Associations of functional HLA class I groups with HIV viral load in a heterogeneous cohort.

Adrian G. ZUCCO[1], Marc BENNEDBÆK[2], Christina EKENBERG[1], Migle GABRIELAITE[3], Preston LEUNG[1], Mark N. POLIZZOTTO[4], Virginia KAN[5], Daniel D. MURRAY[1], Jens D. LUNDGREN[1] and Cameron R. MACPHERSON[1] for the INSIGHT START study group.

[1]PERSIMUNE Center of Excellence, Rigshospitalet, Copenhagen, Denmark

[2]Virus Research and Development Laboratory, Virus and Microbiological Special Diagnostics, Statens Serum Institut, Copenhagen, Denmark

[3]Center for Genomic Medicine, Copenhagen University Hospital, Copenhagen, Denmark.

[4]Clinical Hub for Interventional Research, College of Health and Medicine, The Australian National University, Canberra, Australia

[5]George Washington University, Veterans Affairs Medical Center, Washington D.C, U.S.A.

Corresponding author

Adrian Gabriel Zucco, MSc, PhD student

Tel: +45 35 45 57 75

mail: adrian.gabriel.zucco@regionh.dk

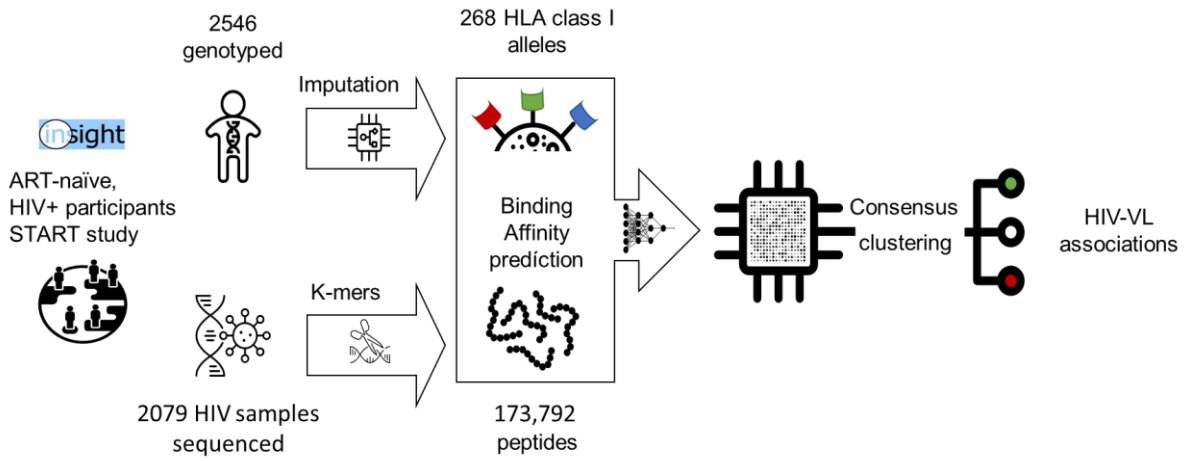
Rigshospitalet, Copenhagen University Hospital

Centre of Excellence for Health, Immunity and Infections (CHIP) & PERSIMUNE

Blegdamsvej 9, DK-2100 Copenhagen Ø, Denmark

Abstract

Human Leucocyte Antigen (HLA) class I alleles are the main host genetic factors involved in controlling HIV-1 viral load (VL). Nevertheless, HLA diversity has proven a significant challenge in association studies. We assessed how accounting for binding affinities of HLA class I alleles to HIV-1 peptides facilitate association testing of HLA with HIV-1 VL in a heterogeneous cohort from the Strategic Timing of AntiRetroviral Treatment (START) study. We imputed HLA class I alleles from host genetic data (2,546 HIV+ participants) and sampled immunopeptidomes from 2,079 host-paired viral genomes (targeted amplicon sequencing). We predicted HLA class I binding affinities to HIV-1 and unspecific peptides, grouping alleles into functional clusters through consensus clustering. These functional HLA class I clusters were used to test associations with HIV VL. We identified four clades totalling 30 HLA alleles accounting for 11.4% variability in VL. We highlight HLA-B*57:01 and B*57:03 as functionally similar but yet overrepresented in distinct ethnic groups, showing when combined a protective association with HIV+ VL (log, β -0.25; adj. p-value < 0.05). We further demonstrate only a slight power reduction when using unspecific immunopeptidomes, facilitating the use of the inferred functional HLA groups in other studies. The outlined computational approach provides a robust and efficient way to incorporate HLA function and peptide diversity, aiding clinical association studies in heterogeneous cohorts. To facilitate access to the proposed methods and results we provide an interactive application for exploring data.



Introduction

The Human Leucocyte Antigen (HLA) is a critical component of the host immune response. HLA Class I alleles mediate the anti-viral response through the presentation of intracellular viral peptides for recognition by Cytotoxic T cells. This mechanism is critical to the host's defence against diverse pathogens, which is why it is among the most genetically diverse regions in the human genome as evidenced by its association with our variable response to infectious disease [1]. In the context of HIV, HLA alleles are the canonical host genetic factors associated with viral load (VL)[2], altogether contributing up to 12% of its variability [3]. Viral diversity, on the other hand, is thought to explain two to four times as much (20%-46%) [4]. The combined relevance of both host and viral genomics suggests the necessity for the simultaneous analysis of both in the context of infectious disease[5]. Different approaches have been used to study this interaction such as genome-to-genome [6] or peptidome-wide associations [7]. In addition, when considering the interplay between host HLA alleles and viral peptides, their association could be explained functionally in terms of epitope binding and presentation.

Analysis of genetic variance within the HLA region and its associations with clinical outcomes is challenged by population-dependent distributions and the diversity of HLA haplotypes. Reaching statistical power within this region is difficult and while accounting for the effects of this important immunological region is necessary, it is often ignored altogether. Yet, the function of HLA is conserved as a core component of the immune response [8], [9]. This suggests there may be shared functional features within and between populations, even in the presence of genotypic plasticity. Understanding the structure of this mutual information could prove relevant for the study of host-pathogen interactions where high mutation rates are observed such as in HIV [10].

Immunopeptidomes have been used to estimate functional similarities between HLA alleles in what was initially denominated HLA supertypes [11]. These functional groups are based on the propensity of various alleles to bind similar sets of peptides. The algorithms used to estimate these binding profiles have gradually improved over time [12], [13]. However, the consideration of HLA functional groups in the context of Genome-Wide Association Studies (GWAS) has largely been overlooked [14].

In this study, we considered the ability of HLA proteins to functionally bind peptides leveraging the cost-effectiveness of high-throughput genotyping as opposed to directly assaying HLA function. We provide a framework based on state-of-the-art computational methods for the study of predicted immunopeptidomes and functional HLA groups in the context of HIV-1 infection. In doing so, we demonstrate the increased statistical power that is gained by moving away from a purely genetic approach to a functional one with implications for studying the immune response either directly or as a confounder. We processed HIV immunopeptidomes that incorporate intra-host viral diversity and imputed HLA alleles from the same geographically diverse cohort of people living with HIV (PLWH) and antiretroviral therapy (ART) naïve participants from the Strategic Timing of Anti-Retroviral Treatment (START) trial [16]. Interactive results of this work and complementary data are available via a web application (https://persimune-health-informatics.shinyapps.io/PAW2022Zucco__HLA_HIV_INSIGHT/).

Methods

Ethics

Host and viral samples in this study were extracted and analyzed from participants in the START clinical trial (NCT00867048) [16], conducted by the International Network for Strategic Initiatives in Global HIV Trials (INSIGHT) and the Community Programs for Clinical Research on AIDS (CPCRA). Written consent for the study

and genetic analyses were obtained from the participants and approved by participants' site ethics review committees.

HLA class I alleles

Imputation of HLA class I alleles (HLA-A, HLA-B and HLA-C) for 2546 genotyped ART-naïve, HIV+ participants was performed with HIBAG at 4-digit resolution [17]. Full details of the imputation process and quality control are described in previous publications [2], [18]. A multi-ethnic pre-trained model was used for imputation with a minimum out-of-bag accuracy of over 90% for all loci. HLA diversity was measured by the inverse Simpson index based on the HLA allele frequencies per locus and country. This index represents the complement of the probability that two participants would have the same HLA allele for a selected locus in a country.

Immunopeptidomes and binding affinity prediction

Plasma samples were obtained for a subset of 2079 ART-naïve, HIV+ participants from 21 countries enrolled in the START study. Viral RNA was sequenced, paired-end, using Illumina MiSeq and covered two amplicons in the HXB2 genome positioned 1485-5058 and 5967-9517. The sample preparation, library preparation, sequencing procedure, and detailed quality controls have been described previously [19]. Raw reads were fragmented into 27-mers using KAT [20] and those with a count higher than 1 were translated into peptide sequences of 9 amino acids length to fit the mean length of HLA Class I epitopes. Peptides were mapped to 10 major HIV proteins (Asp, Gag-Pol, Nef, Vpr, Vpu, gp160, Vif, Pr55, Rev, Tat) from NCBI RefSeq NC001802.1 (Supplementary file 1) using BLAST 2.8.1 (blastp-short). Hits with an E-value > 1E-05 were excluded to remove low-quality k-mers by considering exact matches. To compare HIV peptidomes with random sequences, a set of half-million 9-mers were generated by processing the same number of random protein sequences from Uniprot. Binding affinities for 268 class I HLA alleles to both HIV and random peptidomes using NetMHCpan 4.0 [15]. Three different immunopeptidome subsets were generated by (i) selecting peptides from the top 10% binding affinities, (ii) 10% most variable peptides, and (iii) peptides that potentially bind to at least 10% of the alleles using <500nM as general threshold [12].

Consensus clustering

Hierarchical clustering was implemented using two different linkage functions. Average linkage was used for measuring relationships between HLA alleles represented by dissimilarity defined as cosine, correlation and Euclidean distances. For clustering based on Ward linkage, cosine and correlation distances were corrected by the square root to satisfy the triangular inequality necessary to operate in Euclidean space [21]. To generate an ensemble of clustering solutions, we employed consensus clustering to mitigate bias from the subset, distance metric, or chosen linkage function [22]. A consensus matrix (C_{ij}) of size (n x n) was built where each element is the number of times an i^{th} allele clustered together with a j^{th} allele [23] at a varying total number of clusters selected (3 to 160). The consensus matrix was then processed by hierarchical clustering with average linkage after transforming the values into dissimilarity scores ($1 - C_{ij}$). Clustering was performed in Python 3.7.1 using the Scipy library [24].

Statistical analyses

Associations of $\log_{10}$-transformed VL with each node of the consensus tree were tested by linear regression and adjusted by sex, self-reported race, and country. Tested HLA alleles had to be present in more than 10% of the participants. Multiple testing was controlled by a Benjamini-Hochberg procedure using a q-value < 0.05 to identify associations. Analyses were performed in R v3.6.0 [25].

Data visualization and availability

Consensus clustering dendrograms and association coefficients were depicted using Interactive Tree of Life (iTOL) [26] and tanglegrams, which were generated by the *dendextend* R package [27]. We provide a flexible visualization of peptide-to-HLA binding profiles across viral proteins. Data downloads and access to supplementary materials are made available through the app (https://persimune-health-informatics.shinyapps.io/PAW2022Zucco__HLA_HIV_INSIGHT/).

Results

Baseline characteristics for the START cohort

We used baseline data and the genotypes of 2,546 participants from the START trial. Next-generation sequencing of HIV samples was retrieved for a subset of 2,079 participants. All participants included in the trial were asymptomatic HIV+ and ART-naïve with two CD4+ cell counts >500/µL at least 14 days apart within 60 days of enrollment in the trial. Baseline characteristics for study participants can be found in Table 1.

Prediction of patient-derived HIV immunopeptidomes

Our approach for immunopeptidome generation initially yielded 9.88×10^7 peptide 9-mers. After filtering and mapping with BLAST to ten major HIV proteins, a total of 173,792 9-mers were considered. This accounted for a 99.22% coverage across the reference proteome. Among the final list, we observed 136 best-defined CTL/CD8+ epitopes from Los Alamos (version 2019-11-20; Supplementary file 2).

Data exploration

To facilitate the exploration of the results, a web application was developed (Figure 1) providing tools to navigate interactively the global and local diversity of imputed HLA alleles, HIV subtype-derived peptides, and their corresponding binding profiles. Shannon and Simpson diversity indices can be directly visualized on the world map highlighting the geographical diversity of the cohort. This interactive tool facilitates the further examination of specific HLA allele, HIV subtype, or peptide frequencies for the entire cohort and options to explore details at a country level.

Higher HLA-A diversity is associated with a lower mean viral load per country

The diversity of HLA class I alleles measured in terms of the inverse Simpson index was calculated using the HLA allele frequencies per locus and country for comparison. A negative univariate correlation ($R = -0.65$ $p = 0.0018$) between HLA-A diversity and mean HIV $\log_{10}$(viral load) per country was found. This indicates that countries in our cohort with a high diversity of HLA-A alleles would show lower levels of HIV viral load in the population represented in terms of mean $\log_{10}$(VL). No significant correlation was observed for HLA-B and HLA-C respectively (Figure 2).

Consensus clustering of HIV-derived immunopeptidomes improves statistical power compared to random immunopeptidomes

Consensus trees were generated based on a half-million random peptides from Uniprot and the predicted HIV immunopeptidome under the same methodology. The correlation between both trees was 0.985 indicating high similarity. Minor differences were found in allele distances (Figure S1) with only a few major structural changes such as the displacement of B*35:20 from a predominantly B*15 node (unspecific immunopeptidome) to one dominated by B*35 alleles (HIV immunopeptidome). Another two examples are HLA-B*15:13 and B*15:58 which moved from a node of predominantly HLA-C alleles (unspecific

immunopeptidome) to a node of HLA-B alleles (HIV immunopeptidome). These small differences, when combined, were translated into an increase in statistical power (Figure S2). Overall, when assessed in a multivariate model including all HLA class I alleles as covariates the combined percentage of explained variance accounted for 11.44% of the VL after adjustment for sex, self-reported race, and country. Hence, our clusters of HLA function accounted for HIV-VL variance to a similar degree as reported by Bartha et al., 2017 [3] for genotypic evidence in a comparatively pure European cohort. This highlights that accounting for HLA function is both sufficient and better to explain variability in VL in heterogeneous populations.

HLA class I functional nodes associations with HIV-VL

Associations between the HLA functional nodes and the measurements of HIV-VL taken at study entry were tested using linear regression. Four nodes with a consistent effect size were associated with HIV-VL (Figure 3). These nodes were composed of 30 HLA class I alleles of which 11 were observed in participants with measurable VL at study entry. Two functional nodes were associated with a lower VL, one group composed of HLA-B*57:01, B*58:01, B*57:02, and B*57:03 (β -0.25, q-value 7.02E-06) and the second group composed of a pair of HLA-C*08 alleles, HLA-C*08:04 and C*08:01 (β -0.29, q-value 0.042). In contrast, two nodes showed an association with higher VL, one cluster composed of six HLA-B*44 alleles, B*44:05, B*44:08, B*44:04, B*44:03, B*44:02, B*44:27 (β 0.15, q-value 0.003) and a mixed group composed of 16 alleles: B*35:20, B*35:16, B*35:10, B*35:43, B*35:08, B*35:19, B*35:41, B*35:01, B*35:17, B*35:05, B*44:06, B*56:03, B*53:01, B*15:08 and B*15:11 (β 0.13, q-value 0.048). From these alleles, only two (HLA-B*57:01 and B*57:03) were associated with HIV-VL when tested at the individual allele level (Figure 3).

Discussion

In this study, we implemented a computational approach based on consensus clustering of HLA alleles using predicted immunopeptidomes to explore associations of functional HLA groups in a geographically diverse cohort of PLWH. We defined functional similarity as differences in epitope binding profiles between HLA alleles and performed HLA imputation on patients enrolled in the START study. We combined the host genetic information with viral genomics through the prediction of immunopeptidomes. Viral sequences derived from a subset of participants were processed into peptides to predict binding affinities to 268 HLA class I alleles which allowed us to generate distance matrices of HLA alleles. Consensus clustering allowed us to agglomerate HLA alleles into nodes by their functionality. Four nodes were found to be associated with HIV-VL, implicating alleles that could not be detected when performing independent allele-specific tests alone. These four nodes accounted for a total of 30 HLA alleles of which 11 were observed in our cohort.

The effects differed among the four nodes associated with HIV-VL. One node containing four well-characterized alleles HLA-B*57:01, B*58:01, B*57:02, and B*57:03 showed an association with lower HIV-VL indicating a protective effect. This effect has been previously reported on an individual allele level [28]; we confirmed and observed shared binding affinity profiles that now indicate their common mode of function. While having a similar effect in HIV-VL, these four alleles are represented at different frequencies between populations. For example, HLA-B*57:01 carriers are of European descent compared to B*57:03 which is predominant among those of African descent [2]. The similarity in function was captured by our analysis despite the differences in allele frequencies. A second protective node contained two alleles, HLA-C*08:04 and C*08:01, which have been reported in an admixed population and detected after adjusting for multiple factors such as CD4+ and CD8+ counts [29]. We showed that such association could not only be detected in our study but that functional clustering provided a clear and simple method to do it with greater mechanistic insight. We report a node of six HLA-B*44 alleles, in this case, associated with a higher VL. This finding contradicts a previous study in a Chinese cohort [30] in which a few alleles of the cluster were found to have

a protective effect. Given the large effect of viral diversity on VL, this could be a result of regional adaptation that could not be accounted for in our study due to the lack of participants from the same region. While this suggests a weakness in our study due to the underrepresentation of some ethnicities, divergence in our results from studies on homogeneous populations could be a useful tool to detect localized effects. However, data from a cohort with sufficient Chinese representation is needed to confirm the findings.

Defining discrete nodes of HLA alleles is challenging as distances between alleles are based on predicted binding affinities and measured continuously. Therefore, metrics to define a fixed number of clusters failed to suggest a robust threshold for cutting the trees obtained by consensus clustering. For this reason, the final HIV-associated groups of alleles were determined by assessing the trees hierarchically from leaves (individual HLA alleles) to roots (broad HLA nodes), thus optimizing groups based on their association with VL when the combination of alleles in all child nodes supported it. In this way, the methodology implemented provides an advantage to traditional HLA association studies by adaptively increasing the signal of functional groups allowing to uncover associations that could not be found at the individual allele level due to the number of statistical comparisons or sample size [14]. Grouping or clustering of variables to increase statistical power and limit multiple testing is common practice in epidemiological studies. The functional clustering presented here, however, is based on the assumption that the majority of HLA functionality is mediated by peptide binding affinities. This serves to facilitate study design by simplifying decisions on the size and diversity of the cohort while allowing a biological interpretation of the results. For example, it enables the inference of the effect of unobserved HLA alleles on HIV-VL given their common function to those observed. These are notable challenges across clinical studies attempting to measure the effect on the immune response. Altogether, the functional clustering approach presented is neither specific to HIV nor viral load and may be transferred to study HLAs and outcomes in other host-pathogen interactions.

Similar computational approaches were proposed to explore HLA class I molecules and their role in HIV-1 infection: These studies mainly focused on genome-to-genome approaches [6], techniques to find new epitopes in European participants using a peptidome approach [7] or to explore the interactions of HLA class I molecules to other ligands [31]. In contrast, we focused on the host factors, the HLA class I alleles, accounting for the high diversity of the cohort and inclusion of a larger number of viral sequences to predict immunopeptidomes and expand on genome-to-genome analyses previously performed in our same cohort [19]. While implementations of HLA clustering in the context of HIV-1 have been developed based on a Bayesian framework [13] these were limited by the diversity of their data and assumptions of the clustering parameters. Extra considerations must be taken when clustering predicted immunopeptidomes to avoid bias introduced by predominantly low binding affinity predictions, which can affect the reliability of the computed distances between HLA alleles. To solve this, we used consensus clustering by computing different subsets of immunopeptidomes and avoiding bias from non-binders when calculating distances between HLA class I alleles under multiple linkage functions. The ensemble of implemented techniques avoids bias due to high dimensionality and has shown success when applied to clustering of biological data [22].

We propose diverse methodological improvements in the analysis of host-viral genetics. From the host genetics perspective, we showed that imputed HLA class I alleles can be used to calculate HLA diversity without requiring full HLA sequences. From the viral genetics perspective, we incorporated viral sequences using a fast k-mer approach to avoid generating a consensus sequence, especially for pathogens of high genetic variability and to take into account the intra-host viral diversity [32]. When combining host and viral information through predicted binding affinities, we showed that the differences between clustering based on specific HIV and random immunopeptidomes facilitate a slightly higher resolution of the trees. We suggest that the unspecific HLA clusters from the latter approach could be used for other infectious phenotypes. This

is likely due to the diversity of viral genomes present in our dataset, and this may not be the case for smaller, more homogeneous cohorts. However, it suggests that our immunopeptidomes could work as a proxy for other similarly diverse studies that do not have access to paired viral genomes. Alternatively, new immunopeptidomes may be proposed based entirely on synthetic data at the risk of increased type-II error.

While we demonstrate the utility of the implemented methodology on HIV, there is also a clear road to extend it to the analysis of other pathogens eliciting class-I immune responses. Using this common functional framework, there is scope to analyze the variation of HLA structure in populations and within the context of multiple pathogens. Another possible application could be as a screening tool for the most effective peptides either for targeting populations carrying specific allelic distributions or maximizing coverage of vaccines across geographies. Finally, we propose that the focused consideration of HLA function in terms of peptide binding affinities provides a promising approach to inform modern vaccine design. This would also be relevant in the advent of mRNA vaccines coming to fruition after decades of research [33] as they take advantage of proper peptide selection. As an example, the need for a new class of tools that account for both variable immunogenic coverage and clinically relevant mutations has been highlighted during the recent SARS-CoV-2 pandemic [34]. To facilitate open research, the results of this work are made available for common use via a web application (https://persimune-health-informatics.shinyapps.io/PAW2022Zucco__HLA_HIV_INSIGHT/)

Acknowledgements

We would like to thank all participants in the START trial and all trial investigators. See N Engl J Med 2015; 373:795–807[16] for the complete list of START investigators.

Author contributions

A.G.Z., J.D.L. and C.R.M conceived the study. A.G.Z., M.B, M.G. and C.E prepared the data. A.G.Z performed the statistical and computational analyses. A.G.Z. and C.R.M drafted the manuscript. All authors contributed to data interpretation, critically revised the manuscript, and approved the final version.

Conflict of interest

There are no conflicts of interest to disclose.

Sources of funding

This study was supported by the Danish National Research Foundation (DNRF126) and the National Institute of Allergy and Infectious Diseases, Division of Clinical Research and Division of AIDS (National Institutes of Health grants UM1-AI068641, UM1-AI120197 and U01-AI136780). The START trial was supported by the National Institute of Allergy and Infectious Diseases, National Institutes of Health Clinical Center, National Cancer Institute, National Heart, Lung, and Blood Institute, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institute of Mental Health, National Institute of Neurological Disorders and Stroke, National Institute of Arthritis and Musculoskeletal and Skin Diseases, Agence Nationale de Recherches sur le SIDA et les Hépatites Virales (France), National Health and Medical Research Council (Australia), National Research Foundation (Denmark), Bundes Ministerium für Bildung und Forschung (Germany), European AIDS Treatment Network, Medical Research Council (United Kingdom), National Institute for Health Research, National Health Service (United Kingdom), and the University of Minnesota. Antiretroviral drugs were donated to the central drug repository by AbbVie, Bristol-Myers Squibb, Gilead Sciences, GlaxoSmithKline/ViiV Healthcare, Janssen Scientific Affairs, and Merck.

References

[1] C. Tian *et al.*, "Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections," *Nature Communications*, vol. 8, no. 1, p. 599, Sep. 2017, doi: 10.1038/s41467-017-00257-5.

[2] C. Ekenberg *et al.*, "Association Between Single-Nucleotide Polymorphisms in HLA Alleles and Human Immunodeficiency Virus Type 1 Viral Load in Demographically Diverse, Antiretroviral Therapy–Naive Participants From the Strategic Timing of AntiRetroviral Treatment Trial," *J Infect Dis*, vol. 220, no. 8, pp. 1325–1334, Sep. 2019, doi: 10.1093/infdis/jiz294.

[3] I. Bartha, P. J. McLaren, C. Brumme, R. Harrigan, A. Telenti, and J. Fellay, "Estimating the Respective Contributions of Human and Viral Genetic Variation to HIV Control," *PLOS Computational Biology*, vol. 13, no. 2, p. e1005339, Feb. 2017, doi: 10.1371/journal.pcbi.1005339.

[4] C. Fraser *et al.*, "Virulence and Pathogenesis of HIV-1 Infection: An Evolutionary Perspective," *Science*, vol. 343, no. 6177, Mar. 2014, doi: 10.1126/science.1243727.

[5] P. J. McLaren and M. Carrington, "The impact of host genetic variation on infection with HIV-1," *Nature Immunology*, vol. 16, no. 6, pp. 577–583, Jun. 2015, doi: 10.1038/ni.3147.

[6] I. Bartha *et al.*, "A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control," *eLife Sciences*, vol. 2, p. e01123, Oct. 2013, doi: 10.7554/eLife.01123.

[7] J. Arora, P. J. McLaren, N. Chaturvedi, M. Carrington, J. Fellay, and T. L. Lenz, "HIV peptidome-wide association study reveals patient-specific epitope repertoires associated with HIV control," *PNAS*, vol. 116, no. 3, pp. 944–949, Jan. 2019, doi: 10.1073/pnas.1812548116.

[8] A. Aflalo and L. H. Boyle, "Polymorphisms in MHC class I molecules influence their interactions with components of the antigen processing and presentation pathway," *International Journal of Immunogenetics*, vol. n/a, no. n/a, doi: 10.1111/iji.12546.

[9] S. Buhler, J. M. Nunes, and A. Sanchez-Mazas, "HLA class I molecular variation and peptide-binding properties suggest a model of joint divergent asymmetric selection," *Immunogenetics*, vol. 68, no. 6, pp. 401–416, Jul. 2016, doi: 10.1007/s00251-016-0918-x.

[10] J. M. Carlson, A. Q. Le, A. Shahid, and Z. L. Brumme, "HIV-1 adaptation to HLA: a window into virus–host immune interactions," *Trends in Microbiology*, vol. 23, no. 4, pp. 212–224, Apr. 2015, doi: 10.1016/j.tim.2014.12.008.

[11] A. Sette and J. Sidney, "HLA supertypes and supermotifs: a functional perspective on HLA polymorphism," *Current Opinion in Immunology*, vol. 10, no. 4, pp. 478–482, Aug. 1998, doi: 10.1016/S0952-7915(98)80124-6.

[12] M. Thomsen, C. Lundegaard, S. Buus, O. Lund, and M. Nielsen, "MHCcluster, a method for functional clustering of MHC molecules," *Immunogenetics*, vol. 65, no. 9, pp. 655–665, Sep. 2013, doi: 10.1007/s00251-013-0714-9.

[13] C. van Dorp and C. Kesmir, "Estimating HLA disease associations using similarity trees," *bioRxiv*, p. 408302, Sep. 2018, doi: 10.1101/408302.

[14] A. E. Kennedy, U. Ozbek, and M. T. Dorak, "What has GWAS done for HLA and disease associations?," *International Journal of Immunogenetics*, vol. 44, no. 5, pp. 195–211, Oct. 2017, doi: 10.1111/iji.12332.

[15] V. Jurtz, S. Paul, M. Andreatta, P. Marcatili, B. Peters, and M. Nielsen, "NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data," *The Journal of Immunology*, vol. 199, no. 9, pp. 3360–3368, Nov. 2017, doi: 10.4049/jimmunol.1700893.

[16] Insight Start Study Group, "Initiation of Antiretroviral Therapy in Early Asymptomatic HIV Infection," *New England Journal of Medicine*, vol. 373, no. 9, pp. 795–807, Aug. 2015, doi: 10.1056/NEJMoa1506816.

[17] X. Zheng *et al.*, "HIBAG—HLA genotype imputation with attribute bagging," *The Pharmacogenomics Journal*, vol. 14, no. 2, pp. 192–200, Apr. 2014, doi: 10.1038/tpj.2013.18.

[18] C. Ekenberg *et al.*, "The association of human leukocyte antigen alleles with clinical disease progression in HIV-positive cohorts with varied treatment strategies," *AIDS*, vol. 35, no. 5, pp. 783–789, Apr. 2021, doi: 10.1097/QAD.0000000000002800.

[19] M. Gabrielaite *et al.*, "Human immunotypes impose selection on viral genotypes through viral epitope specificity," *The Journal of Infectious Diseases*, no. jiab253, May 2021, doi: 10.1093/infdis/jiab253.

[20] D. Mapleson, G. Garcia Accinelli, G. Kettleborough, J. Wright, and B. J. Clavijo, "KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies," *Bioinformatics*, vol. 33, no. 4, pp. 574–576, Feb. 2017, doi: 10.1093/bioinformatics/btw663.

[21] S. van Dongen and A. J. Enright, "Metric distances derived from cosine similarity and Pearson and Spearman correlations," *arXiv:1208.3145 [cs, stat]*, Aug. 2012, Accessed: Jan. 11, 2019. [Online]. Available: http://arxiv.org/abs/1208.3145

[22] T. Ronan, Z. Qi, and K. M. Naegle, "Avoiding common pitfalls when clustering biological data," *Sci. Signal.*, vol. 9, no. 432, pp. re6–re6, Jun. 2016, doi: 10.1126/scisignal.aad1932.

[23] A. Strehl and J. Ghosh, "Cluster Ensembles — a Knowledge Reuse Framework for Combining Multiple Partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, Mar. 2003, doi: 10.1162/153244303321897735.

[24] P. Virtanen *et al.*, "SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python," *arXiv e-prints*, p. arXiv:1907.10121, Jul. 2019.

[25] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2019. [Online]. Available: https://www.R-project.org/

[26] I. Letunic and P. Bork, "Interactive Tree Of Life (iTOL) v4: recent updates and new developments," *Nucleic Acids Res*, vol. 47, no. W1, pp. W256–W259, Jul. 2019, doi: 10.1093/nar/gkz239.

[27] T. Galili, "dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering," *Bioinformatics*, vol. 31, no. 22, pp. 3718–3720, Nov. 2015, doi: 10.1093/bioinformatics/btv428.

[28] P. J. R. Goulder and B. D. Walker, "HIV and HLA Class I: An Evolving Relationship," *Immunity*, vol. 37, no. 3, pp. 426–440, Sep. 2012, doi: 10.1016/j.immuni.2012.09.005.

[29] H. Valenzuela-Ponce *et al.*, "Novel HLA class I associations with HIV-1 control in a unique genetically admixed population," *Scientific Reports*, vol. 8, no. 1, p. 6111, Apr. 2018, doi: 10.1038/s41598-018-23849-7.

[30] X. Zhang *et al.*, "HLA-B*44 Is Associated with a Lower Viral Set Point and Slow CD4 Decline in a Cohort of Chinese Homosexual Men Acutely Infected with HIV-1," *Clin Vaccine Immunol*, vol. 20, no. 7, pp. 1048–1054, Jul. 2013, doi: 10.1128/CVI.00015-13.

[31] B. J. Debebe *et al.*, "Identifying the immune interactions underlying HLA class I disease associations," *eLife*, vol. 9, p. e54558, Apr. 2020, doi: 10.7554/eLife.54558.

[32] J. Fellay and V. Pedergnana, "Exploring the interactions between the human and viral genomes," *Hum Genet*, Nov. 2019, doi: 10.1007/s00439-019-02089-3.

[33] N. A. C. Jackson, K. E. Kester, D. Casimiro, S. Gurunathan, and F. DeRosa, "The promise of mRNA vaccines: a biotech and industrial perspective," *npj Vaccines*, vol. 5, no. 1, Art. no. 1, Feb. 2020, doi: 10.1038/s41541-020-0159-8.

[34] G. Liu, B. Carter, and D. K. Gifford, "Predicted Cellular Immunity Population Coverage Gaps for SARS-CoV-2 Subunit Vaccines and Their Augmentation by Compact Peptide Sets," *Cell Systems*, vol. 12, no. 1, pp. 102-107.e4, Jan. 2021, doi: 10.1016/j.cels.2020.11.010.

Table 1. Demographic characteristics of the START cohort at study entry.

Characteristics	Genotyped participants (n=2,546)
Median age (IQR) (years)	36 (29–45)
Sex [*n* (%)]	
Female	511 (20.1)
Male	2035 (79.9)
Race/ethnic group [*n* (%)]	
Black	577 (22.7)
Hispanic	498 (19.6)
Asian	26 (1.0)
White	1404 (55.2)
Other	41 (1.6)
Geographical region [*n* (%)]	
Africa	343 (13.5)
Asia	0 (0)
Australia and New Zealand	96 (3.8)
Europe and Israel	1148 (45.1)
Latin America	499 (19.6)
United States and Canada	460 (18.1)
Mode of HIV-infection [*n* (%)]	
Sexual contact	
MSM	1633 (64.1)
With a person of the opposite sex	751 (29.5)
Injection-drug use	45 (1.8)
Other	117 (4.6)
Median time since HIV diagnosis (IQR) (years)	1 (0–3)
ART-naïve [*n* (%)]	2546 (100)
Median CD4+ T-cell count (IQR) (cells/µl)	651 (585–759)
Median HIV viral load (IQR) (copies/ml)	14 833 (3503–46 000)

IQR, interquartile range; MSM, men who have sex with men; ART, Antiretroviral treatment.

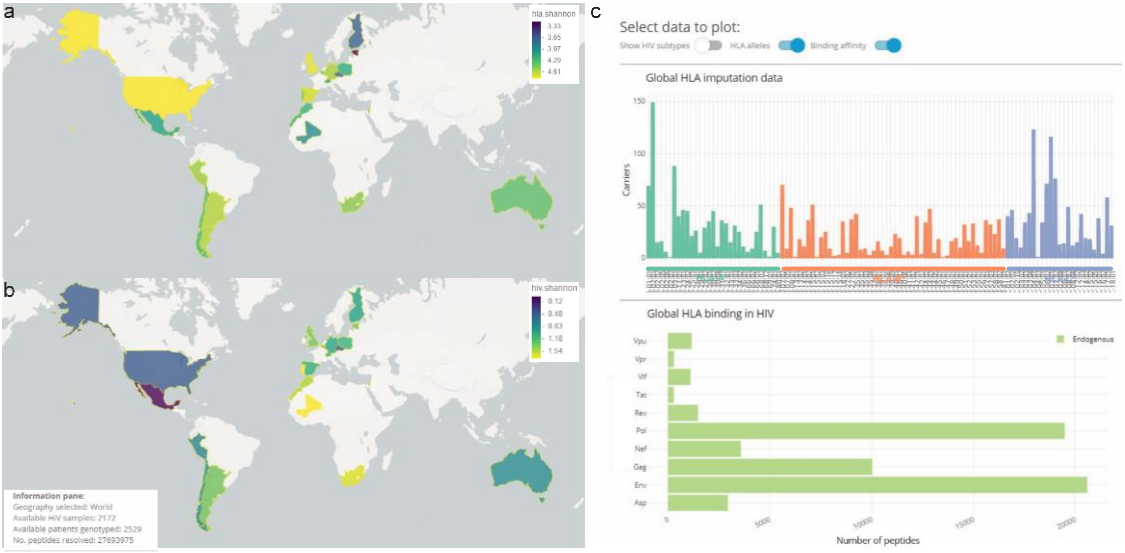


Figure 1. Web application demonstrating a subset of available data visualizations.

Panel (a) and (b) illustrates the global diversity of imputed HLA alleles in Shannon and Simpson indices, respectively. Darker (navy) colors indicate higher diversity. Panel (c) showcases the interactive component to further examine specific HLA allele, an HIV subtype and/or peptide frequencies for the selected cohort with options to explore individual countries in detail.

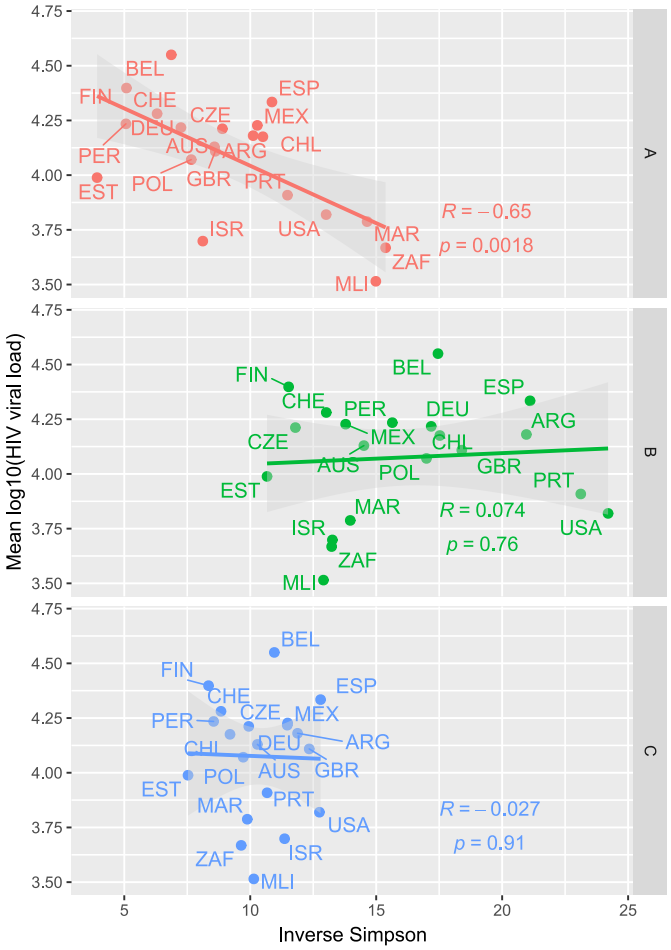


Figure 2. Correlations between HIV viral load and HLA diversity per country for three HLA loci.

Mean $\log_{10}$(HIV-VL) was depicted (y-axis) as dot plots against HLA diversity measured in terms of Simpson index (x-axis) for three HLA class I loci (HLA-A, -B, -C) in 21 countries. The noted R corresponds to a Pearson correlation with its corresponding p-value.

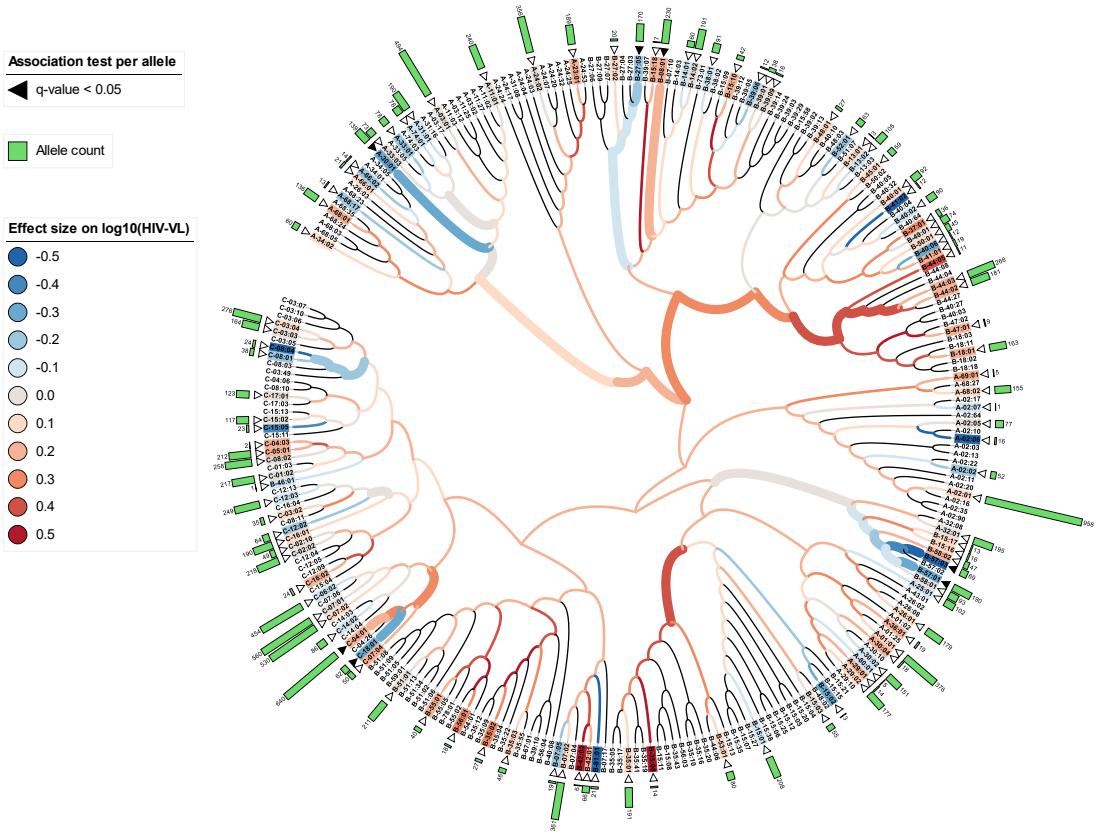


Figure 3. Dendrogram of 268 HLA class I alleles based on consensus clustering of predicted binding affinities to HIV peptides.

Predicted binding affinities to 173,792 HIV peptides were used to calculate the HLA allele distances used for consensus clustering and represented as a dendrogram through hierarchical clustering. Associations with $\log_{10}$(HIV-VL) of each node (HLA functional node) and leaves (HLA alleles) in the dendrogram were tested and adjusted by sex, self-reported race, and country. Associations were defined by an adjusted p-value (Benjamini-Hochberg) < 0.05 and are represented as thick branches for nodes and black triangles for leaves. White triangles indicate HLA alleles detected in our cohort. The effect of the respective associations is color-coded from protective effect (blue) to detrimental (red). On the outer ring, HLA allele counts are depicted as green bars. An interactive version of the tree can be found in the provided web application.