Trans-ancestral fine-mapping of MHC reveals key amino acids associated with spontaneous clearance of hepatitis C in HLA-DQβ1

Authors

Ana Valencia, Candelaria Vergara,
Chloe L. Thio, ..., Alessandro Sette,
David L. Thomas, Priya Duggal

Correspondence

pduggal@jhu.edu

CellPress

Trans-ancestral fine-mapping of MHC reveals key amino acids associated with spontaneous clearance of hepatitis C in HLA-DQβ1

Ana Valencia,[1,2,24] Candelaria Vergara,[3,24] Chloe L. Thio,[1] Nicolas Vince,[4] Venceslas Douillard,[4] Alba Grifoni,[5] Andrea L. Cox,[1] Eric O. Johnson,[6] Alex H. Kral,[6] James J. Goedert,[7] Alessandra Mangia,[8] Valeria Piazzolla,[8] Shruti H. Mehta,[3] Gregory D. Kirk,[1,3] Arthur Y. Kim,[9] Georg M. Lauer,[10] Raymond T. Chung,[10] Jennifer C. Price,[11] Salim I. Khakoo,[12] Laurent Alric,[13] Matthew E. Cramp,[14] Sharyne M. Donfield,[15] Brian R. Edlin,[16] Michael P. Busch,[17] Graeme Alexander,[18] Hugo R. Rosen,[19] Edward L. Murphy,[17] Genevieve L. Wojcik,[3] Mary Carrington,[20,21,22] Pierre-Antoine Gourraud,[4] Alessandro Sette,[5,23] David L. Thomas,[1] and Priya Duggal[3,*]

Abstract

Spontaneous clearance of acute hepatitis C virus (HCV) infection is associated with single nucleotide polymorphisms (SNPs) on the MHC class II. We fine-mapped the MHC region in European (n = 1,600; 594 HCV clearance/1,006 HCV persistence) and African (n = 1,869; 340 HCV clearance/1,529 HCV persistence) ancestry individuals and evaluated HCV peptide binding affinity of classical alleles. In both populations, HLA-DQβ1Leu26 (p value$_{Meta}$ = 1.24 × 10^{-14}) located in pocket 4 was negatively associated with HCV spontaneous clearance and HLA-DQβ1Pro55 (p value$_{Meta}$ = 8.23 × 10^{-11}) located in the peptide binding region was positively associated, independently of HLA-DQβ1Leu26. These two amino acids are not in linkage disequilibrium (r^2 < 0.1) and explain the SNPs and classical allele associations represented by rs2647011, rs9274711, *HLA-DQB1*03:01*, and *HLA-DRB1*01:01*. Additionally, HCV persistence classical alleles tagged by HLA-DQβ1Leu26 had fewer HCV binding epitopes and lower predicted binding affinities compared to clearance alleles (geometric mean of combined IC$_{50}$ nM of persistence versus clearance; 2,321 nM versus 761.7 nM, p value = 1.35 × 10^{-38}). In summary, MHC class II fine-mapping revealed key amino acids in HLA-DQβ1 explaining allelic and SNP associations with HCV outcomes. This mechanistic advance in understanding of natural recovery and immunogenetics of HCV might set the stage for much needed enhancement and design of vaccine to promote spontaneous clearance of HCV infection.

Introduction

An estimated 71.1 million people worldwide have chronic hepatitis C viral (HCV) infection,[1,2] which can result in liver cirrhosis, hepatocellular cancer, and liver failure.[3] Not all acute infections become chronic since approximately 30% of infected individuals spontaneously clear HCV infection. Spontaneous clearance is associated with sex, race, and host genetics.[4,5] Host genetic variants in the Interferon Lamba 3 and Interferon Lambda 4 (*IFNL4, IFNL3*) genes, the MHC region, and G protein-coupled re-

ceptor 158 (*GPR158*) gene[5–9] are all associated with HCV spontaneous clearance. Interestingly, the association in the MHC class II region which spans 305.2 kb across the *HLA-DQB1/HLA-DQA1* and *HLA-DQA2* genes is consistent across populations of different genetic ancestries and demographic histories.[9–16] Specifically, the main genetic signals are localized 48.5 kb upstream of *HLA-DQA2* and 33.3 kb upstream of the *HLA-DQB1* gene;[5,6] however, the causal variants for this association remain elusive.

The tight linkage disequilibrium (LD) across disease-associated MHC haplotype and the highly polymorphic

[1]Johns Hopkins University, School of Medicine, Baltimore, MD 21205, USA; [2]Universidad Pontificia Bolivariana, Medellín, Antioquia 050031, Colombia; [3]Johns Hopkins University, Bloomberg School of Public Health, Baltimore, MD 21205, USA; [4]Université de Nantes, CHU Nantes, Inserm, Centre de Recherche en Transplantation et Immunologie, UMR 1064, ITUN, Nantes 44000, France; [5]Center for infectious Diseases and Vaccine Research, La Jolla Institute for Immunology, La Jolla, CA 92037, USA; [6]GenOmics, Bioinformatics, and Translational Research Center, RTI International, Research Triangle Park, NC 27709, USA; [7]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA; [8]Liver Unit, Medical Sciences Department, Fondazione "Casa Sollievo della Sofferenza" IRCCS, 71013 San Giovanni Rotondo, Italy; [9]Division of Infectious Diseases, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA; [10]Liver Center and Gastrointestinal Division, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA; [11]Division of Gastroenterology, Department of Medicine, School of Medicine, University of California, San Francisco, CA 94143, USA; [12]University of Southampton, Southampton General Hospital, Southampton SO16 6YD, UK; [13]Internal Medicine-Department of Digestive Diseases, Rangueil Hospital, Toulouse University, 1, 31400 Toulouse, France; [14]South West Liver Unit, Plymouth PL6 8DH, UK; [15]Rho, Inc., Durham, NC 27713, USA; [16]SUNY Downstate College of Medicine, Brooklyn, NY 11203, USA; [17]University of California San Francisco and Vitalant Research Institute, San Francisco, CA 94118, USA; [18]UCL Institute for Liver and Digestive Health, The Royal Free Hospital, Pond St, Hampstead, London NW3 2QG, UK; [19]University of Colorado, Aurora, CO 80045, USA; [20]Basic Science Program, Frederick National Laboratory for Cancer Research, National Cancer Institute, Frederick, MD 21702, USA; [21]Laboratory of Integrative Cancer Immunology, Center for Cancer Research, National Cancer Institute, Bethesda, MD 20892, USA; [22]Ragon Institute of MGH, MIT and Harvard, Cambridge, MA 02139, USA; [23]Department of Medicine, Division of Infectious Diseases and Global Public Health, University of California, San Diego (UCSD), La Jolla, CA 92093, USA
[24]These authors contributed equally
*Correspondence: pduggal@jhu.edu
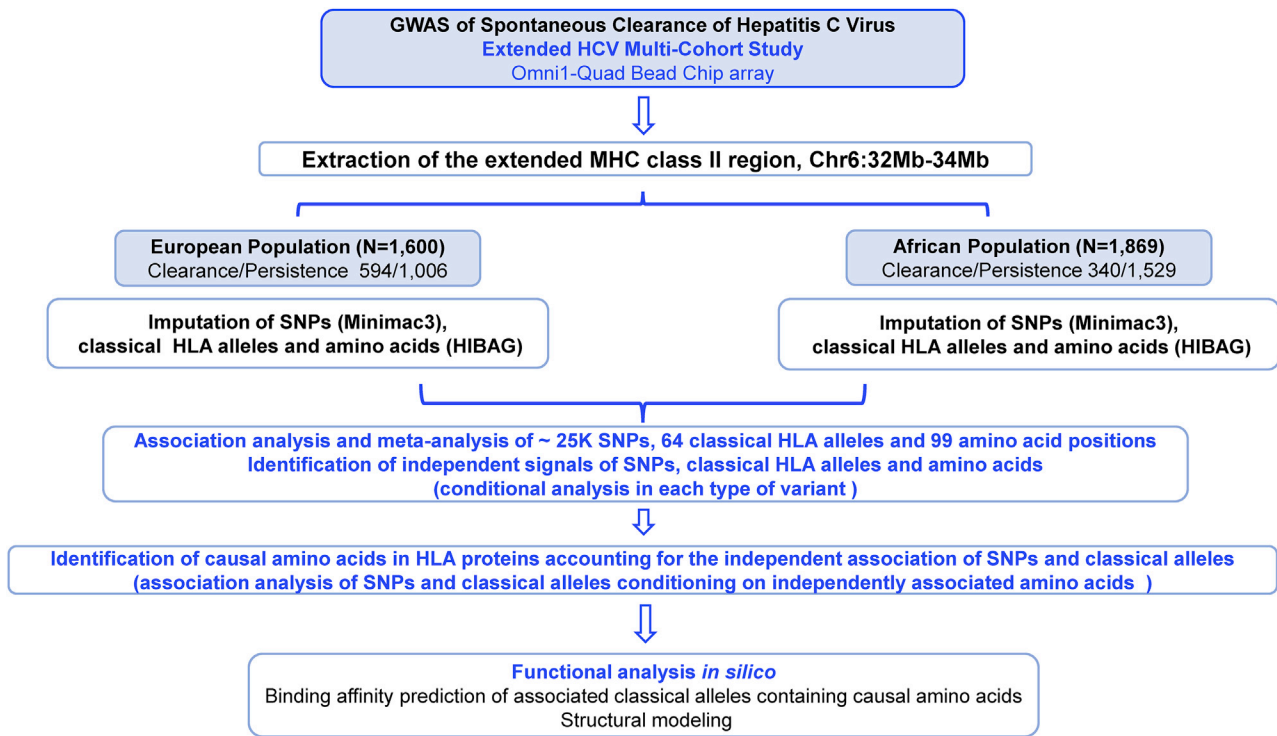https://doi.org/10.1016/j.ajhg.2022.01.001
© 2022

Figure 1. Outline of the trans-ancestral fine mapping strategy used to identify causal variants in the MHC region associated with HCV spontaneous clearance in European and African ancestry populations

nature of associated variants makes determining the contribution of specific classical HLA alleles and amino acid residues to the spontaneous clearance of HCV infection challenging. To identify putative causal amino acid positions and residues associated with HCV spontaneous clearance and their potential impact on peptide presentation by HLA molecules, we investigated the HLA region in individuals of African and European ancestry with HCV spontaneous clearance and persistence. We identified key amino acids that were present in both European and African ancestry individuals, suggesting they are not specific to ancestry populations and may be targeted across diverse groups. We also predicted the peptide binding affinity of HLA class II molecules identified as relevant for HCV clearance or persistence.

Material and methods

Samples
The study design is presented in Figure 1. The study included 3,469 individuals participating in the Extended HCV Genetics Consortium.[5,6,17] This is a multi-site international consortium including multiple studies from Europe and United States in which HCV infection outcomes were ascertained. They include ALIVE (AIDS Link to the Intravenous Experience),[18] BBAASH (Baltimore Before and After Acute Study of Hepatitis),[19] BAHSTION (Boston Acute Hepatitis C Virus Study: Transmission, Immunity and Outcomes Network),[13] Cramp and colleagues' study,[20] HGDS (Hemophilia Growth and Development Study),[21] Mangia and colleagues' study,[22] MHCS (Multicenter Hemophilia Cohort

Study) and MHCS-II,[23] REVELL (Correlates of Resolved Versus Low-Level Viremic Hepatitis C Infection in Blood Donors) study,[24] the Swan Project,[25] the Toulouse, France cohort,[15] WIHS (Women's Interagency HIV Study),[26] the United Kingdom Drug Use cohort,[27] and the Urban Health Study (UHS)[28,29] as described in detail in elsewhere.[5,6,17] These studies were selected because they had well-established hepatitis C virus (HCV) outcomes, as previously described,[5,6,17] available DNA, and IRB approval for genetic testing. Information about HIV infection status was also obtained in the included individuals since it is a determinant of HCV clearance. All subjects gave written informed consent for this study, approved by each Institutional Review Board.[13] The complete sample corresponded to 1,600 individuals of European ancestry (594 with HCV clearance/1,006 with HCV persistence) and 1,869 of African ancestry (340 with clearance/1,529 with persistence). Detailed information about the study cohorts is available in supplemental material and methods and distribution of the analyzed individuals by ancestry group, sex, and HIV infection status is presented in Table 1.

Genotyping, imputation of SNPs, classical HLA alleles, and amino acids
Our initial dataset consisted of genotypes obtained using the Illumina Omni1-Quad BeadChip array (Illumina)[5,6] and processed using standard genome-wide association study protocols for quality control[6] (dbGaP accession numbers dbGaP:phs000454.v1.p1 and dbGaP:phs000248.v1.p1). We extracted genotyped markers from a 2 Mb segment of the MHC corresponding to the extended class II region[30] delimited by chr6:32,000,000–34,000,000 (coordinates based on the Genome Reference Consortium Human Build 37, GRCh37/hg19). SNP imputation was done using Minimac3[31] software through the publicly available Michigan Imputation Server.[6]

Table 1. Demographic characteristics of the analyzed studies by genetically determined ancestry groups

Genetically determined ancestry group	N	HCV infection persistence:clearance	(+) HIV infection (%)	Female sex (%)
African ancestry	1,869	1,529:340	38.6	33.3
European ancestry	1,600	1,006:594	16.1	32.1
Total	3,469	2,535:934	28.2	32.8

As part of the quality control, we excluded SNPs with imputation r^2 < 0.3 and minor allele frequency (MAF) < 1%.

In each ancestry group, we used the method implemented in HIBAG (HLA Imputation using attribute BAGging)[32] to impute classical HLA alleles from genotyped/imputed SNPs for *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1*, and *HLA-DRB1* loci producing genotypes at two field-resolution for each individual. This method takes advantage of the extended haplotype structure within the MHC and makes predictions of classical HLA alleles by averaging HLA-type posterior probabilities over an ensemble of classifiers built on bootstrap samples. Classifiers are trained from large databases of individuals with known HLA and SNP genotypes within the MHC region.[32] In order to capture the appropriate genetic background, we used a paired ancestry group of classifiers for each target population. For European ancestry we used pre-built classifiers computed from individuals of European ancestry from the HLARES database[32] included with the software, and for individuals of African ancestry we built classifiers based on the HLA types and genotypes from the "Consortium on Asthma among African-ancestry Populations in the Americas" (CAAPA, dbGaP Study Accession: dbGaP:phs001123.v1.p1)[33,34] (supplemental material and methods). For the imputation of classical HLA class II alleles and amino acids, we used markers that intersected between our target populations and those included in the classifiers for each ancestry group, corresponding to 3,569 SNPs for European ancestry and 991 for African ancestry.

For association analyses, we excluded classical HLA genotypes imputed at two-field resolution with a posterior probability lower than 0.5[35,36] and alleles with MAF lower than 1%. Details of HLA imputation performance of classical alleles for the different HLA genes are presented in Table S1.

Translation of classical HLA alleles to amino acid residues in the antigen binding domain of the HLA proteins was performed based on the classical HLA alleles imputed by HIBAG and the peptide sequence of alleles available in the HLA database using the HIBAG software.[32] For association analyses with HCV spontaneous clearance, we included all amino acid polymorphic positions with residues with a frequency greater than 1% (Figure 1).

Statistics

We performed association analyses with each type of variant (SNPs, classical HLA alleles, and amino acid residues in the proteins) and conditional analyses to identify independent SNPs, classical HLA alleles, and amino acid signals within each type of variant. Finally, we used conditional analyses across the variant types to identify which amino acids explained the independent associations (Figure 1).

Association analyses and meta-analysis of each type of variant

For each type of variant that met quality control thresholds, we tested for association with HCV spontaneous clearance using an additive logistic regression model adjusting for HIV infection status and the first two principal components using PLINK 2.0 alpha version[37,38] for SNPs, and HIBAG[32] for HLA classical allele and

amino acid residue in the proteins (supplemental material and methods). Population-specific PCs were calculated using the *smartpca* program in EIGENSOFT[32,39] separately in the European and the African ancestry populations (excluding 1000 Genomes reference populations) using 16,142 genomic independent markers. In each ancestry group, individuals grouped in a tight cluster, and PC1 and PC2 explained most of the variance in the data (Figure S1). We extended the analysis to consider HIV status and applied the same regression model and meta-analysis to those variants in individuals with and without HIV independently in each population. In the European ancestry population, we tested associations with 22,259 genotyped/imputed SNPs; 54 classical alleles in the genes *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, and *HLA-DPB1* and 205 residues (and combinations of residues) in 99 amino acid positions in these proteins. For the African ancestry population, we evaluated associations with 25,565 SNPs, 33 classical HLA alleles in the genes *HLA-DRB1* and *HLA-DQB1*, and 155 residues and their combinations. Classical HLA alleles and amino acids in *HLA-DQA1* and *HLA-DPB1* were not imputed in the African ancestry population since data for those genes was not available in the reference panel used for imputation (Table S1).

For each type of variant, we analyzed each ancestry group separately and then performed a fixed effect inverse-variance weighted meta-analysis, using METAL.[40] A locus-specific p value threshold of significance for the meta analysis of this chromosome region would be adequate (p value$_{Meta}$ 2.3 × 10^{-6}), but we chose to use a more conservative population-specific genome-wide p value threshold of significance (p value$_{Meta}$) of 2.05 × 10^{-7} for the SNPs in the meta-analysis,[6] and a p value$_{Meta}$ of 7.8 × 10^{-4} (64 classical HLA alleles tested in the meta-analysis) and 5.0 × 10^{-4} (99 amino acid positions tested in the meta-analysis) was used for classical HLA alleles and amino acids, respectively.

Identification of independent signals for each type of variants

We performed iterative conditional analyses using additive logistic models in each type of variant and meta-analyzed the results from the two populations, in each iteration step. In brief, we first identified the most strongly associated variant in the meta-analysis (lowest p value$_{Meta}$) and performed forward iterative conditional logistic regression to identify other independent signals. The significance threshold for the conditional analysis was set at a p value$_{Meta}$ < 5 × 10^{-6} for SNPs and a p value$_{Meta}$ < 1.0 × 10^{-3} for classical HLA alleles and amino acid residues given the number of tests for each type of variant. If another variant of the same type was significantly associated in the conditional analysis, we introduced this variant in the model as a covariate to evaluate the effect of the remaining variants, until no variant was significant.

Identification of amino acids accounting for independent signals

We performed conditional analyses with customized scripts in R[41] and PLINK 2.0 alpha version[37,38] to determine the amino acid residues accounting for the independently associated SNPs and classical HLA alleles by using iterative additive logistic regression models until the association of the top SNPs (or classical HLA alleles) were diminished and no other SNP (or classical HLA allele)
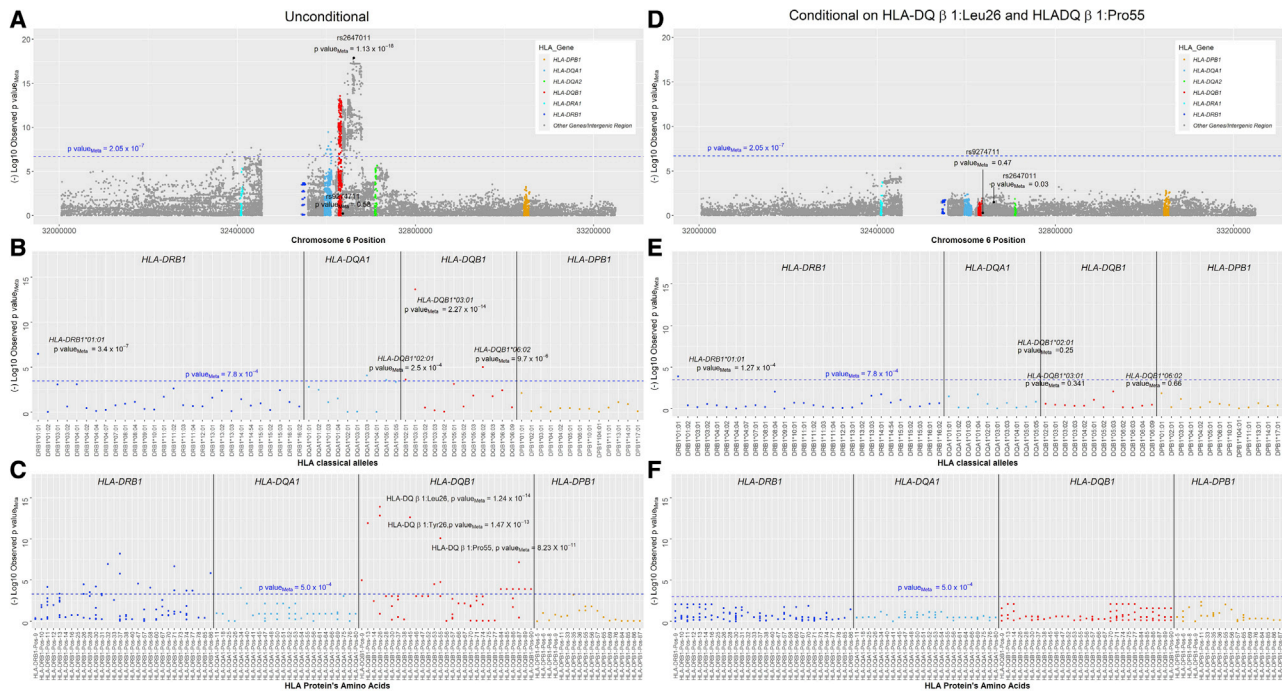
Figure 2. Results of the fixed-effect meta-analyses in 932 individuals with HCV spontaneous clearance and 2,532 persons with persistent HCV infection from African and European ancestry populations
(A) Association results of the unconditional analysis of SNPs.
(B) Association results of the unconditional analysis of HLA classical alleles.
(C) Association results of the unconditional analysis of amino acid residues.
(D–F) Results of the same variants after conditioning on HLADRβ1Leu26 and HLADRβ1Pro55. Each point corresponds to the p value$_{Meta}$ in each type of variant. The blue line and the p value$_{Meta}$ colored in blue represent the threshold level of fixed effects meta-analysis significance corrected by multiple comparisons.

was significantly associated in the region (Figure 2) at the conditional p value$_{Meta}$ threshold. In the conditional models, we included every associated amino acid as a covariate as well as HIV infection status and two principal components.[42] We prioritized the amino acids with the strongest associations and those identified as independently associated and their combinations. For each analysis we meta-analyzed the results from the two populations in each step.

Epitope prediction of associated classical HLA class II alleles containing causal amino acids

We selected the classical class II HLA alleles positively (clearance) and negatively (persistence) associated with HCV spontaneous clearance that represented the associated amino acids of interest. To investigate the capabilities of associated alleles to bind epitopes derived from the entire HCV polyprotein sequence (supplemental material and methods), we then performed T cell epitope prediction analyses. Using tools available in IEDB-Analysis Resources (see web resources),[43] we analyzed the HCV polyprotein sequence (GenBank:AFE48416.1) in combination with the associated clearance and persistence HLA alleles. Predictions were performed by applying the NetMHCIIpan 3.2 algorithm[44] and extracting all the predicted IC$_{50}$ values for each possible 15-mer spanning the entire HCV proteome (supplemental material and methods). IC$_{50}$ values in nanoMolar (nM) represent the concentration of the test peptide, which will displace 50% of a standard peptide from the HLA molecule in question. The lower the predicted IC$_{50}$ values are, the stronger the binding peptide-HLA allele combination is. To be considered an epitope, a cutoff of binding of ≤ 1000 nM has been applied, as pre-

viously reported.[45] The geometric mean ratio (described in supplemental material and methods) has been calculated only in peptides predicted to be epitopes in at least one of the classical HLA class II alleles analyzed in this study. The higher the geometric mean ratio the stronger is the binding affinity in clearance alleles versus persistence ones.

We aimed to address the global prediction patterns of these alleles using a PCA on the entire set of predicted peptides per allele after logarithmic transformation using the "prcomp" function in R.[41] We then performed a more in-depth analysis, considering separately the predicted binding distribution of each specific class II allele where each dot represents the predicted value of the specific peptide derived from the HCV proteome. Combined clearance IC$_{50}$ values (all clearance alleles) versus combined persistence (all persistent alleles) were compared using Mann-Whitney U test. Statistical analyses and graphs were performed using Prism 8 (GraphPad Software). A p value < 0.05 was considered as significant in this analysis.

Location of amino acid residues on HLA proteins

To delineate the physical location of the causal amino acids in the tertiary structure of the HLA proteins, we used X-ray structures from the Protein Data Bank (PDB).[46] This provides structural visualization of HLA-DQα1/DQβ1 heterodimer localizing the amino acids that account for the signal in MHC class II region. The accession code for the selected structure was PDB: 1jk8[47] for HLA-DQα1/DQβ1 and was visualized using PyMOL v. 2.3.2 (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC).

Study approval

Each individual study cohort obtained consent for genetic testing from their governing Institutional Review Board (IRB) and the Johns Hopkins School of Medicine Institutional Review Board approved the overall analysis and written informed consent was received from participants prior to inclusion in the study.

Results

Unconditional and conditional association analyses and meta-analyses

Unconditional and conditional association analysis with SNPs
There were 1,219 SNPs significantly associated with HCV spontaneous clearance in a 305.3 kb genomic segment of the MHC class II region (chr6:32,376,360–32,681,631) harboring the genes *HLA-DRA1*, *HLA-DRB1*, *HLA-DQA1*, and *HLA-DQB1* (Figure 2A, Table S2). The lead SNP (rs2647011, C>A, p value$_{Meta}$ = 1.13×10^{-18}) is in the intergenic region between *HLA-DQB1* and *HLA-DQA2* (Figure 2A). This SNP, located 1,333 bp from rs2647006, which was detected as the top signal in a previous trans-ethnic GWAS[6] and both SNPs are in LD in European (r^2 = 0.96) and African ancestry populations (r^2 = 0.97).

In the analysis conditioning on rs2647011, we detected a second independent association with rs9274711 (p value$_{Meta}$ conditional on rs2647011 = 4.18×10^{-6}, African population OR T allele = 0.71 [95% CI:0.59–0.86], p value = 4.1×10^{-4}, European population OR = 0.72 [95% CI:0.58–0.89], p value = 2.7×10^{-3}) (Figure S2B). This variant is not associated in the individual SNP analysis (p value$_{Meta}$ unconditional = 0.58, African population OR T allele = 0.99 [95% CI: 0.83–1.16], p value = 0.56, European population OR T allele = 0.94 [95% CI:0.77–1.15], p value = 0.83), as shown in Figure S2A and Figure 2. Further analyses determined that allele frequency of the C allele of rs2647011 is 0.37 and 0.39, in European and African ancestry, respectively. Similarly, the frequency of the T allele of rs9274711 is 0.26 and 0.22 those populations, respectively. Both SNPs are not in linkage disequilibrium (r^2 = 0.19 and r^2 = 0.17 in European and African ancestry populations, respectively). Moreover, a formal test of SNP×SNP interaction was not significant in any of the populations (p value rs2647011 × rs9274711 = 0.12 and p value = 0.86 in Europeans and African ancestry populations, respectively). Interestingly, rs2647011A allele (favorable for HCV clearance) is frequently in the same haplotype as rs9274711 T allele (non favorable for HCV clearance, Haplotype Frequency AT = 0.25 and 0.21 for European and African ancestry, respectively). The complementary alleles form a more frequent haplotype CA = 0.37 and 0.39 for European and African ancestry, respectively. Thus, it is plausible that the independent effect of the rs9274711T allele is only present when we condition on rs2647011. When including these two variants in the association analysis, no other SNP was significantly associated in the MHC Class II region in the meta-analysis (Figure S2C).

Unconditional and conditional association analysis with classical HLA alleles

SNP data were then used to impute HLA alleles. The strongest signal of association was *HLA-DQB1*03:01* (p value$_{Meta}$ = 2.27×10^{-14}), while associations with lower statistical significance were found with *HLA-DRB1*01:01* (p value$_{Meta}$ = 3.4×10^{-7}), *HLA-DQB1*06:02* (p value$_{Meta}$ = 9.7×10^{-6}), *HLA-DQB1*05:01* (p value$_{Meta}$ = 7.3×10^{-4}), *HLA-DQB1*02:01* (p value$_{Meta}$ = 2.5×10^{-4}), *HLA-DQA1*03:03* (p value$_{Meta}$ = 8.5×10^{-5}), *HLA-DQA1*05:01* (p value$_{Meta}$ = 2.6×10^{-4}), and *HLA-DQA1*05:05* (p value$_{Meta}$ = 4.4×10^{-4}) (Figure 2B, Table S3).

In the conditional analysis of classical HLA alleles, *HLA-DQB1*03:01* and *HLA-DRB1*01:01* (p value$_{Meta}$ conditional on *HLA-DQB1*03:01* = 5.61×10^{-9}) were independently associated with HCV spontaneous clearance. Including those two alleles in the model eliminated all residual association of classical HLA alleles (Figure S3).

Unconditional and conditional association analysis with amino acid residues

We then identified specific amino acid residues in the HLA-DQβ1 and HLA-DRβ1 proteins associated with HCV spontaneous clearance with high significance (Figure 2C, Table S4). In HLA-DQβ1, the strongest association was HLA-DQβ1-26; among the three possible residues (Leu, Tyr, and Gly), the HLA-DQβ1Leu26 was associated with 35%–45% lower likelihood of clearance (p value$_{Meta}$ = 1.24×10^{-14}), whereas HLA-DQβ1Tyr26 was associated with 64%–81% higher likelihood of HCV spontaneous clearance (p value$_{Meta}$ = 1.47×10^{-13}); this finding was consistent across ancestry groups (Tables 2 and S4).

The conditional analysis of amino acids also identified two independent signals, corresponding to the lead amino acid residue HLA-DQβ1Leu26 and HLA-DQβ1Pro55 (p value$_{Meta}$ unconditional = 8.23×10^{-11}, p value$_{Meta}$ conditional on HLA-DQβ1Leu26 = 2.8×10^{-4}). After including these two amino acid residues in the model, no other amino acids in the MHC class II region were significantly associated (Figures 2F and S4).

For the independently associated SNPs, classical alleles, and amino acids, an analysis stratified by HIV infection status demonstrated no difference in associations between those with and without HIV (Table S5). We also ran the conditional analysis with HLA-DQβ1Tyr26 and identified two residual signals given by HLA-DQβ1Ala71 (p value$_{Meta}$ = 9.94×10^{-8}) and by HLA-DRβ1Glu28 (p value$_{Meta}$ = 1.64×10^{-6}). After conditioning on HLA-DQβ1Tyr26 and HLA-DQβ1Ala71, we observed no additional amino acids associated in the MHC class II region.

Amino acid residues explain SNP and classical HLA allele associations with HCV clearance

After analyzing models including SNPs and classical HLA alleles and conditioning on the associated amino acid residues and combination of residues (Tables S6 and S7), we found that HLA-DQβ1Leu26 and HLA-DQβ1Pro55 reduced the independent SNP association of rs2647011 (p value$_{Meta}$ = 0.03)

Table 2. Results of the unconditional association analysis and fixed effect meta-analysis of the two amino acids in HLA-DQβ1 protein that explain the association in the MHC class II region, HLA-DQβ1Leu26, and HLA DQβ1Pro55 and alternatives HLA-DQβ1Tyr26 and HLA DQβ1Ala71 in the European and African ancestry populations

Amino acid position	Effect allele/ non effect allele	European ancestry population (N = 1,600)				African ancestry population (N = 1,869)				p value$_{Meta}$
		Frequency effect allele				Frequency effect allele				
		Clearance (n = 594)	Persistence (n = 1,006)	OR (95% CI)	p value	Clearance (n = 340)	Persistence (n = 1,529)	OR (95% CI)	p value	
26*	L/G,Y	0.45	0.55	0.64 (0.54–0.75)	8.63×10^{-8}	0.35	0.49	0.55 (0.45–0.68)	2.61×10^{-8}	1.25×10^{-14}
26	L,Y/G	0.77	0.78	0.92 (0.76–1.11)	3.96×10^{-1}	0.67	0.7	0.86 (0.7–1.07)	1.89×10^{-1}	1.29×10^{-1}
26*	Y/G,L	0.31	0.23	1.64 (1.36–1.98)	1.62×10^{-7}	0.32	0.21	1.81 (1.45–2.27)	1.81×10^{-7}	1.47×10^{-13}
55	L/R,P	0.09	0.14	0.59 (0.45–0.79)	3.04×10^{-4}	0.07	0.08	0.82 (0.56–1.21)	3.20×10^{-1}	9.00×10^{-4}
55*	L,P/R	0.56	0.52	1.25 (1.07–1.48)	6.10×10^{-3}	0.43	0.36	1.42 (1.16–1.74)	7.57×10^{-4}	1.69×10^{-5}
55*	P/R,L	0.47	0.38	1.51 (1.28–1.77)	8.79×10^{-7}	0.36	0.27	1.6 (1.29–1.99)	2.11×10^{-5}	8.23×10^{-11}
71	D,K,T/A	0.83	0.86	0.76 (0.61–0.95)	1.73×10^{-2}	0.76	0.8	0.81 (0.64–1.02)	7.67×10^{-2}	3.23×10^{-3}
71	K/A,D,T	0.09	0.14	0.59 (0.44–0.79)	3.43×10^{-4}	0.07	0.08	0.82 (0.56–1.2)	3.17×10^{-1}	1.00×10^{-3}
71	T/A,D,K	0.67	0.64	1.06 (0.88–1.27)	5.49×10^{-1}	0.6	0.61	0.9 (0.73–1.12)	3.47×10^{-1}	8.33×10^{-1}
71	D/A,K,T	0.03	0.03	0.92 (0.58–1.46)	7.21×10^{-1}	0.09	0.08	1 (0.7–1.43)	9.98×10^{-1}	7.96×10^{-1}

*Asterisk indicates the amino acids significantly associated in the fixed-effect unconditional meta-analyses. Single letter and three letter amino acid code: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; T, Thr; S, Ser; Y, Tyr; V, Val; W, Trp.

and rs9274711 (p value$_{Meta}$ = 0.47) (Figure 2D), indicating that the amino acids likely explain the association of rs2647011. Similarly, conditioning on those two key amino acids diminished the association of *HLA-DQB1*03:01* (p value$_{Meta}$ = 0.341) and reduced the association with *HLA-DRB1*01:01* (p value$_{Meta}$ = 1.27 × 10^{-4}) (Figure 2E) further underscoring the role of those amino acids explaining the strong association in this region. Conditioning on HLA-DQβ1Tyr26 and HLA-DQβ1Ala71 also reduced the classical HLA associations, but to a lesser extent (rs2647011 = 3.98 × 10^{-3}; rs9274711 = 0.03; *HLA-DQB1*03:01* = 0.68, and *HLA-DRB1*01:01* = 1.86 × 10^{-3}) (Table S6).

In summary, two amino acids in HLA-DQβ1 protein, HLA-DQβ1Leu26 and HLA DQβ1Pro55 (or alternatively HLA-DQβ1Tyr26 and HLA DQβ1Ala71), explain the SNP and classical allele associations with HCV spontaneous clearance in the MHC class II region.

Epitope predictions of associated classical HLA class II alleles containing causal amino acids

HLA class II binding predictions were performed with classical HLA alleles for clearance (*DRB1*01:01*, *DQA1*03:03/DQB1*03:01*, *DQA1*05:05/DQB1*03:01*) and for persistence (*DQA1*01:02/DQB1*06:02* and *DQA1*05:01/DQB1*02:01*). *HLA-DQB1* clearance/persistence groups of associated alleles were stratified by the putative causal amino acid in position 26 of HLA-DQβ1, as described above and shown in Table 3.

Clearance alleles were separated from the persistence alleles by two principal components derived from the binding prediction, chiefly PCA1 (66% variance, Figure 3A). The separation was augmented when analyzing only the alleles of the *HLA-DQ* locus (Figure 3B). Significantly higher frequencies of predicted epitopes were observed in the amino acid sequences of clearance compared to persistence alleles, both in terms of the fractions predicted to bind with an IC$_{50}$ value of ≤1000 nM and the geometric means of the combined IC$_{50}$ values of the clearance allele group (geometric mean: 761.7 nM) versus the persistent group (2,321 nM, p value = 1.35 × 10^{-38} by Mann-Whitney U test) (Figure 3C, Table S8). The observation was maintained when considering only the *HLA-DQ* locus which showed higher frequencies of predicted epitopes (260 versus 141) and binding affinity compared to persistence alleles (geometric mean: 1,107 nM versus 2,321 nM, p value = 4.07 × 10^{-18}).

Overall, classical HLA class II alleles clustered based on their predicted binding pattern capabilities. Alleles positively associated with HCV spontaneous clearance had more predicted HLA binders, potentially recognized by T cells.

Location of amino acid residues accounting for the association on MHC class II

Amino acid in position 26 of the HLA-DQβ1 chain (HLA-DQβ26) is located in the floor of the putative peptide

Table 3. Amino acid positions and residues associated with HCV spontaneous clearance, significantly associated HLA class II alleles containing them and location of the amino acids in the pockets of each protein

HLA classical allele associated with HCV clearance/persistence[&]	Amino acid residues associated with HCV clearance/persistence and location in pockets of HLA-DQβ1 and HLA-DRβ1 proteins											
Pocket in the protein	P6	P4	P4					P1	P1	P1	P1	P1
HLA-DQB1	9	13	26	45	53	55	84	85	86	87	89	90
DQB1*03:01	Y	A*	Y*	E*	L	P+	Q	L	E	L	T	T
DQB1*06:02[&]	F*&	G&	L*+&	G&	Q&	R*&	E&	V&	A&	F*&	G&	I&
DQB1*02:01[&]	Y	G&	L*+&	G*&	L	L*	Q	L	E	L	T	T
	P6	P4	P4	P6			P1	P7	P9	P4	P4	P1
HLA-DRB1	11	13	26	28	30	31	32	47	58	71	74	86
DRB1*01:01	L	F	L	E	C	I	Y	Y	A&	R	A	G

Amino acids in alleles are coded in a single-letter code. All amino acids in the table reached significance level in the meta-analysis and were associated with clearance except for those with the & symbol which were associated with persistence. Asterisks (*) indicate those that accounted the association of the HLA allele and those with + symbol accounted for the observed association with SNP in MHC class II region. Single letter and three letter amino acid code: A, Ala; C, Cys; E, Glu; F, Phe; G, Gly; I, Ile; L, Leu; P, Pro; Q, Gln; R, Arg; S, Ser; Y, Tyr. Abbreviations: P: Pocket.

binding groove in the second hypervariable region as part of the peptide-binding pocket P4 (Table 3, Figure 4) and can harbor a hydrophobic Leucine versus a hydrophilic polar non-charged tyrosine (and in a very small frequency glycine). Likewise, HLA-DQβ55 forms part of a peptide-binding region in the HLA-DQ molecule, and the change from arginine to proline implies a physio-chemical shift from hydrophilic to hydrophobic, but we cannot verify the effect of these changes on the affinity of the HLA molecule.

Discussion

This study maps the association of HLA class II with HCV spontaneous clearance to two amino acid positions in the peptide binding regions of HLA-DQβ. The finding corresponds with HCV peptide class II binding predictions and is sustained in persons of both African and European ancestry. Greater understanding of how some individuals spontaneously clear HCV infection might be useful for designing urgently needed vaccines that promote spontaneous clearance of HCV infection.[48–50]

The results of our study confirm the association of HCV spontaneous clearance with classical HLA alleles and SNPs that we and others have already reported. For example, Duggal et al. described the association of HLA-DQB1* 03:01 with HCV spontaneous clearance in a European ancestry population.[5,9] Similarly, this allele has being identified in other studies with individuals of European ancestry,[9,51–53] two ethnically distinct Chinese populations,[54] and in families from Egypt.[55] While those studies provided unmistakable evidence of an association with the HLA locus, no prior study has been able to identify the causal variants. In this study, we identify two single amino acid residues in the peptide binding groove of the protein HLA-DQβ1 (Leu26 and Pro55) that explain the as-

sociation of both individual SNPs and classical HLA alleles with HCV clearance.

The amino acid residues within the peptide-binding site of the β-chain of HLA-DQβ1 molecule play an important role in the selectivity of peptide binding. For example, the presence of Leu26 in pocket 4 in the HLA-DQβ1 chain that we found to be associated with HCV clearance also has been identified in traits such as the anticentromere autoantibody response in a small group of European individuals with systemic sclerosis,[56] as well as autoantibody responses such as anti-ro (SS-A) response in patients of European and African ancestry with systemic lupus erythematosus (SLE) or Sjögren's syndrome,[57] and anti-ribosomal P antibodies in another large multiethnic cohort of patients with SLE including individuals of European and Asian ancestry.[58] Similarly, an association with rs1130380, the SNP that causes the amino acid substitution Arg55Pro in the HLA-DQβ1 was previously detected in a genome-wide association study for chronic HCV infection and healthy controls in a Japanese cohort.[59] However, the specific association analysis of the amino acid residue and HCV clearance was not implemented in the Japanese cohort as we did in our analysis.

Both codons we identified as associated with HCV clearance form part of a peptide-binding groove in the HLA-DQ molecule. This finding suggests that HLA-DQ binding may be critical to the net effectiveness of HCV immune responses. This inference is supported by our formal, independent HLA binding analysis. Indeed, the HLA-DQ haplotypes linked to clearance (HLA-DQA1*03:03/HLA-DQB1* 03:01 and HLA- DQA1*05:05/HLA- DQB1*03:01) bound a greater number of predicted HCV peptides and showed binding with higher affinity compared to those linked to persistence (HLA-DQA1*01:02/HLA-DQB1*06:02 and HLA-DQA1*05:01/HLA-DQB1*02:01). Additionally, this finding agrees with Cramp et al. who demonstrated more robust CD4[+] T cell responses in those who were HLA-DQB1*
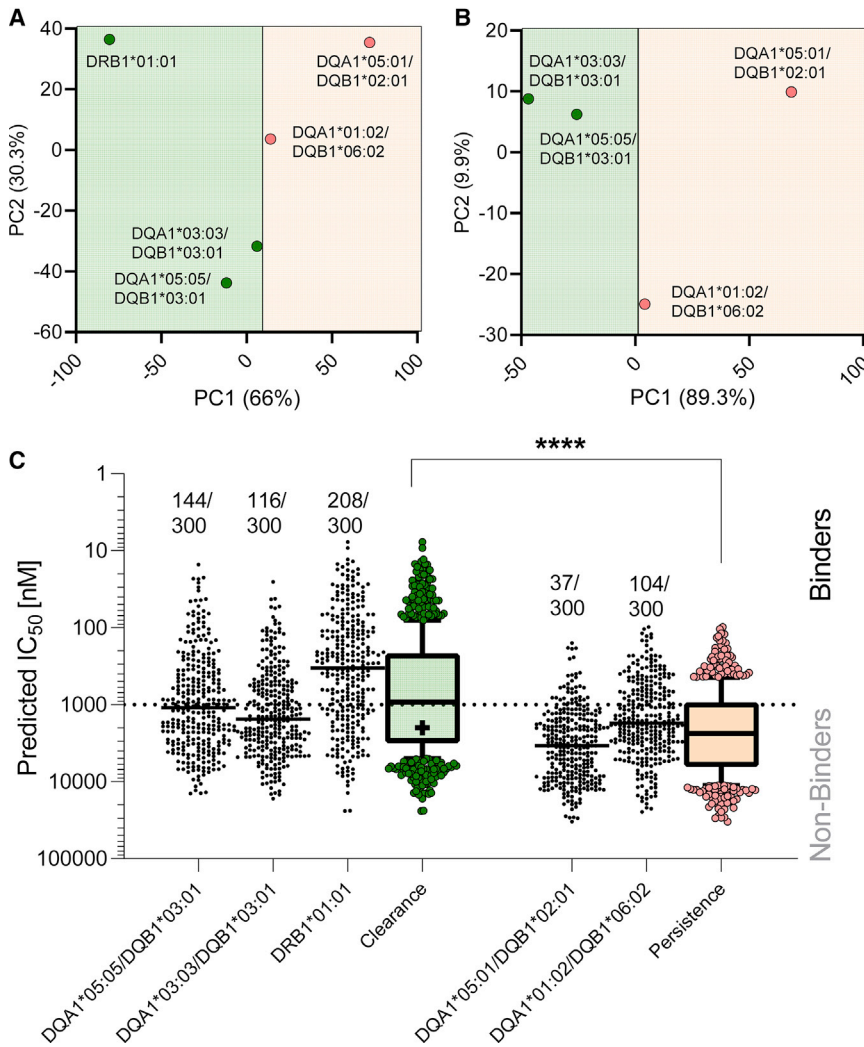
Figure 3. T cell prediction analysis of HLA class II alleles associated with HCV spontaneous clearance

Predicted IC_{50} values of all the possible 15-mer peptides encompassing the entire HCV proteome in combination with HLA class II clearance (*HLA-DRB1*01:01*, *HLA-DQA1*03:03/HLA-DQB1*03:01*, *HLA-DQA1*05:05/HLA-DQB1*03:01*) and persistent alleles (*HLA-DQA1*01:02/HLA-DQB1*06:02* and *HLA-DQA1*05:01/HLA-DQB1*02:01*).

(A) Principal component analysis (PCA) of the entire prediction distribution based on clearance (green) or persistence alleles (red).

(B) PCA based on HLA class II alleles related to the *HLA-DQ* locus only.

(C) Predicted IC_{50} values for each peptide-HLA classical allele combination and for the combined clearance (green) or persistence alleles (red). Epitopes were defined as peptides with a predicted $IC_{50} \leq 1,000$ nM (above dotted line, denoted as "binders") and the numbers of predicted epitopes are shown as a fraction of those considered for each of the HLA class II classical alleles in analysis. Combined data in (C) are expressed as whiskers 10–90 percentile, the plus sign representing the mean. ****p value < 0.0001 by Mann-Whitney U test.

03:01 positive, as well as a strong association with spontaneous resolution of HCV infection.[60] Likewise, Kovacs et al. demonstrated that *HLA-DQB1*03:01* was associated with increased CD8$^+$ T cell activation.[61]

Even though several amino acids forming classical HLA alleles were associated with HCV spontaneous clearance, their association was highly dependent on HLA-DQβ1 Leu26 and Pro55 as well as the association of the classical HLA alleles itself. The direction of the effect of the association of each residue in position 26 of HLA-DQβ1 is highly concordant with the associated classical allele containing them. Leu26 tags the two classical HLA alleles negatively associated with clearance (*HLA-DQB1*06:02* and *HLA-DQB1*02:01*) and Tyr 26 is contained in the classical HLA allele that favors HCV spontaneous clearance, *HLA-DQB1*03:01*. Furthermore, HLA-DQβ26 is the main amino acid position that is different between the clearance and persistence alleles including in our epitope binding assay. Consequently, it is reasonable to understand the association considering this differential affinity for HCV peptides. Although additional work would be needed to prove the inference, there are no experimental models of HCV clear-

ance. Nonetheless, *ex vivo* studies that directly demonstrate differential binding would confirm this inference.

We present consistent fine-mapping results across ancestral populations. This study was limited by the lack of HCV genotypes, which is an additional factor associated with HCV spontaneous clearance, and lack of imputation of the *HLA-DQA* alleles and binding affinity assays in the African populations restricting the generalization of the associations. As with other complex diseases, the penetrance of these HLA alleles is not 100%, other genes also play a role[62,63] and HCV outcome may be mediated or interact with age and initial HCV viral load.

We used HIBAG,[32] a robust imputation program that references large databases with high genetic diversity and has demonstrated sucess in multi-ethnic benchmarking studies.[33,64–66] Moreover, we confirmed that the allelic frequency of the HIBAG imputed alleles is in concordance with that from other populations reported in large databases such as the Allele Net Frequency Database[67] and Immuno Polymorphism Database,[68] which supports the accuracy of the imputation. Additionally, we performed a sensitivity analysis in a subsample of 31 individuals by estimating the HLA classical alleles using HLA-HD, an HLA imputation method based on next generation sequencing data (NGS).[69] We observed high imputation accuracy for *HLA-DRB1* (97.6%), *HLA-DQA1* (100%), reaching 98% for
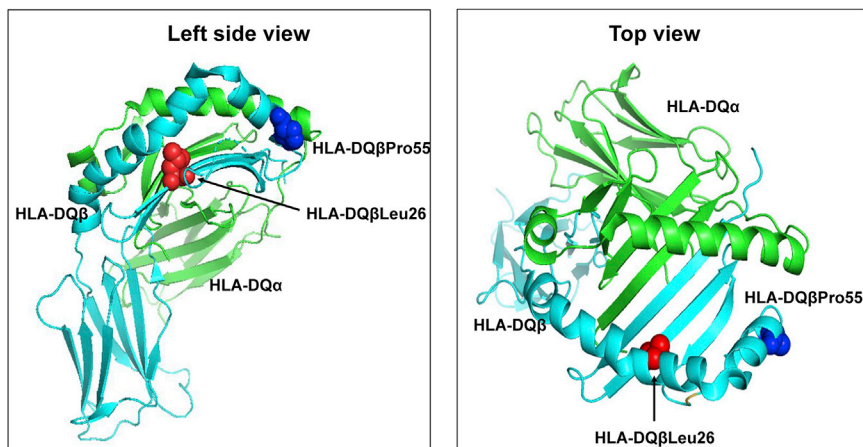
Figure 4. 3D ribbon models for HLA-DQ protein

Molecular structure of the extracellular region of HLA-DQ molecule displaying the peptide binding groove along with the associated amino acid residues that totally accounted for the observed signal in the MHC class II region in European and African ancestry groups. The Protein Data Bank: 1jk8 was used to generate the 3D structures HLA-DQ using PyMOL (The PyMOL Molecular Graphics System, v.2.0 Schrödinger, LLC). Abbreviations: Pro, proline; Leu, leucine.

HLA-DQB1 and *HLA-DPB1* genes (Table S9). The observed that differences in 2 out of 3 discordant alleles were at the second field of resolution (e.g., *DRB1*13:01*, *DRB1*13:02*) and likely reflect differences in the reference sequences used and their ability to capture less frequent alleles. None of the three discordant alleles resulted in changes to the associated top amino acids. We also included HIV infection status in the model to account for any effect of HIV infection in HCV clearance.

In summary, fine mapping of the MHC region associated with the HCV spontaneous clearance revealed the key amino acids in the HLA-DQβ1 protein that explain the association in persons of both European and African ancestry. The mechanistic advance is that we have mapped an association of a large genomic region to two amino acids in HLA-DQβ1 and specifically to their binding with HCV peptides. This advance in understanding of natural recovery might set the stage for much needed enhancement of HCV vaccine efforts which are urgently needed as new HCV infections in the United States have tripled in the past 10 years.

Data and code availability

The datasets used in this study are available at dbGaP accession numbers dbGaP:phs000454.v1.p1 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000454.v1.p1) and dbGaP: phs000248.v1.p1 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000248.v1.p1). The reference panel derived from the "Consortium on Asthma among African-ancestry Populations in the Americas" is avalaible at dbGaP accession number dbGaP:phs001123.v1.p1 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001123.v1.p1). The HCV polyprotein sequence is available at GenBank: AFE48416.1 (https://www.ncbi.nlm.nih.gov/protein/AFE48416.1). Molecular structure used to generate HLA-DQ 3D structures is available at PDB:1jk8 (https://www.rcsb.org/structure/1jk8).

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2022.01.001.

Declaration of interests

J.C.P. has received research grant support to her institution from Gilead Sciences, Merck, and Abbvie and has served on an advisory board for Gilead Sciences and Theratechnolgies. S.H.M. have received speaker fees from Gilead Sciences not related to this work. A.H.K. serves on the Data Monitoring Committee for Kintor Pharmaceutical.

Web resources

GenBank, https://www.ncbi.nlm.nih.gov/genbank/
HLA database used for imputation, http://hla.alleles.org/alleles/p_groups.html

IEDB-Analysis Resources, http://tools.iedb.org/main
PyMOL Molecular Graphics System program, https://pymol.org/2
RCSB Protein Data Bank, http://www.rcsb.org/pdb/home/home.do

References

1. Anonymous Global Hepatitis Report 2017 (2017). Geneva: World Health Organization. https://www.who.int/publications/i/item/global-hepatitis-report-2017.

2. Polaris Observatory HCV Collaborators (2017). Global prevalence and genotype distribution of hepatitis C virus infection in 2015: a modelling study. Lancet Gastroenterol. Hepatol. 2, 161–176.

3. Lavanchy, D. (2009). The global burden of hepatitis C. Liver Int. 29 (Suppl 1), 74–81.

4. Thomas, D.L., Astemborski, J., Rai, R.M., Anania, F.A., Schaeffer, M., Galai, N., Nolt, K., Nelson, K.E., Strathdee, S.A., Johnson, L., et al. (2000). The natural history of hepatitis C virus infection: host, viral, and environmental factors. JAMA 284, 450–456.

5. Duggal, P., Thio, C.L., Wojcik, G.L., Goedert, J.J., Mangia, A., Latanich, R., Kim, A.Y., Lauer, G.M., Chung, R.T., Peters, M.G., et al. (2013). Genome-wide association study of spontaneous resolution of hepatitis C virus infection: data from multiple cohorts. Ann. Intern. Med. 158, 235–245.

6. Vergara, C., Thio, C.L., Johnson, E., Kral, A.H., O'Brien, T.R., Goedert, J.J., Mangia, A., Piazzolla, V., Mehta, S.H., Kirk, G.D., et al. (2019). Multi-Ancestry Genome-Wide Association Study of Spontaneous Clearance of Hepatitis C Virus. Gastroenterology 156, 1496–1507.e7.

7. Prokunina-Olsson, L., Muchmore, B., Tang, W., Pfeiffer, R.M., Park, H., Dickensheets, H., Hergott, D., Porter-Gill, P., Mumy, A., Kohaar, I., et al. (2013). A variant upstream of IFNL3 (IL28B) creating a new interferon gene IFNL4 is associated with impaired clearance of hepatitis C virus. Nat. Genet. 45, 164–171.

8. Aka, P.V., Kuniholm, M.H., Pfeiffer, R.M., Wang, A.S., Tang, W., Chen, S., Astemborski, J., Plankey, M., Villacres, M.C., Peters, M.G., et al. (2014). Association of the IFNL4-ΔG Allele With Impaired Spontaneous Clearance of Hepatitis C Virus. J. Infect. Dis. 209, 350–354.

9. Huang, H., Duggal, P., Thio, C.L., Latanich, R., Goedert, J.J., Mangia, A., Cox, A.L., Kirk, G.D., Mehta, S., Aneja, J., et al. (2017). Fine-mapping of genetic loci driving spontaneous clearance of hepatitis C virus infection. Sci. Rep. 7, 15843, 2.

10. McKiernan, S.M., Hagan, R., Curry, M., McDonald, G.S., Kelly, A., Nolan, N., Walsh, A., Hegarty, J., Lawlor, E., and Kelleher, D. (2004). Distinct MHC class I and II alleles are associated with hepatitis C viral clearance, originating from a single source. Hepatology 40, 108–114.

11. Thio, C.L., Gao, X., Goedert, J.J., Vlahov, D., Nelson, K.E., Hilgartner, M.W., O'Brien, S.J., Karacki, P., Astemborski, J., Carrington, M., and Thomas, D.L. (2002). HLA-Cw*04 and hepatitis C virus persistence. J. Virol. 76, 4792–4797.

12. Hraber, P., Kuiken, C., and Yusim, K. (2007). Evidence for human leukocyte antigen heterozygote advantage against hepatitis C virus infection. Hepatology 46, 1713–1721.

13. Kim, A.Y., Kuntzen, T., Timm, J., Nolan, B.E., Baca, M.A., Reyor, L.L., Berical, A.C., Feller, A.J., Johnson, K.L., Schulze zur Wiesch, J., et al. (2011). Spontaneous control of HCV is associated with expression of HLA-B 57 and preservation of targeted epitopes. Gastroenterology 140, 686–696.e1.

14. Kuniholm, M.H., Kovacs, A., Gao, X., Xue, X., Marti, D., Thio, C.L., Peters, M.G., Terrault, N.A., Greenblatt, R.M., Goedert, J.J., et al. (2010). Specific human leukocyte antigen class I and II alleles associated with hepatitis C virus viremia. Hepatology 51, 1514–1522.

15. Alric, L., Fort, M., Izopet, J., Vinel, J.P., Charlet, J.P., Selves, J., Puel, J., Pascal, J.P., Duffaut, M., and Abbal, M. (1997). Genes of the major histocompatibility complex class II influence the outcome of hepatitis C virus infection. Gastroenterology 113, 1675–1681.

16. Fanning, L.J., Levis, J., Kenny-Walsh, E., Wynne, F., Whelton, M., and Shanahan, F. (2000). Viral clearance in hepatitis C (1b) infection: relationship with human leukocyte antigen class II in a homogeneous population. Hepatology 31, 1334–1337.

17. Wojcik, G.L., Thio, C.L., Kao, W.H., Latanich, R., Goedert, J.J., Mehta, S.H., Kirk, G.D., Peters, M.G., Cox, A.L., Kim, A.Y., et al. (2014). Admixture analysis of spontaneous hepatitis C virus clearance in individuals of African descent. Genes Immun. 15, 241–246.

18. Vlahov, D., Muñoz, A., Anthony, J.C., Cohn, S., Celentano, D.D., and Nelson, K.E. (1990). Association of drug injection patterns with antibody to human immunodeficiency virus type 1 among intravenous drug users in Baltimore, Maryland. Am. J. Epidemiol. 132, 847–856.

19. Cox, A.L., Netski, D.M., Mosbruger, T., Sherman, S.G., Strathdee, S., Ompad, D., Vlahov, D., Chien, D., Shyamala, V., Ray, S.C., and Thomas, D.L. (2005). Prospective evaluation of community-acquired acute-phase hepatitis C virus infection. Clin. Infect. Dis. 40, 951–958.

20. Cramp, M.E., Carucci, P., Underhill, J., Naoumov, N.V., Williams, R., and Donaldson, P.T. (1998). Association between HLA class II genotype and spontaneous clearance of hepatitis C viraemia. J. Hepatol. 29, 207–213.

21. Hilgartner, M.W., Donfield, S.M., Willoughby, A., Contant, C.F., Jr., Evatt, B.L., Gomperts, E.D., Hoots, W.K., Jason, J., Loveland, K.A., McKinlay, S.M., et al. (1993). Hemophilia growth and development study. Design, methods, and entry data. Am. J. Pediatr. Hematol. Oncol. 15, 208–218.

22. Mangia, A., Gentile, R., Cascavilla, I., Margaglione, M., Villani, M.R., Stella, F., Modola, G., Agostiano, V., Gaudiano, C., and Andriulli, A. (1999). HLA class II favors clearance of HCV infection and progression of the chronic liver damage. J. Hepatol. 30, 984–989.

23. Goedert, J.J., Chen, B.E., Preiss, L., Aledort, L.M., and Rosenberg, P.S. (2007). Reconstruction of the hepatitis C virus epidemic in the US hemophilia population, 1940-1990. Am. J. Epidemiol. 165, 1443–1453.

24. Tobler, L.H., Bahrami, S.H., Kaidarova, Z., Pitina, L., Winkelman, V.K., Vanderpool, S.K., Guiltinan, A.M., Cooper, S., Busch, M.P., and Murphy, E.L. (2010). A case-control study of factors associated with resolution of hepatitis C viremia in former blood donors (CME). Transfusion 50, 1513–1523.

25. Edlin, B.R., Shu, M.A., Winkelstein, E., Des Jarlais, D.C., Busch, M.P., Rehermann, B., O'Brien, T.R., Talal, A.H., Tobler, L.H., Zeremski, M., and Beeder, A.B. (2009). More rare birds, and the occasional swan. Gastroenterology 136, 2412–2414.

26. Kuniholm, M.H., Gao, X., Xue, X., Kovacs, A., Marti, D., Thio, C.L., Peters, M.G., Greenblatt, R.M., Goedert, J.J., Cohen,

M.H., et al. (2011). The relation of HLA genotype to hepatitis C viral load and markers of liver fibrosis in HIV-infected and HIV-uninfected women. J. Infect. Dis. *203*, 1807–1814.

27. Khakoo, S.I., Thio, C.L., Martin, M.P., Brooks, C.R., Gao, X., Astemborski, J., Cheng, J., Goedert, J.J., Vlahov, D., Hilgartner, M., et al. (2004). HLA and NK cell inhibitory receptor genes in resolving hepatitis C virus infection. Science *305*, 872–874.

28. Kral, A.H., Bluthenthal, R.N., Lorvick, J., Gee, L., Bacchetti, P., and Edlin, B.R. (2001). Sexual transmission of HIV-1 among injection drug users in San Francisco, USA: risk-factor analysis. Lancet *357*, 1397–1401.

29. Kral, A.H., Lorvick, J., Gee, L., Bacchetti, P., Rawal, B., Busch, M., and Edlin, B.R. (2003). Trends in human immunodeficiency virus seroincidence among street-recruited injection drug users in San Francisco, 1987-1998. Am. J. Epidemiol. *157*, 915–922.

30. Horton, R., Wilming, L., Rand, V., Lovering, R.C., Bruford, E.A., Khodiyar, V.K., Lush, M.J., Povey, S., Talbot, C.C., Jr., Wright, M.W., et al. (2004). Gene map of the extended human MHC. Nat. Rev. Genet. *5*, 889–899.

31. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet. *48*, 1284–1287.

32. Zheng, X., Shen, J., Cox, C., Wakefield, J.C., Ehm, M.G., Nelson, M.R., and Weir, B.S. (2014). HIBAG–HLA genotype imputation with attribute bagging. Pharmacogenomics J. *14*, 192–200.

33. Vince, N., Limou, S., Daya, M., Morii, W., Rafaels, N., Geffard, E., Douillard, V., Walencik, A., Boorgula, M.P., Chavan, S., et al.; CAAPA (2020). Association of HLA-DRB1*09:01 with tIgE levels among African-ancestry individuals with asthma. J. Allergy Clin. Immunol. *146*, 147–155.

34. Daya, M., Rafaels, N., Brunetti, T.M., Chavan, S., Levin, A.M., Shetty, A., Gignoux, C.R., Boorgula, M.P., Wojcik, G., Campbell, M., et al.; CAAPA (2019). Association study in African-admixed populations across the Americas recapitulates asthma risk loci in non-African populations. Nat. Commun. *10*, 880–887.

35. Levin, A.M., Adrianto, I., Datta, I., Iannuzzi, M.C., Trudeau, S., McKeigue, P., Montgomery, C.G., and Rybicki, B.A. (2014). Performance of HLA allele prediction methods in African Americans for class II genes HLA-DRB1, -DQB1, and -DPB1. BMC Genet. *15*, 72, 72.

36. Khor, S.S., Yang, W., Kawashima, M., Kamitsuji, S., Zheng, X., Nishida, N., Sawai, H., Toyoda, H., Miyagawa, T., Honda, M., et al. (2015). High-accuracy imputation for HLA class I and II genes based on high-resolution SNP data of population-specific references. Pharmacogenomics J. *15*, 530–537.

37. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7–8.

38. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

39. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.

40. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190–2191.

41. R Core Team (2013). A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. http://www.r-project.org/.

42. Tian, C., Hromatka, B.S., Kiefer, A.K., Eriksson, N., Noble, S.M., Tung, J.Y., and Hinds, D.A. (2017). Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. Nat. Commun. *8*, 599, 5.

43. Dhanda, S.K., Mahajan, S., Paul, S., Yan, Z., Kim, H., Jespersen, M.C., Jurtz, V., Andreatta, M., Greenbaum, J.A., Marcatili, P., et al. (2019). IEDB-AR: immune epitope database-analysis resource in 2019. Nucleic Acids Res. *47* (W1), W502–W506.

44. Jensen, K.K., Andreatta, M., Marcatili, P., Buus, S., Greenbaum, J.A., Yan, Z., Sette, A., Peters, B., and Nielsen, M. (2018). Improved methods for predicting peptide binding affinity to MHC class II molecules. Immunology *154*, 394–406.

45. Peters, B., Nielsen, M., and Sette, A. (2020). T Cell Epitope Predictions. Annu. Rev. Immunol. *38*, 123–145.

46. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

47. Lee, K.H., Wucherpfennig, K.W., and Wiley, D.C. (2001). Structure of a human insulin peptide-HLA-DQ8 complex and susceptibility to type 1 diabetes. Nat. Immunol. *2*, 501–507.

48. Thomas, D.L. (2019). Global Elimination of Chronic Hepatitis. N. Engl. J. Med. *380*, 2041–2050.

49. Bailey, J.R., Barnes, E., and Cox, A.L. (2019). Approaches, Progress, and Challenges to Hepatitis C Vaccine Development. Gastroenterology *156*, 418–430.

50. Cox, A.L. (2020). Challenges and Promise of a Hepatitis C Virus Vaccine. Cold Spring Harb. Perspect. Med. *10*, a036947. https://doi.org/10.1101/cshperspect.a036947.

51. Minton, E.J., Smillie, D., Neal, K.R., Irving, W.L., Underwood, J.C., James, V.; and Members of the Trent Hepatitis C Virus Study Group (1998). Association between MHC class II alleles and clearance of circulating hepatitis C virus. J. Infect. Dis. *178*, 39–44.

52. Thio, C.L., Thomas, D.L., Goedert, J.J., Vlahov, D., Nelson, K.E., Hilgartner, M.W., O'Brien, S.J., Karacki, P., Marti, D., Astemborski, J., and Carrington, M. (2001). Racial differences in HLA class II associations with hepatitis C virus outcomes. J. Infect. Dis. *184*, 16–21.

53. Harris, R.A., Sugimoto, K., Kaplan, D.E., Ikeda, F., Kamoun, M., and Chang, K.M. (2008). Human leukocyte antigen class II associations with hepatitis C virus clearance and virus-specific CD4 T cell response among Caucasians and African Americans. Hepatology *48*, 70–79.

54. Huang, J., Xu, R., Wang, M., Liao, Q., Huang, K., Shan, Z., You, Q., Li, C., Rong, X., and Fu, Y. (2019). Association of HLA-DQB1*03:01 and DRB1*11:01 with spontaneous clearance of hepatitis C virus in Chinese Li ethnicity, an ethnic group genetically distinct from Chinese Han ethnicity and infected with unique HCV subtype. J. Med. Virol. *91*, 1830–1836.

55. El-Bendary, M., Neamatallah, M., Esmat, G., Kamel, E., Elalfy, H., Besheer, T., Eldeib, D., Eladl, A.H., El-Setouhy, M., El-Gilany, A.H., and El-Waseef, A. (2016). Associations of human leucocyte antigen class II-DQB1 alleles with hepatitis C virus

infection in Egyptian population: a multicentre family-based study. J. Viral Hepat. *23*, 961–970.

56. Reveille, J.D., Owerbach, D., Goldstein, R., Moreda, R., Isern, R.A., and Arnett, F.C. (1992). Association of polar amino acids at position 26 of the HLA-DQB1 first domain with the anticentromere autoantibody response in systemic sclerosis (scleroderma). J. Clin. Invest. *89*, 1208–1213.

57. Reveille, J.D., Macleod, M.J., Whittington, K., and Arnett, F.C. (1991). Specific amino acid residues in the second hypervariable region of HLA-DQA1 and DQB1 chain genes promote the Ro (SS-A)/La (SS-B) autoantibody responses. J. Immunol. *146*, 3871–3876.

58. Arnett, F.C., Reveille, J.D., Moutsopoulos, H.M., Georgescu, L., and Elkon, K.B. (1996). Ribosomal P autoantibodies in systemic lupus erythematosus. Frequencies in different ethnic groups and clinical and immunogenetic associations. Arthritis Rheum. *39*, 1833–1839.

59. Miki, D., Ochi, H., Takahashi, A., Hayes, C.N., Urabe, Y., Abe, H., Kawaoka, T., Tsuge, M., Hiraga, N., Imamura, M., et al. (2013). HLA-DQB1*03 confers susceptibility to chronic hepatitis C in Japanese: a genome-wide association study. PLoS ONE *8*, e84226.

60. Cramp, M.E., Rossol, S., Carucci, P., Williams, R., Naoumov, N.V., and Donaldson, P.T. (1998). The influence of the HLA class II allele DQB1*0301 on HCV specific T helper cell responses. J. Hepatol. *28*, 94.

61. Kovacs, A.A.Z., Kono, N., Wang, C.H., Wang, D., Frederick, T., Operskalski, E., Tien, P.C., French, A.L., Minkoff, H., Kassaye, S., et al. (2020). Association of HLA Genotype With T-Cell Activation in Human Immunodeficiency Virus (HIV) and HIV/Hepatitis C Virus-Coinfected Women. J. Infect. Dis. *221*, 1156–1166.

62. Thomas, D.L., Thio, C.L., Martin, M.P., Qi, Y., Ge, D., O'Huigin, C., Kidd, J., Kidd, K., Khakoo, S.I., Alexander, G., et al. (2009). Genetic variation in IL28B and spontaneous clearance of hepatitis C virus. Nature *461*, 798–801.

63. Vergara, C., Duggal, P., Thio, C.L., Valencia, A., O'Brien, T.R., Latanich, R., Timp, W., Johnson, E.O., Kral, A.H., Mangia, A., et al. (2020). Multi-ancestry fine mapping of interferon lambda and the outcome of acute hepatitis C virus infection. Genes Immun. *21*, 348–359.

64. Adebamowo, S.N., Adeyemo, A.A.; and ACCME Research Group as part of the H3Africa Consortium (2019). Classical HLA alleles are associated with prevalent and persistent cervical high-risk HPV infection in African women. Hum. Immunol. *80*, 723–730.

65. Shen, J., Mehrotra, D.V., Dorr, M.B., Zeng, Z., Li, J., Xu, X., Nickle, D., Holzinger, E.R., Chhibber, A., Wilcox, M.H., et al. (2020). Genetic Association Reveals Protection against Recurrence of *Clostridium difficile* Infection with Bezlotoxumab Treatment. MSphere *5*, e00232-20. https://doi.org/10.1128/mSphere.00232-20.

66. Degenhardt, F., Wendorff, M., Wittig, M., Ellinghaus, E., Datta, L.W., Schembri, J., Ng, S.C., Rosati, E., Hübenthal, M., Ellinghaus, D., et al. (2019). Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles. Hum. Mol. Genet. *28*, 2078–2092.

67. Gonzalez-Galarza, F.F., McCabe, A., Santos, E.J.M.D., Jones, J., Takeshita, L., Ortega-Rivera, N.D., Cid-Pavon, G.M.D., Ramsbottom, K., Ghattaoraya, G., Alfirevic, A., et al. (2020). Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. Nucleic Acids Res. *48* (D1), D783–D788.

68. Robinson, J., Barker, D.J., Georgiou, X., Cooper, M.A., Flicek, P., and Marsh, S.G.E. (2020). IPD-IMGT/HLA Database. Nucleic Acids Res. *48* (D1), D948–D955.

69. Kawaguchi, S., Higasa, K., Shimizu, M., Yamada, R., and Matsuda, F. (2017). HLA-HD: An accurate HLA typing algorithm for next-generation sequencing data. Hum. Mutat. *38*, 788–797.

Supplemental information

Trans-ancestral fine-mapping of MHC reveals

key amino acids associated with spontaneous

clearance of hepatitis C in HLA-DQβ1

Ana Valencia, Candelaria Vergara, Chloe L. Thio, Nicolas Vince, Venceslas Douillard, Alba Grifoni, Andrea L. Cox, Eric O. Johnson, Alex H. Kral, James J. Goedert, Alessandra Mangia, Valeria Piazzolla, Shruti H. Mehta, Gregory D. Kirk, Arthur Y. Kim, Georg M. Lauer, Raymond T. Chung, Jennifer C. Price, Salim I. Khakoo, Laurent Alric, Matthew E. Cramp, Sharyne M. Donfield, Brian R. Edlin, Michael P. Busch, Graeme Alexander, Hugo R. Rosen, Edward L. Murphy, Genevieve L. Wojcik, Mary Carrington, Pierre-Antoine Gourraud, Alessandro Sette, David L. Thomas, and Priya Duggal

Trans-Ancestral Fine-Mapping of MHC Reveals Key Amino Acids Associated with Spontaneous Clearance of Hepatitis C in HLA-DQβ1

Supplemental Materials and Methods

Tables and Figures
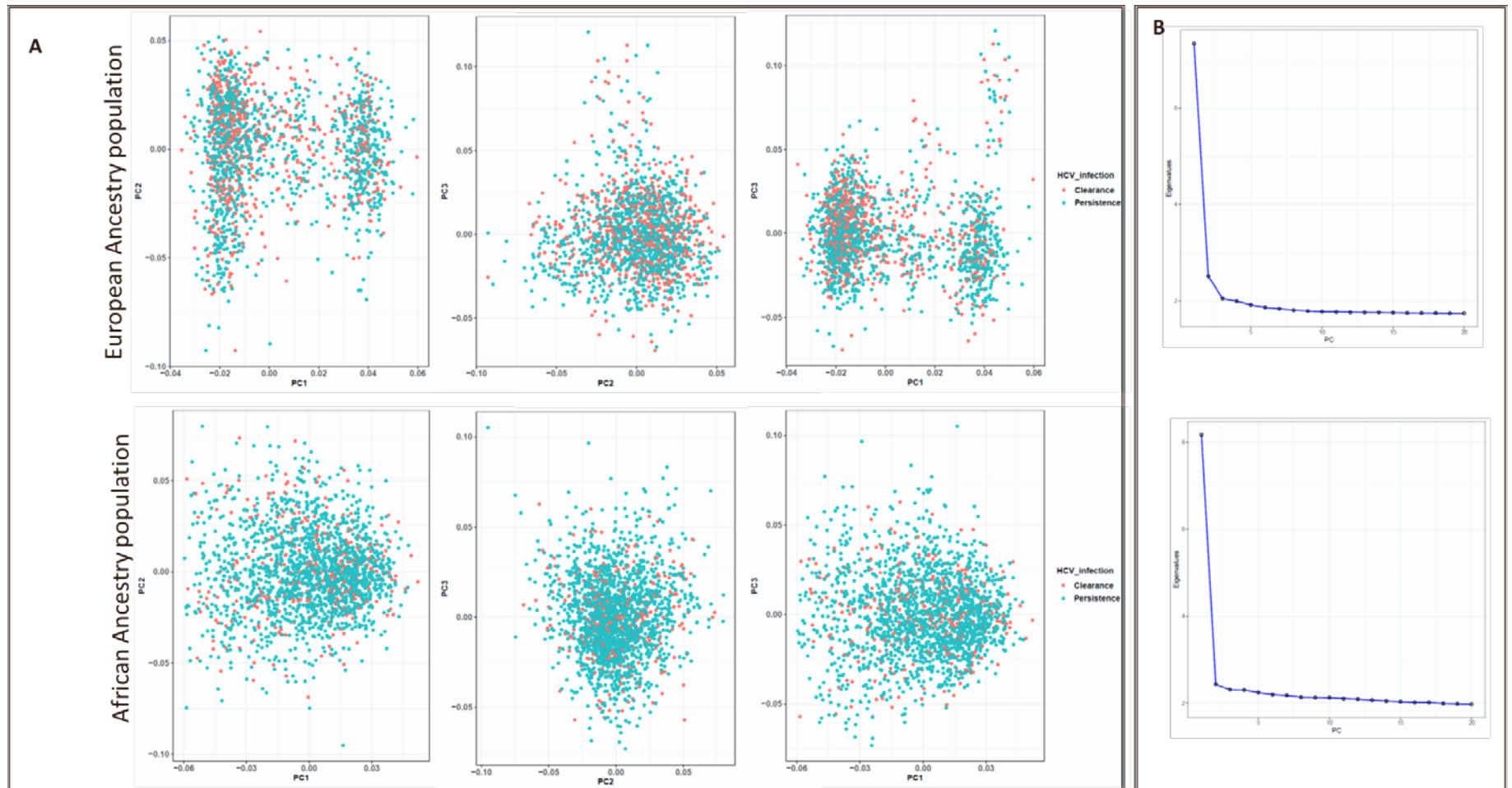
Figure S1: Principal components analysis of the European and African ancestry populations. (A) Scatterplot of PC1, PC2 and PC3. (B) Eigenvalues of each principal component. PC1 and PC2 explains most of the variance in the data for the European and African ancestry populations.

A

Unconditional

rs2647011
p value$_{Meta}$ = 1.13 x 10^{-18}

HLA_Gene
- HLA-DPB1
- HLA-DQA1
- HLA-DQA2
- HLA-DQB1
- HLA-DRA1
- HLA-DRB1
- Other Genes/Intergenic Region

p value$_{Meta}$ = 2.05 x 10^{-7}

rs9274711
p value$_{Meta}$ = 0.58

B

Conditional on rs2647011

HLA_Gene
- HLA-DPB1
- HLA-DQA1
- HLA-DQA2
- HLA-DQB1
- HLA-DRA1
- HLA-DRB1
- Other Genes/Intergenic Region

rs9274711
p value$_{Meta}$ = 4.18 x 10^{-6}

p value$_{Meta}$ conditional = 5 x 10^{-6}

C

Conditional on rs2647011 and rs9274711

HLA_Gene
- HLA-DPB1
- HLA-DQA1
- HLA-DQA2
- HLA-DQB1
- HLA-DRA1
- HLA-DRB1
- Other Genes/Intergenic Region

p value$_{Meta}$ conditional = 5 x 10^{-6}

Figure S2. Results of the meta-analysis of the association of SNPs in the MHC class II region with spontaneous clearance of HCV infection in European and African ancestry in the unconditional analysis and after several steps of conditional analysis on the top SNPs. (A) Association results of the unconditional analysis. (B) Conditional analysis on rs2647011. (C) Conditional analysis on rs2647011 and rs9274711. After conditioning on rs2647011 and rs9274711 no SNPs showed significant association with HCV spontaneous clearance. Each point corresponds to the p value$_{Meta}$ for each SNP. The blue line and the p value$_{Meta}$ colored in blue represent the threshold level of fixed effects meta-analysis significance corrected by multiple comparisons in the unconditional and conditional analyses.

A

HLA-DRB1　　　　　HLA-DQA1　　　　HLA-DQB1　　　　HLA-DPB1

Unconditional

HLA-DQB1*03:01
p value$_{Meta}$ = 2.27 x 10^{-14}

HLA-DRB1*01:01
p value$_{Meta}$ = 3.4 x 10^{-7}

HLA-DQB1*06:02
p value$_{Meta}$ = 9.7 x 10^{-6}

p value$_{Meta}$ = 7.8 x 10^{-4}

HLA-DQB1*02:01
p value$_{Meta}$ = 2.5 x 10^{-4}

B

HLA-DRB1　　　　　HLA-DQA1　　　　HLA-DQB1　　　　HLA-DPB1

Conditional on *HLA-DQB1*03:01*

HLA-DRB1*01:01 , p value$_{Meta}$ = 5.6 X 10^{-9}

p value$_{Meta}$ conditional = 1.0 x 10^{-3}

C

HLA-DRB1　　　　　HLA-DQA1　　　　HLA-DQB1　　　　HLA-DPB1

Conditional on *HLA-DQB1*03:01* and *HLA-DRB1*01:01*

p value$_{Meta}$ conditional = 1.0 x 10^{-3}

(-) Log10 Observed p value$_{Meta}$

HLA classical alleles

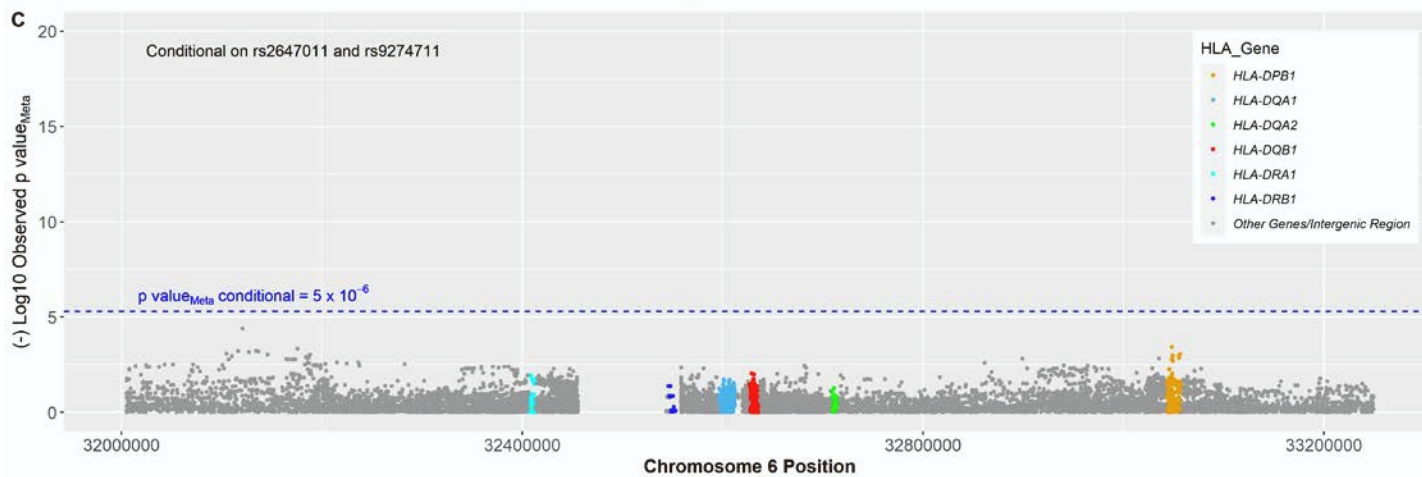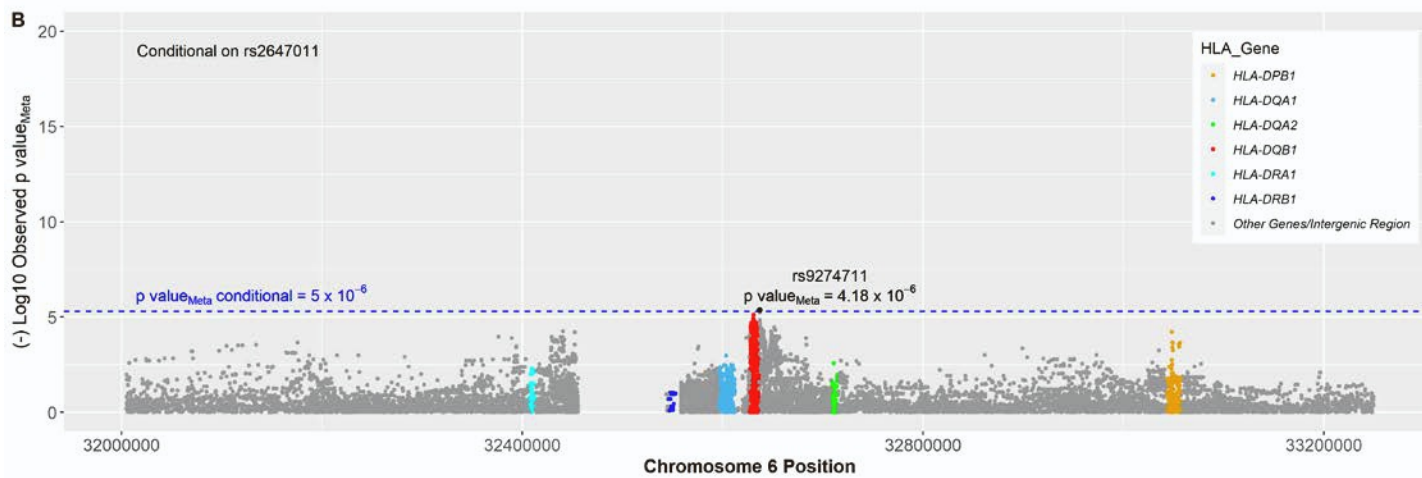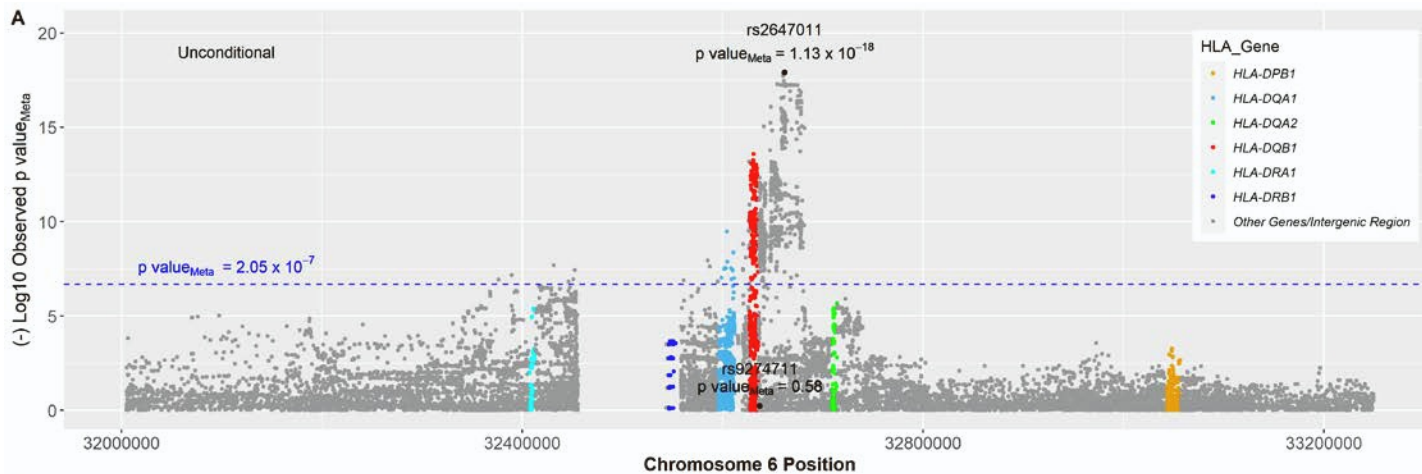**Figure S3**. Results of the meta-analysis of the association of HLA class II alleles with spontaneous clearance of HCV infection in European and African ancestry in the unconditional analysis and after several steps of conditioning on top HLA classical alleles. (A) Association results of the unconditional analysis. (B) Conditional analysis on *HLA-DQB1*03:01*. (C) Conditional analysis on *HLA-DQB1*03:01* and *HLA-DRB1*01:01*. After conditioning on *HLA-DQB1*03:01* and *HLA-DRB1*01:01* no HLA alleles showed significant association with HCV spontaneous clearance. Each point corresponds to the p value$_{Meta}$ for each HLA classical allele. The blue line and the p value$_{Meta}$ colored in blue represent the threshold level of fixed effects meta-analysis significance corrected by multiple comparisons in the unconditional and conditional analyses.
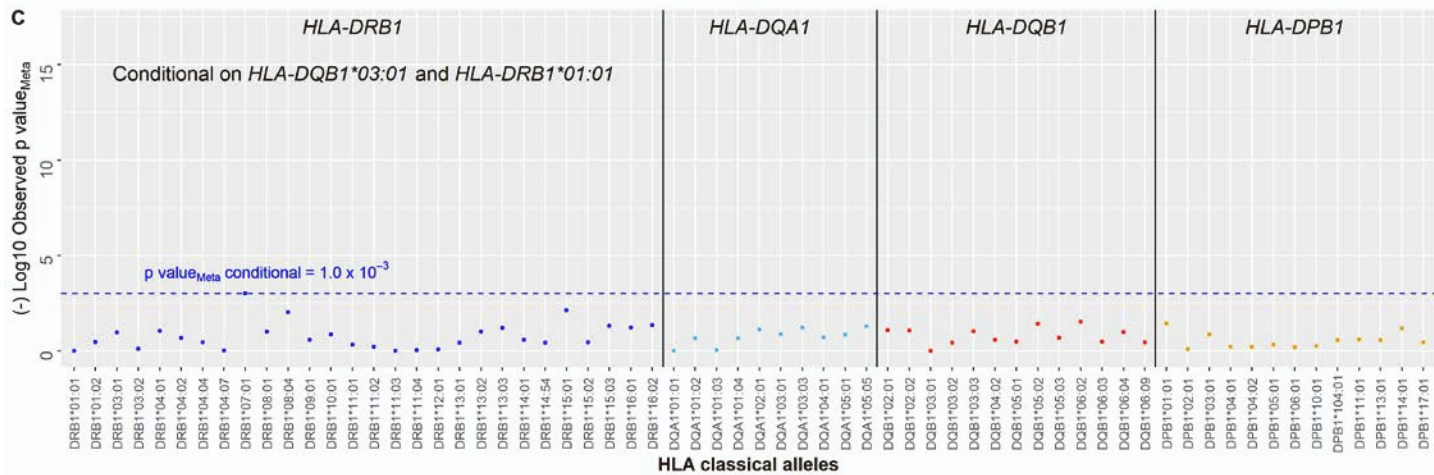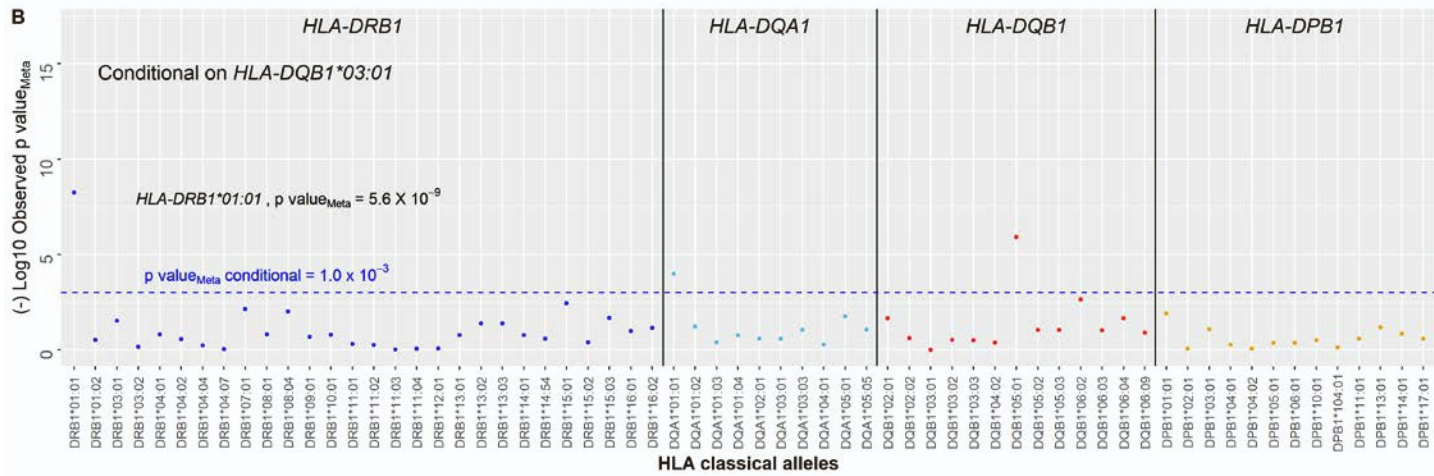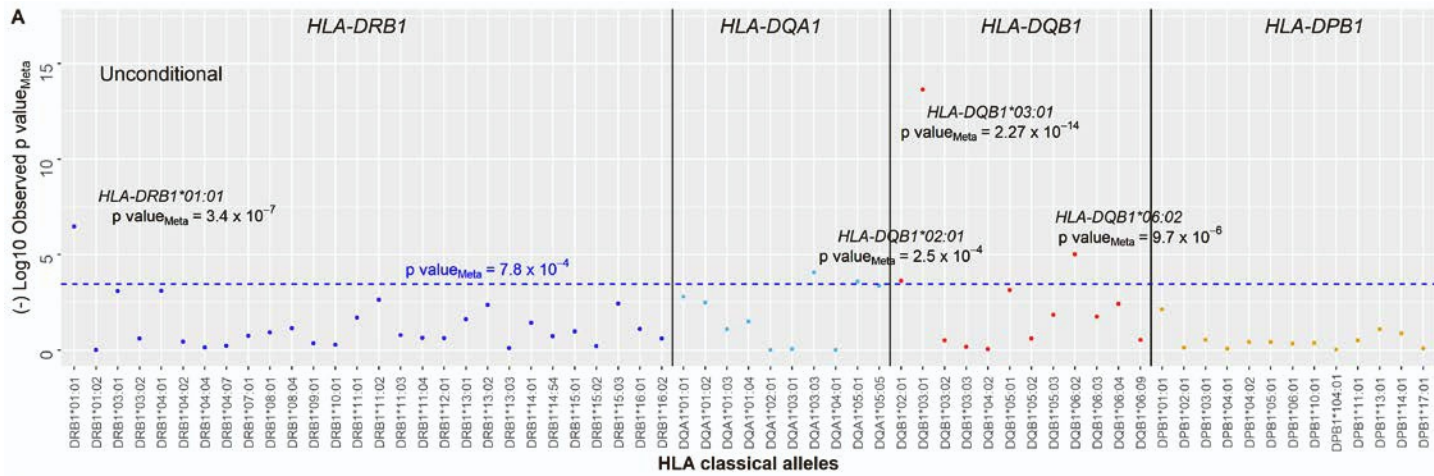
A

HLA-DRB1 | HLA-DQA1 | HLA-DQB1 | HLA-DPB1

Unconditional

HLA-DQ β 1:Leu26, p value$_{Meta}$ = 1.24 x 10^{-14}

HLA-DQ β 1:Tyr26, p value$_{Meta}$ = 1.47 X 10^{-13}

HLA-DQ β 1:Pro55, p value$_{Meta}$ = 8.23 X 10^{-11}

p value$_{Meta}$ = 5.0 x 10^{-4}

(-) Log10 Observed p value$_{Meta}$

HLA Protein's Amino Acids

B

HLA-DRB1 | HLA-DQA1 | HLA-DQB1 | HLA-DPB1

Conditional on HLA-DQ β 1:Leu26

HLA-DQ β 1:Pro55, p value$_{Meta}$ = 2.8 x 10^{-4}

p value$_{Meta}$ conditional = 1.0 x 10^{-3}

(-) Log10 Observed p value$_{Meta}$

HLA Protein's Amino Acids

C

HLA-DRB1 | HLA-DQA1 | HLA-DQB1 | HLA-DPB1

Conditional on HLA-DQ β 1:Leu26 and HLA-DQ β 1:Pro55

p value$_{Meta}$ conditional = 1.0 x 10^{-3}

(-) Log10 Observed p value$_{Meta}$

HLA Protein's Amino Acids

Figure S4. Results of the meta-analysis of the association of amino acid residues in HLA class II proteins with HCV spontaneous clearance in European and African ancestry populations in the unconditional analysis and after several steps of conditioning on top amino acid residues. (A) Association results of the unconditional analysis. (B) Conditional analysis on HLA-DQβ1Leu26. (C) Conditional analysis on HLA-DQβ1Leu26 and HLA-DQβ1Pro55. After conditioning on HLA-DQβ1Leu26 and HLA-DQβ1Pro55 no amino acid residue showed significant association with HCV spontaneous clearance. Each point corresponds to the p value$_{Meta}$ for each amino acid. The blue line and the p value$_{Meta}$ colored in blue represent the threshold level of fixed effects meta-analysis significance corrected by multiple comparisons in the unconditional and conditional analyses.
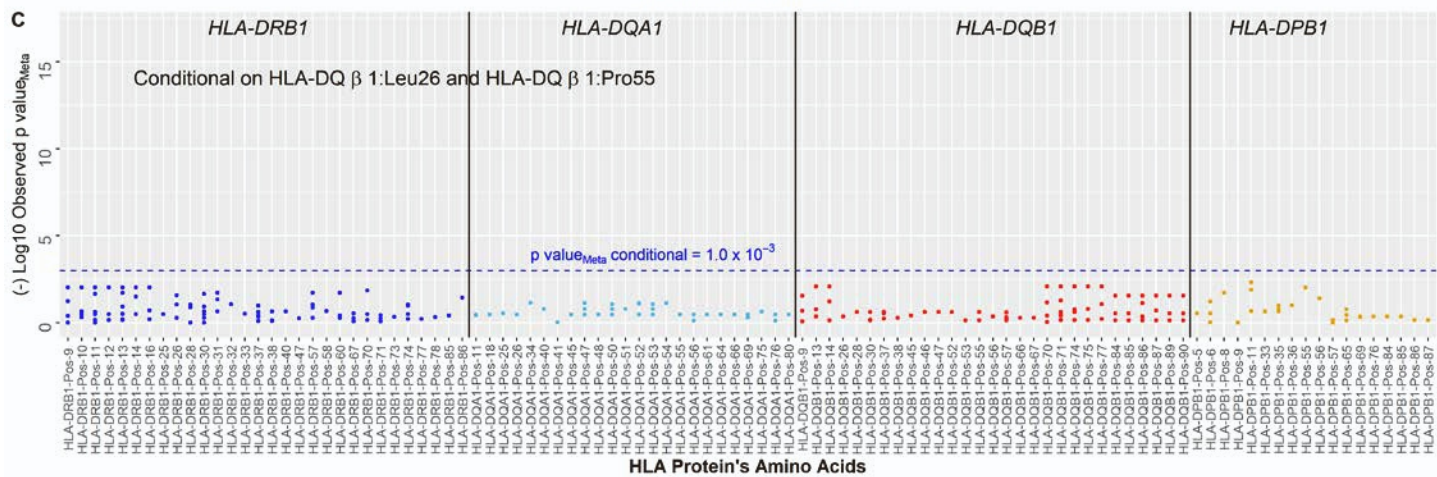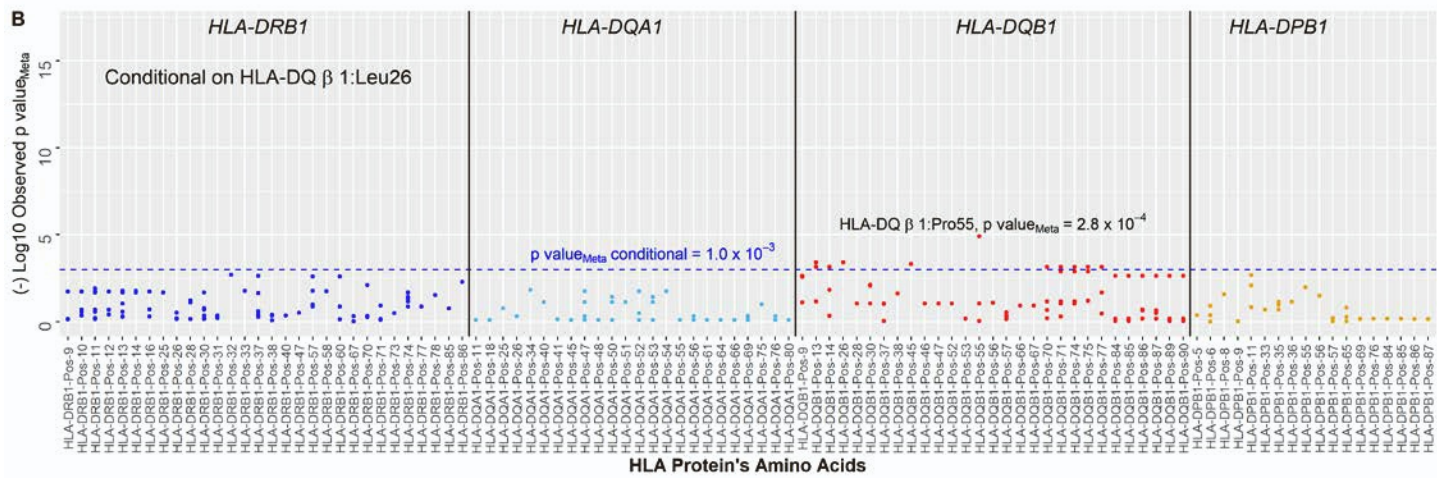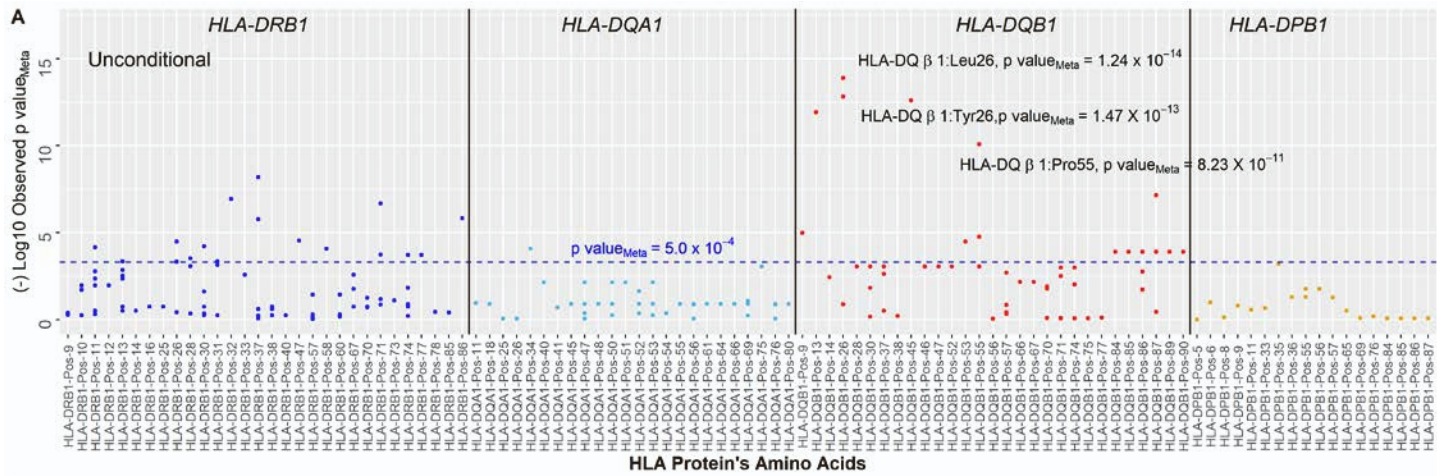
	Imputed HLA class II genes			
European ancestry (N=1,600)	*HLA-DRB1*	*HLA-DQA1*	*HLA-DQB1*	*HLA-DPB1*
No. of individuals with imputed HLA alleles after quality control procedures (% from the total of imputable genotypes)	1,440 (90)	1,582 (98.9)	1,600 (100)	1,543 (96.4)
HCV persistent infection (controls)	912	997	1006	967
HCV spontaneous clearance (cases)	528	585	594	576
Number of unique HLA alleles to test	19	10	12	13
Number of amino acid positions/number of residues or combinations tested	28/63	24/39	29/74	18/29
African ancestry (N=1,864)				
No. of individuals with imputed HLA alleles after quality control procedures (% from the total of imputable genotypes)	1,429 (76.7)	0	1,792 (96.2)	0
HCV persistent infection (controls)	1,179	0	1,467	0
HCV spontaneous clearance (cases)	254	0	330	0
Number of unique HLA alleles to test	20	0	13	0
Number of amino acid positions/number of residues or combinations tested	30/84	0	29/71	0

Table S1. Efficiency and quality of imputation of HLA class II classical alleles and amino acids residues per ancestry group.

Table S2 (Excel file)

Table S2. SNPs significantly associated with HCV spontaneous clearance in the MHC class II region. Abbreviations: Chr: Chromosome; BP: position in base pairs; OR: odds ratio for HCV spontaneous clearance; CI: Confidence Interval; p value$_{Meta}$: p value of the meta-analysis of European and African ancestry populations.

HLA Gene	Effect Allele	European Ancestry (N=1,600)				African Ancestry (N=1,869)				p value$_{Meta}$
		Frequency Effect Allele		OR (95%CI)	p value	Frequency Effect Allele		OR (95%CI)	p value	
		Clearance (n=594)	Persistence (n=1,006)			Clearance (n=340)	Persistence (n=1,529)			
DRB1	DRB1*01:01	0.10	0.06	1.68 (1.3-2.2)	2.0×10^{-4}	0.06	0.03	2.34 (1.5-3.8)	4.7×10^{-4}	3.4×10^{-7}
DQA1	DQA1*03:03	0.08	0.04	1.85 (1.4-2.5)	8.5×10^{-5}	NA	NA	NA	NA	8.5×10^{-5}
	DQA1*05:01	0.09	0.13	0.61 (0.5-0.8)	2.6×10^{-4}	NA	NA	NA	NA	2.6×10^{-4}
	DQA1*05:05	0.19	0.16	1.41(1.2-1.7)	4.4×10^{-4}	NA	NA	NA	NA	4.4×10^{-4}
DQB1	DQB1*02:01	0.09	0.13	0.60 (0.5-0.8)	1.9×10^{-4}	0.06	0.07	0.76 (0.5-1.1)	1.4×10^{-1}	2.5×10^{-4}
	DQB1*03:01	0.27	0.20	1.61 (1.3-1.9)	7.6×10^{-8}	0.27	0.18	1.75 (1.4-2.1)	5.9×10^{-8}	2.3×10^{-14}
	DQB1*05:01	0.13	0.10	1.39 (1.1-1.7)	4.2×10^{-3}	0.13	0.1	1.24 (1.0-1.6)	5.5×10^{-2}	7.3×10^{-4}
	DQB1*06:02	0.10	0.12	0.71 (0.6-0.9)	3.4×10^{-3}	0.15	0.21	0.7 (0.5-0.9)	8.9×10^{-4}	9.7×10^{-6}

Table S3. HLA alleles class II significantly associated with HCV spontaneous clearance in European and African ancestry populations. Abbreviations: OR: odds ratio for HCV spontaneous clearance; CI: Confidence Interval; p value$_{Meta}$: p value of the meta-analysis of European and African ancestry populations; NA: Not available.

Table S4 (excel file)

Table S4. Results of the association analysis and meta-analysis of amino acids residues on HLA class II genes with spontaneous clearance of HCV infection in European and African ancestry groups. Amino acids are named in a single-letter code, those with asterisk (*) are significantly associated with HCV spontaneous clearance. Abbreviations: OR: odds ratio; CI: confidence interval; p value$_{Meta}$: p value of the meta-analysis of both ancestry groups; NA: Not available. Single letter and three letter amino acid code: A:Ala, C:Cys, D:Asp, E:Glu, F:Phe, G:Gly, H:His, I:Ile, K:Lys, L:Leu, M:Met, N:Asn, P:Pro, Q:Gln, R:Arg, T:Thr, S:Ser, Y:Tyr, V:Val, W:Trp.

Marker/Gene (Effect Allele)	HIV infection status	European Ancestry n (HIV -) = 1,342 n (HIV +) = 258		African Ancestry n (HIV -) = 1,142 n (HIV +) = 722		p value$_{Meta}$
		OR	p value	OR	p value	
rs2647011 (A)	HIV (-)	1.23	2.89×10^{-8}	1.26	2.06×10^{-5}	3.28×10^{-12}
	HIV (+)	1.43	1.30×10^{-4}	1.31	2.78×10^{-5}	2.69×10^{-8}
rs9274711 (T)	HIV (-)	0.97	7.39×10^{-1}	0.88	3.74×10^{-1}	3.96×10^{-1}
	HIV (+)	1.08	7.16×10^{-1}	1.02	9.12×10^{-1}	7.79×10^{-1}
*DQB1(DQB1*03:01)*	HIV (-)	1.52	3.03×10^{-5}	1.90	1.95×10^{-6}	3.23×10^{-10}
	HIV (+)	1.96	2.94×10^{-3}	1.61	3.49×10^{-3}	5.36×10^{-5}
*DQB1(DQB1*06:02)*	HIV (-)	0.70	7.18×10^{-3}	0.73	3.76×10^{-2}	7.06×10^{-4}
	HIV (+)	0.54	6.77×10^{-2}	0.60	5.70×10^{-3}	9.28×10^{-4}
*DRB1 (DRB1*01:01)*	HIV (-)	1.65	1.43×10^{-3}	1.77	9.41×10^{-2}	4.42×10^{-4}
	HIV (+)	1.69	1.46×10^{-1}	3.02	1.92×10^{-3}	6.82×10^{-4}
HLA-DQβ1 Leu26	HIV (-)	0.84	1.45×10^{-5}	0.57	5.53×10^{-5}	3.24×10^{-9}
	HIV (+)	0.75	3.27×10^{-3}	0.52	8.10×10^{-5}	9.18×10^{-7}
HLA-DQβ1 Pro55	HIV (-)	1.42	1.86×10^{-4}	1.64	4.50×10^{-4}	3.00×10^{-7}
	HIV (+)	1.77	6.06×10^{-3}	1.57	1.21×10^{-2}	3.15×10^{-4}

Table S5. Unconditional association analysis of independently associated SNPs, MHC classical alleles and amino acids by HIV infection status. Abbreviations: OR: odds ratio; p value$_{Meta}$: p value of the meta-analysis of both ancestry groups.

Table S6 (excel file)

Table S6. Results of the unconditional association analysis and fixed effect meta-analysis of the top associated SNPs, HLA class II classical alleles and amino acid residues and after conditioning with the same type of variant and across different type of variants. Abbreviations: p value$_{Meta}$: p value of the meta-analysis of both ancestry groups. Single letter and three letter amino acid code: A:Ala, C:Cys, D:Asp, E:Glu, F:Phe, G:Gly, H:His, I:Ile, K:Lys, L:Leu, M:Met, N:Asn, P:Pro, Q:Gln, R:Arg, T:Thr, S:Ser, Y:Tyr, V:Val, W:Trp.

Table S7 (excel file)

Table S7. Results of the unconditional association analysis and fixed effect meta-analysis of SNPs, HLA alleles and amino acid residues in the MHC class II and after conditioning with amino acids associated with high significance. Abbreviations: p value$_{Meta}$: p value of the meta-analysis of both ancestry groups. Single letter and three letter amino acid code: A:Ala, C:Cys, D:Asp, E:Glu, F:Phe, G:Gly, H:His, I:Ile, K:Lys, L:Leu, M:Met, N:Asn, P:Pro, Q:Gln, R:Arg, T:Thr, S:Ser, Y:Tyr, V:Val, W:Trp.

Table S8 (excel file)

Table S8. List of peptides spanning the HCV proteome and their relative IC$_{50}$ predicted values for each HLA class II alleles associated with viral clearance or persistence. Peptide sequences are listed in combination with localization, derived protein and predicted IC$_{50}$ value for each of the HLA class II analyzed. Peptides are defined as epitopes if the IC$_{50}$ values was $\leq$ 1000nM in at least one of the HLA class II alleles analyzed. The ratio is calculated only on defined epitopes considering the geometric mean of the IC$_{50}$ values calculated in persistence alleles versus clearance alleles. The higher the ratio, the stronger the binding affinity in clearance HLA alleles versus persistence HLA alleles. Single letter and three letter amino acid code: A:Ala, C:Cys, D:Asp, E:Glu, F:Phe, G:Gly, H:His, I:Ile, K:Lys, L:Leu, M:Met, N:Asn, P:Pro, Q:Gln, R:Arg, T:Thr, S:Ser, Y:Tyr, V:Val, W:Trp.

Gene	All			African Ancestry (N=18)			European Ancestry (N=13)		
	Total alleles (N)	Concordant alleles (N)	Accuracy (%)	Total alleles (N)	Concordant alleles (N)	Accuracy (%)	Total alleles (N)	Concordant alleles (N)	Accuracy (%)
HLA-DRB1	42	41	97.6	26	25	96.2	16	16	100.0
HLA-DQA1	22	22	100.0	NA	NA	NA	22	22	100.0
HLA-DQB1	50	49	98.0	32	31	96.9	18	18	100.0
HLA-DPB1	24	23	98.0	NA	NA	NA	24	23	95.8

Table S9. Results of accuracy of the imputation with HIBAG and HLA-HD in all individuals and in individuals of European and African ancestry.

Materials and Methods

Samples

<u>Study Cohorts</u>

Boston Area HCV Study: Transmission, Immunity, Outcomes Network (BAHSTION). Subjects were enrolled in Massachusetts General Hospital, Brigham and Women's Hospital and Lemuel Shattuck Hospital (USA). HCV spontaneous clearance was determined by 2 negative tests separated by at least six months, and no treatment was provided prior to case status determination.

Cramp et al. These individuals were recruited from patients seen at the South West Liver Unit and at King's College Hospital (London). HCV spontaneous clearance was determined by being both HCV antibody positive and HCV RNA seronegative over a period of at least 18 months with no treatment. HCV persistence was determined by both HCV RNA and HCV antibody positive. The clearance and persistence groups were comparable in terms of age, sex, estimated duration of infection and route of infection[1].

Hemophilia Growth and Development Study (HGDS). The HGDS is a multicenter natural history study of adolescents and children with hemophilia and its complications. HCV spontaneous clearance was defined as HCV antibody positive and HCV RNA non-detectable for two or more study measurements separated by six months. HCV persistence was defined as HCV antibody positive and HCV RNA above the limit of detection for two or more study measurements. HCV treatments given prior to the case determination were not considered[2].

Mangia et al. This cohort corresponds to HCV mono-infected patients with HCV positive antibody test followed up at the Research laboratory of IRCCS "Casa Sollievo della Sofferenza" (Italy). Determination of HCV spontaneous clearance was made by repeated HCV RNA assays

performed every 6 months during an observational period of 18 months with no treatment The original study was approved by the Ethic Committee of the above IRCCS in Italy[3].

AIDS Link to the Intravenous Experience (ALIVE). The ALIVE cohort is based in Baltimore (USA) and includes individuals who inject drugs with follow up visits every six months. HCV spontaneous clearance was determined by has having two serum samples in which HCV RNA was <50 IU/ml and having HCV antibody positive; subjects with HCV antibodies and RNA in serum were considered as persistently infected. No HCV treatments were given prior to the case determination. Approval was granted by the Johns Hopkins Bloomberg School of Public Health[4].

Baltimore Before and After Acute Study of Hepatitis (BBAASH). This study is based in Baltimore (USA) and comprises a prospective cohort of people who inject drugs and at risk of HCV infection[5]. In this study, HCV infection was confirmed by HCV antibody seroconversion for all subjects. HCV spontaneous clearance was defined as the absence of detectable viremia in at least two serum or plasma specimens obtained >300 days after initial viremia and >60 days apart. HCV persistence was defined as continuous viremia <300 days after initial viremia. No subjects were included who underwent therapy prior to determination of HCV outcome. The study was approved by the Institutional Review Board at the Johns Hopkins School of Medicine[5].

Multicenter Hemophilia Cohort Studies (MHCS I and II). The MHCS consists of hemophilic subjects (mostly with hemophilia A) recruited in 16 sites located in the United States, Greece, Germany, and Austria with a high percentage of HCV positive individuals (70%). The Second Multicenter Hemophilia Cohort Study (MHCS-II) includes subjects >13 years of age with an inherited coagulation disorder such as hemophilia A or B, deficiencies in other factors such as V or XI, and von Willebrand who were exposed to HCV[6].

Correlates of Resolved Versus Low-Level Viremic Hepatitis C Infection in Blood Donors study (REVELL). This cohort is a case control study of serologically confirmed HCV positive blood donors identified within a network of 17 blood banks in Western and Southern USA. HCV persistence was defined as being HCV antibody positive and RNA+ both at an index seropositive donation and on 1-2 follow-up samples collected 1-3 years later[7]. HCV spontaneous clearance was determined by negative test by replicate. No HCV therapy was administered while determining the diagnosis. The Committee on Human Research at the University of California, San Francisco, approved the study protocol[7].

The Swan Project. This study recruits persons who inject illicit drugs at a community-based field site on the Lower East Side of Manhattan (USA). The determination of HCV spontaneous clearance/persistence was made with a single test using the discriminatory HCV transcription mediated amplification assay (GenProbe Inc., San Diego, CA). No HCV treatments were given prior to case determination. The study is approved by the institutional review boards of SUNY Downstate College of Medicine and Weill Cornell Medical College. Written informed consent is obtained from all study participants[8].

Toulouse, France cohort. This study includes individuals born in the south of France with HCV spontaneous clearance/persistence determined by antibody and RNA determinations as described extensively in reference[9].

Women's Interagency HIV Study (WIHS). This is a multicenter, collaborative, prospective study for women at risk for HIV infection with several recruiting centers at USA including New York City/Bronx, Brooklyn, NY; Washington DC, Northern California; Los Angeles County/Southern California; Chicago. The study includes HIV+, HIV- women and those previously diagnosed with clinical AIDS or low CD4+ cell counts. HCV spontaneous clearance

was defined as one HCV RNA at entry into WIHS with positive HCV antibodies. No HCV treatments were given prior to case determination[10].

United Kingdom Drug Use cohort. This cohort is based in London at a Hepatology clinic at the Kings College. Subjects identified with HCV spontaneous clearance had two negative PCR reactions at least six months apart and those with HCV persistence had at least two positive PCR reactions. No treatment was given prior to this assessment. Most individuals are self-reported Caucasian, HIV negative, and injection drug users[11].

Urban Health Study. This is a serial, cross-sectional, sero-epidemiological study of injection drug users in the San Francisco Bay Area (USA). HCV spontaneous clearance or persistence was determined based on antibodies and Hepatitis C Viral RNA test was done on those individuals with positive antibody tests. The present analysis included mostly genetically determined African Ancestry individuals[12, 13].

Genotyping, imputation of SNPs, classical HLA alleles and amino acids

The reference population to build the pre-fit classifiers for the European ancestry population used in the software HIBAG included between 1,624 and 2,572 individuals (depending on the HLA locus) from diverse, though mainly European, or European-ancestry individuals from HLARES database obtained from GlaxoSmithKline clinical trials as described in reference 14. SNP from the MHC region in the reference population were typed using the Illumina 1M and 1M Duo platform. 7,976 SNPs with less than 10% missing data were used to build the classifiers. Classical HLA alleles data was generated by Conexio Genomics, HistoGenetics (Ossining, NY, USA) and LabCorp (Burlington, NC, USA) using the SBT, SSO and SSP methodologies for *HLA-A*, *B*, *C*, *DRB1*, *DQA1*, *DQB1* and *DPB1* genes[14].

Classifiers used for the imputation of classical HLA alleles in the African ancestry population were built based on whole genome sequencing data from 880 participating in the "Consortium on Asthma among African-ancestry Populations in the Americas" (CAAPA)[15] with SNP genotyped in the MHC region and typed for classical *HLA* alleles such as *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1 and HLA-DQB1*[16]. As input for the HIBAG algorithm we selected independent SNPs in this reference population determined by pruning out the markers having a linkage disequilibrium (LD) r^2 value > 0.8. After pruning, we obtained 24,504 SNPs with a genotyping rate > 98% and minor allele frequency > 0.001 to use for the construction of the classifiers. On the other hand, to select SNP of our African ancestry group to input in HIBAG, we used genotyped SNPs and imputed SNPs. Allelic dosages of the imputed SNPs were transformed to hard calls using PLINK 2.00 alpha version[17, 18] and imputed markers with uncertainty greater than 0.1 were treated as missing. Out of the 25,565 imputed/genotyped SNPs, 14,620 were in the intersection with the reference panel and were used as input for building the classifiers. For each population we imputed classical HLA alleles and translated to amino acids in cases and controls together.

Statistical analysis

Association analyses and meta-analysis of each type of variants

For the association analysis we encoded each HLA classical allele and amino acid residue as a bi-allelic marker, corresponding to the presence or absence of each allele and for multi allelic amino acid positions we also generated a variable for each possible combination of the residues for that position using HIBAG[14].

Identification of independent signals for each type of variants

In the model of the conditional analyses for each type of variant, we generated dummy variables for each classical HLA allele and each residue in amino acid positions to which we added the top associated variants of each type one by one in an additive logistic regression model including HIV infection status as well as the first two principal components after PCA. We then evaluated the effect of this conditioning on the association of the other variants of the same type. Data formatting of the classical HLA alleles and amino acids was implemented with customized scripts in R[19] and conditional association analyses were performed using PLINK 2.00 alpha version[17, 18].

Epitope prediction of associated classical HLA class II alleles containing causal amino acids

In order to determine pairs of *HLA-DQA1/HLA-DQB1* classical HLA alleles to include in this analysis, we estimated haplotypes of these 2 genes in the European ancestry population using a maximum likelihood estimate of haplotype probabilities using the haplo.stats package (https://www.mayo.edu/research/labs/statistical-genetics-genetic-epidemiology/software) implemented in R[19]. Haplotypes with highest frequencies containing classical HLA alleles with positive and negative effect in HCV spontaneous clearance were included in the analysis. Predicted epitopes associated with clearance or persistence have been selected based on the ratio calculated as following, where *p* are the persistence alleles and *c* are the clearance alleles:

$$Geomean\ Ratio = \frac{\sqrt[pn]{Allele\ p1 * Allele\ p2 * Allele\ pn}}{\sqrt[cn]{Allele\ c1 * Allele\ c2 * Allele\ cn}}$$

Results

Unconditional and conditional association analyses and meta-analyses

<u>*Unconditional and conditional association analysis with amino acids residues*</u>.

Besides HLA-DQβ1Leu26 and HLA-DQβ1Tyr26, other amino acids were also associated with HCV spontaneous clearance with similar significance, including HLA-DQβ1Glu45 (p value$_{Meta}$= 2.4 x 10^{-13}), HLA-DQβ1Ala13 (p value$_{Meta}$= 1.18 x 10^{-12}). HLA-DQβ1Leu53, HLA-DQβ1Pro55 and HLA-DQβ1Phe87 were also associated with less strength of association (Table S4). Their association was notably diminished in the conditional analysis with HLA-DQβ1Leu26 excepting HLA-DQβ1Pro55 and they were not significantly associated after double conditional analysis with HLA-DQβ1Leu26 and HLA-DQβ1Pro55 (Figure S4).

Conditional analysis on associated residues in amino acid positions 13, 45, 53, 55 and 87 did not have a major effect on the association of HLA-DQβ1Leu26. Only HLA-DQβ1Phe9 was able to eliminate the significance of the association of HLA-DQβ1Leu26 (Table S7).

Amino acid residues explain SNP and HLA classical allele associations with HCV clearance

The individual conditional models for other amino acid residues associated with high significance are presented in Table S7. The change in the association of rs2647011 and rs9274711 observed when conditioning with other amino acids was not as significant as it was in the conditional model with HLA-DQβ1Leu26 (Tables S6 and S7). The association of *HLA-DQB1*03:01* was eliminated in the conditional analysis of HLA-DQβ1Ala13 and HLA-DQβ1Glu45 and decreased with HLA-DQβ1 Tyr9 and HLA-DQβ1Pro55 (Table S7).

The association with the classical allele *HLA-DQB1*06:02* was abrogated when conditioning in each of the associated amino acid positions: HLA-DQβ1Tyr9, HLA-DQβ1Ala13, HLA-DQβ1Glu45, HLA-DQβ1Pro55 and HLA-DQβ1Phe87. The association of *HLA-DRB1*01:01* was only reduced with conditional analysis with amino acid residues in HLA-DRβ1 but was widely reduced when conditioning in HLA-DQβ1Leu26 (Table S7). The association of

HLA-DQB1*02:01 was reduced in the conditional analysis with amino acid residues in positions 13, 26, 45, 55 of HLA-DQβ1.

Epitope prediction of classical HLA class II alleles associated with HCV clearance.

The list of peptide sequences used for the prediction analyses with the IC_{50} prediction values for each HLA class II allele analyzed as well as additional information related to the location, specific protein and ratio between persistence/clearance alleles calculated on peptides predicted to be epitopes in at least one of the alleles studied is presented in Table S8.

References

1. Kim, A.Y., Kuntzen, T., Timm, J., Nolan, B.E., Baca, M.A., Reyor, L.L., Berical, A.C., Feller, A.J., Johnson, K.L., Schulze zur Wiesch, J. et al. (2011). Spontaneous control of HCV is associated with expression of HLA-B 57 and preservation of targeted epitopes. Gastroenterology 140, 686-696.e1.

2. Hilgartner, M.W., Donfield, S.M., Willoughby, A., Contant, C.F.,Jr, Evatt, B.L., Gomperts, E.D., Hoots, W.K., Jason, J., Loveland, K.A., McKinlay, S.M. (1993). Hemophilia growth and development study. Design, methods, and entry data. Am. J. Pediatr. Hematol. Oncol. 15, 208-218.

3. Mangia, A., Gentile, R., Cascavilla, I., Margaglione, M., Villani, M.R., Stella, F., Modola, G., Agostiano, V., Gaudiano, C., Andriulli, A. (1999). HLA class II favors clearance of HCV infection and progression of the chronic liver damage. J. Hepatol. 30, 984-989.

4. Vlahov, D., Munoz, A., Anthony, J.C., Cohn, S., Celentano, D.D., Nelson, K.E. (1990). Association of drug injection patterns with antibody to human immunodeficiency virus type 1 among intravenous drug users in Baltimore, Maryland. Am. J. Epidemiol. 132, 847-856.

5. Cox, A.L., Netski, D.M., Mosbruger, T., Sherman, S.G., Strathdee, S., Ompad, D., Vlahov, D., Chien, D., Shyamala, V., Ray, S.C. et al. (2005). Prospective evaluation of community-acquired acute-phase hepatitis C virus infection. Clin. Infect. Dis. 40, 951-958.

6. Goedert, J.J., Chen, B.E., Preiss, L., Aledort, L.M., Rosenberg, P.S. (2007). Reconstruction of the hepatitis C virus epidemic in the US hemophilia population, 1940-1990. Am. J. Epidemiol. 165, 1443-1453.

7. Tobler, L.H., Bahrami, S.H., Kaidarova, Z., Pitina, L., Winkelman, V.K., Vanderpool, S.K., Guiltinan, A.M., Cooper, S., Busch, M.P., Murphy, E.L. (2010). A case-control study of factors

associated with resolution of hepatitis C viremia in former blood donors (CME). Transfusion 50, 1513-1523.

8. Edlin, B.R., Shu, M.A., Winkelstein, E., Des Jarlais, D.C., Busch, M.P., Rehermann, B., O'Brien, T.R., Talal, A.H., Tobler, L.H., Zeremski, M. et al. (2009). More rare birds, and the occasional swan. Gastroenterology 136, 2412-2414.

9. Alric, L., Fort, M., Izopet, J., Vinel, J.P., Charlet, J.P., Selves, J., Puel, J., Pascal, J.P., Duffaut, M., Abbal, M. (1997). Genes of the major histocompatibility complex class II influence the outcome of hepatitis C virus infection. Gastroenterology 113, 1675-1681.

10. Kuniholm, M.H., Gao, X., Xue, X., Kovacs, A., Marti, D., Thio, C.L., Peters, M.G., Greenblatt, R.M., Goedert, J.J., Cohen, M.H. et al. (2011). The relation of HLA genotype to hepatitis C viral load and markers of liver fibrosis in HIV-infected and HIV-uninfected women. J. Infect. Dis. 203, 1807-1814.

11. Khakoo, S.I., Thio, C.L., Martin, M.P., Brooks, C.R., Gao, X., Astemborski, J., Cheng, J., Goedert, J.J., Vlahov, D., Hilgartner, M. et al. (2004). HLA and NK cell inhibitory receptor genes in resolving hepatitis C virus infection. Science 305, 872-874.

12. Kral, A.H., Bluthenthal, R.N., Lorvick, J., Gee, L., Bacchetti, P., Edlin, B.R. (2001). Sexual transmission of HIV-1 among injection drug users in San Francisco, USA: risk-factor analysis. Lancet 357, 1397-1401.

13. Kral, A.H., Lorvick, J., Gee, L., Bacchetti, P., Rawal, B., Busch, M., Edlin, B.R. (2003). Trends in human immunodeficiency virus seroincidence among street-recruited injection drug users in San Francisco, 1987-1998. Am. J. Epidemiol. 157, 915-922.

14. Zheng, X., Shen, J., Cox, C., Wakefield, J.C., Ehm, M.G., Nelson, M.R., Weir, B.S. (2014). HIBAG--HLA genotype imputation with attribute bagging. Pharmacogenomics J. 14, 192-200.

15. Daya, M., Rafaels, N., Brunetti, T.M., Chavan, S., Levin, A.M., Shetty, A., Gignoux, C.R., Boorgula, M.P., Wojcik, G., Campbell, M. et al. (2019). Association study in African-admixed populations across the Americas recapitulates asthma risk loci in non-African populations. Nat. Commun. 10, 880-7.

16. Vince, N., Limou, S., Daya, M., Morii, W., Rafaels, N., Geffard, E., Douillard, V., Walencik, A., Boorgula, M.P., Chavan, S. et al. (2020). Association of HLA-DRB1 *09:01 with tIgE levels among African-ancestry individuals with asthma. J. Allergy Clin. Immunol. 146, 147-155.

17. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7-8. eCollection 2015.

18. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559-575.

19. R Core Team. (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.