weak binding threshold, which included 30 9-mers and 38 8/10/11-mers), we calculated that anchor residue escape in 9-mer epitopes was predicted to reduce peptide-HLA binding by a median of 8.9-fold (IQR, 3.7- to 52.1-fold) compared to 1.25-fold (IQR, 0.80- to 1.99-fold) at nonanchor sites (Fig. 4A). Among 8/10/11-mer epitopes, anchor residue escape was predicted to reduce peptide binding a median of 8.9-fold (IQR, 1.9- to 114.3-fold) compared to 1.0-fold (IQR, 0.87- to 1.56-fold) at nonanchor sites (Fig. 4B). Overall, 36 out of 212 (17%) within-epitope escape mutations were predicted to "completely" abrogate peptide-HLA binding, defined as epitopes which bound with an $IC_{50}$ of <1,000 nM affinity in their nonadapted form but an $IC_{50}$ of >1,000 nM affinity in their escaped form (data not shown).

Downloaded from http://jvi.asm.org/ on November 13, 2018 by guest

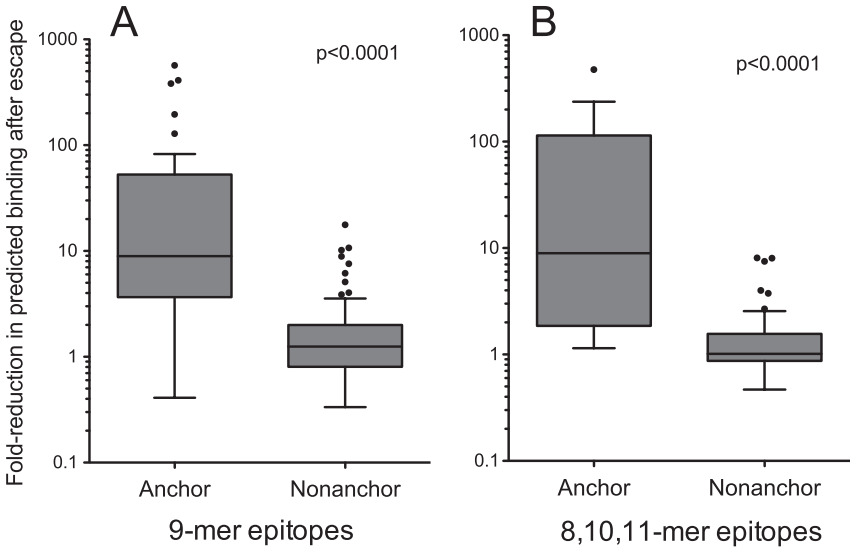


**FIG 4** Fold changes in predicted peptide-HLA binding affinities following escape at anchor residues. Fold changes in predicted peptide-HLA binding affinities as a consequence of escape at HLA-specific anchor (versus nonanchor) residues within 9-mer (A) and 8-, 10-, and 11-mer (B) epitopes are shown as Tukey box plots.

**Correlates of protective immunity revealed by analysis of population-level immune escape pathways.** Expression of specific HLA class I alleles is associated with differential rates of HIV-1 disease progression (27, 62, 89), but the precise mechanisms underlying these effects remain incompletely known (6, 56, 70, 93). Given that HLA-associated polymorphisms identify viral sites under strong *in vivo* selection by an HLA allele at the population level, analysis of their frequency, distribution, statistical strength, sequence conservation, and other characteristics allows us to identify characteristics that discriminate protective HLA allele-associated immune pressures from nonprotective ones. It is important to emphasize that, although population-level HLA-associated polymorphisms may allow us to infer specific properties, tendencies, or attributes that render certain HLA alleles particularly effective at controlling HIV, by no means are we proposing that selection of escape mutations is beneficial at the individual patient level.

HLA-associated protective effects were defined according to published hazard ratios for progression to AIDS (HR-AIDS) derived from a large longitudinal natural history cohort (89; see also Materials and Methods). We investigated 33 interrelated features of viral sites under HLA-mediated selection as potential correlates of protective immunity. These 33 features were divided into two broad categories: those related to their frequency/distribution (overall, and within known epitopes) and those related to the statistical strength of selection and/or mutational constraints on the sites themselves (Table 1). Where appropriate, analyses were undertaken at proteome-wide and individual-protein levels. HLA-associated polymorphisms from the subtype-level analysis were used.

Of the 33 variables investigated, we identified 12 potential correlates of HLA-mediated protective immunity against disease progression ($P < 0.05$, $q < 0.2$) (Table 1). Among the strongest correlates of protection was the significant inverse relationship between total number of sites under selection by a given HLA and its HR-AIDS (Spearman's $R = -0.41$; $P = 0.0024$) (Fig. 5A), indicating that protective HLA alleles exert substantial *in vivo* pressure on a greater number of sites across the HIV proteome. This observation was almost exclusively driven by HLA-A and -B alleles (Spearman's $R = -0.48$; $P = 0.0016$) (data not shown). When analyzed at the individual protein level, this inverse relationship was strongest for Gag and remained significant for all proteins except Nef, where the trend persisted but was not significant (Table 1).

Notably, no significant correlations were observed between HR-AIDS and median sequence conservation of sites under selection (where the sequence conservation of a given site was defined as the proportion of the total cohort harboring the consensus residue at this position [102]). This remained true when the bias induced by HLA-mediated selection on these sites was addressed by excluding persons expressing these alleles from the calculation of conservation (Table 1). Similarly, no significant correlations were observed between HR-AIDS and the median number of covarying codons per site under selection. However, significant inverse relationships were observed between an HLA allele's HR-AIDS and the median odds ratio of selection for mutations in Gag ($R = -0.6$, $P = 0.0012$) (Fig. 5B) and, to a lesser extent, in Pol ($R = -0.34$, $P = 0.039$) but not overall or in other HIV-1 proteins.

Another highly significant proteome-wide correlate of protection was the tendency of protective alleles to select escape mutations at HLA anchor sites within known epitopes (Spearman's $R = -0.47$; $P = 0.0052$) (Fig. 5C). This effect appeared to be largely constrained to Pol and Gag ($P = 0.0015$ and $P = 0.065$, respectively). The median number of sites under selection per epitope in Gag, but not elsewhere in the viral genome, was also modestly inversely correlated with HR-AIDS (Table 1).

To attempt to identify which of these features represented the strongest independent correlates of protection, we investigated a variety of multivariate models incorporating four of our top "hits" (median OR of selection in Gag, median OR of selection in Pol, number sites under selection proteome-wide, and percentage of selected sites occurring at anchor residues). Unfortunately due to

Downloaded from http://jvi.asm.org/ on November 13, 2018 by guest

**TABLE 1** Identifying correlates of HLA allele-associated protective immunity

| Feature | Category | No. of HLA alleles analyzed | Spearman *R* | *P* value | *q* value[a] |
|---|---|---|---|---|---|
| Median OR of selection in Gag | Strength/constraint | 26 | −0.60 | 0.0012 | **0.04** |
| No. of sites under selection (proteome-wide) | Frequency/distribution | 53 | −0.41 | 0.0024 | **0.04** |
| Proportion of selected sites occurring at HLA anchor residues (proteome-wide [%]) | Frequency/distribution | 34 | −0.47 | 0.0052 | **0.04** |
| No. of sites under selection in Gag | Frequency/distribution | 53 | −0.38 | 0.0054 | **0.04** |
| No. of sites under selection in accessory proteins | Frequency/distribution | 53 | −0.37 | 0.0058 | **0.04** |
| No. of selected sites occurring at HLA anchor residues (proteome-wide) | Frequency/distribution | 44 | −0.35 | 0.018 | **0.10** |
| No. of sites under selection within/near epitopes (proteome-wide) | Frequency/distribution | 53 | −0.31 | 0.025 | **0.11** |
| No. of active epitopes in Gag | Frequency/distribution | 53 | −0.30 | 0.027 | **0.11** |
| Median no. of sites under selection per epitope in Gag | Frequency/distribution | 22 | −0.46 | 0.032 | **0.12** |
| Median OR of selection in Pol | Strength/constraint | 38 | −0.34 | 0.039 | **0.13** |
| No. of sites under selection in Env | Frequency/distribution | 53 | −0.28 | 0.045 | **0.13** |
| No. of sites under selection in Pol | Frequency/distribution | 53 | −0.28 | 0.046 | **0.13** |
| Median conservation of selected sites in Nef | Strength/constraint | 38 | −0.29 | 0.079 | 0.20 |
| Median conservation of selected sites in accessory proteins | Strength/constraint | 45 | 0.24 | 0.11 | 0.25 |
| Median conservation of selected sites in Env | Strength/constraint | 32 | 0.28 | 0.13 | 0.28 |
| No. of ites under selection within/near epitopes in Gag | Frequency/distribution | 28 | −0.27 | 0.16 | 0.33 |
| No. of active epitopes (proteome-wide) | Frequency/distribution | 53 | −0.17 | 0.23 | 0.41 |
| Median no. of covarying codons per selected site (proteome-wide) | Strength/constraint | 51 | −0.17 | 0.23 | 0.41 |
| Median no. of sites under selection per epitope (proteome-wide) | Frequency/distribution | 44 | −0.18 | 0.24 | 0.41 |
| Median conservation of selected sites in Nef | Strength/constraint | 51 | 0.16 | 0.25 | 0.42 |
| Proportion of sites under selection in Gag (%) | Frequency/distribution | 51 | −0.14 | 0.33 | 0.50 |
| Proportion of sites under selection in Env (%) | Frequency/distribution | 51 | −0.13 | 0.35 | 0.50 |
| Proportion of sites under selection in Nef (%) | Frequency/distribution | 51 | 0.13 | 0.36 | 0.50 |
| Proportion of sites under selection in accessory proteins (%) | Frequency/distribution | 51 | −0.13 | 0.37 | 0.50 |
| No. of sites under selection in Nef | Frequency/distribution | 53 | −0.12 | 0.39 | 0.50 |
| Median OR of escape in Nef | Strength/constraint | 38 | 0.14 | 0.40 | 0.50 |
| Median no. of covarying codons per site under selection (Gag only) | Strength/constraint | 26 | −0.084 | 0.68 | 0.81 |
| Median OR of selection in accessory proteins | Strength/constraint | 45 | 0.059 | 0.70 | 0.81 |
| Median OR of selection in Env | Strength/constraint | 32 | −0.067 | 0.72 | 0.81 |
| Proportion of sites under selection in Pol (%) | Frequency/distribution | 51 | −0.0069 | 0.96 | 0.99 |
| Median conservation of selected sites in Gag | Strength/constraint | 26 | −0.0068 | 0.97 | 0.99 |
| Median OR of selection (proteome-wide) | Strength/constraint | 50 | −0.0024 | 0.99 | 0.99 |
| Median conservation of selected sites in Pol | Strength/constraint | 39 | 0.0016 | 0.99 | 0.99 |

[a] Shown in boldface are *q* values of <0.2, indicating potential correlates of HLA-mediated protective immunity against disease progression.

the strong interdependency of these variables and the fact that not all variables were available for all HLA alleles, multivariate analyses were not robust to changes in model selection procedures. The only variable that consistently emerged as a significant independent predictor of protection was the number of sites under selection by a given HLA across the HIV-1 proteome (data not shown).

## DISCUSSION

HLA-associated polymorphisms identify HIV-1 sites that are under consistently strong immune pressure *in vivo* and whose sequence diversity is largely driven by HLA. As such, the well-powered nature of this analysis combined with its proteome-wide focus allowed us to study the frequency, distribution, and characteristics of viral sites under HLA-mediated selection. The proportion of population-level sequence variation attributable, at least in part, to HLA (defined as the proportion of sites in a given protein harboring at least one HLA-associated polymorphism) differs markedly by protein and is independent of the protein's overall conservation. For example, whereas the majority of variable Nef codons are attributable, at least in part, to HLA class I-mediated pressures, this explains only a quarter of Vpu sequence variation, implying the existence of other host factors in driving evolution of this viral gene (40, 52, 112) (to provide context, nearly 40% of variable sites in the highly conserved $p24^{Gag}$ protein are associated with HLA).

Building on recent advances (26, 69), we demonstrated significant variation in the proportion of differential escape among closely related HLA subtypes, implicating differences in TCR repertoire and/or the interaction between TCR and the epitope-HLA complex as the underlying mechanisms in at least some cases (123). Widespread differential escape is consistent with reported differences in HIV-1 disease progression between HLA alleles differing by as little as 1 amino acid (43, 68, 86) and also supports the notion that analysis of HLA-associated polymorphisms provides a novel perspective from which to discriminate both quantitative and qualitative differences in HLA-restricted immune responses. Of interest, HLA allele subtype pairs that exhibit the highest level of differential escape were those with the greatest intratype sequence divergence, raising the intriguing hypothesis that differential escape may be a consequence of active positive selection at these loci.

Analysis of the distribution of HIV-1 escape mutations within known CTL epitopes revealed that abrogation of peptide-HLA

Downloaded from http://jvi.asm.org/ on November 13, 2018 by guest

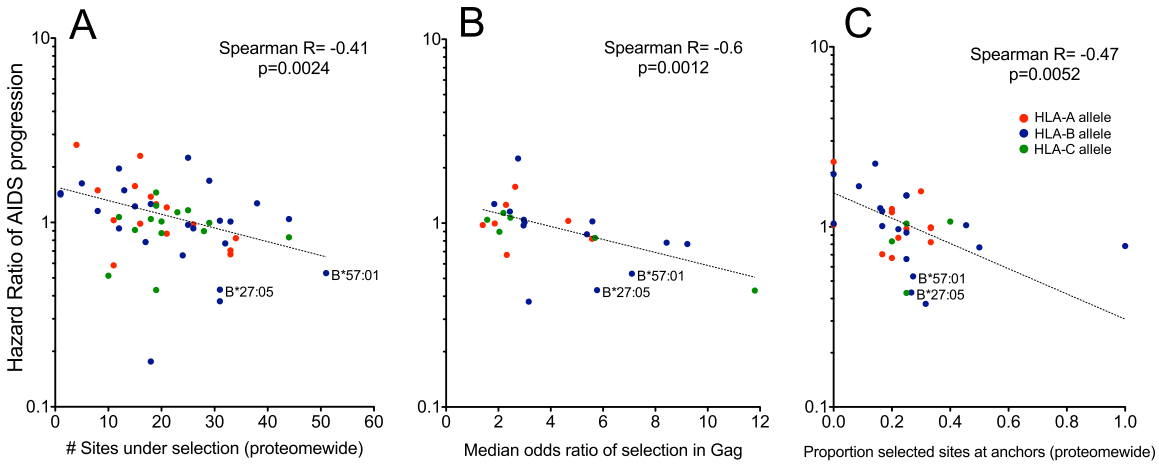


**FIG 5** Population-level immune escape pathways in HIV-1 reveal correlates of protective immunity: selected results. Individual HLA-A, -B, and -C alleles are as indicated. Significant inverse correlations were observed between the total number of sites under immune selection by a given HLA across the entire HIV-1 proteome (A), the median odds ratio of escape in Gag (B), and the percent escape at HLA anchor residues (C) and published hazard ratios of AIDS progression. HLA-B*57:01 and B*27:05 are labeled for special interest.

binding via anchor residue mutation represented a major escape mechanism and that such mutations incurred, on average, a nearly 10-fold reduction in predicted HLA binding affinity. Among 9-mer epitopes, the statistical strength of escape mutations at anchor residues was higher than for nonanchor residues. Mutations occurring at nonanchor positions within the epitope were less frequently detected, and those at residues immediately flanking the N and C termini of the epitope were less frequent still (although still significantly higher than in extraepitopic zones, at least for HLA-B). Although it is tempting to conclude that abrogation of peptide-HLA binding represents HIV's favored escape mechanism, a potential limitation should be noted. It is possible that detection of HLA-associated polymorphisms at the population level is inherently biased toward detection of "straightforward" escape mutations such as those directly affecting binding to HLA, while escape mutations resulting from more complex and/or "secondary" interactions (e.g., escape from populations of TCRs capable of interacting with HLA-bound peptide) may be more challenging to detect. Indeed, numerous examples exist where an escape variant selected in one individual is highly recognized by another's immune response (38, 47, 63, 90).

Despite these caveats, recent large-scale studies of immune escape indicate that data sets of the present size have sufficient power to detect even rare escape mutations (26), as well as to detect cases where escape from one HLA allele represents the susceptible form for another (19). The high frequency of anchor residue escape has implications for proposed vaccine strategies aimed at immunizing with both wild-type and common variants to channel HIV evolution down unconventional pathways (7, 13, 19). While still valid, such strategies should consider excluding epitope variants that are experimentally verified to not bind their restricting HLA, unless such variants overlap key epitopes restricted by other alleles and/or lead to the creation of novel epitopes relevant to immune control (4).

Analysis of HLA-associated polymorphisms identified at the population level also provides a unique perspective from which to identify, albeit indirectly, specific characteristics that differentiate protective from nonprotective HLA-restricted CTL responses. Such sequence-based methods offer an alternative perspective to traditional *in vitro* and/or *ex vivo* assays such as gamma interferon (IFN-γ) enzyme-linked immunosorbent spot (ELISpot) assays (115), where, for a variety of reasons (including use of synthetic consensus peptides at supraphysiological concentrations that bypass antigen-processing machinery, detection of a single cytokine readout only, inability to detect previous effective responses due to escape in the autologous virus, and others [121]), responses detected *in vitro* may not accurately discriminate effective responses occurring *in vivo*. Indeed, a recent large-scale analysis of >2,000 chronically clade C-infected persons identified differential immune escape as a better predictor of average per-HLA plasma viral load than responses measured by IFN-γ ELISpot assay (26).

Importantly, analysis of population-level HLA-associated polymorphisms as markers of *in vivo* immune pressure by no means argues that CTL escape is protective at the individual level; indeed, escape mutations in individuals have been linked to higher viral load and disease progression (21, 26, 39, 49). At the population level, HLA-associated polymorphic sites identify HLA-restricted CTL responses that are consistently mounted in a substantial number of individuals expressing that allele—responses that are strong enough to drive the virus to escape in reproducible ways. As such, these sites represent the total potential of common CTL responses restricted by a given HLA allele to effectively target HIV; analysis of their distribution and characteristics can therefore help illuminate correlates of protection. Translation of these findings to the individual level is more complex. First, not all individuals expressing a given HLA will mount all possible CTL responses. Broadly generalizing, from a viral control perspective, the ideal would be to mount an effective immune response that the virus is incapable of escaping. Less ideal (but more the norm) would be to generate an effective response that the virus ultimately is able to escape (yielding losses of immune control that in some cases may be offset by viral fitness costs). Least ideal would be the inability to mount the response in the first place. At the individual level, therefore, escape mutations can be regarded as genetic relics of formerly effective CTL responses, where, in general, the nega-

Downloaded from http://jvi.asm.org/ on November 13, 2018 by guest

tive consequences of escape are still preferable to the inability to mount responses at all (26).

Our proteome-wide and protein-specific analyses shed new light on the long-standing debate around whether CTL response breadth is (31, 101) or is not (1) a correlate of protection. For example, the number of Env-specific CD8 T-cell responses in chronic infection, as measured by IFN-γ ELISpot assay, has been associated with poorer HIV control at the individual level (57, 67), suggesting that not all CTL responses are equally effective (29, 58, 108). Our results indicate that protective alleles (most notably those of the HLA-A and -B loci) have the potential to exert immune pressure on a larger number of sites across the viral proteome and that selection breadth potential is significantly associated with protection for all proteins except Nef (although a trend remained for the latter). Furthermore, the number of associated sites across the proteome per HLA allele was the only variable that consistently emerged as an independent correlate of protection in multivariate analyses. Our observations highlight the utility of using population-level HLA-associated polymorphisms to preferentially discriminate the impact of HLA-restricted immune responses that are most effective at controlling HIV-1 replication *in vivo* and suggest that the ability of an HLA allele to mount broad CTL selection pressures to a wide variety of proteins beyond Gag—including Env—represents a general correlate of protection.

The negative results of this study are also illuminating. The lack of correlation between HR-AIDS and the density of the amino acid covariation network associated with any given site under immune selection suggests that targeting structural or functional "network hubs" (i.e., those exhibiting higher-order evolutionary constraints within a given HIV protein [33]) is not independently protective. Similarly, the lack of correlation between HR-AIDS and average sequence conservation of sites under HLA selection indicates that the ability to target constrained sites is not in itself protective: it may depend on where these sites are in the HIV-1 proteome and how strongly they are targeted. Indeed, analysis at the individual protein level revealed that the statistical strength of selection on conserved sites (as measured by the odds ratio of association) in Gag and, to a lesser extent, Pol but not other viral proteins was a significant correlate of protection. Since a high odds ratio reflects both the frequency of selection among persons expressing the HLA allele and the likelihood of reversion in persons who lack the allele, it may be interpreted as an index of exceptionally strong selection at mutationally constrained sites. Our results thus highlight the strength of immune selection on conserved sites within structural and functional HIV-1 proteins as an important correlate of protection.

Gag was also confirmed as a particularly effective immune target. The number of sites under active immune selection in Gag, the number of active Gag epitopes (i.e., those harboring evidence of immune selection), and the average number of sites under selection per Gag epitope (a marker of diversity of selection pressures on a given epitope) also represented significant correlates of protection. Although an HLA allele's ability to consistently mount immune pressure on conserved Gag codons may appear an intuitive correlate of protection, the factors that constrain the virus to escape at those specific sites remain unclear. That the median odds ratio of escape in Gag correlated significantly with both the number of sites under selection and the proportion of anchor residue escape in the same protein and that in 9-mers the odds ratio of escape was significantly higher at anchor sites suggest that a hallmark of HLA-associated protection is the ability to consistently mount intense CTL responses from which the virus is compelled to escape (hence, the high odds of escape) but can only do so at a limited number of positions (hence, the position specificity) and at a biologically relevant fitness cost (hence, sequence conservation in persons lacking the allele). Furthermore, if the anchor residue is otherwise highly conserved, escape at this position could indicate that CTL targeting the epitope are highly polyclonal and/or cross-reactive, thus limiting escape options to those that abrogate epitope presentation. If so, this supports the clonal composition and/or diversity of an epitope-specific CTL repertoire as a correlate of protection (28, 58) and supports vaccine strategies seeking to stimulate CTL responses to both wild-type and commonly occurring escape variants at nonanchor sites (11, 42, 109), most notably in epitopes where escape at anchor residues is likely to incur a fitness cost. Indeed, a recent analysis limited to Pol epitopes suggested that mutations which disrupt peptide binding to protective HLA alleles (particularly HLA-A alleles) may be have greater fitness cost than escape from nonprotective HLA alleles (85).

Although results provide potentially useful information to guide future vaccine and immunogenicity studies, some limitations merit mention. We employed bioinformatics and statistical approaches to infer general mechanisms of immune evasion and its consequences for epitope presentation *in vivo*, but individual predictions will require rigorous experimental verification prior to consideration as potential immunogens. Furthermore, although well-powered analyses of HLA-associated polymorphisms identify the majority of escape pathways commonly driven by a given HLA (26), they may still underestimate the amount of immune selection imposed on HIV at the population level. For example, as our method detects only HIV polymorphisms whose frequencies differ statistically among persons having or lacking a particular HLA allele, it is unable to detect immune pressures directed against viral codons so highly conserved that the benefits of escape would not outweigh the fitness costs. The B*57-associated KF11 epitope provides a potential example of such a phenomenon: when presented by B*57:01 (but not B*57:03) in context of subtype B, a highly conserved, cross-reactive TCR repertoire is induced, from which the virus escapes with great difficulty (123). Furthermore, although such well-powered analyses can identify extremely rare pathways (26), they may still underestimate the amount of immune selection imposed on HIV at the individual level; indeed, the extent to which unique immune responses (and their escape pathways) occur and affect disease progression remains unclear. Our method may also preferentially detect straightforward escape pathways (such as those directly impacting peptide binding to HLA) over those that may be more "multifactorial" and/or variable among individuals (e.g., TCR escape that could occur at a variety of intraepitope positions depending on the individual's T-cell repertoire); the extent of this potential bias could be quantified by longitudinally analyzing escape sites in seroconverter cohorts (18, 47, 53). Moving forward, it will be essential to translate indirect correlates of protection at the HLA allele level to specific features of protective HLA-restricted CTL responses at the individual level.

The present study confirms the differential influence of HLA class I-mediated selection pressures on population-level sequence diversity of individual HIV-1 proteins and the common occur-

Downloaded from http://jvi.asm.org/ on November 13, 2018 by guest

rence of differential escape among closely related alleles (19, 24–26, 99, 104). That the majority of HLA-associated polymorphisms occur outside known epitopes suggests that many epitopes remain undiscovered; indeed, HLA-associated polymorphisms represent an excellent tool to guide epitope discovery as they are unbiased by consensus sequences or limited knowledge of binding motifs (5, 13, 17, 92). This study also extends our understanding of the proportion of population-level HIV-1 diversity attributable to HLA selection pressures, identifies abrogation of HLA-peptide binding as a predominant escape mechanism, suggests a potential evolutionary mechanism underlying differential escape between allele group members, and provides an extensive reference of proteome-wide HLA-mediated escape pathways.

Importantly, our hypothesis that HLA-associated polymorphisms mark sites under active immune selection by HLA alleles *in vivo* affords a novel perspective from which to identify features of such responses potentially relevant to vaccine design. Specifically, the tendency for protective alleles to exert selection pressures on a larger number of viral sites highlights selection breadth as a general correlate of protection. The tendency of HIV-1 to evade immune pressure by protective HLA alleles via mutations at anchor residues and/or at multiple sites within epitopes supports CTL repertoire composition and/or diversity as a correlate of protection (28, 58) although this remains controversial (82). The average sequence conservation of sites under selection was not a general correlate of protection, indicating that the factors underpinning protective immune responses are more complex than simply the ability to target conserved regions. However, that the ability to strongly mount immune pressure on conserved sites in Gag and Pol was protective extends the evidence supporting Gag as a particularly effective immune target (14, 36, 67, 124) and suggests that Gag (and Pol)-derived peptides possess characteristics which render them inherently more effective as immune targets: rapid presentation of Gag-derived and, to a lesser extent, Pol-derived peptides from incoming virions (106, 107) provides an intriguing explanation for this phenomenon. Taken together, our results suggest that vaccine strategies aimed at stimulating broad CTL responses toward non-Nef epitopes and, in particular, intense polyclonal responses that would force escape at conserved anchor positions, especially within constrained Gag and Pol epitopes, may be particularly effective at controlling HIV-1. Analyses of population-level escape pathways for other viral pathogens (such as hepatitis C virus) (45, 97, 118) may similarly illuminate relevant correlates of protective cellular immunity.

## ACKNOWLEDGMENTS

We gratefully acknowledge the technical, laboratory, and database staff at the BC Centre for Excellence in HIV/AIDS in Vancouver, Canada, and the Microsoft Research High Performance Computing team in Redmond, WA.

J.M.C., C.J.B., and Z.L.B. conceived and designed the study. J.M.C., C.J.B., E.M., J.L., M.A.B., N.P., C.E.D, D.H., R.A., and Z.L.B. contributed ideas, developed analytical methods, and/or performed data analysis and interpretation. A.Q.L., C.K.S.C, L.A.C., D.J.H.F.K., and Z.L.B. collected and/or analyzed experimental data. S.A.R., G.N., M.C., S.M., P.R.H., and M.J. contributed cohorts and data. J.M.C. and Z.L.B. wrote the paper. All authors critically reviewed the manuscript.

This work was supported by operating grants from the Canadian Institutes for Health Research (CIHR) (MOP-93536 and HOP-115700) to Z.L.B./M.A.B. In addition this project has been funded in part by NIAID grants AI27670, AI36214, and AI064086 to University of California, San Diego (UCSD; R. Haubrich) and in whole or in part with federal funds from the Frederick National Laboratory for Cancer Research, under contract number HHSN261200800001E. This Research was supported in part by the Intramural Research Program of the NIH, Frederick National Lab, Center for Cancer Research. C.J.B. is supported by a Vanier Canada Graduate Scholarship from the CIHR. E.M. is supported by a Master's Scholarship from the Canadian Association of HIV Research and Abbott Virology. M.A.B. holds a Canada Research Chair, Tier 2, in Viral Pathogenesis and Immunity. J.M.C., J.L., N.P., C.E.D., and D.H. are employees of Microsoft Corporation. Z.L.B. is the recipient of a CIHR New Investigator Award and a scholar award from the Michael Smith Foundation for Health Research. The ACTG Human DNA Repository is supported by grants AI068636 and RR024975. The following is a list of ACTG sites that participated in both A5142 and A5128 protocols, along with their grant numbers: Northwestern University (sites 2701, 2702, and 2705), Clinical Trials Unit (CTU) grant AI 069471; University of Minnesota (sites 1501, 1504, and 1505), CTU grant AI 27661; Vanderbilt University (site 3652), CTU grant AI-069439; Indiana University (sites 2601 and 2603), CTU grant AI25859 and GCRC grant MO1RR000750; University of Miami School of Medicine (site 901), CTU grant AI069477; University of Cincinnati (site 2401), CTU grant AI-069513; University of Alabama (sites 5801 and 5802), CTU grant 1 U01 AI069452-01 and GCRC grant M01 RR-00032; University of Southern California (site 1201), CTU grant AI27673; Cornell CTU (site 30329, 7803, and 7804), CTU grants AI069419-01 and CTSC RR024996; The Ohio State University (Site 2301), CTU grant AI069474; University of Rochester (sites 1101, 1102, 1107, and 1108), CTU grant AI69411 and GCRC grant 5-MO1 RR00044; Univeristy of North Carolina—Chapel Hill (site 3201), CFAR grant AI50410, CTU grant AI69423-01, and GCRC grant RR00046; University of Pittsburgh (site 1001 and 1008), CTU grant AI69494-01; Duke University Medical Center (site 1601), CTU grant 1U01-AI069484; Harvard/BMC CTU (sites 103, 104, and 107), CTU grant AI069472, CFAR grant AI060354, and GCRC grant RR02635; Durban International CTU (site 11201), grant UOIA138858; Case Western Reserve University (sites 2501, 2503, and 2508), CTU grant AI 069501; University of Pennsylvania, Philadelphia (sites 6201 and 6206), CTU grant AI 69467-01 and CFAR grant 5-P30-AI-045008-07; Colorado ACTU (site 6101), CTU grant AI069450 and GCRC grant RR00051; University of Texas Medical Branch—Galveston (site 6301), CTU grant AI32782; Johns Hopkins University (site 201), CTU grant AI-69465 and GCRC grant RR-00052; University of California, Los Angeles (sites 601 and 603), CTU grant AI069424; University of California, Davis Medical Center (site 3852), CTU grant AI38858-09S1; University of Maryland, Institute of Human Virology (site 4651), CTU grant AI069447-01; Washington University in St. Louis (site 2101), CTU grant AI069495; University of California, San Francisco (site 801), CTU grant AI069502-01; Stanford University (sites 501, 505, and 506), CTU grant AI069556; University of California, San Diego (site 701), grant AI069432; Beth Israel Medical Center (site 2851), CTU grant AI46370; New York University/New York City Health and Hospitals Corp. at Bellevue Hospital Center (site 401), CTU grant AI069532 and GCRC grant M01-RR00096; The Miriam Hospital (site 2951), CTU grant AI69472; University of Texas Southwestern Medical Center at Dallas (site 3751), CTU grant AI046376-05; University of Hawaii at Manoa and Queen's Medical Center (site 5201), CTU grant AI34853; University of Washington, Seattle (site 1401), CTU grant AI069434.

The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## REFERENCES

1. **Addo MM, et al.** 2003. Comprehensive epitope analysis of human immunodeficiency virus type 1 (HIV-1)-specific T-cell responses directed

Downloaded from http://jvi.asm.org/ on November 13, 2018 by guest

against the entire expressed HIV-1 genome demonstrate broadly directed responses, but no correlation to viral load. J. Virol. **77**:2081–2092.
2. **Akaike H.** 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. **19**:716–723.
3. **Allen TM, et al.** 2005. Selective escape from CD8+ T-cell responses represents a major driving force of human immunodeficiency virus type 1 (HIV-1) sequence diversity and reveals constraints on HIV-1 evolution. J. Virol. **79**:13239–13249.
4. **Almeida CA, et al.** 2011. 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.
5. **Almeida CA, et al.** 2010. Exploiting knowledge of immune selection in HIV-1 to detect HIV-specific CD8 T-cell responses. Vaccine **28**:6052–6057.
6. **Almeida JR, et al.** 2007. Superior control of HIV-1 replication by $CD8^+$ T cells is reflected by their avidity, polyfunctionality, and clonal turnover. J. Exp. Med. **204**:2473–2485.
7. **Altfeld M, Allen TM.** 2006. Hitting HIV where it hurts: an alternative approach to HIV vaccine design. Trends Immunol. **27**:504–510.
8. **Ammaranond P, et al.** 2005. A new variant cytotoxic T lymphocyte escape mutation in HLA-B27-positive individuals infected with HIV type 1. AIDS Res. Hum. Retroviruses **21**:395–397.
9. **Avila-Rios S, et al.** 2009. Unique features of HLA-mediated HIV evolution in a Mexican cohort: a comparative study. Retrovirology **6**:72.
10. **Bansal A, et al.** 2010. CD8 T cell response and evolutionary pressure to HIV-1 cryptic epitopes derived from antisense transcription. J. Exp. Med. **207**:51–59.
11. **Barouch DH, et al.** 2010. Mosaic HIV-1 vaccines expand the breadth and depth of cellular immune responses in rhesus monkeys. Nat. Med. **16**:319–323.
12. **Berger CT, et al.** 2010. Viral adaptation to immune selection pressure by HLA class I-restricted CTL responses targeting epitopes in HIV frameshift sequences. J. Exp. Med. **207**:61–75.
13. **Bhattacharya T, et al.** 2007. Founder effects in the assessment of HIV polymorphisms and HLA allele associations. Science **315**:1583–1586.
14. **Borghans JA, Molgaard A, de Boer RJ, Kesmir C.** 2007. HLA alleles associated with slow progression to AIDS truly prefer to present HIV-1 p24. PLoS One **2**:e920. doi:10.1371/journal.pone.0000920.
15. **Borrow P, et al.** 1997. Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. Nat. Med. **3**:205–211.
16. **Brackenridge S, et al.** 2011. An early HIV mutation within an HLA-B*57-restricted T cell epitope abrogates binding to the killer inhibitory receptor 3DL1. J. Virol. **85**:5415–5422.
17. **Brockman MA, et al.** 2012. Uncommon pathways of immune escape attenuate HIV-1 integrase replication capacity. J. Virol. **86**:6913–6923.
18. **Brumme ZL, et al.** 2008. 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.
19. **Brumme ZL, et al.** 2007. Evidence of Differential HLA Class I-Mediated Viral Evolution in Functional and Accessory/Regulatory Genes of HIV-1. PLoS Pathog. **3**:e94. doi:10.1371/journal.ppat.0030094.
20. **Brumme ZL, et al.** 2009. HLA-associated immune escape pathways in HIV-1 subtype B Gag, Pol and Nef proteins. PLoS One **4**:e6687. doi:10.1371/journal.pone.0006687.
21. **Brumme ZL, et al.** 2008. Human leukocyte antigen-specific polymorphisms in HIV-1 Gag and their association with viral load in chronic untreated infection. AIDS **22**:1277–1286.
22. **Buranapraditkun S, et al.** 2011. A novel immunodominant $CD8^+$ T cell response restricted by a common HLA-C allele targets a conserved region of Gag HIV-1 clade CRF01_AE infected Thais. PLoS One **6**:e23603. doi:10.1371/journal.pone.0023603.
23. **Cale EM, Bazick HS, Rianprakaisang TA, Alam SM, Letvin NL.** 2011. Mutations in a dominant Nef epitope of simian immunodeficiency virus diminish TCR:epitope peptide affinity but not epitope peptide:MHC class I binding. J. Immunol. **187**:3300–3313.
24. **Carlson J, Kadie C, Mallal S, Heckerman D.** 2007. Leveraging hierarchical population structure in discrete association studies. PLoS One **2**:e591. doi:10.1371/journal.pone.0000591.
25. **Carlson JM, et al.** 2008. Phylogenetic dependency networks: inferring patterns of CTL escape and codon covariation in HIV-1 Gag. PLoS Comput. Biol. **4**:e1000225. doi:10.1371/journal.pcbi.1000225.
26. **Carlson JM, et al.** 2012. Widespread impact of HLA restriction on immune control and escape Pathways of HIV-1. J. Virol. **86**:5230–5243.
27. **Carrington M, O'Brien SJ.** 2003. The influence of HLA genotype on AIDS. Annu. Rev. Med. **54**:535–551.
28. **Chen H, et al.** 2012. TCR clonotypes modulate the protective effect of HLA class I molecules in HIV-1 infection. Nat. Immunol. **13**:691–700.
29. **Chen H, et al.** 2009. Differential neutralization of human immunodeficiency virus (HIV) replication in autologous CD4 T cells by HIV-specific cytotoxic T lymphocytes. J. Virol. **83**:3138–3149.
30. **Chopera DR, Wright JK, Brockman MA, Brumme ZL.** 2011. Immune-mediated attenuation of HIV-1. Future Virol. **6**:917–928.
31. **Chouquet C, et al.** 2002. Correlation between breadth of memory HIV-specific cytotoxic T cells, viral load and disease progression in HIV infection. AIDS **16**:2399–2407.
32. **Cotton LA, et al.** 2012. HLA class I sequence-based typing using DNA recovered from frozen plasma. J. Immunol. Methods **382**:40–47.
33. **Dahirel V, et al.** 2011. Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. Proc. Natl. Acad. Sci. U. S. A. **108**:11530–11535.
34. **Dempster AP, Laird NM, Rubin DB.** 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Series B Stat. Methodol. **39**:1–38.
35. **Draenert R, et al.** 2004. Immune selection for altered antigen processing leads to cytotoxic T lymphocyte escape in chronic HIV-1 infection. J. Exp. Med. **199**:905–915.
36. **Edwards BH, et al.** 2002. Magnitude of functional $CD8^+$ T-cell responses to the gag protein of human immunodeficiency virus type 1 correlates inversely with viral load in plasma. J. Virol. **76**:2298–2305.
37. **Falk K, Rotzschke O, Stevanovic S, Jung G, Rammensee HG.** 1991. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature **351**:290–296.
38. **Feeney ME, et al.** 2005. HIV-1 viral escape in infancy followed by emergence of a variant-specific CTL response. J. Immunol. **174**:7524–7530.
39. **Feeney ME, et al.** 2004. 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.
40. **Fellay J, et al.** 2009. Common genetic variation and the control of HIV-1 in humans. PLoS Genet. **5**:e1000791. doi:10.1371/journal.pgen.1000791.
41. **Felsenstein J.** 2005. PHYLIP (phylogeny inference package) version 3.6. Department of Genome Sciences, University of Washington, Seattle, WA.
42. **Fischer W, et al.** 2007. Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. Nat. Med. **13**:100–106.
43. **Gao X, et al.** 2001. Effect of a single amino acid change in MHC class I molecules on the rate of progression to AIDS. N. Engl. J. Med. **344**:1668–1675.
44. **Gaschen B, et al.** 2002. Diversity considerations in HIV-1 vaccine selection. Science **296**:2354–2360.
45. **Gaudieri S, et al.** 2006. Evidence of viral adaptation to HLA class I-restricted immune pressure in chronic hepatitis C virus infection. J. Virol. **80**:11094–11104.
46. **Gnanakaran S, et al.** 2011. Recurrent signature patterns in HIV-1 B clade envelope glycoproteins associated with either early or chronic infections. PLoS Pathog. **7**:e1002209. doi:10.1371/journal.ppat.1002209.
47. **Goonetilleke N, et al.** 2009. 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.
48. **Goulder PJ, et al.** 2000. Differential narrow focusing of immunodominant human immunodeficiency virus gag-specific cytotoxic T-lymphocyte responses in infected African and caucasoid adults and children. J. Virol. **74**:5679–5690.
49. **Goulder PJ, et al.** 1997. Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS. Nat. Med. **3**:212–217.
50. **Goulder PJ, Watkins DI.** 2004. HIV and SIV CTL escape: implications for vaccine design. Nat. Rev. Immunol. **4**:630–640.
51. **Haas DW, et al.** 2003. A multi-investigator/institutional DNA bank for AIDS-related human genetic studies: AACTG Protocol A5128. HIV Clin. Trials **4**:287–300.
52. **Hasan Z, et al.** 2012. Minor contribution of HLA class I-associated

Downloaded from http://jvi.asm.org/ on November 13, 2018 by guest

selective pressure to the variability of HIV-1 accessory protein Vpu. Biochem. Biophys. Res. Commun. **421**:291–295.
53. **Henn MR, et al.** 2012. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog. **8**:e1002529. doi:10.1371/journal.ppat.1002529.
54. **Holdsworth R, et al.** 2009. The HLA dictionary 2008: a summary of HLA-A, -B, -C, -DRB1/3/4/5, and -DQB1 alleles and their association with serologically defined HLA-A, -B, -C, -DR, and -DQ antigens. Tissue Antigens **73**:95–170.
55. **Honeyborne I, et al.** 2006. Motif inference reveals optimal CTL epitopes presented by HLA class I alleles highly prevalent in southern Africa. J. Immunol. **176**:4699–4705.
56. **Horton H, et al.** 2006. Preservation of T cell proliferation restricted by protective HLA alleles is critical for immune control of HIV-1 infection. J. Immunol. **177**:7406–7415.
57. **Huang KH, et al.** 2011. Progression to AIDS in South Africa is associated with both reverting and compensatory viral mutations. PLoS One **6**:e19018. doi:10.1371/journal.pone.0019018.
58. **Iglesias MC, et al.** 2011. Escape from highly effective public CD8$^+$ T-cell clonotypes by HIV. Blood **118**:2138–2149.
59. **Iversen AK, et al.** 2006. Conflicting selective forces affect T cell receptor contacts in an immunodominant human immunodeficiency virus epitope. Nat. Immunol. **7**:179–189.
60. **John M, et al.** 2010. 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.
61. **Johnson VA, et al.** 2011. 2011 Update of the drug resistance mutations in HIV-1. Top. Antivir. Med. **19**:156–164.
62. **Kaslow RA, et al.** 1996. Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection. Nat. Med. **2**:405–411.
63. **Kaul R, et al.** 2001. CD8$^+$ lymphocytes respond to different HIV epitopes in seronegative and infected subjects. J. Clin. Invest. **107**:1303–1310.
64. **Kawashima Y, et al.** 2009. Adaptation of HIV-1 to human leukocyte antigen class I. Nature **458**:641–645.
65. **Kelleher AD, et al.** 2001. Clustered mutations in HIV-1 gag are consistently required for escape from HLA-B27-restricted cytotoxic T lymphocyte responses. J. Exp. Med. **193**:375–386.
66. **Kiepiela P, et al.** 2004. Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. Nature **432**:769–775.
67. **Kiepiela P, et al.** 2007. CD8$^+$ T-cell responses to different HIV proteins have discordant associations with viral load. Nat. Med. **13**:46–53.
68. **Kloverpris HN, et al.** 2012. HIV control through a single nucleotide on the HLA-B locus. J. Virol. **86**:11493–11500.
69. **Kloverpris HN, et al.** 2012. HLA-B*57 micropolymorphism shapes HLA allele-specific epitope immunogenicity, selection pressure, and HIV immune control. J. Virol. **86**:919–929.
70. **Kosmrlj A, et al.** 2010. Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection. Nature **465**:350–354.
71. **Koup RA.** 1994. Virus escape from CTL recognition. J. Exp. Med. **180**:779–782.
72. **Larder BA, Kemp SD.** 1989. Multiple mutations in HIV-1 reverse transcriptase confer high-level resistance to zidovudine (AZT). Science **246**:1155–1158.
73. **Leslie A, et al.** 2005. Transmission and accumulation of CTL escape variants drive negative associations between HIV polymorphisms and HLA. J. Exp. Med. **201**:891–902.
74. **Leslie A, et al.** 2006. 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.
75. **Listgarten J, et al.** 2008. Statistical resolution of ambiguous HLA typing data. PLoS Comput. Biol. **4**:e1000016. doi:10.1371/journal.pcbi.1000016.
76. **Llano A, Frahm N, Brander C.** 2009. How to optimally define optimal cytotoxic T lymphocyte epitopes in HIV infection, p 3–24. *In* Yusim K, et al (ed), HIV molecular immunology. Los Alamos National Laboratory, Los Alamos, NM.
77. **Mallal SA.** 1998. The Western Australian HIV Cohort Study, Perth, Australia. J. Acquir. Immune Defic. Syndr. Hum. Retrovirol. **17**(Suppl 1):S23–S27.
78. **Marsh SGE, Parham P, Barber LD.** 2000. The HLA factsbook. Academic Press, London, United Kingdom.
79. **Matthews PC, et al.** 2011. HLA-A*7401-mediated control of HIV viremia is independent of its linkage disequilibrium with HLA-B*5703. J. Immunol. **186**:5675–5686.
80. **Matthews PC, et al.** 2009. HLA Footprints on HIV-1 are associated with inter-clade polymorphisms and intra-clade phylogenetic clustering. J. Virol. **83**:4605–4615.
81. **Matthews PC, et al.** 2008. Central role of reverting mutations in HLA associations with human immunodeficiency virus set point. J. Virol. **82**:8548–8559.
82. **Mendoza D, et al.** 2012. HLA B*5701-positive long-term nonprogressors/elite controllers are not distinguished from progressors by the clonal composition of HIV-specific CD8$^+$ T cells. J. Virol. **86**:4014–4018.
83. **Miles JJ, Douek DC, Price DA.** 2011. Bias in the αβ T-cell repertoire: implications for disease pathogenesis and vaccination. Immunol. Cell Biol. **89**:375–387.
84. **Moore CB, et al.** 2002. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. Science **296**:1439–1443.
85. **Mostowy R, et al.** 2012. Estimating the fitness cost of escape from HLA presentation in HIV-1 protease and reverse transcriptase. PLoS Comput. Biol. **8**:e1002525. doi:10.1371/journal.pcbi.1002525.
86. **Ngumbela KC, et al.** 2008. Targeting of a CD8 T cell env epitope presented by HLA-B*5802 is associated with markers of HIV disease progression and lack of selection pressure. AIDS Res. Hum. Retroviruses **24**:72–82.
87. **Nickle DC, et al.** 2007. Coping with viral diversity in HIV vaccine design. PLoS Comput. Biol. **3**:e75. doi:10.1371/journal.pcbi.0030075.
88. **Nielsen M, et al.** 2007. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. PLoS One **2**:e796. doi:10.1371/journal.pone.0000796.
89. **O'Brien SJ, Gao X, Carrington M.** 2001. HLA and AIDS: a cautionary tale. Trends Mol. Med. **7**:379–381.
90. **Oxenius A, et al.** 2004. 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.
91. **Parker KC, et al.** 1992. Sequence motifs important for peptide binding to the human MHC class I molecule, HLA-A2. J. Immunol. **149**:3580–3587.
92. **Payne RP, et al.** 2010. Efficacious early antiviral activity of HIV Gag- and Pol-specific HLA-B 2705-restricted CD8$^+$ T cells. J. Virol. **84**:10543–10557.
93. **Pereyra F, et al.** 2010. The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. Science **330**:1551–1557.
94. **Phillips RE, et al.** 1991. Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. Nature **354**:453–459.
95. **Price DA, et al.** 1997. Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. Proc. Natl. Acad. Sci. U. S. A. **94**:1890–1895.
96. **Price DA, et al.** 2004. T cell receptor recognition motifs govern immune escape patterns in acute SIV infection. Immunity **21**:793–803.
97. **Rauch A, et al.** 2009. Divergent adaptation of hepatitis C virus genotypes 1 and 3 to human leukocyte antigen-restricted immune pressure. Hepatology **50**:1017–1029.
98. **Robinson J, et al.** 2011. The IMGT/HLA database. Nucleic Acids Res. **39**:D1171–D1176.
99. **Rolland M, et al.** 2010. Amino-acid co-variation in HIV-1 Gag subtype C: HLA-mediated selection pressure and compensatory dynamics. PLoS One **5**:e12463. doi:10.1371/journal.pone.0012463.
100. **Rolland M, et al.** 2011. Increased breadth and depth of cytotoxic T lymphocytes responses against HIV-1-B Nef by inclusion of epitope variant sequences. PLoS One **6**:e17969. doi:10.1371/journal.pone.0017969.
101. **Rolland M, et al.** 2008. Broad and Gag-biased HIV-1 epitope repertoires are associated with lower viral loads. PLoS One **3**:e1424. doi:10.1371/journal.pone.0001424.
102. **Rolland M, Nickle DC, Mullins JI.** 2007. HIV-1 group M conserved elements vaccine. PLoS Pathog. **3**:e157. doi:10.1371/journal.ppat.0030157.
103. **Rolland M, et al.** 2011. Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. Nat. Med. **17**:366–371.
104. **Rousseau CM, et al.** 2008. HLA class-I driven evolution of human

Downloaded from http://jvi.asm.org/ on November 13, 2018 by guest

immunodeficiency virus type 1 subtype C proteome: immune escape and viral load. J. Virol. **82**:6434–6446.
105. **Sabbaj S, et al.** 2003. Cross-reactive CD8$^+$ T cell epitopes identified in US adolescent minorities. J. Acquir. Immune Defic. Syndr. **33**:426–438.
106. **Sacha JB, et al.** 2007. Gag-specific CD8$^+$ T lymphocytes recognize infected cells before AIDS-virus integration and viral protein expression. J. Immunol. **178**:2746–2754.
107. **Sacha JB, et al.** 2007. Pol-specific CD8$^+$ T cells recognize simian immunodeficiency virus-infected cells prior to Nef-mediated major histocompatibility complex class I downregulation. J. Virol. **81**:11703–11712.
108. **Saez-Cirion A, et al.** 2007. HIV controllers exhibit potent CD8 T cell capacity to suppress HIV infection ex vivo and peculiar cytotoxic T lymphocyte activation phenotype. Proc. Natl. Acad. Sci. U. S. A. **104**:6776–6781.
109. **Santra S, et al.** 2010. Mosaic vaccines elicit CD8$^+$ T lymphocyte responses that confer enhanced immune coverage of diverse HIV strains in monkeys. Nat. Med. **16**:324–328.
110. **Schuler MM, Nastke MD, Stevanovikc S.** 2007. SYFPEITHI: database for searching and T-cell epitope prediction. Methods Mol. Biol. **409**:75–93.
111. **Sidney J, Peters B, Frahm N, Brander C, Sette A.** 2008. HLA class I supertypes: a revised and updated classification. BMC Immunol. **9**:1. doi:10.1186/1471-2172-9-1.
112. **Snoeck J, Fellay J, Bartha I, Douek DC, Telenti A.** 2011. Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints. Retrovirology **8**:87. doi:10.1186/1742-4690-8-87.
113. **Stewart-Jones GB, et al.** 2005. Crystal structures and KIR3DL1 recognition of three immunodominant viral peptides complexed to HLA-B*2705. Eur. J. Immunol. **35**:341–351.
114. **Storey JD, Tibshirani R.** 2003. Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. U. S. A. **100**:9440–9445.
115. **Streeck H, Frahm N, Walker BD.** 2009. The role of IFN-gamma Elispot assay in HIV vaccine research. Nat. Protoc. **4**:461–469.
116. **Taylor BS, Hammer SM.** 2008. The challenge of HIV-1 subtype diversity. N. Engl. J. Med. **359**:1965–1966.
117. **Theodossis A, et al.** 2010. Constraints within major histocompatibility complex class I restricted peptides: presentation and consequences for T-cell recognition. Proc. Natl. Acad. Sci. U. S. A. **107**:5534–5539.
118. **Timm J, et al.** 2007. Human leukocyte antigen-associated sequence polymorphisms in hepatitis C virus reveal reproducible immune responses and constraints on viral evolution. Hepatology **46**:339–349.
119. **Veerassamy S, Smith A, Tillier ER.** 2003. A transition probability model for amino acid substitutions from blocks. J. Comput. Biol. **10**:997–1010.
120. **Woods CK, et al.** 2012. Automating HIV drug resistance genotyping with RECall, a freely accessible sequence analysis tool. J. Clin. Microbiol. **50**:1936–1942.
121. **Yang OO.** 2003. Will we be able to "spot" an effective HIV-1 vaccine? Trends Immunol. **24**:67–72.
122. **Yokomaku Y, et al.** 2004. Impaired processing and presentation of cytotoxic-T-lymphocyte (CTL) epitopes are major escape mechanisms from CTL immune pressure in human immunodeficiency virus type 1 infection. J. Virol. **78**:1324–1332.
123. **Yu XG, et al.** 2007. Mutually exclusive T-cell receptor induction and differential susceptibility to human immunodeficiency virus type 1 mutational escape associated with a two-amino-acid difference between HLA class I subtypes. J. Virol. **81**:1619–1631.
124. **Zuniga R, et al.** 2006. Relative dominance of Gag p24-specific cytotoxic T lymphocytes is associated with human immunodeficiency virus control. J. Virol. **80**:3122–3125.

Downloaded from http://jvi.asm.org/ on November 13, 2018 by guest