Human immunotypes impose selection on viral genotypes through viral epitope specificity

Migle Gabrielaite[1,#], Marc Bennedbæk[2,#], Adrian G. Zucco[2], Christina Ekenberg[2], Daniel D. Murray[2], Virginia L. Kan[3], Giota Touloumi[4], Linos Vandekerckhove[5], Dan Turner[6], James Neaton[7], H. Clifford Lane[8], Sandra Safo[9], Alejandro Arenas-Pinto[10], Mark N. Polizzotto[11], Huldrych F. Günthard[12,13], Jens D. Lundgren[2], Rasmus L. Marvig[1,*]; for the INSIGHT START trial group

[1.]Centre for Genomic Medicine, Copenhagen University Hospital, Copenhagen, Denmark, [2.]Centre of Excellence for Health, Immunity and Infections, Department of Infectious Diseases, Rigshospitalet, University of Copenhagen, Denmark, [3.]Veterans Affairs Medical Center, The George Washington University School of Medicine and Health Sciences, Washington, District of Columbia, USA, [4.]Department of Hygiene, Epidemiology and Medical Statistics, Medical School, National and Kapodistrian University of Athens, Athens, Greece, [5.]HIV Cure Research Center, Department of Internal Medicine and Pediatrics, Faculty of Medicine and Health Sciences, Ghent University; Ghent University Hospital, Ghent, Belgium, [6.]Crusaid Kobler AIDS Center, Tel-Aviv Sourasky Medical Center, Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel, [7.]School of Public Health, University of Minnesota, Minneapolis, USA, [8.]Division of Clinical Research, National Institute of Allergy and Infectious Diseases, NIH, Bethesda, Maryland, USA, [9.]Division of Biostatistics, University of Minnesota, Minneapolis, USA, [10.]MRC Clinical Trials Unit, University College London, London, United Kingdom, [11.]Kirby Institute for Infection and Immunity, University of New South Wales, Sydney, Australia, [12.]Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zürich, Zürich, Switzerland, [13.]Institute of Medical Virology, University of Zürich, Zürich, Switzerland.

* Correspondence to Rasmus L. Marvig rasmus.lykke.marvig@regionh.dk

# Contributed equally

Summary

We analyzed HIV and human host genetics in concert to resolve the interplay of viral and human genetics during infection. We showed that HIV genetic variants associated with different HLA alleles, and that genetic interactions affected viral load.

Abstract

Background: Understanding the genetic interplay between human hosts and infectious pathogens is crucial for how we interpret virulence factors. Here, we tested for associations between HIV and host genetics, and interactive genetic effects on viral load (VL) in HIV+ ART-naive clinical trial participants.

Methods: HIV genomes were sequenced and the encoded amino acid (AA) variants were associated with VL, human single nucleotide polymorphisms (SNPs) and imputed HLA alleles, using generalized linear models with Bonferroni correction.

Results: Human (388,501 SNPs) and HIV (3,010 variants) genetic data was available for 2,122 persons. Four HIV variants were associated with VL (p-values<1.66×10^{-5}). Twelve HIV variants were associated with a range of 1–512 human SNPs (p-value<4.28×10^{-11}). We found 46 associations between HLA alleles and HIV variants (p-values<1.29×10^{-7}). We found HIV variants and immunotypes when analyzed separately, were associated with lower VL, whereas the opposite was true when analyzed in concert. Epitope binding prediction showed HLA alleles to be weaker binders of associated HIV AA variants relative to alternative variants on the same position.

Conclusions: Our results show the importance of immunotype specificity on viral antigenic determinants, and the identified genetic interplay puts emphasis that viral and human genetics should be studied in the context of each other.

Keywords: HIV; GWAS; genome-wide association study; viral genomics; host genomics; genome-to-genome analysis; viral load

Introduction

Human leukocyte antigens (HLA) present peptides from invading pathogens to help the generation of an immune response [1]. Genetic variants of HLA genes, i.e. different alleles, show differences in peptide presentation which translates into non-identical immune response towards a given pathogen, and some HLA alleles will be selectively beneficial for the host to fight infection. Vice versa, pathogens are genetically diverse, and genetic variants are selected to evade the human immune response [2], but also to optimize other phenotypes such as antiretroviral therapy resistance [3–5]. Accordingly, it has been found that the genetic variation of HIV-1 is partly shaped by which HLA is encountered within the host [6]. Similarly, HLAs expressed by the host affect the replicative rate of the virus: some HLAs or even their functional groups are known to be protective while others are associated with worse outcomes [7,8]. The proportion of variation in the HIV genome attributable directly to the interaction with the host genetics remains unclear. Furthermore, the molecular specificity of the interplay between viral genetic escape and host genomic control remains poorly defined [9].

In principle, the ability of HIV to replicate in the body can be quantified by the plasma viral load (VL): the key determinant to identify the risk of disease progression and death [3]. Large variations in VL are observed in HIV+ persons and a sizable fraction hereof in early HIV infection is assumed to be due to variation in viral replicative ability which is also influenced by the host immune response [10]. A fraction of variance in HIV replication and disease progression is attributable to viral [10–14] or host genetics [15–18]. However, viral and host genetic interactions are poorly understood [9]. Prior HIV and host genetic association studies encompassed data from persons with similar genetic background or individual subtypes of HIV [19–21]. Expanding to a cohort with multiple viral subtypes and diverse host demographics could elucidate specific factors that influence HIV pathogenesis or generalize previous findings from more homogenous cohorts.

In this study we performed four analyses (Figure 1): 1) Association of HIV AA variants with VL to estimate the effect of viral genetic variation on VL in HIV infection, 2) Association of HIV AA variants with human single nucleotide polymorphisms (SNPs) to identify genome-to-genome associations, 3) Association of HIV AA variants with HLA alleles to identify HLA alleles with the strongest association with HIV genetic variants, and 4) bivariate association and interaction of HIV AA variants and HLA alleles on VL to resolve the interplay of viral and human genetics during the HIV infection. These associations were performed in a demographically diverse cohort of treatment-naive HIV-positive persons with different HIV subtypes from the Strategic Timing of Antiretroviral Treatment (START) cohort [22].

Materials and methods

Study population

Samples analysed in this study were derived from participants from the START trial, conducted by the International Network for Strategic Initiatives in Global HIV Trials, which included 4,685 antiretroviral therapy-naive and asymptomatic participants [22] who had a $CD4^+$ cell count >500cells/µl at baseline, no history of AIDS or of antiretroviral treatment and were above 18-years-old at study entry. For this study, we included participants which 1) had baseline plasma samples with a VL measurement ≥1,000 copies/mL and viral genomes sequenced by next-generation sequencing (N=3,785) and 2) consented to human genotyping (N=2,546). All samples were derived from participants who provided written informed consent, and the study was reviewed by participant site ethics review committees (NCT00867048) [22].

HIV sequencing, alignment and variant calling

The detailed laboratory procedure is described in the supplementary materials and methods; in short, viral RNA was extracted from plasma and amplified using two amplicons spanning 7,125nt of the 9,719nt (74%) HIV genome and sequenced by next-generation sequencing. Reads were first

aligned against the GRCh37.p13 human reference genome with Bowtie2 v2.2.8 to retain read pairs with neither read aligning to the human genome. Virvarseq [23] was used for alignment of cleaned reads and AA variant calling with the HXB2 HIV-1 reference genome (GenBank accession number K03455.1) with default parameters. Virvarseq alignment, realignment and AA variant calling was performed per gene basis (*rev* and *tat* genes were split into separate analyses for each reading frame). The median, minimum, maximum, 1st and 3rd quartile read alignment coverage and depth for HIV genes is visualized in Supplementary Figure 1. The two amplicons did not completely cover the genes *gag*, *pol* and *tat*, therefore, variant calls for the *gag* gene were considered from HXB2 reference genome AA position 240, for the *pol* gene until AA position 983 and for the *tat* gene from position 47. No sequencing was available for the genes *vpr* and *vif*. Minimum Phred quality scores for all called codons were >20. Variant calling was performed only for samples that had at least 10% of the gene covered by aligned reads with minimum 10-fold coverage. Variants were called if there was a minimum of 10-fold coverage and a minimum of 50% aligned reads supported the variant, i.e. we used the consensus call for the viral quasispecies.

Genotyping and quality control

2,546 study participants were genotyped by Affymetrix Axiom SNP-array enriched with markers related to immune dysfunction consisting of 388,501 probes after excluding the following: 1) non-monoallelic variants, 2) variants with >20% missingness, 3) <0.05 minor allele frequency and 4) variants which deviated from the Hardy-Weinberg equilibrium ($p<1\times10^{-6}$). The detailed procedure for host genotyping and HLA allele imputation is described elsewhere [15]. Only HLA alleles with ≥1% frequency in the population were used in the association analyses: 66 Class I HLA alleles (20 HLA-A, 29 HLA-B and 17 HLA-C alleles) and 64 Class II HLA alleles (14 HLA-DPB1, 13 HLA-DQA1, 14 HLA-DQB1 and 23 HLA-DRB1 alleles). Individuals with 1) >20% missingness, 2) outlying heterozygosity rates (more than ±3SD), 3) cryptic relatedness (PI Hat>0.2) [24] and 4) missing baseline VL measurements were excluded.

Statistical analysis

The association analysis was performed with PLINK2 [25]. A dominant linear regression model was fitted to assess associations of VL with HIV AA variants, and of VL with the bivariate association and interaction between HIV AA variants and HLA alleles. A generalized linear model with logit link function was applied for genome-to-genome and HIV AA variants with HLA allele association analysis. VL measurements were $\log_{10}$ transformed prior to analysis. Bonferroni-correction for multiple testing with an adjusted threshold of 0.05 was used for all associations except the VL association with the interaction between viral and host genetics. Age, sex assigned at birth and the first four principal components of the genetic data from both human and viral populations were used as covariates to account for population structures. The principal component analysis was performed using EIGENSOFT v6.1.4 [26,27]. R [28] was used for explained variance analysis and visualizations. The explained variance was analyzed by comparing the variable's sum of squares with the total sum of squares from ANOVA analysis.

Epitope binding prediction

Epitope binding was predicted with NetMHCpan v4.0 [29] considering all the 9-mer peptides with 9 nt window around the position of interest using the default binding affinity thresholds.

Results

Study participants

A total of 2,172 participants consented for genetic testing and had matching HIV genomes sequenced. Three participants did not have baseline VL measured, 46 had outlying heterozygosity and one had cryptic relatedness, leaving 2,122 participants in the study, spanning 5 continents and 23 countries (Table 1).

HIV amino acid variants

A total of 16,040 AA variants were called with VirVarSeq. The AA variants were filtered based on allele frequency: ≥5% and ≤95% of the study population. If more than one AA variant was observed in a position, the most abundant variant was designated as a reference and excluded from further analysis to reduce redundancy. This reduced the number of HIV AA variants to 3,010. The AA variant distribution across the HIV genome was 215 (7%) variants in *gag*, 1,041 (35%) in *pol*, 80 (2%) in *tat*, 144 (5%) in *rev*, 110 (4%) in *vpu*, 1,184 (39%) in *env* and 236 (8%) in *nef* genes.

Association of HIV amino acid variants with viral load

Of the 3,010 analyzed HIV AA variants, four were associated with VL (Bonferroni p-value<1.66×10^{-5}). All HIV AA variants tested for association with VL are visualized in a Manhattan plot (Supplementary Figure 2). The four variants were Pol980E (integrase), Tat53R, Rev7D and Env571W (gp41), details of these AA variants and their sequencing coverage are listed in Supplementary Table 1 and 2. The effect size of the four HIV AA variants ranged from -0.084 to +0.097 $\log_{10}$ copies/mL. ANOVA of the four HIV AA variants shows that they account for 8.2% of the total variation in $\log_{10}$ VL. Pearson's correlation coefficient of VL with the number of aligned reads was 0.21.

HIV AA variant association with host SNPs

Association tests between viral AA variants and host genome-wide SNPs resulted in a total of 12 HIV AA variants which were associated with a range of 1–512 human SNPs (Bonferroni threshold p-value<4.28×10^{-11}) (Table 2). Most associations were in the genes *gag* (N=5) and *nef* (N=5) with the strongest association between Gag242N HIV AA variant and the human rs1248395773 SNP (p-value=1.22×10^{-79}). All but one association were in the HLA Class I gene region (chr6:29,910,247–31,324,939). The association between Gag242N and a host SNP outside the HLA Class I gene region (chr15:50785016) was in high linkage disequilibrium (LD) (R^2≥0.90) with several associated SNPs from the HLA gene region (Supplementary Table 3).

Association of HIV amino acid variants with HLA alleles

As all the associations in the previous analysis were in the HLA gene region, we next tested the associations between the HIV AA variants and the imputed HLA alleles. We identified 46 associations (Bonferroni threshold p-value<1.29×10^{-7}) between 33 different HLA alleles (27 Class I and 6 Class II) and 31 HIV AA variants (Figure 2). The 12 previously identified associated HIV AA variants were also associated with HLA alleles. As in the previous analysis, most associations were in the genes *gag* (N=12) and *nef* (N=17). Furthermore, several HLA alleles were in LD while no HIV AA variants were in LD (Figure 3A, Supplementary Table 4). Associated HLA alleles showed different prevalence across participants' race; however, all of the associated alleles were present in the major subpopulations (White, Black and Hispanic participants) (Supplementary Table 5).

We further compared our findings to a study by Bartha *et al.* 2013 [19] that performed a similar analysis although not including HLA Class II alleles. Of the 39 associations between HLA Class I alleles and HIV AA variants identified in this study, 17 associations overlapped with the 53 associations between HLA Class I alleles and HIV AA variants identified by Bartha *et al.* (Figure 3B). Like our analysis, Bartha *et al.* also identified most associations to *gag* and *nef*; nonetheless, the largest relative overlap between studies was for *pol* (5 of 14 associations overlapped). Moreover, out of total 34 HLA Class I alleles associated with HIV AA variants in at least one study, 17 HLA alleles were associated with HIV AA variants in both studies.

Viral load association with the interaction between viral and host genetics

To disentangle the effect of the HLA allele presence together with the associated HIV AA variants, we performed an interaction analysis for the 46 previously identified HLA-HIV associations with VL. The effect sizes with 95% confidence intervals together with the p-values of the estimates are reported in Supplementary Table 6. In three cases the associated HLA allele and HIV AA variant interaction association with VL was below our p-value threshold (p-value<5×10^{-2}) (Table 3). In all three cases, the individual effects of the HLA alleles resulted in lower VL (p-value<5×10^{-2}), Gag302K and Pol432R

had no effect on VL, and Gag242N showed lower VL (p-value<5×10^{-2}). However, the interaction between the HLA alleles and the associated HIV AA variant in all three cases showed higher VL (p-value<5×10^{-2}) (Figure 4). Moreover, the interaction between B*57:01 and Gag242N AA variant did not meet our p-value threshold (p-value=8.5×10^{-2}); however, it showed the same interaction tendency as B*57:03 allele interaction with Gag242N AA variant (Supplementary Table 8). Both B*57:01 and B*57:03 HLA alleles are known to belong to the same functional group. Furthermore, we predicted the binding affinities in these four cases (Supplementary Table 7 and 8). Epitopes with the associated HIV AA variants were weak or non-binders; vice versa epitopes with other than the HIV AAs on the same position were strong binders.

The three associations with interaction effects on VL were used to investigate the explained variance in $\log_{10}$ VL. We first applied the model of VL with the three HLA alleles and the associated HIV AA variants as covariates, and HLA alleles together with HIV AA variants explained 2.0% of the total variance. When the interactions were added as covariates to the model, the explained variance was 3.6%.

Discussion

The host and viral genome interaction and its effect on HIV VL in demographically diverse populations is poorly understood. In this study, we performed four analyses using baseline data from ART-naive START trial participants to explore the interaction of viral and host genetics together with the impact it has on the VL which is a proxy for disease progression. While we found little signal in the associations between HIV AA variants and VL, we identified both host SNPs and HLA alleles which were associated with HIV AA variants. Finally, we showed that the interaction of several of these associations was affecting the VL.

The first association analysis with VL against HIV AA variants showed a weak signal, with only four associations below the p-value threshold. Similar results were observed in a study on persons infected with subtype B by Bartha *et al.*, 2013 [19]. Only the HIV AA variant Env571W is located in a region previously described as being associated with the replicative fitness of HIV [30,31]. However, while these associations pass the stringent p-value threshold requirements, further functional laboratory validation is necessary to confirm any biological function. From the low number of HIV AA variants associated with VL, it is apparent that single variants in the HIV genome cannot sufficiently explain differences in VL between individuals. The analysis of associations between VL and HIV AA variants is inherently limited, as VL correlates with the ability to detect HIV AA variants: low sequencing coverage occurred more frequently in low VL samples.

The second association analysis with HIV AA variants against human SNPs showed strong associations between HIV AA variants and SNPs in the HLA region. Particularly, Gag242N was associated with 511 SNPs in the Class I HLA region. Interestingly, none of the associations were observed in the gene *env* which might be due to the hypervariable nature of some regions of this gene and that adaptation to the host immune system happens shortly after the initial infection which makes it difficult to capture both adapted and non-adapted viral genotypes in the data set [32]. Ultimately, the associations in the host genome were only observed in the HLA gene region. The same observation of the associated SNPs being in the HLA gene region was made by Bartha *et al.*, 2013. To improve the sensitivity and biological interpretability of the analysis, we next tested directly for associations between HLA alleles and HIV AA variants.

Of 66 tested Class I HLA alleles, 27 were associated with one or several HIV AA variants, which could be corresponding to the $CD8^+$ T-cell mediated response primarily interacting with HIV. Our analysis adds to the previous findings of Bartha *et al.*, 2013 [19] that were based on a smaller and demographically more homogeneous cohort of primarily European descent, and 17 associations between HIV AA and Class I HLA alleles overlap between our and the study of Bartha *et al*. The

overlapping associations were located in *gag*, *pol* and *nef* genes which suggests these HIV genome regions as likely being under high host pressure. Several factors might explain differences in the associations identified across studies: 1) The START cohort is larger and more demographically heterogeneous than the cohort from Bartha *et al.* where only participants with Western European ancestry and carrying subtype B HIV were included, 2) this analysis included both Class I and Class II HLA alleles while Class II HLA alleles were not part of the Bartha *et al.* study, and 3) Bartha *et al.* did not include a part of the *env* gene (coding for GP120) while in this analysis the beginning of the *gag* gene, the end of the *pol* gene, the beginning of the *tat* gene, and the genes *vif* and *vpr* were not sequenced. Our study supports the findings from Kinloch *et al.* [33], that many escape pathways are not subtype B specific, but rather shared among different subtypes.

Only 6 out of 64 tested HLA Class II alleles had associations with p-values below the Bonferroni threshold. Furthermore, several associated HLA Class II alleles were in LD with other associated HLA alleles which could indicate that HLA Class II alleles have less effect on HIV diversity as the primary immune response is mediated through $CD8^+$ T-cells, and $CD4^+$ cells expressing HLA Class II molecules play a lesser role during the infection [34]. However, from the 46 identified associations between HLA alleles and HIV AA variants (Figure 3A) we show that a heterogeneous group of HLA alleles is associated with a diverse group of HIV AA variants which might be essential during the HIV infection for viral adaptation to the host.

The interaction analysis between HIV AA variants with the associated HLA alleles and their influence on VL revealed that, while several HLA alleles are linked to lower VL and the HIV AA variants associate with no or even a slightly lower VL, the interaction between them associated with higher VL; that way the protective effect of the HLA allele is diluted when analyzed with the corresponding viral genotype. Arora *et al.,* 2019 [35] recently showed that B*57:01 and B*57:03 HLA alleles share similar clusters of HIV-associated epitopes. While predominantly found in different races, our analysis shows that these HLA alleles and their interaction with the associated HIV AA variants show

the same impact on VL which further supports the observed interaction effect as a selected biological trait. This observation is further supported by the predicted epitope binding affinity of the four associations where the associated HLA allele had a lower binding affinity to the epitope containing the associated HIV AA variant than the other observed AA variants in the position of interest.

While only a fraction of variance is explained using the three associated HLA alleles and HIV AA variants together with their interaction, the observation that the addition of the interaction to the model explained almost twice as much variance indicates that the interaction between the viral and host genomes is essential when analyzing viral evolution and adaptation.

Our study has several limitations. First, while it is a demographically diverse and large cohort, we did not have consent for genotyping from Asian START clinical trial participants. Furthermore, larger cohorts that expand on the demographic diversity are needed for the more sensitive association identification in the association of VL to HIV AA variants analyzed in concert with interactive HLA alleles as the analysis of participants with less common HLA alleles did not have enough power. Moreover, because of the study design no follow-up HIV genomes were included in the study which could have provided further support for the identified associations as the phenomena of the viral adaptation to the host immunotype. SNP-array genotyping of the host genome and HIV amplicon sequencing instead of whole genome sequencing limited the amount of genetic variants that could be identified for association analysis; however our SNP arrays were enriched for immune genes of the human genome [15] and the majority of the coding region in the HIV genome was covered. Similarly, the HLA allele information was imputed with high accuracy and not based on conventional HLA typing, thus slightly increasing the uncertainty in our study. In addition, we analyzed only HIV consensus sequences, and accordingly we did not address the intra-host diversity which would provide more detailed insights at the individual level. Finally, because of the nature of the cohort, no

participants with low $CD4^+$ cell count and low VL were included in the analysis and, therefore, HLA alleles from HIV controllers will not be well-represented in our analysis.

In conclusion, by using a demographically diverse cohort of treatment-naive HIV-positive persons with different HIV subtypes, we were able to identify and quantify the influence that viral and host genetics have on VL. We observed that HIV AA variants alone explained little to no variation in VL. Instead, we found that a diverse group of HLA Class I alleles, and to less extent also HLA Class II, impose selection on HIV genomes for variants primarily in *gag* and *nef* genes, and that VL was influenced by the interplay of viral and human genetics. Accordingly, HIV AA variants and the host HLA alleles should be analyzed and interpreted together, highlighting that HIV virulence factors are best studied in the context of the host.

Acknowledgments

We would like to thank all participants in the START trial and all trial investigators (see N Engl J Med 2015; 373:795-807 for the complete list of START investigators). The START trial was supported by the National Institute of Allergy and Infectious Diseases, Division of AIDS (United States) (NIH Grants UM1-AI068641, UM1-AI120197 and 1U01-AI36780), 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 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. All figures were partly or completely created using BioRender.

Funding

The work was supported by the Danish National Research Foundation (grant 126).

Conflict of interest statement

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest.

Author contributions

R.L.M. and J.D.L. conceived the study; R.L.M. and J.D.L. supervised the project; M.G., M.B., A.G.Z., R.L.M and J.D.L. designed the analysis plan; M.G. and M.B. conducted the analysis; M.G. and M.B. wrote the manuscript; M.G., M.B., A.G.Z., C.E., D.D.M., V.L.K., G.T., L.V., D.T., J.N., H.C.L., S.S., A.A.P., M.N.P., H.F.G., J.D.L.and R.L.M. revised the analysis plan, and reviewed and edited the manuscript.

Ethics declaration

Ethics Samples included in this study were derived from participants who consented in the clinical trial, START (NCT00867048) (22), run by the International Network for Strategic Initiatives in Global HIV Trials (INSIGHT). The study was approved by the institutional review board or ethics committee at each contributing center, and written informed consent was obtained from all participants. All informed consents were reviewed and approved by participant site ethics review committees.

This work was presented at a virtual CROI 2020; Poster 00272

Correspondence:

Correspondence to Migle Gabrielaite, migle.gabrielaite@regionh.dk, Marc Bennedbæk, marc.bennedbaek@regionh.dk and Rasmus L. Marvig rasmus.lykke.marvig@regionh.dk

Figures and Tables

Figure 1. Schematic visualization of the four association analyses performed in this study. AA (amino acid); SNP (single nucleotide polymorphism).

Figure 2. Associations between HIV amino acid (AA) variants and HLA alleles projected on HIV genome with the strongest association for each HIV AA variant included. Red horizontal line marks the Bonferroni p-value threshold ($p=1.29\times10^{-7}$). The x-axis denotes gene AA positions. Circle size denotes the number of significantly associated host HLA alleles.

Figure 3. (A) Network of the associated HLA alleles and HIV amino acid (AA) variants where line thickness represents the association's strength and red arrows mark HLA alleles in linkage disequilibrium (LD). The opacity of the arrows corresponds to squared correlation coefficient (R^2). (B) Comparison between the HLA type and HIV AA associations identified in this analysis and Bartha *et al.* 2013 study.

Figure 4. Violin plots of the three HIV amino acid (AA) variant and HLA allele associations which interaction had a strong (p-value<0.05) effect on viral load (VL) (A-C), and the Gag242N and B*57:01 association which did not have a strong interaction effect on VL but shows the same tendency as Gag242N and B*57:03 association (D).

Table 1. Characteristics of 2,122 participants from the START trial included in this analysis. IQR (interquartile range)

Characteristic	**START participants after quality control, No. (%) (N=2,122)**
Age median (IQR), years	36 (29–45)
Sex	
Female	365 (17%)
Male	1,757 (83%)
Self-reported race	
Asian	21 (1%)
Black	432 (20%)
Hispanic	382 (18%)
White	1,250 (59%)
Other	37 (2%)
Geographic region	
Africa	244 (11%)
Australia	85 (4%)
Europe and Israel	1,030 (49%)

Latin America	392 (18%)
United States	371 (17%)
Mode of HIV infection	
Injection drug use	34 (2%)
Sex with same sex	1,445 (68%)
Sex with opposite sex	625 (29%)
Other	23 (1%)
Subtype	
A	24 (1%)
B	1,364 (64%)
C	101 (5%)
AB	35 (2%)
AE	17 (1%)
AG	38 (2%)
BC	85 (4%)
BF	113 (5%)
BG	11 (1%)
Other	67 (3%)

Unable	268 (13%)
Time since HIV diagnosis, median (IQR), years	1.08 (0.44–2.84)
Recent infection (within 6 months)	163 (7.7%)
$CD4^+$ cell count, median (IQR), cells/μL	643 (583–744)
HIV load, median (IQR), copies/mL	19,450 (6393–51,732)

Table 2. 12 HIV amino acid (AA) variants significantly associated with varying numbers of host single nucleotide polymorphisms (SNPs). The information about host and viral variant location in the genome together with the lowest p-value is reported.

HIV AA variant	Lowest p-value	Number of associated host SNPs	Region of associated host SNPs
GAG242N	1.22×10^{-79}	512	chr6:29818568–32626272; chr15:50785016
GAG302K	1.35×10^{-16}	8	chr6:30993440–31326410
GAG357S	3.17×10^{-16}	48	chr6:31005726–31812038
GAG397R	1.13×10^{-32}	186	chr6:29607476–30222020
GAG403K	2.78×10^{-19}	4	chr6:29830074–29924779
POL432R (RT277R)[a]	6.26×10^{-39}	134	chr6:29611885–30711805
POL837I (INT122I)[b]	1.85×10^{-11}	1	chr6:31316609
NEF071K	1.38×10^{-31}	150	chr6:30861729–31879158
NEF092R	4.70×10^{-19}	8	chr6:29785827–30026290
NEF102H	8.79×10^{-16}	1	chr6:29785827–30026290
NEF105R	9.38×10^{-12}	3	chr6:31242649–31245821

NEF135F	6.36×10^{-21}	8	ch6:29735280–29992261

[a]RT = Reverse Transcriptase; [b]INT = Integrase

Table 3. Summary of HLA alleles, the associated HIV amino acid (AA) variants and their interaction effect on viral load which passed the p-value threshold. The variables in bold represent association tests below the p-value threshold. CI (confidence interval).

Variable	Effect size (95% CI)	p-value
Gag242N	**-0.136 (-0.220 – -0.053)**	**1.4×10^{-3}**
B*57:03	**-0.591 (-0.945 – -0.237)**	**1.1×10^{-3}**
Gag242N×B*57:03	**0.577 (0.163 – 0.991)**	**6.4×10^{-3}**
Gag302K	0.005 (-0.124 – 0.135)	9.3×10^{-1}
B*14:01	**-0.419 (-0.712 – -0.125)**	**5.2×10^{-3}**
Gag302K×B*14:01	**0.540 (0.185 – 0.896)**	**3.0×10^{-3}**
Pol432R	-0.015 (-0.082 – 0.051)	6.5×10^{-1}
A*03:01	**-0.188 (-0.324 – -0.051)**	**7.3×10^{-3}**
Pol432R×A*03:01	**0. 226 (0.068 – 0.384)**	**5.1×10^{-3}**

Bibliography

1. Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V, Balloux F. Pathogen-driven selection and worldwide HLA class I diversity. Curr Biol. **2005**; 15(11):1022–1027.

2. Hertz T, Nolan D, James I, John M, Gaudieri S, Phillips E, et al. Mapping the landscape of host-pathogen coevolution: HLA class I binding and its relationship with evolutionary conservation in human and viral proteins. J Virol. **2011**; 85(3):1310–1321.

3. Fraser C, Lythgoe K, Leventhal GE, Shirreff G, Hollingsworth TD, Alizon S, et al. Virulence and pathogenesis of HIV-1 infection: an evolutionary perspective. Science. **2014**; 343(6177):1243727.

4. Zanini F, Puller V, Brodin J, Albert J, Neher RA. In vivo mutation rates and the landscape of fitness costs of HIV-1. Virus Evol. **2017**; 3(1):vex003.

5. Voronin Y, Holte S, Overbaugh J, Emerman M. Genetic drift of HIV populations in culture. PLoS Genet. **2009**; 5(3):e1000431.

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

7. Naranbhai V, Carrington M. Host genetic variation and HIV disease: from mapping to mechanism. Immunogenetics. **2017**; 69(8–9):489–498.

8. Zucco AG, Bennedbaek M, Ekenberg C, Gabrielaite M, Murray DD, MacPherson C, et al. Functional clustering and association of HLA class I alleles to viral load in HIV-positive and ART-naive participants from the INSIGHT START study. HIV Med. **2019**; 20:181.

9. Fellay J, Pedergnana V. Exploring the interactions between the human and viral genomes. Hum Genet. **2020**; 139(6–7):777–781.

10. Bartha I, McLaren PJ, Brumme C, Harrigan R, Telenti A, Fellay J. Estimating the respective contributions of human and viral genetic variation to HIV control. PLoS Comput Biol. **2017**;

13(2):e1005339.

11. Alizon S, Wyl V von, Stadler T, Kouyos RD, Yerly S, Hirschel B, et al. Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load. PLoS Pathog. **2010**; 6(9):e1001123.

12. Bertels F, Marzel A, Leventhal G, Mitov V, Fellay J, Günthard HF, et al. Dissecting HIV Virulence: Heritability of Setpoint Viral Load, CD4+ T-Cell Decline, and Per-Parasite Pathogenicity. Mol Biol Evol. **2018**; 35(1):27–37.

13. Blanquart F, Wymant C, Cornelissen M, Gall A, Bakker M, Bezemer D, et al. Viral genetic variation accounts for a third of variability in HIV-1 set-point viral load in Europe. PLoS Biol. **2017**; 15(6):e2001855.

14. Bachmann N, Turk T, Kadelka C, Marzel A, Shilaih M, Böni J, et al. Parent-offspring regression to estimate the heritability of an HIV-1 trait in a realistic setup. Retrovirology. **2017**; 14(1):33.

15. Ekenberg C, Tang M-H, Zucco AG, Murray DD, MacPherson CR, Hu X, 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. **2019**; 220(8):1325–1334.

16. Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, Weale M, et al. A whole-genome association study of major determinants for host control of HIV-1. Science. **2007**; 317(5840):944–947.

17. International HIV Controllers Study, Pereyra F, Jia X, McLaren PJ, Telenti A, Bakker PIW de, et al. The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. Science. **2010**; 330(6010):1551–1557.

18. Fellay J, Ge D, Shianna KV, Colombo S, Ledergerber B, Cirulli ET, et al. Common genetic variation and the control of HIV-1 in humans. PLoS Genet. **2009**; 5(12):e1000791.

19. Bartha I, Carlson JM, Brumme CJ, McLaren PJ, Brumme ZL, John M, et al. A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. elife. **2013**; 2:e01123.

20. Power RA, Davaniah S, Derache A, Wilkinson E, Tanser F, Gupta RK, et al. Genome-Wide Association Study of HIV Whole Genome Sequences Validated using Drug Resistance. PLoS ONE. **2016**; 11(9):e0163746.

21. Tough RH, McLaren PJ. Interaction of the host and viral genome and their influence on HIV disease. Front Genet. **2018**; 9:720.

22. INSIGHT START Study Group, Lundgren JD, Babiker AG, Gordin F, Emery S, Grund B, et al. Initiation of antiretroviral therapy in early asymptomatic HIV infection. N Engl J Med. **2015**; 373(9):795–807.

23. Verbist BMP, Thys K, Reumers J, Wetzels Y, Van der Borght K, Talloen W, et al. VirVarSeq: a low-frequency virus variant detection pipeline for Illumina sequencing using adaptive base-calling accuracy filtering. Bioinformatics. **2015**; 31(1):94–101.

24. Marees AT, Kluiver H de, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res. **2018**; 27(2):e1608.

25. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. **2015**; 4:7.

26. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. **2006**; 2(12):e190.

27. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. **2006**; 38(8):904–909.

28. R: The R Project for Statistical Computing [Internet]. [cited 2020 Feb 10]. Available from: https://www.r-project.org/

29. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J Immunol. **2017**; 199(9):3360–3368.

30. Sloan EA, Kearney MF, Gray LR, Anastos K, Daar ES, Margolick J, et al. Limited nucleotide changes in the Rev response element (RRE) during HIV-1 infection alter overall Rev-RRE activity and Rev multimerization. J Virol. **2013**; 87(20):11173–11186.

31. Sherpa C, Rausch JW, Le Grice SFJ, Hammarskjold M-L, Rekosh D. The HIV-1 Rev response element (RRE) adopts alternative conformations that promote different rates of virus replication. Nucleic Acids Res. **2015**; 43(9):4676–4686.

32. Andrews SM, Zhang Y, Dong T, Rowland-Jones SL, Gupta S, Esbjörnsson J. Analysis of HIV-1 envelope evolution suggests antibody-mediated selection of common epitopes among Chinese former plasma donors from a narrow-source outbreak. Sci Rep. **2018**; 8(1):5743.

33. Kinloch NN, Lee GQ, Carlson JM, Jin SW, Brumme CJ, Byakwaga H, et al. Genotypic and Mechanistic Characterization of Subtype-Specific HIV Adaptation to Host Cellular Immunity. J Virol. **2019**; 93(1).

34. Blankson JN, Siliciano RF. MHC class II genotype and the control of viremia in HIV-1-infected individuals on highly active antiretroviral therapy. J Clin Invest. **2001**; 107(5):549–551.

35. Arora J, McLaren PJ, Chaturvedi N, Carrington M, Fellay J, Lenz TL. HIV peptidome-wide association study reveals patient-specific epitope repertoires associated with HIV control. Proc Natl Acad Sci USA. **2019**; 116(3):944–949.

List of abbreviations

AA: amino acid

AIDS: acquired immune deficiency syndrome

ANOVA: analysis of variance

ART: antiretroviral treatment

HIV: human immunodeficiency virus

HLA: human leukocyte antigen

LD: linkage disequilibrium

nt: nucleotide

SD: standard deviation

SNP: single nucleotide polymorphism

START: Strategic Timing of Antiretroviral Treatment

VL: viral load

Figure 1

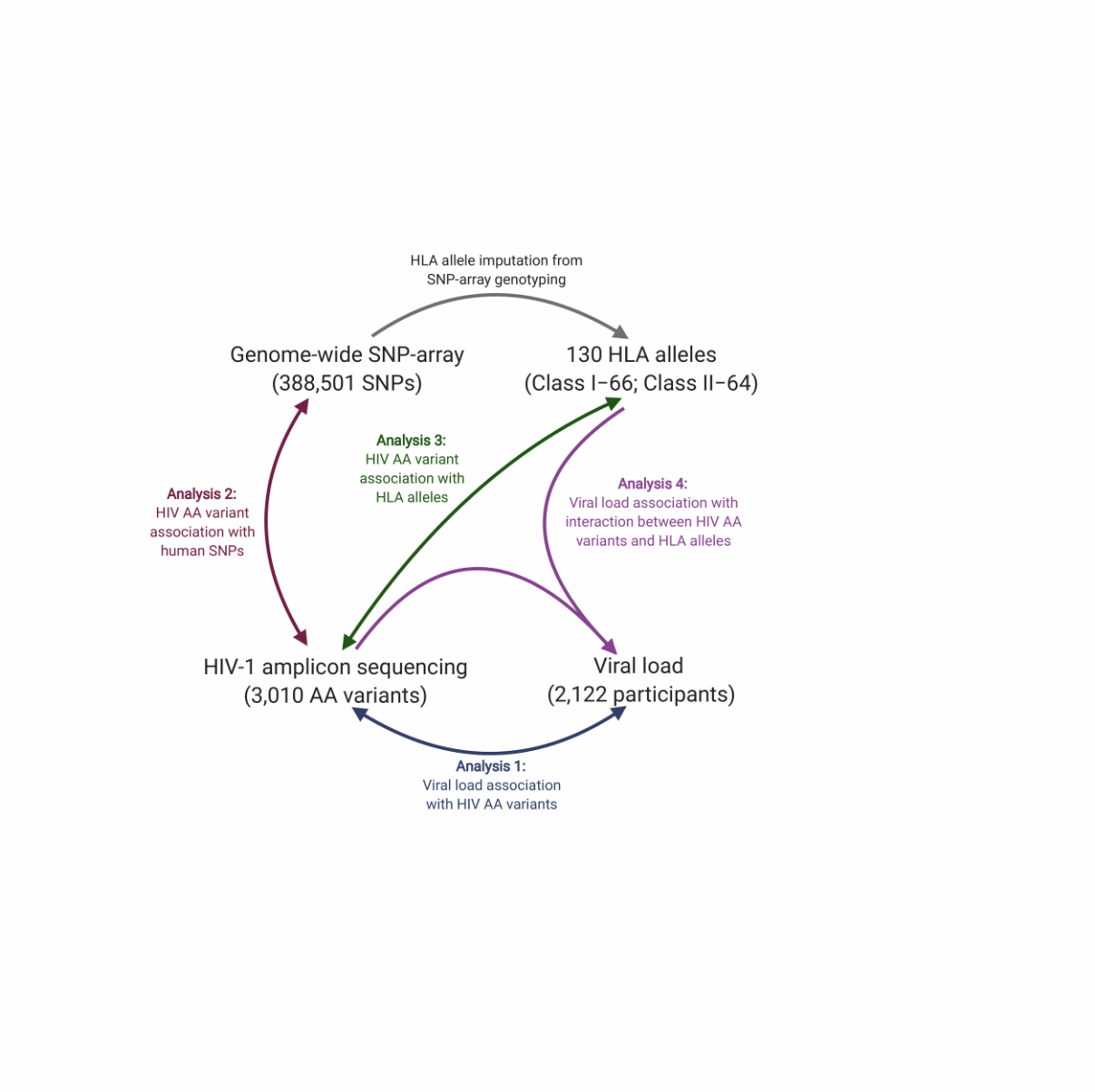

Figure 2

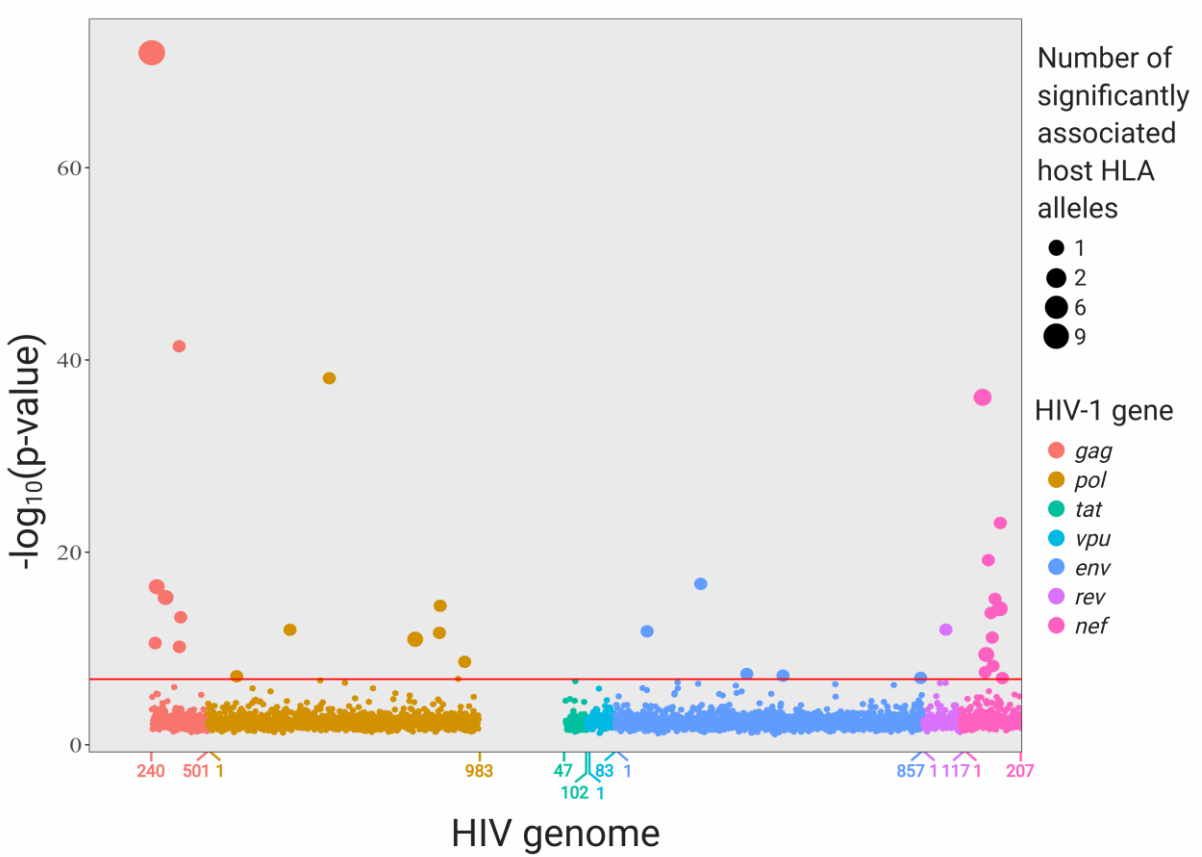

Figure 3

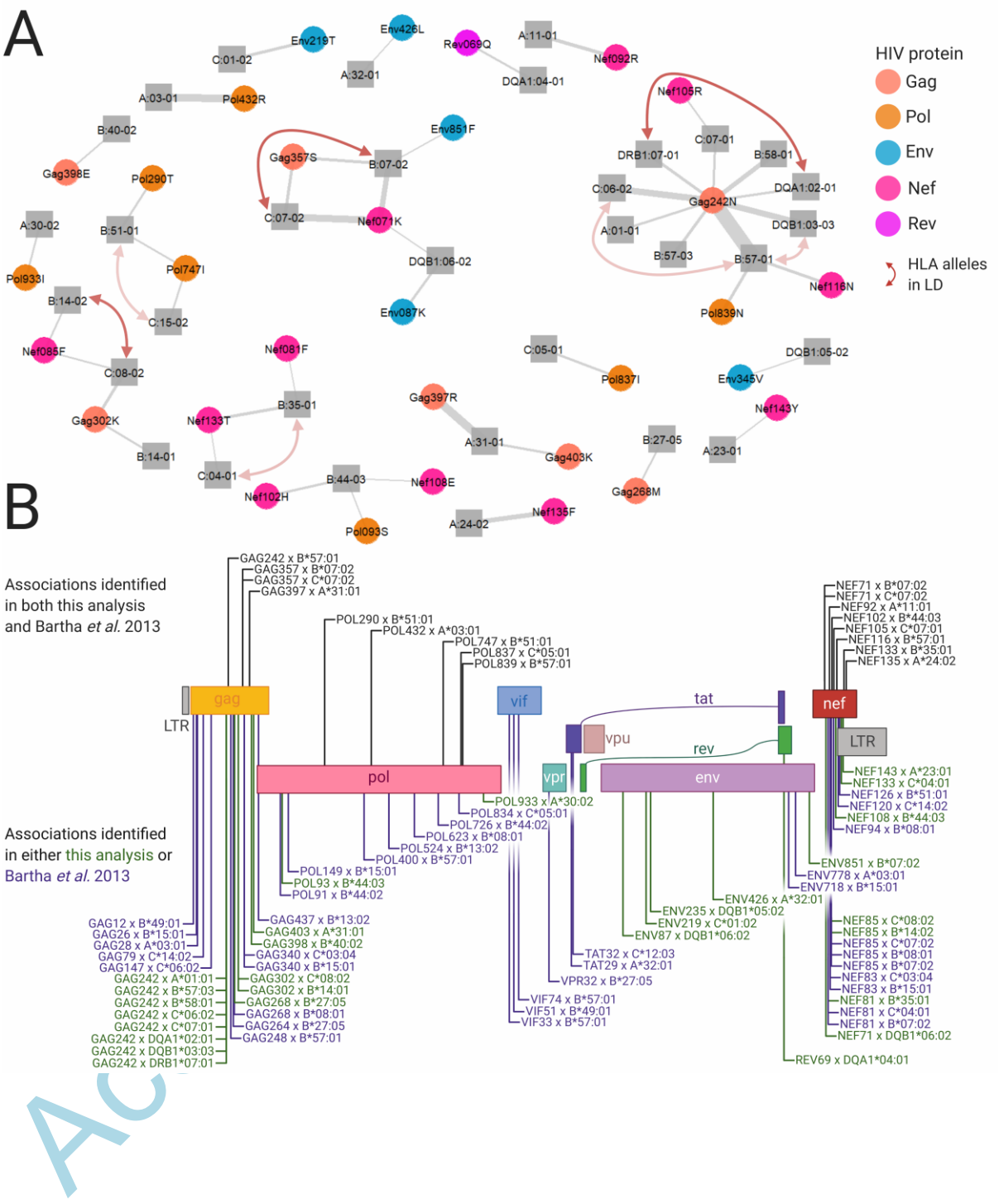

Figure 4