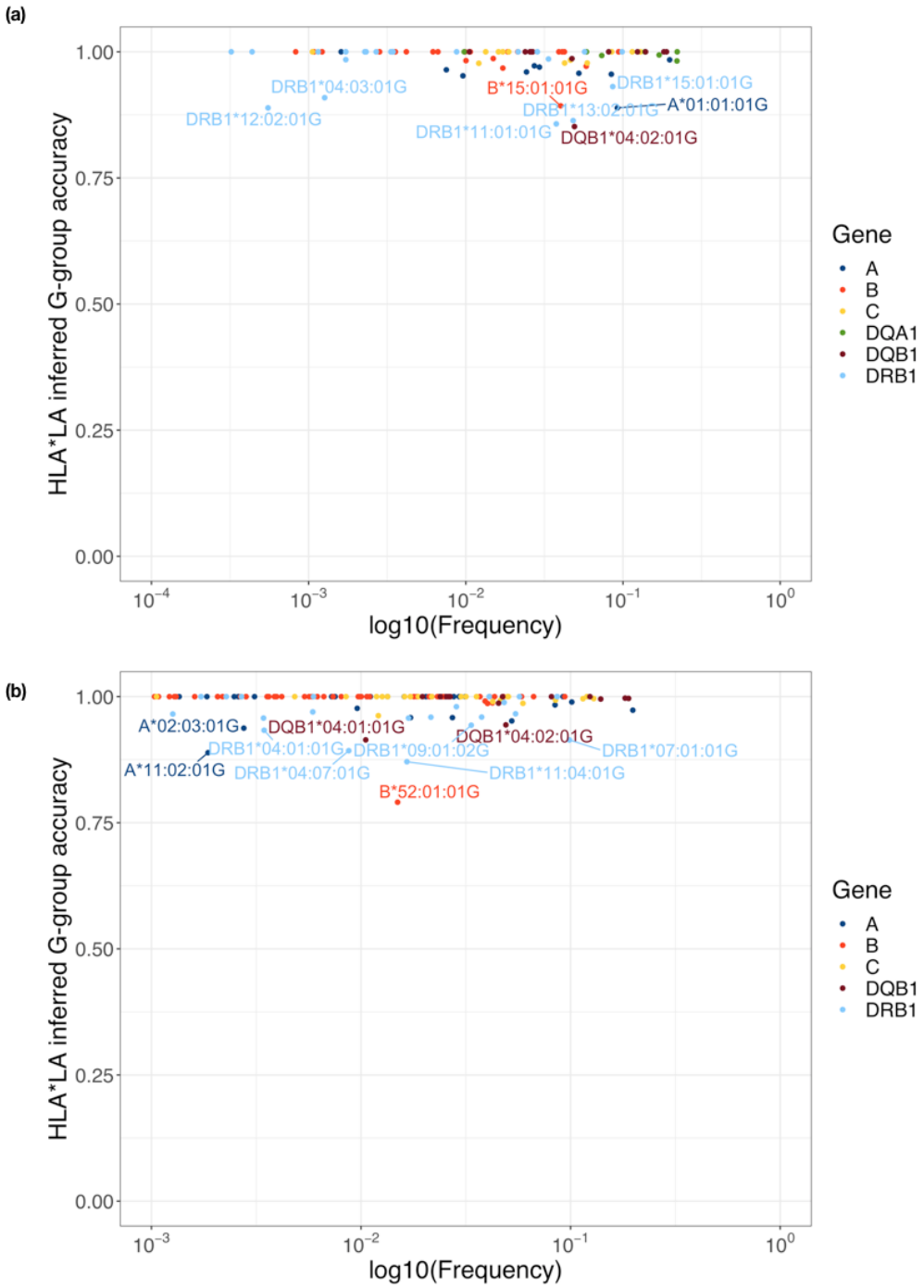
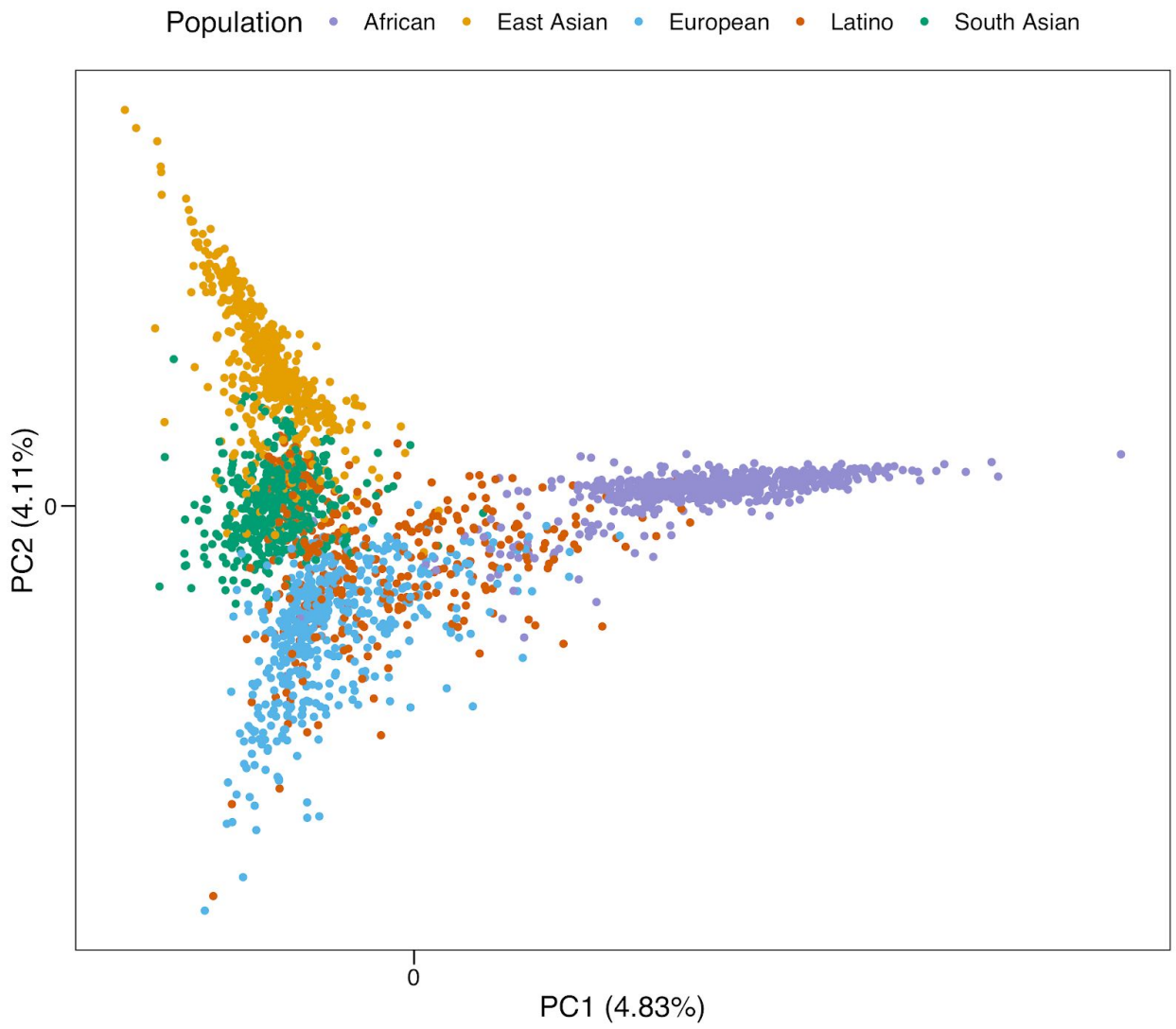
Supplementary Figure 1. HLA*LA typing accuracy of all observed G-group alleles against gold-standard sequencing based typing. The x-axis is the log10(allele frequency) observed in all 21,546 subjects. The y-axis is the accuracy of HLA*LA inferred G-group alleles compared against the sequencing based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

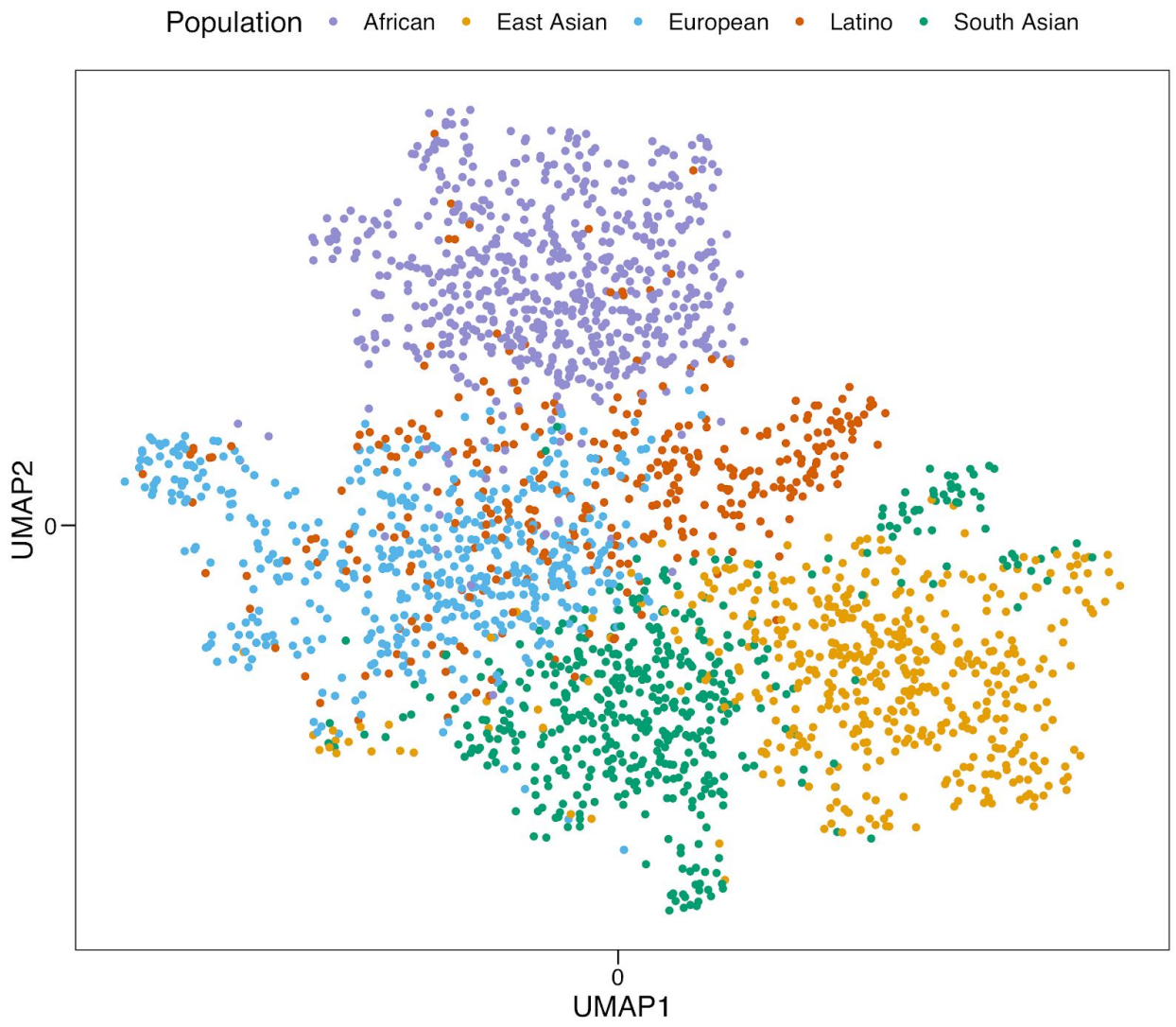


Supplementary Figure 2. Dimension reduction of the pairwise IBD distance between 1000 Genomes samples using MHC region markers only with population labels. (**a**) The first two principal components show separation of continental groups (European, African, and East Asian) as well as the admixed nature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

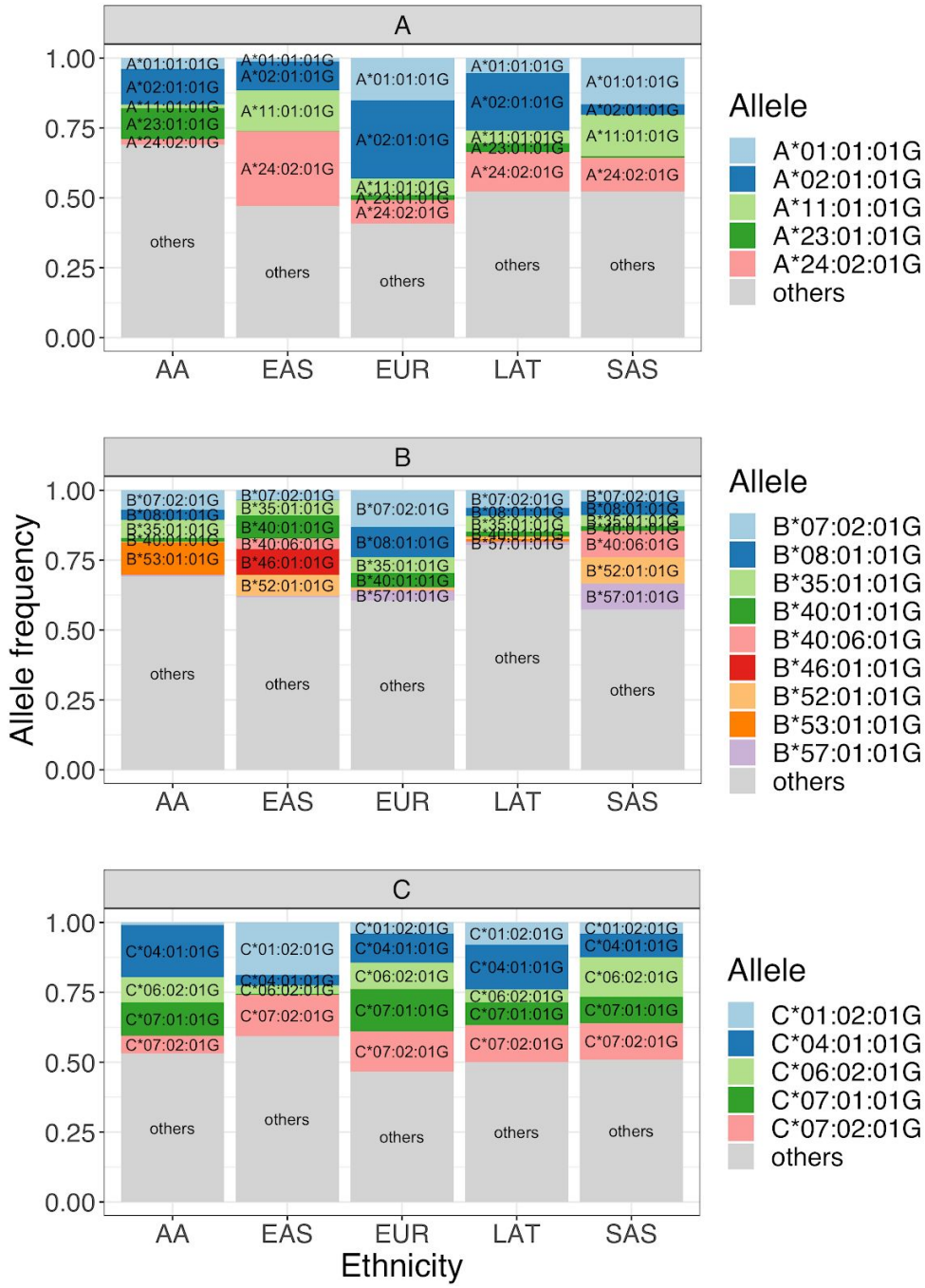
(a)

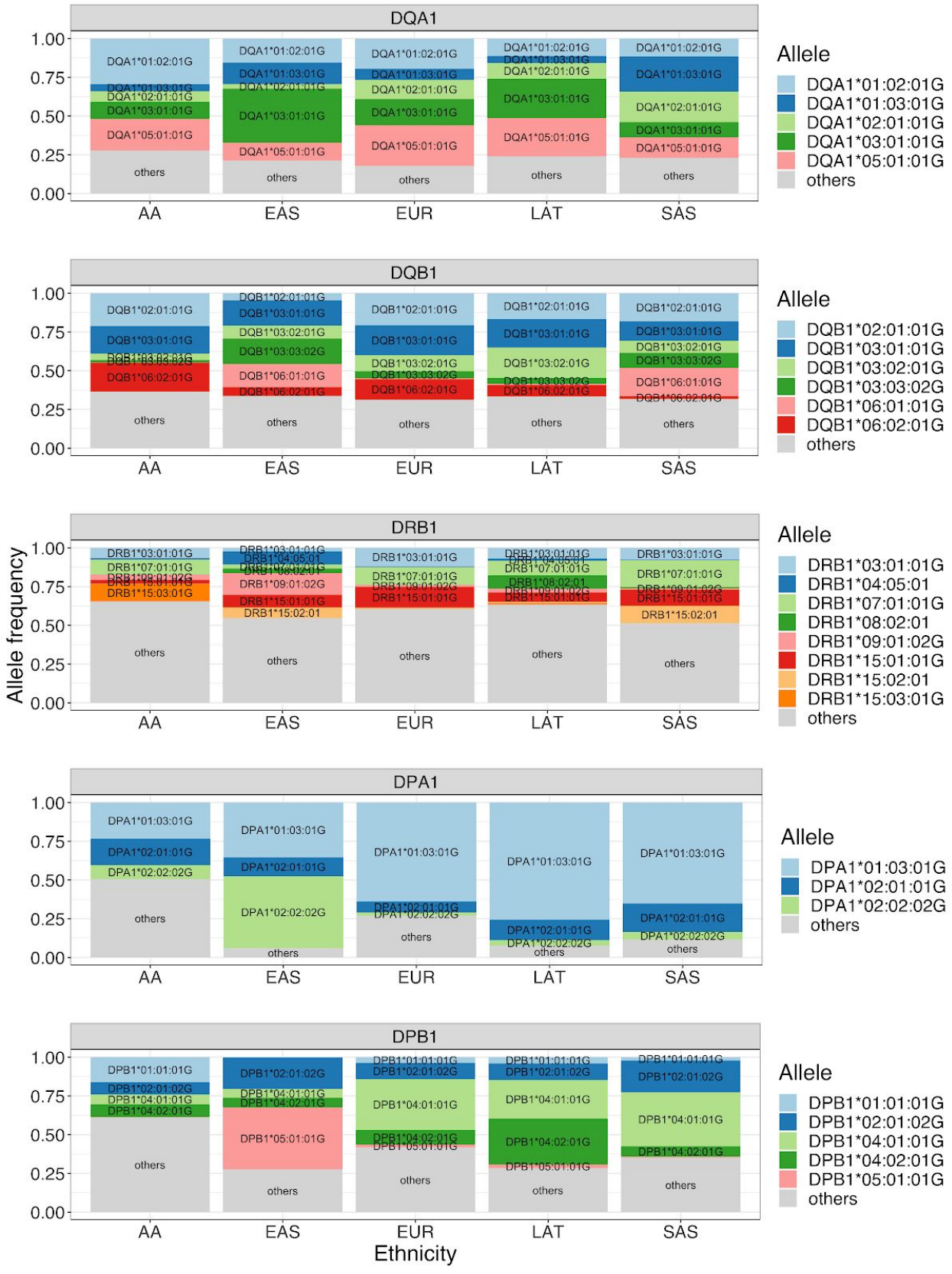


(b)


Population
African
East Asian
European
Latino
South Asian
UMAP2
0
0
UMAP1

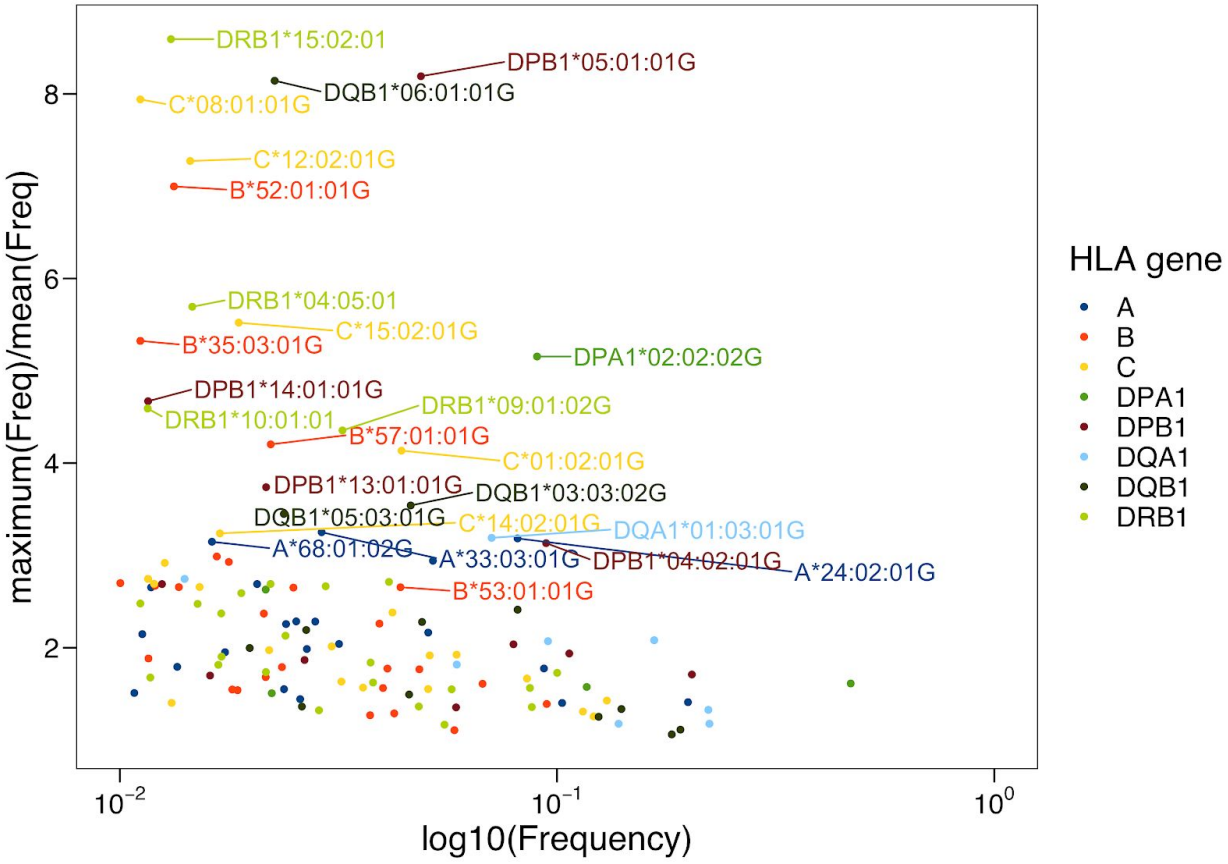
Supplementary Figure 3. Diversity of eight classical HLA genes in the constructed multi-ethnic MHC reference panel. Each gene is stratified by six populations (AA = Admixed African; EAS = East Asian; EUR = European; LAT = Latino; SAS = South Asian). The top two most common alleles within each classical gene of each population are plotted across all panels. Alleles that have frequencies greater than 1% are also labelled in the bar plots.



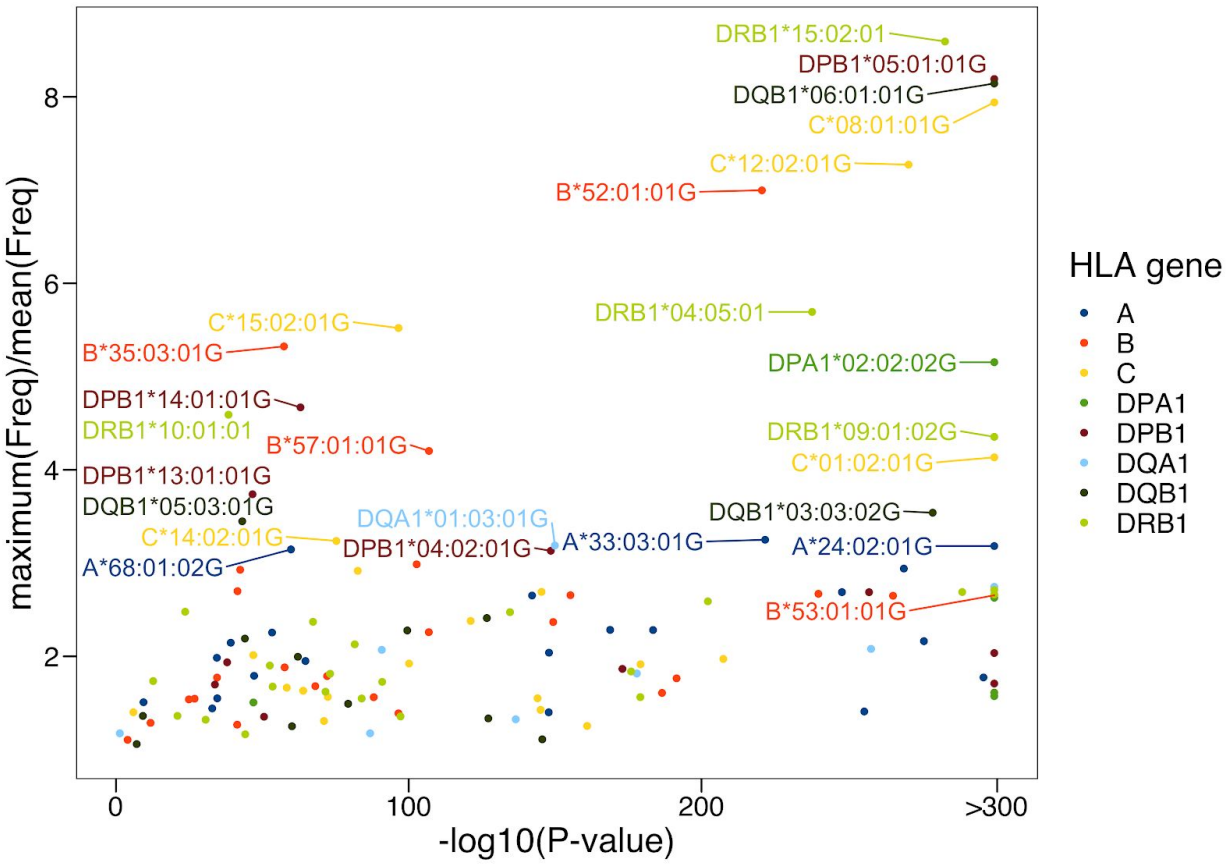

DQA1
DQB1
DRB1
DPA1
DPB1
Allele frequency
Ethnicity
AA
EAS
EUR
LAT
SAS
Allele
DQA1*01:02:01G
DQA1*01:03:01G
DQA1*02:01:01G
DQA1*03:01:01G
DQA1*05:01:01G
DQB1*02:01:01G
DQB1*03:01:01G
DQB1*03:02:01G
DQB1*03:03:02G
DQB1*06:01:01G
DQB1*06:02:01G
DRB1*03:01:01G
DRB1*04:05:01
DRB1*07:01:01G
DRB1*08:02:01
DRB1*09:01:02G
DRB1*15:01:01G
DRB1*15:02:01
DRB1*15:03:01G
DPA1*01:03:01G
DPA1*02:01:01G
DPA1*02:02:02G
DPB1*01:01:01G
DPB1*02:01:02G
DPB1*04:01:01G
DPB1*04:02:01G
DPB1*05:01:01G
others

Supplementary Figure 4. 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 obtained from a 3x5 genotype contingency table.

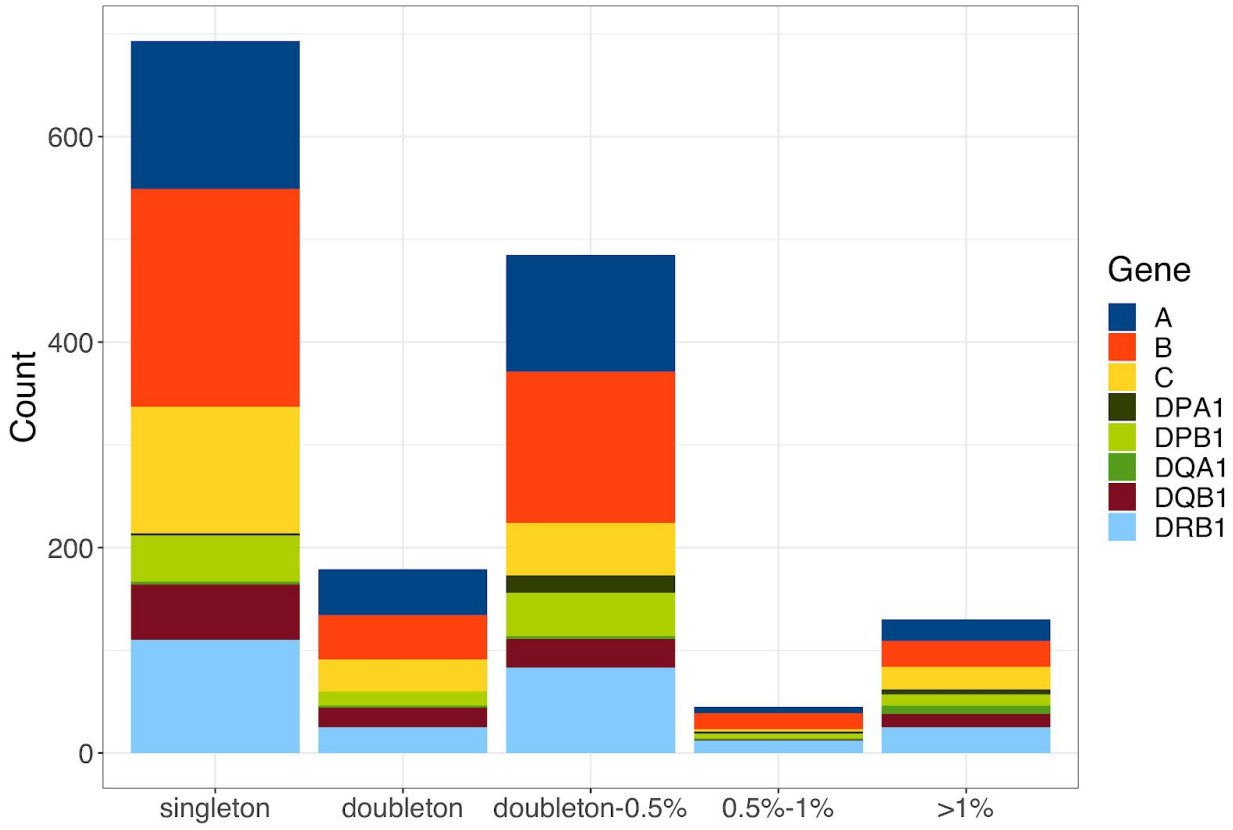
(**a**)



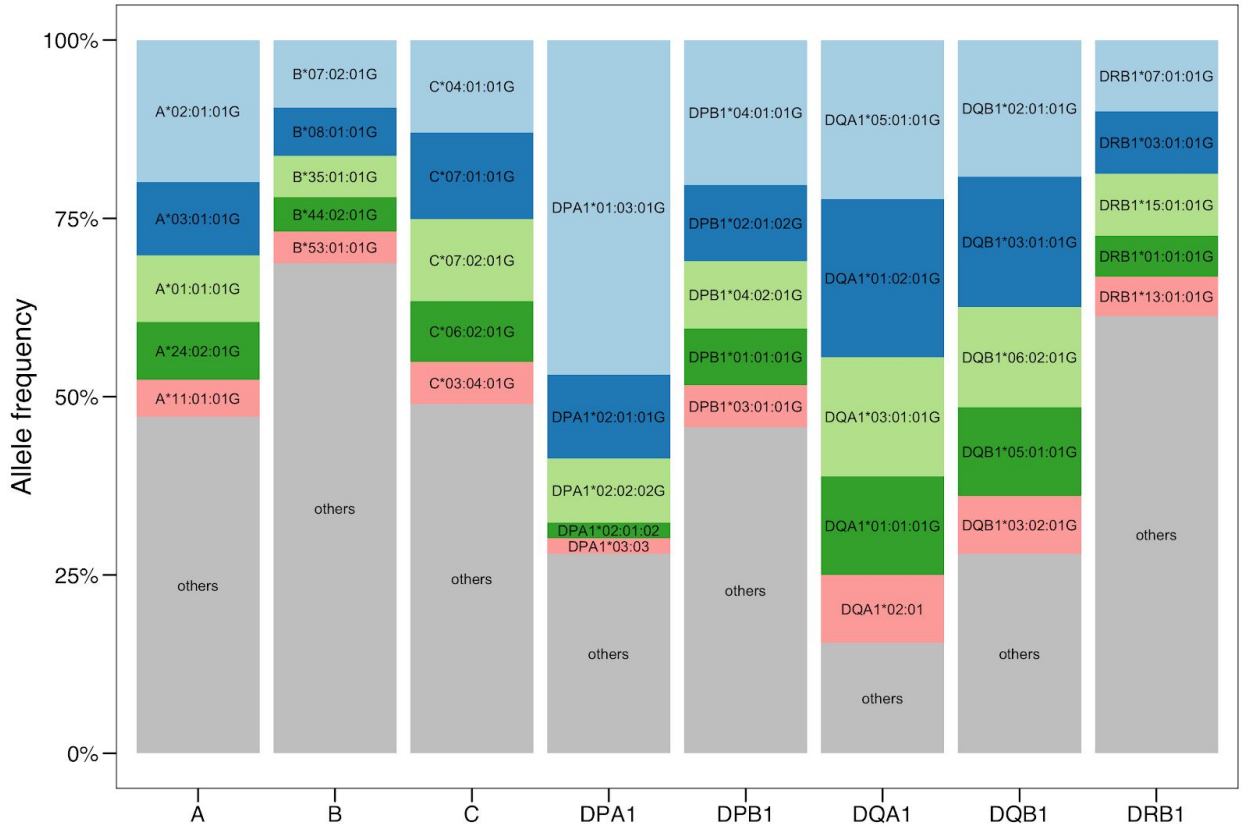
(**b**)


DRB1*15:02:01
DPB1*05:01:01G
DQB1*06:01:01G
C*08:01:01G
C*12:02:01G
B*52:01:01G
DRB1*04:05:01
C*15:02:01G
B*35:03:01G
DPA1*02:02:02G
DPB1*14:01:01G
DRB1*10:01:01
B*57:01:01G
DRB1*09:01:02G
C*01:02:01G
DPB1*13:01:01G
DQB1*05:03:01G
DQA1*01:03:01G
DQB1*03:03:02G
C*14:02:01G
DPB1*04:02:01G
A*33:03:01G
A*24:02:01G
A*68:01:02G
B*53:01:01G
maximum(Freq)/mean(Freq)
8
6
4
2
0
100
200
>300
-log10(P-value)
HLA gene
A
B
C
DPA1
DPB1
DQA1
DQB1
DRB1

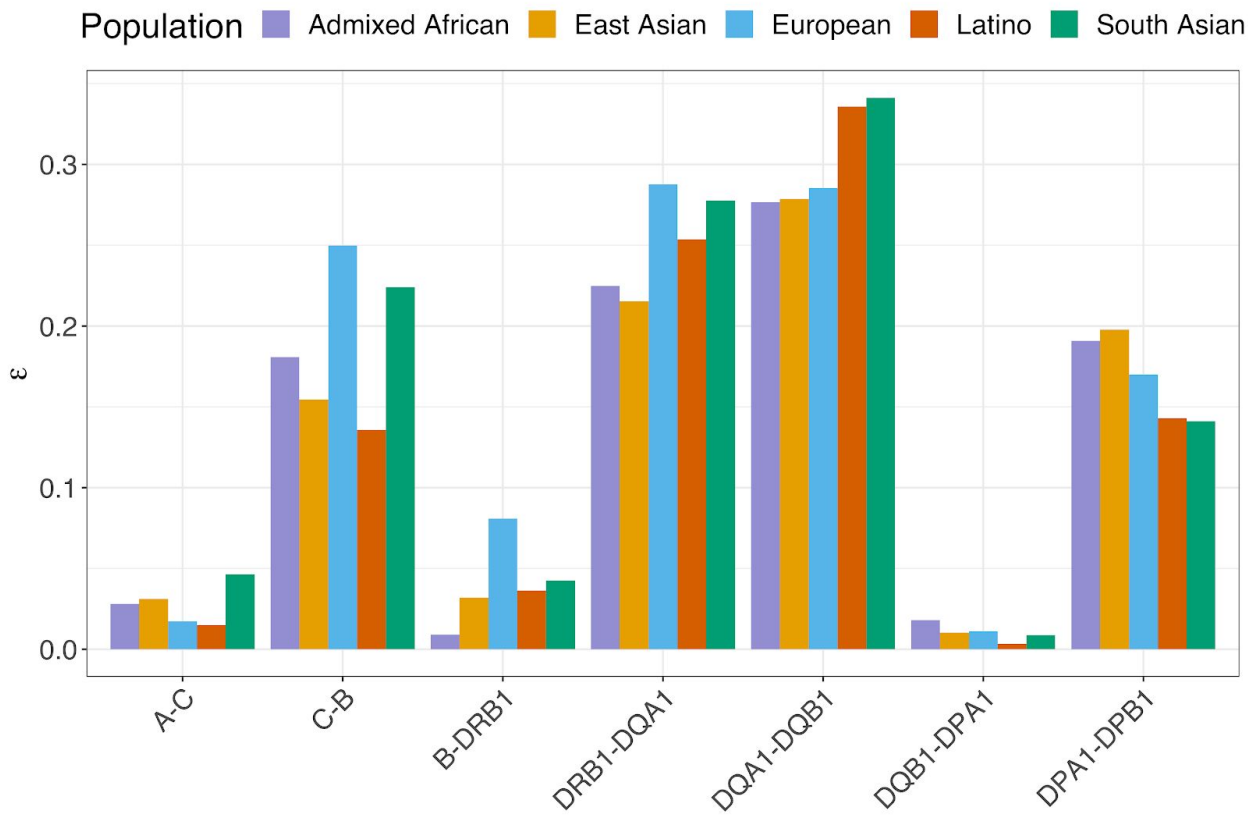
Supplementary Figure 5. Reported G-group alleles stratified by overall frequency among 21,546 individuals.



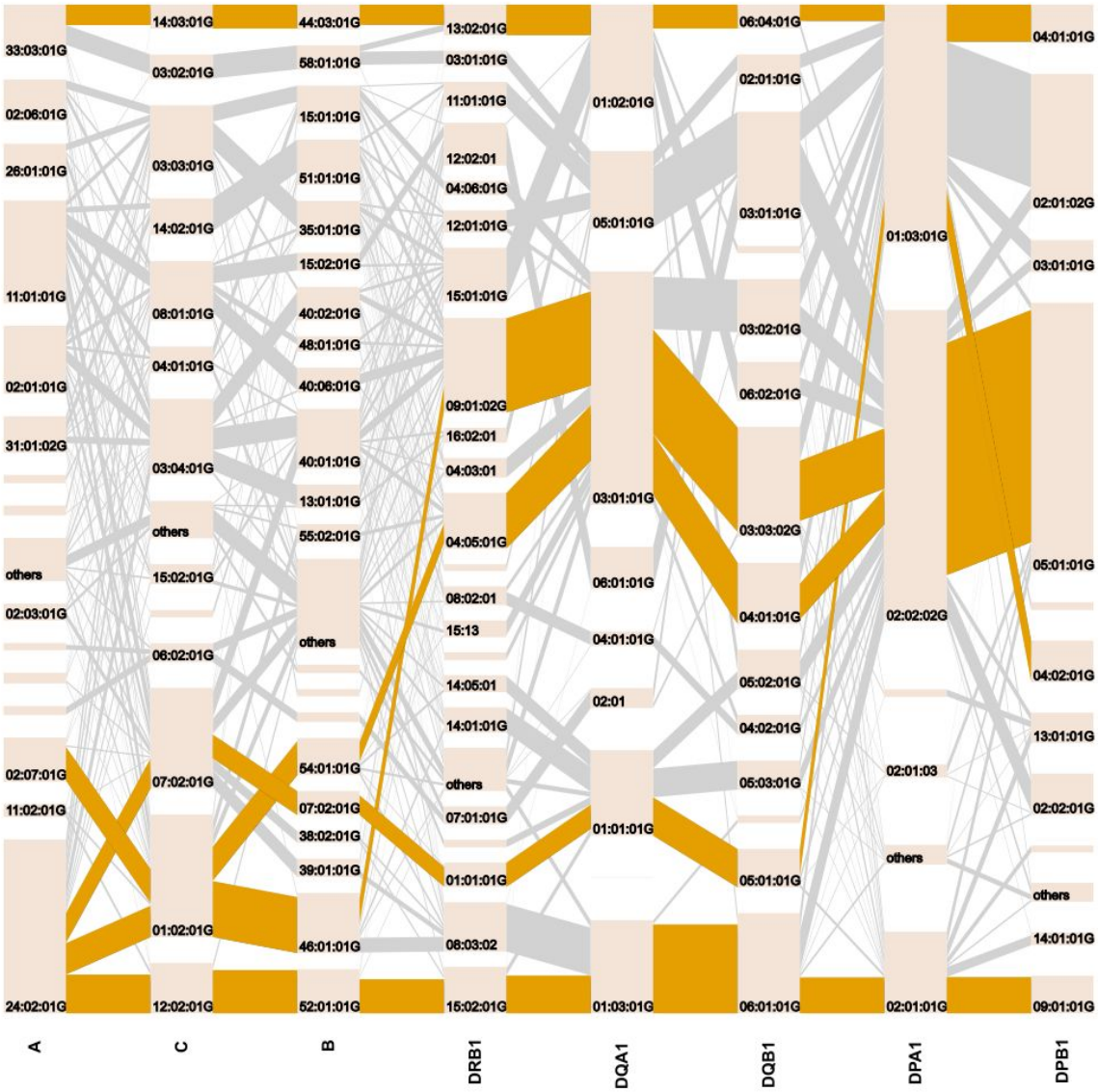
Supplementary Figure 6. Allele diversity of eight classical *HLA* genes in global populations. For each gene, the top five most frequent alleles across all populations are shown (light blue=most frequent; dark blue=second frequent; light green=third frequent; dark green=fourth frequent; red=fifth frequent; gray= all other alleles).



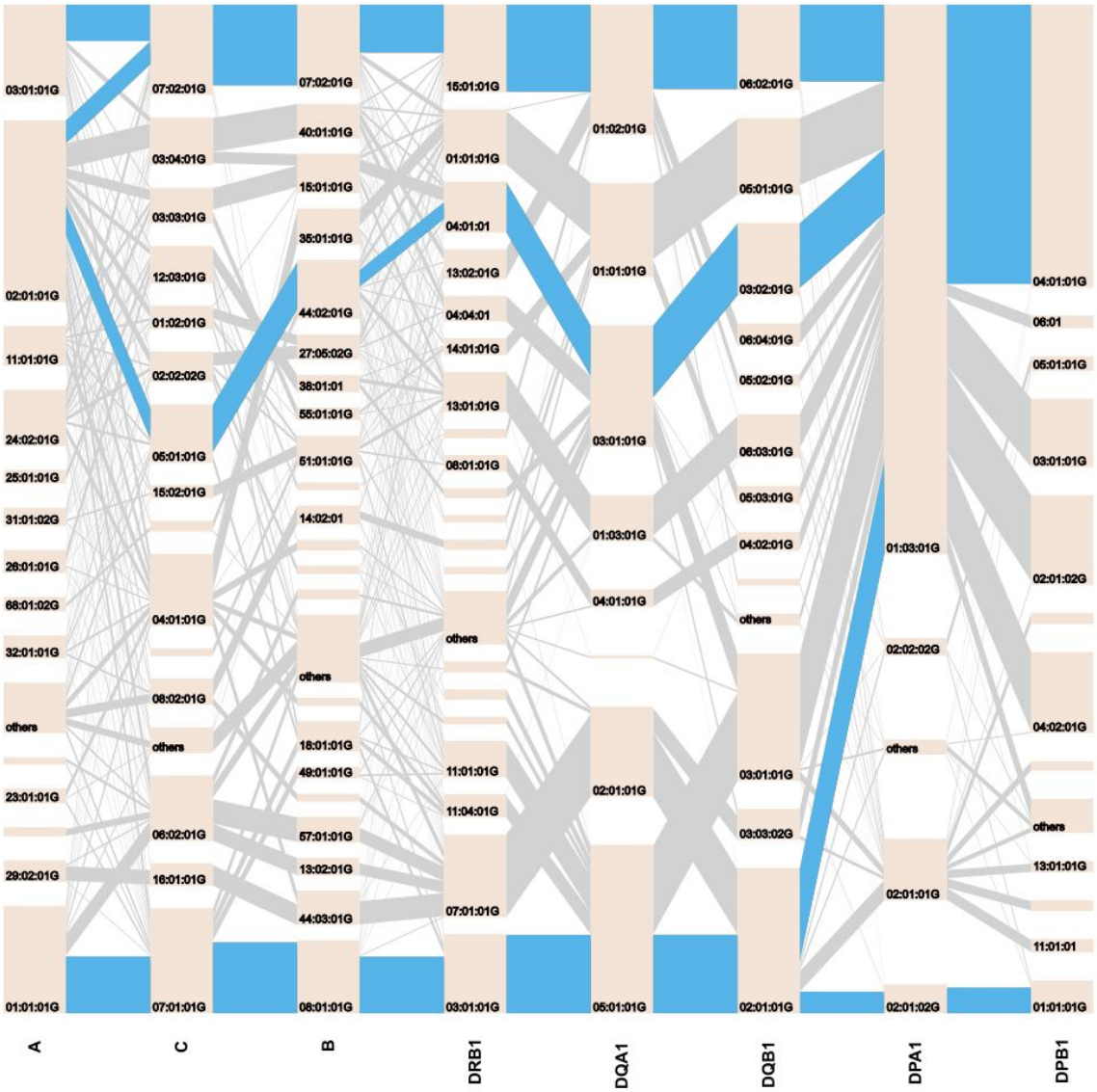
Supplementary Figure 7. Pairwise LD measurement index, ε, among five ancestral populations. The value ε evaluates the LD structures between two multiallelic variants and ranges between no LD (0) and perfect LD (1). Higher ε were observed between *DQA1*, *DQB1*, and *DRB1;* between *DPA1* and *DPB1*; and between *B* and *C*.



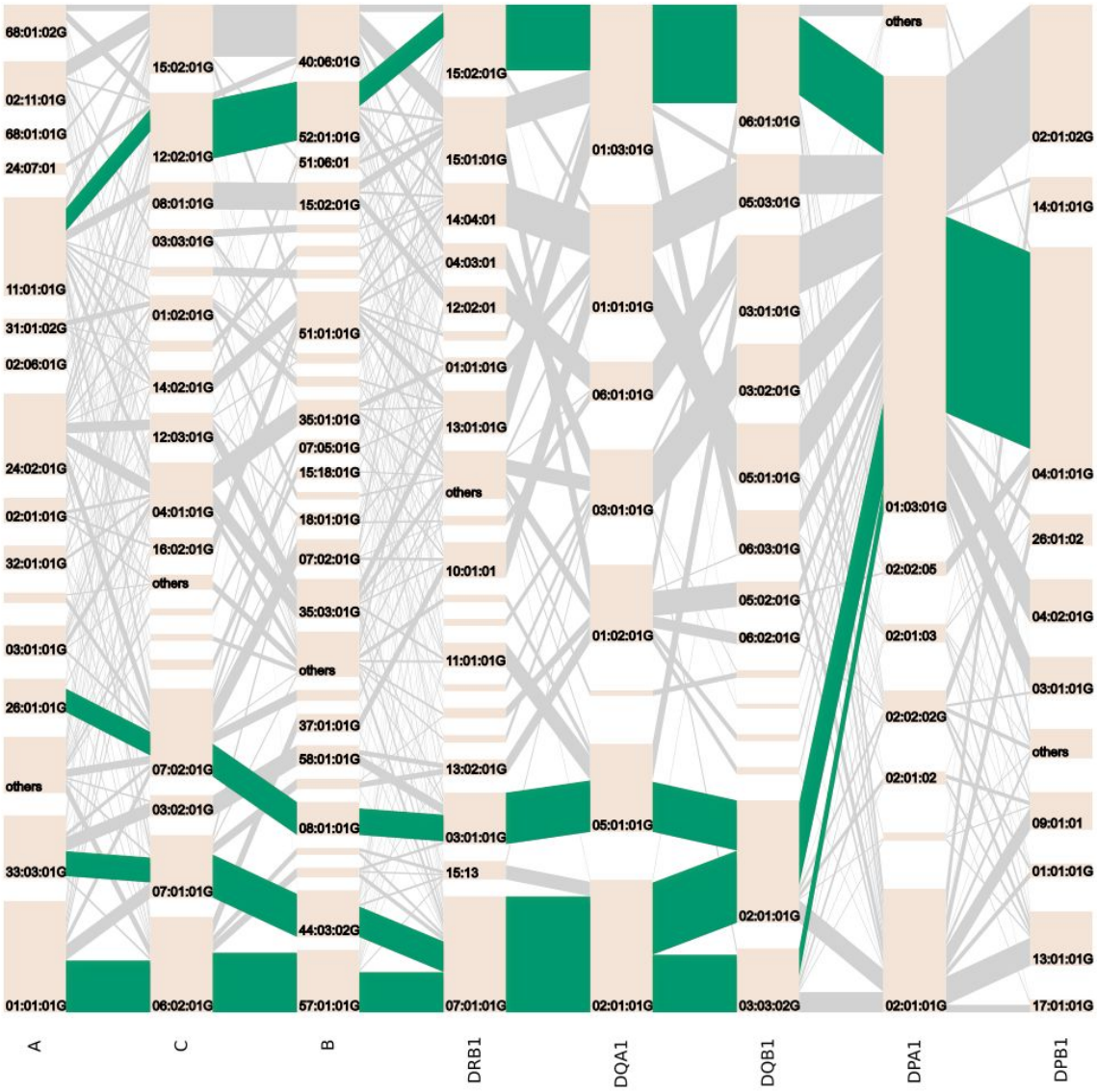
Supplementary Figure 8. Haplotype structure for eight classical HLA genes in East Asians. The haplotype structures of the eight classical HLA genes. The tile in a bar represents an *HLA* allele, and its height corresponds to the frequencies of the *HLA* allele. All long-range *HLA* haplotypes with frequency >1% are bolded and highlighted in yellow.



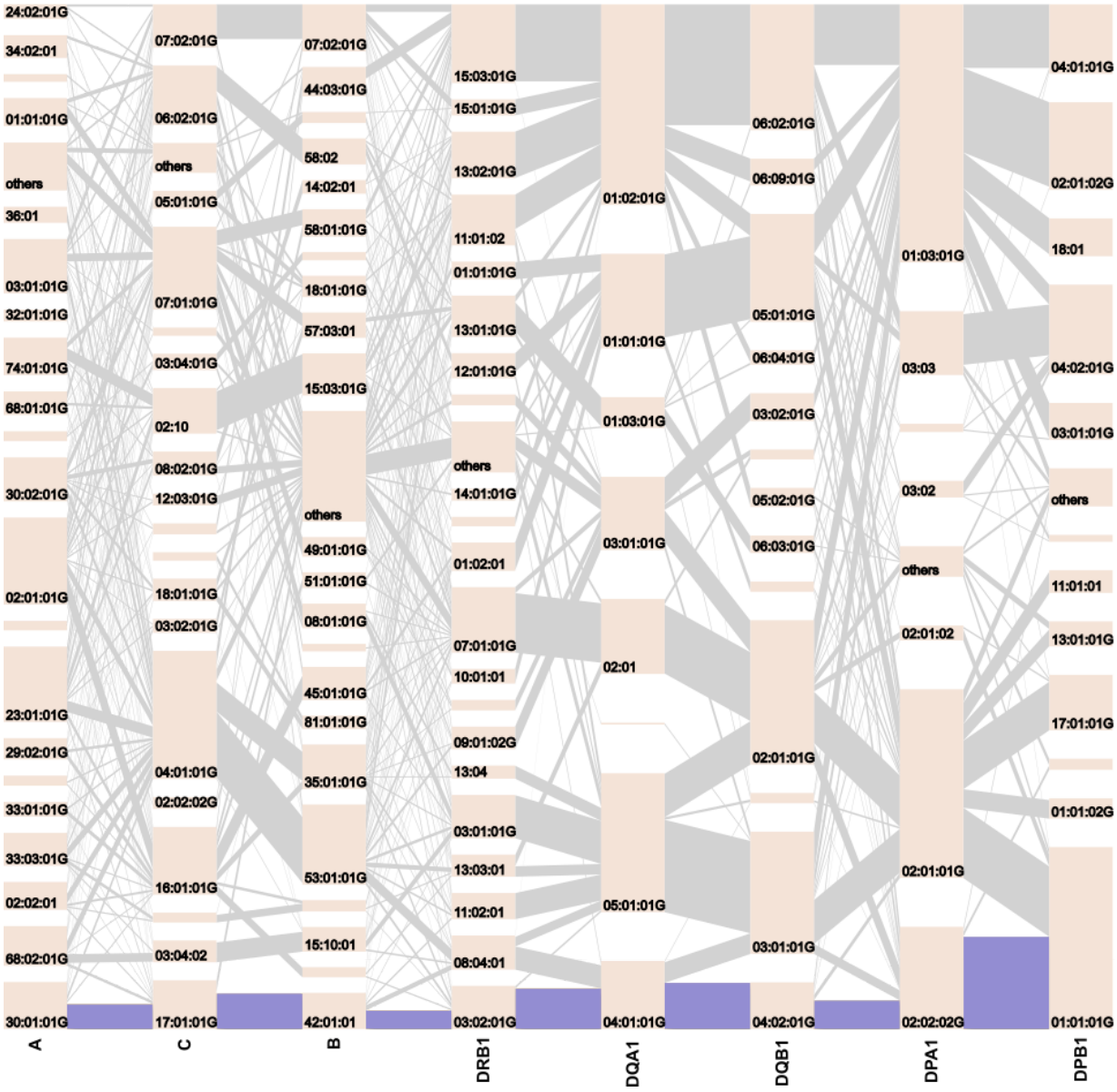
Supplementary Figure 9. Haplotype structure for eight classical HLA genes in Europeans. The haplotype structures of the eight classical HLA genes. The tile in a bar represents an *HLA* allele, and its height corresponds to the frequencies of the *HLA* allele. All long-range *HLA* haplotypes with frequency >1% are bolded and highlighted in blue.



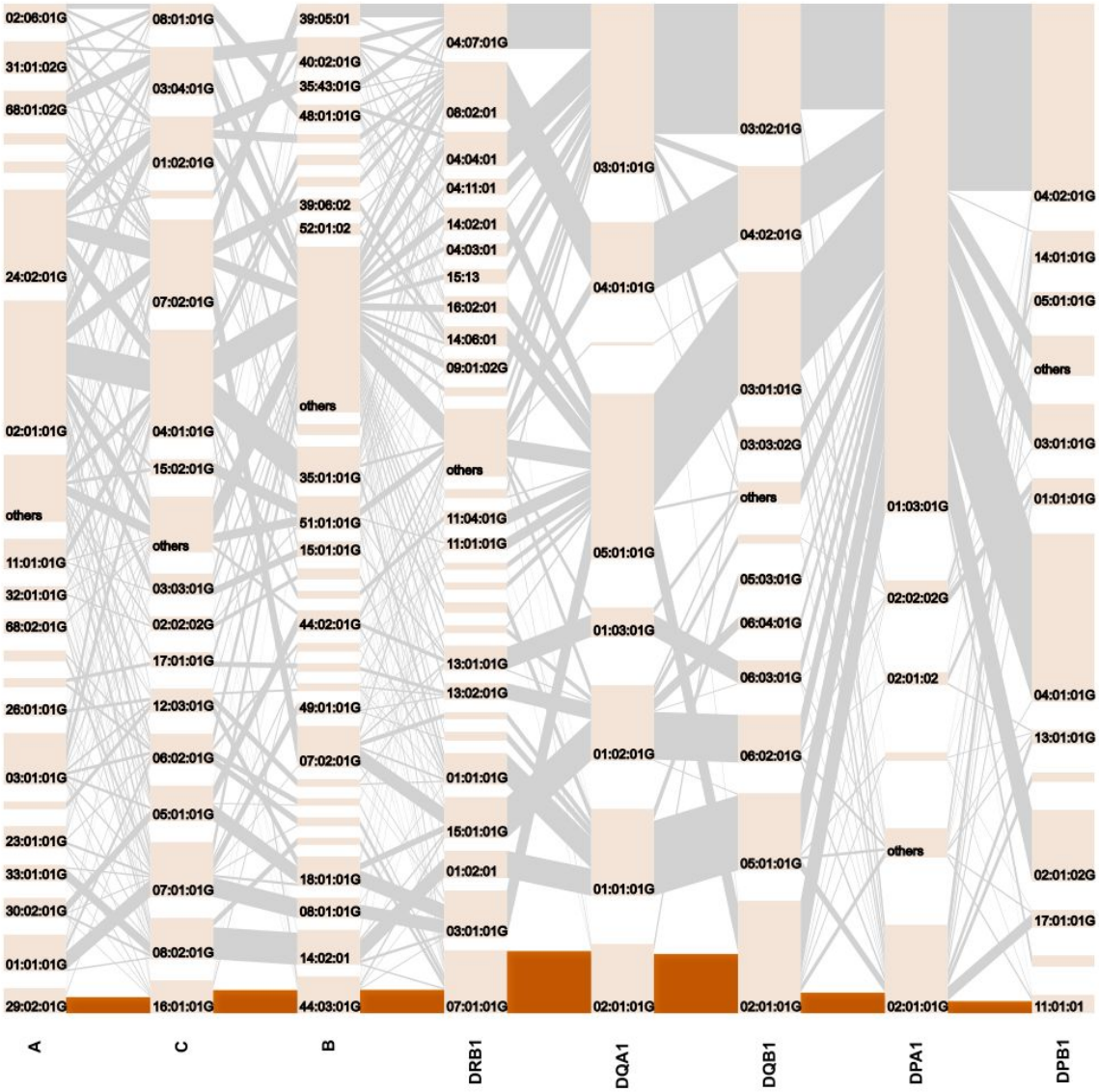
Supplementary Figure 10. Haplotype structure for eight classical HLA genes in South Asians. The haplotype structures of the eight classical HLA genes. The tile in a bar represents an *HLA* allele, and its height corresponds to the frequencies of the *HLA* allele. All long-range *HLA* haplotypes with frequency >1% are bolded and highlighted in green.



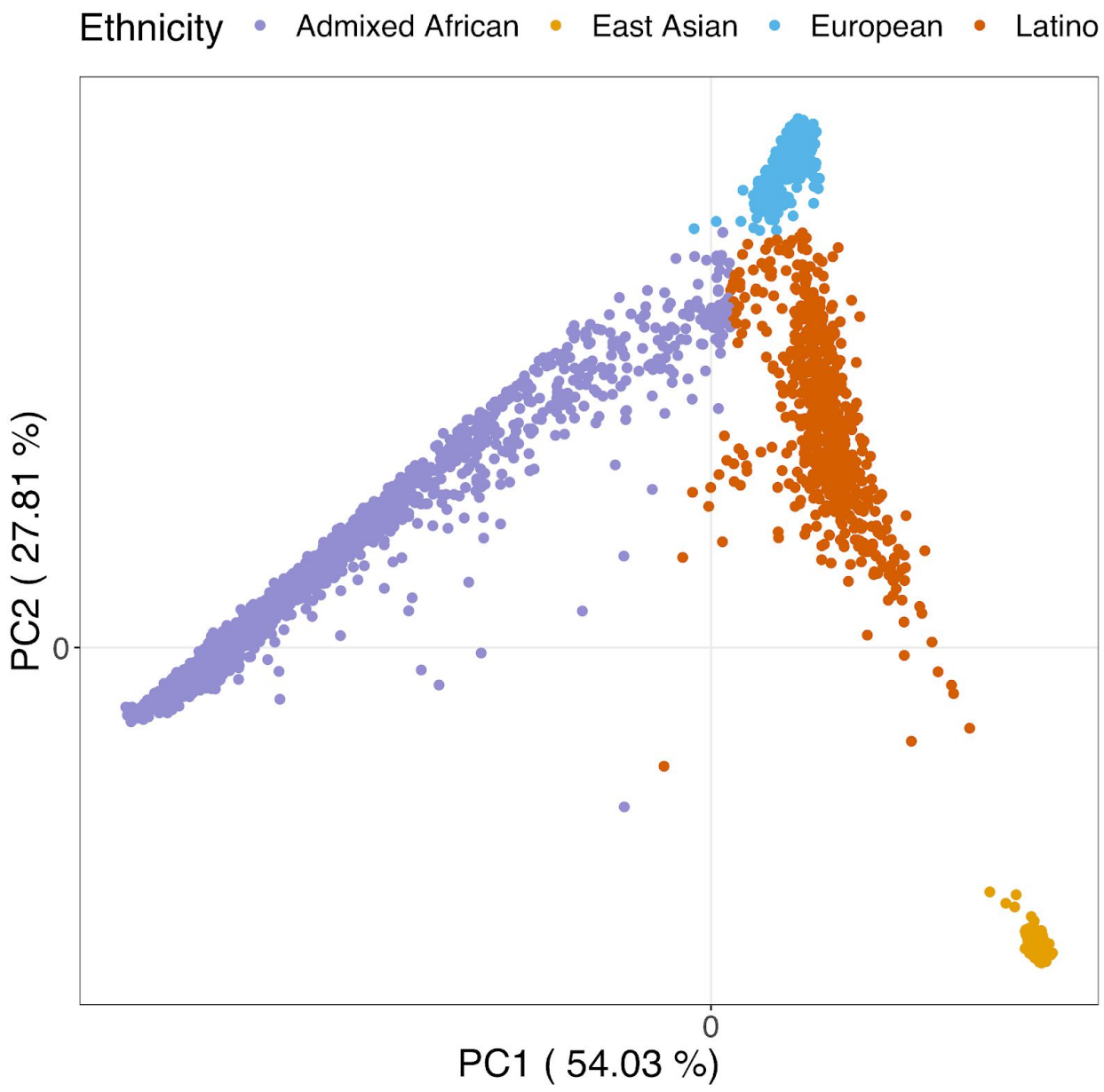
Supplementary Figure 11. Pairwise LD and haplotype structure for eight classical HLA genes in Admixed Africans. The haplotype structures of the eight classical HLA genes. The tile in a bar represents an *HLA* allele, and its height corresponds to the frequencies of the *HLA* allele. All long-range *HLA* haplotypes with frequency >1% are bolded and highlighted in purple.



Supplementary Figure 12. 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. The haplotype A29::DP11(frequency = 0.74%) is bolded and highlighted in orange.

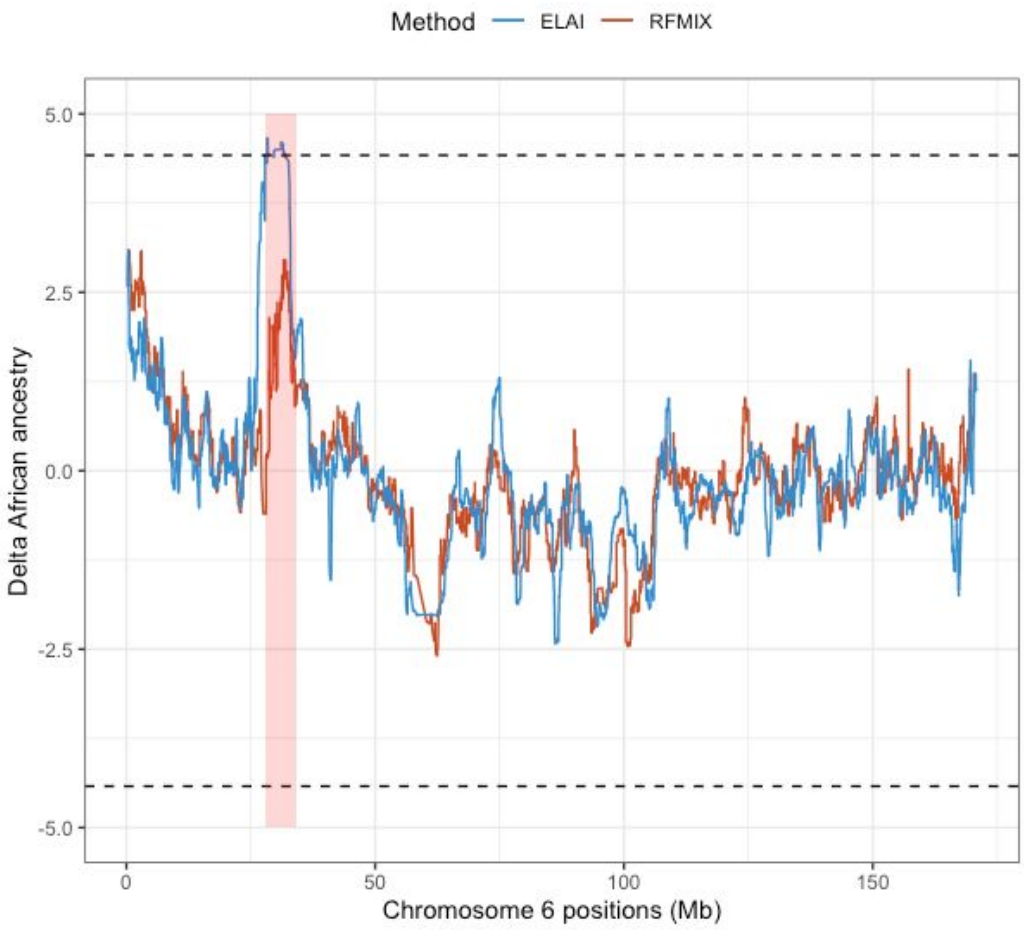


Supplementary Figure 13. Principal component analysis (PCA) of the MESA GWAS samples. The genotypes projected onto the first two principal components using SNP weights precomputed from samples in the 1000 Genomes Phase 3 project using SNPweights. Four diverse ancestries were collected in the MESA project - Admixed African (purple); East Asian (yellow); European (blue) and Latino (orange).

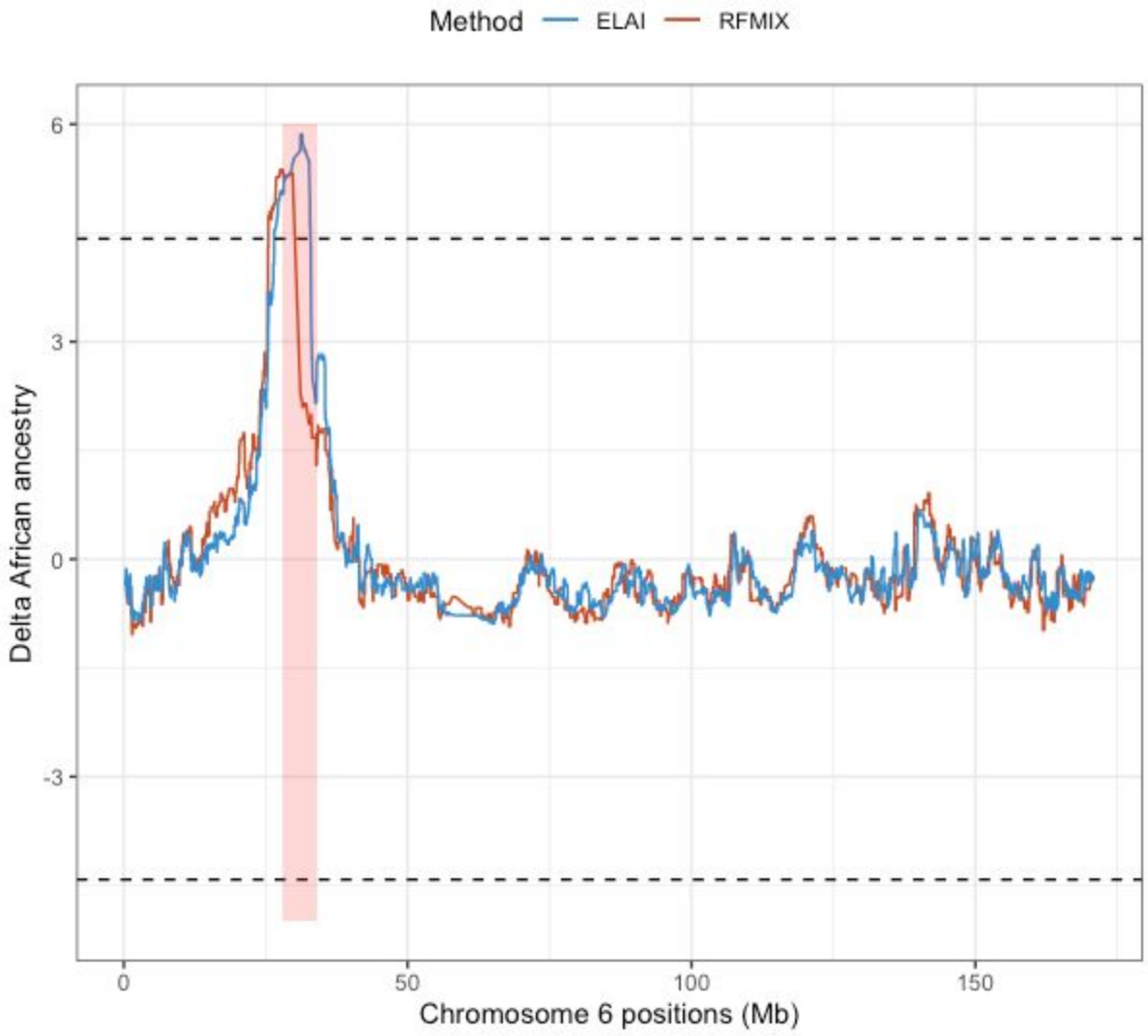


Supplementary Figure 14. Deviation from average genome-wide ancestry in Admixed African and Latino populations. The x-axis is the genomic position of Chromosome 6. The y-axis shows the local African ancestry deviation measure inferred at a given position for (**a**) Admixed Africans and (**b**) Latinos. The MHC region (chr6:28Mb-34Mb) is highlighted in red shading. Local ancestries were estimated using RFMix[1] (red) and ELAI[2] (blue). The ancestry deviation measure is the difference between African ancestry at a given genomic position with respect to the genome-wide average estimated by ADMIXTURE [3] with K=3, normalized by the standard deviation of the ancestry estimate. The dashed line indicates the genome-wide significance threshold at ± 4.42 standard deviation of the ancestry estimate deviated from the genome-wide average[4].

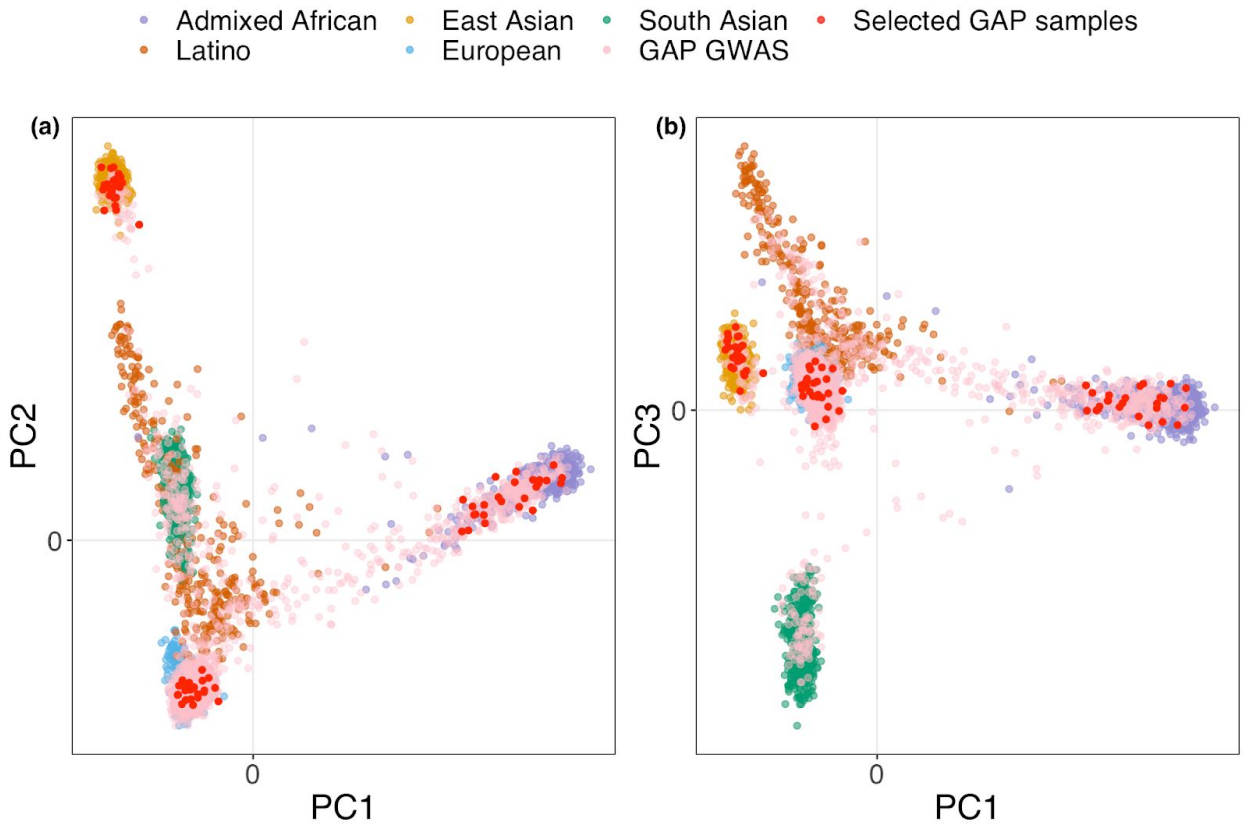
(a)



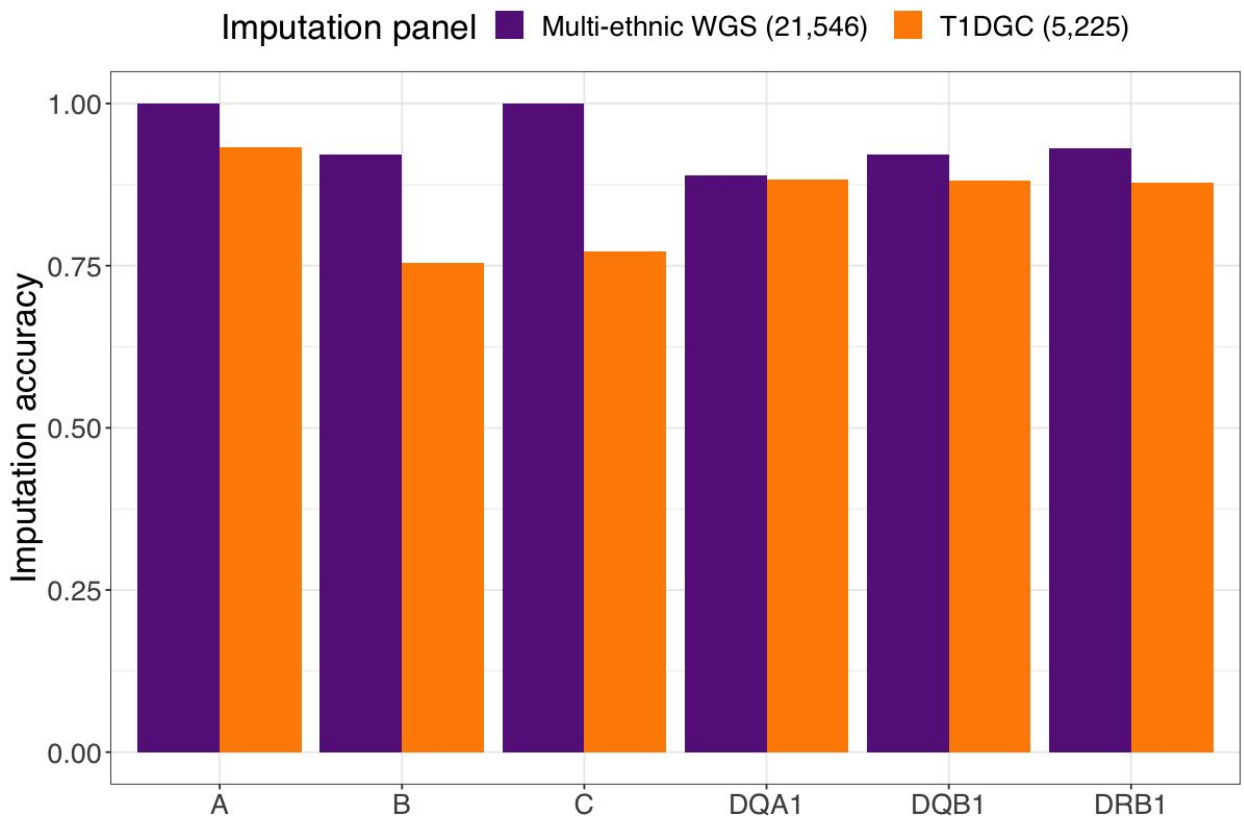
(b)


Method
ELAI
RFMIX
Delta African ancestry
6
3
0
-3
0
50
100
150
Chromosome 6 positions (Mb)

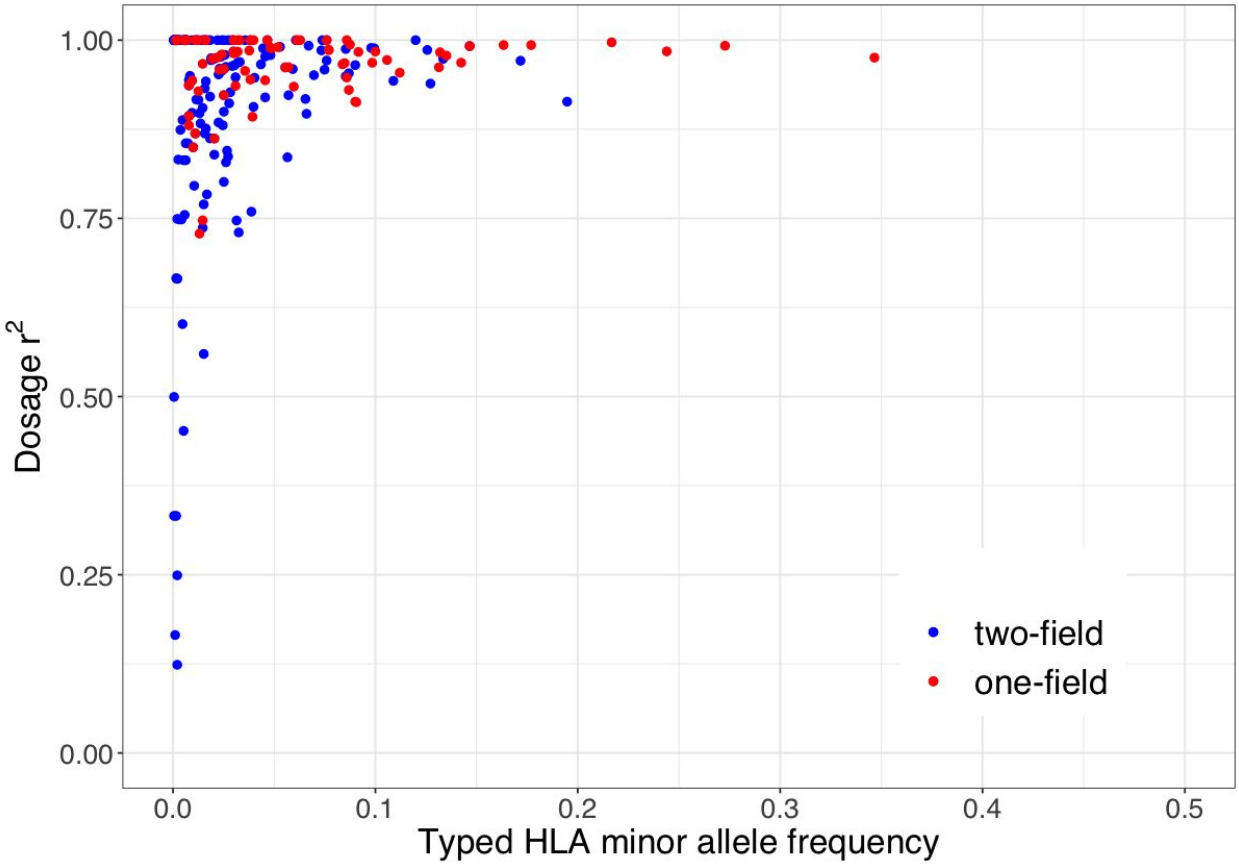
Supplementary Figure 15. Principal component analysis (PCA) of the GAP registry GWAS samples. GAP samples are plotted with five 1000 Genomes Phase 3 populations. (**a**) First and second principal components. (**b**) First and third principal components. We randomly selected 25 individuals with African ancestry (purple); 25 with East Asian ancestry (yellow) and 25 with European ancestry (blue) for undergoing gold-standard *HLA* typing (red).



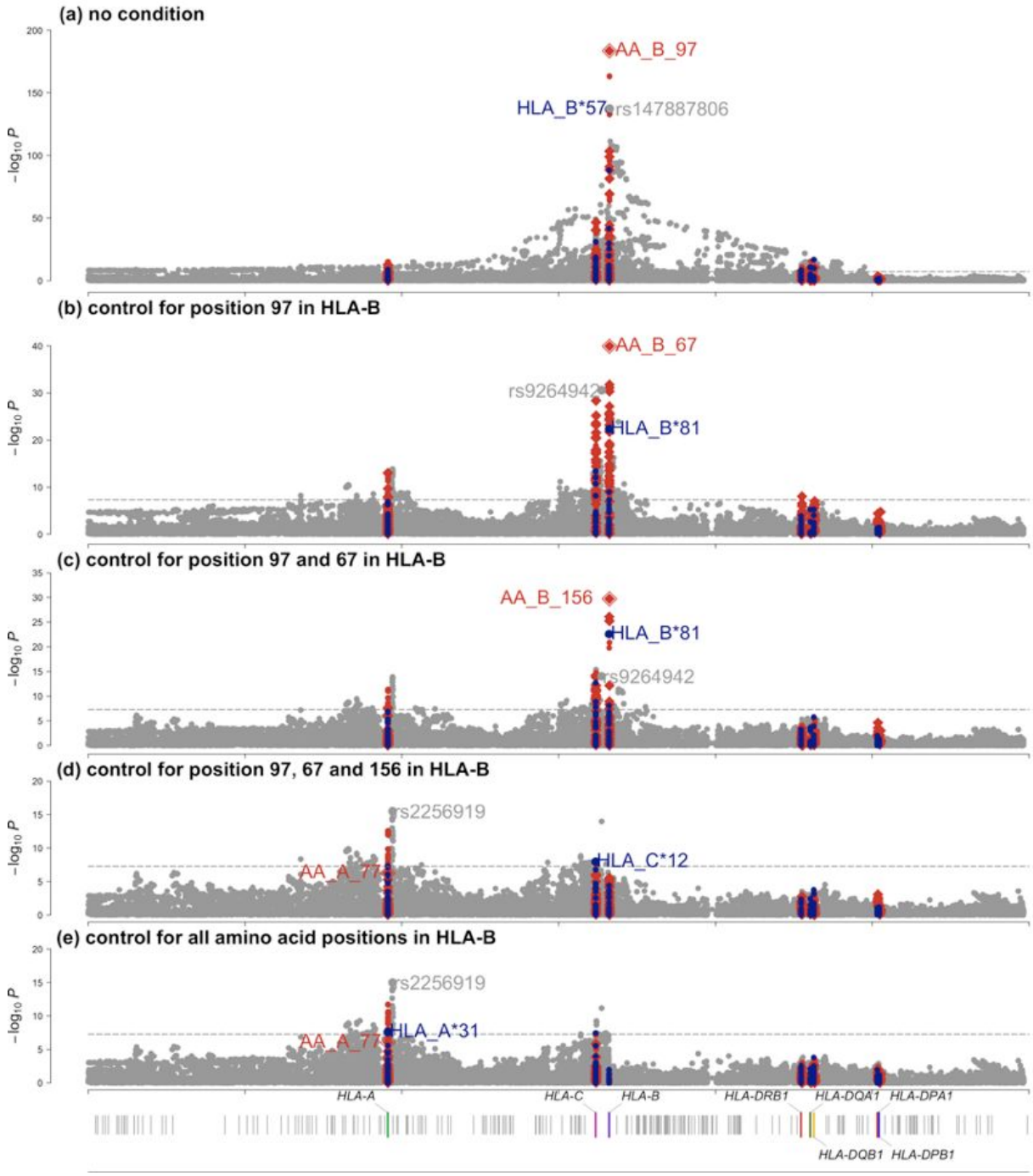
Supplementary Figure 16. Imputation accuracy of GaP registry GWAS data. Imputation accuracy measured as the genotype concordance for G-group classical HLA alleles measured in the 75 subjects included in the GAP registry. Accuracy is compared between the whole multi-ethnic HLA reference (dark purple, sample size = 21,546), The Type 1 Diabetes Genetics Consortium[5] (T1DGC) panel that consists of European individuals only (orange, sample size = 5,225).



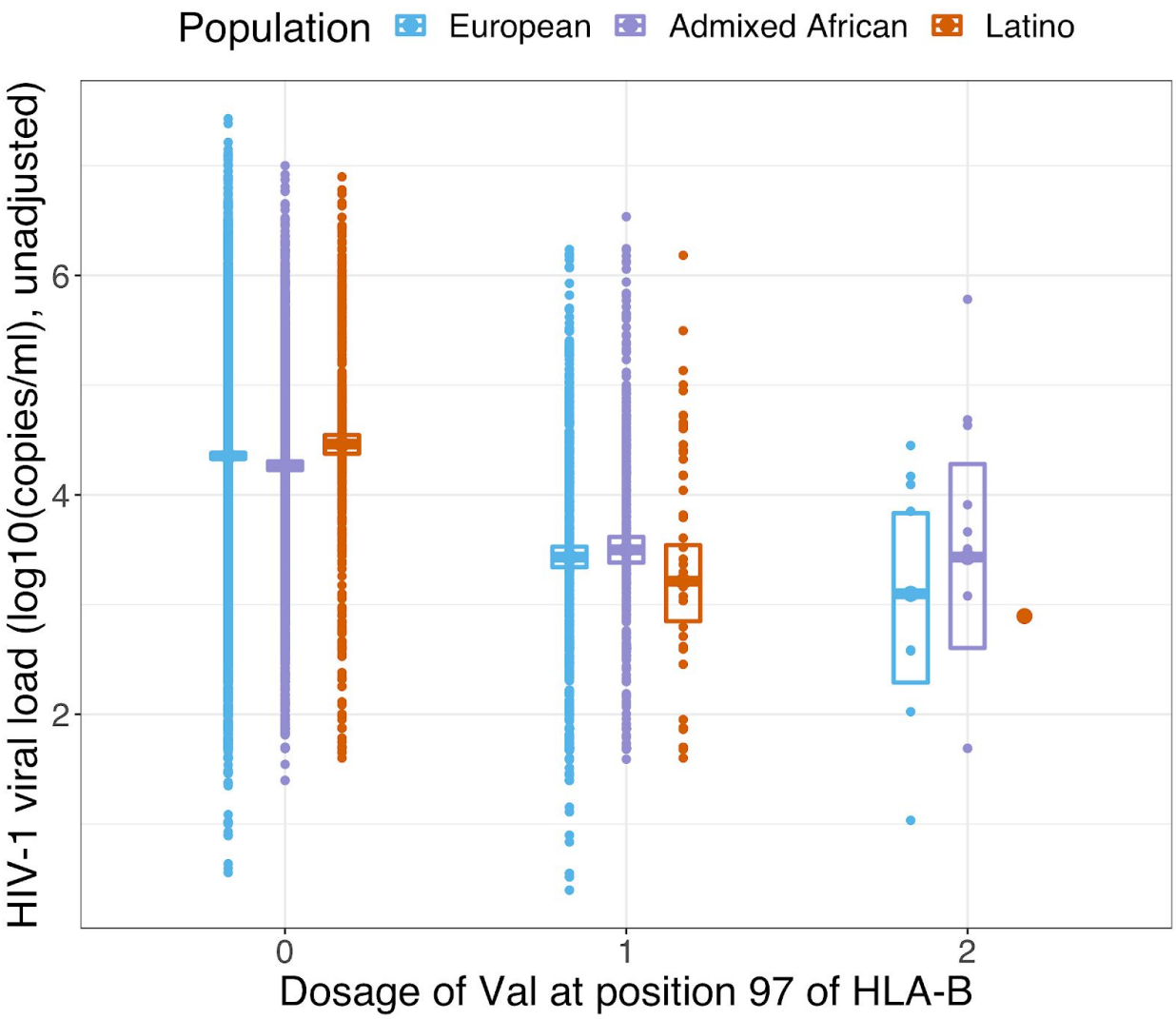
Supplementary Figure 17. Correlation between imputed and typed dosage (Dosage r^2) of classical *HLA* alleles in 1,067 Admixed African HIV-1 samples. The x-axis shows the minor allele frequency observed in the SBT dataset. Blue points show two-field *HLA* alleles. Red points show one-field *HLA* alleles.



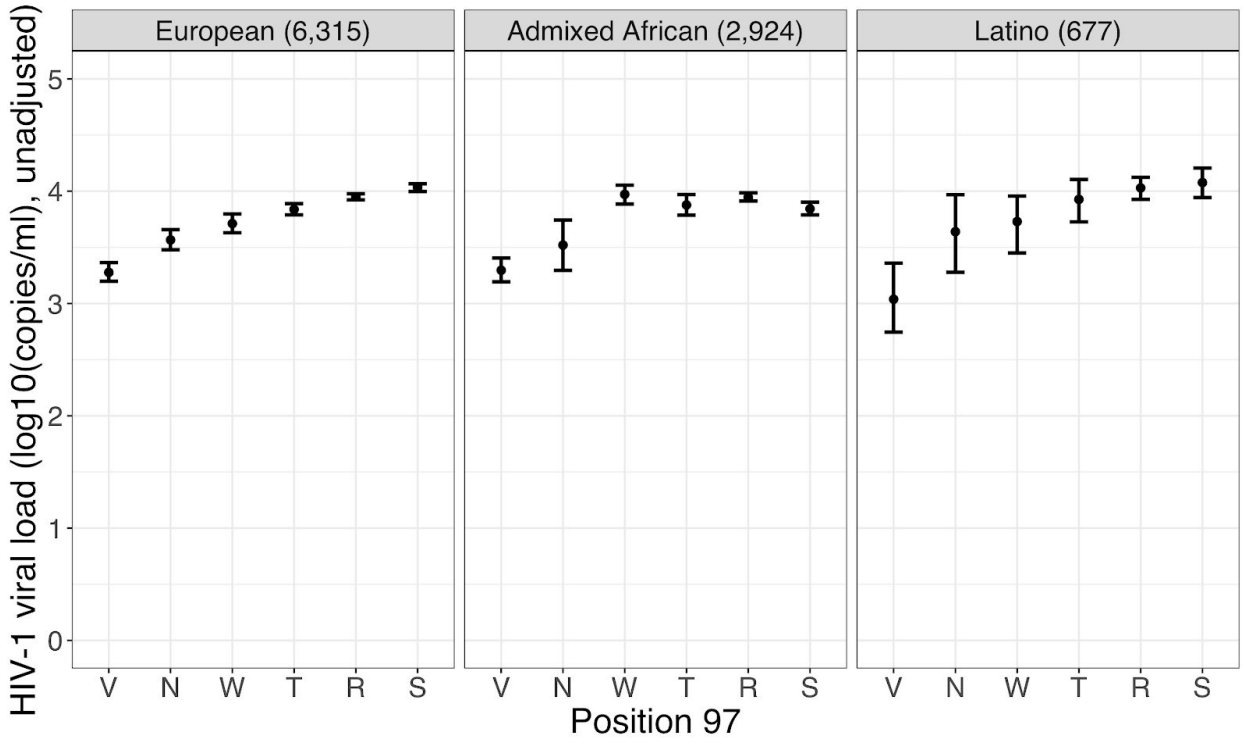
Supplementary Figure 18. Association tests within the MHC to HIV-1 viral load. The x-axis shows the genomic positions of chromosome 6 (build 37), and the y-axis is the -log10 (P-value) obtained from the regression analyses for SNPs (gray), classical *HLA* alleles (blue) and amino acids (red). The dashed black line indicates the genome-wide significance threshold ($P = 5 \times 10^{-8}$). For biallelic markers, results were calculated by a linear regression including sex, cohort-specific principal components and ancestry indicator as covariates (circle). Association at amino acid positions with more than two residues was calculated using a multi-degree-of-freedom omnibus test including the same covariates (diamond). The top associated amino acid, classical *HLA* allele and SNPs are annotated in the figure. (**a**) Of all variants tested, the top hit maps to amino acid position 97 in HLA-B. (**b**) Subsequent conditional analysis controlling for all residues at position 97 in HLA-B revealed an independent association at position 67 in HLA-B. (**c**) Results conditioned on position 97 and 67 in HLA-B showed a third signal at position 156 in HLA-B. (**d**) Results conditioned on position 97, 67 and 156 in HLA-B showed position 77 in HLA-A has the strongest association signal outside HLA-B among all amino acid positions. (**e**) Results conditioned on all amino acid positions in HLA-B. Notably, amino acid positions were more significant than any single SNP or classical *HLA* allele in each conditional analysis for the three amino acid positions in HLA-B.


(a) no condition
AA_B_97
HLA_B*57
rs147887806
(b) control for position 97 in HLA-B
AA_B_67
rs9264942
HLA_B*81
(c) control for position 97 and 67 in HLA-B
AA_B_156
HLA_B*81
rs9264942
(d) control for position 97, 67 and 156 in HLA-B
rs2256919
AA_A_77
HLA_C*12
(e) control for all amino acid positions in HLA-B
rs2256919
AA_A_77
HLA_A*31
HLA-A
HLA-C
HLA-B
HLA-DRB1
HLA-DQA1
HLA-DPA1
HLA-DQB1
HLA-DPB1

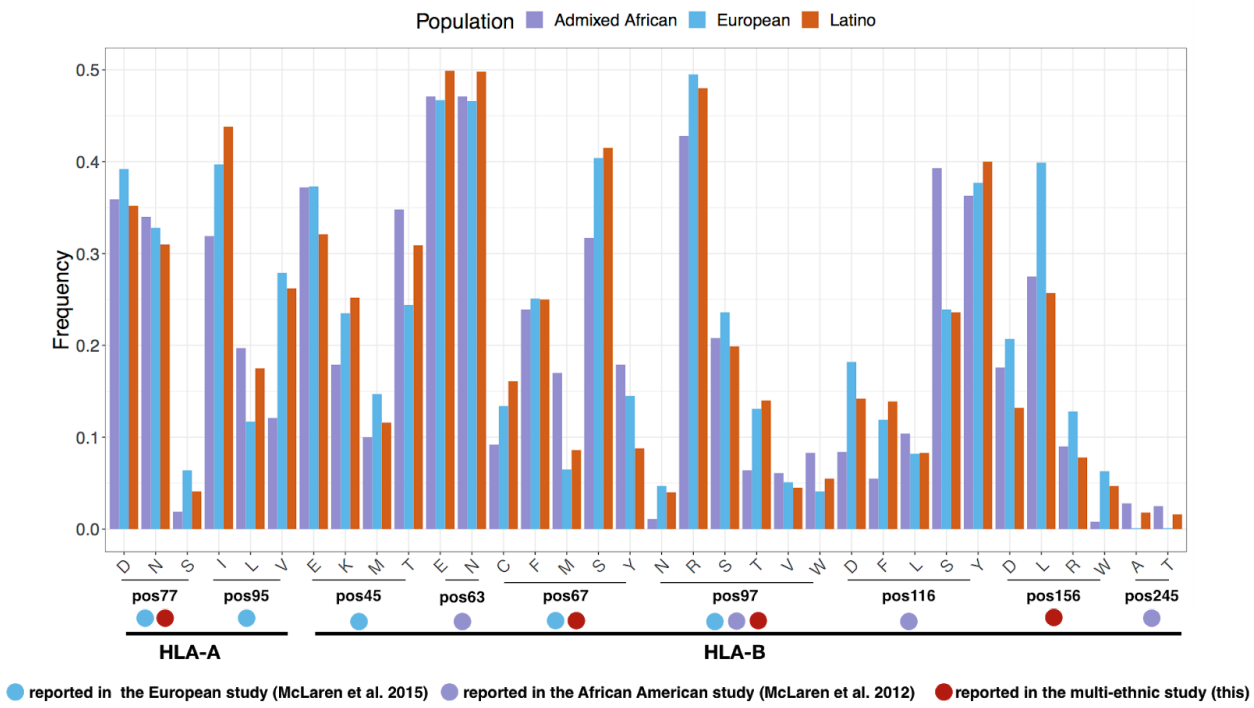
Supplementary Figure 19. Effect of the most protective residue 97Val in HLA-B across diverse populations. Change in log10 HIV-1 set point viral load (spVL, RNA copies per milliliter) of individual amino acid residues at position 97 in HLA-B. Points show the mean of log10(spVL) observed in each population with 95% confidence interval (whiskers). Residues range from strongly protective (i.e., viral load decreasing) to risk (viral load increasing) in the overall population.



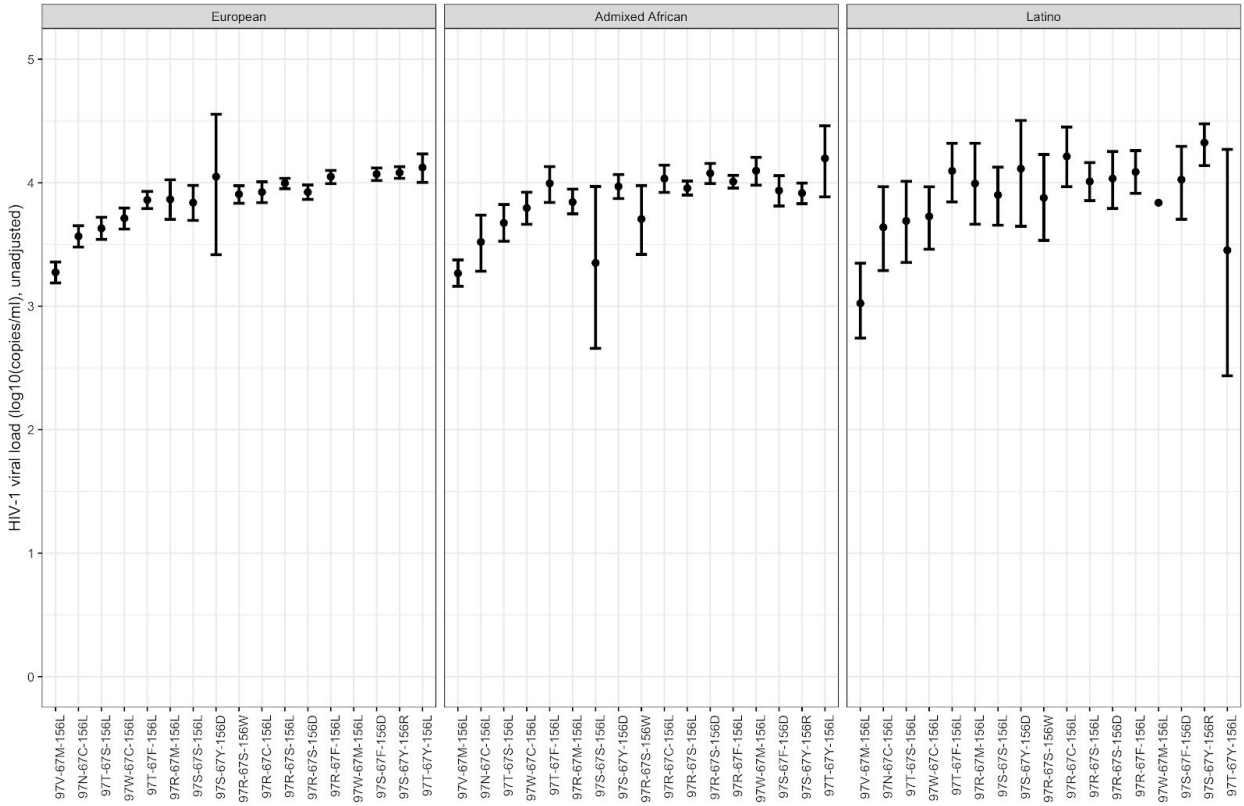
Supplementary Figure 20. Effect on set point viral load of individual residues at position 97 in HLA-B. Mean set point viral load (spVL, RNA copies per milliliter) and its standard error of all six residues at position 97 in HLA-B in three populations independently. Residues are ranked from the most protective to the most risky in the overall population.



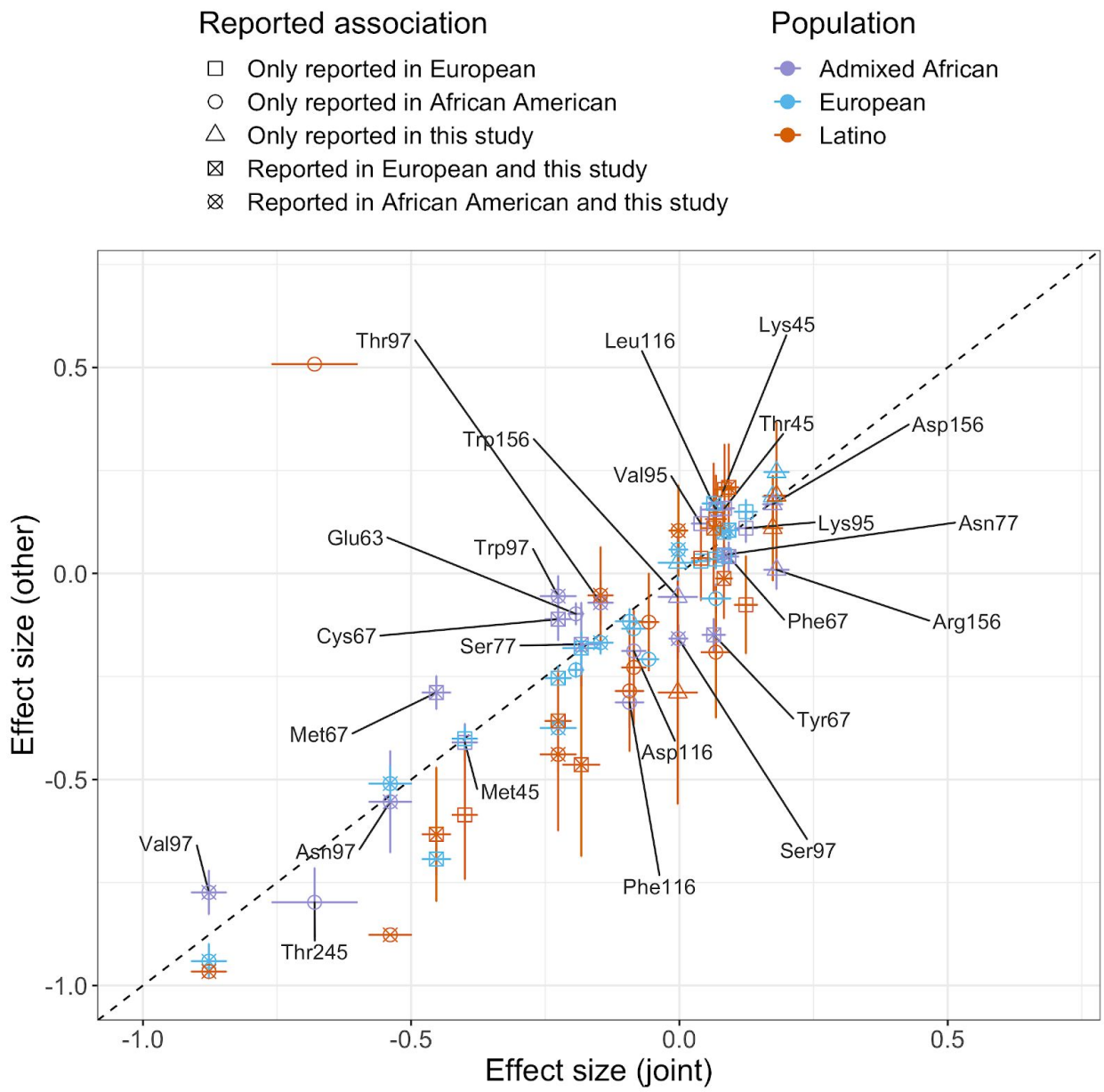
Supplementary Figure 21. Frequencies in all previously and current reported amino acid positions that are associated with HIV-1 viral load.



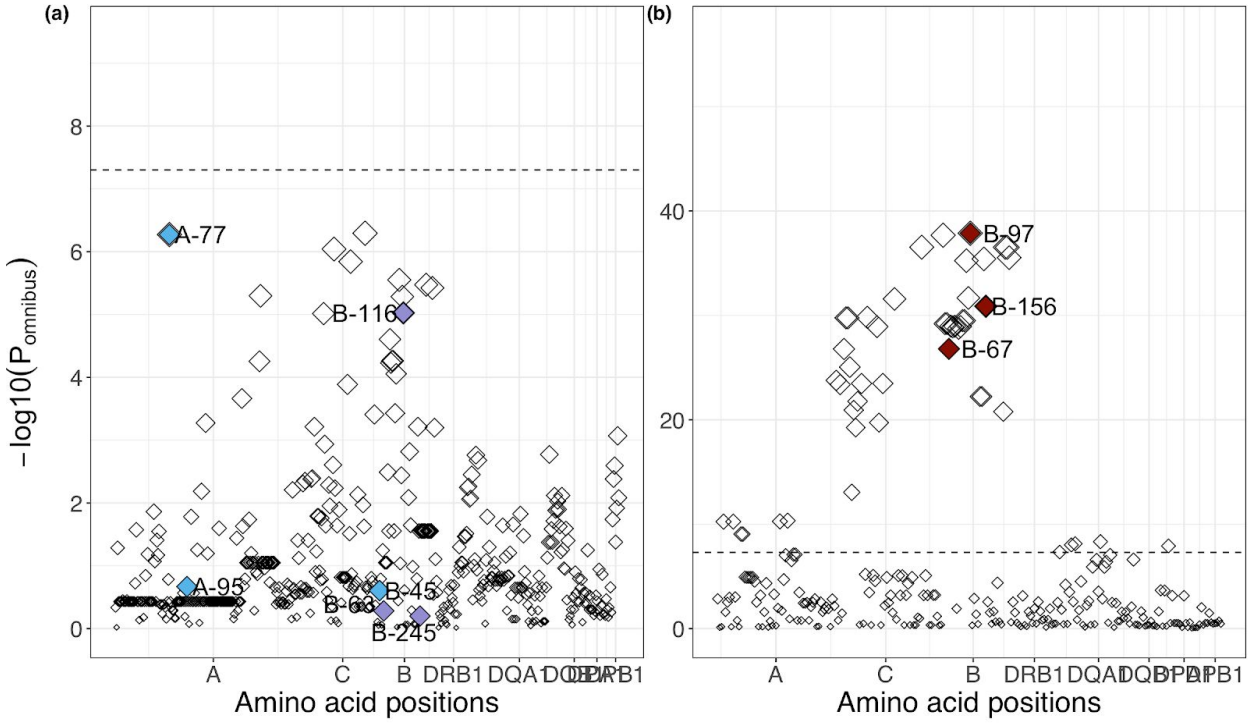
Supplementary Figure 22. Effect on set point viral load of individual haplotypes formed by amino acid positions 97, 67 and 156 in HLA-B. Mean viral load and its standard error of the haplotypes with frequency > 1% formed by the three independently associated amino acid positions (97, 67 and 156 in HLA-B) in three populations independently.



Supplementary Figure 23. Effect on set point viral load of individual amino acid residues at each position reported in this and previous work[6,7]. 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 and its standard error in individual populations (purple = Admixed American; blue = European and orange = Latino). Different shapes indicate which study an amino acid was reported in.



Supplementary Figure 24. Conditional analysis of other previously reported independently associated amino acid positions. Manhattan plots of amino acid positions in the six classical HLA genes. Each point shows a single amino acid position and its omnibus P-value after controlling for (**a**) independent positions that are associated with spVL in this study (position 97, 67 and 156 in HLA-B); and (**b**) independent positions that are only reported in previous studies[6,7] and not in the presented work (position 45, 63 and 116 in HLA-B and position 77, 95 in HLA-A). Independently associated amino acid positions that are only reported in the European population[7] are shown in blue. Independently associated amino acid positions that are only reported in the African American population[6] are shown in purple. Independently associated amino acid positions identified in this study are shown in red.



References

1. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
2. Zhou, Q., Zhao, L. & Guan, Y. Strong Selection at MHC in Mexicans since Admixture. *PLoS Genet.* **12**, e1005847 (2016).
3. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
4. Bhatia, G. *et al.* Genome-wide scan of 29,141 African Americans finds no evidence of directional selection since admixture. *Am. J. Hum. Genet.* **95**, 437–444 (2014).
5. Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One* **8**, e64683 (2013).
6. 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).
7. 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).