JOURNAL OF VIROLOGY, Feb. 2011, p. 1310–1321 Vol. 85, No. 3
0022-538X/11/$12.00 doi:10.1128/JVI.01966-10

Mapping the Landscape of Host-Pathogen Coevolution: HLA Class I Binding and Its Relationship with Evolutionary Conservation in Human and Viral Proteins▽†‡

Tomer Hertz,[1]§ David Nolan,[2] Ian James,[2] Mina John,[2] Silvana Gaudieri,[2,3] Elizabeth Phillips,[2] Jim C. Huang,[1] Gonzalo Riadi,[1,4] Simon Mallal,[2]* and Nebojsa Jojic[1]*

Microsoft Research, One Microsoft Way, Redmond, Washington 98052[1]; Institute for Immunology and Infectious Diseases, Royal Perth Hospital and Murdoch University, Murdoch 6150, Western Australia, Australia[2]; School of Anatomy and Human Biology, Centre for Forensic Science, University of Western Australia, Australia[3]; and Fundación Ciencia para la Vida, Avenida Zañartu 1482, Ñuñoa, Santiago, Chile[4]

Received 16 September 2010/Accepted 9 November 2010

The high diversity of HLA binding preferences has been driven by the sequence diversity of short segments of relevant pathogenic proteins presented by HLA molecules to the immune system. To identify possible commonalities in HLA binding preferences, we quantify these using a novel measure termed "targeting efficiency," which captures the correlation between HLA-peptide binding affinities and the conservation of the targeted proteomic regions. Analysis of targeting efficiencies for 95 HLA class I alleles over thousands of human proteins and 52 human viruses indicates that HLA molecules preferentially target conserved regions in these proteomes, although the arboviral *Flaviviridae* are a notable exception where nonconserved regions are preferentially targeted by most alleles. HLA-A alleles and several HLA-B alleles that have maintained close sequence identity with chimpanzee homologues target conserved human proteins and DNA viruses such as *Herpesviridae* and *Adenoviridae* most efficiently, while all HLA-B alleles studied efficiently target RNA viruses. These patterns of host and pathogen specialization are both consistent with coevolutionary selection and functionally relevant in specific cases; for example, preferential HLA targeting of conserved proteomic regions is associated with improved outcomes in HIV infection and with protection against dengue hemorrhagic fever. Efficiency analysis provides a novel perspective on the coevolutionary relationship between HLA class I molecular diversity, self-derived peptides that shape T-cell immunity through ontogeny, and the broad range of viruses that subsequently engage with the adaptive immune response.

Human leukocyte antigen (HLA) molecules and viruses are thought to be locked in an evolutionary arms race, where viruses adapt to evade HLA-restricted immune responses and HLA alleles evolve to optimize the fitness of human populations in the face of a wide range of pathogen species as well as the genetic variation within each pathogenic species. HLA diversity has been driven and maintained by heterozygote advantage (25), which is most evident in geographical regions with greater pathogen diversity (51), and by frequency-dependent selection, in which low-frequency allelic variants gain advantage in an environment of shifting pathogen selection (58). In turn, the selective pressures of HLA-restricted immune responses on pathogens are evident in a range of immune evasion strategies employed by viruses and encoded in their genomes, such as the ability of large DNA viruses (e.g., herpesviruses) to "hide" by inhibiting antigen presentation (61) and mimicking host peptides (39, 60) or the ability of RNA viruses to "run" through rapid evolution of genetic diversity (22, 35, 40, 43, 52, 53).

We along with others have explored the rapid viral adaptation to HLA-restricted immune responses using sequence analyses and have detected statistically significant associations between host HLA alleles and specific amino acid polymorphisms of human immunodeficiency virus (HIV) and hepatitis C virus (HCV) (4, 8, 28, 29, 43, 62). These findings have informed and directed experimentation which has, for example, confirmed that some of these HLA allele-specific viral polymorphisms are due to abrogation of HLA binding or peptide processing (17, 28, 29, 34, 62). In contrast, there is a paucity of direct evidence linking HLA evolution to the selective pressure of pathogens as the reproductive advantage for humans operates on a long timescale (5). Limited direct evidence from a set of 34 oncoproteins and HIV Nef suggests that HLA alleles might preferentially target evolutionarily conserved peptides (12, 23). As functionally important sites on proteins tend to be evolutionarily conserved (12, 26, 64), immune surveillance of conserved ligands focuses immune resources to genomic areas in humans and pathogens where mutations might alter function (26, 57) or incur a fitness cost (28, 29, 66).

The recent availability of large curated databases of genetic

* Corresponding author. Mailing address for S. Mallal: Institute for Immunology and Infectious Diseases, Royal Perth Hospital and Murdoch University, Wellington Street, Murdoch 6150, Western Australia, Australia. Phone: 61 8 9224 2899. Fax: 61 8 9224 2920. E-mail: S.Mallal@iiid.com.au. Mailing address for N. Jojic: Microsoft Research, One Microsoft Way, Redmond, WA 98052. Phone: (425) 497-8401. Fax: (425) 936-7329. E-mail: jojic@microsoft.com.

§ Present address: Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109.

† Supplemental material for this article may be found at http://jvi.asm.org/.

▽ Published ahead of print on 17 November 2010.

‡ The authors have paid a fee to allow immediate free access to this article.

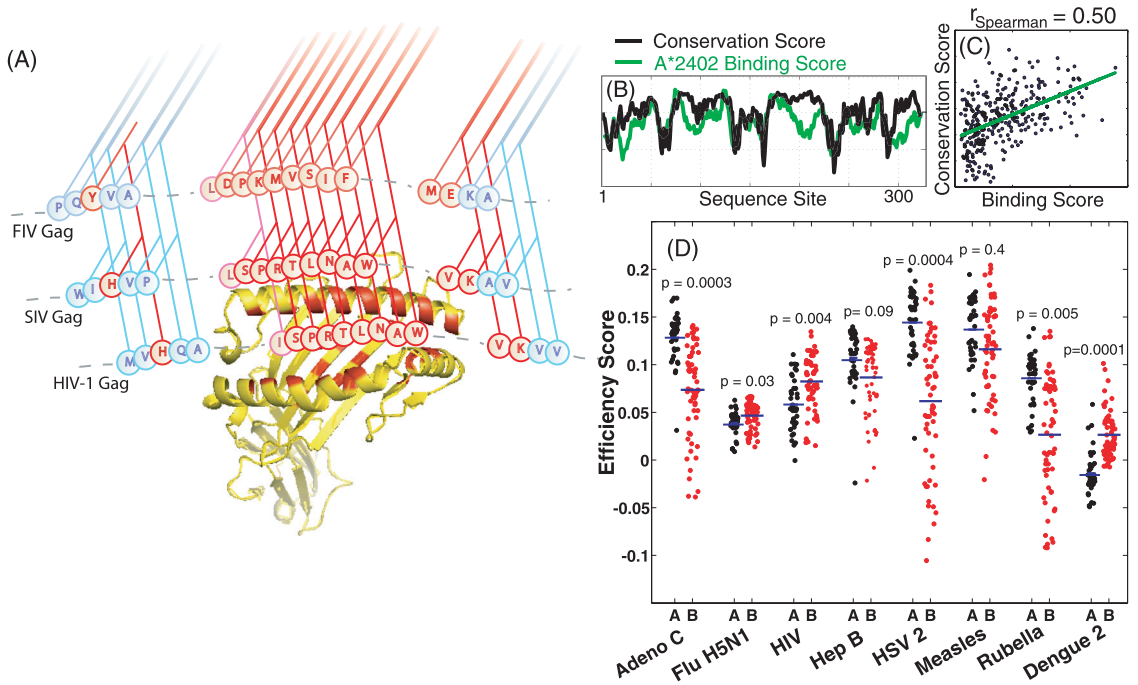

FIG. 1. Computing allele efficiency scores. (A) A representative illustration of an MHC molecule (yellow) binding to a segment of the HIV-1 Gag protein, showing comparisons with the phylogenetically related simian immunodeficiency virus (SIV) Gag and feline immunodeficiency virus (FIV) Gag proteins. While the topology of the phylogenetic tree is shared among the protein sites, the evolutionary rates, indicated by the variation of color of the branches from red (most conserved) to blue (least conserved), may vary dramatically. (B and C) The allele targeting efficiency score (r) for a given protein (herpesvirus-1 capsid triplex subunit 1 in this case) is defined by the rank correlation coefficient between site conservation scores (evolutionary rate) and HLA site binding scores (an average binding energy for peptides containing the site), along the protein (HLA-A*2402 here), assessed at each amino acid position. (D) Distributions of HLA-A and HLA-B locus efficiency scores for a range of human viruses. Each point represents an HLA allele-specific efficiency for the relevant full-length viral proteome. Blue bars represent locus means. P values indicate significance of locus differences tested using mixed effects analysis (Fig. 5 and Table 1). Adeno, adenovirus; Flu, influenza virus; Hep B, hepatitis B virus; HSV 2, herpes simplex virus type 1.

sequences has aided in the investigation of evolutionary relationships between human and pathogen genetic diversity. These databases enable studies of evolutionary conservation using sequence variation (2). In addition, the experimental determination of tens of thousands of HLA binding affinity measurements (48) has allowed robust estimation of binding affinities for a wide range of HLA-peptide combinations (38). These data allow direct investigation into the relationship between HLA binding and target sequence conservation, as well as into the differences in these patterns across viral species and different HLA alleles.

Here, we examined the relationships between HLA class I molecules and a large selection of pathogen-derived and self-derived HLA peptide ligands, comparing the likelihood of a given HLA molecule to bind a given peptide, and the relative conservation of that peptide sequence (Fig. 1A). We term the tendency of a given HLA molecule to bind to conserved regions of a protein its "targeting efficiency." Using this approach, we explored the variability of HLA alleles in their ability to target conserved regions of human and viral proteins. Finally, we explored the functional relevance of HLA targeting efficiency and found that targeting efficiency is associated with improved disease outcome for HIV infection and dengue virus and accounts in part for interindividual variation in HIV viral load in predictive models.

These findings suggest that the relationships between HLA binding preference and evolutionary conservation of target sequences provide a central basis around which balancing selection of both host and pathogen genetic diversity may be better understood, as first proposed by Hughes and others (23, 64). Our interspecies approach is complementary to previous intraspecies studies of HLA-allele specific viral polymorphisms (27, 60, 63), which have more statistical power in the variable than the conserved elements of pathogen genomes, and provides a novel tool with which HLA pathogen coevolution can be examined.

MATERIALS AND METHODS

Conservation scores. Conservation scores for an analyzed protein were computed using the ConSeq server (http://conseq.tau.ac.il/), which estimates the conservation score $C(i)$ for each protein site i using a phylogenetic tree built from a set of homologous sequences. The tree is used to infer the evolutionary rates (log probabilities of substitution) for each site along the given protein (2) (Fig. 1A). Conservation scores were computed only for proteins which had at least five homologous proteins in the UniRef100 database (release 11.0). For computational efficiency, no more than 100 homologues were used. Homologous proteins were aligned using the MUSCLE program, version 3.6. The site conservation scores are computed in the context of the entire protein as ConSeq uses aligned homologues to compute the phylogeny in which the targeted protein resides. This analysis also incorporated adjustment for variables that may affect P values, including various numbers of protein homologues in conservation score computations, as well as various protein length distributions. Comparisons of the effects of using inter- versus intraspecies homologous proteins for estimating the evolutionary rates are provided for HIV in Table S6 in the supplemental material.

Binding scores. The binding scores are based on experimental measurements characterizing individual HLA-peptide interactions, as catalogued in the Immune Epitope Database (IEDB) (48, 50), as well as known HLA-peptide binding configurations (32). Binding energies of HLA-peptide complexes were systematically estimated using the adaptive double-threading (ADT) structure-based approach (32) for estimating the binding energy of a major histocompatibility complex (MHC)-peptide complex. The method estimates the 50% inhibitory concentrations ([IC_{50}s] a measure of the binding affinity) after threading both target peptides and HLA proteins (in particular, the known contact residues shown in red on the HLA structure in Fig. 1A) onto solved HLA-peptide complex structures. Here, we have focused on 9-mer peptide targets as the vast majority of known HLA class I epitopes are of this length.

The model parameters were fit to log IC_{50}s obtained from the IEDB for ~34,000 experiments covering 35 HLA-A and HLA-B alleles. We excluded all HIV epitopes from these data for training in order to avoid a possible bias in the analyses of HIV viral load data. The ADT model can provide estimates for HLA molecules other than the limited number on which it was trained by threading the arbitrary HLA sequence onto the structure of another similar HLA protein and using the estimated model parameters, which generalize for the entire HLA allelic family. We analyzed 95 HLA-A and HLA-B alleles from a Caucasian population in Australia (43), a cohort which provided a total coverage of 95% of HLA-A alleles and 90% of HLA-B alleles in Europe. We note that of these 95 alleles, empirical binding data were provided for only 35 alleles. Due to lack of sufficient experimental measurements for HLA-C alleles in the IEDB, we did not consider these alleles in the current study. The binding score at a given position along the protein is the sum of binding energies for nine overlapping peptides, which measures the probability that the site will be visible to immune surveillance. The binding energy model provides an energy estimate, $E_a(e) = E_a(e_1,e_2,\ldots,e_9)$, where a denotes the index of an HLA allele, and e is $(e_1,e_2,\ldots,e_9)$ the 9-mer peptide. The model is fit to the logarithm of the IC_{50} measurements for different allele-peptide combinations. Therefore, the probability of peptide presentation is proportional to $e^{-Ea(e)}$, and high energy indicates low presentation probability and vice versa. In order to estimate the log probability of presentation of a single site in a protein, presentation probabilities of all peptides straddling that site need to be considered. As an estimate that is robust to prediction errors, we define the binding score $B(i)$ for the i-th amino acid in the sequence $s = (s_1,s_2,\ldots s_N)$ of an arbitrary N-long protein whose segments may be presented by an HLA molecule as follows: $B_a(i) = -\sum_{j=i-8}^{i} E_a(s_j, s_{j+1}, \ldots, s_{j+8})$. The predicted binding energies are highly correlated with the true experimentally measured binding energies (Spearman correlation of >0.75), and for some alleles the accuracy of prediction for our method and for other prediction methods (see, for example, references 3, 7, 16, 31, 44, 45, and 49) is believed to be within the accuracy of the IC_{50} measurement error. A recent analysis of various prediction methods can be found in Nielsen et al. (46).

Computing allele efficiency scores. The allele efficiency score r is defined as the Spearman correlation coefficient of the binding score and the conservation score for a given protein. A positive score indicates preferential targeting of conserved regions, and a negative score indicates preferential targeting of variable regions.

Statistical analyses. Statistical comparisons were corrected for the potential lack of independence of measurements resulting from the hierarchical structure of the groups, such as correlations generated by similarities within viral families and within HLA supertype classes. The sampling procedure for these analyses, along with a detailed analysis of potential sources of bias in our analysis and a detailed discussion of the statistical power of the methods used, is available in the supplemental material.

Data sets. We obtained binding data for training the HLA binding predictor from the IEDB resource, consisting of experimentally measured binding affinities (IC_{50}s) for ~34,000 HLA-peptide pairs, spanning 35 HLA-A and HLA-B alleles. For the viral targeting analysis, we collected the sequences of the most commonly studied human and plant viruses from the NCBI virus database. A full list is provided in Table S2 in the supplemental material. NCBI GenBank accession number information for individual viral proteins is available in Table S8 in the supplemental material. We based our analysis of the human proteins with disease-associated mutations on a version of the Online Mendelian Inheritance in Man (OMIM) database (McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD, and NCBI, National Library of Medicine, Bethesda, MD, as of February 2007 [http://www.ncbi.nlm.nih.gov/omim]). A control group of randomly sampled proteins was obtained from the UniRef50 database, release 11.0 (http://www.pir.uniprot.org/search/textSearch_NR5.shtml), which contains representatives of natural protein clusters with >50% interclass similarity. The random sample did not include any human proteins. BLAST queries were performed against the UniRef100 database.

HLA typing. A total of 200 Western Australian HIV Cohort participants provided samples with consent for HLA typing. HLA class I genotyping was resolved to the four-digit level based on the sequence of exon 2 and 3 (exon 2/3) and using standard sequence-based typing (SBT). Allele and heterozygote ambiguities between alleles not identical in exon 2/3 were resolved using alternative primers.

Comparison of efficiency scores in human proteins and human and plant viruses. In order to test the hypothesis that there is a correlation between conservation and binding patterns in proteins known to interact with HLA molecules, we examined the relationship between the efficiency scores of 95 HLA-A and HLA-B alleles (which account for more than 90% of all HLA alleles in European Caucasians) for target proteins spanning 4,761 human proteins derived from the OMIM database and 52 human viruses (see Table S2A in the supplemental material). We also examined these same correlations in 70 plant viruses (see Table S2B), as well as a control sample of 3,800 random proteins obtained from the UniRef50 database expecting that we would observe significant, but weaker, efficiency scores because of the sharing of conservation patterns between the proteins that typically do not interact with the immune system and the human proteins and human viruses that do. These control groups were also used to ascertain the significance of targeting efficiencies and variations in targeting efficiencies on human proteins and human viruses.

Additional analysis to detect potential bias. In order to verify that our conclusions were not dependent on the use of a specific predictor, we also used the NetMHCpan online prediction method (3) to replicate our results on Gag efficiency scores and HIV viral load correlations. Negative correlations with viral load were also obtained using this method (data not shown). In addition, we also conducted an analysis of HIV efficiency scores based on experimentally determined T-cell epitope maps and compared this analysis to the one obtained using predicted epitope maps. We found more evidence for HLA targeting of conserved regions when we used the measured epitope maps than when using the predicted ones, indicating that our analysis may underestimate the extent of this phenomenon (see supplemental material and Table S7).

RESULTS

Targeting efficiency. We introduce a novel score termed "targeting efficiency" to quantify the relationship between HLA binding and conservation of target peptides. The HLA allele targeting efficiency score is defined to be the Spearman rank correlation coefficient between binding scores and conservation scores for amino acids along a given protein. Positive scores denote a preference for binding conserved regions while negative scores indicated a preference to bind to variable regions. The process of calculating a targeting efficiency score for the example pair of the herpesvirus-1 capsid triplex subunit 1 protein and the HLA-A*2402 molecule is illustrated in Fig. 1B and C. In addition to calculating targeting efficiency scores for individual proteins, the approach also allowed an efficiency score to be calculated for an entire proteome by concatenating all protein sequences of the given pathogen for which data were available.

HLA molecules preferentially target conserved areas of human proteins. In order to examine whether HLA molecules preferentially target conserved targets on human proteins, we analyzed a set of 4,761 proteins spanning the entire OMIM database of human-disease associated proteins (http://www.ncbi.nlm.nih.gov/omim/).This analysis showed that both HLA-A and HLA-B molecules tend to preferentially bind to conserved areas of human proteins (Fig. 2), as first proposed by Hughes and Hughes in 1995 in an analysis of 34 oncoproteins (23). These results are also in agreement with those reported by Yeager et al. (64). The average targeting efficiencies were significantly higher for both HLA-A ($P < 10^{-76}$) and HLA-B alleles ($P < 10^{-8}$) for OMIM proteins than for a random sample of 3,800 proteins (UniRef50 database) (Fig. 3). Similar targeting preferences were obtained from a smaller

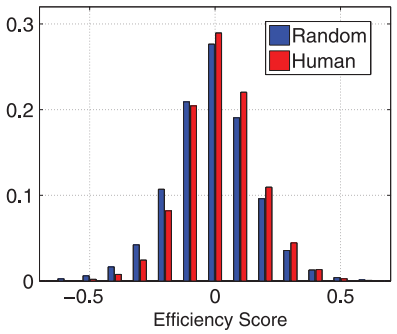

FIG. 2. Comparative histograms of human and random protein efficiency scores. These histograms demonstrate preferential targeting of evolutionarily conserved self-peptides by class I HLA molecules. The distribution of human efficiency scores is statistically significantly higher than that of random proteins ($P = 5.4 \times 10^{-77}$).

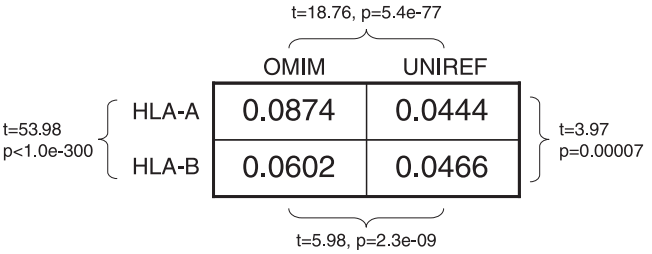

FIG. 3. Mean efficiency scores of 4,761 human (OMIM) proteins and 3,800 random (UniRef50) proteins. The *P* values are based on mixed effects analysis (see the supplemental material).

comparison using 300 randomly selected non-disease-related human proteins and 300 randomly selected UniRef proteins (see Table S3 in the supplemental material), verifying that this property is not specific to disease-associated human proteins. Importantly, the reported differences in efficiency distributions are not a consequence of the differences in the raw scores of the two parameters used to compute them as neither the binding scores ($P = 0.96$) nor the conservation scores ($P = 0.99$) were distributed differently in human and UniRef random protein sets. Additionally, significantly higher binding scores were observed at disease-associated mutation sites ($P < 10^{-10}$ over 2,825,558 sites), providing further evidence of HLA coevolution with its targets.

We were also interested in exploring the relationship between efficiency scores and the assignment of HLA alleles to loci and to supertype groups (described in reference 29; see also Table S1 in the supplemental material). Therefore, we examined the efficiency scores of HLA-A versus HLA-B alleles for both the human and random protein sets described above. We found a marked preference for HLA-A to target conserved regions of human proteins (HLA-A $\gg$ HLA-B, $P < 10^{-300}$) compared with the random UniRef50 sample (HLA-B $>$ HLA-A, $P = 0.00007$). However, further analysis indicated that while the distribution of efficiency scores appears more uniformly positive for the HLA-A allele families (Fig. 4), a limited number of HLA-B alleles also preferentially bind conserved human protein sequences. For example, the B58 supertype (e.g., HLA-B57 and -B58 alleles) and a subgroup of HLA-B7 supertype alleles (HLA-B55 and -B56) have higher efficiency scores than most other B alleles for human proteins (Fig. 4). Interestingly, all of these alleles form a distinct HLA cluster closely associated with chimpanzee Patr-B alleles (see Fig. S2 in the supplemental material). Conversely, HLA-B alleles from the B44, B27, and B7 supertype families (other than B55 and B56) have low or negative average allele efficiencies for human proteins (Fig. 4). These results indicate that grouping HLA alleles by their abilities to target conserved regions of human proteins does not strictly follow their supertype classification, but it does reflect their phylogenetic history as alleles that are more likely to target conserved areas of human proteins cluster with chimpanzee Patr-B alleles (41).

Evolutionary conservation and HLA-viral interactions. To determine if binding preferences for viruses follow a trend similar to the one observed for human proteins, we computed the HLA efficiency scores for 52 human viruses and 70 plant viruses (listed in Tables S2A and S2B in the supplemental material). Figure 5 (see also Fig. S3) shows the distribution of efficiency scores as heat maps in which the viruses (*x* axis) are grouped by their families, and the HLA allelic variants (*y* axis) are organized into HLA supertype families characterized by their HLA peptide binding preferences (56). Figure 1D also presents the distribution of efficiency scores for HLA-A and HLA-B alleles for a selection of human viruses. Importantly, neither the binding scores ($P = 0.87$) nor the conservation scores ($P = 0.59$) were distributed differently in the two viral groups (human- and plant-infecting), consistent with previous reports (30). However, differences emerge when allele targeting efficiency is considered, suggesting that the HLA system has been optimized through coevolution with viruses to recognize functionally important protein regions that are relevant to pathogen threats. Most interestingly, these groupings revealed patterns of efficiency variation over different HLA alleles and different viral families (Fig. 5), indicating a possible functional importance of the existence of long-lasting allelic lineages for the HLA-A but not HLA-B allele locus (24, 41) and possible evidence of specialization of the HLA loci. To assess this further, we investigated the distribution of efficiency scores for HLA-A and HLA-B loci according to viral genome composition and viral species, as described in the supplemental material (22, 28, 29, 62).

HLA-A alleles preferentially target conserved regions of DNA viruses. In our analysis of DNA viruses, we found that both HLA-A and HLA-B loci demonstrated a preference for targeting evolutionarily conserved regions of human DNA viruses relative to plant DNA virus species ($P < 0.01$) (Fig. 6). Moreover, there was also a striking general preference for HLA-A alleles to target conserved regions of human DNA viruses compared with HLA-B alleles ($P < 10^{-13}$). This was most notable for the *Herpesviridae* and *Adenoviridae* (Table 1), with significant differences between HLA-A and HLA-B loci for 8 of 10 viral proteomes assessed. Since these viruses have emerged via distinct lineages through vertebrate evolution (13, 42), these observations are not readily explained by sequence similarity between these viral families. Yet the results are consistent with coevolutionary relationships between herpesviruses and adenoviruses and their human and ancestral primate hosts (13). Moreover, we found a linear correlation between

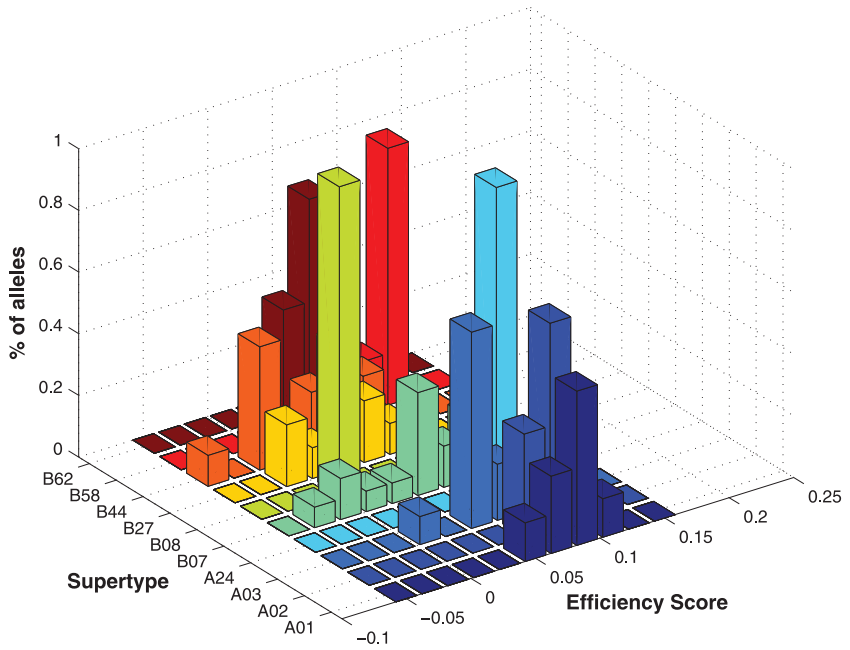

FIG. 4. Allele efficiency scores for OMIM human proteins by HLA supertype groups. HLA efficiency scores of 95 HLA alleles grouped by supertypes are shown for the a set of 4,761 human proteins that form the OMIM database. As can be seen, HLA-A alleles have higher efficiency scores than HLA-B alleles, with the exception of the B58 supertype.

HLA allele efficiencies for the analyzed human proteins and these DNA viruses (as shown for cytomegalovirus in Fig. 7A), as well as strong correlations among overall efficiency scores for the proteomes of DNA viruses (Fig. 7C). These observations are consistent with previous evidence that these viruses exploit host peptide mimicry as a means of host immune evasion (39, 60). We suggest that herpesviruses, which establish persistent but nonprogressive infection in the vast majority of human hosts, may evade the immune system (18) by exploiting the "holes" in the T-cell repertoire that are created by negative thymic selection. The similarity of herpesvirus and adenovirus proteome sequences to self-peptides would also be anticipated to induce more specific, less cross-reactive T-cell responses (10, 18), which may contribute to the progressive inflation of herpesvirus-specific T cells with a restricted T-cell receptor (TCR) repertoire during human ageing (33).

HLA-B alleles target RNA viruses more efficiently. In contrast to results obtained for human protein and DNA viral protein targeting by HLA class I, we found that HLA-B alleles had higher efficiency scores for RNA viruses (Fig. 6B) ($P < 10^{-16}$), with preferential targeting of evolutionarily conserved viral proteins by HLA-B noted for 23/34 (68%) of RNA viruses assessed (Table 1). Further scrutiny revealed a spectrum of targeting efficiency profiles across the range of HLA-B-virus pairs. At one end of the spectrum we found a general trend toward positive HLA-B efficiency scores for the human-adapted *Paramyxoviridae* (including respiratory viruses such as respiratory syncytial virus, parainfluenza virus, and metapneumovirus, as well as measles and mumps viruses) and the *Picornaviridae* (predominantly rhinovirus and enterovirus species). These RNA viruses exhibit a diverse range of "high-efficiency" HLA-B specificities (Fig. 5), consistent with the existence of a host-pathogen evolutionary relationship that is relatively specific for the HLA-B locus. Preferential and efficient targeting of these highly infectious, but nonpersistent, RNA viruses by the highly polymorphic HLA-B locus could also provide a potential mechanism for the observed rapid evolution of these viral species, characterized by the emergence of transient viral mutations (which may provide an HLA context-specific selection advantage) that are then purged by purifying selection (22, 52).

At the other end of this spectrum, the arboviral *Flaviviridae* are a dramatic counter-example to the general observation that HLA molecules preferentially target evolutionarily conserved proteomic regions. As shown in Fig. 5, most HLA-A and HLA-B molecules preferentially targeted nonconserved protein sequences from arboviral flaviviruses.

Comparing HLA efficiency profiles of pathogens provides additional evidence for host pathogen coevolution. Targeting efficiency scores for two different protein groups are often correlated across different HLA alleles. For example, we found a strong negative correlation between allele efficiency scores for flaviviruses and the efficiency scores for human proteins or double-stranded DNA (dsDNA) viruses (Fig. 7A and C). The finding that HLA-A and HLA-B alleles that target conserved regions of human (self) proteins tend to target nonconserved regions of the dengue virus suggests that these viruses seek to reduce similarity to self-peptides to the extent that they are ignored by the immune system (18), thereby exploiting positive T-cell selection to their advantage. This evolutionary strategy also focuses HLA binding on relatively nonconserved viral proteome regions that are likely to have functional and genomic plasticity. We also find correlations between unrelated viruses (Fig. 7C), suggesting that these pathogens share common strategies to evade HLA-restricted immune surveillance. In contrast, negative correlations between unrelated vi-

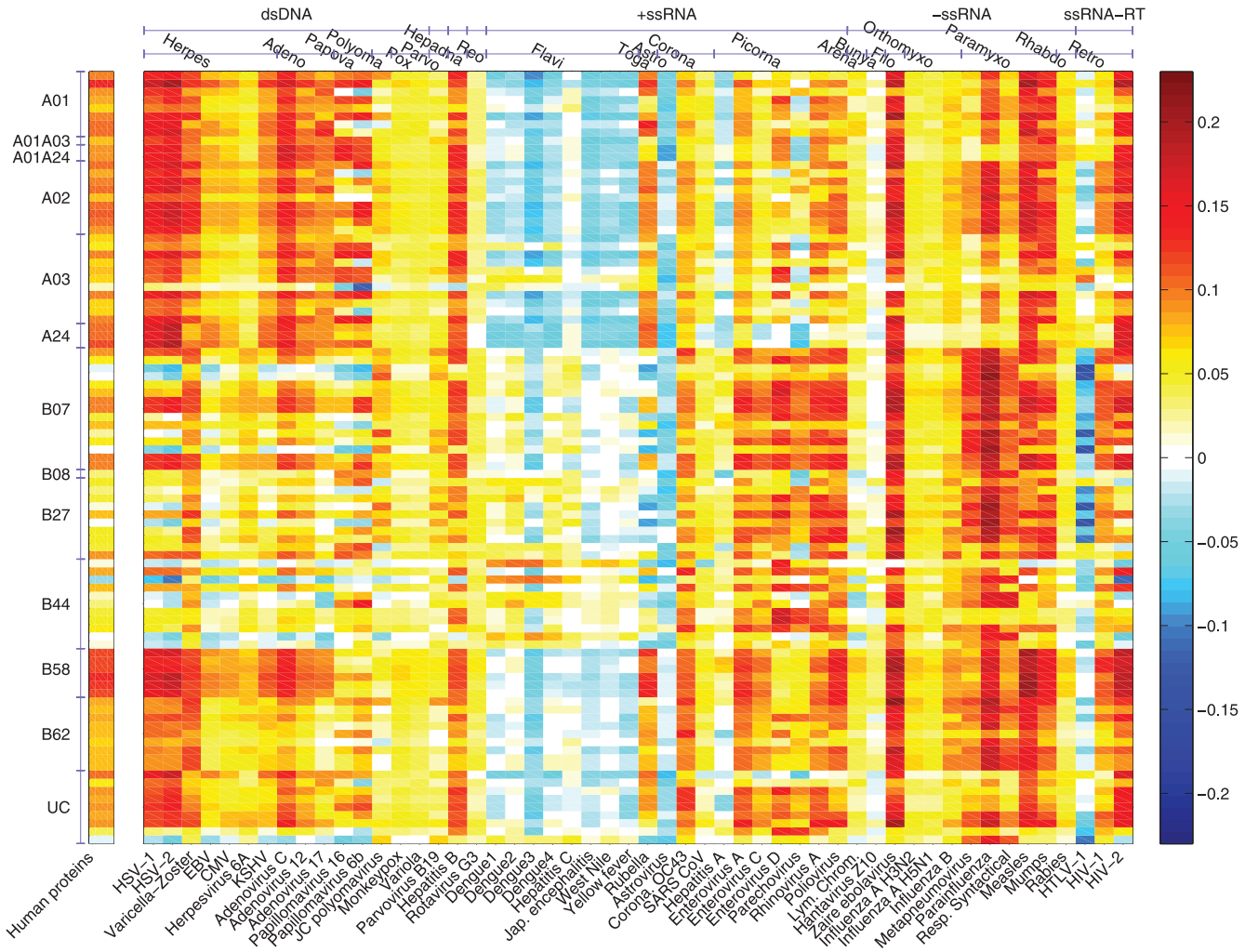

FIG. 5. Heat map distribution of allele efficiencies for human viruses and human proteins (*x* axis) by HLA supertype families (*y* axis). A matrix of efficiency scores computed for each of the 95 HLA alleles studied for 52 human viruses and a set of human proteins. Each entry in this efficiency matrix represents the efficiency score of a specific HLA allele (*y* axis) for a specific viral proteome. HLA alleles were grouped by supertypes, and human viruses were grouped by viral families and by Baltimore classification. Average efficiency scores over a large set of human proteins are presented in the bar to the left of the matrix. Distinct patterns of targeting efficiency can be observed for both HLA alleles (grouped by supertype or loci) and for different viral groups and families. UC, unclassified alleles that have not been assigned to supertypes; HSV-1, herpes simplex virus type 1; EBV, Epstein-Barr virus; CMV, cytomegalovirus; KSHV, Kaposi's sarcoma-associated herpesvirus; SARs-CoV, severe acute respiratory syndrome coronavirus; HTLV-1, human T-cell leukemia virus type 1; ssRNA, single-stranded RNA; RT, reverse transcriptase.

ruses may indicate that these pathogens exploit the weaknesses in the immune surveillance created by overadaptation of the immune system to one virus or the other.

We found that these correlations cannot be explained by the sequence (dis)similarities among viruses: when the same correlation factors are computed for simulated HLA binding preferences, they are not as strong as they are for the 95 HLA alleles analyzed here, which cover true binding properties for over 90% of the European population. This again suggests a prominent role for HLA targeting efficiency in shaping HLA and pathogen coevolution (Fig. 7B; see also the supplemental material).

HLA-disease associations and targeting efficiency. The findings presented thus far point to coadaptive relationships between HLA allelic diversity and human viruses but do not address the functional relevance of these observations. It is apparent from these analyses that relationships between virus species and host HLA diversity are highly specific, indicating specialized roles for HLA loci and for HLA allelic variants within these loci. We thus considered whether HLA targeting efficiencies for a given pathogen (with subsequent possible consequences for HLA-restricted immune responses) are associated with altered infectious disease outcomes. We focused on the specific examples of HIV disease progression, HIV viral load, and the incidence of dengue hemorrhagic fever (DHF).

HLA targeting efficiency and susceptibility to dengue hemorrhagic fever. Dengue hemorrhagic fever is a severe clinical manifestation of a secondary dengue flavivirus infection. Previous studies have suggested that the pathogenesis of this syndrome involves cross-reactive T-cell responses (15), which may be enriched in the context of low-affinity interactions between HLA class I and conserved viral peptides (as described for

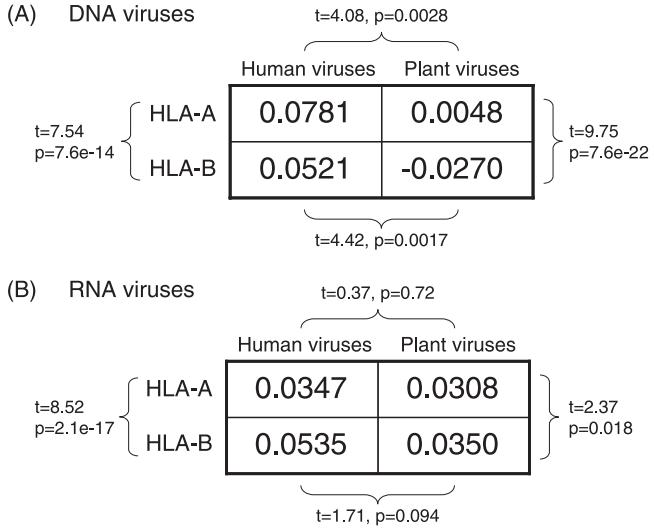

FIG. 6. Comparison of efficiency scores of human and plant DNA (A) viruses and human and plant RNA viruses (B) utilizing mixed effects analysis. The *P* values are based on mixed effects analysis (see supplemental material).

dengue viruses in Fig. 5; see also Fig. S7 in the supplemental material) (10, 18). In support of this model, we found that HLA alleles known to be associated with susceptibility to dengue hemorrhagic fever (11) appeared to be less likely to target evolutionarily conserved proteomic regions than alleles that confer resistance to this disease ($P = 0.05$) (Fig. 8).

Using protein-specific HLA targeting efficiency for predicting HIV-1 viral load and HIV disease progression. We next analyzed the effect of HLA targeting efficiency for specific HIV-1 proteins on HIV-1 viral load in a study population of 191 HIV clade B-infected, treatment-naive individuals from the Western Australian HIV cohort for whom viral loads (plasma HIV RNA level) and full HLA typing were available (43). We computed HLA targeting efficiency for each individual by averaging binding scores over each patient's specific HLA-A and HLA-B repertoire, thus approximating the aggregate ability of patient-specific HLA alleles to differentiate between conserved and variable targets.

While the above analysis of viral targeting involved full proteome targeting efficiencies (Fig. 5), the underlying analysis of residue-specific binding and conservation scores calculated for individual viral proteins and for specific HLA alleles allows for a more focused exploration of the landscape of HLA-pathogen interactions based on targeting of particular individual proteins (see, for example, Fig. S4 to S6 in the supplemental material). Analysis of protein targeting efficiencies has the potential for broader use in the investigation of viral immunity relevant to natural infection as well as vaccine design and evaluation, particularly when indicators of functional immunity are available, as is the case in HIV-1 infection.

While overall HLA allele efficiencies toward HIV proteins correlated negatively (but not significantly) with log viral load ($P = 0.24$), certain individual protein targeting efficiencies showed significant correlations with viral load. For example, HLA-B locus efficiency in targeting Gag protein alone is more strongly correlated with viral load ($r = -0.19$; $P = 0.009$), consistent with experimental evidence that HLA-B-restricted cytotoxic T lymphocyte (CTL) responses to Gag epitopes play a significant role in determining the natural history of HIV infection (6, 35, 36, 54). We then analyzed the distribution of HLA allele efficiency scores for HIV-1 Gag according to their known associations with HIV progression (9, 19) and found that protective HLA alleles tend to rank more highly in targeting efficiency of conserved Gag proteomic regions (Fig. 9).

We also investigated the combined effect of efficiency scores for individual proteins and HLA loci on log viral load using multivariate regression. This analysis used the efficiency scores of all nine HIV proteins, taking into account proteasomal cleavage, and was performed on multiple test/train splits of the data (see supplemental material). Here, we found that HIV efficiencies alone account for 7.0% of the log viral load variance ($P = 3.67 \times 10^{-4}$; correlation, 0.27), gender alone explains 7.0%, and the patient's ethnic group alone shows no significant explanatory power. The HLA efficiencies and gender combined explain 11.2% of log viral load variance. The explanatory power of efficiency scores was not attributable to a single HLA allele or HLA supertype effect nor to confounding demographic effects such as gender and race (described in the supplemental material).

Incorporating proteasomal cleavage, HLA binding, and evolutionary conservation: presentation efficiency. HLA binding is not the sole determinant of potential immune targets. The processing of intracellular antigens relies on relatively monomorphic and evolutionarily conserved proteins to optimize peptide cleavage prior to HLA binding (47), thus providing an additional mechanism for ligand selection. We therefore examined the potential influence of proteasomal cleavage on targeting efficiency by using the NetChop algorithm (46, 55). We analyzed targeting efficiency across human proteins and viral species using a restricted data set of peptides with a high probability of appropriate C terminus cleavage. We found that proteasomal cleavage restriction has indeed coevolved with HLA binding, and cleavage is also directed toward conserved targets. As shown in Fig. S7 in the supplemental material, the distribution of HLA targeting efficiencies remained similar to those identified in Fig. 5, indicating that relationships between HLA binding preference and evolutionary conservation are preserved among peptide targets that are selected via the antigen-processing complex.

DISCUSSION

In this study, we have found that HLA class I molecules preferentially sample conserved regions of human proteins and many viral families, as initially hypothesized by Hughes and Hughes (23). We uncovered a striking exception in the arboviral *Flaviviridae* species, where HLA molecules preferentially target nonconserved regions. This methodology provides a capacity to map the landscape of host-virus interactions from a novel perspective and also allows for closer examination of these effects at the viral protein level (see Fig. S4 to S6 in the supplemental material), providing a platform for comparative analyses of the complex coevolutionary relationships that exist between viruses and their human hosts.

These findings also provide evidence for the evolution of HLA class I locus and allelic specialization, suggesting a partial

TABLE 1. Comparison of HLA-A and HLA-B allele efficiencies for 52 viruses[a]

Virus Group	Virus	Mean A-B Efficiency	SE	P-Value
dsDNA	Herpes simplex type 1	0.0638	0.0181	0.001
	Herpes simplex type 2	0.0802	0.0208	0.0004
	Varicella-Zoster	0.0014	0.0096	0.8868
	CMV	0.0447	0.0118	0.0005
	Human herpesvirus 6A	0.0253	0.0095	0.0107
	KSHV	-0.0116	0.0068	0.0938
	EBV	0.0441	0.0121	0.0007
	Adenovirus.C	0.0548	0.0138	0.0003
	Adenovirus type 12	0.0276	0.0088	0.003
	Adenovirus type 17	0.0594	0.0117	<.0001
	Papillomavirus	0.0381	0.0120	0.0027
	Papillomavirus2	0.0134	0.0160	0.407
	JC.polyomavirus	-0.0001	0.0109	0.9962
	Monkeypox	0.0033	0.0060	0.5822
	Variola	0.0074	0.0055	0.1841
dsDNA-RT	Hepatitis.B	0.0161	0.0094	0.0937
ssDNA	Parvovirus.B19	-0.0045	0.0075	0.5483
dsRNA	Rotavirus.G3	-0.0133	0.0046	0.0056
+ssRNA	Dengue1	-0.0271	0.0096	0.0074
	Dengue2	-0.0383	0.0089	0.0001
	Dengue3	-0.0249	0.0131	0.064
	Dengue4	-0.0453	0.0090	<.0001
	Hepatitis.C	-0.0041	0.0099	0.6834
	Jap.encephalitis	-0.0328	0.0067	<.0001
	West.Nile	-0.0356	0.0062	<.0001
	Yellow.fever	-0.0206	0.0074	0.0078
	Rubella	0.0565	0.0192	0.0051
	Astrovirus	-0.0236	0.0077	0.0039
	Coronavirus.OC43	-0.0193	0.0066	0.0051
	SARS.coronavirus	-0.0084	0.0076	0.2783
	Hepatitis.A	-0.0279	0.0087	0.0024
	Enterovirus.A	-0.0270	0.0080	0.0015
	Enterovirus.C	-0.0464	0.0078	<.0001
	Enterovirus.D	-0.0348	0.0116	0.0042
	Parechovirus	-0.0696	0.0165	0.0001
	Rhinovirus.A	-0.0300	0.0086	0.0011
	Poliovirus	-0.0436	0.0114	0.0004
	Lym.Chrom.	0.0126	0.0091	0.1725
-ssRNA	Hantavirus.Z10	-0.0152	0.0055	0.0079
	Zaire.ebolavirus	-0.0095	0.0151	0.5338
	Influenza.A.H3N2	-0.0163	0.0057	0.0069
	Influenza.A.H5N1	-0.0105	0.0048	0.0335
	Influenza.B	-0.0129	0.0062	0.0448
	Metapneumovirus	-0.0528	0.0096	<.0001
	Parainfluenza	-0.0533	0.0158	0.0016
	Respiratory.syncytial	-0.0365	0.0083	0.0001
	Measles	0.0145	0.0169	0.395
	Mumps	0.0064	0.0127	0.6191
	Rabies	0.0219	0.0060	0.0007
ssRNA-RT	HIV	-0.0324	0.0106	0.0039
	HIV.2	0.0334	0.0228	0.1504
	T.lymphotropic	0.0577	0.0143	0.0002

[a] Results are color coded as follows: blue denotes viruses in which the efficiency score of HLA-A alleles is higher than the efficiency of HLA-B alleles, red denotes cases where HLA-B alleles have higher efficiency scores, darker shades represent differences which were found to be statistically significant ($P < 0.05$) using a mixed-effects analysis, and lighter shades denote nonsignificant differences.

division of labor between the coinherited HLA-A and HLA-B loci. While molecules encoded in both loci participate in surveillance of various proteins, the HLA-A locus and certain HLA-B alleles appear to have a particularly important role in surveillance of evolutionarily conserved regions of the human proteome (14). This finding is specific to human (rather than randomly selected) proteins and is even more evident at sites of disease-associated mutation, suggesting optimization of ligand selection through human (and ancestral vertebrate) evolution.

Further evidence of partial HLA specialization can also be found through analyses of HLA-viral interactions as HLA alleles that target conserved elements from the human protein repertoire also target conserved regions of human-adapted DNA viruses. In this respect, our findings are supported by other studies (18, 39, 60) indicating that these ancient DNA viruses exploit holes in the repertoire of reactive T cells created through thymic selection, thereby evading effective immune surveillance by maintaining similarity to self-peptides. We extend these observations to show that the extent to which individual HLA alleles are adapted to bind conserved human protein elements is highly correlated with their targeting efficiencies toward DNA viruses. We also find that HLA-B alleles tend to more efficiently target conserved regions of RNA viruses. These results are in keeping with those of Prugnolle et al., who noted that relationships between pathogen diversity

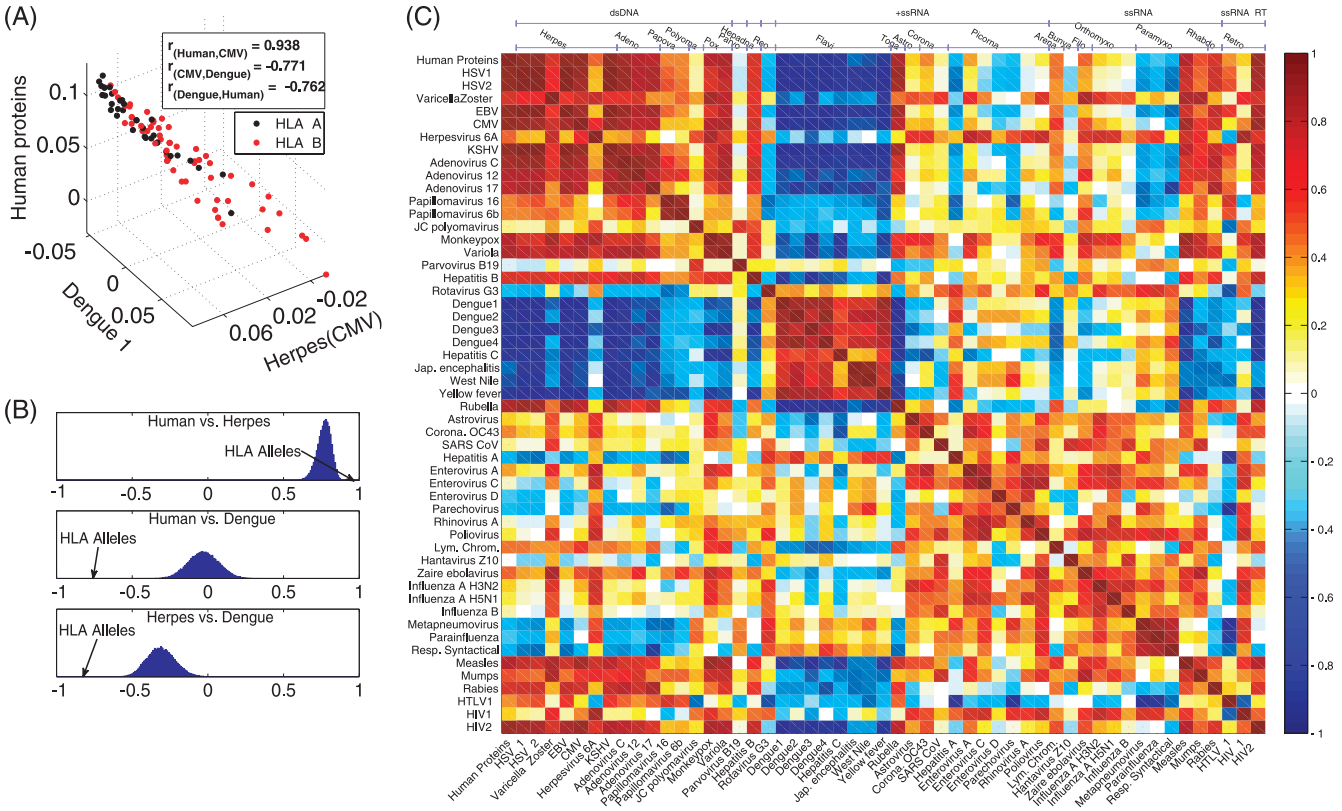

FIG. 7. Correlations between allele efficiency scores for human proteins and human viruses. (A) Correlation (Spearman rank) between allele efficiency scores for human proteins, cytomegalovirus (herpesvirus), and dengue virus (flavivirus) (each representing one column in Fig. 5). (B) Frequency distributions of correlation coefficients between proteomes derived from randomized HLA alleles (n = 10,000) compared with actual values from panel A, as indicated with arrows. (C) Correlation matrix of efficiency scores: human and viral proteomes (the three scores from panel A are dots with appropriate intensity in this matrix). The extent to which HLA efficiency scores are correlated between human viruses, as well as self-peptides (extreme left column), is represented here according to Spearman rank correlation coefficient values. For abbreviations, see the legend of Fig. 5.

and balancing selection are particularly evident at the HLA-B locus (51). They are also supported by the findings of McAdam et al. (41) and by Hughes et al. (24), who found that a large number of HLA-B alleles are products of small-scale recombination events and that the HLA-B locus evolves much more rapidly than the HLA-A locus, suggesting that these two loci have been subject to different types of natural selection over long periods of time in response to different pathogenic threats. Our results are also in line with evidence of more effective HLA context-specific purifying selection followed by reversion in RNA viruses than in DNA viruses, as reported by Hughes et al. (22).

It is important to emphasize that these preferences exhibited by HLA alleles are not evident when either HLA binding energies or evolutionary conservation of target peptides is considered in isolation but only when these factors are considered together. This is in keeping with the findings of Istrail et al. (30), who conducted genome-wide analyses of binding preferences of HLA supertypes and found no meaningful differences in the tendency of HLA alleles to bind human proteins over proteins from other organisms.

Viewed from the perspective of viral evolution, these data suggest that viral species choose distinct adaptive pathways under HLA-restricted immune selection (1, 40). This is most dramatically illustrated for the arboviral *Flaviviridae* species, in which variable rather than conserved proteomic regions are the preferred targets for HLA binding. Evolution toward the "extinction" of predicted HLA targets in the dengue virus genome has been noted previously (21). In this context, it is interesting that dengue virus infection actively promotes (rather than downregulates) TAP (transporter associated with antigen processing)-dependent antigen processing and HLA class I cell surface expression during flavivirus infection (20), indicating that the flaviviruses employ immune evasion strategies that are the opposite of those of many DNA viral species. This particular adaptive strategy may be influenced by the fact that arboviral flaviviruses must maintain the ability to infect arthropod vectors as well as vertebrate hosts (including nonhuman primates) without significant genomic adaptation (37, 59).

The HLA targeting efficiency scores may also prove a useful tool for predicting patient response to infections, as illustrated by the examples of disease outcomes in dengue virus and HIV-1 infections. These scores provide an example of a novel, numeric, and real-valued representation of an HLA molecule, which can be utilized to quantify similarities and differences between HLA molecules based on a target preference function. Such a projection allows identification of common target-

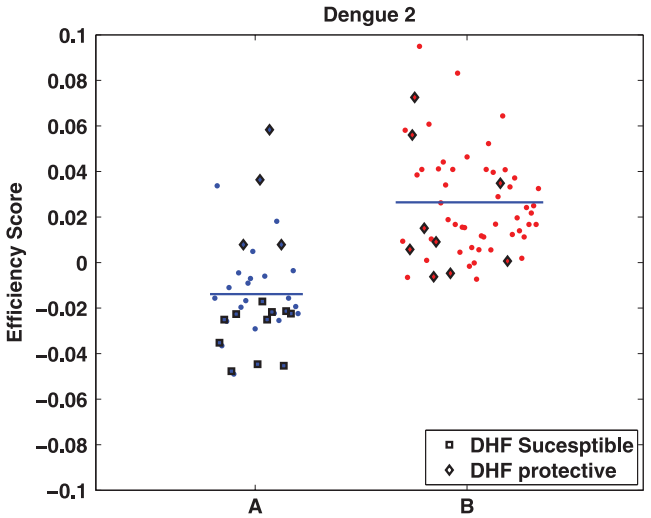

FIG. 8. Targeting efficiencies for dengue virus (serotype 2, whole proteome), for all 95 analyzed HLA alleles. Each dot marks the efficiency score of a single HLA allele. Alleles are sorted by loci. Blue bars represent average locus efficiencies. HLA alleles previously associated with hemorrhagic fever are marked by squares, and those associated with protection are indicated with diamonds. Differences between the two groups were found to be significant ($P = 0.05$).

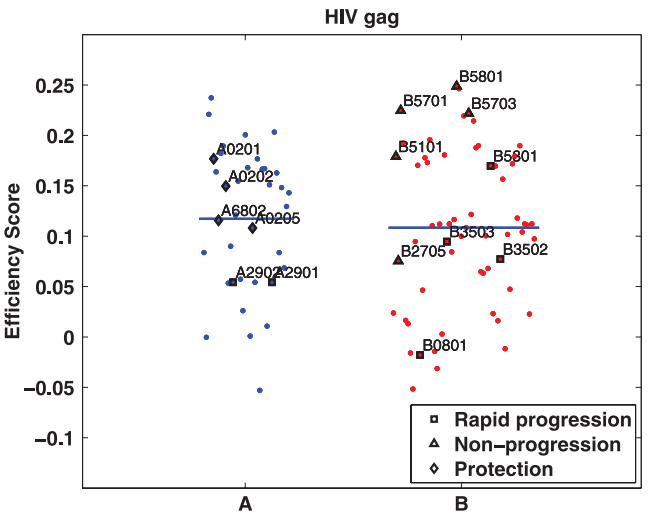

FIG. 9. Targeting efficiencies for HIV-1 Gag protein for all 95 analyzed HLA alleles. Each dot marks the efficiency score of a single HLA allele. Alleles are sorted by loci. Blue bars represent average loci efficiencies. HLA alleles previously associated with slow HIV disease progression are marked by triangles, and those associated with rapid disease progression are indicated with squares. Alleles that have been associated with protection from infection are marked by diamonds.

ing characteristics among patients with different HLA types, thus potentially increasing statistical power in the analysis of patient cohorts. This representation is similar in concept to HLA supertypes (56), which was previously the only method for classifying HLA alleles while attempting to retain biologically meaningful differences. Further studies will be required to investigate these attributes, but it is notable that relationships between HLA targeting efficiency and HLA supertype classifications are by no means uniform, as evidenced in the HIV viral load analysis as well as in the data shown in Fig. 4 and 5 and in Fig. S4 in the supplemental material.

However, multiple factors contribute to disease expression in the context of viral infection, and HLA class I binding is only one of many necessary but not sufficient, genetically determined factors involved in antigen processing and the subsequent generation of pathogen-specific immune responses. To investigate the potential influence of one of these factors, we examined the effect of proteasomal cleavage on HLA targeting efficiency. We found that proteasomal cleavage restriction was also directed toward conserved targets (see Fig. S7) but that the tendency of HLA alleles to target conserved regions remained as strong even when only the peptides which were likely cleavage targets were considered. This suggests that both HLA-peptide binding and proteasomal cleavage have been co-optimized to target conserved regions.

Previous studies of HLA allele-specific viral polymorphisms (27–29, 60, 63) have shown that adaptive interactions between individual human hosts and autologous viral populations are unique and highly dynamic, involving the evolution of HLA-specific CTL escape mutations that are known to influence the natural history of viral infection (43). We therefore offer that the methods described here, designed to investigate the broad patterns of host-pathogen coevolution across multiple viruses, complement other approaches that examine one virus at a time, such as studies that reveal host-virus adaptation by assessing HLA-associated viral polymorphisms (27, 60, 63) or phylogeny (28, 29).

Our analyses using a diverse array of HLA alleles and viral proteomes suggest that, in general, HLA-A preferentially targets DNA viruses and that HLA-B preferentially targets RNA viruses while both HLA-A and -B alleles tend to bind to nonconserved regions in arboviral flaviruses. It must be emphasized that these broad observations identify trends and will not generalize to all the viruses and the individual proteins or epitopes within those viruses.

Although viruses typically encode thousands of amino acids, most of the responding $CD4^+$ and $CD8^+$ T cells recognize a tiny fraction of the potential antigenic determinants (65). This serves to maximize the efficiency of clonal recruitment and activation for a highly specific and avid antiviral response. More than 90% of CD8 T-cell immunodominance is thought to be explained by HLA-peptide binding affinity, as only ~1% of peptides form a complex with HLA class I molecules with sufficient stability to be presented in adequate numbers to activate naïve $CD8^+$ T cells (65). While within-host immunodominance may underpin efficient primary and secondary responses to some acute viral infections, the extreme dominance of a few or even single clonotypes associated with some persistent viruses and vaccine-induced responses is problematic if those immunodominant responses are not protective. The study of targeting efficiency in such infections may help clarify, in part, virus-specific immunodominance patterns and the strategies different groups of viruses have taken to counteract these. This has significant implications for vaccine immunogen design as the efficacy of an immunodominant vaccine-induced response is likely to be improved if directed against determinants that have high targeting efficiency and are

functionally important to the virus rather than determinants that reproduce the counter-evolutionary strategies of the virus.

In conclusion, this study has taken advantage of recent advances in large-scale genome sequencing, HLA binding measurements, and curation, along with the availability of computationally intensive analysis techniques, to address the hypothesis that HLA class I-restricted peptide sampling is preferentially targeted to evolutionarily conserved, functionally important regions of human and viral proteomes. The data support this view and also provide support for balancing selection of HLA class I allelic diversity (particularly at the HLA-B locus) anchored on this property in response to the challenges provided by diverse human viruses. The approach provides a novel perspective on the ongoing coevolutionary relationships between HLA class I polymorphism, adaptive T-cell immunity, and the self-peptides and viruses that engage with these systems.

ACKNOWLEDGMENTS

We thank Matthew Care for providing the OMIM data set and Sarel Fleishman, Chen Yanover, Noah Zaitlen, Felipe Veloso, David Holmes, and David Heckerman for useful discussions. We also thank Jacob John for information regarding plant viruses and Manuel Reyes Gomez for retraining the adaptive double-threading method on the data that excludes HIV epitopes, as well as Itay Mayrose for help with using the ConSeq server. We thank Rachel Tompa and Renee Ireton for editing the manuscript. We thank the anonymous reviewers for useful comments.

REFERENCES

1. **Antoniou, A. N., and S. J. Powis.** 2008. Pathogen evasion strategies for the major histocompatibility complex class I assembly pathway. Immunology **124:**1–12.
2. **Berezin, C., et al.** 2004. ConSeq: the identification of functionally and structurally important residues in protein sequences. Bioinformatics **20:**1322–1324.
3. **Bhasin, M., and G. P. Raghava.** 2007. A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes. J. Biosci. **32:**31–42.
4. **Bhattacharya, T., et al.** 2007. Founder effects in the assessment of HIV polymorphisms and HLA allele associations. Science **315:**1583–1586.
5. **Borghans, J. A. M., J. B. Beltman, and R. J. De Boer.** 2004. MHC polymorphism under host-pathogen coevolution. Immunogenetics **55:**732–739.
6. **Borghans, J. A. M., A. Molgaard, R. J. de Boer, and C. Kesmir.** 2007. HLA alleles associated with slow progression to AIDS truly prefer to present HIV-1 p24. PLoS One **2:**e920.
7. **Buus, S., et al.** 2003. Sensitive quantitative predictions of peptide-MHC binding by a "query by committee" artificial neural network approach. Tissue Antigens **62:**378–384.
8. **Carlson, J., C. Kadie, S. Mallal, and D. Heckerman.** 2007. Leveraging hierarchical population structure in discrete association studies. PLoS One **2:**e591.
9. **Carrington, M., and S. J. O'Brien.** 2003. The influence of HLA genotype on AIDS. Annu. Rev. Med. **54:**535–551.
10. **Chao, D. L., M. P. Davenport, S. Forrest, and A. S. Perelson.** 2005. The effects of thymic selection on the range of T cell cross-reactivity. Eur. J. Immunol. **35:**3452–3459.
11. **Chaturvedi, U. C., R. Nagar, and R. Shrivastava.** 2006. Dengue and dengue haemorrhagic fever: implications of host genetics. FEMS Immunol. Med. Microbiol. **47:**155–166.
12. **da Silva, J.** 1998. Conservation of cytotoxic T lymphocyte (CTL) epitopes as a host strategy to constrain parasite adaptation: evidence from the nef gene of human immunodeficiency virus 1 (HIV-1). Mol. Biol. Evol. **15:**1259–1268.
13. **Davison, A. J., M. Benko, and B. Harrach.** 2003. Genetic content and evolution of adenoviruses. J. Gen. Virol. **84:**2895–2908.
14. **De Tomaso, A. W., et al.** 2005. Isolation and characterization of a protochordate histocompatibility locus. Nature **438:**454–459.
15. **Dong, T., et al.** 2007. High pro-inflammatory cytokine secretion and loss of high avidity cross-reactive cytotoxic T-cells during the course of secondary dengue virus infection. PLoS One **2:**e1192.
16. **Donnes, P., and O. Kohlbacher.** 2006. SVMHC: a server for prediction of MHC-binding peptides. Nucleic Acids Res. **34:**W194–W197.
17. **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.
18. **Frankild, S., R. J. de Boer, O. Lund, M. Nielsen, and C. Kesmir.** 2008. Amino acid similarity accounts for T cell cross-reactivity and for "holes" in the T cell repertoire. PLoS One **3:**e1831.
19. **Haynes, B. F., G. Pantaleo, and A. S. Fauci.** 1996. Toward an understanding of the correlates of protective immunity to HIV infection. Science **271:**324–328.
20. **Hershkovitz, O., et al.** 2008. Dengue virus replicon expressing the nonstructural proteins suffices to enhance membrane expression of HLA class I and inhibit lysis by human NK cells. J. Virol. **82:**7666–7676.
21. **Hughes, A. L.** 2001. Evolutionary change of predicted cytotoxic T cell epitopes of dengue virus. Infect. Genet. Evol. **1:**123–130.
22. **Hughes, A. L., and M. A. K. Hughes.** 2007. More effective purifying selection on RNA viruses than in DNA viruses. Gene **404:**117–125.
23. **Hughes, A. L., and M. K. Hughes.** 1995. Self peptides bound by HLA class I molecules are derived from highly conserved regions of a set of evolutionarily conserved proteins. Immunogenetics **41:**257–262.
24. **Hughes, A. L., M. K. Hughes, and D. I. Watkins.** 1993. Contrasting roles of interallelic recombination at the HLA-A and HLA-B loci. Genetics **133:**669–680.
25. **Hughes, A. L., and M. Nei.** 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335:**167–170.
26. **Hughes, A. L., et al.** 2003. Widespread purifying selection at polymorphic sites in human protein-coding loci. Proc. Natl. Acad. Sci. U. S. A. **100:**15754–15757.
27. **Huseby, E. S., et al.** 2005. How the T cell repertoire becomes peptide and MHC specific. Cell **122:**247–260.
28. **Irausquin, S. J., and A. L. Hughes.** 2010. Conflicting selection pressures target the NS3 protein in hepatitis C virus genotypes 1a and 1b. Virus Res. **147:**202–207.
29. **Irausquin, S. J., and A. L. Hughes.** 2008. Distinctive pattern of sequence polymorphism in the NS3 protein of hepatitis C virus type 1b reflects conflicting evolutionary pressures. J. Gen. Virol. **89:**1921–1929.
30. **Istrail, S., et al.** 2004. Comparative immunopeptidomics of humans and their pathogens. Proc. Natl. Acad. Sci. U. S. A. **101:**13268–13272.
31. **Jacob, L., and J. P. Vert.** 2008. Efficient peptide-MHC-I binding prediction for alleles with few known binders. Bioinformatics **24:**358–366.
32. **Jojic, N., M. Reyes-Gomez, D. Heckerman, C. Kadie, and O. Schueler-Furman.** 2006. Learning MHC I–peptide binding. Bioinformatics **22:**e227–e235.
33. **Karrer, U., et al.** 2003. Memory inflation: Continuous accumulation of antiviral $CD8^+$ T cells over time. J. Immunol. **170:**2022–2029.
34. **Kelleher, A. D., 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.
35. **Kiepiela, P., et al.** 2004. Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. Nature **432:**769–775.
36. **Kiepiela, P., et al.** 2007. $CD8^+$ T-cell responses to different HIV proteins have discordant associations with viral load. Nat. Med. **13:**46–53.
37. **Kuno, G., and G. J. J. Chang.** 2005. Biological transmission of arboviruses: reexamination of and new insights into components, mechanisms, and unique traits as well as their evolutionary trends. Clin. Microbiol. Rev. **18:**608–637.
38. **Lin, H. H., S. Ray, S. Tongchusak, E. L. Reinherz, and V. Brusic.** 2008. Evaluation of MHC class I peptide binding prediction servers: Applications for vaccine research. BMC Immunol. **9:**8.
39. **Louzoun, Y., T. Vider, and M. Weigert.** 2006. T-cell epitope repertoire as predicted from human and viral genomes. Mol. Immunol. **43:**559–569.
40. **Lucas, M., U. Karrer, A. Lucas, and P. Klenerman.** 2001. Viral escape mechanisms: escapology taught by viruses. Int. J. Exp. Pathol. **82:**269–286.
41. **McAdam, S. N., et al.** 1994. A uniquely high level of recombination at the HLA-B locus. Proc. Natl. Acad. Sci. U. S. A. **91:**5893–5897.
42. **McGeoch, D. J., F. J. Rixon, and A. J. Davison.** 2006. Topics in herpesvirus genomics and evolution. Virus Res. **117:**90–104.
43. **Moore, C. B., et al.** 2002. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. Science **296:**1439–1443.
44. **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.
45. **Nielsen, M., C. Lundegaard, O. Lund, and C. Kesmir.** 2005. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. Immunogenetics **57:**33–41.
46. **Nielsen, M., et al.** 2003. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci. **12:**1007–1017.
47. **Paulsson, K. M.** 2004. Evolutionary and functional perspectives of the major histocompatibility complex class I antigen-processing machinery. Cell. Mol. Life Sci. **61:**2446–2460.
48. **Peters, B., et al.** 2006. A community resource benchmarking predictions of peptide binding to MHC-I molecules. PLoS Comput. Biol. **2:**e65.
49. **Peters, B., and A. Sette.** 2005. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. BMC Bioinformatics **6:**132.

50. **Peters, B., et al.** 2005. The design and implementation of the immune epitope database and analysis resource. Immunogenetics **57:**326–336.
51. **Prugnolle, F., et al.** 2005. Pathogen-driven selection and worldwide HLA class I diversity. Curr. Biol. **15:**1022–1027.
52. **Pybus, O. G., et al.** 2007. Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. Mol. Biol. Evol. **24:**845–852.
53. **Rousseau, C. M., et al.** 2008. HLA class I-driven evolution of human immunodeficiency virus type 1 subtype C proteome: immune escape and viral load. J. Virol. **82:**6434–6446.
54. **Sacha, J. B., et al.** 2007. Gag-specific $CD8^+$ T lymphocytes recognize infected cells before AIDS-virus integration and viral protein expression. J. Immunol. **178:**2746–2754.
55. **Saxova, P., S. Buus, S. Brunak, and C. Kesmir.** 2003. Predicting proteasomal cleavage sites: a comparison of available methods. Int. Immunol. **15:**781–787.
56. **Sidney, J., B. Peters, N. Frahm, C. Brander, and A. Sette.** 2008. HLA class I supertypes: a revised and updated classification. BMC Immunol. **9:**1.
57. **Subramanian, S., and S. Kumar.** 2006. Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. BMC Genomics **7:**306.
58. **Takahata, N., and M. Nei.** 1990. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics **124:**967–978.
59. **Vasilakis, N., et al.** 2007. Potential of ancestral sylvatic dengue-2 viruses to re-emerge. Virology **358:**402–412.
60. **Vider-Shalit, T., V. Fishbain, S. Raffaeli, and Y. Louzoun.** 2007. Phase-dependent immune evasion of herpesviruses. J. Virol. **81:**9536–9545.
61. **Vossen, M. T. M., E. M. Westerhout, C. Soderberg-Naucler, and E. J. H. J. Wiertz.** 2002. Viral immune evasion: A masterpiece of evolution. Immunogenetics **54:**527–542.
62. **Westover, K. M., and A. L. Hughes.** 2007. Evolution of cytotoxic T-lymphocyte epitopes in hepatitis B virus. Infect. Genet. Evol. **7:**254–262.
63. **Wong, P., G. M. Barton, K. A. Forbush, and A. Y. Rudensky.** 2001. Dynamic tuning of T cell reactivity by self-peptide-major histocompatibility complex ligands. J. Exp. Med. **193:**1179–1187.
64. **Yeager, M., M. Carrington, and A. L. Hughes.** 2000. Class I and class II MHC bind self peptide sets that are strikingly different in their evolutionary characteristics. Immunogenetics **51:**8–15.
65. **Yewdell, J. W.** 2006. Confronting complexity: real-world immunodominance in antiviral $CD8^+$ T cell responses. Immunity **25:**533–543.
66. **Yu, D., M. C. Silva, and T. Shenk.** 2003. Functional map of human cytomegalovirus AD169 defined by global mutational analysis. Proc. Natl. Acad. Sci. U. S. A. **100:**12396–12401.