Classification:

Major: Biological Sciences

Minor: Genetics

HIV Peptidome-Wide Association Study Reveals Patient-Specific Epitope Repertoires Associated with HIV Control

Short Title: Epitope Repertoires Associated with HIV Control

Jatin Arora[1], Paul J. McLaren[2,3], Nimisha Chaturvedi[4,5], Mary Carrington[6,7], Jacques Fellay[4,5] & Tobias L. Lenz[1,*]

Affiliation Affiliations:

[1] Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, 24306 Plön, Germany. [2] JC Wilt Infectious Diseases Research Center, National HIV and Retrovirology Laboratory, Public Health Agency of Canada, R3E 0W3, Winnipeg, Canada. [3] Department of Medical Microbiology and Infectious Diseases, University of Manitoba, R3E 0J9, Winnipeg Canada. [4] Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland. [5] Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland. [6] Cancer and Inflammation Program, Leidos Biomedical Research, Frederick National Laboratory, Frederick, MD 21702, USA. [7] Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02139-3583, USA

Corresponding Author:

Tobias L. Lenz

Max Planck Institute for Evolutionary Biology

August-Thienemann-Str. 2

Plön 24306, Germany

Phone: +49 4522 763 228

Email: lenz@post.harvard.edu

Keywords:
Human Leukocyte Antigen (HLA) | HIV | Evolution | Adaptive Immunity

Abstract:
Genetic variation in the peptide-binding groove of the highly polymorphic human leukocyte antigen (HLA) class I molecules has repeatedly been associated with HIV-1 control and progression to AIDS, accounting for up to 12% of the variation in HIV-1 set point viral load (spVL). This suggests a key role in disease control for HLA presentation of HIV-1 epitopes to cytotoxic T cells. However, a comprehensive understanding of the relevant HLA-bound HIV epitopes is still elusive. Here we describe a peptidome-wide association study (PepWAS) approach that integrates HLA genotypes and spVL data from 6,311 HIV-infected patients to interrogate the entire HIV-1 proteome (3,252 unique peptides) for disease-relevant peptides. This PepWAS approach predicts a core set of epitopes associated with spVL, including previously characterized epitopes but also several novel disease-relevant peptides. More importantly, each patient presents only a small subset of these predicted core epitopes through their individual HLA-A and -B variants. Eventually, the individual differences in these patient-specific epitope repertoires account for the variation in spVL that was previously associated with HLA genetic variation. PepWAS thus enables a comprehensive functional interpretation of the robust but little understood association between HLA and HIV-1 control, prioritizing a short list of disease-associated epitopes for the development of targeted therapy.

Significance Statement:

Individual differences in HIV-1 control and progression to AIDS have been pinpointed to genetic variation in the Human Leukocyte Antigen (HLA), coding for antigen-presenting molecules. However, our understanding of the corresponding antigens is still incomplete. Here we developed a new approach that combines HLA genotypes and viral load data of HIV infected individuals to screen the entire HIV proteome for disease-relevant peptides. Our PepWAS approach identified a limited manageable core set of peptides, accounting for the entire variation in viral load previously associated with genetic variation in the HLA. This core set of disease-relevant antigens thus provides a functional link between HLA genetic variation and HIV-1 control, confirming several known antigens, but also prioritizing novel antigens as new therapeutic targets.

\body

HLA class I proteins are thought to play a critical role in immune recognition of HIV-1 by presenting endogenously processed viral peptides at the surface of infected cells to cytotoxic T cells, in order to trigger destruction of the infected cells (1). Indeed, genetic variation in the HLA region has repeatedly been identified as the major genetic determinant of HIV-1 control in genome-wide association studies (2, 3). Most recently, McLaren *et al.* (4) fine-mapped the entire HLA's association with HIV-1 control and disease progression to five independent amino acid residues in the peptide binding groove of the HLA-B and HLA-A molecules. These five residues alone accounted for 12.3% of the variation in viral load, suggesting a major role for specific HLA-presented viral epitopes in HIV-1 control. However, our understanding of the disease-relevant viral epitopes is still incomplete, hampered by the economically hardly feasible challenge of employing a full-factorial experimental assay to screen the entirety of the HIV-1 peptidome for binding by all relevant HLA alleles. Therefore, we developed a novel computational analysis approach that identifies and prioritizes disease-associated peptides based on individual HLA genotype and disease phenotype information. Our approach uses established computational algorithms to predict for each individual whether a given peptide is bound by the individual's HLA variants, and then uses regression analysis on the disease phenotype (here HIV set point viral load) to estimate whether the ability to bind the peptide is non-randomly associated with the disease phenotype. This approach is analogous to a genetic association study, except that it incorporates one additional layer by translating genetic variation into functional variation (HLA variant-specific peptide binding). Importantly, this approach does not simply define all peptides bound by a risk HLA variant as risk peptides. *Instead, for each peptide it integrates the disease effect of all HLA variants that are able to bind the peptide and thus estimates a peptide-specific association with disease.* Since most peptides are bound by several HLA variants, integrating the effect of all binding HLA variants is essential (**Fig. 1**). For instance, a peptide can have no association with disease, even if it is bound by the highest risk variant, simply because it is also bound by several other non-associated (or even protective) variants. Ultimately, our approach identifies a list of peptides with varying associations to disease, which can directly inform therapy development by prioritizing global as well as patient-specific candidate epitopes. As a

proof-of-concept, we analyze here a unique dataset of 6,311 individuals of European ancestry with chronic HIV-1 infection (***SI Appendix*, Table S1)**. Screening the entire HIV-1 peptidome for candidate epitopes, we identify a comprehensive list of peptides that explain the well-established association between HLA genetic variation and HIV-1 control, including several previously uncharacterized epitopes as novel candidates for targeted therapy.

Results and Discussion:

Our analyses are based on a large dataset of HIV-infected individuals (4) that includes both pre-treatment level of set point viral load (spVL) as a correlate of disease progression (5) and imputed HLA genotypes (4-digit allele resolution). We focused on the two HLA loci (HLA-B and HLA-A) reported to have independent associations with HIV-1 control and disease progression (4). Potential HLA-bound peptides were identified using an established computational algorithm that is based on empirical training data (6) and integrates several complementary prediction methods in a consensus approach, outperforming comparable algorithms (6, 7). Such algorithms have been used in a wide spectrum of HLA-related studies ranging from vaccine design to cancer evolution and HIV disease genetics (8–10). Without *a-priori* selection, we screened all possible 9mer HIV-1 peptides (N = 3,252) in a sliding window across the entire HIV-1 M group subtype B reference proteome (11) against all represented HLA-B and HLA-A alleles (344,712 HLA:peptide complexes), and identified 214 and 173 distinct HIV-1 peptides predicted to be bound by one or more of the represented HLA-B and HLA-A alleles, respectively.

In order to evaluate the significance of the predicted peptide repertoires, we interrogated several layers of empirical evidence (see ***SI Appendix*, Supporting Text)**. We observed an enrichment for previously known HIV-1 epitopes **(*SI Appendix*, Fig. S1A)**, a correlation between an HLA-B allele's effect on viral load and the number of HIV-1 peptides it is predicted to bind (***SI Appendix*, Fig. S1B)**, and detected previously reported viral escape mutations (***SI Appendix*, Fig. S1C)**. Following these independent layers of evidence that our analysis pipeline predicts disease-relevant binding of HLA to HIV-1 peptides, we subsequently refer to the entire predicted set of HLA-bound peptides as *predicted epitopes*, highlighting the point that not all of them have been experimentally validated.

Next, we tested whether the patient-specific repertoire of predicted HIV-1 epitopes, defined by the number of peptides predicted to be bound by the specific HLA allele combination of the patient, was associated with spVL. For this, we ran a linear regression across the 6,311 HIV-1 patients, with spVL as dependent variable and the patient-specific number of bound peptides as predicting variable, together with other covariates (see methods). We first focused on the effect of peptides bound by HLA-B, and used only the known CTL epitopes from the Los Alamos HIV Molecular Immunology Database (12), of which 80 were represented among the 214 predicted HLA-B bound epitopes. The individual number of these known CTL epitopes bound by patient-specific HLA-B variants accounted for only 1.8% of the individual variation in spVL (**Fig. 2**). In order to evaluate this association, we then included all predicted HLA-bound HIV-1 epitopes (N = 214) in the analysis, including the previously known CTL epitopes as well as any other HLA-B-bound peptide from the HIV-1 proteome. Interestingly, the total number of all predicted HLA-B-bound epitopes per patient accounted for 5.3% variation in spVL (**Fig. 2**), suggesting that the Los Alamos CTL epitope dataset is not yet fully saturated with regard to disease-relevant peptides. However, the accounted variation was still lower than the 11.4% variation associated with genetic variation at HLA-B in previous genotype-based studies, suggesting that the total predicted epitope repertoire still included peptides irrelevant for the association between HLA and HIV-1. This is supported by a previous study, which showed that not all HLA-bound peptides are epitopes targeted by CD8+ T-cells (13). We thus aimed to refine the repertoire of predicted HLA-bound HIV-1 epitopes further to comprise only disease-relevant epitopes. For this, we calculated the epitope-specific association with spVL by running a separate linear regression for each predicted epitope and recording R^2 and β-coefficient as measures of the epitope's effect on spVL. This is analogous to the approach of a genome-wide association study (GWAS), where each genetic variant is tested for its association with a given trait, except that here we focus on functional protein variation (peptide binding by a patient's HLA molecules) rather than genetic variation. Following this analogy, we term our approach *peptidome-wide association study* (PepWAS). Of 214 HIV-1 epitopes predicted to be bound by HLA-B, 132 accounted for nominal variation (adjusted R^2 value > 0) in spVL, 74 of which were negatively and 58 positively associated with spVL (β-coefficients ranging from -0.1 to 0.77; ***SI Appendix*, Table S2)**. Importantly,

we do not require statistical significance at this point as this is a candidate screen and we thus aim to minimize the number of false negatives. Subsequently, we designate the nominally associated epitopes as *disease-associated* predicted epitopes, even though their effects are not necessarily independent as they were tested with separate regression models. An analogous investigation of peptide binding by HLA-A alleles revealed an additional 74 disease-associated epitopes (***SI Appendix*, Table S3)**.

Having refined the predicted HIV-1 epitope repertoire to only disease-associated predicted epitopes, we then tested whether this subset accounted for a larger fraction of the variation in spVL than the total predicted HIV-1 epitope repertoire. Indeed, the patients' ability to bind a smaller or larger fraction of the HLA-B-specific disease-associated predicted epitopes accounted for 11.4% of the variation in spVL (**Fig. 2**). Similarly, for HLA-A, the total number of predicted HIV-1 epitopes bound by individual HLA-A genotypes accounted for 0.3% of the variation, while disease-associated predicted epitopes accounted for 1.4% of the variation in spVL. On average, a patient's HLA-B allele pair bound 16.2 ±7 (SD) disease-associated predicted HIV-1 epitopes, while its HLA-A alleles bound significantly less (6.6 ±6.5; Paired Wilcox rank sum test, $P < 0.0001$; ***SI Appendix*, Fig. S6)**. This quantitative difference in peptide presentation might contribute to the stronger spVL-association of HLA-B compared to HLA-A, as a larger number of presented peptides should more likely lead to a more efficient CD8 T cell response, as has indeed been observed for HLA-B compared to HLA-A (14). HLA-C-bound epitopes did not show any significant association with spVL, mirroring the lack of independent genetic associations for HLA-C in the latest GWAS (4). Predicted disease-associated epitopes of HLA-B and HLA-A together accounted for 12.2% of the variation in HIV-1 viral load, approximately corresponding to the 12.3% variation previously attributed to all independent genetic associations in the entire HLA (**Fig. 2A**).

Interestingly, the *Env* protein showed the largest number of disease-associated predicted epitopes, with both positive and negative effects. Among the disease-associated predicted HLA-B-bound epitopes, *Env*-derived epitopes alone accounted for 6.4% of variation in spVL, the highest among all HIV-1 proteins (**Fig. 3A**). In addition to already known *Env*-derived CD8+ T-cell targeted epitopes associated with lower viral load and disease control e.g. RIKQIINMW, HRLRDLLLI (13), ERYLKDQQL (15), our analysis revealed

previously undescribed HLA-epitope complexes e.g. B*57:01-STQLFNSTW, -NSTWFNSTW, or -RGWEALKYW showing strong associations with lower viral load (**Fig. 3C**). The potential importance of the predicted *Env* epitopes is quite surprising, since the high genetic variability of the *Env* protein across different HIV-1 isolates suggests that the virus could readily evolve escape variants in this protein. However, a previous study has already established that sequence conservation alone is not a reliable predictor of protective epitopes, instead highlighting structural conservation as the more important feature (13). More intriguingly, we found that the protective *Env* epitopes predicted through our PepWAS approach are significantly enriched for residues that are associated with broadly neutralizing antibodies (bNAbs; OR = 1.5, P = 0.036, ***SI Appendix*, Fig. S7)**, suggesting that they represent parts of the Env protein that can be efficiently targeted in both antibody therapy as well as in HLA-mediated CTL response.

Notably, several of the represented HLA alleles were predicted to bind both negatively and positively disease-associated epitopes (***SI Appendix*, Tables S4 and S5)**, i.e. epitopes bound by the same HLA allele did not necessarily have the same effect on viral load. This can be explained by the fact that a given epitope can be bound by several different HLA alleles with very distinct disease association (see schematic in **Fig. 1**). This is also in agreement with a previous study showing that viral control is mediated by specific immunogenic epitopes which could be restricted by HLA alleles other than already known ones (13).

HLA molecule variants are known to bind peptide repertoires with distinct anchor motifs, based on the composition of their peptide-binding groove (16). This entailed the possibility that our PepWAS approach is merely identifying distinct groups of peptides per HLA variant, thus translating the known HLA variant-specific effect on viral load into peptide group-specific effects. While still helpful in guiding epitope research, this would provide only limited knowledge-gain compared to the HLA allele-specific associations known from previous work (4). In order to test for this possibility, we performed a cluster analysis on the predicted disease-associated epitopes bound by HLA-B (N = 132) and analyzed cluster-specific motifs and HLA allele binding patterns. Intriguingly, among the ten most dominant epitope clusters, each exhibiting a distinct peptide motif, nine were defined by multiple HLA-B alleles **(Fig. 4)**, some of them even belonging to different supertypes (***SI***

Appendix*, Table S7)**. All of these clusters included both novel and previously described epitopes, and three of them were defined by both risk- and protection-conferring alleles. Furthermore, all HLA variants bound peptides of multiple dominant clusters; e.g. B*57:01 is associated with 3 dominant clusters, each showing a distinct peptide motif, but all showing a strong preference for amino acid 'W' at anchor position 9 **(Fig. 4)**. Overall, the cluster analysis shows that our PepWAS approach identifies groups of peptides with distinct motifs that are different from HLA variant-specific binding motifs (see also schematic in **Fig. 1**). Generally, the 24 disease-associated epitopes predicted to be bound by HLA-B*57:01 (but some of these also by other alleles), accounted for the highest level of variation in spVL, even though they derived from 5 different HIV-1 proteins (**Fig. 3B, C and *SI Appendix*, Fig. S8)**. One of these epitopes, the well characterized HIV-1 *Gag* epitope ISPRTLNAW (belonging to the dark purple cluster in **Fig. 4**), slightly exceeded the effect of all other epitopes (**Fig. 3B**), in concordance with experimental evidence (17). Other HLA-B alleles, including the B*08, B*44, and B*51 types, were also included in our dataset, and their predicted epitope repertoires roughly followed their disease-association known from previous studies (SI Appendix*, Fig. S1B;** allele-specific associations and number of bound peptides are given in ***SI Appendix*, Table S4)**.

Mechanistically, a negative association between predicted HIV-1 epitopes and viral load is intuitive and likely resulting from the peptides' immunodominant role in CTL response and their escape mutations leading to significant fitness costs for the virus. However, a number of the predicted HIV-1 epitopes exhibited a positive association with viral load, indicating that they confer lower disease protection relative to the bulk of the peptides. They likely represent peptide variants that fail to elicit an efficient CTL response or can readily mutate with negligible fitness effects, thus allowing viral escape from HLA presentation at no cost for the virus. Indeed, the most risk-associated predicted *Vpu* epitope, IPIVAIVAL (***SI Appendix*, Fig. S8F;** belonging to the largest, grey cluster in **Fig. 4**), includes an anchor residue that exhibits significant variation in primary HIV-1 clones and is involved in mediating immune-evasion through down-regulation of HLA-C (18), whose high expression has been implicated in HIV control (19). The lack of significant associations between predicted HLA-C bound epitopes and viral load in our analysis might indicate that previously observed viral control associated with HLA-C is not mediated

through specific peptide presentation of HLA-C. However, more research is required to fully understand the role of HLA-C in viral control (18).

So far, our analysis was based on the HIV-1 genome reference sequence. Though widely used for research, focusing on this sequence accession may restrict our findings. We thus repeated the entire analysis using the HIV-1 proteome consensus sequence from the Los Alamos database, which incorporates major variation across different HIV-1 strains. The results remained qualitatively the same (***SI Appendix*, Fig. S9 and Table S8)**. However, HIV is well known to exhibit substantial within-host evolution (20, 21) and it is easily conceivable that the ability of a patient's HLA variants to bind HIV epitopes is significantly affected by genetic variation in the patient's HIV population (22). We therefore also analyzed patient-specific autologous HIV-1 sequence information, which was available for a small subset of patients, covering 8 of the 10 HIV-1 proteins (***SI Appendix*, Table S6)**. For 4 of the 8 proteins (*Gag*, *Pol*, *Vif* and *Nef*) we found that the proportion of variation in spVL associated with HLA-bound epitope repertoires changed when predicting epitopes from autologous sequences instead of from the reference sequence. In all 4 cases, the variation associated with predicted autologous epitopes was higher than when using their homologs from the reference sequence (***SI Appendix*, Table S6)**, suggesting that our PepWAS approach might be able to explain more variation in spVL than a standard GWAS if autologous sequences were available for a larger fraction of infected individuals.

PepWAS relies on computational algorithms for the prediction of binding affinities between HLA variants and peptides, and is thus inherently limited by their accuracy and specificity. For instance, the empirical data used to train currently established HLA class I algorithms contains mainly 9mer peptides, even though HLA class I molecules can occasionally bind slightly shorter or longer peptides. Such peptides might therefore be missed by current prediction algorithms. On the other hand, the current setup does in fact identify 9mer cores of larger known epitopes. For instance, the here predicted protective 9mer *Gag* epitope 'STLQEQIGW' resides within the previously described 10mer *Gag* epitope TW10 (**Fig. 3B**). Furthermore, this limitation is likely to be alleviated as more training data is becoming available.

Overall, our findings reveal a functional basis of the robustly established association between HLA genes and HIV-1 infection outcome. We show that both quantity and quality

of HLA-bound HIV epitopes contribute to controlling a patient's viral load. Our data also suggests a more important role for *Env* protein-derived epitopes than previously thought. Ultimately, our PepWAS approach of combining computational HLA-specific epitope prediction with disease phenotype validation provides a promising avenue for identification and prioritization of novel epitopes. As such, it complements existing empirical essays for the development of targeted therapy. Noteworthy, by involving a functional layer (peptide binding), the PepWAS approach enables the detection of disease-relevant properties that are shared among several genetic variants (overlap in peptide binding among HLA alleles). Such shared properties would be undetectable by GWAS, because of its focus on distinct genetic variants instead of function, and should therefore lead to higher sensitivity in the PepWAS approach compared to GWAS. Furthermore, the PepWAS approach allows to account for individual variation in the pathogen proteome if autologous sequence information is available, potentially further increasing sensitivity. As such it may be applied to any HLA-associated complex disease.

Material and Methods:

For detailed information on Material and Methods see *SI Appendix* Supporting Methods available online.

Samples and Genotype data:

We analyzed HLA genotype data and set point viral load (spVL) measurements of 6,311 subjects chronically infected with HIV-1. The original data and thorough quality check are described in detail in McLaren *et al.* (4) and explained briefly in Supporting Methods.

HLA binding affinity for HIV-1 epitopes:

We used the NCBI accession NC_001802.1 as the reference sequence for the HIV-1 proteome (M group subtype B). The algorithm NetMHCcons-1.1 was used to predict HLA allele-specific binding affinities for all 9mer peptides generated from the entire HIV-1 proteome, applying the default affinity rank threshold for 'strongly bound' peptides (rank < 0.5).

Association with viral load:

The association of an allele or a peptide with viral load (spVL) was calculated using a linear regression model corrected for population covariates following McLaren *et al.* (4). Covariates included the first five principle components of SNP variation and the cohort identity (all adopted from McLaren *et al.* (4)). Variation in viral load attributable to a given variable (allele or epitope) was calculated as the difference between adjusted-R^2 values of the model with variable and covariates and the model with covariates only, following McLaren et al. (4). The variable's regression coefficient was used as the measure of its effect on viral load.

Clustering of HLA-B-specific predicted epitopes:

Position-associated entropy was calculated for all HLA-B-bound disease-associated epitopes (N = 132) and used for visualization in a non-metric multidimensional scaling plot as well as for density-based clustering.

HLA binding of peptides from autologous HIV-1 sequences:

We analyzed autologous HIV-1 sequences from Bartha et al. (23). Autologous sequences were available for 8 of 10 HIV-1 proteins (only Gp41 segment for *Env*), and only for a small subset of patients in our cohorts (***SI Appendix*, Table S6**).

Acknowledgements:

Patient and HIV sequence data was collected and generously provided by the International Collaboration for the Genomics of HIV. This project was supported by the Emmy Noether Programme of the Deutsche Forschungsgemeinschaft (DFG grant LE 2593/3-1 to T.L.L.). Furthermore, this project has been funded in whole or in part with federal funds from the Frederick National Laboratory for Cancer Research, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This Research was also supported in part by the Intramural Research Program of the NIH, Frederick National Lab, Center for Cancer Research.

Competing interests:

The authors declare no conflict of interest.

References:

1. Goulder PJR, Walker BD (2012) HIV and HLA Class I: An Evolving Relationship. *Immunity* 37(3):426–440.
2. Fellay J, et al. (2007) A whole-genome association study of major determinants for host control of HIV-1. *Science (80-)* 317(5840):944–947.
3. Pereyra F, et al. (2010) The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science (80-)*. doi:10.1126/science.1195271.
4. McLaren PJ, et al. (2015) Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load. *Proc Natl Acad Sci* 112(47):14658–14663.
5. Mellors JW, et al. (1996) Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science (80-)* 272(5265):1167–1170.
6. Karosiene E, Lundegaard C, Lund O, Nielsen M (2012) NetMHCcons: A consensus method for the major histocompatibility complex class i predictions. *Immunogenetics* 64(3):177–186.
7. Zhang H, Lundegaard C, Nielsen M (2009) Pan-specific MHC class I predictors: A benchmark of HLA class I pan-specific prediction methods. *Bioinformatics* 25(1):83–89.
8. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N (2015) Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* 160(1–2):48–61.
9. Strønen E, et al. (2016) Targeting of cancer neoantigens with donor-derived T cell receptor repertoires. *Science (80-)* 352(6291):1337–1341.
10. Košmrlj A, et al. (2010) Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection. *Nature* 465(7296):350–354.
11. Martoglio B (1997) Signal peptide fragments of preprolactin and HIV-1 p-gp160 interact with calmodulin. *EMBO J* 16(22):6636–6645.
12. Yusim K, et al. (2009) HIV molecular immunology. *Los Alamos, New Mex Los Alamos Natl Lab Theor Biol Biophys*:3–24.
13. Pereyra F, et al. (2014) HIV Control Is Mediated in Part by CD8+ T-Cell Targeting

of Specific Epitopes. *J Virol* 88(22):12937–12948.

14. Kiepiela P, et al. (2004) Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* 432(7018):769–775.

15. Borrow P, Lewicki H, Hahn BH, Shaw GM, Oldstone MBA (1994) Virus-Specific CD8+ Cytotoxic T-Lymphocyte Activity Associated with Control of Viremia in Primary Human Immunodeficiency. 68(9):6103–6110.

16. Falk K, Rötzschke O, Stevanović S, Jung G, Rammensee HG (1991) Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* 351(6324):290–296.

17. Llano A, Williams A, Olvera A, Silva-Arrieta S, Brander C (2013) Best-Characterized HIV-1 CTL Epitopes: The 2013 Update. *HIV Mol Immunol 2013*:3–25.

18. Apps R, et al. (2016) HIV-1 Vpu Mediates HLA-C Downregulation. *Cell Host Microbe* 19(5):686–695.

19. Thomas R, et al. (2009) HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. *Nat Genet* 41(12):1290–1294.

20. Cotton LA, et al. (2014) Genotypic and Functional Impact of HIV-1 Adaptation to Its Host Population during the North American Epidemic. *PLoS Genet* 10(4). doi:10.1371/journal.pgen.1004295.

21. Li G, et al. (2015) An integrated map of HIV genome-wide variation from a population perspective. *Retrovirology* 12(1). doi:10.1186/s12977-015-0148-6.

22. Kawashima Y, et al. (2009) Adaptation of HIV-1 to human leukocyte antigen class I. *Nature* 458(7238):641–645.

23. Bartha I, et al. (2013) A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *Elife* 2013(2):1–16.

24. McLaren PJ, Carrington M (2015) The impact of host genetic variation on infection with HIV-1. *Nat Immunol* 16(6):577–583.

Figure legends:

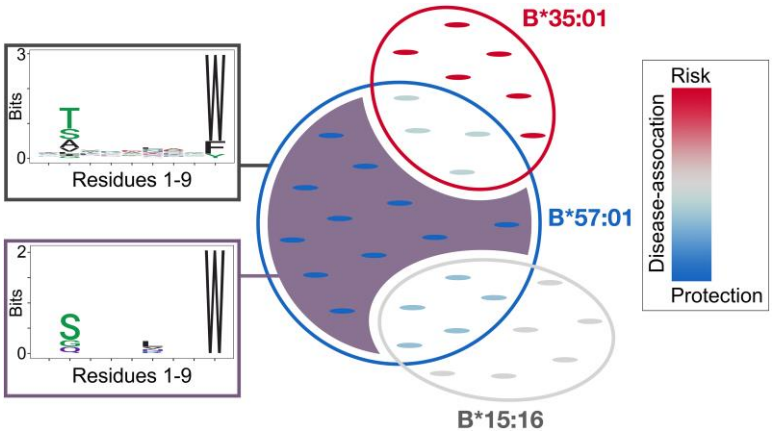
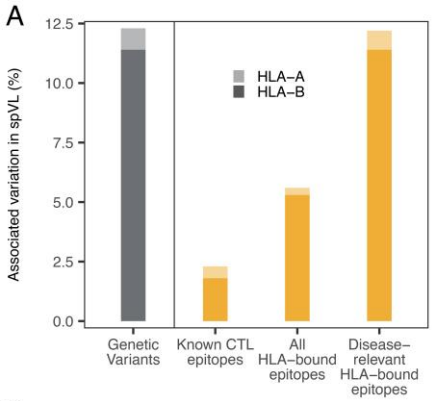


Fig. 1. Schematic for determining peptide-specific associations through PepWAS. Disease-associated peptides are identified by integrating the different disease-associations of the different HLA alleles that are predicted to bind them. Some peptides will only be bound by one HLA allele, thus drawing their disease-association directly from the disease-association of that allele (e.g. peptides in the purple shaded area, bound only by HLA-B*57:01). However, many peptides will be bound by several HLA alleles, which can have quite distinct, possibly even opposing disease associations (e.g. peptides in overlap of *B57:01 and *B35:01). In this case, the disease-association of the peptide derives from the disease-associations of each of the binding HLA alleles as well as their frequencies in the dataset. The novel peptidome-wide association study (PepWAS) approach differentiates these distinct sets of peptides and identifies both specific peptides and epitope motifs with distinct disease-association (e.g. distinct motif of purple shaded peptides, corresponding to the dark purple cluster in Fig. 5). Circles depict repertoires of peptides (small pointed ovals) predicted to be bound by the given HLA allele. Overlap of circles defines sets of peptides bound by both HLA alleles. Color of circles and peptides depicts disease-association of corresponding HLA alleles and peptides, respectively, from blue (protective) to red (risk). The number of peptides in this schematic does not correspond to the actual number of peptides observed for these HLA alleles. In reality, the overlap among HLA alleles is substantially more complex than depicted in this simplified schematic.

A



B

Model parameters	HLA-B	HLA-A	HLA-B + HLA-A
Covariates	–	–	–
Covariates + Known CTL epitopes	1.8 (4.1×10^{-27})	0.5 (1.7×10^{-8})	2.3 (1.8×10^{-33})
Covariates + All HIV-1 epitopes	5.3 (2.1×10^{-77})	0.3 (1.8×10^{-6})	5.7 (4.9×10^{-82})
Covariates + Disease-associated epitopes	11.4 (5.1×10^{-170})	1.4 (2.9×10^{-20})	12.2 (5.3×10^{-180})

Fig. 2. Variation in viral load associated with predicted epitope repertoires bound by HLA-B and HLA-A.

Among HIV patients (N = 6,311), the proportion of variation (estimated as adjusted ΔR^2) in set point viral load (spVL) associated with the patient-specific number of predicted HLA-bound HIV-1 epitopes is shown separately for HLA-B and HLA-A, and for different epitope sets. (**A**) Previously, 11.4% and 0.9% of the variation in spVL had been associated with independent genetic variants in HLA-B and HLA-A, respectively (grey bars; data from ref. 4). Here we instead calculated the variation in spVL associated with individual HLA-bound HIV epitope repertoires (yellow bars), based on known CTL epitopes from Los Alamos HIV Molecular Immunology Database, all HLA-bound HIV epitopes, and only the disease-associated HIV epitopes (the latter corresponding to 99.2% of the variation previously associated with HLA genetic variation). (**B**) Variation associated with different sets of predicted epitopes. *P*-values (in parentheses) indicate the improvement over null model (covariates only: first five PCs and cohort group). Number of disease-associated predicted epitopes is 132 for HLA-B, and 74 for HLA-A, respectively.

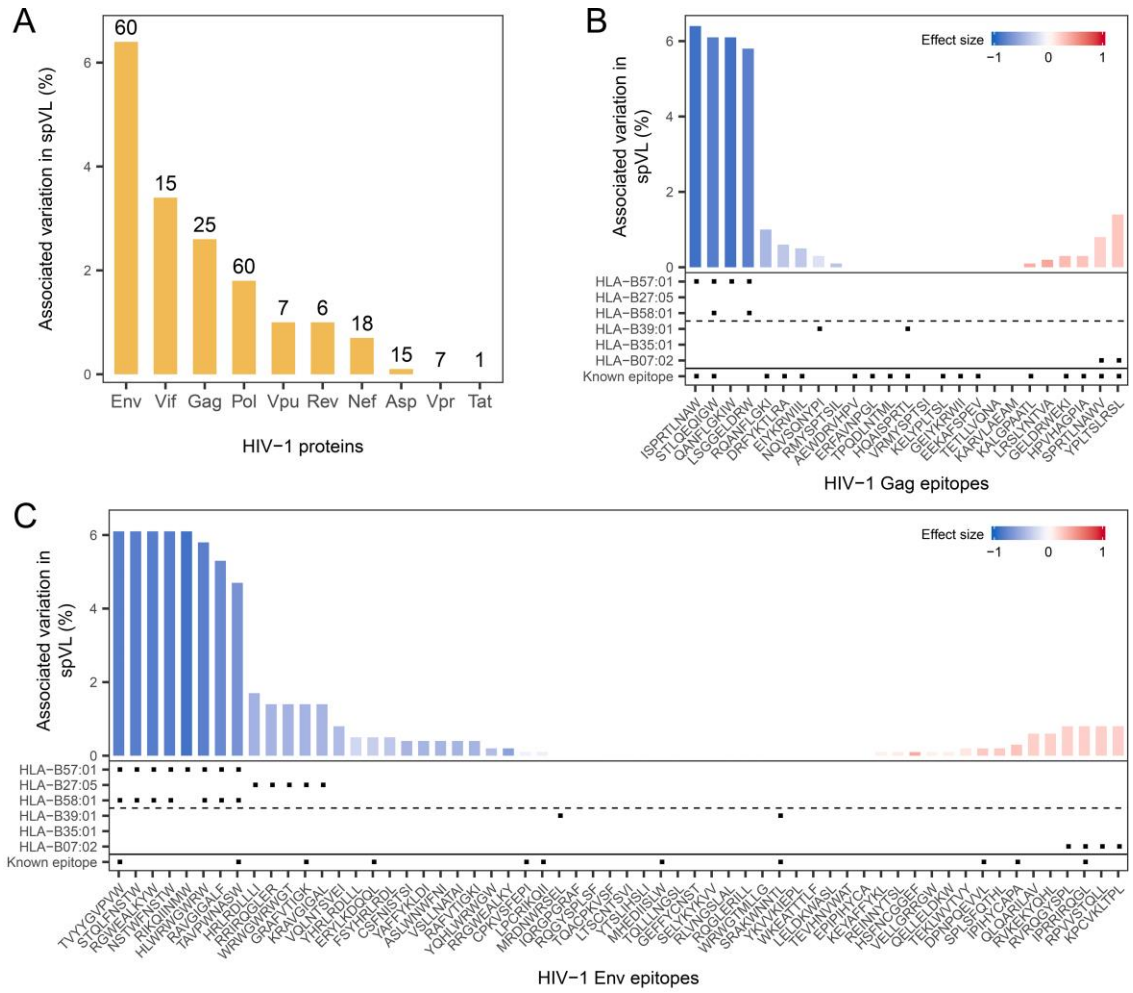


Fig. 3. Epitope- and protein-specific association with viral load.

(**A**) Percent of variation in spVL associated with all predicted epitopes of a given HIV-1 protein. Absolute number of predicted HLA-B bound epitopes per protein is shown above the bars. (**B-C**) Predicted HLA-B-bound epitopes accounted for varied levels of variation in set point viral load (spVL). Height of the bar represents the fraction of variation in spVL associated with each epitope, while the color reflects each epitope's effect on spVL, ranging from protection (blue) to risk (red). Note that epitope effects are estimated separately and are thus not independent. *Gag* (**B**) and *Env* (**C**) proteins are shown as representative examples, together with information on predicted binding for 3 protective and 3 risk HLA-B alleles highlighted in a recent review (24) and whether peptides are known epitopes in Los Alamos HIV database. All other HIV-1 proteins are shown in ***SI Appendix*, Fig. S8**.

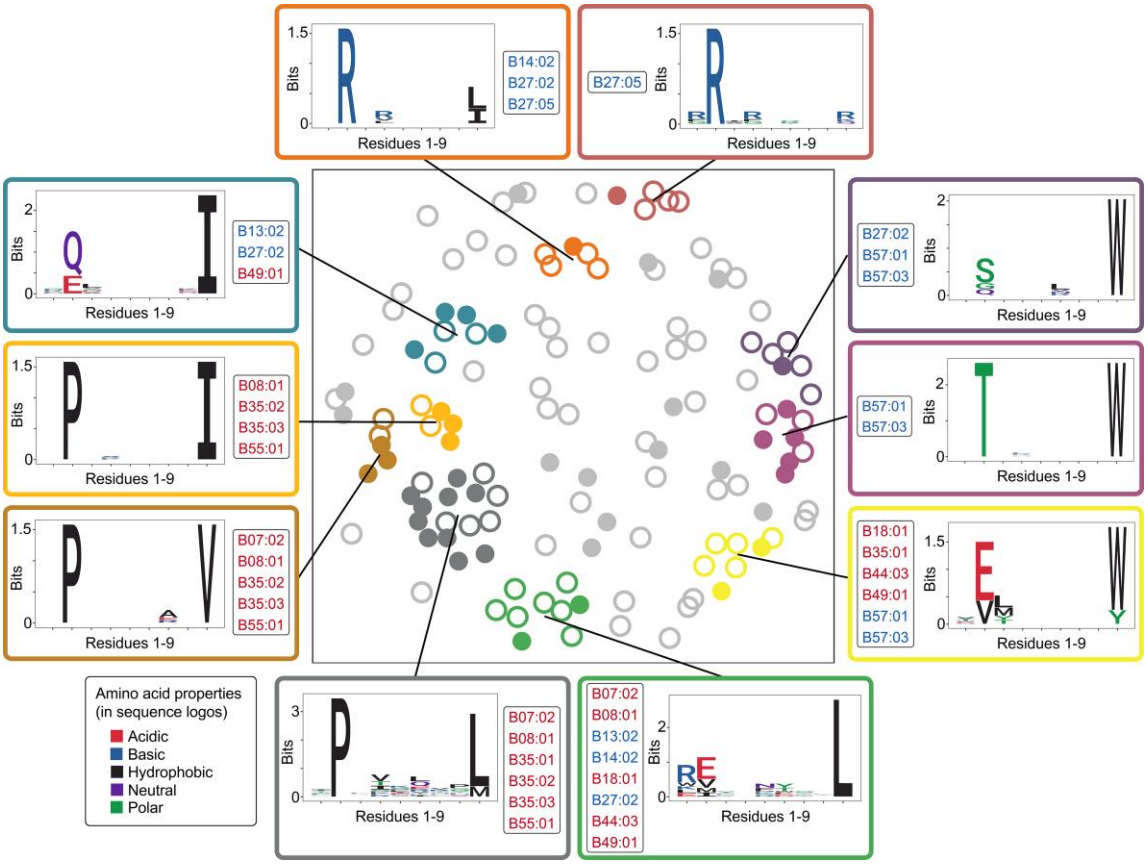


Fig. 4. Clusters of disease-associated epitopes.

Non-metric multidimensional scaling (NMDS) was used to visualize the pairwise distance between predicted HLA-B-bound disease-associated epitopes, which revealed 10 dominant clusters. Each circle represents an HLA-B bound disease- associated epitope (N = 132). Filled circles represent known CTL epitopes from the Los Alamos HIV Molecular Immunology Database (N = 45), while open circles represent previously uncharacterized disease-associated predicted epitopes. Cluster-specific motif and HIV-1 associated HLA-B alleles (N = 16) binding the cluster's epitopes are shown. The coloring of the allele names indicates disease-association of the specific alleles.