Supplementary Information

A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response

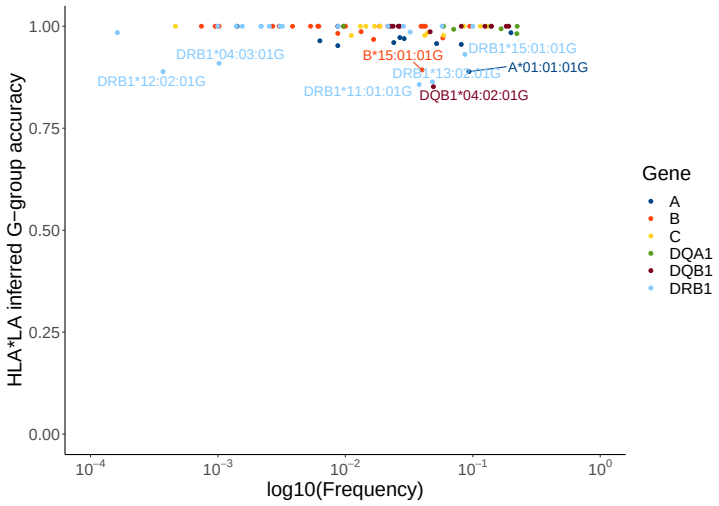
Luo et al. 2021

Contents

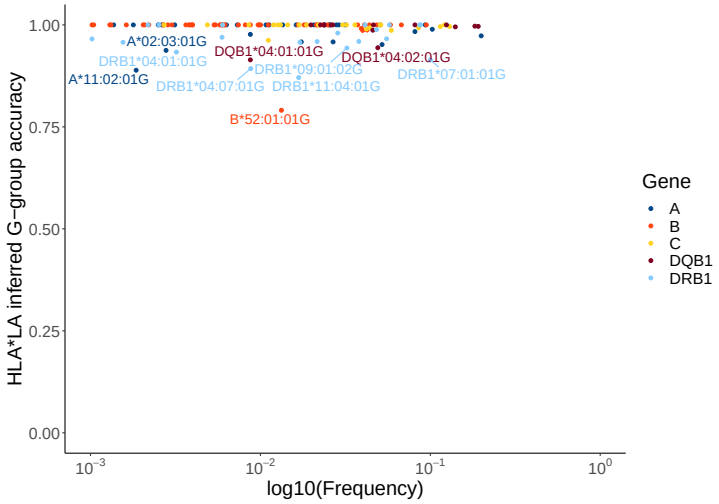
1. Supplementary Figures

Supplementary Figure 1: HLA*LA typing accuracy of all observed G-group alleles against gold-standard sequencing based typing. The x-axis is the log10(allele frequency) observed in all 21,546 individuals. The y-axis is the accuracy of HLA*LA inferred G-group alleles compared against the sequencing based typing. (a) The 288 East Asian individuals from the JPN cohort. (b) The 955 multi-ancestry individuals from the 1000 Genomes Project.

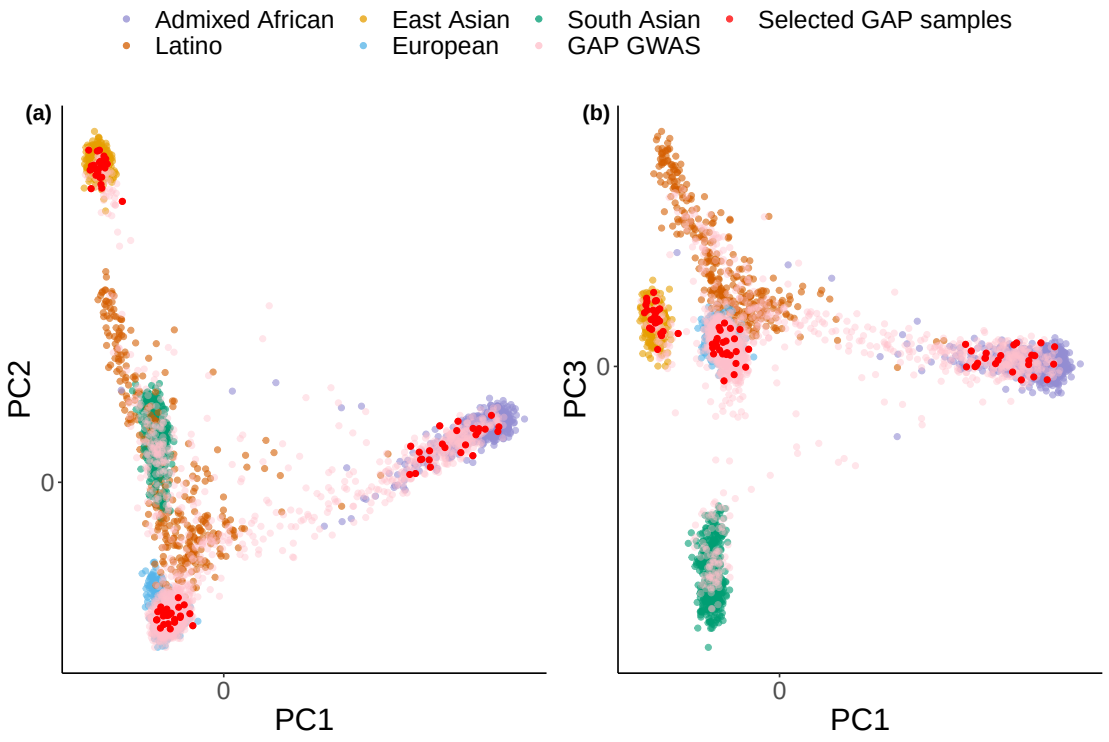
(a) JPN cohort



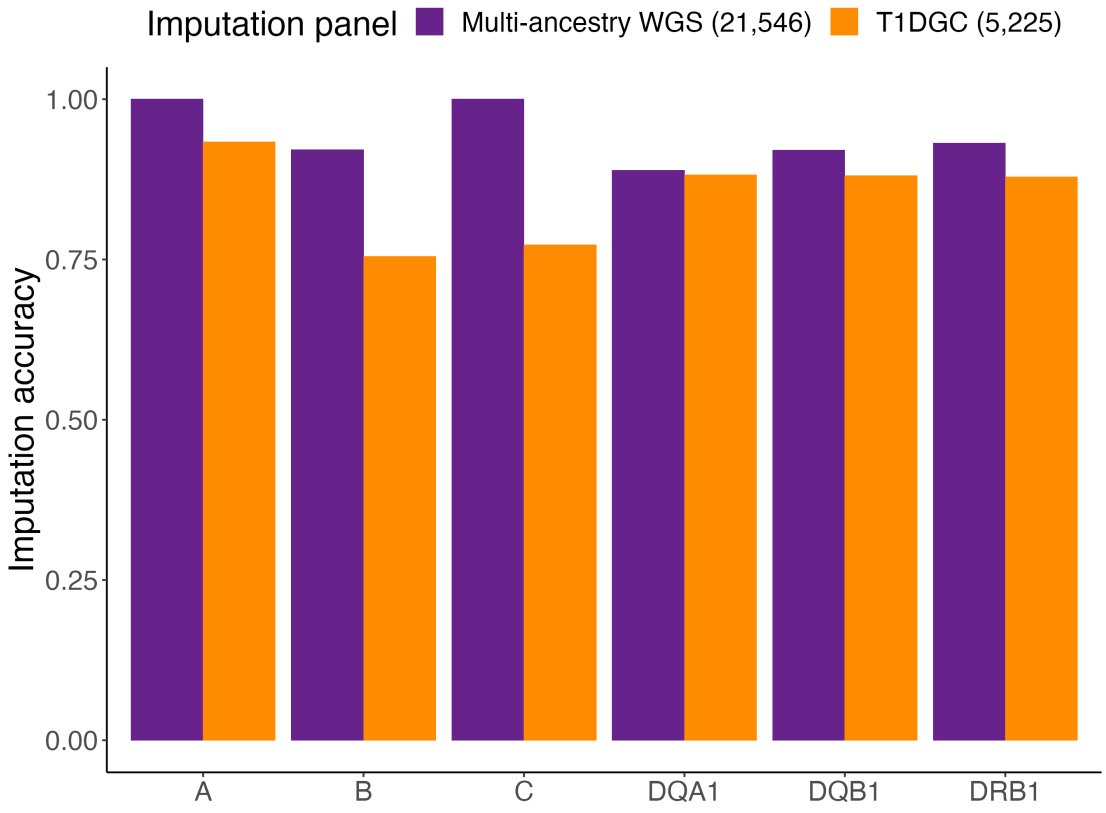
(b) 1000 Genomes Project



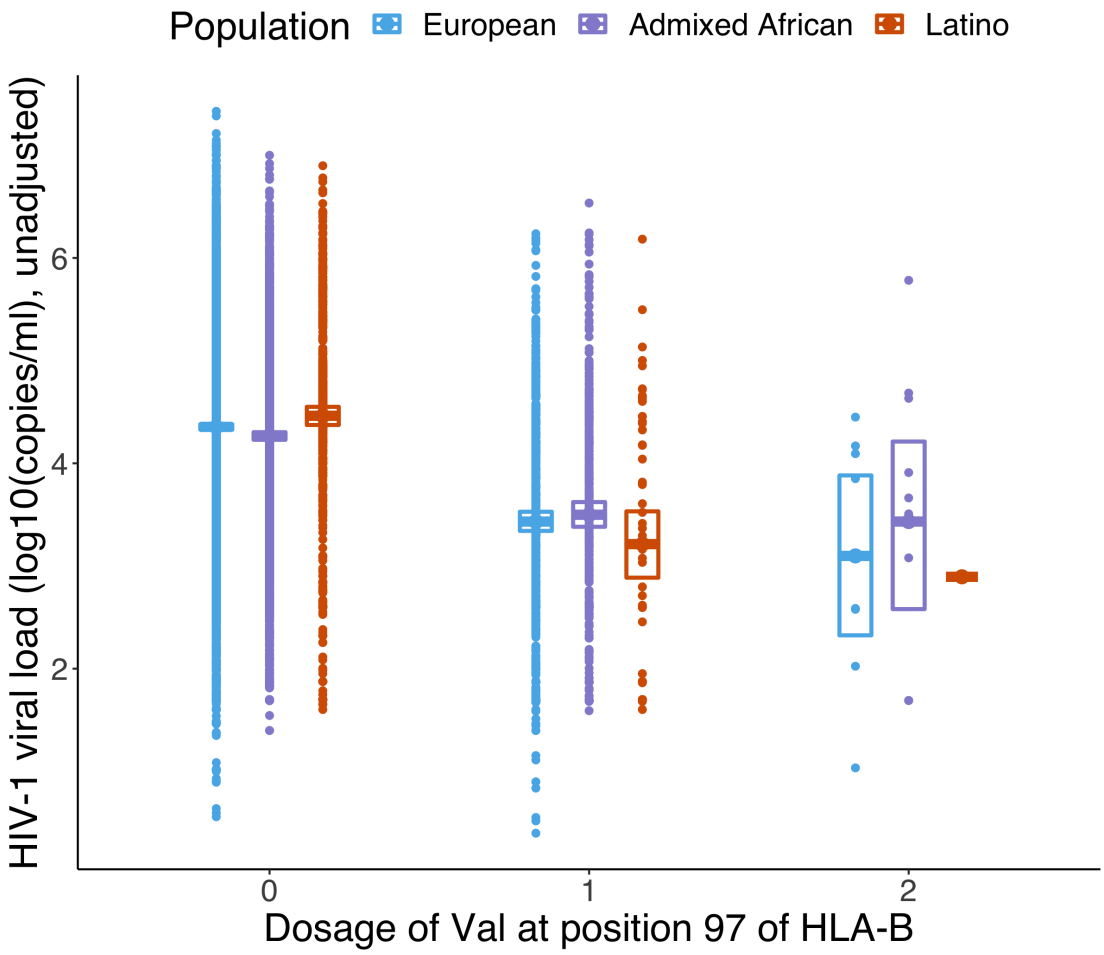
Supplementary Figure 2: Principal component analysis (PCA) of the GAP registry GWAS samples. GAP samples are plotted with five 1000 Genomes Phase 3 populations. (a) First and second principal components. (b) First and third principal components. We randomly selected 25 individuals with African ancestry (purple); 25 with East Asian ancestry (yellow) and 25 with European ancestry (blue) for undergoing gold-standard *HLA* typing (red).



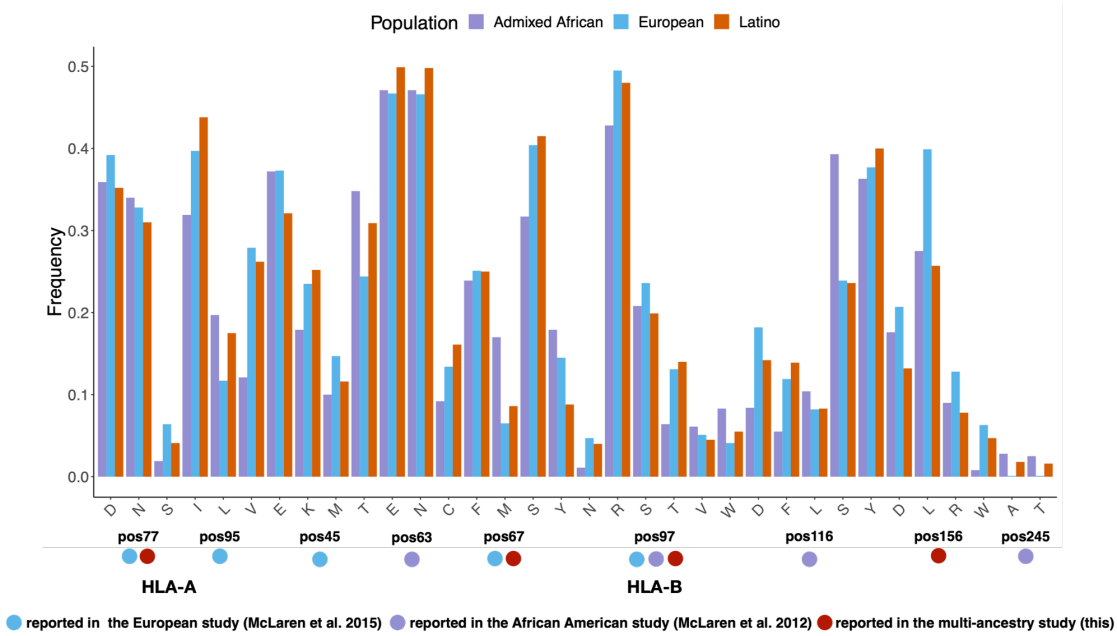
Supplementary Figure 3: Imputation accuracy of GaP registry GWAS data. Imputation accuracy measured as the genotype concordance for G-group classical *HLA* alleles measured in the 75 individuals included in the GAP registry. Accuracy is compared between the whole multi-ancestry HLA reference (dark purple, sample size = 21,546), The Type 1 Diabetes Genetics Consortium5 (T1DGC) panel that consists of European individuals only (orange, sample size = 5,225).



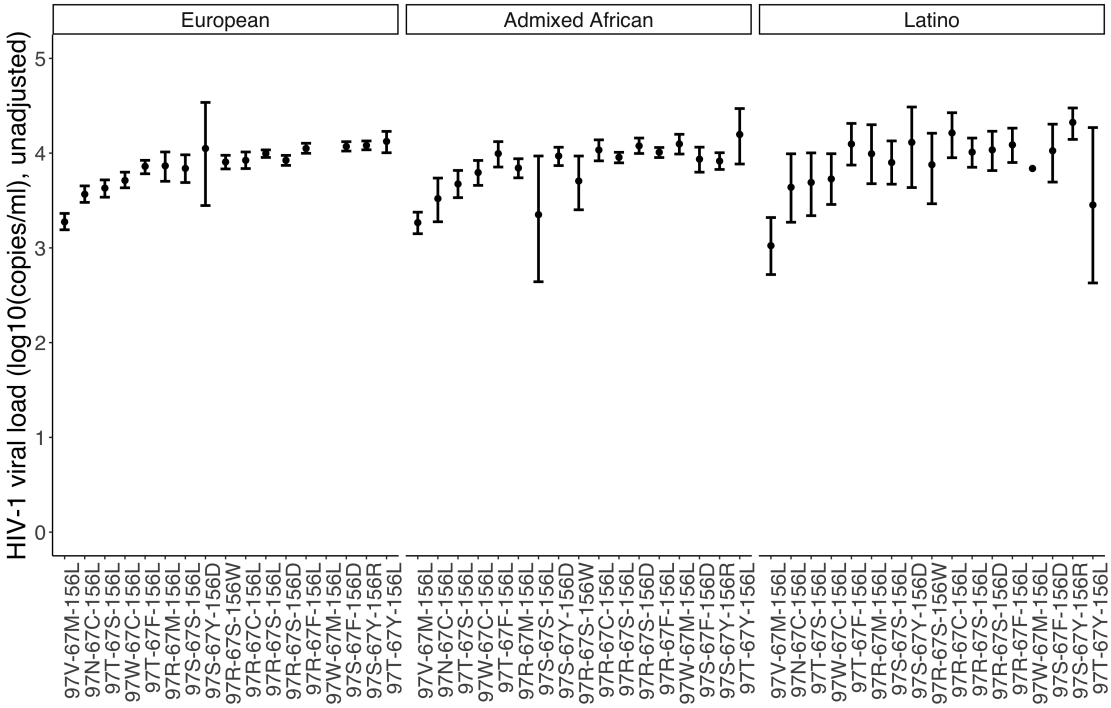
Supplementary Figure 4: Effect of the most protective residue 97Val in HLA-B across diverse populations. Change in log10 HIV-1 set point viral load (spVL, RNA copies per milliliter) of individual amino acid residues at position 97 in HLA-B. Boxplots show the 25th, mean, and 75th percentile of the distribution, with whiskers extending to the largest (and smallest). Residues range from strongly protective (i.e., viral load decreasing) to risk (viral load increasing) in the overall population.



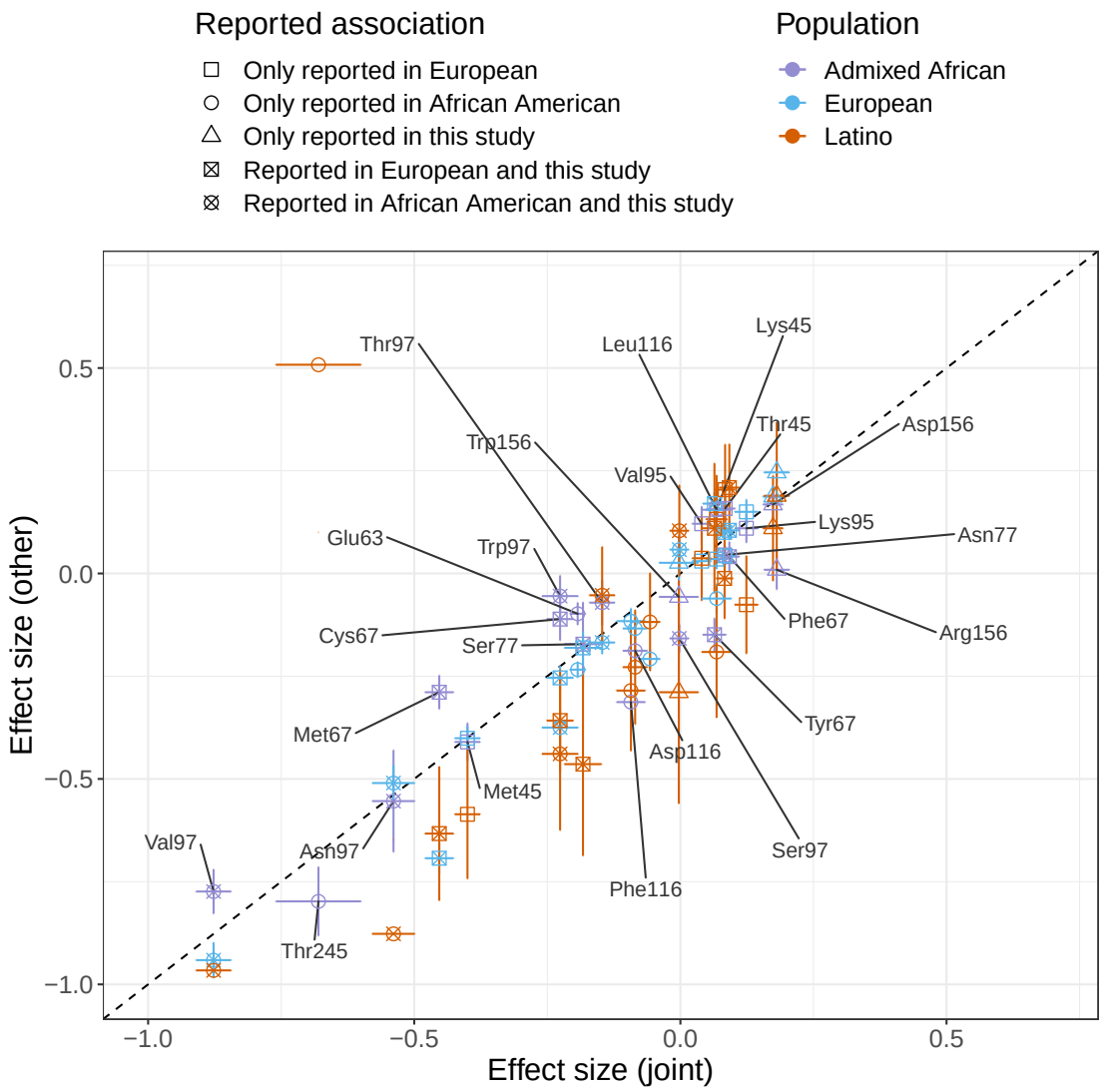
Supplementary Figure 5: Frequencies in all previously and current reported amino acid positions that are associated with HIV-1 viral load.



Supplementary Figure 6: Effect on set point viral load of individual haplotypes formed by amino acid positions 97, 67 and 156 in HLA-B. Mean viral load and its standard error of the haplotypes with frequency > 1% formed by the three independently associated amino acid positions (97, 67 and 156 in HLA-B) in three populations independently. Data are presented as mean values ± standard errors. There are 3,901 Admixed African, 7,455 European and 677 Latino independent samples included in the analysis.

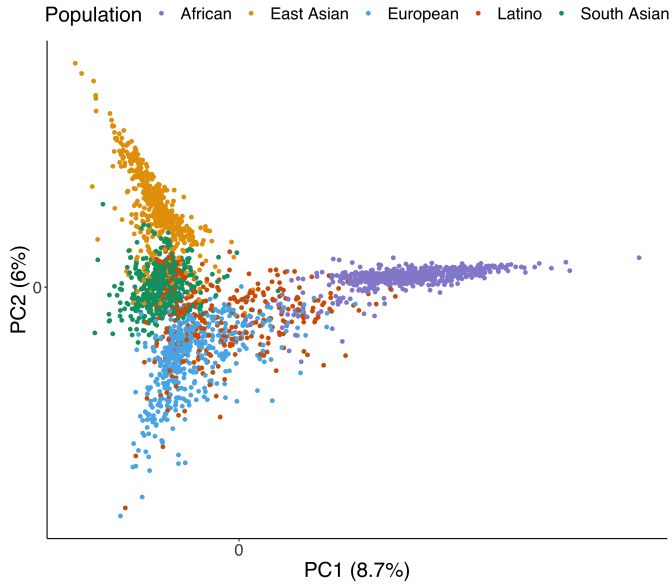


Supplementary Figure 7: Effect on set point viral load of individual amino acid residues at each position reported in this and previous work [1, 2]. Results were calculated per allele using linear regression models. The x-axis shows the effect size and its standard errors in the joint analysis, and the y-axis shows the effect size (mean) ± standard error in individual populations (purple = Admixed American (N=3,901 independent samples), ; blue = European (N=7,455 independent samples) and orange = Latino (N=677 independent samples)). Different shapes indicate which study an amino acid was reported in.

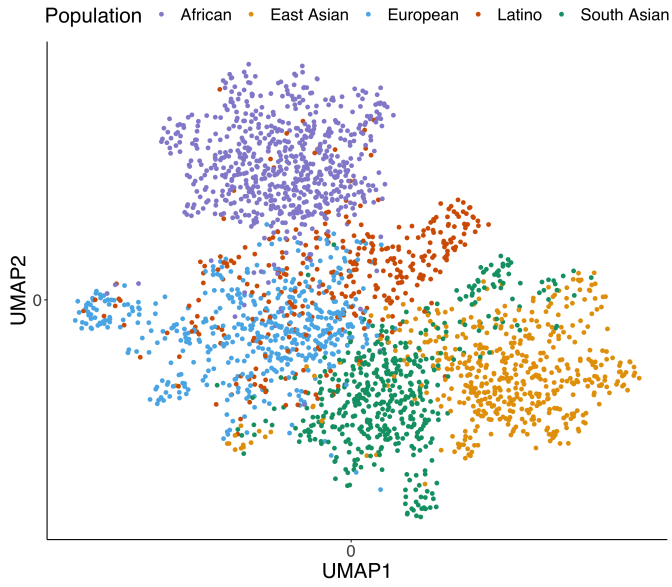


Supplementary Figure 8: Global diversity of the MHC region of the samples included in the 1000 Genomes Project. (a) Dimension reduction of the pairwise IBD distance between 2,504 samples included in the 1000 Genomes Project using MHC region markers only with population labels. The first two principal components show separation of continental groups (European, African, and East Asian) as well as the admixed nature of the Latino and South Asian samples. (b) UMAP on the same pairwise IBD distance forms distinct clusters for all five populations.

(a) Principal Component

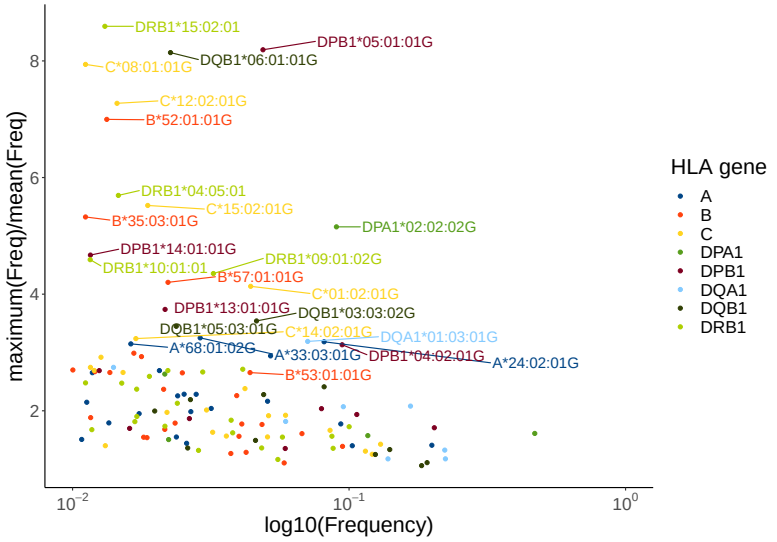


(b) UMAP

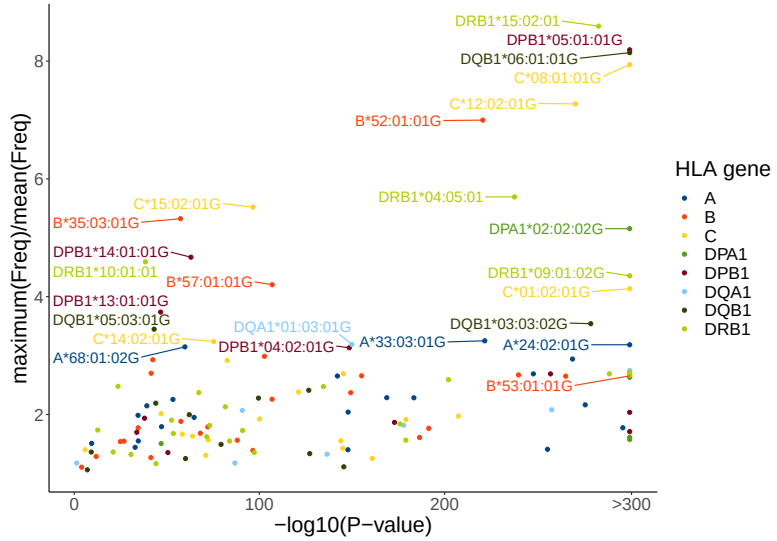


Supplementary Figure 9: Heterogeneity of observed common G-group alleles (frequency > 1%). The y-axis reports the ratio of the maximum frequency of a given allele among five continental populations to its overall frequency. The x-axis shows (a) the log10 frequency of all G-group alleles with frequencies >1% observed in the overall population; and (b) the -log10(P-value) of chi-square statistics (two-sided) obtained from a 3×5 genotype contingency table.

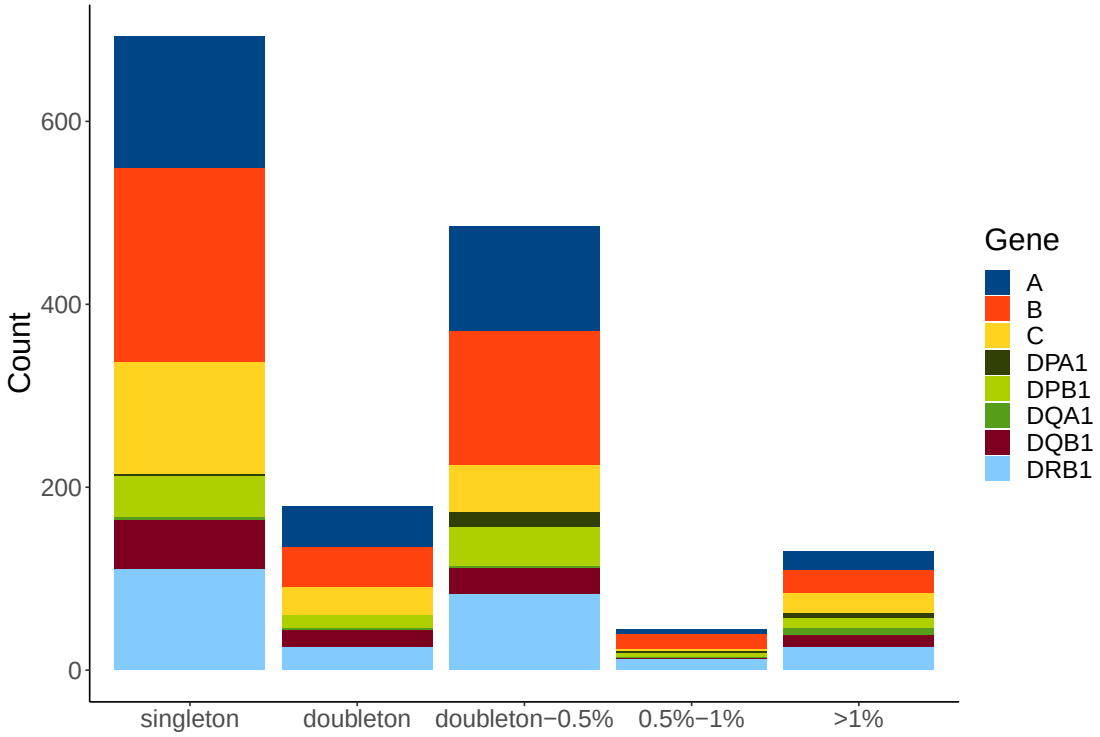
(a) Overall G-group allele frequency



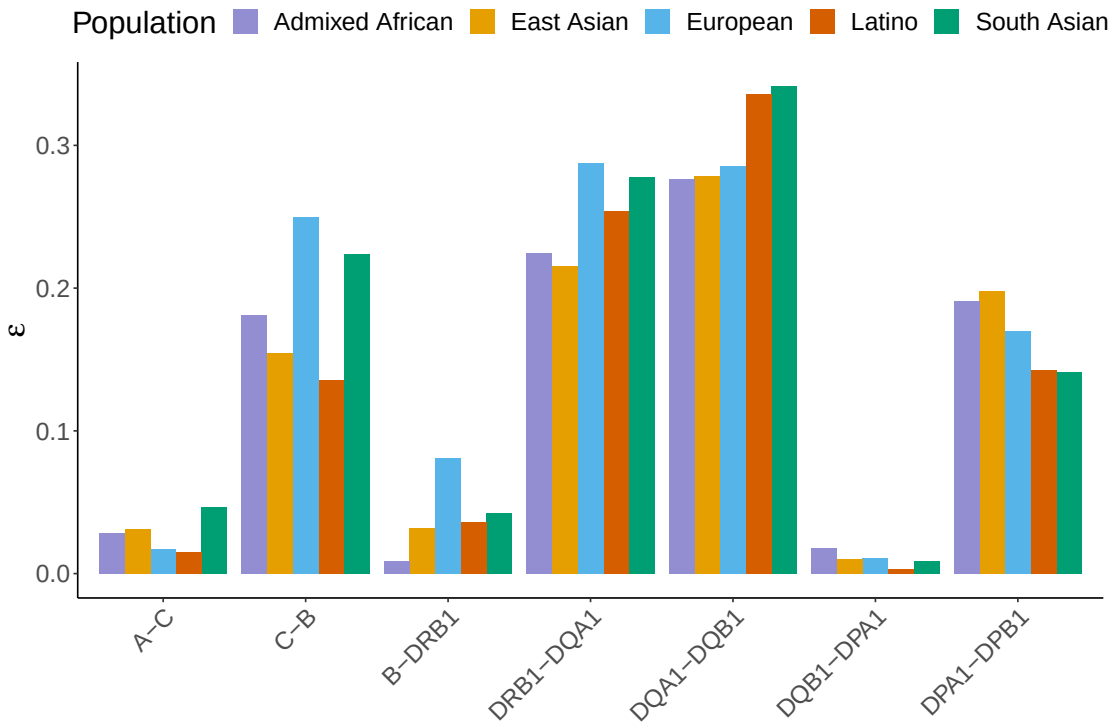
(b) Heterogeneity P-value among five continental populations



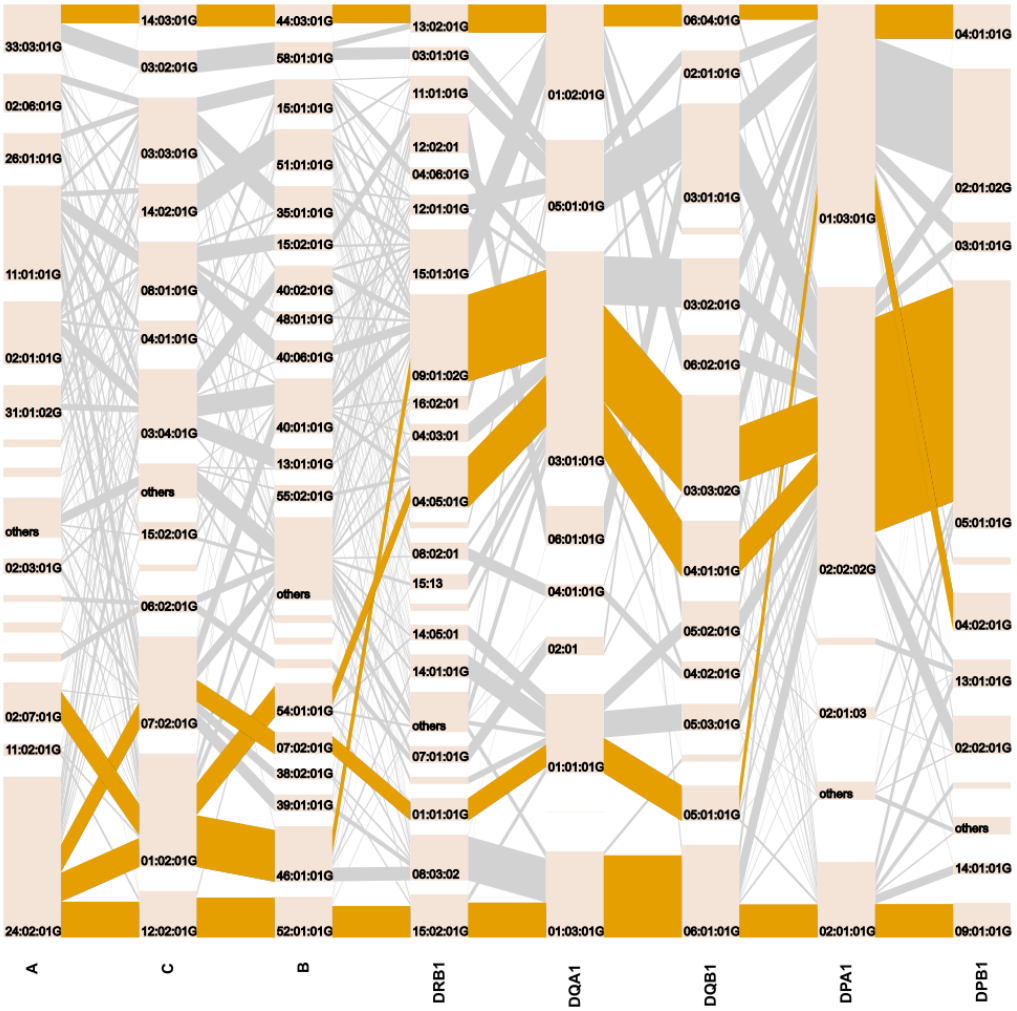
Supplementary Figure 10: Number of reported G-group alleles stratified by overall frequency among 21,546 individuals.



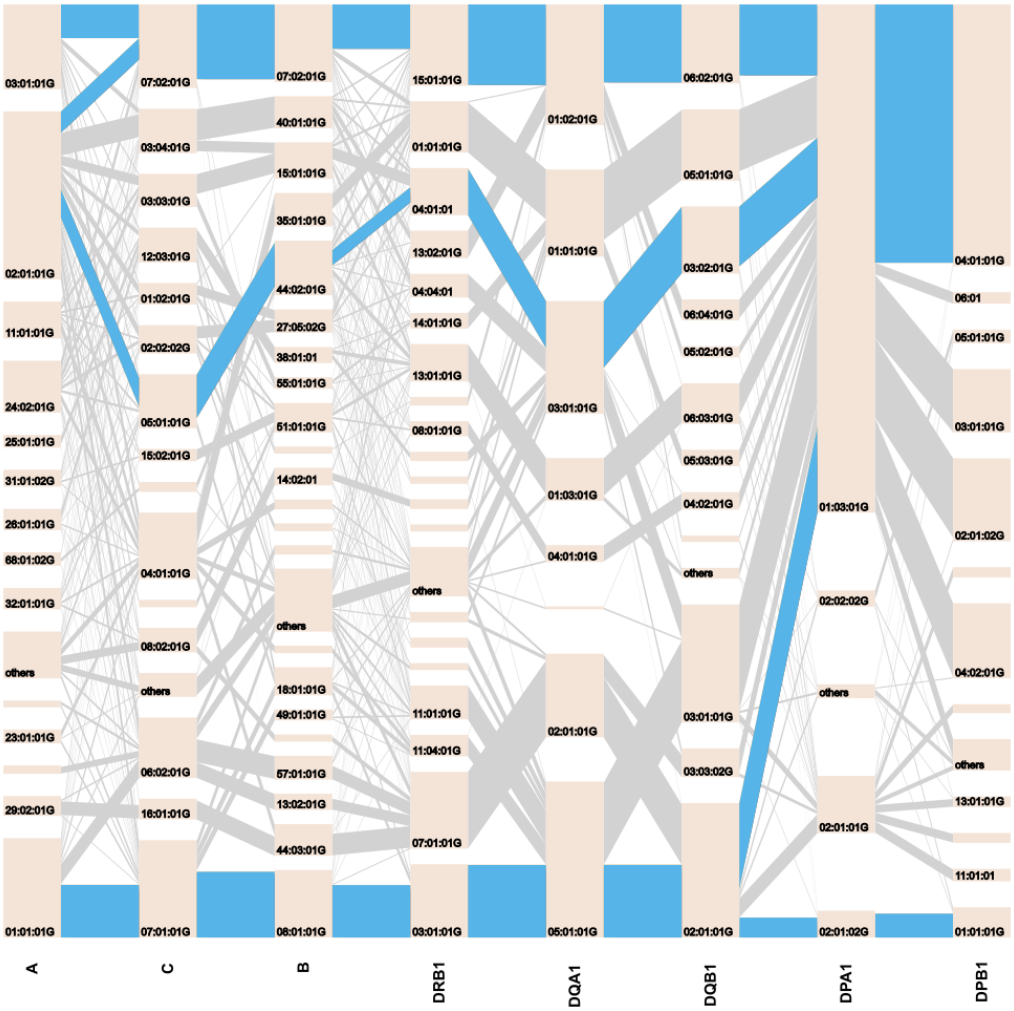
Supplementary Figure 11: Pairwise LD measurement index, ϵ, among five ancestral populations. The value ϵ evaluates the LD structures between two multiallelic variants and ranges between no LD (0) and perfect LD (1). Higher were observed between *DQA1*, *DQB1*, and *DRB1*; between *DPA1* and *DPB1*; and between *B* and *C*.



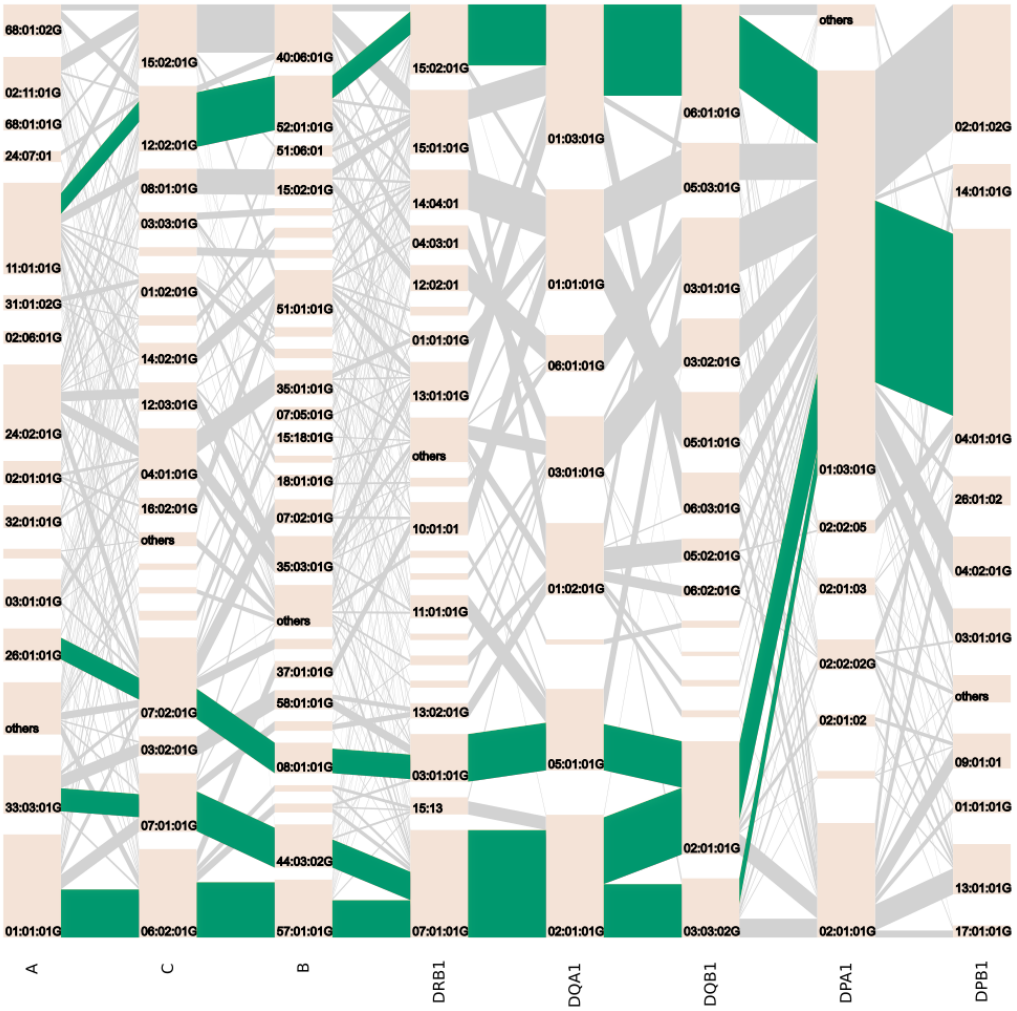
Supplementary Figure 12: Haplotype structure for eight classical HLA genes in East Asians. The haplotype structures of the eight classical HLA genes. The tile in a bar represents an *HLA* allele, and its height corresponds to the frequencies of the *HLA* allele. All long-range *HLA* haplotypes with frequency >1% are in bold and highlighted in yellow.



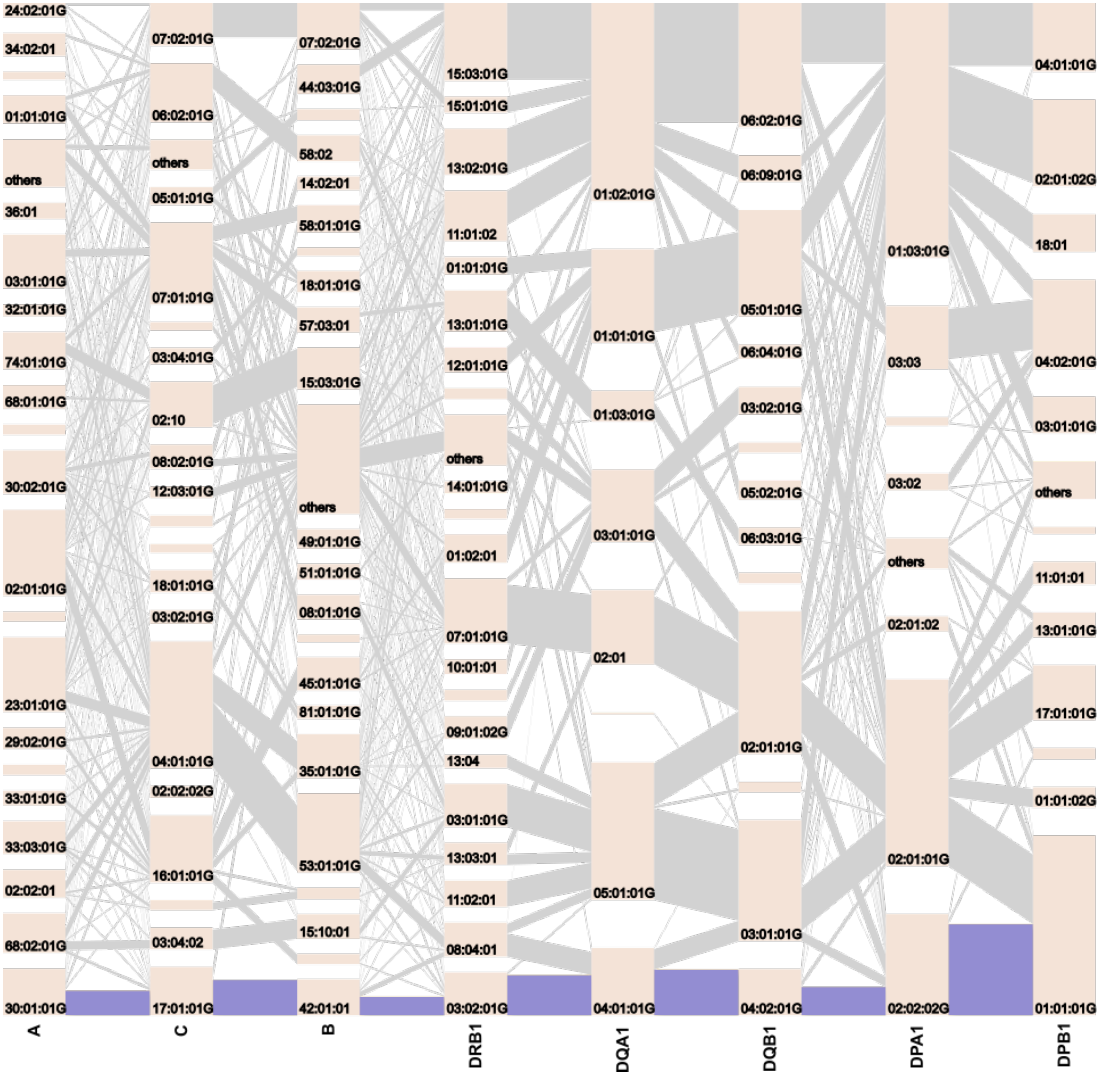
Supplementary Figure 13: Haplotype structure for eight classical HLA genes in Europeans. The haplotype structures of the eight classical HLA genes. The tile in a bar represents an *HLA* allele, and its height corresponds to the frequencies of the *HLA* allele. All long-range *HLA* haplotypes with frequency >1% are bolded and highlighted in blue.



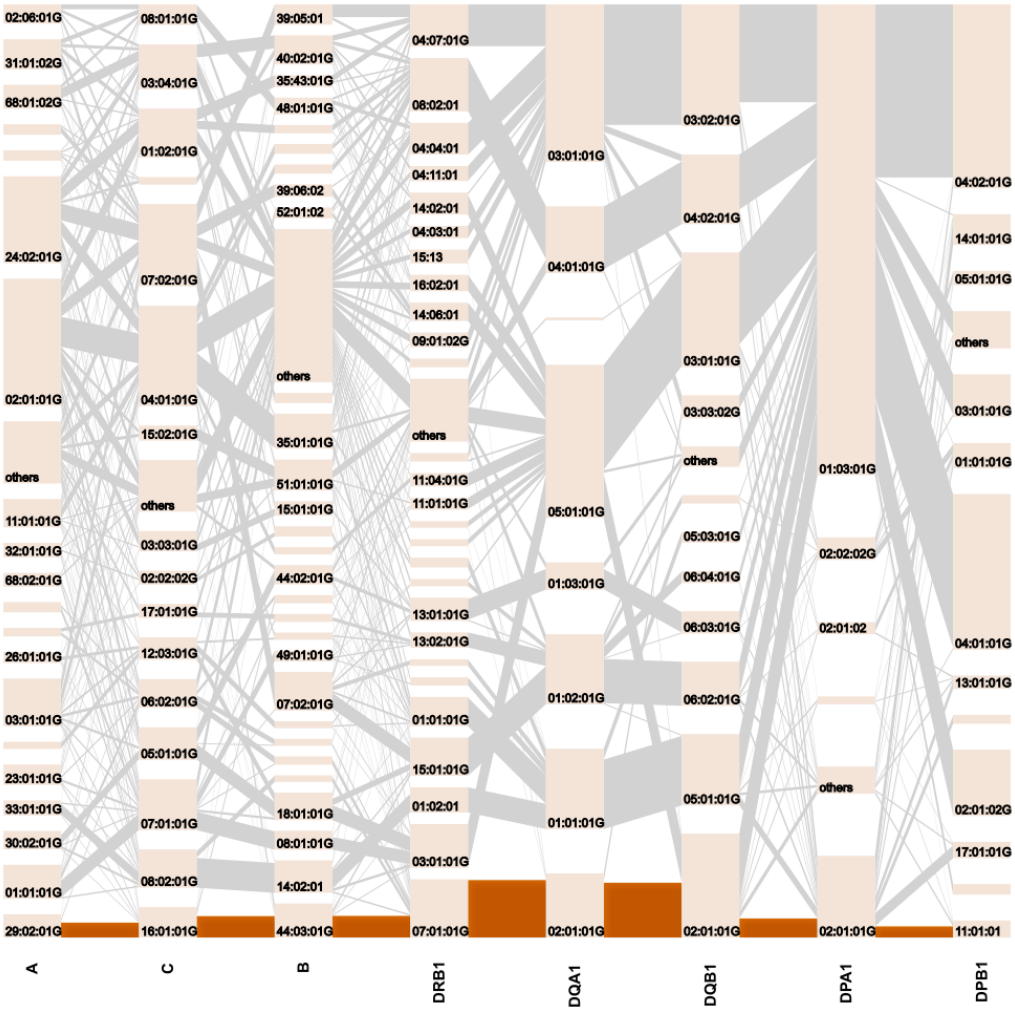
Supplementary Figure 14: Haplotype structure for eight classical HLA genes in South Asians. The haplotype structures of the eight classical HLA genes. The tile in a bar represents an *HLA* allele, and its height corresponds to the frequencies of the *HLA* allele. All long-range *HLA* haplotypes with frequency >1% are in bold and highlighted in green.



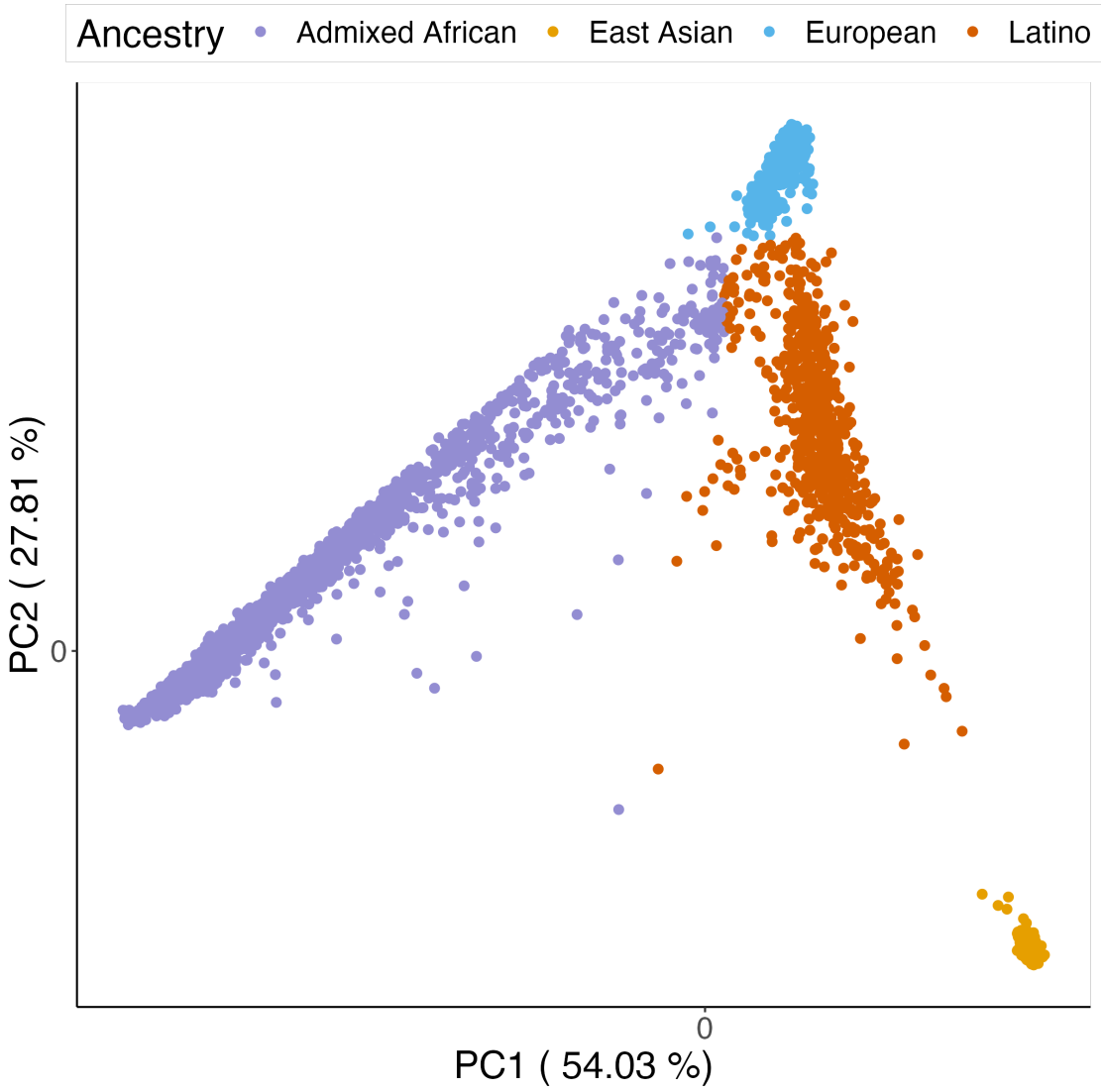
Supplementary Figure 15: Haplotype structure for eight classical HLA genes in Admixed Africans. The haplotype structures of the eight classical HLA genes. The tile in a bar represents an *HLA* allele, and its height corresponds to the frequencies of the *HLA* allele. All long-range *HLA* haplotypes with frequency >1% are in bold and highlighted in purple.



Supplementary Figure 16: Haplotype structure for eight classical HLA genes in Latinos. The haplotype structures of the eight classical HLA genes. The tile in a bar represents an *HLA* allele, and its height corresponds to the frequencies of the *HLA* allele. All long-range *HLA* haplotypes with frequency >1% are in bold and highlighted in orange.

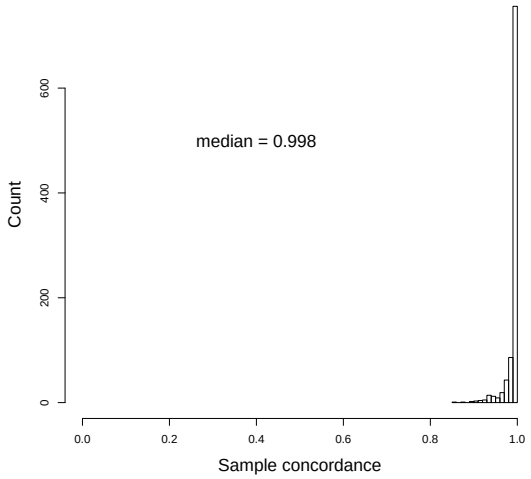


Supplementary Figure 17: Principal component analysis (PCA) of the MESA GWAS samples. The genotypes projected onto the first two principal components using SNP weights precomputed from samples in the 1000 Genomes Phase 3 project using SNPweights. Four diverse ancestries were collected in the MESA project - Admixed African (purple); East Asian (yellow); European (blue) and Latino (orange).

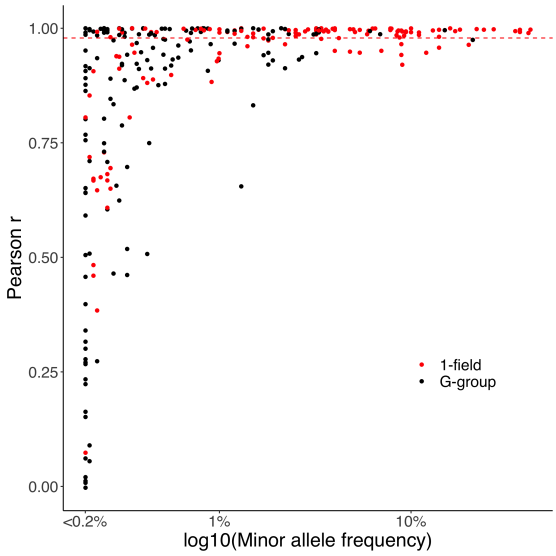


Supplementary Figure 18: HLA imputation performance evaluation using Beagle v.4 and Minimac4. Using 2,291 genotyped data from the GaP registry, we evaluated the concordance rate of inferred dosage at all HLA variations between two different imputation methods - Beagle v4 and Minimac4. (a) A histogram of per sample concordance rate using all imputed HLA variations. (b) Correlation between Beagle v4 and Mimimac4 imputed dosage (Pearson r) of classical *HLA* alleles at 1-field and G-group resolution. The x-axis shows the log10 scaled minor allele frequencies.

(a) Per sample concordance



(b) Per allele dosage correlation



2. Supplementary Tables

Supplementary Table 1: Summary of individual cohort size included in the project prior to quality control. A description of all whole-genome sequencing cohorts used to construct the reference panel. 1KG = 1000 Genomes Project; COPDGene = The chronic Obstructive Pulmonary Disease Gene study; EST = Estonian Biobank; JHS = Jackson Heart Study; MESA = Multi-Ethnic Study of Atherosclerosis; JPN (JBIC) = Japan Biological Informatics Consortium; JPN (BBJ) = Biobank Japan Project

Cohort	Sample size	Sex (% male)	Age (range)	Genome build	Sequencing depth in MHC region
1KG	2,504	NA	NA	GRCh38	46.1
COPDGene	10,623	47%	45-80	GRCh38	35.0
EST	2,244	50%	18-79	GRCh37	26.1
JHS	3,027	64%	65-84	GRCh37	29.6
MESA	4,620	53%	45-84	GRCh38	36.9
JPN (JBIC)	295	21.6%	16-68	GRCh37	26.0
JPN (BBJ)	1,025	41.4%	NA	GRCh37	18.1

Supplementary Table 2: Number of samples within each ancestry group included in the reference panel after quality control. AA represents individuals of Admixed African ancestry; EAS represents individuals of East Asian ancestry; EUR represents individuals of European ancestry; LAT represents individuals of Native American ancestry; SAS represents individuals of South Asian ancestry.

Ancestry	Number of samples
AA	7,849
EAS	2,069
EUR	10,187
LAT	952
SAS	489
Total	21,546

Supplementary Table 3: Description of sample quality control. Samples have been excluded from the reference panel if they have coverage < 20x in any of the eight HLA classical genes included or failed genome-wide quality control (when information is provided).

Criteria	Number of excluded samples
coverage <20x	426
genome-wide QC	2,540
Total	2,792

Supplementary Table 4: Concordance rate between HLA*PRG inferred and sequencing based *HLA* alleles at one-field, amino acid and G-group resolution within 288 individuals from the JPN cohort.

Resolution	A	B	C	DQA1	DQB1	DRB1
one-field	0.991	0.991	0.998	0.995	0.998	0.984
amino acid	0.998	1.000	0.999	0.996	1.000	0.998
G-group	0.963	0.984	0.990	0.994	0.984	0.980

Supplementary Table 5: Concordance rate between HLA*LA inferred and sequencing based typing *HLA* alleles at one-field, amino acid (AA) and G-group resolution of 955 individuals included in the 1000 Genomes Project.

	A			B			C			DQB1			DRB1		
Ancestry	**1-field**	**AA**	**G-group**	**1-field**	**AA**	**G-group**	**1-field**	**AA**	**G-group**	**1-field**	**AA**	**G-group**	**1-field**	**AA**	**G-group**
AFR	0.988	0.968	0.974	0.962	0.999	0.904	0.991	0.994	0.918	0.985	0.999	0.962	0.968	0.997	0.933
EAS	0.998	0.925	0.929	0.951	0.998	0.925	0.994	0.992	0.966	0.996	1.000	0.987	0.976	0.992	0.923
EUR	0.991	0.983	0.986	0.988	0.999	0.972	0.997	0.997	0.985	0.995	1.000	0.989	0.988	0.998	0.971
LAT	0.990	0.961	0.961	0.987	0.998	0.917	0.997	0.997	0.977	0.995	1.000	0.987	0.974	0.994	0.909

Supplementary Table 6: Total number of markers included in the multi-ancestry MHC and the European-only T1DGC reference panel.

	classical *HLA* alleles	amino acids	HLA SNPs	intergenic SNPs	Total
multi-ancestry WGS (N=21,546)	640	4,513	10,923	38,398	54,474
T1DGC (N=5,225)	424	1,246	985	5,698	8,353

Supplementary Table 7: Concordance rate between imputed and sequence based typing classical *HLA* alleles at 1-filed, 2-field and G-group resolution among 955 individuals included in the 1000 Genomes Project. We kept 6,007 markers overlapping with the Global Genotyping Array for imputation. We also removed 955 overlapping samples included in the multi-ancestry reference panel before imputation.

		A			B			C			DQB1			DRB1		
	Ancestry	**1-field**	**2-field**	**G-group**	**1-field**	**2-field**	**G-group**	**1-field**	**2-field**	**G-group**	**1-field**	**2-field**	**G-group**	**1-field**	**2-field**	**G-group**
Multi-ancestry	**AFR**	0.988	0.974	0.974	0.962	0.930	0.906	0.991	0.944	0.962	0.985	0.962	0.933	0.968	0.953	0.933
	EAS	0.998	0.933	0.929	0.951	0.929	0.925	0.996	0.972	0.987	0.996	0.987	0.921	0.974	0.921	0.923
	EUR	0.991	0.988	0.988	0.988	0.974	0.972	0.997	0.985	0.989	0.997	0.989	0.969	0.986	0.969	0.971
	LAT	0.990	0.961	0.961	0.987	0.917	0.917	0.997	0.977	0.987	0.995	0.987	0.909	0.974	0.909	0.909
T1DGC	**AFR**	0.977	0.950	0.950	0.936	0.901	0.898	0.982	0.892	0.863	0.971	0.944	0.944	0.950	0.918	0.839
	EAS	0.994	0.916	0.912	0.919	0.878	0.873	0.993	0.927	0.923	0.994	0.989	0.989	0.957	0.841	0.841
	EUR	0.994	0.994	0.969	0.986	0.963	0.963	0.998	0.991	0.946	0.991	0.986	0.986	0.986	0.972	0.972
	LAT	0.995	0.951	0.917	0.966	0.826	0.813	1.000	0.964	0.948	0.992	0.990	0.990	0.982	0.881	0.870

Supplementary Table 8: Imputation accuracy evaluation of the GaP registry samples. Imputation accuracy is evaluated between imputed and sequencing based typing *HLA* alleles in (a) three class I genes; and (b) three class II genes at 1-field, 2-field and G-group resolution of 75 individuals from three ancestry groups (Admixed African (AA); EAS (East Asian) and European (EUR)) included in the GaP registry.

(a) class I genes

		A			B			C		
	Ancestry	**1-field**	**2-field**	**G-group**	**1-field**	**2-field**	**G-group**	**1-field**	**2-field**	**G-group**
Multi-ancestry	**AA**	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.778	1.000
	EAS	1.000	1.000	1.000	0.953	0.953	0.848	1.000	1.000	1.000
	EUR	1.000	1.000	1.000	0.968	0.973	0.973	1.000	0.929	1.000
T1DGC	**AA**	1.000	0.950	0.950	0.895	0.833	0.799	1.000	0.722	0.750
	EAS	1.000	0.929	0.929	0.967	0.762	0.696	1.000	0.857	0.923
	EUR	1.000	1.000	1.000	0.986	0.943	0.943	1.000	0.929	0.923

(b) class II genes

		DQA1			DQB1			DRB1		
	Ancestry	**1-field**	**2-field**	**G-group**	**1-field**	**2-field**	**G-group**	**1-field**	**2-field**	**G-group**
Multi-ancestry	AA	0.992	0.726	0.875	1.000	0.792	0.950	1.000	0.882	0.917
	EAS	1.000	1.000	0.889	1.000	1.000	1.000	1.000	1.000	0.980
	EUR	1.000	0.731	0.875	1.000	0.833	0.909	0.969	0.913	0.913
T1DGC	AA	1.000	0.718	0.857	0.980	0.741	0.889	1.000	0.824	0.833
	EAS	1.000	0.670	0.889	1.000	0.894	0.983	0.987	0.863	0.922
	EUR	1.000	0.731	0.875	1.000	0.833	0.909	1.000	0.900	0.900

Supplementary Table 9: Imputation accuracy between imputed and next-generation sequencing based *HLA* alleles at amino acid level of 75 individuals from the GaP registry. Imputation accuracy are stratified by amino acid positions inside (exon 2 and 3 for HLA class I genes and exon 2 for HLA class II genes) and outside peptide binding site (PBS). AA=Admixed African; EAS = East Asian; and EUR=European. (a) per class I genes; (b) per class II genes; and (c) average accuracy.

(a) Imputation accuracy for each class I gene

	A			B			C		
Ancestry	**All**	**inside PBS**	**outside PBS**	**All**	**inside PBS**	**outside PBS**	**All**	**inside PBS**	**outside PBS**
Admixed African	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	99.9%	100.0%	99.9%
East Asian	100.0%	100.0%	100.0%	99.7%	99.5%	100.0%	100.0%	100.0%	100.0%
European	100.0%	100.0%	100.0%	99.8%	99.7%	99.9%	100.0%	100.0%	100.0%

(b) Imputation accuracy for each class II gene

	DPB1			DQB1			DRB1		
Ancestry	**All**	**inside PBS**	**outside PBS**	**All**	**inside PBS**	**outside PBS**	**All**	**inside PBS**	**outside PBS**
Admixed African	99.9%	99.9%	99.9%	98.1%	99.5%	97.3%	98.4%	99.9%	97.6%
East Asian	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	99.5%	99.9%	99.3%
European	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	98.4%	99.7%	97.8%

(c) Average imputation accuracy across all six classical HLA genes

	AA	EAS	EUR
all	99.4%	99.9%	99.7%
inside PBS	99.9%	99.9%	99.9%
outside PBS	99.1%	99.9%	99.6%

Supplementary Table 10: HIV GWAS cohorts descriptions. AA=Admixed African; EUR=European; LAT=Latino.

Cohort abbreviation	Number of samples	markers in the MHC region	Ancestry	Sex (% female)	Name of cohort that approved the study protocol
cs1	380	581	AA	28.5	The International HIV Controllers Study & The AIDS Clinical Trials Group
cs2	379	2,026	AA	34.1	The International HIV Controllers Study & The AIDS Clinical Trials Group
cs3	412	1,908	AA	29.4	The International HIV Controllers Study & The AIDS Clinical Trials Group
iav	405	2,147	AA	41.2	The Internatinoal AIDS Vaccine Initiative
ldg	850	507	AA	19.4	The AIDS Linked to the IntraVenous Experience (ALIVE) Cohort, The Multicenter AIDS Cohort Study (MACS) & The Multicenter Hemophilia Cohort Studies (MHCS)
mdo	587	1,834	AA	4.4	Center for HIV/AIDS Vaccine Immunology (CHAVI)
pum	358	419	AA	100	The Pumwani Sex Workers Cohort, University of Nairobi
uhs	530	3,198	AA	28.1	Urban Health Study: Genetics Cohort
acs	408	294	EUR	9.1	The Amsterdam Cohort Studies on HIV infection and AIDS
cs1_eur	505	497	EUR	88.5	The International HIV Controllers Study & The AIDS Clinical Trials Group
cs2_eur	578	1,816	EUR	11.8	The International HIV Controllers Study & The AIDS Clinical Trials Group
cs3_eur	580	1,582	EUR	10.3	The International HIV Controllers Study & The AIDS Clinical Trials Group
ecs	1,507	463	EUR	27.1	Center for HIV/AIDS Vaccine Immunology (EuroCHAVI)
fgp	962	259	EUR	17.2	The nonprogressor Genomics of Resistance to Immunodeficiency Virus Study & The ANRS PRIMO cohort
lgd	1,476	553	EUR	1.6	The AIDS Linked to the IntraVenous Experience (ALIVE) Cohort, The Multicenter AIDS Cohort Study (MACS) & The Multicenter Hemophilia Cohort Studies (MHCS)
md1	768	1,803	EUR	0	The Multicenter AIDS Cohort Study
md2	422	461	EUR	0	The Multicenter AIDS Cohort Study
uhs_eur	239	2,982	EUR	16.7	Urban Health Study: Genetics Cohort
his1	190	2,330	LAT	23.2	The International HIV Controllers Study & The AIDS Clinical Trials Group
his2	186	2,250	LAT	18.3	The International HIV Controllers Study & The AIDS Clinical Trials Group
his3	301	2,293	LAT	81.4	The International HIV Controllers Study & The AIDS Clinical Trials Group

Supplementary Table 11: Imputation accuracy of 1,067 African American individuals. Imputation accuracy are measured between imputed and sequencing based *HLA* alleles with minor allele frequency >0.5% at one-field, amino acid and G-group resolution of 1,067 HIV-1 infected African American individuals included in the McLaren et al. 2012[1]

	one-field	amino acid	G-group
A	0.959	0.991	0.956
B	0.920	0.992	0.924
C	0.969	0.985	0.940
DPB1	0.910	1.000	0.930
DQA1	0.900	0.998	0.979
DQB1	0.986	0.972	0.971
DRB1	0.975	0.997	0.900

Supplementary Table 12: Independently associated amino acid positions in HLA proteins identified by step-wise forward conditional analysis. Joint P-value (P) and conditional P-values (P_{cond}) and cumulative variance explained from the haplotype analysis using one-sided F-test are reported both in the joint analysis and in the individual populations. AA represents Admixed Africans; EUR represents Europeans and LAT represents Latinos.

Gene	Pos	Joint			AA			EUR			LAT		
		P	P_{cond}	variance explained	P	P_{cond}	variance explained	P	P_{cond}	variance explained	P	P_{cond}	variance explained
B	**97**	2.86×10^{-184}	NA	0.091	9.17×10^{-44}	NA	0.077	6.07×10^{-142}	NA	0.109	1.97×10^{-8}	NA	0.147
B	**67**	1.29×10^{-99}	1.08×10^{-40}	0.112	5.29×10^{-16}	1.22×10^{-8}	0.097	1.44×10^{-112}	1.01×10^{-26}	0.132	2.07×10^{-6}	0.460	0.166
B	**156**	1.07×10^{-24}	1.92×10^{-30}	0.129	6.06×10^{-5}	3.07×10^{-24}	0.138	4.53×10^{-24}	4.36×10^{-4}	0.134	0.363	0.315	0.175
A	**77**	1.90×10^{-13}	5.35×10^{-7}	0.140	6.31×10^{-2}	0.182	0.137	2.70×10^{-12}	3.21×10^{-6}	0.149	0.115	5.76×10^{-2}	0.263

Supplementary Table 13: The top ten pairs of polymorphic amino acid positions in HLA-B among 7,260 combinations (sorted by P-value). For each pair of amino acid positions, we used the groups of residues occurring at these positions (degree of freedom) to estimate effect size and calculated for each of these models the delta deviance (sum of squares) in risk prediction and its P-values compared to the null model using one-sided F-test.

Gene	Amino acid positions	Degree of freedom	Sum of squares	P-value	Variance explained
B	67,97	16	1072.14	4.75×10^{-218}	0.112
B	45,97	13	1027.66	7.87×10^{-211}	0.105
B	63,97	9	1007.18	2.98×10^{-210}	0.104
B	74,97	9	1006.16	5.10×10^{-210}	0.103
B	80,97	12	1002.44	4.20×10^{-206}	0.103
B	97,245	6	972.00	1.41×10^{-205}	0.103
B	82,97	9	985.34	2.67×10^{-205}	0.101
B	83,97	9	985.34	2.67×10^{-205}	0.101
B	97,156	13	1002.24	4.45×10^{-205}	0.103
B	97,143	7	972.52	1.45×10^{-204}	0.103

Supplementary Table 14: The top ten triplets of polymorphic amino acid positions in HLA-B among 287,980 combinations (sorted by P-value). For each triplets of amino acid positions, we used the groups of residues occurring at these positions (degree of freedom) to estimate effect size and calculated for each of these models the delta deviance (sum of squares) in risk prediction and its P-values compared to the null model using one-sided F-test.

Gene	Amino acid positions	Degree of freedom	Sum of squares	P-value	Variance explained
B	67,97,156	26	1224.88	5.68×10^{-244}	0.129
B	67,97,245	18	1183.64	1.15×10^{-241}	0.126
B	67,97,143	19	1184.26	7.07×10^{-241}	0.126
B	67,97,147	19	1184.26	7.07×10^{-241}	0.126
B	74,97,156	19	1172.17	4.22×10^{-238}	0.121
B	63,97,245	11	1121.45	3.00×10^{-234}	0.118
B	63,97,156	19	1155.14	3.39×10^{-234}	0.120
B	45,97,245	14	1133.78	4.77×10^{-234}	0.118
B	74,97,245	11	1119.80	7.21×10^{-234}	0.117
B	63,97,143	12	1123.17	1.29×10^{-233}	0.119

Supplementary Table 15: Effects of individual amino acids within HLA proteins. For each amino acid position highlighted in this or previous studies (45, 63, 67, 97, 116, 156 and 245 in HLA-B and 77 and 95 in HLA-A) from individual population (European (EUR) results are reported by McLaren et al. 2015 [2] [PMID: 22718199]; African American (AA) results are reported by McLaren et al. 2012 [1] [PMID: 22718199]). 1 indicates an amino acid position has been reported; 0 indicates otherwise. We tested a multiallelic association between the HIV-1 viral load and a haplotype matrix with covariates, including sex, study-specific PCs, and a categorical variable indicating a population. P-values are obtained using one-sided F-test. Ref represents the reference residue.

			Unadjusted allele frequency				Publication			Joint			Admixed African			European			Latino		
Gene	pos	amino acid	AA	EUR	LAT	Joint	This	EUR	AA	beta	se	P	beta	se	P	beta	se	P	beta	se	P
B	97	N	0.012	0.046	0.036	0.035	1	1	1	-0.542	0.040	6.96E-42	-0.562	0.123	4.87E-06	-0.523	0.043	2.18E-34	-0.840	0.194	1.55E-05
B	97	R	0.428	0.495	0.480	0.479	1	1	1	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref
B	97	S	0.207	0.239	0.195	0.228	1	1	1	-0.004	0.018	8.48E-01	-0.148	0.034	1.14E-05	0.049	0.021	2.23E-02	0.108	0.082	1.84E-01
B	97	T	0.068	0.127	0.142	0.110	1	1	1	-0.147	0.024	0.000	-0.074	0.054	0.171	-0.176	0.027	1.48E-10	-0.001	0.090	9.91E-01
B	97	V	0.065	0.050	0.054	0.055	1	1	1	-0.880	0.033	0.000	-0.770	0.055	0.000	-0.943	0.042	2.66E-110	-0.972	0.160	1.31E-09
B	97	W	0.079	0.039	0.056	0.052	1	1	1	-0.213	0.033	-0.018	0.049	0.718	0.272	-0.377	0.046	4.92E-16	-0.429	0.153	5.18E-03
B	67	C	0.091	0.132	0.167	0.121	1	1	0	-0.222	0.025	1.57E-19	-0.085	0.050	9.10E-02	-0.265	0.028	1.99E-21	-0.233	0.094	1.34E-02
B	67	F	0.246	0.255	0.230	0.251	1	1	0	0.089	0.019	2.55E-06	0.056	0.035	1.07E-01	0.094	0.022	1.58E-05	0.300	0.079	1.38E-04
B	67	M	0.170	0.064	0.091	0.097	1	1	0	-0.445	0.027	1.19E-59	-0.258	0.040	6.61E-11	-0.688	0.038	1.90E-71	-0.408	0.130	1.74E-03
B	67	S	0.317	0.404	0.415	0.376	1	1	0	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref
B	67	Y	0.180	0.145	0.098	0.154	1	1	0	0.060	0.022	7.11E-03	-0.132	0.039	6.79E-04	0.163	0.027	1.07E-09	0.128	0.120	2.86E-01
B	156	D	0.176	0.209	0.153	0.197	1	0	0	0.172	0.019	1.10E-18	0.188	0.036	2.24E-07	0.165	0.023	3.16E-13	0.199	0.095	3.66E-02
B	156	L	0.275	0.399	0.257	0.351	1	0	0	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref
B	156	R	0.092	0.130	0.075	0.116	1	0	0	0.180	0.024	8.92E-14	0.027	0.048	5.79E-01	0.224	0.028	4.97E-16	0.336	0.141	1.70E-02
B	156	W	0.008	0.062	0.026	0.044	1	0	0	-0.005	0.037	8.90E-01	-0.078	0.151	6.06E-01	0.006	0.039	8.76E-01	-0.152	0.239	5.25E-01
B	245	A	0.028	0.001	0.018	0.026	0	0	1	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref
B	245	T	0.026	0.000	0.016	0.009	0	0	1	-0.684	0.080	1.84E-17	-0.783	0.086	1.44E-19	-0.627	0.366	8.67E-02	0.283	0.278	3.09E-01
A	77	D	0.645	0.605	0.651	0.619	1	1	0	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref
A	77	N	0.335	0.330	0.308	0.331	1	1	0	0.083	0.016	3.28E-07	0.045	0.030	1.32E-01	0.100	0.020	4.94E-07	-0.012	0.098	9.02E-01
A	77	S	0.019	0.064	0.041	0.048	1	1	0	-0.183	0.035	2.04E-07	-0.172	0.102	8.97E-02	-0.181	0.039	3.08E-06	-0.464	0.223	3.80E-02
B	45	E	0.372	0.373	0.321	0.370	0	1	0	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref
B	45	K	0.178	0.237	0.269	0.221	0	1	0	0.068	0.020	6.19E-04	0.130	0.038	7.07E-04	0.047	0.023	3.96E-02	0.039	0.073	5.87E-01
B	45	M	0.104	0.145	0.112	0.131	0	1	0	-0.402	0.024	3.20E-62	-0.434	0.047	3.48E-20	-0.384	0.028	1.35E-43	-0.614	0.120	2.87E-07
B	45	T	0.348	0.244	0.290	0.277	0	1	0	0.084	0.019	7.44E-06	0.138	0.032	1.82E-05	0.054	0.023	1.79E-02	0.094	0.072	1.88E-01
A	95	I	0.319	0.397	0.438	0.374	0	1	0	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref
A	95	L	0.197	0.117	0.175	0.147	0	1	0	0.124	0.022	2.01E-08	0.110	0.035	1.70E-03	0.150	0.030	5.61E-07	-0.076	0.119	5.23E-01
A	95	V	0.121	0.279	0.262	0.227	0	1	0	0.040	0.018	3.04E-02	0.121	0.042	3.70E-03	0.030	0.021	1.64E-01	0.037	0.103	7.19E-01
B	116	D	0.089	0.182	0.156	0.153	0	0	1	-0.084	0.023	2.45E-04	0.009	0.051	8.59E-01	-0.115	0.025	4.72E-06	-0.308	0.100	2.09E-03
B	116	F	0.056	0.117	0.141	0.100	0	0	1	-0.092	0.027	7.29E-04	-0.093	0.063	1.39E-01	-0.097	0.030	1.36E-03	-0.426	0.102	2.69E-05
B	116	L	0.101	0.082	0.091	0.088	0	0	1	0.072	0.028	1.15E-02	0.363	0.049	1.17E-13	-0.040	0.035	2.52E-01	-0.351	0.123	4.39E-03
B	116	S	0.388	0.240	0.236	0.202	0	0	1	-0.055	0.019	3.97E-03	0.201	0.031	1.53E-10	-0.185	0.023	1.37E-15	-0.266	0.086	2.10E-03
B	116	Y	0.363	0.377	0.400	0.374	0	0	1	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref
B	63	E	0.467	0.465	0.505	0.467	0	0	1	-0.191	0.015	1.09E-36	-0.097	0.027	2.38E-04	-0.231	0.018	2.57E-39	-0.246	0.053	2.99E-06
B	63	N	0.471	0.466	0.498	0.470	0	0	1	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref	ref

Supplementary Table 16: Effect estimates and frequencies for the three amino acids in HLA-B associated with HIV-1 viral load. Estimated effects for haplotypes of three amio acids in HLA-B. The effect size and its standard error from a multivariate regression is listed for each population, taking the most frequent haplotype as the reference (that is, giving that haplotype an effect size of 0). Unadjusted haplotype frequencies are given in each and overall populations. AA represents Admixed African; EUR represents European and LAT represents Latino populations, respectively. Heterogeneity p-value (P(het)) of each haplotype is calculated using two-sided F-statistics with two degrees of freedom (**Methods**). Effect size and its standard error in each population are listed only for haplotypes that show evidence of heterogeneity (P-value < 0.05 /26 Bonferroni-corrected for multiple tests, bolded). Ref represents the reference haplotype.

HLA-B at position			Effect size (standard error)					Unadjusted allele frequency				Classical HLA-*B* allele
97	**67**	**156**	**AA**	**EUR**	**LAT**	**Joint**	**P_het**	**AA**	**EUR**	**LAT**	**Joint**	
V	M	L				-0.921 (0.036)	0.031	0.056	0.049	0.059	0.051	*B*57:01;B*57:03*
S	Y	L				-0.886 (0.082)	0.236	0.026	NA	NA	0.008	*B*81:01*
S	C	L				-0.776 (0.212)	0.700	NA	0.001	NA	0.001	
R	Y	L				-0.759 (0.134)	0.798	0.009	NA	NA	0.003	*B*39:10*
V	M	R				-0.645 (0.123)	0.869	0.010	NA	NA	0.003	*B*57:02*
N	C	L				-0.554 (0.041)	0.257	0.012	0.046	0.037	0.035	*B*27:05*
T	S	L				-0.436 (0.041)	0.041	0.028	0.039	0.056	0.037	*B*13:02;B*52:01*
W	C	L				-0.397 (0.041)	0.581	0.030	0.039	0.054	0.037	*B*14:01;B*14:02*
S	S	L				-0.252 (0.066)	0.013	0.002	0.014	0.070	0.013	*B*40:02*
R	S	W				-0.177 (0.038)	0.618	0.009	0.062	0.028	0.044	*B*15:01;B*15:10;B*15:16*
S	S	D				-0.171 (0.101)	0.237	0.007	0.004	NA	0.005	*B*41:02*
T	F	L				-0.125 (0.036)	0.001	0.030	0.059	0.073	0.051	*B*51:01;B*78:01*
R	M	L				-0.125 (0.045)	0.375	0.061	0.014	0.028	0.029	*B*15:16;B*58:01*

R	C	L				-0.078 (0.039)	0.055	0.042	0.039	0.060	0.041	*B*15:10;B*15:16;B*39:10*
R	S	D	0.165 (0.056)	-0.07 (0.034)	-0.153 (0.173)	-0.019 (0.028)	**0.002**	0.075	0.108	0.084	0.097	*B*37:01;B*44:02;B*45:01*
R	S	L				Ref		0.191	0.176	0.197	0.180	*B*15:03;B*15:10;B*18:01; B*39:10;B*40:01;B*44:03; B*44:03;B*49:01;B*50:01*
S	Y	D				0.015 (0.055)	0.884	0.059	NA	0.017	0.019	*B*42:01;B*42:02*
S	Y	R	-0.06 (0.055)	0.037 (0.033)	-0.002 (0.187)	0.022 (0.027)	**0.007**	0.080	0.124	0.070	0.108	*B*07:02;B*07:05*
R	F	R				0.034 (0.117)	0.302	NA	0.005	NA	0.004	
S	F	D				0.041 (0.031)	0.218	0.034	0.095	0.042	0.074	*B*08:01*
R	F	L				0.045 (0.027)	0.730	0.182	0.095	0.113	0.122	*B*35:01;B*53:01*
T	F	D				0.081 (0.207)	0.842	NA	0.002	NA	0.001	
T	C	L				0.086 (0.096)	0.583	0.002	0.006	0.014	0.006	*B*39:10*
W	M	L				0.098 (0.064)	0.268	0.046	NA	NA	0.014	*B*58:02*
W	C	R				0.127 (0.243)	1.000	0.003	NA	NA	0.001	
T	Y	L				0.176 (0.058)	0.207	0.005	0.021	NA	0.016	
R	Y	D				0.285 (0.28)	1.000	0.002	NA	NA	0.001	

Supplementary Table 17: Effect sizes of additive and non-additive effects for the three reported amino acid positions (97, 67 and 156) in HLA-B associated with HIV-1 viral load. Additive and non-additive effects are shown for haplotype formed by amino acids 97, 67 and 156 in HLA-B with a frequency >0.5% . P-values were obtained from one-sided F-test indicate the significance of improvement in fit by amino acid-specific models after sequentially including the additive and the non-additive term for a given amino acid. Effect sizes and standard errors (s.e.) are given for a purely additive scenario and for a non-additive scenario; in the latter case, effect sizes are given separately for heterozygotes (het) and homozygotes (hom).

HLA-B position			**Additive Model**		**Non-additive Model (add. + non-add. components)**	
97	**67**	**156**	**Effect size (s.e.)**	**P**	**Het effect (s.e.)**	**Hom effect (s.e.)**
V	M	L	-0.921 (0.036)	0.002	-1.35 (0.13)	-0.51 (0.13)
S	Y	L	-0.886 (0.082)	0.288	-0.4 (0.5)	-1.44 (0.5)
N	C	L	-0.554 (0.041)	0.007	-0.91 (0.13)	-0.2 (0.13)
T	S	L	-0.436 (0.041)	0.150	-0.63 (0.13)	-0.27 (0.13)
W	C	L	-0.397 (0.041)	0.285	-0.59 (0.16)	-0.21 (0.16)
S	S	L	-0.252 (0.066)	0.808	-0.15 (0.29)	-0.36 (0.29)
R	S	W	-0.177 (0.038)	0.237	-0.32 (0.11)	-0.05 (0.11)
S	S	D	-0.171 (0.101)	0.101	-1.03 (0.5)	0.71 (0.5)

T	F	L	-0.125 (0.036)	0.115	-0.32 (0.1)	0.02 (0.1)
R	M	L	-0.125 (0.045)	0.957	-0.11 (0.13)	-0.13 (0.13)
R	C	L	-0.078 (0.039)	0.229	-0.21 (0.12)	0.05 (0.12)
R	S	D	-0.019 (0.028)	0.677	-0.04 (0.06)	-0.03 (0.06)
S	Y	D	0.015 (0.055)	0.849	-0.01 (0.16)	0.04 (0.16)
S	Y	R	0.022 (0.027)	0.626	0 (0.06)	0.05 (0.06)
S	F	D	0.041 (0.031)	0.701	0 (0.07)	0.06 (0.07)
R	F	L	0.045 (0.027)	0.481	0.07 (0.05)	0 (0.05)
T	F	D	0.081 (0.207)	NA	0.14 (0.29)	0 (0.29)
T	C	L	0.086 (0.096)	0.820	0.16 (0.5)	-0.02 (0.5)
W	M	L	0.098 (0.064)	0.945	0.12 (0.19)	0.15 (0.19)
T	Y	L	0.176 (0.058)	0.415	-0.05 (0.25)	0.37 (0.25)

Supplementary Table 18: Number of classical *HLA* alleles observed in each ancestry observed in the multi-ancestry MHC reference panel. AA represents individuals with Admixed African; EAS represents individuals of East Asian ancestry; EUR represents individuals of European ancestry; LAT represents individuals of Native American ancestry; SAS represents individuals of South Asian ancestry.

	A	B	C	DQA1	DQB1	DRB1	DPA1	DPB1
AA	164	220	108	13	62	145	25	73
EAS	84	93	51	10	26	70	20	36
EUR	201	230	134	13	77	155	23	70
LAT	85	125	44	10	31	66	20	38
SAS	35	52	26	8	19	44	17	23
Total	330	443	230	17	115	255	27	115

Supplementary Table 19: Hardy Weinberg test results of the eight classical HLA genes. P-values are derived from Hardy-Weinberg exact test. P-values < 0.05 are considered to be significant (sig) indicated by *).

	Admixed American		European		East Asian		Latino		South Asian	
	P-value	**Sig.**	**P-value**	**Sig.**	**P-value**	**Sig.**	**P-value**	**Sig.**	**P-value**	**Sig.**
A	3.50E-02	*	8.05×10^{-1}	NS	6.66×10^{-16}	***	3.71×10^{-2}	*	4.17×10^{-1}	NS
C	3.24×10^{-1}	NS	6.16×10^{-3}	**	1.01×10^{-2}	*	8.61×10^{-1}	NS	2.13×10^{-1}	NS
B	1.69×10^{-1}	NS	8.12×10^{-5}	***	1.52E-11	***	1.78×10^{-1}	NS	5.13×10^{-1}	NS
DRB1	1.82×10^{-5}	***	$< 2.22 \times 10^{-16}$	***	8.77×10^{-12}	***	4.42×10^{-5}	***	6.11×10^{-5}	***
DQA1	2.48×10^{-1}	NS	2.08×10^{-2}	*	2.48×10^{-1}	NS	6.55×10^{-1}	NS	1.85×10^{-1}	NS
DQB1	1.67×10^{-1}	NS	1.63×10^{-1}	NS	5.38×10^{-4}	***	8.00×10^{-1}	NS	6.24×10^{-1}	NS
DPA1	1.22×10^{-3}	**	1.72×10^{-1}	NS	9.56×10^{-2}	NS	2.69×10^{-1}	NS	8.75×10^{-3}	**
DPB1	4.86×10^{-1}	NS	5.42×10^{-1}	NS	8.28×10^{-3}	**	5.28×10^{-2}	NS	3.99×10^{-1}	NS

Supplementary Table 20: Ambiguity in G-group resolution identified among the 75 individuals included in the GaP registry. In total there are 4/146 G-group alleles that have non-unique 2-field translation. The numbers in red shows the false 2-field inference based on the G-group → 2-field reduction. AF = allele frequency; EUR = European; EAS = East Asian; AA = Admixed African.

G-group allele	2-field alleles	EUR	AF_EUR	EAS	AF_EAS	AA	AF_AA	Total alleles	AF_ALL
C*07:01:01G	C*07:01	6	12.0%	1	2.0%	5	10.0%	13	8.7%
	C*07:06	0	0.0%	1	2.0%	0	0.0%	1	0.7%
	C*07:18	1	2.0%	0	0.0%	6	12.0%	7	4.7%
C*17:01:01G	C*17:01	1	2.0%	0	0.0%	4	8.0%	5	3.3%
	C*17:03	0	0.0%	0	0.0%	1	2.0%	1	0.7%
DQB1*02:01:01G	DQB1*02:01	3	6.0%	4	8.0%	3	6.0%	10	6.7%
	DQB1*02:02	3	6.0%	3	6.0%	6	12.0%	12	8.0%
DQB1*03:01:01G	DQB1*03:01	10	20.0%	9	18.0%	4	8.0%	23	15.3%
	DQB1*03:19	0	0.0%	0	0.0%	2	4.0%	2	1.3%

Supplementary Table 21: The most common haplotypes and their frequencies for 2-field alleles that belong to the same G-groups in the 75 individuals from the GaP registry. In total there are nine 2-field alleles that belong to four G-group alleles. We evaluated the most common haplotypes and their frequencies of which these nine 2-field alleles belong in Admixed African (AA), East Asian (EAS), European (EUR) and all 75 individuals from the GaP registry with sequencing-based typing *HLA* allele information.

G-group allele	2-field alleles	Most common haplotype (all)	Freq (all)	Most common haplotype (AA)	Freq (AA)	Most common haplotype (EAS)	Freq (EAS)	Most common haplotype (EUR)	Freq (EUR)
C*07:01:01G	C*07:01	A*01:01~B*08:01~C*07:01	2.59%	A*29:02~B*49:01~C*07:01	4.55%	A*32:01~B*08:01~C*07:01	2.13%	A*01:01~B*08:01~C*07:01	8.33%
	C*07:06	A*33:03~B*44:03~C*07:06	0.86%	NA	NA	A*33:03~B*44:03~C*07:06	2.78%	NA	NA
	C*07:18	A*01:01~B*58:01~C*07:18	0.86%	A*03:01~B*47:03~C*07:18	2.27%	NA	NA	A*11:01~B*58:01~C*07:18	2.78%
C*17:01:01G	C*17:01	A*30:01~B*42:01~C*17:01	2.59%	A*30:01~B*42:01~C*17:01	6.82%	NA	NA	A*02:01~B*41:01~C*17:01	2.78%
	C*17:03	A*34:02~B*41:03~C*17:03	0.86%	A*34:02~B*41:03~C*17:03	2.27%	NA	NA	NA	NA
DQB1*02:01:01G	DQB1*02:01	DPB1*04:01~DQB1*02:01 ~DRB1*03:01	2.83%	DPB1*18:01~DQB1*02:01 ~DRB1*03:01	5.26%	DPB1*04:02~DQB1*02:01 ~DRB1*12:02	3.13%	DPB1*04:01~DQB1*02:01 ~DRB1*03:01	5.56%
	DQB1*02:02	DPB1*01:01~DQB1*02:02 ~DRB1*07:01	1.89%	DPB1*01:01~DQB1*02:02 ~DRB1*07:01	5.26%	DPB1*04:01~DQB1*02:02 ~DRB1*07:01	3.13%	DPB1*03:01~DQB1*02:02 ~DRB1*07:01	2.78%
DQB1*03:01:01G	DQB1*03:01	DPB1*04:01~DQB1*03:01 ~DRB1*11:04	2.83%	DPB1*02:01~DQB1*03:01 ~DRB1*08:04	5.26%	DPB1*05:01~DQB1*03:01 ~DRB1*03:01	6.25%	DPB1*04:01~DQB1*03:01 ~DRB1*11:04	8.33%
	DQB1*03:19	DPB1*01:01~DQB1*03:19 ~DRB1*08:04	0.94%	DPB1*105:01~DQB1*03:19 ~DRB1*11:01	2.63%	NA	NA	NA	NA

Supplementary Table 22: A summary of all available sequence-based typing (SBT) of *HLA* alleles in three different cohorts used for imputation accuracy evaluation.

Cohort	Available HLA genes	Size	Highest Resolution	Method
GaP	A,B, C, DRB1, DQA1, DQB1, DRB1, DPB1	75	G-group for DQA1; 3-field for A, B, C, DPB1, DQB1 and DRB1	NGS and PCR-sequence-specific oligonucleotide probe sequencing (PCR-SSOP) for HLA-DQA1
1000 Genomes	A, B, C, DQB1, DRB1	955	G-group (provided in 2-field resolution)	Sanger sequencing
HIV-1 (AA individuals only)	A, B, C, DQA1, DQB1, DRB1, DPB1	1,067	G-group (provided in 2-field resolution)	sequencing exons 2 and 3 and/or single-stranded conformation polymorphism PCR

Supplementary Table 23: **TOPMed study specific omics support information.**

TOPMed Accession #	TOPMed Project	Parent Study	TOPMed Phase	Omics Center	Omics Support
phs000951	COPD	COPDGene	1	NWGC	3R01HL089856-08S1
phs000951	COPD	COPDGene	2	Broad Genomics	HHSN268201500014C
phs000964	JHS	JHS	1	NWGC	HHSN268201100037C
phs001416	MESA	MESA	2	Broad Genomics	3U54HG003067-13S1

3. Supplementary Note

3.1 Deep-coverage whole-genome sequencing cohort descriptions

Study participants for constructing the multi-ancestry MHC reference panel were from the Jackson Heart Study (JHS, $N = 3,024$), the Multi-Ethnic Study of Atherosclerosis cohort (MESA, $N = 4,260$), the Chronic Obstructi ve Pulmonary Disease Gene study (COPDGene, $N = 10,674$), the Estonian Biobank (EST, $N = 2,244$), the Japan Biological Informatics Consortium (JPN, $N = 295$), the Biobank Japan (JPN, $N = 1,025$) and the 1000 Genomes Project (1KG, $N = 2,504$). Each study was previously approved by respective institutional review boards (IRBs), including for the generation of whole genome sequencing (WGS) data and association with phenotypes. All participants provided written consent. Details of each cohort included in the reference panel construction is described below.

3.1.1 Jackson Heart Study (JHS)

The Jackson Heart Study (JHS) (`https://www.jacksonheartstudy.org`) is a large, population-based observational study evaluating the etiology of cardiovascular, renal, and respiratory diseases among African Americans residing in the three counties (Hinds, Madison, and Rankin) that make up the Jackson, Mississippi metropolitan area. Data and biologic materials have been collected from 5,306 participants, including a nested family cohort of 1,498 members of 264 families. The age at enrollment for the unrelated cohort was 35-84 years; the family cohort included related individuals > 21 years old. Participants provided extensive medical and social history, had an array of physical and biochemical measurements and diagnostic procedures, and provided genomic DNA during a baseline examination (2000-2004) and two follow-up examinations (2005-2008 and 2009-2012). The study population is characterized by a high prevalence of diabetes, hypertension, obesity, and related disorders. Annual follow-up interviews and cohort surveillance are ongoing, and preparation for a fourth examination is in progress.

3.1.2 Multi-Ethnic Study of Atherosclerosis (MESA)

The Multi-Ethnic Study of Atherosclerosis (MESA) is a study of the characteristics of subclinical cardiovascular disease and the risk factors that predict progression to clinically overt cardiovascular disease or progression of the subclinical disease. MESA consisted of a diverse, population-based sample of an initial 6,814 asymptomatic men and women aged 45-84. 38 percent of the recruited participants were white, 28 percent African American, 22 percent Hispanic, and 12 percent Asian, predominantly of Chinese descent. Participants were recruited from six field centers across the United States: Wake Forest University, Columbia University, Johns Hopkins University, University of Minnesota, Northwestern University and University of California - Los Angeles. Each participant received an extensive physical exam and determination of coronary calcification, ventricular mass and function, flow-mediated endothelial vasodilation, carotid intimal-medial wall thickness and presence of echogenic lucencies in the carotid artery, lower extremity vascular insufficiency, arterial wave forms, electrocardiographic (ECG) measures, standard coronary risk factors, sociodemographic factors, lifestyle factors, and psychosocial factors. Selected repetition of subclinical disease measures and risk factors at follow-up visits allowed study of the progression of disease. Participants are being followed for identification and characterization of cardiovascular disease events, including acute myocardial infarction and other forms of coronary heart disease (CHD), stroke, and congestive heart failure; for cardiovascular disease interventions; and for mortality. The first examination took place over two years, from July 2000 - July 2002. It was followed by four examination periods that were 17-20 months in length. Participants have been contacted every 9 to 12 months throughout the study to assess clinical morbidity and mortality.

3.1.3 The Chronic Obstructive Pulmonary Disease Gene (COPDGene) study

Eligible individuals in The Chronic Obstructive Pulmonary Disease Gene (COPDGene) Study (NCT00608764, `www.copdgene.org`) were of non-Hispanic white (NHW) or African-American (AA) ancestry, aged 45-80 years old, with at least 10 pack-years of smoking and no diagnosed lung disease other than COPD or asthma [3]. IRB approval was obtained at all study centers, and all study participants provided

written informed consent.

M.H.C. was supported by NHLBI grants R01HL113264, R01HL137927, and R01HL135142. The COPDGene project (NCT00608764) was supported by Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health.

COPD Foundation Funding
The COPDGene® project is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Novartis, Pfizer, Siemens and Sunovion.

COPDGene® Investigators – Core Units
Administrative Center: James D. Crapo, MD (PI); Edwin K. Silverman, MD, PhD (PI); Barry J. Make, MD; Elizabeth A. Regan, MD, PhD
Genetic Analysis Center: Terri Beaty, PhD; Ferdouse Begum, PhD; Peter J. Castaldi, MD, MSc; Michael Cho, MD; Dawn L. DeMeo, MD, MPH; Adel R. Boueiz, MD; Marilyn G. Foreman, MD, MS; Eitan Halper-Stromberg; Lystra P. Hayden, MD, MMSc; Craig P. Hersh, MD, MPH; Jacqueline Hetmanski, MS, MPH; Brian D. Hobbs, MD; John E. Hokanson, MPH, PhD; Nan Laird, PhD; Christoph Lange, PhD; Sharon M. Lutz, PhD; Merry-Lynn McDonald, PhD; Margaret M. Parker, PhD; Dandi Qiao, PhD; Elizabeth A. Regan, MD, PhD; Edwin K. Silverman, MD, PhD; Emily S. Wan, MD; Sungho Won, Ph.D.; Phuwanat Sakornsakolpat, M.D.; Dmitry Prokopenko, Ph.D.
Imaging Center: Mustafa Al Qaisi, MD; Harvey O. Coxson, PhD; Teresa Gray; MeiLan K. Han, MD, MS; Eric A. Hoffman, PhD; Stephen Humphries, PhD; Francine L. Jacobson, MD, MPH; Philip F. Judy, PhD; Ella A. Kazerooni, MD; Alex Kluiber; David A. Lynch, MB; John D. Newell, Jr., MD; Elizabeth A. Regan, MD, PhD; James C. Ross, PhD; Raul San Jose Estepar, PhD; Joyce Schroeder, MD; Jered Sieren; Douglas Stinson; Berend C. Stoel, PhD; Juerg Tschirren, PhD; Edwin Van Beek, MD, PhD; Bram van Ginneken, PhD; Eva van Rikxoort, PhD; George Washko, MD; Carla G. Wilson, MS;
PFT QA Center, Salt Lake City, UT: Robert Jensen, PhD

Data Coordinating Center and Biostatistics, National Jewish Health, Denver, CO: Douglas Everett, PhD; Jim Crooks, PhD; Camille Moore, PhD; Matt Strand, PhD; Carla G. Wilson, MS
Epidemiology Core, University of Colorado Anschutz Medical Campus, Aurora, CO: John E. Hokanson, MPH, PhD; John Hughes, PhD; Gregory Kinney, MPH, PhD; Sharon M. Lutz, PhD; Katherine Pratte, MSPH; Kendra A. Young, PhD
Mortality Adjudication Core: Surya Bhatt, MD; Jessica Bon, MD; MeiLan K. Han, MD, MS; Barry Make, MD; Carlos Martinez, MD, MS; Susan Murray, ScD; Elizabeth Regan, MD; Xavier Soler, MD; Carla G. Wilson, MS
Biomarker Core: Russell P. Bowler, MD, PhD; Katerina Kechris, PhD; Farnoush Banaei-Kashani, Ph.D.

COPDGene® Investigators – Clinical Centers
Ann Arbor VA: Jeffrey L. Curtis, MD; Carlos H. Martinez, MD, MPH; Perry G. Pernicano, MD.
Baylor College of Medicine, Houston, TX: Nicola Hanania, MD, MS; Philip Alapat, MD; Mustafa Atik, MD; Venkata Bandi, MD; Aladin Boriek, PhD; Kalpatha Guntupalli, MD; Elizabeth Guy, MD; Arun Nachiappan, MD; Amit Parulekar, MD;
Brigham and Women's Hospital, Boston, MA: Dawn L. DeMeo, MD, MPH; Craig Hersh, MD, MPH; Francine L. Jacobson, MD, MPH; George Washko, MD
Columbia University, New York, NY: R. Graham Barr, MD, DrPH; John Austin, MD; Belinda D'Souza, MD; Gregory D.N. Pearson, MD; Anna Rozenshtein, MD, MPH, FACR; Byron Thomashow, MD Duke University Medical Center, Durham, NC: Neil MacIntyre, Jr., MD; H. Page McAdams, MD; Lacey Washington, MD
HealthPartners Research Institute, Minneapolis, MN: Charlene McEvoy, MD, MPH; Joseph Tashjian, MD
Johns Hopkins University, Baltimore, MD: Robert Wise, MD; Robert Brown, MD; Nadia N. Hansel, MD, MPH; Karen Horton, MD; Allison Lambert, MD, MHS; Nirupama Putcha, MD, MHS
Los Angeles Biomedical Research Institute at Harbor UCLA Medical Center, Torrance, CA: Richard Casaburi, PhD, MD; Alessandra Adami, PhD; Matthew Budoff, MD; Hans Fischer, MD; Janos Porszasz, MD, PhD; Harry Rossiter, PhD; William Stringer, MD
Michael E. DeBakey VAMC, Houston, TX: Amir Sharafkhaneh, MD, PhD; Charlie Lan, DO Minneapolis VA: Christine Wendt, MD; Brian Bell, MD

Morehouse School of Medicine, Atlanta, GA: Marilyn G. Foreman, MD, MS; Eugene Berkowitz, MD, PhD; Gloria Westney, MD, MS.
National Jewish Health, Denver, CO: Russell Bowler, MD, PhD; David A. Lynch, MB
Reliant Medical Group, Worcester, MA: Richard Rosiello, MD; David Pace, MD
Temple University, Philadelphia, PA: Gerard Criner, MD; David Ciccolella, MD; Francis Cordova, MD; Chandra Dass, MD; Gilbert D'Alonzo, DO; Parag Desai, MD; Michael Jacobs, PharmD; Steven Kelsen, MD, PhD; Victor Kim, MD; A. James Mamary, MD; Nathaniel Marchetti, DO; Aditi Satti, MD; Kartik Shenoy, MD; Robert M. Steiner, MD; Alex Swift, MD; Irene Swift, MD; Maria Elena Vega-Sanchez, MD
University of Alabama, Birmingham, AL: Mark Dransfield, MD; William Bailey, MD; Surya Bhatt, MD; Anand Iyer, MD; Hrudaya Nath, MD; J. Michael Wells, MD
University of California, San Diego, CA: Joe Ramsdell, MD; Paul Friedman, MD; Xavier Soler, MD, PhD; Andrew Yen, MD
University of Iowa, Iowa City, IA: Alejandro P. Comellas, MD; Karin F. Hoth, PhD; John Newell, Jr., MD; Brad Thompson, MD
University of Michigan, Ann Arbor, MI: MeiLan K. Han, MD, MS; Ella Kazerooni, MD; Carlos H. Martinez, MD, MPH
University of Minnesota, Minneapolis, MN: Joanne Billings, MD; Abbie Begnaud, MD; Tadashi Allen, MD
University of Pittsburgh, Pittsburgh, PA: Frank Sciurba, MD; Jessica Bon, MD; Divay Chandra, MD, MSc; Carl Fuhrman, MD; Joel Weissfeld, MD, MPH
University of Texas Health Science Center at San Antonio, San Antonio, TX: Antonio Anzueto, MD; Sandra Adams, MD; Diego Maselli-Caceres, MD; Mario E. Ruiz, MD

3.1.4 Description for the 1,320 Japanese individuals (JPN)

Genomic DNA of 295 Japanese individuals were obtained from Epstein-Barr virus transformed B-lymphoblast cell lines of unrelated Japanese individuals established by the Japan Biological Informatics Consortium (JBIC, `http://www.jbic.or.jp/english`). WGS analysis was conducted as described elsewhere [4]. Among the 295 individuals, 21.6% are female, and all participants are 16-84 years old at enrollment (mean = 51.3). Briefly, WGS library was constructed using the TruSeq DNA PCR-Free Library Preparation Kit (Illumina) according to the manufacturer's

protocols. All individuals were sequenced using 2×150-bp paired end reads on a HiSeq X Five (Illumina). The sequence reads were converted to the FASTQ format using `bcl2fastq2` (version 2.17.1.14) and trimmed to clip Illumina adapters using `Trimmomatic` (version 0.36). They were aligned to the reference human genome with the decoy sequence (GRCh37/hg19, human_g1k_v37_decoy) using `BWA-MEM` (version 0.7.5a).

A separate 1,025 Japanese individuals were enrolled from BioBank Japan Project (BBJ), as described elsewhere [5]. They were affected with any of the five diseases (acute myocardial infarction, drug eruption, colorectal cancer, breast cancer, prostate cancer). WGS data of the individuals were obtained from the National Bioscience Database Center (NBDC) Human Database (`https://humandbs.biosciencedbc.jp/en/`, ID: hum0014) and Japanese Genotype-phenotype Archive (JGA; `https://www.ddbj.nig.ac.jp/jga/index.html`, ID: JGAS00000000114). 41.4% of these individuals are female. WGS analysis was conducted as described elsewhere [5]. Briefly, WGS library was constructed using the TruSeq Nano DNA Library Preparation Kit (Illumina) according to the manufacturer's protocols. All individuals were sequenced using 2×160-bp paired end reads on a HiSeq 2500 (Illumina). The sequence reads were converted to the FASTQ format using `bcl2fastq2` (version 2.17.1.14) and trimmed to clip Illumina adapters using `Trimmomatic` (version 0.36). They were aligned to the reference human genome with the decoy sequence (GRCh37/hg19, human_g1k_v37_decoy) using `BWA-MEM` (version 0.7.5a).

3.1.5 Description of the 2,244 Estonian individuals (EST)

The 2,244 Estonian individuals were enrolled from the Estonian Biobank of the Estonian Genome Center. All samples followed a PCR-free sample preparation. Libraries sequenced on the Illumina HiSeq X Ten at 30x coverage. The details of sample selection and processing were described previously[6].

3.1.6 1000 Genomes Project (1KG)

All deep-coverage whole genome sequencing data of the 1000 Genomes Project as described in [7] were downloaded from `ftp://ftp.1000genomes.ebi.ac.uk/vol1/`

ftp/data_collections/1000G_2504_high_coverage. A subset of 1,267 individuals covers 14 populations and four major continental groups underwent Sequence-based typing (SBT) for typing HLA three class I genes (*HLA-A*, *HLA-B* and *HLA-C*) and two class II genes (*HLA-DRB1* and *HLA-DQB1*) at G-group resolution[8]. The SBT HLA genotype was downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20140725_hla_genotypes/20140702_hla_diversity.txt.

3.2 Construction of a multi-ancestry HLA reference panel

3.2.1 Read mapping

Whole genome-sequencing reads alignment was performed by each cohort. Due to changes in the informatics pipeline over the course of different projects, two different versions of human reference were used:

- The primary GRCh37 primary assembly with human herpesvirus and the concatenated decoy sequences (GRCh37/hg19, human_g1k_v37_decoy). Cohorts mapped to this version of reference includes the JHS, EST, and JPN.
- The primary GRCh38 assembly. Cohorts mapped to this version of reference includes the MESA, the COPDGene study and 1GK.

Genomic coordinate of SNPs reported in GRCh38 was lifted down using *liftOver* software[9] to GRCh37.

3.2.2 Inference of HLA classical alleles from whole-genome sequencing

To performed HLA typing using whole-genome sequencing data, we first used samtools (version 1.3) to extract the extend MHC region (chr6:25,000,000-35,000,000) and all unmapped reads from aligned reads:

```
samtools view -h $inputfile.bam  6:25000000-35000000 \
-O bam -o $inputfile.MHC.bam
samtools view -f0x4 -h $inputfile.bam -O bam -o $inputfile.MHC_unmapped.bam
```

We then merged the extended MHC region and all unmapped reads with:

```
samtools merge -h $sample.MHC.bam $sample\_merged.bam \
$sample.MHC.bam $sample.unmapped.bam
samtools sort $sample\_merged.bam -O b -o $sample.bam
samtools index $sample.bam
```

We used HLA*PRG [10] to perform HLA typing in six classical HLA genes (HLA-*A*, *-B*, *-C*, *-DQA1*, *-DQB1*, *-DRB1*) in the earlier collection of the two cohorts (JHS and EST) when HLA*LA [11] was not available. We then performed HLA typing in eight classical HLA genes using HLA*LA (https://github.com/DiltheyLab/HLA-LA/blob/master/) for all samples included in the JPN, COPDGene, MESA, 1000G and the 1000 Genomes Project (HLA-*A*, *-B*, *-C*, *-DQA1*, *-DQB1*, *-DRB1*, *-DPA1*, *-DPB1*).

```
    perl5.8.9 HLA-PRG-LA.pl \
  --BAM $sample.bam \
  --graph PRG_MHC_GRCh38_withIMGT \
  --sampleID $sample \
  --maxThreads $n
```

Briefly, both HLA*PRG and HLA*LA constructs a directed graph in which alternative alleles, insertions and deletions are represented as alternative paths using whole-genome sequencing data.

3.2.3 Variant calling

To construct a HLA imputation we used variants (SNPs and Indels) called within the extend MHC region (chr6:25,000,000-35,000,000) either from vcfs provided by each individual cohort (MESA, COPDGene and 1KG) or performed using mapped reads described in Section 3.2.1 following GATK [12] (version 3.6) best practices (EST, JHS and JPN).

Briefly, we first used HaplotypeCaller to perform per-sample variant calling with the following command:

```
java -jar GenomeAnalysisTK.jar \
  -T HaplotypeCaller -R Homo_sapiens_assembly19.fasta \
  -I $sample.bam \
```

```
  --emitRefConfidence GVCF \
  -L 6:25000000-35000000 \
  -o $sample.g.vcf.gz \
  -variant_index_type LINEAR \
  -variant_index_parameter 128000
```

Then we combined multiple sample *gvcfs* using `GenotypeGVCFs`:

```
java -jar GenomeAnalysisTK.jar \
  -T GenotypeGVCFs \
  -R Homo_sapiens_assembly19.fasta \
  --variant sample.list \
  -o all.vcf.gz \
  --max_alternate_alleles 2
```

We next unified all variants called in individual cohort using `CombineVariants`.

3.2.4 Variant-level quality control

Post variant calling, we applied the following quality control (QC) criteria to each variant that had been called:

- variant must overlap with sites discovered in 1000 Genomes Phase III release [7].
- in the extended MHC region (chr6:28-34Mb)
- SNPs only
- minor allele frequency $\geq 0.5\%$ in each and combined cohort
- HWE p-value $\leq 1 \times 10^{-20}$ in each cohort
- missingness rate $\leq 10\%$
- variants within the eight classical HLA gene body

These QC criteria have been applied both on individual cohorts. We did not filter variants that are out of Hardy-Weinberg equilibrium due to extreme high linkage disequilibrium in this region[13]. In total, there are 38,398 SNPs included in the reference panel.

3.2.5 Sample-level quanlity control

Since individual cohort were sequenced at different depth, to ensure quality of inferred HLA types, we first restrict samples that have more than 20x in each of the eight HLA genes. The average coverage in the MHC region is summarized in **Supplementary Table 1**. Next we excluded samples that failed genome-wide QC. In total, there are 21,546 individuals included in the reference panel.

References

[1] McLaren, P. J. *et al.* Fine-mapping classical HLA variation associated with durable host control of HIV-1 infection in african americans. *Hum. Mol. Genet.* **21**, 4334–4347 (2012).

[2] McLaren, P. J. *et al.* Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 14658–14663 (2015).

[3] Regan, E. A. *et al.* Genetic epidemiology of COPD (COPDGene) study design. *COPD* **7**, 32–43 (2010).

[4] Okada, Y. *et al.* Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of japanese. *Nat. Commun.* **9**, 1631 (2018).

[5] Hirata, J. *et al.* Genetic and phenotypic landscape of the major histocompatibilty complex region in the japanese population. *Nat. Genet.* (2019).

[6] Mitt, M. *et al.* Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur. J. Hum. Genet.* **25**, 869–876 (2017).

[7] 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

[8] Gourraud, P.-A. *et al.* HLA diversity in the 1000 genomes dataset. *PLoS One* **9**, e97282 (2014).

[9] Hinrichs, A. S. *et al.* The UCSC genome browser database: update 2006. *Nucleic Acids Res.* **34**, D590–8 (2006).

[10] Dilthey, A. T. *et al.* High-Accuracy HLA type inference from Whole-Genome sequencing data using population reference graphs. *PLoS Comput. Biol.* **12**, e1005151 (2016).

[11] Dilthey, A. T. *et al.* HLA*LA-HLA typing from linearly projected graph alignments. *Bioinformatics* **35**, 4394–4396 (2019).

[12] McKenna, A. *et al.* The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

[13] Graffelman, J., Jain, D. & Weir, B. A genome-wide study of Hardy-Weinberg equilibrium with next generation sequence data. *Hum. Genet.* **136**, 727–741 (2017).