

# 1 **A high-resolution HLA reference panel capturing global population diversity**

## 2 **enables multi-ethnic fine-mapping in HIV host response**

3 Yang Luo<sup>1,2,3,4,5</sup>, Masahiro Kanai<sup>5,4,6,7,8</sup>, Wanson Choi<sup>9</sup>, Xinyi Li<sup>10</sup>, Kenichi Yamamoto<sup>8,11</sup>, Kotaro  
4 Ogawa<sup>8,12</sup>, Maria Gutierrez-Arcelus<sup>1,2,3,4,5</sup>, Peter K. Gregersen<sup>13</sup>, Philip E. Stuart<sup>14</sup>, James T.  
5 Elder<sup>14,15</sup>, Jacques Fellay<sup>16,17</sup>, Mary Carrington<sup>18,19</sup>, David W. Haas<sup>20,21</sup>, Xiuqing Guo<sup>22</sup>, Nicholette  
6 D. Palmer<sup>23</sup>, Yii-Der Ida Chen<sup>22</sup>, Jerome. I. Rotter<sup>22</sup>, Kent. D. Taylor<sup>22</sup>, Stephen. S. Rich<sup>24</sup>, Adolfo  
7 Correa<sup>25</sup>, James G. Wilson<sup>26</sup>, Sekar Kathiresan<sup>5,27,28</sup>, Michael H. Cho<sup>29</sup>, Andres Metspalu<sup>30</sup>, Tonu  
8 Esko<sup>5,30</sup>, Yukinori Okada<sup>8,31</sup>, Buhm Han<sup>32</sup>, NHLBI Trans-Omics for Precision Medicine (TOPMed)  
9 Consortium, Paul J. McLaren<sup>33,34</sup>, Soumya Raychaudhuri<sup>1,2,3,4,5,35</sup>

10 1 Center for Data Sciences, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA  
11 2 Division of Rheumatology, Immunology, and Immunity, Brigham and Women's Hospital, Harvard  
12 Medical School, Boston, MA, USA  
13 3 Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA  
14 4 Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA  
15 5 Broad Institute of MIT and Harvard, Cambridge, MA, USA  
16 6 Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA  
17 7 Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, USA  
18 8 Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, 565-0871,  
19 Japan  
20 9 Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul, South Korea  
21 10 Committee on Genetics, Genomics, and Systems Biology, University of Chicago, Chicago, IL, USA  
22 11 Department of Pediatrics, Osaka University Graduate School of Medicine, Osaka 565-0871, Japan  
23 12 Department of Neurology, Osaka University Graduate School of Medicine, Osaka 565-0871, Japan  
24 13 The Robert S. Boas Center for Genomics and Human Genetics, Feinstein Institute for Medical  
25 Research, North Shore LIJ Health System, Manhasset, NY, USA  
26 14 Department of Dermatology, University of Michigan, Ann Arbor, Michigan, USA  
27 15 Ann Arbor Veterans Affairs Hospital, Ann Arbor, Michigan, USA  
28 16 Precision Medicine Unit, Lausanne University Hospital and University of Lausanne, Lausanne,  
29 Switzerland  
30 17 School of Life Sciences, EPFL, Lausanne, Switzerland  
31 18 Basic Science Program, Frederick National Laboratory for Cancer Research, Frederick, Maryland,  
32 USA  
33 19 Ragon Institute of MGH, MIT and Harvard, Boston, Massachusetts, USA  
34 20 Vanderbilt University Medical Center, Nashville, TN, USA  
35 21 Meharry Medical College, Nashville, TN, USA  
36 22 The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The  
37 Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA USA  
38 23 Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, North Carolina, USA  
39 24 Center for Public Health Genomics, University of Virginia School of Medicine Charlottesville, VA, USA

40 25 Medicine, University of Mississippi Medical Center, MS, USA  
41 26 Physiology and Biophysics, University of Mississippi Medical Center, MS, USA  
42 27 Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA  
43 28 Cardiology Division of the Department of Medicine, Massachusetts General Hospital, Boston, MA,  
44 USA  
45 29 Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital,  
46 Harvard Medical School, Boston, MA, USA  
47 30 Estonian Genome Center, Institute of Genomics, University of Tartu, Estonia  
48 31 Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka  
49 University, Suita 565-0871, Japan  
50 32 Department of Medicine, Seoul National University College of Medicine, Seoul, Korea  
51 33 J.C. Wilt Infectious Diseases Research Centre, National Microbiology Laboratories, Public Health  
52 Agency of Canada, Winnipeg, Canada  
53 34 Department of Medical Microbiology and Infectious Diseases, University of Manitoba, Winnipeg,  
54 Canada  
55 35 Centre for Genetics and Genomics Versus Arthritis, University of Manchester, Manchester, UK

56 **Correspondence to**

57 Yang Luo  
58 Harvard New Research Building  
59 77 Avenue Louis Pasteur, Suite 255  
60 Boston, MA 02115  
61 [yangluo@broadinstitute.org](mailto:yangluo@broadinstitute.org)  
62 Tel: 617-525-4468

63 Soumya Raychaudhuri  
64 Harvard New Research Building  
65 77 Avenue Louis Pasteur, Suite 250  
66 Boston, MA 02115  
67 [soumya@broadinstitute.org](mailto:soumya@broadinstitute.org)  
68 Tel: 617-525-4484 Fax: 617-525-4488

## 69 Abstract

70 **Defining causal variation by fine-mapping can be more effective in multi-ethnic genetic**  
71 **studies, particularly in regions such as the MHC with highly population-specific**  
72 **structure. To enable such studies, we constructed a large (N=21,546) high resolution HLA**  
73 **reference panel spanning five global populations based on whole-genome sequencing**  
74 **data. Expectedly, we observed unique long-range HLA haplotypes within each population**  
75 **group. Despite this, we demonstrated consistently accurate imputation at G-group**  
76 **resolution (94.2%, 93.7%, 97.8% and 93.7% in Admixed African (AA), East Asian (EAS),**  
77 **European (EUR) and Latino (LAT)). We jointly analyzed genome-wide association studies**  
78 **(GWAS) of HIV-1 viral load from EUR, AA and LAT populations. Our analysis pinpointed**  
79 **the MHC association to three amino acid positions (97, 67 and 156) marking three**  
80 **consecutive pockets (C, B and D) within the HLA-B peptide binding groove, explaining**  
81 **12.9% of trait variance, and obviating effects of previously reported associations from**  
82 **population-specific HIV studies.**

## 83 Main

84 The HLA genes located within the MHC region encode proteins that play essential roles in  
85 immune responses including antigen presentation. They account for more heritability than all  
86 other variants together for many diseases<sup>1-4</sup>. It also has more reported GWAS trait associations  
87 than any other locus<sup>5</sup>. The extended MHC region spans 6Mb on chromosome 6p21.3 and  
88 contains more than 260 genes<sup>6</sup>. Due to population-specific positive selection it harbors  
89 unusually high sequence variation, longer haplotypes than most of the genome, and haplotypes

90 that are specific to individual ancestral populations<sup>7,8</sup>. Consequently, the MHC is among the  
91 most challenging regions in the genome to analyze. Advances in HLA imputation have enabled  
92 population-specific association and fine-mapping studies of this locus<sup>2,9-12</sup>. But despite large  
93 effect sizes, fine-mapping in multiple populations simultaneously is challenging without a single  
94 large and high-resolution multi-ethnic reference panel. This has caused confusion in some  
95 instances. For example, defining the driving HLA alleles may inform the design of antigenic  
96 peptides for vaccines<sup>13,14</sup> for HIV-1, which led to 770,000 deaths in 2018 alone<sup>15</sup>. However,  
97 multiple risk HLA risk alleles have been independently reported in different populations<sup>1,10,16</sup>, and  
98 it is not clear if they represent truly population-specific signals or are confounded by linkage.

## 99 Results

### 100 Performance evaluation of inferred classical HLA alleles

101 To build a large-scale multi-ethnic HLA imputation reference panel, we used high-coverage  
102 whole genome sequencing (WGS) datasets<sup>17-21</sup> from the Japan Biological Informatics  
103 Consortium<sup>20</sup>, the BioBank Japan Project<sup>18</sup>, the Estonian Biobank<sup>22</sup>, the 1000 Genomes Project  
104 (1KG)<sup>21</sup> and a subset of studies in the TOPMed program (**Supplementary Note**,  
105 **Supplementary Table 1-2**). To perform HLA typing using WGS data, we extracted reads  
106 mapped to the extended MHC region (chr6:25Mb-35Mb) and unmapped reads from 24,338  
107 samples. We applied a population reference graph<sup>23-25</sup> for the MHC region to infer classical  
108 alleles for three HLA class I genes (HLA-*A*, -*B* and -*C*) and five class II genes (HLA-*DQA1*,  
109 -*DQB1*, -*DRB1*, -*DPA1*, -*DPB1*) at G-group resolution, which determines the sequences of the  
110 exons encoding the peptide binding groove. We required samples to have >20x coverage  
111 across all HLA genes (**Supplementary Table 1, 3**). After quality control our panel included  
112 21,546 individuals: 10,187 EUR, 7,849 AA, 2,069 EAS, 952 LAT and 489 SAS.

113 To assess the accuracy of the WGS *HLA* allele calls, we compared the inferred *HLA* classical  
114 alleles to gold standard sequence-based typing (SBT) in 955 1KG subjects and 288 Japanese  
115 subjects and quantified concordance. In both cohorts we observed slightly higher average  
116 accuracy for class I genes, obtaining 99.0% (one-field, formally known as two-digit), 99.2%  
117 (amino acid) and 96.5% (G-group resolution), than class II genes, obtaining 98.7% (one-field),  
118 99.7% (amino acid) and 96.7% (G-group resolution, **Methods, Supplementary Figure 1,**  
119 **Supplementary Tables 4-5, Extended Data 1**).

## 120 HLA diversity

121 To quantify MHC diversity, we calculated identity-by-descent (IBD) distances<sup>26</sup> between all  
122 individuals using 38,398 MHC single nucleotide polymorphisms (SNPs) included in the  
123 multi-ethnic HLA reference panel (N=21,546) and applied principal component analysis (PCA,  
124 **Methods**). PCA distinguished EUR, EAS and AA as well as the admixed LAT and SAS samples  
125 (**Figure 1a, Supplementary Figure 2**). This reflected widespread *HLA* allele frequency  
126 differences between populations (**Figure 1b-c, Supplementary Figure 3**). Of 130 unique  
127 common (frequency > 1%) G-group alleles, 129 demonstrated significant differences of  
128 frequencies across populations (4 degree-of-freedom Chi-square test, p-value < 0.05/130,  
129 **Supplementary Figure 4**). The only exception was *DQA1\*01:01:01G* which was nominally  
130 significant (unadjusted p-value = 0.047). These differences may be related to adaptive selection.  
131 For example, the *B\*53:01:01G* allele is enriched in Admixed Africans (11.7% in AA versus 0.3%  
132 in others) and it has been previously associated with malaria protection<sup>27,28</sup>. Consistent with  
133 previous reports<sup>29,30</sup>, we observed that *HLA-B* had the highest allelic diversity (n=443) while

134 HLA-*DQA1* had the least ( $n=17$ , **Supplementary Figure 5-6, Supplementary Table 6,**  
135 **Extended Data 1**).

136 To understand the haplotype structure of HLA between pairs of HLA genes we calculated a  
137 multiallelic linkage disequilibrium (LD) measurement index<sup>31-33</sup>,  $\epsilon$ , which is 0 when there is no  
138 LD and 1 when there is perfect LD (**Figure 2a**). We observed higher  $\epsilon$  between *DQA1*, *DQB1*,  
139 and *DRB1*; between *DPA1* and *DPB1*; and between *B* and *C* (**Supplementary Figure 7**). The  
140 heterogeneity between different populations was underscored by the presence of  
141 population-specific common (frequency >1%) high resolution long-range haplotypes  
142 (HLA-A~C~B~DRB1~DQA1~DQB1~DPA1~DPB1, **Figure 2b, Supplementary Figure 8-12,**  
143 **Extended Data 2, Methods**). The most common within-population haplotype was A24::DP6  
144 (HLA-A\*24:02:01G~C\*12:02:01G~B\*52:01:01G~DRB1\*15:02:01G~DQA1\*01:03:01G~DQB1\*06  
145 :01:01G~DPA1\*02:01:01G~DPB1\*09:01:01G) found at a frequency of 3.61% in EAS  
146 (**Supplementary Figure 8**). This haplotype is strongly associated with immune-mediated traits  
147 such as HIV<sup>34</sup> and ulcerative colitis<sup>35</sup> in Japanese individuals. The next most common haplotype  
148 was the well-described European-specific ancestral haplotype A1::DP1 or 8.1<sup>36,37</sup> (  
149 frequency=2.76%,  
150 HLA-A\*01:01:01G~C\*07:01:01G~B\*08:01:01G~DRB1\*03:01:01G~DQA1\*05:01:01G~DQB1\*02:  
151 01:01G~DPA1\*02:01:02G~DPB1\*01:01:01G, **Supplementary Figure 9**). This haplotype is  
152 associated with diverse immunopathological phenotypes in the European population, including  
153 systemic lupus erythematosus<sup>38</sup>, myositis<sup>39</sup> and several other conditions<sup>36</sup>. We observed  
154 long-range haplotypes in admixed populations including A1::DP4 in SAS (frequency=1.86%,  
155 **Supplementary Figure 10**), A30::DP1 in AA (frequency=1.18%,  
156 HLA-A\*30:01:01G~C\*17:01:01G~B\*42:01:01:G~DRB1\*03:02:01G~DQA1\*04:01:01G~DQB1\*04

157 :02:01G~DPA1\*02:02:02G~DPB1\*01:01:01G , **Supplementary Figure 11**), and A29::DP11 in  
158 LAT (frequency=0.74%,  
159 HLA-A\*29:02:01G~C\*16:01:01G~B\*44\*03:01:G~DRB1\*07:01:01G~DQA1\*02:01:01G~DQB1\*0  
160 2:01:01G~DPA1\*02:01:01G~DPB1\*11:01:01G, **Supplementary Figure 12**).

161 These haplotypes also have associations with multiple diseases: for example C\*06:02~B\*57:01  
162 is associated with psoriasis<sup>40</sup> and A\*30:01~C\*17:01~B\*42:01 is associated with HIV<sup>41</sup>.

### 163 HLA selection signature

164 Previous studies have suggested that recent natural selection favors African ancestry in the  
165 HLA region in admixed populations<sup>42-45</sup>. To test this hypothesis in our data, we obtained WGS  
166 data from a subset of individuals within two admixed populations (1,832 AA and 594 LAT,  
167 determined by the first three global principal components, **Supplementary Figure 13**,  
168 **Supplementary Note**). Admixed individuals have genomes that are a mosaic of different  
169 ancestries. If genetic variations or haplotypes from an ancestral population are advantageous,  
170 then they are under selection and are expected to have higher frequency than by chance. Using  
171 ELAI<sup>46</sup>, we quantified how much the ancestry proportions differed within the MHC from the  
172 genome-wide average. In AA, we observed that the average genome-wide proportion of African  
173 ancestry was 74.5%, compared to 78.0% in the extended MHC region, corresponding to a 3.42  
174 (95% CI: 3.35-3.49) standard deviation increase. In LAT, we observed 5.76% African ancestry  
175 genome-wide versus 16.0% in the extended MHC region, representing an increase of 4.23  
176 (95% CI: 4.14-4.31) standard deviations (**Methods, Supplementary Figure 14**). To ensure our  
177 results are robust to different local ancestry inference methods, we applied an alternative  
178 method called RFMix<sup>47</sup> and observed a similarly consistent MHC-specific excess of African  
179 ancestry in LAT, and also an excess in AA that was more modest (**Supplementary Figure 14**).



180 Construction of a multi-ethnic HLA reference panel and its performance evaluation

181 Next, we constructed a multi-ethnic HLA imputation reference panel based on classical HLA  
182 alleles and 38,398 genomic markers in the extended MHC region using a novel HLA-focused  
183 pipeline HLA-TAPAS (HLA-Typing At Protein for Association Studies). Briefly, HLA-TAPAS can  
184 handle HLA reference panel construction (*MakeReference*); HLA imputation (*SNP2HLA*) and  
185 HLA association (*HLAassoc*) (**Methods, URLs**). Compared to a widely used HLA reference  
186 panel with European-only individuals (The Type 1 Diabetes Genetics Consortium<sup>48</sup>, T1DGC),  
187 this new reference panel has a six-fold increase in the number of observed *HLA* alleles and  
188 non-*HLA* genomic markers (**Supplementary Table 7**). We noted the difference in observed  
189 classical *HLA* alleles is mainly due to the inclusion of diverse populations rather than its size;  
190 after downsampling the reference panel to be the same size as T1DGC (N=5,225), there was  
191 still a three-fold increase in observed alleles (**Figure 3a**).

192 To empirically assess imputation accuracy of our reference panel, we first used the publicly  
193 available gold-standard *HLA* types (*HLA-A*, *-B*, *-C*, *-DRB1* and *-DQB1*) of 1,267 diverse samples  
194 from AA, EAS, EUR and LAT included in 1KG. We removed 955 overlapping samples within the  
195 reference panel, and to ensure a representative analysis we kept 6,007 markers overlapping  
196 with the *Global Genotyping Array* SNPs. Across the five genes, the average G-group resolution  
197 accuracies were 94.2%, 93.7%, 97.8% and 93.7% in AA, EAS, EUR and LAT (**Figure 3b-c**,  
198 **Supplementary Table 8, Methods, Extended Data 3**). Compared to the T1DGC panel, our  
199 multi-ethnic reference panel showed the most improvement for individuals of non-European  
200 descent; we obtained 4.27%, 2.96%, 2.90% and 1.05% improvement at G-group resolution for  
201 AA, EAS, LAT, and EUR individuals, respectively (**Figure 3d**). Increased diversity was



202 responsible for the improvement; downsampling the reference panel to the same size as the  
203 T1DGC panel still yielded superior performance (**Figure 3d**). To validate our panel further, we  
204 imputed *HLA* alleles into a multi-ethnic cohort of 2,291 individuals from the Genotype and  
205 Phenotype (GaP) registry genotyped on the ImmunoChip array. We obtained SBT *HLA* type  
206 information for six classical class I and class II loci (*HLA-A*, *-B*, *-C*, *-DQA1*, *-DQB1*, *-DRB1*) in 75  
207 samples with diverse ancestral background (25 EUR, 25 EAS and 25 AA, **Supplementary**  
208 **Figure 15, Methods**). Average accuracies were 99.0%, 95.7% and 97.0% for EUR, EAS and  
209 AA respectively when comparing SBT *HLA* alleles at G-group resolution (**Methods, Extended**  
210 **Data 3**). Similar to the 1KG analysis, the multi-ethnic reference panel showed significant  
211 improvement for individuals with non-European descent (6.3% and 11.1% improvement for EAS  
212 and African individuals respectively at G-group resolution), and a more modest 2% improvement  
213 in EUR (**Supplementary Figure 16, Supplementary Table 9**).

214 Fine-mapping causal variants of HIV jointly in three populations in the MHC region  
215 Next we investigated MHC effects within human immunodeficiency virus type 1 (HIV-1) set point  
216 viral load. Upon primary infection with HIV-1, the set point viral load is reached after the immune  
217 system has developed specific cytotoxic T lymphocytes (CTL) that are able to partially control  
218 the virus. It has been well-established that the set point viral load (spVL) varies in the infected  
219 population and positively correlates with rate of disease progression<sup>49</sup>. Previous studies  
220 suggested that HIV-1 infection has a strong genetic component, and specific HLA class I alleles  
221 explain the majority of genetic risk<sup>10,50</sup>. The existence of multiple independent, ancestry-specific,  
222 risk-associated alleles has been reported in both European<sup>1,10</sup> and African American<sup>16</sup>  
223 populations. However, without a multi-ethnic reference panel it has not been possible to  
224 determine if these signals are consistent across different ancestral groups.

225 To define the MHC allelic effects shared across multiple populations, we applied our multi-ethnic  
226 MHC reference panel to 7,445 EUR, 3,901 AA and 677 LAT HIV-1 infected subjects (**Methods**,  
227 **Supplementary Table 10**). Imputation resulted in 640 classical HLA alleles, 4,513 amino acids  
228 in HLA proteins and 49,321 SNPs in the extended MHC region for association and fine-mapping  
229 analysis. We confirmed 96.6% imputation accuracy of two-field (or four-digit) resolution with a  
230 minor allele frequency > 0.5% in this cohort by comparing imputed classical alleles to the SBT  
231 alleles in a subset of 1,067 AA subjects<sup>16</sup>(**Supplementary Figure 17, Extended Data 3**).

232 We next tested SNPs, amino acid positions and classical *HLA* alleles across the MHC for  
233 association to spVL. We performed this jointly in EUR, AA and LAT population using a linear  
234 regression model with sex, principal components and ancestry as covariates (**Methods**). In  
235 agreement with previous studies, we found the strongest spVL-associated classical *HLA* allele  
236 is *B\*57* (effect size = -0.84,  $P_{binary} = 8.68 \times 10^{-144}$ ). This corresponded to a single residue Val97  
237 in HLA-B that tracks almost perfectly with *B\*57* ( $r^2 = 0.995$ ) and showed the strongest  
238 association of any single residue (effect size = -0.84,  $P_{binary} = 5.99 \times 10^{-145}$ , **Supplementary**  
239 **Figure 18**).

240 Then to determine which amino acid positions have independent association with spVL, we  
241 tested each of the amino acid positions by grouping haplotypes carrying a specific residue at  
242 each position in an additive model<sup>2,9</sup> (**Methods**). We found the strongest spVL-associated amino  
243 acid variant in HLA-B is as previously reported<sup>1,10,16</sup> at position 97 (**Figure 4a-b, Supplementary**  
244 **Table 11**) which strikingly explains 9.06% of the phenotypic variance. Position 97 in HLA-B was  
245 more significant ( $P_{omnibus} = 2.86 \times 10^{-184}$ ) than any single SNP or classical *HLA* allele, including

246 **B\*57 (Supplementary Figure 18, Extended Data 4)**. Of the six allelic variants  
247 (Val/Asn/Trp/Thr/Arg/Ser) at this position, the Val residue conferred the strongest protective  
248 effect (effect size = -0.88,  $P = 9.32 \times 10^{-152}$ , **Supplementary Figure 19**) relative to the most  
249 common residue Arg (frequency = 47.8%). All six amino acid alleles have consistent  
250 frequencies and effect sizes across the three population groups (**Figure 5a-b, Supplementary**  
251 **Figure 20**).

252 We next wanted to test whether there were other independent effects outside of position 97 in  
253 HLA-B. After accounting for the effects of amino acid 97 in HLA-B using a conditional haplotype  
254 analysis (**Methods**), we observed a significant independent association at position 67 in HLA-B  
255 ( $P_{omnibus} = 2.82 \times 10^{-39}$ , **Figure 4c-d, Supplementary Table 11**). Considering this might be an  
256 artifact of forward search, we exhaustively tested all possible pairs of polymorphic amino acid  
257 positions in HLA-B. Of 7,260 pairs of amino acid positions, none obtained a better  
258 goodness-of-fit than the pair of positions 97 and 67, which collectively explained 11.2% variance  
259 in spVL (**Figure 5e, Supplementary Table 12**). At position 67, Met67 residue shows the most  
260 protective effect (effect size = -0.44,  $P = 1.19 \times 10^{-59}$ ) among the five possible amino acids  
261 (Cys/Phe/Met/Ser/Tyr) relative to the most common residue Ser (frequency =10.0%).

262 Conditioning on positions 97 and 67 revealed an additional association at position 156 in HLA-B  
263 ( $P_{omnibus} = 1.92 \times 10^{-30}$ , **Figure 4e-f, Supplementary Table 11**). In agreement with the  
264 stepwise conditional analysis, when we tested all 287,980 possible combinations of three amino  
265 acid positions in HLA-B, the most statistically significant combination of amino acids sites is 67,  
266 97 and 156 ( $P = 5.68 \times 10^{-244}$ , **Supplementary Table 13**). These three positions explained  
267 12.9% of the variance (**Figure 5e**). At position 156, residue Arg shows the largest risk effect

268 (effect size = 0.180,  $P = 8.92 \times 10^{-14}$ ) among the four possible allelic variants

269 (Leu/Arg/Asp/Trp), relative to the most common residue Leu (frequency = 35.1%).

270 These amino acid positions mark three consecutive pockets within the HLA-B peptide-binding

271 groove (**Figure 5c**). Position 97 is located in the C-pocket and has an important role in

272 determining the specificity of the peptide-binding groove<sup>51,52</sup>. Position 67 is in the B-pocket, and

273 Met67 side chains occupy the space where larger B-pocket anchors reside in other

274 peptide-MHC structures; its presence limits the size of potential peptide position P2 side

275 chains<sup>52</sup>. Amino acid position 156 is part of the D-pocket and influences the conformation of the

276 peptide-binding region<sup>53</sup>. These results are consistent with the observation that in HLA-*B\*57*, the

277 single most protective spVL-associated one-field allele (a single change at position 156 from

278 Leu → Arg or equivalently HLA-*B\*57:03* → HLA-*B\*57:02*) leads to an increased repertoire of

279 HIV-specific epitope<sup>41,54</sup>.

280 Despite differences in the power to detect associations due to differences in allele frequencies

281 (**Supplementary Figure 21**), we observed generally consistent effects of individual residues

282 across populations (**Figure 5d, Supplementary Figure 22-23, Supplementary Table 14**).

283 There are 26 unique haplotypes defined by the amino acids at positions 67, 97 and 156 in

284 HLA-B (**Table 1, Supplementary Table 15**). When we tested for effect size heterogeneity by

285 ancestry for each of these haplotypes (**Methods**), we observed only 2 of 26 haplotypes showed

286 heterogeneity (F-test P-value < 0.05/26), possibly due to different interplay between genetic and

287 environmental variation at population-level. These results support the concept that these

288 positions mediate HIV-1 viral load in diverse ancestries.

289 To assess whether there were other independent MHC associations outside HLA-B, we  
290 conditioned on all amino acid positions in HLA-B and observed associations at HLA-A, including  
291 at position 77 in HLA-A ( $P_{omnibus} = 9.10 \times 10^{-7}$ , **Figure 4g-h, Supplementary Table 11**), the  
292 classical *HLA* allele *HLA-A\*31* ( $P_{binary} = 2.45 \times 10^{-8}$ ) and the *rs2256919* promoter SNP (  
293  $P_{binary} = 3.10 \times 10^{-16}$ , **Supplementary Figure 18**). These associations argue for an effect at  
294 HLA-A, but larger studies and functional studies will be necessary to define the driving effects.

## 295 Discussion

296 In our study we demonstrated accurate imputation with a single large reference panel for HLA  
297 imputation. We have shown how this reference panel can be used to impute genetic variation at  
298 eight *HLA* classical genes accurately across a wide range of populations. Accurate imputation in  
299 multi-ethnic studies is essential for fine-mapping.

300 We showed the utility of this approach by defining the alleles that best explain HIV-1 viral load in  
301 infected individuals. Our work implicates three amino acid positions (97, 67 and 156) in HLA-B  
302 in conferring the known protective effect of HLA class I variation on HIV-1 infection. Combining  
303 all alleles at these three positions explained 12.9% of the variance in spVL (**Figure 5e**). These  
304 positions all fall within the peptide-binding groove of the respective MHC protein (**Figure 5c**),  
305 indicating that variation in the amino acid content of the peptide-binding groove is the major  
306 genetic determinant of HIV control. Supported by experimental studies<sup>54–57</sup>, positions highlighted  
307 in our work indicated a structural basis for the HLA association with HIV disease progression  
308 that is mediated by the conformation of the peptide within the class I binding groove. This result  
309 highlights how a study with ancestrally diverse populations can potentially point to causal  
310 variation by leveraging linkage disequilibrium difference between ethnic groups.

311 We note that previous studies have shown position 97 in HLA-B has the strongest association  
312 with HIV-1 spVL or case-control in African American and European populations, but highlighted  
313 different additional signals via conditional analysis (position 45, 67 in HLA-B and position 77, 95  
314 in HLA-A in Europeans<sup>1,10,16</sup> and position 63, 116 and 245 in HLA-B in African Americans<sup>16</sup>).  
315 These signals do not explain the signals we report here; after conditioning on positions 45, 63,  
316 116, 245 of HLA-B and 95 of HLA-A, the association of the four identified amino acids identified  
317 in this study remained significant ( $P < 5 \times 10^{-8}$ ). In contrast, our binding groove alleles explain  
318 these other alleles; conditioning on the four amino acid positions identified in this study  
319 (positions 67, 97 and 156 in HLA-B), all previously reported positions did not pass the  
320 significance threshold ( $P > 5 \times 10^{-8}$ , **Supplementary Figure 24**).

321 Furthermore, defining the effect sizes for *HLA* alleles across different populations is essential for  
322 defining risk of a wide-range of diseases in the clinical setting. There is increasing application of  
323 genome-wide genotyping by patients both by healthcare providers and direct-to-consumer  
324 vendors. The large effects of the MHC region for a wide-range of immune and non-immune  
325 traits, makes it essential to define *HLA* allelic effect sizes essential in multi-ethnic studies in  
326 order to build generally applicable clinical polygenic risk scores for many diseases in diverse  
327 populations<sup>58-61</sup>. Resources like the one we present here will be an essential ingredient in such  
328 studies.

## 329 Methods

### 330 Individuals included in the reference panel

331 Study participants were from the Jackson Heart Study (JHS , N = 3,027), Multi-Ethnic Study of  
332 Atherosclerosis (MESA, N=4,620), Chronic Obstructive Pulmonary Disease Gene (COPDGene)  
333 study (N=10,623), Estonian Biobank (EST, N=2,244), Japan Biological Informatics Consortium  
334 (JPN, N=295), Biobank Japan (JPN, N=1,025) and 1000 Genomes Project (1KG, N=2,504).  
335 Each study was previously approved by respective institutional review boards (IRBs), including  
336 for the generation of WGS data and association with phenotypes. All participants provided  
337 written consent. Further details of cohort descriptions and phenotype definitions are described in  
338 the **Supplementary Note**.

### 339 HLA-TAPAS

340 HLA-TAPAS (HLA-Typing At Protein for Association Studies) is an HLA-focused pipeline that  
341 can handle HLA reference panel construction (*MakeReference*), HLA imputation (*SNP2HLA*),  
342 and HLA association (*HLAassoc*). It is an updated version of the SNP2HLA<sup>48</sup> to build an  
343 imputation reference panel, perform *HLA* classical allele, amino acid and SNP imputation within  
344 the extended MHC region. Briefly, major updates include (1) using PLINK1.9 (**URLs**) instead of  
345 v1.07; (2) using BEAGLE v4.1 (**URLs**) instead of v3 for phasing and imputation; and (3)  
346 including custom R scripts for performing association and fine-mapping analysis at amino acid  
347 level in multiple ancestries. The source code is available for download (**URLs**).



348 Construction of a multi-ethnic HLA reference panel using whole-genome  
349 sequences

350 To construct a multi-ethnic HLA imputation reference panel, we used 24,338 whole-genome  
351 sequences at different depths (**Supplementary Table 1**). Details of the construction using  
352 deep-coverage whole-genome sequencing are described in the **Supplementary Note**. Briefly,  
353 alignment and variant-calling for genomes sequenced by each cohort were performed  
354 independently. We performed local realignment and quality recalibration with the Genome  
355 Analysis Toolkit<sup>62</sup> (GATK; version 3.6) on Chromosome 6:25,000,000-35,000,000. We detected  
356 single nucleotide variants (SNV) and indels using GATK with HaplotypeCaller. To eliminate  
357 false-positive sites called in the MHC region, we restrict our panel to SNVs reported in 1000  
358 Genomes Project<sup>21</sup> only.

359 We next inferred classical HLA alleles at G-group resolution for eight classical HLA genes  
360 (*HLA-A*, *-B*, *-C*, *-DQA1*, *-DQB1*, *-DRB1*, *-DPA1* and *-DPB1*) using a population reference  
361 graph<sup>24,25</sup>. To extend the reference panel versatility, we inferred amino acid variation, one-field  
362 and two-field resolution alleles from the inferred G-group alleles. After removing samples with  
363 low-coverage and failed genome-wide quality control (**Supplementary Table 3**), we constructed  
364 a multi-ethnic HLA imputation reference panel (N=21,546) using the HLA-TAPAS  
365 *MakeReference* module (**URLs, Method**).

## 366 Sequence-based typing of *HLA* alleles

367 Purified DNA from the 75 donors from the GaP registry (at the Feinstein Institute for Medical  
368 Research) was sent to NHS Blood and Transplant, UK, where *HLA* typing was performed.

369 Next-generation sequencing was done for *HLA-A*, *-B*, *-C*, *-DQB1*, *-DPB1* and *-DRB1*.

370 PCR-sequence-specific oligonucleotide probe sequencing was performed for *HLA-DQA1* in all  
371 samples. These typing methods yielded classical allele calls for seven genes at three-field  
372 (*HLA-A*, *-B*, *-C* and *-DQB1*) or G-group resolution (*HLA-DQA1*, *-DPB1* and *-DRB1*).

373 Genomic DNA from the 288 unrelated samples of Japanese ancestry underwent high-resolution  
374 allele typing (three-field alleles) of six classical *HLA* genes (*HLA-A*, *-B* and *-C* for class I; and  
375 *HLA-DRB1*, *-DQA1* and *-DPB1* for class II)<sup>20</sup>.

376 The 1000 Genomes panel consists of 1,267 individuals with information on five *HLA* genes  
377 (*HLA-A*, *-B*, *-C*, *-DQB1*, and *-DRB1*) at G-group resolution among four major ancestral groups  
378 (AA, EAS, EUR and LAT)<sup>7</sup>.

379 We obtained *HLA* typing of the 1,067 African American subjects included in the HIV-1 viral load  
380 study as described previously<sup>16,63</sup>. Briefly, seven classical *HLA* genes (*HLA-A*, *-B*, *-C*, *-DQA1*,  
381 *-DQB1* *-DRB1* and *-DPB1*) were obtained by sequencing exons 2 and 3 and/or single-stranded  
382 conformation polymorphism PCR, and was provided at two-field resolution.

383 Accuracy measure between inferred and sequence-based typing *HLA*

384 genotypes

385 Allelic variants at HLA genes can be typed at different resolutions: one-field HLA types specify  
386 serological activity, two-field HLA types specify the amino acids encoded by the exons of the  
387 HLA gene, and three-field types determine the full exonic sequence including synonymous  
388 variants. G-group resolution determines the sequences of the exons encoding the peptide  
389 binding groove, that is, exons 2 and 3 for class I and exon 2 class II genes. Thus, any  
390 polymorphism occurring in exon 4 of class I gene or exon 3 of class II gene was not defined.  
391 This means many G-group alleles can map to multiple three-field and two-field *HLA* alleles.

392 We calculated the accuracy at each *HLA* gene by summing across the dosage of each correctly  
393 inferred *HLA* allele or amino acid across all individuals ( $N$ ), and divided by the total number of  
394 observations ( $2*N$ ). That is,

395

$$Accuracy(g) = \frac{\sum_i^N D_i(A_{1i,g}) + D_i(A_{2i,g})}{2N} ,$$

396 where  $Accuracy(g)$  represents the accuracy at a classical HLA gene (e.g. HLA-*B*).  $D_i$   
397 represents the inferred dosage of an allele in individual  $i$ , and alleles  $A_{1i,g}$  and  $A_{2i,g}$  represent  
398 the true (SBT) *HLA* types for an individual  $i$ .

399 To evaluate the accuracy between the inferred and validated *HLA* types obtained from SBT at  
400 G-group resolution, we translated the highest resolution specified by the validation data to its  
401 matching G-group resolution based IMGT/HLA database (e.g. HLA-A\*01:01 →  
402 HLA-A\*01:01:01G), and compared it to the primary output from *HLA\*LA* or *HLA-TAPAS*. We

403 also translated all G-group alleles to their matching amino acid sequences, and compared them  
404 against the validation alleles, we referred to this as the amino acid level.

405 To evaluate imputation performance in individual classical *HLA* alleles and amino acids, we  
406 calculated the dosage  $r^2$  correlation between imputed and SBT dosage.

$$407 \quad r^2 = \frac{\left[ \sum_{i=1}^N x_i y_i - \left( \sum_{i=1}^N x_i \right) \left( \sum_{i=1}^N y_i \right) / N \right]^2}{\left[ \left( \sum_{i=1}^N x_i^2 - \left( \sum_{i=1}^N x_i \right)^2 / N \right) \left( \sum_{i=1}^N y_i^2 - \left( \sum_{i=1}^N y_i \right)^2 / N \right) \right]} ,$$

408 where  $x_i$  and  $y_i$  represents the inferred and SBT dosage of an allele in individual  $i$ .  $N$

409 represents the number of individuals.

#### 410 Principal component analysis

411 We performed a principal component analysis of the MHC region based on the  
412 identity-by-descent (IBD) distances between all 21,809 individuals included in the multi-ethnic  
413 reference panel. We computed the IBD distance using Beagle (Version 4.1, **URLs**) and  
414 averaged over 100 runs with all variants (54,474) included in the HLA reference panel. Due to  
415 uneven representation of different ethnicity groups (**Supplementary Table 2**), we applied a  
416 weighted PCA approach, where mean and standard deviation of the IBD matrix within an  
417 ethnicity group are weighted inversely proportional to the sample size.

#### 418 HLA haplotype frequency estimation

419 We applied an expectation-maximization algorithm approach implemented in Hapl-o-Mat<sup>64</sup>  
420 (**URLs**) to estimate HLA haplotype frequency based on eight classical HLA alleles inferred at  
421 G-group resolution. We estimated haplotype frequencies both overall and within five continental  
422 populations (**Extended Data 2**).

## 423 Local ancestry inference

424 To detect local ancestry in admixed samples, we first applied ELAI<sup>46</sup> to chromosome 6 with 1000  
425 Genomes Project<sup>21</sup> as the reference panel. We extracted 63,998 common HapMap3 SNPs  
426 between the WGS (MESA cohort) and the 1000 Genome reference panel. We used the same  
427 set of SNPs for ELAI and RFMix analysis. We applied ELAI<sup>46</sup> to 1,832 African Americans and  
428 594 Latinos. For 1,832 African American individuals included in the study, we used genotypes of  
429 99 CEU and 108 YRI in the 1000 Genome Project as reference panel, assuming admixture  
430 generation to be seven generations ago. We used two upper-layer clusters and 10 lower-layer  
431 clusters in the model. For Latinos, we selected 65 Latinos with Native American (NAT) ancestry  
432 > 75% included in the 1000 Genomes Project identified using the ADMIXTURE analysis<sup>65</sup> and  
433 used these individuals with high NAT, as well as CEU and YRI from 1000 Genomes as  
434 reference panels. We assumed that the admixture time was 20 generations ago. For ELAI, we  
435 used three upper-layer clusters and 15 lower-layer clusters in the model.

436 To address the technical concerns that local ancestry methods are biased by the high LD of  
437 MHC region<sup>66,67</sup>, we performed an alternative method, RFMix<sup>47</sup>, for local ancestry inference that  
438 accounts for high LD and lack of parental reference panels. Similar deviation from genome-wide  
439 ancestry was observed using RFMix (**Supplementary Figure 14**), indicating that the selection  
440 signals we observed here are robust to different inference methods.

#### 441 HLA imputation in the HIV-1 viral load GWAS data in three populations

442 We used genome-wide genotyping data from 12,023 HIV-1 infected individuals aggregated  
443 across more than 10 different cohorts (**Supplementary Table 10**). The details of these samples  
444 and quality control procedures have been described previously<sup>10,68</sup>. Using the HIV-1 viral load  
445 GWAS data, we extracted the genotypes of SNPs located in the extended MHC region  
446 (chr6:28-34Mb, **Supplementary Table 10**). We conducted genotype imputation of one-field,  
447 two-field and G-group classical *HLA* alleles and amino acid polymorphisms of the eight class I  
448 and class II HLA genes using the constructed multi-ethnic HLA imputation reference panel and  
449 the HLA-TAPAS pipeline.

450 After imputation, we obtained the genotypes of 640 classical alleles, 4,513 amino acid positions  
451 of the eight classical HLA genes, and 49,321 SNPs located in the extended MHC region. We  
452 excluded variants with MAF < 0.5% and imputation  $r^2 < 0.5$  for all association studies. In total,  
453 we tested 51,358 variants in our association and fine-mapping study.

#### 454 HLA association analysis

455 For the HIV-1 viral loads of EUR, AA and LAT samples, we conducted a joint haplotype-based  
456 association analysis using a linear regression model under the assumption of additive effects of  
457 the number of HLA haplotypes for each individual. Phased haplotypes at a locus (i.e., HLA  
458 amino acid position) were constructed from the phased imputed genotypes of variants in the  
459 locus (i.e., amino acid change or SNP) and were converted to a haplotype matrix where each  
460 row is observed haplotypes (in the locus), not genotypes.

461 For each amino acid position, we applied a conditional haplotype analysis. We tested a  
462 multiallelic association between the HIV-1 viral load and a haplotype matrix (of the position) with  
463 covariates, including sex, study-specific PCs, and a categorical variable indicating a population.  
464 That is

$$465 \quad y = \beta_0 + \sum_i^{m-1} \beta_{1i} x_i + \sum_j^C \beta_{2j} c_j,$$

466 where  $x_i$  is the amino acid haplotype formed by each of the  $m$  amino acid residues that occur at  
467 that position, and  $c_j$  are the covariates included in the model.

468 To get an omnibus  $P$ -value for each position, we estimated the effect of each amino acid by  
469 assessing the significance of the improvement in fit by calculating the in-model fit, compared to  
470 a null model following an F-distribution with  $m - 1$  degrees of freedom. This is implemented  
471 using an ANOVA test in R as described previously<sup>32,69</sup>. The most frequent haplotype was  
472 excluded from a haplotype matrix as a reference haplotype for association.

473 **For the conditional analysis**, we assumed that the null model consisted of haplotypes as  
474 defined by residues at previously defined amino acid positions. The alternative model is in  
475 addition of another position with  $m$  residues. We tested whether the addition of those amino  
476 acid positions, and the creation of  $k$  additional haplotypes groups, improved on the previous  
477 set. We then assessed the significance of the improvement in the delta deviance (sum of  
478 squares) over the previous model using an F-test. We performed stepwise conditional analysis  
479 to identify additional independent signals by adjusting for the most significant amino acid  
480 position in each step until none met the significance threshold ( $P = 5 \times 10^{-8}$ ). We restricted



481 analysis to haplotypes that have a minimum of 10 occurrences within HLA-B, and removed any  
482 individual with rare haplotypes for the conditional analysis.

483 **For the exhaustive search**, we tested all possible amino acid pairs and triplets for association.  
484 For each set of amino acid positions, we used the groups of residues occurring at these  
485 positions to estimate effect size and calculated for each of these models the delta deviance in  
486 risk prediction and its p-values compared to the null model.

#### 487 Heterogeneity testing of effect sizes

488 We used interaction analyses with models that included haplotype-by-ancestry (*Haplotype x*  
489 *Ancestry*) interaction terms. The fit of nested models was compared to a null model using the  
490 *F*-statistic with two degrees of freedom, for which the association interaction P-value indicated  
491 whether the inclusion of the *Haplotype x Ancestry* interaction terms improved the model fit  
492 compared to the null model that did not include the interaction terms. Interaction P-values for all  
493 haplotypes formed by positions 97, 67 and 156 in HLA-B are listed in **Supplementary Table 15**.  
494 Haplotypes that had a significant Bonferroni-corrected *Haplotype x Ancestry* interaction  
495 heterogeneity P-value ( $P < 0.05/26$ ) were considered to show evidence of significant effect size  
496 heterogeneity between ancestries.

#### 497 URLs

498 HLA-TAPAS, <https://github.com/immunogenomics/HLA-TAPAS>

499 IMGT/HLA, <https://www.ebi.ac.uk/ipd/imgt/hla/>;

500 GATK version 3.6, <https://software.broadinstitute.org/gatk/download/archive>;

501 HLA\*LA, <https://github.com/DiltheyLab/HLA-PRG-LA>;

502 PLINK 1.90, <https://www.cog-genomics.org/plink2>;  
503 Beagle 4.1, [https://faculty.washington.edu/browning/beagle/b4\\_1.html](https://faculty.washington.edu/browning/beagle/b4_1.html);  
504 Hapl-o-Mat, <https://github.com/DKMS/Hapl-o-Mat/>;  
505 1000 Genomes gold-standard HLA types,  
506 [http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data\\_collections/HLA\\_types/](http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HLA_types/)

## 507 Acknowledgements

508 The study was supported by the National Institutes of Health (NIH) TB Research Unit Network,  
509 Grant U19 AI111224-01.

510 The views expressed in this manuscript are those of the authors and do not necessarily  
511 represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of  
512 Health; or the U.S. Department of Health and Human Services.

513 The Genotype and Phenotype (GaP) Registry at The Feinstein Institute for Medical Research  
514 provided fresh, de-identified human plasma; blood was collected from control subjects under an  
515 IRB-approved protocol (IRB# 09-081) and processed to isolate plasma. The GaP is a  
516 sub-protocol of the Tissue Donation Program (TDP) at Northwell Health and a national resource  
517 for genotype-phenotype studies.  
518 [https://www.feinsteininstitute.org/robert-s-boas-center-for-genomics-and-human-genetics/gap-re-](https://www.feinsteininstitute.org/robert-s-boas-center-for-genomics-and-human-genetics/gap-registry/)  
519 [gistry/](https://www.feinsteininstitute.org/robert-s-boas-center-for-genomics-and-human-genetics/gap-registry/)

520 A.M. is supported by Gentransmed grant 2014-2020.4.01.15-0012. ; D.W.H. is supported by  
521 NIH grants AI110527, AI077505, TR000445, AI069439, and AI110527. D.H.S. was supported by

522 R01 HL92301, R01 HL67348, R01 NS058700, R01 AR48797, R01 DK071891, R01 AG058921,  
523 the General Clinical Research Center of the Wake Forest University School of Medicine (M01  
524 RR07122, F32 HL085989), the American Diabetes Association, and a pilot grant from the  
525 Claude Pepper Older Americans Independence Center of Wake Forest University Health  
526 Sciences (P60 AG10484). J.T.E. and P.E.S. were supported by NIH/NIAMS R01 AR042742,  
527 R01 AR050511, and R01 AR063611.

528 For some HIV cohort participants, DNA and data collection was supported by NIH/NIAID AIDS  
529 Clinical Trial Group (ACTG) grants UM1 AI068634, UM1 AI068636 and UM1 AI106701, and  
530 ACTG clinical research site grants A1069412, A1069423, A1069424, A1069503, AI025859,  
531 AI025868, AI027658, AI027661, AI027666, AI027675, AI032782, AI034853, AI038858,  
532 AI045008, AI046370, AI046376, AI050409, AI050410, AI050410, AI058740, AI060354,  
533 AI068636, AI069412, AI069415, AI069418, AI069419, AI069423, AI069424, AI069428,  
534 AI069432, AI069432, AI069434, AI069439, AI069447, AI069450, AI069452, AI069465,  
535 AI069467, AI069470, AI069471, AI069472, AI069474, AI069477, AI069481, AI069484,  
536 AI069494, AI069495, AI069496, AI069501, AI069501, AI069502, AI069503, AI069511,  
537 AI069513, AI069532, AI069534, AI069556, AI072626, AI073961, RR000046, RR000425,  
538 RR023561, RR024156, RR024160, RR024996, RR025008, RR025747, RR025777, RR025780,  
539 TR000004, TR000058, TR000124, TR000170, TR000439, TR000445, TR000457, TR001079,  
540 TR001082, TR001111, and TR024160.

541 Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by  
542 the National Heart, Lung and Blood Institute (NHLBI). See the TOPMed Omics Support Table  
543 (**Supplementary Table 16**) for study specific omics support information. Core support including

544 centralized genomic read mapping and genotype calling, along with variant quality metrics and  
545 filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1;  
546 contract HHSN268201800002I). Core support including phenotype harmonization, data  
547 management, sample-identity QC, and general program coordination were provided by the  
548 TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract  
549 HHSN268201800001I). We gratefully acknowledge the studies and participants who provided  
550 biological samples and data for TOPMed.

551 The COPDGene project was supported by Award Number U01 HL089897 and Award Number  
552 U01 HL089856 from the National Heart, Lung, and Blood Institute. The content is solely the  
553 responsibility of the authors and does not necessarily represent the official views of the National  
554 Heart, Lung, and Blood Institute or the National Institutes of Health. The COPDGene project is  
555 also supported by the COPD Foundation through contributions made to an Industry Advisory  
556 Board comprised of AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Novartis, Pfizer,  
557 Siemens and Sunovion. A full listing of COPDGene investigators can be found at:  
558 <http://www.copdgene.org/directory>

559 The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson  
560 State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the  
561 Mississippi State Department of Health (HHSN268201800015I) and the University of Mississippi  
562 Medical Center (HHSN268201800010I, HHSN268201800011I and HHSN268201800012I)  
563 contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute  
564 on Minority Health and Health Disparities (NIMHD). The authors also wish to thank the staffs  
565 and participants of the JHS.

566 MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung,  
567 and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is  
568 provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159,  
569 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003,  
570 N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164,  
571 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168,  
572 N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420. MESA Family is conducted  
573 and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with  
574 MESA investigators. Support is provided by grants and contracts R01HL071051,  
575 R01HL071205, R01HL071250, R01HL071251, R01HL071258, R01HL071259, by the National  
576 Center for Research Resources, Grant UL1RR033176. The provision of genotyping data was  
577 supported in part by the National Center for Advancing Translational Sciences, CTSI grant  
578 UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease  
579 Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes  
580 Endocrinology Research Center. This project has been funded in whole or in part with federal  
581 funds from the Frederick National Laboratory for Cancer Research, under Contract No.  
582 HHSN261200800001E. The content of this publication does not necessarily reflect the views or  
583 policies of the Department of Health and Human Services, nor does mention of trade names,  
584 commercial products, or organizations imply endorsement by the U.S. Government. This  
585 Research was supported in part by the Intramural Research Program of the NIH, Frederick  
586 National Lab, Center for Cancer Research.

## 587 Author contributions

588 Y. L. and S.R. conceived, designed and performed analyses, wrote the manuscript and  
589 supervised the research. M.K. implemented the omnibus test for the HIV-1 fine-mapping study.  
590 Y.L., W.C., M.K., P.E.S., J.T.E., and B.H. contributed to the development of the HLA-TAPAS  
591 pipeline. X.L. performed the selection analysis. J.T.E, M.G.-A. and P.K.G helped with the GaP  
592 data acquisition. K.Y., K.O., D.W.H., X.G., N.D.P., Y.I.C., J.I.R., K.D.T., S.S.R., A.C., J.G.W.,  
593 S.K., M.H.C., A.M., T.E., and Y.O. contributed to the WGS data acquisition. J.F., M.C. and P.J.M  
594 contributed to the HIV-1 data acquisition. All authors contributed to the writing of the manuscript.

## 595 Competing interests

596 M.H.C. has received consulting or speaking fees from Illumina and AstraZeneca, and grant  
597 support from GSK and Bayer.

## 598 References

- 599 1. International HIV Controllers Study *et al.* The major genetic determinants of HIV-1 control  
600 affect HLA class I peptide presentation. *Science* **330**, 1551–1557 (2010).
- 601 2. Raychaudhuri, S. *et al.* Five amino acids in three HLA proteins explain most of the  
602 association between MHC and seropositive rheumatoid arthritis. *Nat. Genet.* **44**, 291–296  
603 (2012).
- 604 3. Evans, D. M. *et al.* Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis  
605 implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nat.*  
606 *Genet.* **43**, 761–767 (2011).
- 607 4. Snyder, A. *et al.* Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N.*  
608 *Engl. J. Med.* **371**, 2189–2199 (2014).

- 609 5. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association  
610 studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* vol. 47  
611 D1005–D1012 (2019).
- 612 6. Horton, R. *et al.* Gene map of the extended human MHC. *Nat. Rev. Genet.* **5**, 889–899  
613 (2004).
- 614 7. Gourraud, P.-A. *et al.* HLA diversity in the 1000 genomes dataset. *PLoS One* **9**, e97282  
615 (2014).
- 616 8. Robinson, J. *et al.* IPD-IMGT/HLA Database. *Nucleic Acids Res.* **48**, D948–D955 (2020).
- 617 9. Hu, X. *et al.* Additive and interaction effects at three amino acid positions in HLA-DQ and  
618 HLA-DR molecules drive type 1 diabetes risk. *Nat. Genet.* **47**, 898–905 (2015).
- 619 10. McLaren, P. J. *et al.* Polymorphisms of large effect explain the majority of the host genetic  
620 contribution to variation of HIV-1 virus load. *Proc. Natl. Acad. Sci. U. S. A.* **112**,  
621 14658–14663 (2015).
- 622 11. Tian, C. *et al.* Genome-wide association and HLA region fine-mapping studies identify  
623 susceptibility loci for multiple common infections. *Nat. Commun.* **8**, 599 (2017).
- 624 12. Onengut-Gumuscu, S. *et al.* Type 1 Diabetes Risk in African-Ancestry Participants and  
625 Utility of an Ancestry-Specific Genetic Risk Score. *Diabetes Care* **42**, 406–415 (2019).
- 626 13. Matthews, P. C. *et al.* Central role of reverting mutations in HLA associations with human  
627 immunodeficiency virus set point. *J. Virol.* **82**, 8548–8559 (2008).
- 628 14. Wang, Y. Development of a human leukocyte antigen-based HIV vaccine. *F1000Res.* **7**,  
629 (2018).
- 630 15. WHO | Progress reports on HIV. (2020).
- 631 16. McLaren, P. J. *et al.* Fine-mapping classical HLA variation associated with durable host  
632 control of HIV-1 infection in African Americans. *Hum. Mol. Genet.* **21**, 4334–4347 (2012).



- 633 17. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program.  
634 *bioRxiv* 563866 (2019) doi:10.1101/563866.
- 635 18. Okada, Y. *et al.* Deep whole-genome sequencing reveals recent selection signatures linked  
636 to evolution and disease risk of Japanese. *Nat. Commun.* **9**, 1631 (2018).
- 637 19. Mitt, M. *et al.* Improved imputation accuracy of rare and low-frequency variants using  
638 population-specific high-coverage WGS-based imputation reference panel. *Eur. J. Hum.*  
639 *Genet.* **25**, 869–876 (2017).
- 640 20. Hirata, J. *et al.* Genetic and phenotypic landscape of the major histocompatibility complex  
641 region in the Japanese population. *Nat. Genet.* (2019) doi:10.1038/s41588-018-0336-0.
- 642 21. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation.  
643 *Nature* **526**, 68–74 (2015).
- 644 22. Nelis, M. *et al.* Genetic structure of Europeans: a view from the north–east. *PLoS One* **4**,  
645 (2009).
- 646 23. Dilthey, A., Cox, C., Iqbal, Z., Nelson, M. R. & McVean, G. Improved genome inference in  
647 the MHC using a population reference graph. *Nature Genetics* vol. 47 682–688 (2015).
- 648 24. Dilthey, A. T. *et al.* High-Accuracy HLA Type Inference from Whole-Genome Sequencing  
649 Data Using Population Reference Graphs. *PLoS Comput. Biol.* **12**, e1005151 (2016).
- 650 25. Dilthey, A. T. *et al.* HLA\*LA-HLA typing from linearly projected graph alignments.  
651 *Bioinformatics* **35**, 4394–4396 (2019).
- 652 26. Browning, B. L. & Browning, S. R. A fast, powerful method for detecting identity by descent.  
653 *Am. J. Hum. Genet.* **88**, 173–182 (2011).
- 654 27. Hill, A. V. *et al.* Common west African HLA antigens are associated with protection from  
655 severe malaria. *Nature* **352**, 595–600 (1991).
- 656 28. Sanchez-Mazas, A. *et al.* The HLA-B landscape of Africa: Signatures of pathogen-driven

- 657 selection and molecular identification of candidate alleles to malaria protection. *Mol. Ecol.*  
658 **26**, 6238–6252 (2017).
- 659 29. Maiers, M., Gragert, L. & Klitz, W. High-resolution HLA alleles and haplotypes in the United  
660 States population. *Hum. Immunol.* **68**, 779–788 (2007).
- 661 30. Gonzalez-Galarza, F. F. *et al.* Allele frequency net database (AFND) 2020 update:  
662 gold-standard data classification, open access genotype data and new query tools. *Nucleic*  
663 *Acids Res.* **48**, D783–D788 (2020).
- 664 31. Nothnagel, M., Fürst, R. & Rohde, K. Entropy as a measure for linkage disequilibrium over  
665 multilocus haplotype blocks. *Hum. Hered.* **54**, 186–198 (2002).
- 666 32. Okada, Y. *et al.* Construction of a population-specific HLA imputation reference panel and  
667 its application to Graves' disease risk in Japanese. *Nat. Genet.* **47**, 798–802 (2015).
- 668 33. Okada, Y. eLD: entropy-based linkage disequilibrium index between multiallelic sites. *Hum*  
669 *Genome Var* **5**, 29 (2018).
- 670 34. Chikata, T. *et al.* Host-specific adaptation of HIV-1 subtype B in the Japanese population. *J.*  
671 *Virology* **88**, 4764–4775 (2014).
- 672 35. Nomura, E. *et al.* Mapping of a disease susceptibility locus in chromosome 6p in Japanese  
673 patients with ulcerative colitis. *Genes Immun.* **5**, 477–483 (2004).
- 674 36. Price, P. *et al.* The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8,  
675 DR3) with multiple immunopathological diseases. *Immunol. Rev.* **167**, 257–274 (1999).
- 676 37. Horton, R. *et al.* Variation analysis and gene annotation of eight MHC haplotypes: the MHC  
677 Haplotype Project. *Immunogenetics* **60**, 1–18 (2008).
- 678 38. Graham, R. R. *et al.* Visualizing human leukocyte antigen class II risk haplotypes in human  
679 systemic lupus erythematosus. *Am. J. Hum. Genet.* **71**, 543–553 (2002).
- 680 39. Miller, F. W. *et al.* Genome-wide association study identifies HLA 8.1 ancestral haplotype

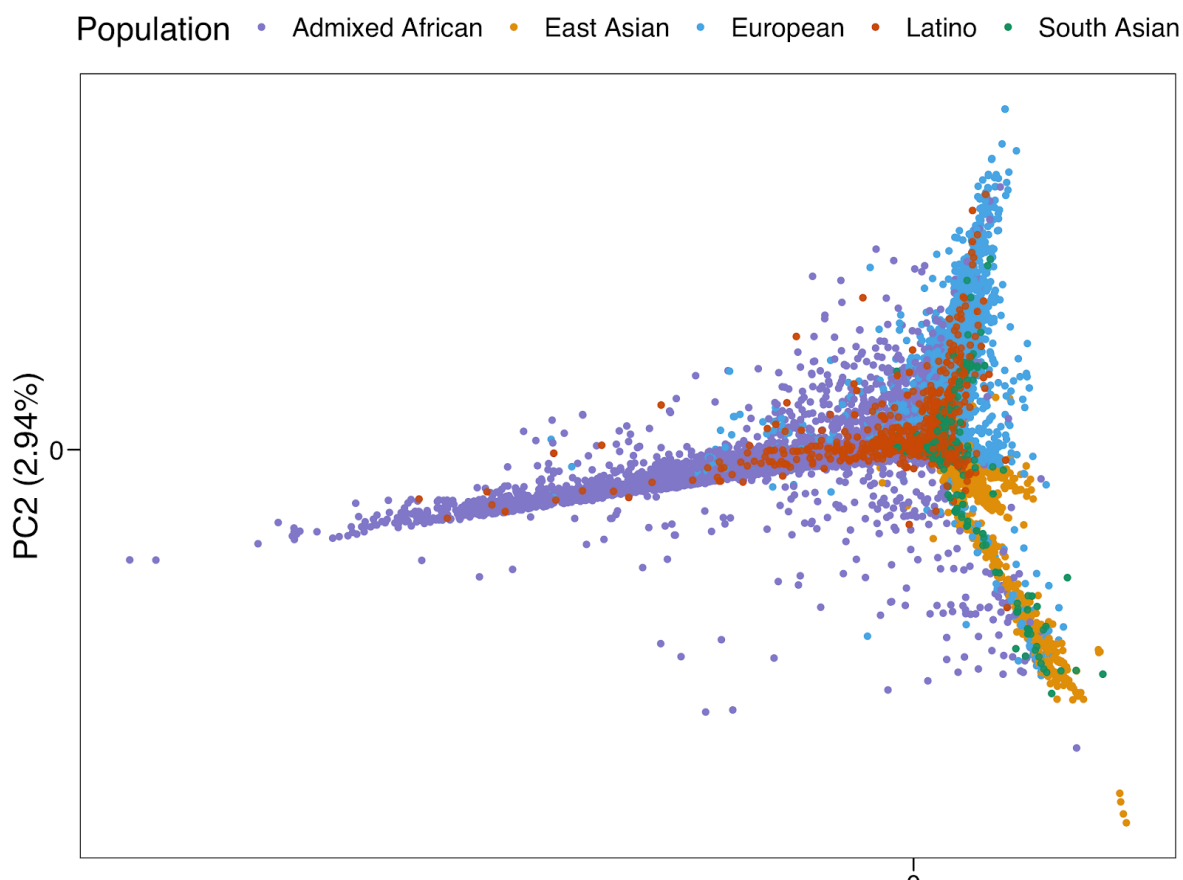
- 681 alleles as major genetic risk factors for myositis phenotypes. *Genes Immun.* **16**, 470–480  
682 (2015).
- 683 40. Haapasalo, K. *et al.* The Psoriasis Risk Allele HLA-C\*06:02 Shows Evidence of Association  
684 with Chronic or Recurrent Streptococcal Tonsillitis. *Infect. Immun.* **86**, (2018).
- 685 41. Kløverpris, H. N. *et al.* HIV control through a single nucleotide on the HLA-B locus. *J. Virol.*  
686 **86**, 11493–11500 (2012).
- 687 42. Salter-Townshend, M. & Myers, S. Fine-Scale Inference of Ancestry Segments Without  
688 Prior Knowledge of Admixing Groups. *Genetics* **212**, 869–889 (2019).
- 689 43. Zhou, Q., Zhao, L. & Guan, Y. Strong Selection at MHC in Mexicans since Admixture. *PLoS*  
690 *Genet.* **12**, e1005847 (2016).
- 691 44. Meyer, D., C Aguiar, V. R., Bitarello, B. D., C Brandt, D. Y. & Nunes, K. A genomic  
692 perspective on HLA evolution. *Immunogenetics* **70**, 5–27 (2018).
- 693 45. Norris, E. T. *et al.* Admixture-enabled selection for rapid adaptive evolution in the Americas.  
694 *bioRxiv* 783845 (2019) doi:10.1101/783845.
- 695 46. Guan, Y. Detecting structure of haplotypes and local ancestry. *Genetics* **196**, 625–642  
696 (2014).
- 697 47. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative  
698 modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**,  
699 278–288 (2013).
- 700 48. Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One*  
701 **8**, e64683 (2013).
- 702 49. Mellors, J. W. *et al.* Quantitation of HIV-1 RNA in plasma predicts outcome after  
703 seroconversion. *Ann. Intern. Med.* **122**, 573–579 (1995).
- 704 50. Bartha, I. *et al.* Estimating the Respective Contributions of Human and Viral Genetic

- 705 Variation to HIV Control. *PLoS Comput. Biol.* **13**, e1005339 (2017).
- 706 51. Blanco-Gelaz, M. A. *et al.* The amino acid at position 97 is involved in folding and surface  
707 expression of HLA-B27. *Int. Immunol.* **18**, 211–220 (2006).
- 708 52. Stewart-Jones, G. B. E. *et al.* Structures of Three HIV-1 HLA-B\*5703-Peptide Complexes  
709 and Identification of Related HLAs Potentially Associated with Long-Term Nonprogression.  
710 *The Journal of Immunology* vol. 175 2459–2468 (2005).
- 711 53. Archbold, J. K. *et al.* Natural micropolymorphism in human leukocyte antigens provides a  
712 basis for genetic control of antigen recognition. *J. Exp. Med.* **206**, 209–219 (2009).
- 713 54. Gaiha, G. D. *et al.* Structural topology defines protective CD8+ T cell epitopes in the HIV  
714 proteome. *Science* **364**, 480–484 (2019).
- 715 55. Macdonald, W. A. *et al.* A naturally selected dimorphism within the HLA-B44 supertype  
716 alters class I structure, peptide repertoire, and T cell recognition. *J. Exp. Med.* **198**,  
717 679–691 (2003).
- 718 56. Klooverpris, H. N. *et al.* HLA-B\*57 Micropolymorphism Shapes HLA Allele-Specific Epitope  
719 Immunogenicity, Selection Pressure, and HIV Immune Control. *Journal of Virology* vol. 86  
720 919–929 (2012).
- 721 57. Carrington, M. & Walker, B. D. Immunogenetics of spontaneous control of HIV. *Annu. Rev.*  
722 *Med.* **63**, 131–145 (2012).
- 723 58. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals  
724 with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
- 725 59. Khera, A. V. *et al.* Polygenic Prediction of Weight and Obesity Trajectories from Birth to  
726 Adulthood. *Cell* **177**, 587–596.e9 (2019).
- 727 60. Torkamani, A. & Topol, E. Polygenic Risk Scores Expand to Obesity. *Cell* vol. 177 518–520  
728 (2019).

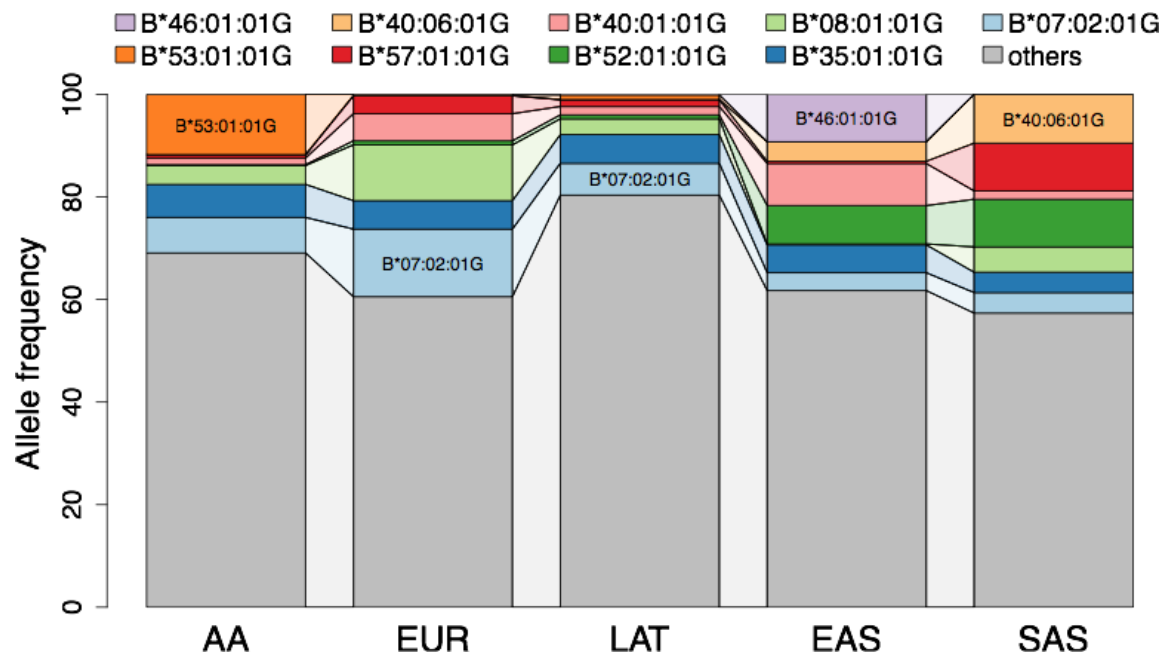
- 729 61. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health  
730 disparities. *Nat. Genet.* **51**, 584–591 (2019).
- 731 62. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome  
732 Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–33 (2013).
- 733 63. Julg, B. *et al.* Possession of HLA class II DRB1\*1303 associates with reduced viral loads in  
734 chronic HIV-1 clade C and B infection. *J. Infect. Dis.* **203**, 803–809 (2011).
- 735 64. Schäfer, C., Schmidt, A. H. & Sauter, J. Hapl-o-Mat: open-source software for HLA  
736 haplotype frequency estimation from ambiguous and heterogeneous data. *BMC*  
737 *Bioinformatics* **18**, 284 (2017).
- 738 65. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in  
739 unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
- 740 66. Price, A. L. *et al.* Long-range LD can confound genome scans in admixed populations.  
741 *American journal of human genetics* vol. 83 132–5; author reply 135–9 (2008).
- 742 67. Pasaniuc, B. *et al.* Analysis of Latino populations from GALA and MEC studies reveals  
743 genomic loci with biased local ancestry estimation. *Bioinformatics* **29**, 1407–1415 (2013).
- 744 68. McLaren, P. J. *et al.* Association study of common genetic variants and HIV-1 acquisition in  
745 6,300 infected cases and 7,200 controls. *PLoS Pathog.* **9**, e1003515 (2013).
- 746 69. Okada, Y. *et al.* Contribution of a Non-classical HLA Gene, HLA-DOA, to the Risk of  
747 Rheumatoid Arthritis. *Am. J. Hum. Genet.* **99**, 366–374 (2016).

## 748 Figures

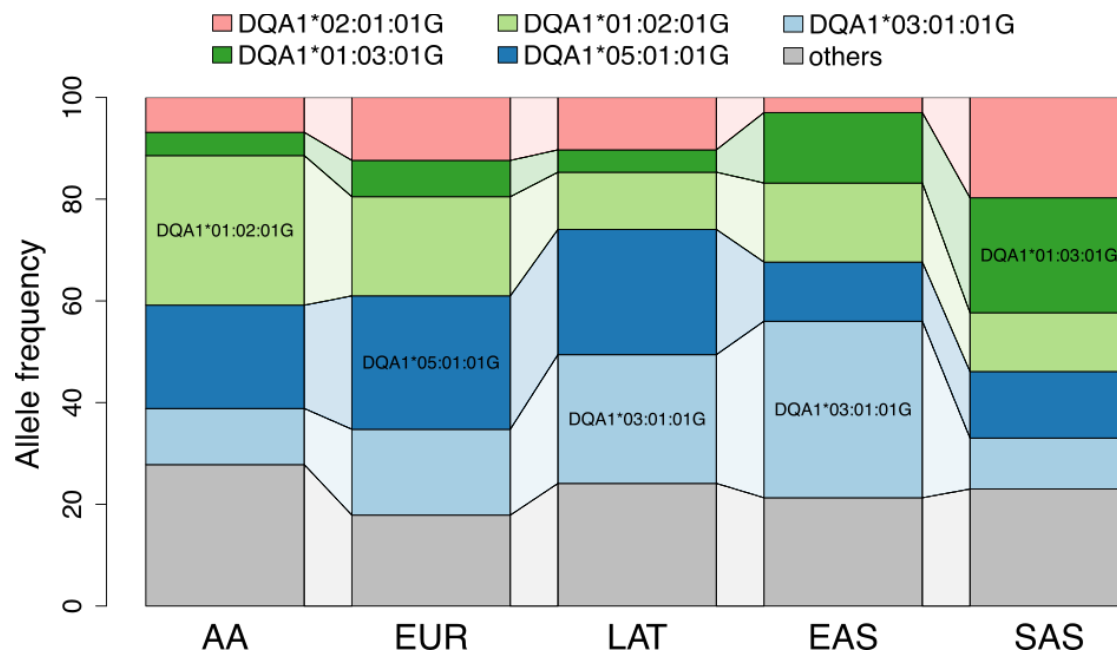
749 **Figure 1. Global diversity of the MHC region.** (a) Principal component analysis of the  
750 pairwise IBD distance between 21,546 samples using MHC region markers. Allele diversity of  
751 (b) HLA-B and (c) HLA-DQA1 among five continental populations (AA=Admixed African;  
752 EUR=European; LAT=Latino; EAS=East Asian; SAS=South Asian). The top two most common  
753 alleles within each population group are named, the remaining alleles are grouped as 'others'.  
754 (a)



755 (b)

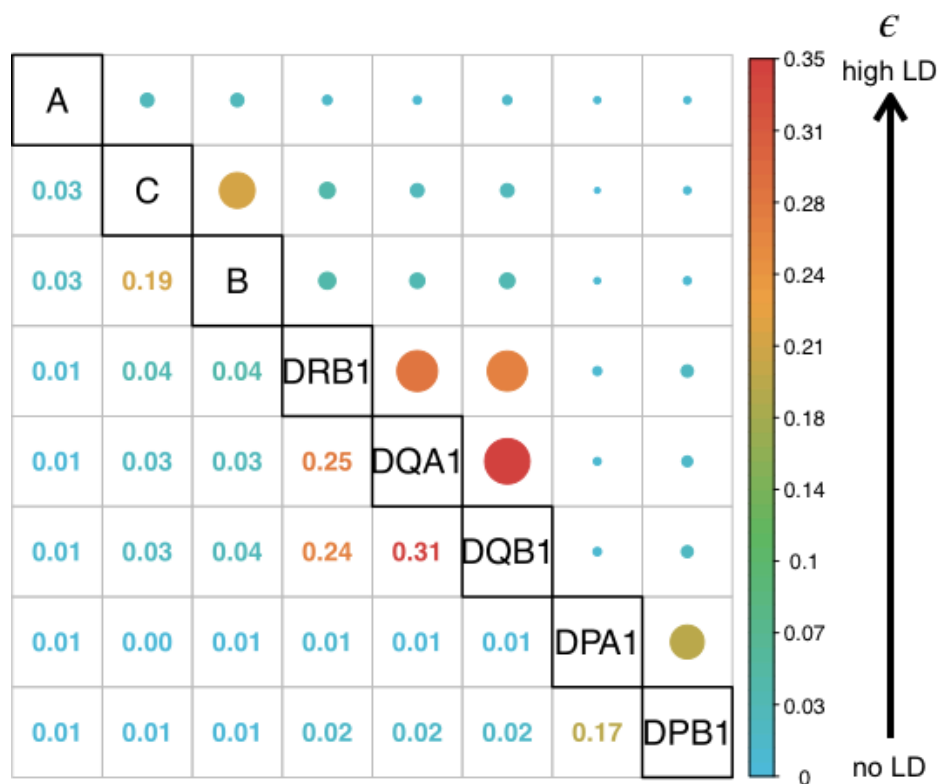


756 (c)

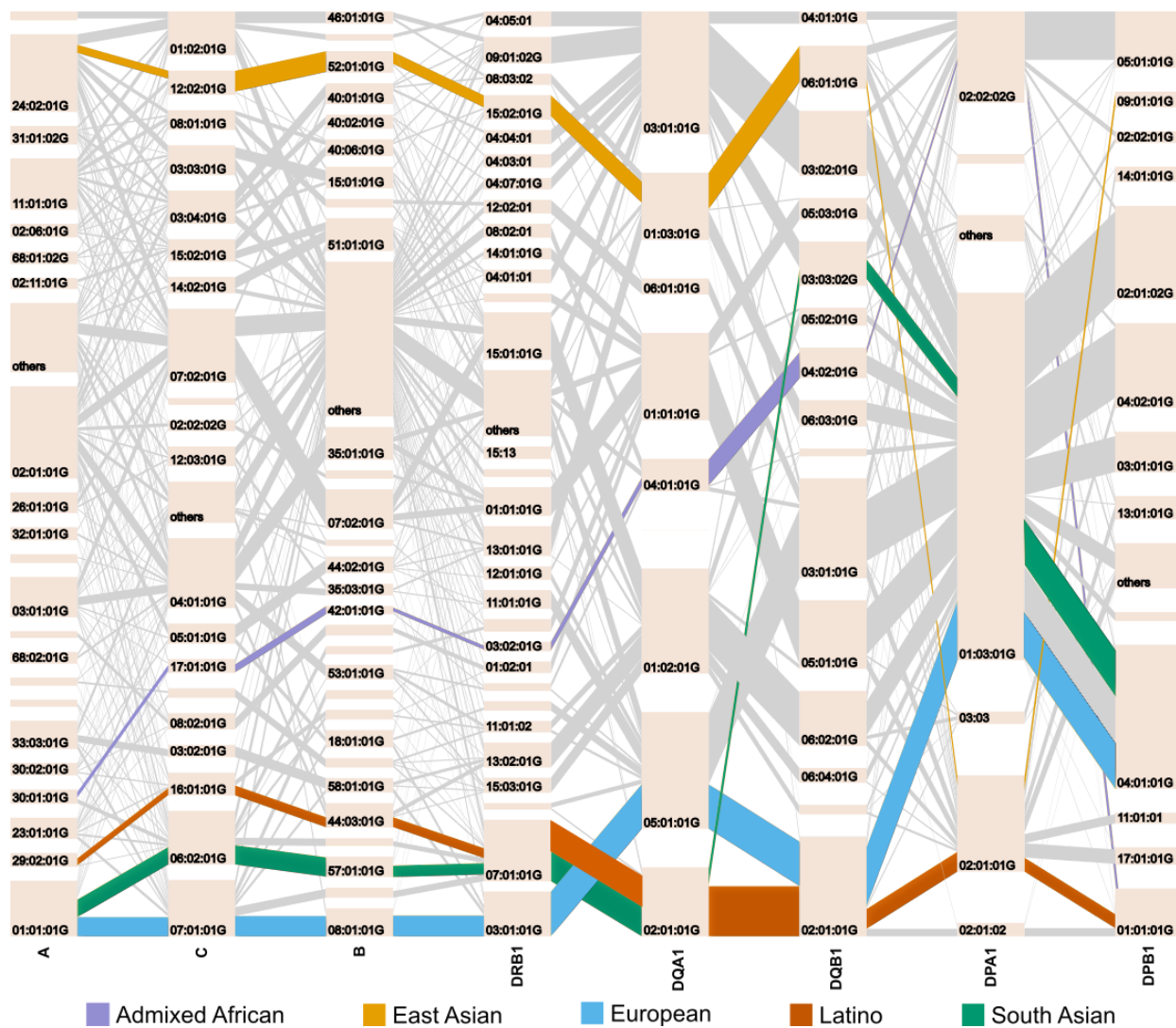




757 **Figure 2. Pairwise LD and haplotype structure for six classical HLA genes in five**  
 758 **population groups.** (a) shows the pairwise normalized entropy ( $\epsilon$ ) measuring the difference of  
 759 the haplotype frequency distribution for linkage disequilibrium and linkage equilibrium among  
 760 five population groups. It takes values between 0 (no LD) to 1 (perfect LD). (b) shows the  
 761 haplotype structures of the eight classical HLA genes in each population. The tile in a bar  
 762 represents an *HLA* allele, and its height corresponds to the frequencies of the *HLA* allele. The  
 763 gray lines connecting between two alleles represent *HLA* haplotypes. The width of these lines  
 764 corresponds to the frequencies of the haplotypes. The most frequent long-range *HLA*  
 765 haplotypes within each population is bolded and highlighted in a color described by the key at  
 766 the bottom.  
 767 (a)

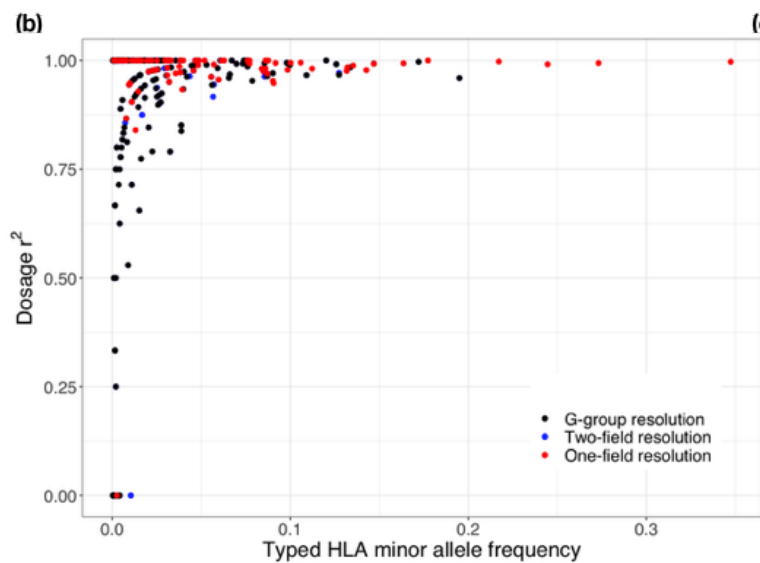
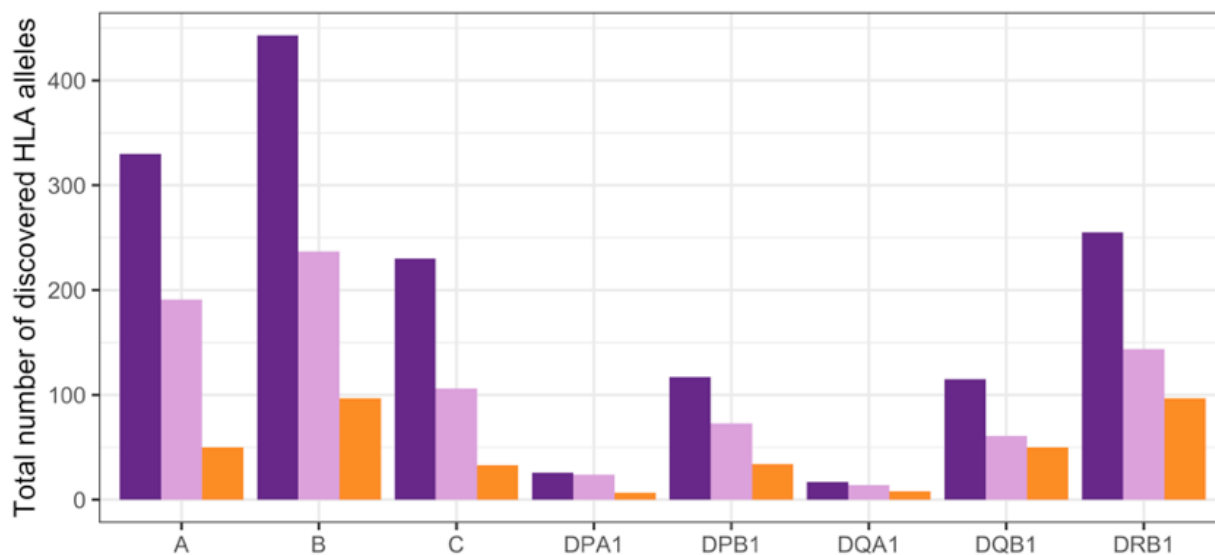


768 (b)



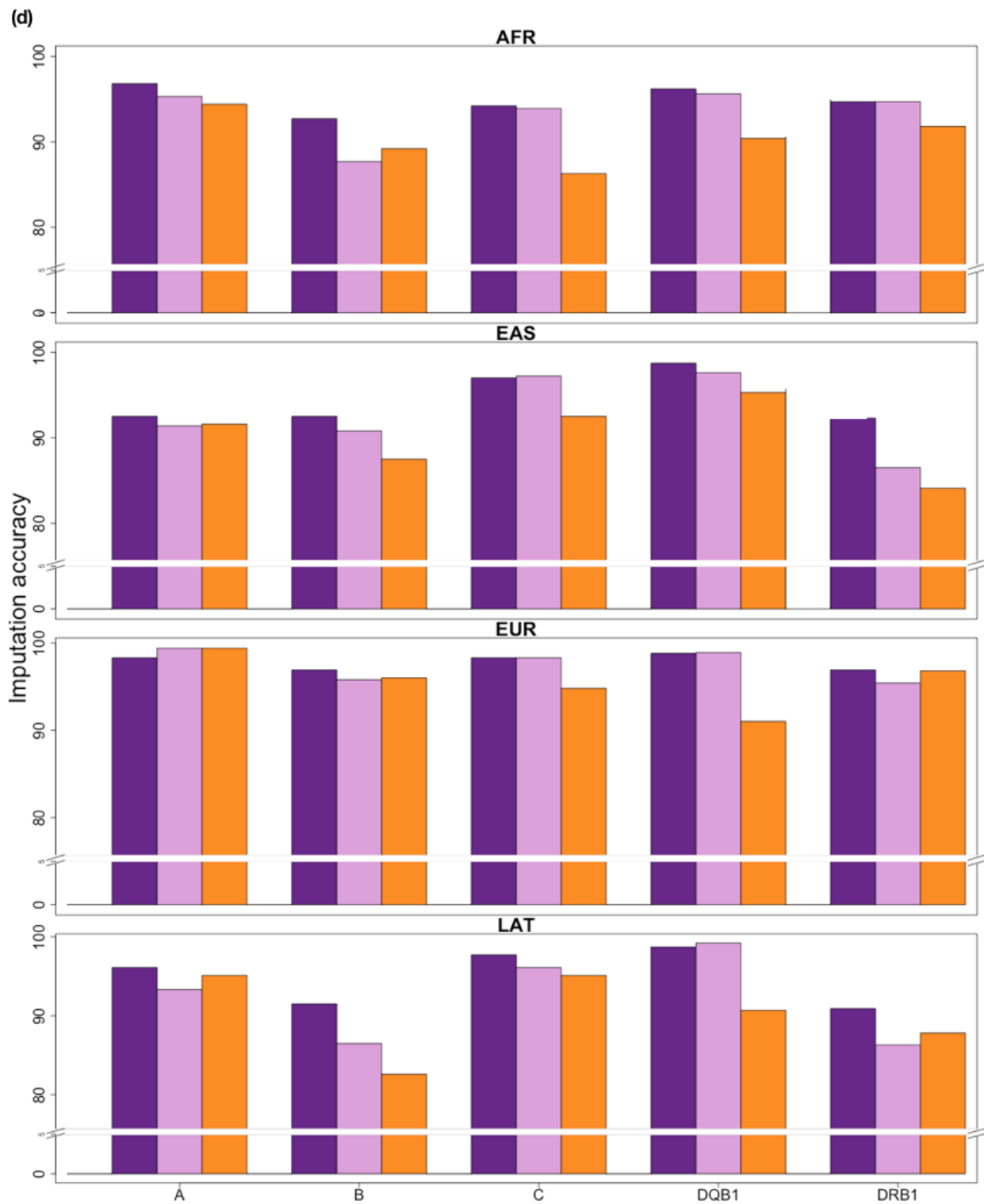
769 **Figure 3. The multi-ethnic HLA reference panel shows improvement in allele diversity and**  
770 **imputation accuracy. (a).** The number of HLA alleles at the two-field resolution included in the  
771 multi-ethnic HLA reference panel (N = 21,546) compared to the European only Type 1 Diabetes  
772 Genetics Consortium<sup>48</sup> (T1DGC) panel (N = 5,225) as well as a subset of the multi-ethnic HLA  
773 panel down-sampled to the same size as T1DGC. **(b).** The correlation between imputed and  
774 typed dosages of classical *HLA* alleles using the multi-ethnic HLA reference panel at one-field  
775 (red), two-field (blue) and G-group resolution (black) of the 955 1000 Genomes subjects. **(c).**  
776 The imputation accuracy for five classical HLA genes at one-field, two-field and G-group  
777 resolution. **(d).** The imputation accuracy at G-group resolution of the 1000 Genomes subjects  
778 stratified by four diverse ancestries when using three different imputation reference panels as  
779 described in **(a)**.

(a) Reference panels **Multi-ethnic WGS (21,546)** **Down-sampled WGS (5,225)** **T1DGC (5,225)**

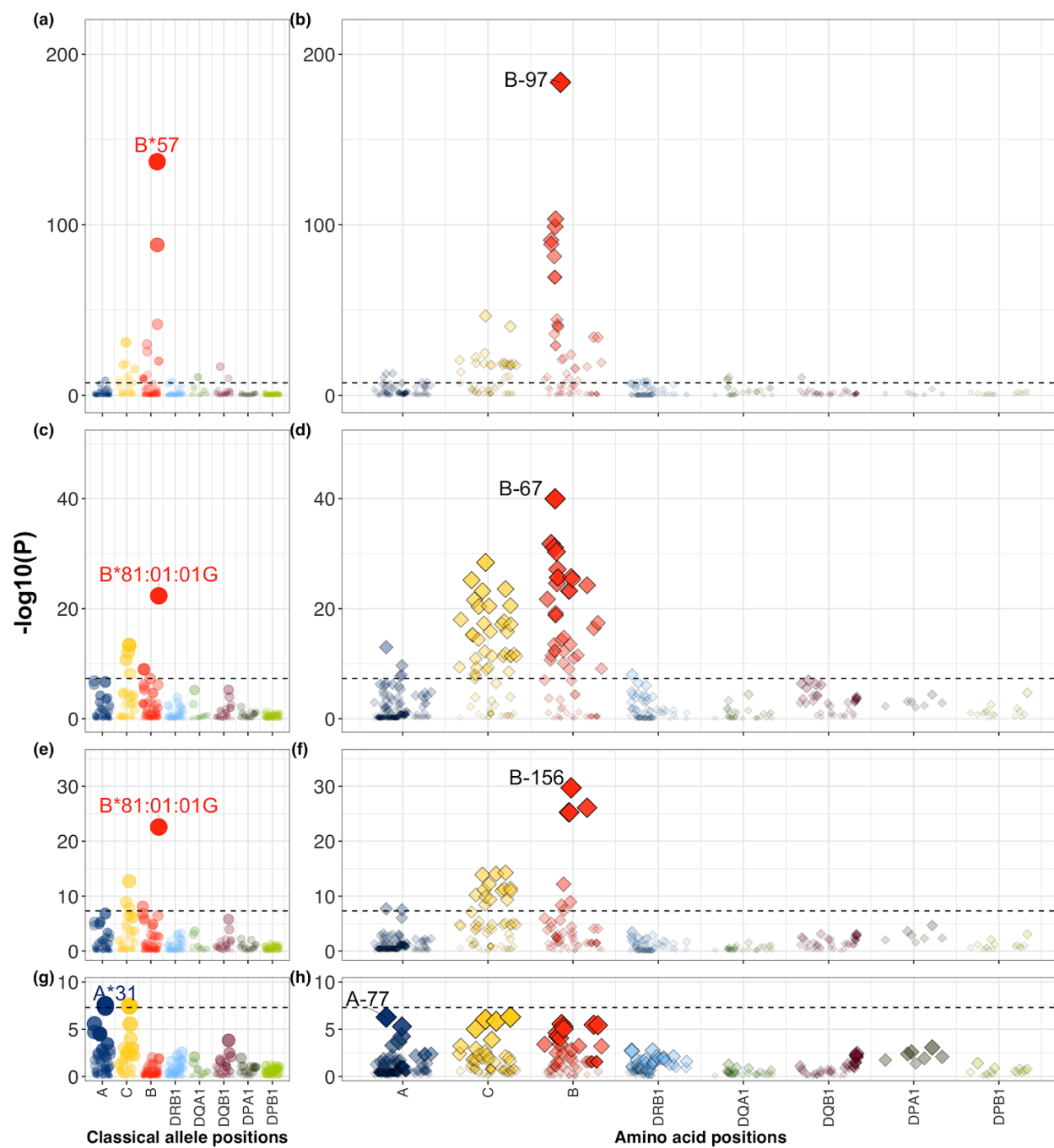


(c)

	one-field	two-field	G-group
<b>A</b>	0.992	0.964	0.963
<b>B</b>	0.972	0.937	0.930
<b>C</b>	0.995	0.969	0.981
<b>DQB1</b>	0.993	0.981	0.933
<b>DRB1</b>	0.976	0.938	0.934

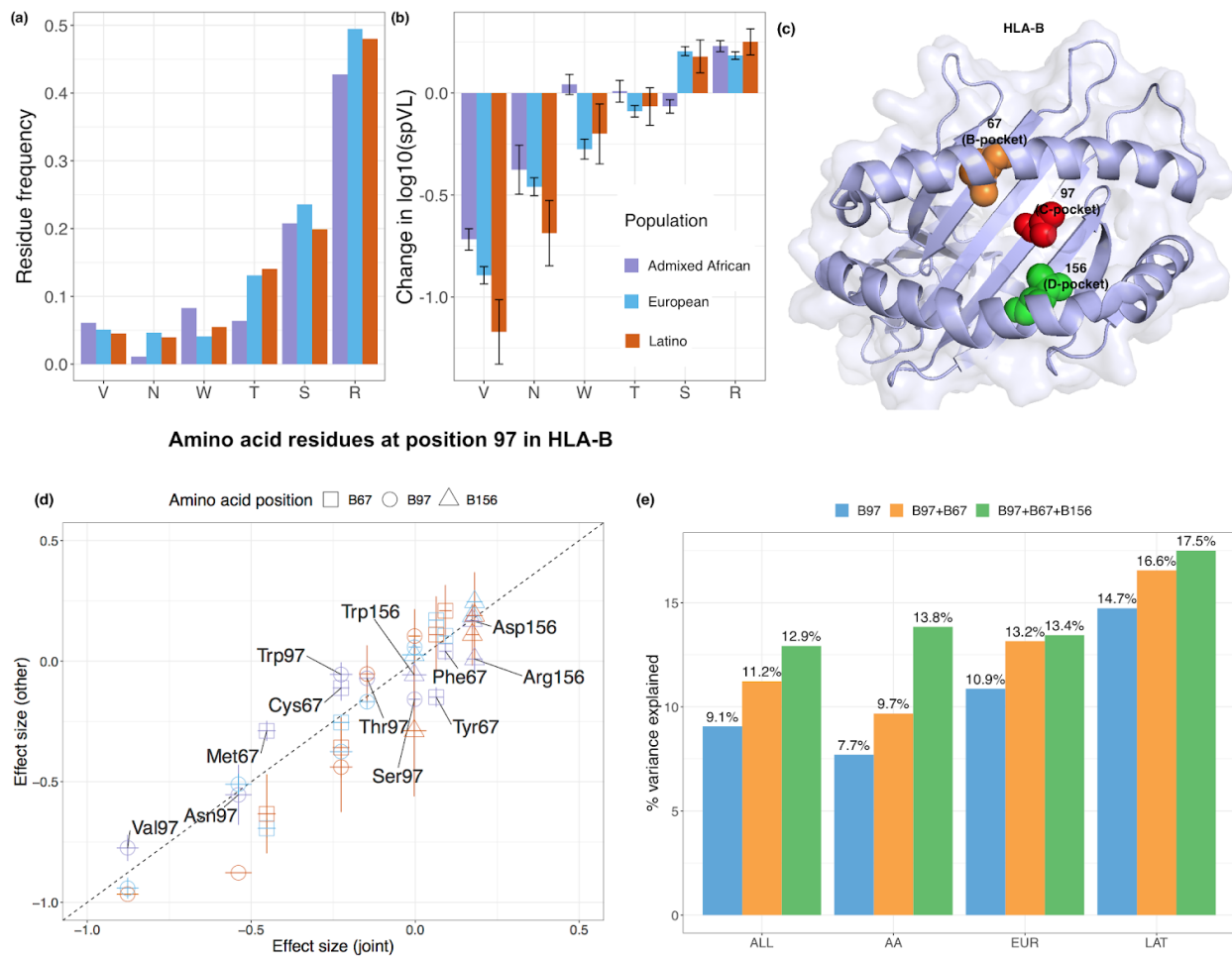


780 **Figure 4. Stepwise conditional analysis of the allele and amino acid positions of classical**  
781 **HLA genes to HIV-1 viral load.** Each circle point represents the linear regression  $-\log_{10}($   
782  $P_{binary})$  for all classical *HLA* alleles. Each diamond point represents  $-\log_{10}(P_{omnibus})$  for the  
783 tested amino acid positions in HLA (blue=HLA-A; yellow=HLA-C; red=HLA-B;  
784 lightblue=HLA-*DRB1*; green=HLA-*DQA1*; purple=HLA-*DQB1*, darkgreen=HLA-*DPA1*;  
785 lightgreen=HLA-*DPB1*). Association at amino acid positions with more than two alleles was  
786 calculated using a multi-degree-of-freedom omnibus test. The dashed blacked line represents  
787 the significance threshold of  $P = 5 \times 10^{-8}$ . Each panel shows the association plot in the process  
788 of stepwise conditional omnibus test. (a) One-field classical allele *B\*57* ( $P = 9.84 \times 10^{-138}$ ) and  
789 (b) amino acid position 97 in HLA-B ( $P_{omnibus} = 2.86 \times 10^{-184}$ ) showed the strongest association  
790 signal. Results conditioned on position 97 in HLA-B showed a secondary signal at (c) classical  
791 allele *B\*81:0101:G* ( $P = 4.53 \times 10^{-23}$ ) and (d) position 67 in HLA-B ( $P_{omnibus} = 1.08 \times 10^{-40}$ ).  
792 Results conditioned on position 97 and 67 in HLA-B showed the same classical allele (e)  
793 *B\*81:0101G* ( $P = 2.70 \times 10^{-23}$ ) and (f) third signal at position 156 in HLA-B (  
794  $P_{omnibus} = 1.92 \times 10^{-30}$ ). Results conditioned on position 97, 67 and 156 int HLA-B showed a  
795 fourth signal at (g) HLA-A\*31 ( $P = 2.45 \times 10^{-8}$ ) and (h) position 77 in HLA-A (  
796  $P_{omnibus} = 5.35 \times 10^{-7}$ ) outside HLA-B.



797 **Figure 5. Location and effect of three independently associated amino acid positions in**  
798 **HLA-B. (a)** Allele frequency of six residues at position 97 in HLA-B among three populations.  
799 **(b)** Effect on spVL (i.e., change in log<sub>10</sub> HIV-1 spVL per allele copy) of individual amino acid  
800 residues at position 97 in HLA-B. Results were calculated per allele using linear regression  
801 models, including gender and principal components within each ancestry as covariates. **(c)**  
802 HLA-B (PDB ID code 2bvp) proteins. Omnibus and stepwise conditional analysis identified three  
803 independent amino acid positions (positions 97 (red), 67 (orange), and 156 (green) in HLA-B.  
804 **(d)** Effect on spVL (i.e., change in log<sub>10</sub> HIV-1 spVL per allele copy) of individual amino acid  
805 residues at each position reported in this and previous work<sup>10,16</sup>. Results were calculated per  
806 allele using linear regression models. The x-axis shows the effect size and its standard errors in  
807 the joint analysis, and the y-axis shows the effect size and its standard error in individual  
808 populations (purple = Admixed American; blue = European and orange = Latino). **(e)** Variance of  
809 spVL explained by the haplotypes formed by different amino acid positions.





## 810 Tables

811 **Table 1. Effect estimates for the haplotypes defined by the three independent amino**  
 812 **acids in HLA-B associated with HIV-1 viral load.** Only haplotypes with >1% frequency in the  
 813 overall population are listed (**Supplementary Table 15**). Classical alleles of HLA-B are grouped  
 814 based on the amino acid residues presented at position 97, 67 and 156 in HLA-B. For each  
 815 haplotype, the multivariate effect is given as an effect size, taking the most frequent haplotype  
 816 (97R-67S-156L) as the reference (effect size = 0). Heterogeneity p-value (P(het)) of each  
 817 haplotype is calculated using a F-statistics with two degrees of freedom (**Methods**). Effect size  
 818 and its standard error in each population are listed only for haplotypes that show evidence of  
 819 heterogeneity (P-value < 0.05 /26, bolded). Unadjusted haplotype frequencies are given in each  
 820 population.

HLA-B amino acid at position			Effect size (standard error)				P(het)	Unadjusted allele frequency				HLA-B allele
97	67	156	AA	EUR	LAT	Joint		AA	EUR	LAT	Joint	
V	M	L				-0.921 (0.036)	0.031	0.056	0.049	0.059	0.051	<i>B*57:01;B*57:03</i>
N	C	L				-0.554 (0.041)	0.257	0.012	0.046	0.037	0.035	<i>B*27:05</i>
T	S	L				-0.436 (0.041)	0.041	0.028	0.039	0.056	0.037	<i>B*13:02;B*52:01</i>
W	C	L				-0.397 (0.041)	0.581	0.03	0.039	0.054	0.037	<i>B*14:01;B*14:02</i>
S	S	L				-0.252 (0.066)	0.013	0.002	0.014	0.07	0.013	<i>B*40:02</i>
R	S	W				-0.177 (0.038)	0.618	0.009	0.062	0.028	0.044	<i>15:10;B*15:16</i>
T	F	L				-0.125 (0.036)	0.001	0.03	0.059	0.073	0.051	<i>B*51:01;B*78:01</i>
R	M	L				-0.125 (0.045)	0.375	0.061	0.014	0.028	0.029	<i>B*15:16;B*58:01</i>
R	C	L				-0.078 (0.039)	0.055	0.042	0.039	0.06	0.041	<i>15:16;B*39:10</i>
R	S	D	0.165 (0.056)	-0.07 (0.034)	-0.153 (0.173)	-0.019 (0.028)	<b>0.002</b>	0.075	0.108	0.084	0.097	<i>44:02;B*45:01</i>
R	S	L				<b>Reference</b>	0.536	0.191	0.176	0.197	0.18	<i>15:10;B*18:01;B*39:10;B*42:01;B*42:02</i>
S	Y	D				0.015 (0.055)	0.884	0.059	NA	0.017	0.019	<i>B*07:02;B*07:05</i>
S	Y	R	-0.06 (0.055)	0.037 (0.033)	-0.002 (0.187)	0.022 (0.027)	<b>0.007</b>	0.08	0.124	0.07	0.108	
S	F	D				0.041 (0.031)	0.218	0.034	0.095	0.042	0.074	<i>B*08:01</i>
R	F	L				0.045 (0.027)	0.73	0.182	0.095	0.113	0.122	<i>B*35:01;B*53:01</i>
W	M	L				0.098 (0.064)	0.268	0.046	NA	NA	0.014	<i>B*58:02</i>
T	Y	L				0.176 (0.058)	0.207	0.005	0.021	NA	0.016	