Adaptive Interactions between HLA and HIV-1: Highly Divergent Selection Imposed by HLA Class I Molecules with Common Supertype Motifs

Mina John, David Heckerman, Ian James, Lawrence P. Park, Jonathan M. Carlson, Abha Chopra, Silvana Gaudieri, David Nolan, David W. Haas, Sharon A. Riddler, Richard Haubrich and Simon Mallal

*J Immunol* 2010; 184:4368-4377; Prepublished online 15 March 2010; doi: 10.4049/jimmunol.0903745

http://www.jimmunol.org/content/184/8/4368

Supplementary Material

http://www.jimmunol.org/content/suppl/2010/03/15/jimmunol.0903745.DC1

References

This article cites 54 articles, 25 of which you can access for free at:

http://www.jimmunol.org/content/184/8/4368.full#ref-list-1

Subscription

Information about subscribing to *The Journal of Immunology* is online at:

http://jimmunol.org/subscription

Permissions

Submit copyright permission requests at:

http://www.aai.org/About/Publications/JI/copyright.html

Email Alerts

Receive free email-alerts when new articles cite this article. Sign up at:

http://jimmunol.org/alerts
Adaptive Interactions between HLA and HIV-1: Highly Divergent Selection Imposed by HLA Class I Molecules with Common Supertype Motifs

Mina John,*† David Heckerman,‡ Ian James,* Lawrence P. Park,* Jonathan M. Carlson,‡ Abha Chopra,* Silvana Gaudieri,*‡§ David Nolan,*† David W. Haas,*†|| Sharon A. Riddler,# Richard Haubrich,** and Simon Mallal*†

Currently, 1.1 million individuals in the United States are living with HIV-1 infection. Although this is a relatively small proportion of the global pandemic, the remarkable mix of ancestries in the United States, drawn together over the past two centuries of continuous population migrations, provides an important and unique perspective on adaptive interactions between HIV-1 and human genetic diversity. HIV-1 is a rapidly adaptable organism and mutates within or near immune epitopes that are determined by the HLA class I genotype of the infected host. We characterized HLA-associated polymorphisms across the full HIV-1 proteome in a large, ethnically diverse national United States cohort of HIV-1-infected individuals. We found a striking divergence in the immunoselection patterns associated with HLA variants that have very similar or identical peptide-binding specificities but are differentially distributed among racial/ethnic groups. Although their similarity in peptide binding functionally clusters these HLA variants into supertypes, their differences at sites within the peptide-binding groove contribute to race-specific selection effects on circulating HIV-1 viruses. This suggests that the interactions between the HLA/HIV peptide complex and the TCR vary significantly within HLA supertype groups, which, in turn, influences HIV-1 evolution.

proteome, which may compensate or precondition primary escape mutations (19–21), are also HLA allele specific (22). Evidence that adaptive changes in HIV accumulate in populations as a function of the frequency of the selecting HLA alleles has been published (23). The extent to which the HLA signatures from these studies and their population dynamics can be generalized more broadly to other populations depends, to a significant extent, on how HLA subtypes differentially distributed among different racial groups, even within allelic families, diverge or converge in their interactions with HIV peptides. There has been limited information about this from population-based studies because low-resolution HLA genotyping has been used, because particular molecular splits of broad HLA types predominate in less admixed cohorts, or because comparisons across cohorts may be confounded by different viral subtypes. The worldwide clade B epidemic includes immunogenetically distinct populations in Asia, the Pacific region, the Caribbean, and particularly Central and South America, where pathogen diversity has been associated with significant HLA allelic diversification within lineages, particularly at the HLA-A locus (24). Human migrations have also led to increasing racial/ethnic diversity in western countries affected by subtype B HIV, exemplified by the United States epidemic (25). The extent to which the selection effects of HLA in distinct ethnic groups leads to population-level divergence and subsequent clustering of viral genome or epitope diversity has not been fully elucidated. Therefore, the United States epidemic presents the unique conditions required to characterize differential HIV-1 adaptation patterns in diverse ethnic groupings in the context of one predominant viral subtype.

Conversely, the imprints of selection in HIV should inform an understanding of diversifying selective forces on HLA peptide-binding specificities and more subtle HLA gene variations affecting TCR recognition of HLA-peptide complexes. The latter may be underestimated in more homogenous populations, in which a spread of related HLA variants is not present. In this study, we used high-resolution HLA genotyping, self-identified race/ethnicity information, sequences of all HIV genes, and published and novel analytical methods to examine convergence and divergence of HLA-associated adaptation patterns in HIV-1. We used a large cohort drawn from 55 centers throughout the country as a snapshot of contemporary United States genetic and HIV-1 diversity and the adaptive interactions between them.

Materials and Methods

Study populations

The 555 study subjects drawn from 55 participating centers across the United States were enrolled in AIDS Clinical Trials Group (ACTG) protocol A5128 (26) and provided DNA under protocol A5128 (27). Protocol A5128 facilitates storage and retrieval of DNA samples from A5128/A3128 participants for ACTG genetics studies. Race/ethnic groups were predefined by standard ACTG protocol, and individuals were self-classified into these categories at study enrollment. Selected analyses also used a comparator population of 245 individuals recruited into the Western Australian (WA) HIV Cohort Study, a population-based, observational cohort study. In both cohorts, patients provided written informed consent to these investigations, and studies received approval from their respective Institutional Review Boards prior to commencement.

HLA genotyping

HLA class I genotypes (at HLA-A, -B, and -C loci) were determined based on locus-specific PCR amplification of exons 2–3 and were all resolved to unambiguous four-digit-level resolution using standard DNA sequence-based typing. Ambiguities were resolved following sequencing with allele-specific subtyping primers. Sequence electropherograms were analyzed using Assign (Conexio Genomics, Applecross, Australia). HLA allele frequencies in the ACTG study population were compared with published data derived from the U.S. National Marrow Donor Program random sampling of >1000 anonymous unrelated subjects self-classified into five major ethnic groups in the United States (7).

HIV-1 sequencing

Standard bulk sequencing, using a nested PCR approach, was used for the United States cohort HIV-1 sequencing. HIV-1 RNA was isolated from pretreatment-stored plasma using lysis buffer. After reverse transcription (Superscript III Reverse Transcriptase; Invitrogen, Carlsbad, CA), two overlapping 5’ and 3’ fragments, ~6 kb in length and spanning the full HIV-1 genome, were amplified by nested PCR. Bidirectional sequencing on all fragments was performed using an ABI 3130XL Analyzer (Hitachi, Singapore), with iterative gap filling using alternative primers. Electropherograms were analyzed and edited using Assign (Conexio Genomics). Standard pre- and post-PCR guidelines were followed, and a customized laboratory information-management system tracked single PCRs with individual sample numbers, including well positions, controls, and other PCR details. Because of the nested PCR approach, contamination was monitored using PCR negative controls for first- and second-round PCRs. The large overlapping region between 5’ and 3’ fragments allowed an inbuilt test of sequencing precision and contamination, and individual sequence fragments were analyzed for phylogenetic relatedness routinely. HIV-1 sequences were aligned against the full genome of reference sequence HXB2 as near full-length genomes or partial-length sequences with intervening gaps. All insertion/deletion mixtures (predominantly in env, but detected at low frequencies in all proteins) resulting in stretches of unalignable International Index of Proteins and Applied Chemistry codes were removed, causing gaps in sequence. Areas not sequenced because of repeated sequencing or amplification failures also created gaps. A table showing the number of sequenced bases, number and proportion of mixtures, and number and size of gaps for gag, pol, and nef per individual is provided in the supplemental data. For the computation of single-site HIV–HLA associations as described below, the United States cohort sequence fragments were combined with previously generated WA HIV Cohort data to generate a total sequence dataset of 800 patient-derived sequences, and HLA genotypes and associations were adjusted for potential phylogenetic clustering across the two cohorts. For the construction of maximum-likelihood phylogenetic trees in this analysis (28), genome alignments considered HIV-1 genes (for gag: n = 746, pol: n = 781, nef: n = 616, and env: n = 742) or 1-kb sequence blocks (covering rev: n = 708, tat: n = 709, vpr: n = 709, and vpu: n = 715). All gene and sequence blocks included in the analysis were required to have >800 nucleotides; however, the median level of coverage within all analyzed non-env genes was 98.2–100% per gene and 79.14% in env.

Computation of single-site HIV–HLA polymorphism associations

Associations between HLA alleles and HIV polymorphisms were tested using methods developed to incorporate viral phylogenetic tree structure and HLA linkage disequilibrium (LD) into correlations and have been used in published studies (15, 29, 30). First, standard maximum-likelihood phylogenetic trees were constructed from sequences covering Gag, Pol, Nef, and Env as separate proteins and 1-kb sequence blocks covering Rev, Tat, Vif, Vpr, and Vpu. For every viral amino acid and HLA allele combination, a likelihood ratio test was used to evaluate whether a model incorporating viral phylogenetic structure and HLA-mediated selection pressure explained the observed data better than did a model assuming neutral evolution according to the tree only. Two generative models for escape response in viral sequences were compared representing the two alternative models; the likelihood of the observations was maximized over the parameters of both models using an Expectation-Maximization algorithm. A p value was generated using a likelihood ratio test based on those likelihoods. For every contingency table for a given amino acid–HLA allele pair, associations were defined as denoting “attraction”: enrichment of the specified amino acid in the specified HLA group, “repulsion”: depletion of the amino acid in the non-HLA allele group, “escape”: depletion of non-amino acid in the HLA group, and “reversion”: enrichment of amino acid in the non-HLA group. Thus, escape/reversion associations identified the reversion/nonadapted/wild-type amino acid and repulsion/attraction associations identified the adapted amino acid for a given HLA-associated amino acid substitution. The numbers of individuals with/without each HLA allele and with/without each viral polymorphism had to exceed an actual value of 3 to limit instability due to the use of large sample approximations and the possibility of misclassification. All comparisons were tested using HLA genotypes at two- and four-digit resolutions to maximize specificity for effects by four-digit subtypes with
different peptide binding but retain sensitivity for effects seen only at the two-digit level because of low-frequency subtypes. A multivariate analysis was used to identify and eliminate HLA associations driven by positive LD between HLA alleles (31). For every viral amino acid at each codon, the HLA allele with the strongest association was added to the list of identified associations. Then, individuals expressing this allele were removed from the dataset, and the analysis was repeated. This standard forward-selection procedure was iterated until no HLA allele yielded an association with uncorrected \( p < 0.05 \). Because there were multiple comparisons made in the analyses, we used \( q \) values, which estimate the false-discovery rate among identified associations compared with a randomly permuted dataset (32). Only associations with \( q \leq 0.2 \), indicating a 20% false-discovery rate, are presented in the paper.

Web resources Source code for the program used to compute HLA–HIV-1 polymorphism associations is available at www.microsoft.com/science.

Correlations between HLA matching and viral sequence similarity

Correlations were tested between pairwise viral sequence dissimilarity and the matching of HLA genotype in individuals carrying those sequences as a more global analysis of the influence of HLA on viral sequence divergence. Analyses were restricted to cases that had \( \geq 90\% \) sequence coverage within the protein window used and to windows for which \( \geq 80 \) cases satisfied this restriction. Dissimilarity of HLA genotypes was measured by the number of nonmatching alleles, counting homozygote alleles as two alleles. Viral sequence dissimilarity was measured by pairwise differences in amino acid using the binary Manhattan metric, which adjusts automatically for missing values. Correlations were tested over full-length Gag, and then localized correlations were tested over sliding intervals of 10 aa across Gag representing an epitope-length window. All analyses were carried out using SPlus 8.0 for HLA-A, -B, and -C loci combined and separately. Because the pairwise dissimilarity scores were not independent, significance was assessed by randomization tests in which HLA genotypes and viral sequences were permuted, and the standard \( R^2 \) was compared with the randomization distribution. Five hundred permutations were used to estimate \( p \) values, thus truncating \( p \) values at 1/500.

Results

The study population consisted of 222 white non-Hispanics (40%), 211 black non-Hispanics (38%), 108 Hispanics (regardless of race; 19.5%), 11 Asians/Pacific Islanders (2%), and 3 others (1 Alaskan Native American, 1 identified as of more than one race; and 1 “other/unknown”). These proportions were comparable with those reported as United States HIV prevalence estimates—white (34.6%), black (46.1%), Hispanic (17.5%), and Asian/Pacific Islander (1.4%)—by the Centers for Disease Control and Prevention based on the national HIV/AIDS Reporting System up to 2006 (25). Other epidemiological and clinical data are provided in Supplemental Table I.

HLA class I genetic diversity

HLA diversity within the study population varied substantially according to race/ethnicity (Figs. 1, 2). Of HLA alleles carried by \( \geq 3\% \) of individuals within at least one ethnic group, 20 HLA-A and 31 HLA-B alleles were identified. Of these, nine (45%) HLA-A and six (18.7%) HLA-B alleles were common to all racial/ethnic groups. The proportion of relatively race-specific genotypes was greater within the HLA-B locus than for HLA-A. Hence, 1 of 13 (7.7%) HLA-A and 5 of 16 (31.3%) HLA-B alleles could be considered relatively specific to the white population. Similarly, 2 of 18 (11.1%) HLA-A and 6 of 16 (37.5%) HLA-B alleles were specific to Hispanics, whereas the black population provided the largest proportion.

![Figure 1](http://www.jimmunol.org/) - HLA allele carriage rate in groups defined by self-identified race/ethnicity for HLA-A (A), HLA-B (B), and HLA-C (C) loci. Only HLA alleles with >3% carriage frequency are included (n = 555).
of race-specific HLA-A alleles (5 of 17, 29.1%) and HLA-B alleles (8 of 16, 50.0%). Phylogenetic trees based on the peptide-binding regions of all HLA class I alleles were constructed (Supplemental Fig. 1). Several groups of HLA alleles that appeared more closely associated within a lineage included highly race-specific alleles, such as HLA-B*5701 and -B*5703, which differ at only two residues in the peptide-binding domain (Supplemental Fig. 2), but are distributed almost exclusively in whites and blacks, respectively (Fig. 2). These trees reproduced many previously noted relationships, although strong support for phylogenetic relationships between HLA-A and -B alleles is limited by high levels of intra-allelic recombination at these loci through human and primate evolution (33).

The gene frequencies of HLA alleles within the study subpopulations were formally compared with frequencies in the only contemporary large-scale population-based sampling of United States ethnic groups available (7) (Fig. 1, Supplemental Table II). After adjustment for the number of comparisons made over each HLA/ethnic group combination, there were no significant differences in the frequency of any HLA-A allele between whites and blacks; however, there was enrichment of HLA-B*5703 in the black study cohort subpopulation (5.07%) compared with blacks in the healthy reference population (0.4%, \( p = 0.000067 \)) (7). Among Hispanics, HLA-A*6803, -B*4201, -C*0305, and -C*1701 were increased in the HIV-infected cohort compared with background rates (\( p = 0.004, p = 0.016, p = 0.042, \) and \( p = 0.00027 \), respectively), whereas HLA-C*0304 was depleted (\( p = 0.022 \)).

**FIGURE 2.** Prevalence of HLA alleles in relation to specific race/ethnic groups in the United States cohort. HLA-B/C genotype combinations in which the relevant HLA-B allele is accompanied by its haplotypic HLA-C allele at a frequency >80% in the cohort are also shown. Supertype groups are in superscript. UC, unclassified supertype.

**HIV-1 genetic diversity**

HIV-1 sequences were identified as clade B in 97% of cases, based on phylogenetic analysis of all HIV genes. There were six minor clusters with bootstrap values \( >70% \) detected in these analyses, all of which had less than five individuals within them, with no consistent sharing of study site, city, or ethnicity to suggest recent transmission clusters. When United States gag, pol, and nef sequences were compared with those from geographically distant individuals from WA (\( n = 245, 87\% \) subtype B), sequences from Australian individuals self-classified as white interdigitated with United States whites and all other subpopulations from the United States (Fig. 3, Supplemental Fig. 3).

**HIV-1 adaptation to HLA**

We carried out a comprehensive analysis of HLA-associated viral polymorphisms across the full HIV-1 proteome, with adjustment for potential phylogenetic relatedness between viral sequences and LD between coinherited HLA alleles. This analysis identified 874
significant associations of unique amino acid polymorphisms and HLA-A (238, 27%), HLA-B (438, 50%), or HLA-C (198, 23%) alleles ($q$ values < 0.2) (Fig. 4A–C, Supplemental Fig. 4 for all full proteome adaptation maps). In all proteins, HLA-B associations were the most common, except for Vpu, in which HLA-A dominated. The majority of associations involved substitutions between the population consensus amino acid and a codominant di-morphism; however, more complex patterns involving minor amino acids at highly polymorphic positions (Env A319T, entropy 0.88) or HLA alleles driving to more than one alternative adaptation (e.g., HLA-A^*0301 Gag K28R/Q and HLA-B^*0801 Nef K94M/Q/E/N) or away from or one susceptible amino acid at the same position (e.g., HLA-C^*0602 Pol A/N709x) were also apparent. As noted in previous studies, associations frequently overlapped with the same or opposing effects (13–18). In the latter case, adaptation for one HLA allele corresponded to reversion for other HLA alleles and vice versa. Over the whole proteome, there were 202 statistically significant associations at sites ≤4 aa flanking experimentally characterized and published CD8 T cell epitopes with matching HLA restriction (34) and 165 additional associations within or near putative epitopes based on the Epipred epitope-prediction algorithm (35). (The CD8 T cell epitope prediction program used is available at http://atom.research.microsoft.com/bio/epipred.aspx.) For those associations with corresponding known or high-probability putative epitopes, mutation was extraepitopic in 132 cases (36%). In all cases matching known T cell escape mutations as classically described, Epipred indicated a low prediction score for the adapted epitope sequence relative to the nonadapted one. There were other patterns of predicted T cell reactivity in relation to HLA-driven changes, including where immunoreactivity was known or predicted for the adapted epitope. The remaining HLA associations not within or flanking known epitopes may represent viral co-variation associated with a primary HLA-driven adaptation; however, other factors that contribute to this proportion may include the relative underrepresentation of known epitopes in polymorphic regions and restricted by less-studied HLA alleles and the false discovery rate of 20%.

The first described naturally occurring variant p24 epitopes, GGKKKYKF and DCKTILKAL, shown to abolish cytotoxicity of variant-expressing cells corresponded to HLA-B^*0801–associated polymorphisms Gag L31F and K331R, respectively, in this study (36). Similarly, many well-characterized mutations observed to evolve and escape CD8 T cell responses experimentally in direct immunological studies of individuals were reproduced in this study as HLA–HIV-1 polymorphism associations (8–12, 37–47) (Supplemental Fig. 4). There was a partial overlap between the associations detected in this study and the specific associations reported by other studies with other populations (13–18). However, complete concordance is not expected, given significant differences in ethnic/HLA distributions, population size, and viral subtype composition. Formal concordance and stratification analyses accounting for these differences and involving more populations are underway.

Differential constraints to HLA-driven adaptation in the HIV proteome

In 95.7% of identified nonadapted and adapted pairs in HLA-driven polymorphisms, the amino acids were at a minimal single-nucleotide distance from each other, with the remainder having two-nucleotide differences. HLA-associated polymorphisms were dominated by a small group of amino acid pairs with similar physicochemical properties: positively charged (K/R: 17.8%), negatively charged (E/D: 14.5%), nonpolar aliphatic (V/I: 5.3%, L/I: 4.6%, L/V: 3.1%), and bulky aromatic residues (Y/F: 3.8%), suggesting that protein function and structure contingent upon these properties would tend to be preserved. At a protein level, the density of HLA associations differed between proteins in the following (descending) order: Nef > Vpr > Rev > Vif > Tat > Gag > Pol > Vpu > Env and at subprotein level: Nef > Vpr > p17 > Vif > p2p7p16 > Rev > p24 > Gag/Pol TF > Vpu > Integrate > Protease > Env > Tat > RT (Supplemental Fig. 5). When average Shannon entropy scores (48) were plotted against HLA association density per protein, Nef and Env had higher variability (lesser constraint), but levels of HLA selection were dissimilar (Supplemental Fig. 5). We detected sparse HLA class I–associated selection in Env, despite the high degree of general variability. Gag and Pol had less HLA selection than Nef in the face of higher constraint. This may imply that the CD8 T cell responses selecting changes in Gag and Pol epitopes (against constraint), and to a lesser degree, Nef epitope (without constraint), are qualitatively stronger. In
contrast, Env is subject to humoral and other selective forces, and insertion/deletion polymorphisms and other complex variability in Env may reduce the power to detect HLA class I associations. The remaining accessory proteins had higher HLA selection than did Gag and Pol, as expected, but the distribution within this group did not indicate a simple direct correlation between entropy and HLA-associated selection. Notably, Vpu was the most polymorphic protein but had the least HLA-associated selection after Env.

**Cumulative HLA-driven adaptation per individual**

In each study subject, we examined all HIV-1 residues in which an HLA association matched their own HLA class I alleles. The residues at these positions were identified as being adapted, nonadapted, or missing if within a sequence gap and counted as a proportion of all HLA-associated sites relevant to the individual. Although the population-level analysis pointed to differences in HLA-associated density across proteins, it is notable that within individuals, the average proportion of all relevant (individual-specific) HLA-associated sites that were in the adapted state was uniformly high across Vif (66%), Env (64%), Nef (58%), Vpr (57%), Gag (55%), Pol (54%), and Tat (40%) and was lower in Rev (28%) and Vpu (27%).

**Selection profiles of HLA alleles within supertypes and allelic lineages**

HLA-associated adaptation in HIV was mapped in a manner analogous to resistance maps for antiretroviral drugs such that the unique immunoselection profile across the whole HIV-1 proteome is shown for every given HLA allele in the cohort (Fig. 4A–C, Supplemental Fig. 4). Then the convergence or divergence in immunoselection patterns associated with different HLA alleles could be visualized. Because the selection imprints on HIV-1 should partly reflect the binding specificity of HLA peptide-binding regions, we examined HLA alleles within defined supertypes (49): more specifically those related in phylogenetic trees based on exon 2 and 3 sequences (Supplemental Fig. 1). The degree of admixture in the United States cohort meant that such alleles were comparable in population frequency, so that divergence in their immunoselection would not be explained by unequal statistical power. Across all of the main relevant groupings, there was some minor overlap in the epitopes targeted by grouped alleles; however, overall, there was a striking lack of overlap in the selection patterns within those epitopes. For example, HLA-B*5701, B*5703, B*5801, and B*5802 are of the B58 supertype and have marked race/ethnic group specificity (Figs. 1, 2). HLA-B*5701/3 and -B*5801 had clustered adaptations around well-known epitopes in p24 (TW10 and IW9), Integrase (SW10), Rev (RY10), and Nef (HW9); however, there were only two single-codon changes, Gag T242N in TW10 and Pol T840A (or integrase T125A), which was common to all three alleles (Fig. 4A). HLA-B*5802, associated with rapid HIV subtype C disease progression, had only one shared adaptation with other alleles in the HLA-B58 group that are associated with slow HIV-1 disease progression (HLA-B*5701-associated Rev T15x) and no shared adaptations with the most closely related HLA-B*5801 (50). Within the B7 supertype, HLA-B*0702 is present in white and black populations; however, HLA-B*8101 and -B*4201 are rare in whites. As with the B58 group, the selection profiles of these alleles were largely nonoverlapping (Fig. 4B). Across selected allele pairs within the B44 supertype (e.g., HLA-B*4001 and -B*4002, HLA-B*4402 and -B4403) that have extremely conserved exon 2 and 3 sequences (Supplemental Fig. 2) and strong phylogenetic support for close lineage (33), there was little evident overlap in the selection these alleles imposed on HIV (Fig. 4C). Differential immunoselection in the Gag KF11 epitope by HLA-B*5701-restricted T cells in subtype B HIV-1 and HLA-B*5703-restricted T cells in subtype C HIV-1 (51) as well as B7 supertype divergence in relation to shared epitopes in HIV-1 subtype C (52), was shown to be associated with strongly divergent patterns of TCR recruitment and qualitatively different T cell responses. The patterns of HIV-1 subtype B adaptations associated with all HLA-A and -B alleles within all defined HLA supertypes and broad allelic families for all HLA-C alleles relevant to this population (Supplemental Fig. 4) suggest that this is a widespread, general phenomenon. As such, subtle HLA gene variations can make a significant impact on viral diversity at the population level, through diversification of viral peptide sampling and perhaps more extensively through differential TCR usage by the same peptides.

**The influence of HLA on viral sequence divergence**

We then sought to reconcile the observations that, on the one hand, race-specific HLA alleles have very divergent viral-adaptation patterns in terms of single-site associations, whereas on the other hand, there was no clustering of viral sequences by race/ethnic group in the phylogenetic trees (Fig. 3). Given the multiple diversifying forces on viral sequences and underlying constraint, we analyzed the correlations between pairwise viral sequence divergence and HLA divergence (or nonmatching of HLA alleles) in individuals carrying those sequences. When whole protein lengths were considered, HLA-A/B/C matching did not correlate significantly with greater viral sequence similarity (e.g., \( p = 0.06 \) in Gag), consistent with the phylogenetic trees. However, when shorter sliding intervals of 10 aa within viral proteins were considered, the localized correlations within a number of intervals became significant, with \( p \) values based on 500 randomizations of the data (Fig. 5A). Thus, the influence of HLA diversity on HIV-1 diversity is demonstrably localized within epitope-length windows, and the correlation plots effectively visualized a landscape of HLA-driven divergent evolution of HIV-1. In contrast to phylogenetic relationships calculated using full-protein sequences and encompassing all influences on diversity, these profiles represent the CD8 T cell view of HIV-1, not as a whole replicating virus or functional proteins but as an assortment of 8-11mer peptides. Given that our analyses captured primary polymorphisms and covarying adaptation patterns, this also suggests that most covariation is highly localized and proximal to the primary adaptation site in HIV-1. This is supported by explicit population-based analyses of amino acid covariation in HIV-1 (22) and is the case in most experimentally characterized examples of HIV-1 immune escape and compensatory mutation (8, 19–21).

Having established the areas of strongest HLA-driven epitope diversity, we split the correlations by race/ethnic group and by HLA loci (Fig. 5B–D, Supplemental Fig. 6). We chose to focus on the example of Gag for these detailed analyses, given the importance of this protein to immunological control and as a common vaccine Ag (53). In these plots, the only strongly significant HLA-B peak (\( -\log p < 2 \)) shared between white and black subpopulations corresponded to clustered HLA-B*57/5801–associated Gag T242N/T247 and G248, as described in previous analyses. Otherwise, the HLA–HIV-1 correlation landscape was strikingly different among all racial/ethnic groups. The absence of the T242N-associated peak in Hispanics may be related, in part, to the complete absence of the B57/58 group in Amerindiandians. There were unique peaks in the Hispanics corresponding to HLA alleles more common in Hispanics and adaptations detected in phylogeny-corrected analyses (e.g., HLA-B*4002 Gag 219P in Hispanics alone) as in other groups (e.g., HLA-B*0702 Gag in whites alone). Peaks not corresponding to specific HLA associations could also be caused by HLA-lineage correlations in this analysis or by HLA-selection effects not sufficiently strong to be detected with
FIGURE 4. Maps of unique HLA-associated adaptations (q value < 0.2) in the HIV-1 genome for selected grouped HLA alleles present in the cohort within the HLA-B58 supertype (A), B7 supertype (B), and B44 supertype (C). The nonadapted (susceptible/revertant) amino acids are displayed above the
the current level of statistical power. In either case, these plots show the extent to which race-specific HLA alleles could shape viral epitope diversity, and, in turn, the CD8 T cell responses that they elicit in distinct ways in different human populations.

**Discussion**

The global diversity of HLA derives from the unique demographic histories of human ancestral groups spanning 100,000–200,000 y, as well as the selection imposed by the diverse microbial environments into which these groups migrated and survived in isolation (5, 6). In this context, the convergence of African, European, Asian, and Amerindian ancestries into a single United States population, as sampled in this study, brings together a diversity of HLA alleles across multiple allelic lineages and with widely differing evolutionary histories. This type of population sampling, with ascertainment not specifically based on race/ethnicity, provides a more relevant sampling of HLA diversity for HIV-1 vaccine and therapeutics research, and the major subpopulations studied herein are comparable in size to those in the largest, national healthy population study of contemporary HLA genetic diversity in the United States (7). Although the different HLA distributions in this study cohort of HIV-infected individuals predominantly reflect ancestry groups, as seen in the background population, the statistically significant enrichment of HLA-B*5703, which is associated with slower HIV-1 disease progression, is notable and could be evidence of HIV-1 infection as a selective influence on the frequency of particular HLA alleles in a HIV-prevalent population within a contemporary timeframe. Given that those individuals with the extremes of slow disease progression are not likely to be enrolled in this cohort, such effects may only be underestimated in this study. However, it is not clear what underpins the differences in other alleles among Hispanics. It is also possible that all of these differences are caused by some differential enrollment of certain groups into the study cohort or the background population study, although the basis for these effects for these particular alleles is unknown.

By examination of the selection effects of these diverse HLA alleles on HIV-1 sequences, aspects of ancestral HLA evolution are apparent. There is a dominance of the HLA-B in selection, as seen in immunological studies of HIV-1 (50) and adding to accumulating evidence of the strongest pathogen-driven balancing selection operating at the HLA-B locus (5, 6). Because most serological/broad allelic families of HLAs and supertypes contain subtypes that traverse human ancestral groups and there is adequate statistical power to compare these subtypes, in this study, we show that there is striking divergence across most alleles. It is HIV-1 itself that reveals the functional differences between HLA alleles in a way that is independent of, and without specific reference to, their

---

**FIGURE 5.** Plots of significance (−logP) of correlations between nonmatching in HLA genotype and viral sequence dissimilarity over sliding windows of 10 aa in HIV-1 Gag. A, The p values determined by randomizations are plotted for correlation analyses of HLA-A, -B, and -C loci combined in all United States cohort individuals. Correlation significance values are plotted for HLA-B locus nonmatching in particular in white (B, n = 222), black (C, n = 209), and Hispanic (D, n = 107) subpopulations.

---

line in blue text, and adapted amino acids are shown below the line in red text. Blanks indicate where only the nonadapted or the adapted amino acid was evident from the leaf associations in the analysis. Locations of published CD8 T cell epitopes are shown as boxes at association sites.
phylogenetic relatedness or supertype. These results suggest that alleles within a supertype are functionally very different in terms of the HLA-peptide-TCR interaction, even those apparently very closely related phylogenetically and within the same two-digit genotype. The parallel selection effects of HLA-B*5701 and -B*5703, which have been a strong focus in HIV research, seem to be the exception rather than the rule. This is consistent with the markedly different effects on HIV-1 disease progression associated with HLA-B*57 (50) and -B35 subtypes (54) and suggests that the HLA allelic variation across the many other HLA groupings, particularly at HLA-B, have a functional evolutionary basis. Therefore, it is important for clinical applications that HLA alleles of a supertype or family are not assumed to elicit qualitatively equal T cell responses, despite promiscuously bound epitopes.

Just as these analyses provide a view of “immunology taught by viruses,” the counterrevolution of HIV-1 provides an approach to “virology taught by the immune system” (55). The large number of HLA associations across all HIV-1 proteins reveals the enormous scope of subtype B HIV-1 adaptability against a wide span of global HLA diversity. Cumulative per-individual adaptation, as a proportion of sites relevant to the autologous HLA repertoire, is high across most HIV-1 proteins, although Vpu stands out as an exception to the rule in this and other respects having high entropy but low density of HLA associations and dominant HLA-A locus selection (Supplemental Fig. 5). Finally, the analyses of HLA–viral sequence dissimilarity correlations show the extent to which overall HLA repertoire drives viral sequence divergence in a global sense, given the degree of constraint and the other immune and nonimmune influences on viral epitope diversity at any one site. Ultimately, it is diversity within the viral peptide that directly affects applications in CD8 T cell immunotherapy or vaccines. Within these immunologically relevant sequence windows, diversity is influenced by HLA to the extent that the landscape of HLA-driven viral diversity differs among race/ethnic groups (Fig. 5). Such divergence could impact HIV vaccine efficacy across these groups or the extrapolation of efficacy in one group based on viral sequence data from another group, if not informed by knowledge of variances in HLA alleles at high resolution and their divergent impacts on viral sequence. It is important to recognize that this divergence is clear within epitope windows and driven by definable changes at single residues (Fig. 4, Supplemental Fig. 4) but is not visible in standard phylogenetic trees (Fig. 3). Given the functional relevance of these changes, this could be important in understanding the optimal CD8 T cell responses in certain populations in which strong population substructures exist (e.g., Asia or Central and South America). Therefore, the host genetic admixture within the United States provides the unique opportunity to characterize and measure these effects and would be enhanced by studies of specifically targeted populations that are likely to shape circulating viral epitopes in distinct ways (56).

Although studies of larger and combined HLA–viral sequence datasets from different countries and regions will benefit from the general increase in statistical power, this study illustrates the value of broad population-based sampling, inclusion of immunogenetically distinct groupings, individual-level ancestry information, and high-resolution HLA genotyping in characterizing the specific interactions underpinning HLA–HIV coevolution. By mapping these interactions as proteome-wide putative immunoselection profiles of individual HLA alleles (Fig. 4, Supplemental Fig. 4), as well as “hotspots” of HLA-driven epitope divergence (Fig. 5), it is hoped that this information can be more readily used for the significant amount of HIV immunology, virology, and vaccine research involving study subjects within the United States, from where these genetic data were directly drawn, as well as in populations within the broader HIV-1 subtype B epidemic in which many of these HLA alleles are represented.

Acknowledgments

We thank the study teams, study sites, and participants in the U.S. Adult ACTG A5142 and A5128 protocols and the WA HIV Cohort Study and colleagues in the Center for Clinical Immunology and Biomedical Statistics. We also thank Susan Herrmann, Shay Leary, and Anthony Fordham for assistance with manuscript preparation.

Disclosures

The authors have no financial conflicts of interest.

References


