

1        **The Pathogen Adaptation of HLA Alleles and the Correlation with**  
2        **Autoimmune Diseases based on the HLA Diversity Resource in the**  
3        **Han Chinese**

4        Shuai Liu<sup>1,3#</sup>, Yanyan Li<sup>1#</sup>, Tingrui Song<sup>1#</sup>, Jingjing Zhang<sup>1,2</sup>, Peng Zhang<sup>1</sup>, Huaxia  
5        Luo<sup>1</sup>, Sijia Zhang<sup>1,3</sup>, Yiwei Niu<sup>1,3</sup>, Tao Xu<sup>4,5\*</sup>, Shunmin He<sup>1,3\*</sup>

6        1 Key Laboratory of Epigenetic Regulation and Intervention, Institute of Biophysics,  
7        Chinese Academy of Sciences, Beijing 100101, China.

8        2 University of Chinese Academy of Sciences, Beijing 100049, China.

9        3 College of Life Sciences, University of Chinese Academy of Sciences, Beijing  
10       100049, China.

11       4 National Laboratory of Biomacromolecules, CAS Centre for Excellence in  
12       Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing  
13       100101, China.

14       5 Shandong First Medical University & Shandong Academy of Medical Sciences,  
15       Jinan 250117 Shandong, China.

16       # These authors contributed equally: Shuai Liu, Yanyan Li, Tingrui Song

17       \*Corresponding Authors: Tao Xu, Shunmin He

18       E-mail: [xutao@ibp.ac.cn](mailto:xutao@ibp.ac.cn); [heshunmin@ibp.ac.cn](mailto:heshunmin@ibp.ac.cn)

19

20       **ABSTRACT**

21       Human leukocyte antigen (HLA) genes play a crucial role in the adaptation of  
22       human populations to the dynamic pathogenic environment. Despite their significance,  
23       investigating the pathogen-driven evolution of HLAs and the implications for  
24       autoimmune diseases presents considerable challenges. Here, we genotyped over  
25       twenty HLA genes at 3-field resolution in 8278 individuals from diverse ethnic  
26       backgrounds, including 4013 unrelated Han Chinese. We focused on the adaptation of  
27       HLAs in the Han Chinese by analysing their binding affinity for various pathogens,  
28       and explored the potential correlations between pathogen adaptation and autoimmune

29 diseases. Our findings reveal that specific HLA alleles like HLA-DRB1\*07:01 and  
30 HLA-DQB1\*06:01, confer strong pathogen adaptability at the sequence level, notably  
31 for *Corynebacterium diphtheriae* and *Bordetella pertussis*. Additionally, alleles like  
32 HLA\*03:02 demonstrate adaptive selection against pathogens like *Mycobacterium*  
33 *tuberculosis* and *Coronavirus* at the gene expression level. Simultaneously, the  
34 aforementioned HLA alleles are closely related to some autoimmune diseases such as  
35 multiple sclerosis (MS). These exploratory discoveries shed light on the intricate  
36 coevolutionary relationships between pathogen adaptation and autoimmune diseases  
37 in the human population. These efforts led to an HLA database at  
38 <http://bigdata.ibp.ac.cn/HLAtyping>, aiding searches for HLA allele frequencies (AF)  
39 across populations.

40 KEYWORDS: HLA genotypes; Han Chinese; Pathogen adaptation; Autoimmune  
41 disease

## 42 **Introduction**

43 The immune process is of paramount importance in the intricate interplay  
44 between the *in vivo* life system and the surrounding pathogenic microbial  
45 environment [1]. The HLA genes located on chromosomal region 6p21.3, in particular  
46 the classical HLA genes (*HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DPA1*,  
47 *HLA-DPBI*, *HLA-DQA1*, *HLA-DQB1*), play a pivotal role in the identification and  
48 presentation of invasive foreign pathogens to the immune system [2]. As a result,  
49 these classical HLA genes are under intense selective pressure for resistance against  
50 pathogens, resulting in the highest degree of polymorphism observed in the human  
51 genome [3,4]. To date, two primary mechanistic theories, negative  
52 frequency-dependent selection (NFDS) and heterozygote advantage (HA), have been  
53 proposed to explain the remarkable diversity of HLA genes in the fluctuating  
54 pathogenic microbial environment [5]. Genetic predisposition to different pathogens  
55 is an area of increasing interest [6]. Although many human HLA resources have been  
56 developed [7–14], there are few studies disclosing the HLA frequency database of the  
57 Han Chinese population with high resolution at the exon level and focusing on human

58 pathogen co-evolution. By exploiting the binding affinity of HLA molecules for  
59 pathogenic peptides, we can gain valuable insights into the evolutionary history of  
60 human adaptation to pathogens [15,16].

61 There has always been a phenomenon of co-adaptation between humans and  
62 pathogens, which has had an important impact on many human diseases, a typical  
63 example being autoimmune diseases. The hypothesis of “pathogen-driven selection”  
64 provides important insights into the mechanisms of the evolution of autoimmune  
65 diseases [17,18]. An ancient genomics study shows that immunity genes have been  
66 strongly affected by both positive and negative selection, and that resistance to  
67 infection has increased the risk of inflammatory disease over the past millennia [19].  
68 Despite this, HLA genes are the most important genes for pathogen adaptation and  
69 autoimmunity, and there are few studies based on them to elucidate the genetic impact  
70 of pathogen adaptation on autoimmune disease in the population.

71 In this study, we performed high-resolution genotyping of HLA alleles in a large  
72 cohort consisting of 4129 unrelated individuals from the NyuWa Genome Project [20],  
73 3202 unrelated individuals from the 1000 Genomes Project (1KGP) [21], and 893  
74 unrelated individuals from the Human Genome Diversity Project (HGDP) [22]. Our  
75 comprehensive analysis extended to characterising the binding affinity of common  
76 HLA alleles in the Han Chinese population for a range of human epidemic pathogens,  
77 and evaluating the influence of pathogen adaptation of HLA alleles on autoimmune  
78 disease. Finally, we explored the adaptive selection on the transcriptional regulation  
79 of the HLA alleles under the pathogen pressure. The analysis of HLA alleles in a large  
80 Han population helps us to understand the adaptation landscape of populations to  
81 pathogens and the relationship between pathogen adaptation and autoimmune disease.

## 82 **Results**

### 83 **High-resolution Genotyping of HLA Genes in Multiracial Populations**

84 In this comprehensive study, we successfully genotyped a total of 31 HLA genes.  
85 Our analysis was grounded in the assessment of genotyping rate, resolution, and  
86 accuracy across 8278 samples from the NyuWa project, the 1KGP, and the HGDP.

87 The genotyping rates of various HLA genes within these cohorts revealed that the  
88 classical HLA genes, including *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQA1*,  
89 *HLA-DQB1*, *HLA-DPA1*, and *HLA-DPBI*, were genotyped in all samples (Figure 1A).  
90 Moreover, all these HLA genes achieved a minimum of 2-field genotyping level  
91 (amino acid resolution), with majority achieving the 3-field genotyping level (exon  
92 sequence resolution) (Figure 1A). Upon comparing our genotyping results with  
93 benchmark genotypes for *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, and *HLA-DQB1* from  
94 the 1KGP, we observed that the genotyping accuracy for these genes exceeded 99% at  
95 the 1-field genotyping level (serotyping level), and had an impressive 94% to 97%  
96 accuracy at the amino acid level. (Figure 1B). In addition, we quantified the genetic  
97 diversity by calculating the heterozygosity of HLA genes with genotyping rates  
98 surpassing 80%. The findings indicate that the seven classic HLA genes, *HLA-A*,  
99 *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, and *HLA-DPBI*, exhibit  
100 exceptionally high heterozygosity, with values exceeding 80% within the population  
101 (Figure 1C). This high level of heterozygosity underscores their critical role in the  
102 adaptation to pathogens [3,4]. The genetic architecture of HLA genes in the Han  
103 Chinese closely resembles that of East Asian population but exhibits significant  
104 differences when compared to European and African populations, in particular, the  
105 most frequent HLA haplotype  
106 (HLA-A\*02:07–HLA-C\*01:02–HLA-B\*46:01–HLA-DRA\*01:01–HLA-DRB1\*09:0  
107 1–HLA-DQA1\*03:02–HLA-DQB1\*03:03–HLA-DPA1\*02:02–HLA-DPBI\*05:01)  
108 in the Han Chinese only accounted for 0.19%, while the most frequent HLA  
109 haplotype  
110 (HLA-A\*02:07–HLA-C\*01:02–HLA-B\*46:01–HLA-DRA\*01:01–HLA-DRB1\*14:5  
111 4–HLA-DQA1\*01:04–HLA-DQB1\*05:02–HLA-DPA1\*02:02–HLA-DPBI\*02:02)  
112 in the East Asian accounted for 0.30%, and the most frequent HLA haplotype  
113 (HLA-A\*03:01–HLA-C\*07:02–HLA-B\*07:02–HLA-DRA\*01:02–HLA-DRB1\*15:0  
114 1–HLA-DQA1\*01:02–HLA-DQB1\*06:02–HLA-DPA1\*01:03–HLA-DPBI\*04:01)  
115 in the European accounted for 0.40%, and the most frequent HLA haplotype  
116 (HLA-A\*30:01–HLA-C\*17:01–HLA-B\*42:01–HLA-DRA\*01:02–HLA-DRB1\*03:0

117 2–HLA-DQA1\*04:01–HLA-DQB1\*04:02–HLA-DPA1\*02:02–HLA-DPB1\*01:01)  
118 in the African accounted for 0.23% (Figure S1A–D).

119 Compared with the Allele Frequency Net Database [23], the HLA genes resource  
120 in this study provides HLA alleles of classification precision up to 3-field of  
121 large-scale of Han Chinese individuals, as well as multi-ethnic populations around the  
122 world. HLA alleles with 2-field or higher genotyping fields can provide information  
123 on the amino acid chain, which could help us complete molecular docking to  
124 determine the ability of HLA molecules to recognise foreign substances, and further  
125 find out the recent adaptation landscapes of HLA alleles in human populations. In  
126 addition, this HLA data resource has filled many missing genotypes of HLA genes in  
127 the northern Han population, providing a more complete perspective to understand the  
128 diversity and adaptive evolution of HLA genes in the Han Chinese population.

### 129 **Pathogen-peptide-binding Landscape of HLA Elements in the Han Chinese** 130 **Population**

131 In order to ascertain potential correlations between HLA-peptide binding affinity  
132 and pathogen adaptation, an investigation was conducted into the impact of HLA  
133 peptide binding affinity on the adaptability of the human immune system in the  
134 context of HIV and HCV infection. This investigation focused in particular on the  
135 associated HLA alleles as reported in studies [24,25]. The results indicate a significant  
136 correlation between an increased HLA binding affinity to HIV or HCV peptides and a  
137 reduced capacity for pathogen invasion and persistence (Figure S2, S3). These  
138 findings reveal a critical aspect of HLA adaptation to these viruses, namely that a  
139 higher affinity confers an enhanced ability to identify and combat these pathogens.  
140 Although HIV and HCV have only been present in the human population for only a  
141 few decades, the population's adaptation to the consensus sequences of other  
142 pathogens can lead to the adaptation of HIV and HCV [26,27]. This discovery has  
143 profound biological significance, as it underscores the strategic use of affinity scores  
144 of HLA molecules for pathogen-derived peptides in the prediction and analysis of  
145 immune responses, thereby enhancing our understanding of host-pathogen interaction  
146 and co-evolution.

147 By evaluating the binding affinity of both the common HLA types (AF  $\geq$  5%)  
148 and the low-frequency HLA types (AF  $\geq$  0.01 and AF  $<$  0.05) with various epidemic  
149 pathogens (Table S1), we observed that the different alleles of *HLA-DRB1*, regardless  
150 of their AF, demonstrated a stronger ability to bind to a wide range of pathogen  
151 antigens (Figure 2A and Figure 2B). This phenomenon is also observed for the rare  
152 *HLA-DRB1* alleles in the population (Figure S4), as well as for *HLA-DRB1* alleles in  
153 other populations (Figure S5 and Figure S6). Specifically, HLA-DRB1\*07:01 (AF =  
154 8.1%) shows a robust binding affinity to *Corynebacterium diphtheriae*, and  
155 HLA-DRB1\*08:03 (AF = 6.2%) exhibits a strong binding affinity to *Clostridium*  
156 *tetani* and *Bacillus anthracis* (Figure 2A). Additionally, HLA-DRB1\*14:54 (AF =  
157 2.7%) displays a notable binding affinity to *Bacillus anthracis* (Figure 2B). It is  
158 important to note that there is still a subtle tendency for the adaptive HLA types to  
159 bind to peptides. The most common peptide type bound by HLA-DRB1\*07:01 is  
160 “LIVS/T/KALKLI/L” (Figure S7A), while the most common peptide type bound by  
161 HLA-DRB1\*08:03 is “YV/KSSI/KK/NKILD” (Figure S7B), and  
162 “YL/II/KK/IKKNIE” is the most common peptide type bound by HLA-DRB1\*12:02  
163 (Figure S7C), and “IKI/NS/EK/SKE/NLL/I” is the most common peptide type bound  
164 by HLA-DRB1\*14:54 (Figure S7D). This preference for HLA peptide affinity  
165 suggests that adaptive HLA types may only target one of the prevalent pathogens with  
166 the preferred peptide, and adaptation to the rest of the pathogens is simply a result of  
167 passive adaptation to a pathogen that also carries the preferred peptide. Furthermore,  
168 HLA-DQB1\*03:01 and HLA-DQB1\*06:01 show a pronounced affinity for specific  
169 extracellular pathogens, such as *Mycobacterium tuberculosis* and *Bordetella pertussis*  
170 (Figure 2A). The affinity of the two *HLA-DQB1* alleles for the three extracellular  
171 pathogen antigenic peptides also showed a clear preference for a particular type of  
172 amino acid sequence (Figure S8). Among the peptides bound by HLA-DQB1\*03:01,  
173 there was a clear preference for the binding of the “VAAAAAAAA” peptide type  
174 (Figure S8A), whereas among the peptides bound by HLA-DQB1\*06:01, the  
175 “VAAAAAAAA” peptide type was also significantly preferred for binding (Figure  
176 S8B). This result suggests that the two HLA alleles that appear to be adapted to one of

177 these three pathogens are also adapted to the other two.

178 **Antagonistic Interactions between HLA Alleles involved in Adaptation and**  
179 **Autoimmune**

180 In order to investigate the genetic correlations between autoimmune diseases and  
181 pathogen adaptation in the human population, a correlation analysis was performed on  
182 the HLA alleles in the Han Chinese population and the autoimmune diseases related  
183 HLA alleles (Table S2). In general, a significant proportion of these *HLA-DRB1*  
184 alleles (HLA-DRB1\*03:01, HLA-DRB1\*04:05, HLA-DRB1\*07:01,  
185 HLA-DRB1\*09:01, HLA-DRB1\*01:01, HLA-DRB1\*13:01, HLA-DRB1\*13:02) and  
186 the HLA-DQB1\*06:01, which confers resistance to pathogens, has also been  
187 identified as the susceptibility alleles for numerous autoimmune diseases (Figure 3,  
188 Figure 4 and Figure S9). The pleiotropy of these HLA alleles provides an antagonistic  
189 balance between protection against foreign pathogens and the risk of autoimmunity  
190 [19]. In comparison to other ethnic groups, the HLA-DRB1\*07:01 and  
191 HLA-DRB1\*13:01 alleles in European and African populations exert a pleiotropic  
192 effect on the correlation between susceptibility to autoimmune diseases and pathogen  
193 adaptation (Figure S10 and Figure S11). Moreover, the high genetic linkage of HLA  
194 genes means that many pathogen-adapted HLA alleles also influence the population  
195 frequencies of autoimmune-related HLA alleles [28]. Indeed, it has been observed that  
196 over 80% of these common HLA alleles (Figure 3) and half of these low-frequency  
197 HLA alleles (Figure 4) show a significant genetic linkage to autoimmune-related HLA  
198 alleles in the Han Chinese population. In addition to the pleiotropic effects of HLA  
199 alleles, the high genetic linkage between HLA alleles represents a further significant  
200 factor contributing to the antagonistic balance between pathogen adaptation and  
201 disease risk. To illustrate, HLA-DRB1\*07:01 exhibits a robust binding affinity for the  
202 antigen of *Corynebacterium diphtheriae*, and it also shows significantly positive  
203 correlations with HLA-DQB1\*02:02 (Pearson correlation coefficient  $r^2 = 0.72$ ; the  
204 genetic distance between two HLA alleles  $DQB1_{cM} - DRB1_{cM} = 0.058cM$ ) and  
205 HLA-DQA1\*02:01 ( $r^2 = 0.99$ ;  $DRB1_{cM} - DQA1_{cM} = 0.014cM$ ). The former is a  
206 risk factor for coeliac disease (CD) [29], while the latter is associated with an

207 increased risk of inflammatory bowel disease (IBD) [30] (Figure 3).

208         Additionally, there are complex scenarios in which HLA alleles that interact with  
209 one another exhibit varying degrees of adaptability to different pathogens.  
210 Collectively, they influence the prevalence of autoimmune-related HLA alleles in  
211 diverse directions within the population. To illustrate, HLA-DQB1\*06 is a risk factor  
212 for MS [31], and has a strong positive correlation with many HLA alleles associated  
213 with pathogen adaptation, including some subtypes such as HLA-DQB1\*06:01 etc.,  
214 as well as the genetically linked HLA-DRB1\*08:03 ( $r^2 = 0.38$ ), and  
215 HLA-DRB1\*15:01 ( $r^2 = 0.49$ ; of the strongest susceptibility effect in MS [32]), also  
216 contains a pathogen-adapted HLA allele HLA-DQB1\*03:01 ( $r^2 = -0.28$ ) with a  
217 significant negative correlation (Figure 3). *Clostridium tetani* which produces tetanus  
218 toxin, and *Bordetella pertussis* which produces pertussis toxin, are significant  
219 inducers of MS [33,34]. Therefore, we postulated that the *Clostridium tetani*, which  
220 had left adaptive genetic signatures on HLA-DRB1\*08:03, and the *Bordetella*  
221 *pertussis*, which had left adaptive genetic signatures on HLA-DQB1\*03:01 (Figure  
222 2A), may have a significant impact on the genetic susceptibility to MS in the Han  
223 Chinese population in the recent evolutionary history. As an additional example,  
224 HLA-DPB1\*02:01, HLA-DPB1\*04:01 and HLA-DPB1\*05:01 represent the alleles at  
225 the *HLA-DPB1* locus, where the first two are positively correlated with  
226 HLA-DPA1\*01:03, while the latter is negatively correlated with HLA-DPA1\*01:03,  
227 which is the protective factor for IBD [30] (Figure 3). IBD has been reported to be  
228 associated with infection with a range of intestinal pathogens, including *Salmonella*  
229 *enterica*, *Epstein-Barr virus*, *Measles virus*, *Mumps virus*, *Rubella virus*,  
230 *Entamoeba histolytica*, *Toxoplasma gondii* etc. [35]. From a certain perspective, IBD  
231 has a more complex genetic structure due to the antagonistic evolution of HLA alleles  
232 and a wide range of intestinal pathogens.

### 233 **Pathogen-driven Adaptation of Transcriptional Regulation of *HLA-C***

234         In our work, we identified the upstream regulatory region of the *HLA-C* gene,  
235 which contains two genetically linked expression quantitative trait loci (eQTLs), 6:  
236 31336302A>G and 6:31337864C>T, as a target of recent positive selection based on



237 integrated haplotype score (iHS) and singleton density score (SDS) analyses (Figure  
238 S12; Figure S13). The derived alleles of these selected eQTLs exhibit a significantly  
239 extended haplotype homozygosity (Figure 4A and Figure 4C) and are characterised by  
240 a higher frequency in the East Asian population compared to the European and  
241 African populations (Figure S14). By making use of the resources on tissue-specific  
242 gene expression provided by the GTEx project, we were able to ascertain that the  
243 favoured alleles in question are prone to promoting the expression of *HLA-C* across a  
244 range of tissues (Figure 4B and Figure 4D). This elevated immune-related  
245 transcriptional level of HLA-C plays an essential role in strengthening the body's  
246 defences against pathogens. In accordance with Ohta's near-neutral theory of  
247 evolution, a mutation is considered nearly neutral if the selection coefficient  $s$  for the  
248 mutation meets the condition  $0.2 < |2Ns| < 4$  [36]. By tracking historical allele  
249 frequencies, it was observed that these two adaptive loci have been subject to a  
250 selection coefficient of 0.0009 ( $2 \times Ne \times s = 2 \times 20000 \times 0.0009 = 36 > 1$  ,  
251 indicating a slightly strong selection pressure) (Figure S15; Figure S16). It is  
252 noteworthy that the eQTL 6:31337864C>T has been identified as a risk genetic factor  
253 for schizophrenia, with the potential for elevated prevalence due to the pressures of  
254 pathogen-driven selection.

255 Furthermore, the impact of recent positive selection on the regulation of *HLA-C*  
256 alleles was examined, and it was determined that HLA-C\*03:02 and HLA-C\*04:01  
257 are significantly linked with favoured the eQTL (Figure 4E). HLA-C\*03:02, which  
258 has a higher allele frequency in East Asians, has been associated with a broad and  
259 potent resistance to various pathogens, especially *Mycobacterium tuberculosis*,  
260 *Salmonella enterica*, *Variola virus*, *Rabies virus*, and *Coronavirus*. In contrast,  
261 HLA-C\*04:01, which has the lowest allele frequency in East Asians, exhibits minimal  
262 resistance to pathogens (Figure 4E). Our findings suggest that the transcriptional  
263 regulation adjustment of HLA-C\*03:02 has contributed to the enhanced general  
264 adaptability of the Han Chinese population to a range of pathogens. The affinity of  
265 HLA-C\*03:02 for different pathogens is significantly different, which may be related  
266 to the specific preference of this allele for antigenic sequences. Therefore, we

267 analysed the motif characteristics of peptides binding to HLA-C\*03:02 and found that  
268 the amino acid fragments that can be bound by this HLA allele are very diverse and  
269 do not show significant motif preferences (Figure S17), which reveals the broad  
270 spectrum of adaptation of this HLA allele to pathogens.

271 It is also possible that *HLA-C*'s adaptation to pathogens at the gene expression  
272 level may affect HLA alleles related to human autoimmune diseases. Although  
273 HLA-C\*04:01 does not show significant correlation with known autoimmune  
274 diseases related HLA alleles, HLA-C\*03:02 makes a difference on four autoimmune  
275 diseases, including IBD, MS, Type 1 diabetes (T1D) and CD, in a way of genetic  
276 linkage (Figure 3). Here involved in four HLA alleles which have a significant  
277 positive correlation with HLA-C\*03:02, consist of HLA-DRB\*03:01 ( $r^2 = 0.29$ ;  
278 susceptible to MS and T1D, protective for IBD), HLA-DQB\*02:01 ( $r^2 = 0.28$ ;  
279 susceptible to CD and T1D, protective for IBD), HLA-DQA1\*05:01 ( $r^2 = 0.29$ ;  
280 susceptible to CD) and HLA-DRB\*03 ( $r^2 = 0.29$ ; susceptible to T1D) (Figure 3 and  
281 Table S2). Therefore, from the perspective of the highly linked nature of the HLA  
282 region, the adaptive selection of HLA alleles in gene expression will also affect the  
283 prevalence of autoimmune diseases in the population.

#### 284 **A High-resolution HLA Allele Frequency Database**

285 In an effort to facilitate data sharing, we have established a comprehensive HLA  
286 allele frequency database (<http://bigdata.ibp.ac.cn/HLAtyping>) that enables users to  
287 search for allele frequencies and homozygosity of HLA genes across various  
288 populations, ranging from serotyping-level to exon-genotyping-level detail. This  
289 database offers a user-friendly interface with an intuitive search function on its  
290 homepage (Figure 5A), whereby users can input the name of the HLA allele to access  
291 relevant information. To further enhance the search experience, the database has been  
292 designed to include customizable options for selecting the genotyping fields (Figure  
293 5B) and specifying the population of interest (Figure 5C). The database provides a  
294 comprehensive range of information for each HLA allele, including allele frequency,  
295 homozygosity and heterozygosity (proportion of carriers but not homozygotes)  
296 (Figure 5D). The allele frequency is a fundamental characteristic of HLA alleles

297 within a population and is instrumental for analysing population genetic structure,  
298 inferring demographic history, and assessing the immune adaptability to pathogenic  
299 antigens (both infectious and tumourous antigens) among human populations. The  
300 proportion of homozygous and heterozygous forms of an HLA allele provides a more  
301 detailed understanding than allele frequency alone, offering insights into the  
302 individual-level adaptation to the pathogenic environments.

### 303 **Discussion**

304 Pathogenic microbes represent an important driving force influencing human  
305 adaptive evolution [6]. It has been widely reported that HLA alleles are associated  
306 with susceptibility to pathogens [37–42] and autoimmune diseases [18,43–47]. The  
307 theory of pathogen-driven HLA diversity has been well developed to explain the  
308 significant association between the variants of HLA alleles and infectious diseases, as  
309 well as the autoimmune diseases [18,40,48]. Fortunately, the recent availability of  
310 large-scale of genome sequences in the human population [20–22] has facilitated the  
311 investigation of evolutionary relationships between HLA and pathogens at the  
312 population level. In light of the findings on host-pathogen co-evolution on  
313 MHC-peptide binding affinity from the studies [15,16], we performed an analysis of  
314 the binding affinity between pathogen peptides and the predominant HLA types in the  
315 Han Chinese population. Our results indicate that a significant number of HLA types  
316 possess varying degrees of binding ability to numerous pathogens (Figure 2), and are  
317 closely associated with the genetic risk of autoimmune diseases due to gene  
318 pleiotropy or genetic linkage (Figure 3).

319 The alleles of *HLA-DRB1* are found a notably stronger binding affinity to various  
320 pathogens than that of other HLA genes (Figure 2). This suggests that *HLA-DRB1*  
321 alleles have played a more substantial role in shaping the historical adaptation of  
322 human population to the intricate and varied pathogenic environment. *HLA-DRB1*  
323 plays a central role in the immune system by presenting peptides derived from  
324 extracellular proteins [49], and has the most diversity in the class II HLA genes [50] A  
325 phylogenetic study found that the *HLA-DRB1* gene clusters human-specific alleles

326 [51], thus we could speculated that *HLA-DRB1* has been in a long-term arms race  
327 with various pathogens and may have contributed to lots of local adaptation of  
328 humans. Furthermore, in this paper, the affinity between HLA and pathogens is  
329 measured by the average of the affinity scores between HLA and different peptides of  
330 pathogen, which describes the relative probability score of an HLA molecule  
331 recognizing a pathogen antigen. Although the strong binding effect of specific  
332 antigenic peptides may be masked, for example, one peptide affinity is 0.85 from a  
333 pathogen, and three peptides with 0.30 for another pathogen and all of them is 0.9,  
334 there is currently no more suitable method.

335       Assessing the impact of pathogen adaptation on the prevalence of autoimmune  
336 diseases in human populations is an important issue. There are two major genetic  
337 factors mediating the interaction between pathogen adaptation and autoimmune  
338 diseases. One is the pleiotropy of HLA alleles. In the Finnish population, several HLA  
339 alleles have been reported to be associated with susceptibility to infectious diseases  
340 and autoimmune diseases, such as *HLA-DQA1\*03:01* and *HLA-DQB1\*03:02* [52].  
341 Another one is the genetic linkage between nearby HLA alleles. Many of HLA-loci  
342 based associations result from linkage disequilibrium between the HLA gene studied  
343 and other HLA genes or non-HLA genes close by [53]. In this work, the gene  
344 pleiotropy and the genetic linkage between HLA genes represent a further avenue of  
345 enquiry in our research on measuring associations between the pathogen-adapted  
346 HLA alleles and the autoimmune-related HLA alleles. However, autoimmune diseases  
347 are also the interactive result of genes and living environment, and involve  
348 confounder factors such as sex bias and infections [54]. Therefore, it is still a  
349 challenge for us to discriminate the effects of HLA alleles on the induction of  
350 autoimmunity, especially when the HLA alleles are affected by pathogen adaptation, it  
351 is more difficult for us to quantify how the effect of specific pathogens on HLA genes  
352 will affect the occurrence and development of autoimmune diseases.

353       While our research, bolstered by clinical data on HIV and HCV infections [24,25],  
354 has identified a significant positive correlation between HLA-peptides binding  
355 affinity and the host's susceptibility to infection, further clinical data encompassing a

356 broader spectrum of pathogens is necessary to fully elucidate the extent of this  
357 association. Additionally, it is important to recognize the variability in pathogen  
358 strains across different geographical regions [55–57]. Consequently, future research  
359 endeavours should prioritize the investigation of human adaptive evolution to  
360 pathogens at the strain level, taking into account the diverse nature of these pathogens  
361 and their impact on human populations.

362 The topic of T cell clonality is also worthy of discussion. Following the  
363 processing of extracellular antigens into peptides and subsequent complexation with  
364 surface class II MHC molecules in on professional antigen-presenting cells such as  
365 dendritic cells, these MHC-peptide complexes are presented and recognised by CD4  
366 helper-T-cells [58]. The clonotype of a T cell population represents a molecular  
367 description of the unique sequences required to produce the T cell's TCR antigen  
368 specificity, as well as the specific V and J genes involved in the composite  
369 rearrangements, following the completion of the selection and maturation process of T  
370 cells in the thymus [59]. This adaptive immune process represents a pivotal step in the  
371 host's recognition of exogenous pathogens. Future research may take this into account,  
372 as well as other immune processes involved in the recognition of exogenous  
373 pathogens.

## 374 **Materials and methods**

### 375 **Data Resource Description**

376 In our study, we utilised genome data from a total of 8278 human individuals for  
377 HLA genotyping. The dataset comprised 4013 unrelated individuals from the NyuWa  
378 Chinese project, 2504 unrelated individuals from the 1000 Genomes Project (1KGP)  
379 (<https://www.internationalgenome.org/>), and 828 unrelated individuals from the  
380 Human Genome Diversity Project (HGDP) (<https://www.internationalgenome.org/>),  
381 as well as 116, 698, 119 related samples in each genome project. The  
382 quality-controlled alignment files (BAM format) for the 1KGP and HGDP samples  
383 were downloaded from their respective repositories, and the whole genome  
384 sequencing data for the NyuWa project samples were processed in alignment with the

385 methods described in the published paper[20]. Subsequently, all reads that had been  
386 assigned to the HLA region (chr6:28,510,120-33,480,577, according to the GRCh38  
387 assembly) were extracted for HLA genotyping. This encompasses reads that  
388 correspond to HLA genes, as well as those in an unmapped state. This approach  
389 ensures the full utilisation of sequencing data information in the HLA region, thereby  
390 reducing the false positive rate of HLA genotyping.

### 391 **HLA Genotyping**

392 The HLA-HD [60] software (<https://www.genome.med.kyoto-u.ac.jp/HLA-HD/>)  
393 is a high-fidelity tool designed for high-resolution and precision HLA genotyping. In  
394 our study, HLA genes were genotyped using the aforementioned software, with all  
395 reads that were potentially derived from the HLA genes realigned to the reference  
396 sequences from the IMGT/HLA database (Release 3.45.0). To evaluate the accuracy  
397 of HLA genotyping, we employed the SBT-PCR-based “Golden Sets” HLA typing  
398 data from 1206 individuals from the 1KGP [7]  
399 ([https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20140725\\_hla\\_genotypes/](https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20140725_hla_genotypes/)). The accuracy at each HLA gene is calculated by summing across the dosage of  
400 each correctly inferred HLA allele across all individuals ( $n$ ), and divided by the total  
401 number of observations ( $2n$ ) [13]. That is,

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

403 where  $Accuracy(g)$  represents the accuracy at a classical HLA gene (for  
404 example,  $HLA-C$ ).  $D_i$  represents the inferred dosage of an allele in individual  $i$ , and  
405 alleles  $A_{1i,g}$  and  $A_{2i,g}$  represent the true (SBT-PCR-based “Golden Sets”) HLA  
406 types for an individual  $i$ .

### 407 **HLA-peptides Binding Affinity Analysis**

408 In our investigation of the adaptability of HLA alleles to various epidemic  
409 pathogens (Table S1), we established a quantitative measure to define the affinity  
410 between an HLA molecule and a pathogen antigen as follows:

$$A = \frac{\sum_{i=1}^N Score_{EL_i}}{N} \quad (2)$$

411 In this equation,  $N$  denotes the total count of peptides derived from the antigen.

412 The term  $Score_{EL_i}$  signifies the predicted binding score (Predicted by  
413 NetMHCpan-4.1 and NetMHCIIpan-4.0 [61]) for the HLA molecule and the peptide  
414 that has significant binding affinity, comprising strong binders if the binding affinity  
415 ranking is in the top 0.5% and weak binders if the binding affinity ranking is in the  
416 top 2%. The quantification of HLA-peptide binding affinity is a well-established  
417 approach for prognosing the pathogen resistance of individual hosts [62].  
418 Consequently, the metric  $A$  serves as a descriptive tool to articulate the likelihood of  
419 a host's immune system effectively identifying and responding to pathogenic threats.

#### 420 **Correlation Analysis of HLA-peptides Affinity and Pathogen Resistance**

421 In order to evaluate the relationship between HLA-peptide binding affinity and  
422 pathogen resistance, we employed the mean value of predicted binding scores for  
423 HLA interactions with a pathogen's peptides as an evaluative metric. In order to gain  
424 insight into the clinical adaptation of HIV and HCV infections in relation to the HLA  
425 genotypes of their hosts, we computed the predicted mean binding scores for the  
426 engagement of HLA molecules with pathogenic peptides, drawing on two studies that  
427 had previously explored this topic. Subsequently, we compared these scores with  
428 various measures of pathogen resistance, including viral setpoint, progression of viral  
429 disease, and the status of viral RNA level, in order to ascertain the efficacy of HLA  
430 genes across different levels of binding affinity.

#### 431 **Reconstruction of HLA Reference Panel**

432 Following the established pipeline for constructing an HLA reference panel  
433 (<https://github.com/immunogenomics/HLA-TAPAS>) [13], we integrated HLA  
434 genotypes and haplotype reference panel that were delineated from SNPs and indels  
435 within the HLA region. The initial step entailed the recapturing of variants that had  
436 been absent from the initial analysis due to the high degree of polymorphism  
437 characteristic of HLA genes. This was accomplished by realigning the sequences of  
438 the genotyped HLA genes to the GRCh38 assembly. Subsequently, variants with a  
439 minor allele frequency exceeding 0.001 within the MHC region were extracted from  
440 the haplotype reference panels of the NyuWa Han Chinese, the 1KGP, and the HGDP.  
441 In the final step, the HLA genotypes were integrated into the aforementioned

442 haplotype reference panels, after which the haplotypes were re-phased using Shapeit4  
443 [63].

#### 444 **Selection Analysis of HLA Regulators**

445 To detect adaptive selection within HLA regulatory regions, we systematically  
446 computed the single density score (SDS), integrated haplotype score (iHS) statistics  
447 for genome-wide detection of recent positive selections and Beta statistics for  
448 long-term balancing selections in the NyuWa Han Chinese population [64–67]. In  
449 order to ensure the reliability of the results, we applied rigorous criteria to common  
450 variants, defined by a minor allele frequency (MAF) of 0.05 or greater. Subsequently,  
451 the aforementioned common variants were employed in the calculation of SDS, iHS  
452 and Beta statistics. Statistics with a false discover rate (FDR) adjusted using the  
453 Benjamini-Hochberg procedure of less than 0.05 ( $|SDS| \geq 3.857$ ;  $|iHS| \geq 4.234$ ;  
454  $Beta \geq 8.41$ ) are considered to indicate significant selection signals. To ascertain the  
455 robustness of potential selective sweeps or balancing selection events, a sliding  
456 window approach was employed, with each segment comprising 100 SNPs and a step  
457 size of 50 SNPs. Intervals with a higher proportion of significant signals within this  
458 framework were considered as indicative of more robust selective events. To  
459 substantiate the implications of these selection signals on phenotypic variation, we  
460 leveraged eQTL data from the GTEx project, which provided evidence for alterations  
461 in gene expression phenotypes at the transcriptional level [68]. The integration of  
462 selection statistics with functional genomics data enabled the potential fitness  
463 consequences of adaptive regulatory changes in the HLA region to be inferred.

#### 464 **Allele Frequency Trajectories Computation**

465 An approximate full-likelihood method [69] was employed to deduce the  
466 selection coefficients and allele frequency trajectories for the alleles under selection.  
467 In this analysis, our focus was on a 5kb haploblock, characterised by infrequent  
468 recombination, which encompasses the two adaptive sites of interest. This region was  
469 selected for the reconstruction of allele frequency trajectories over the last 1000  
470 generations, with the stipulation that the insertions and deletions within the  
471 haploblock were not considered. In preparation for inferring the allele frequency



472 trajectories, the Relate software [70] was employed to estimate the genealogical and  
473 population history of the haploblock. For the Han Chinese population, the effective  
474 population size was set to 20,000, the mutation rate to 1.25e-8, the years per  
475 generation to 28, and the number of sampling times of branch lengths to 100. All  
476 other parameters were configured according to their default settings.

### 477 **Ethical statement**

478 This study was approved by the Medical Research Ethics Committee of Institute  
479 of Biophysics, Chinese Academy of Sciences. All participants provided written  
480 informed consent. The informed consent is used to collect samples for genome studies  
481 conducted by Chinese Academy of Sciences. The consent requires participants to be  
482 30-70 years old patients and healthy people with full capacity. Participants voluntarily  
483 donate blood samples, provide clinical treatment information and sign informed  
484 consent. All their personal information is kept confidential. Participants can choose  
485 not to participate in sample donation, or withdraw at any time.

### 486 **Data Availability**

487 The DNA sequencing data of NyuWa samples used in this study have been  
488 deposited in the Genome Sequence Archive (GSA) in National Genomics Data Centre,  
489 China National Centre for Bioinformation/Beijing Institute of Genomics, Chinese  
490 Academy of Sciences, under accession number HRA004185  
491 (<https://ngdc.cncb.ac.cn/gsa-human/>). These data are available under restricted access  
492 for privacy protection and can be obtained by application on the GSA database  
493 website following the guidance of “Request Data” on this website. These data have  
494 also been deposited in the National Omics Data Encyclopaedia (NODE) of the  
495 Bio-Med Big Data Centre, Shanghai Institute of Nutrition and Health, Chinese  
496 Academy of Sciences, under accession number OEP002803  
497 (<http://www.biosino.org/node>). The user can register and login to this website and  
498 follow the guidance of “Request for Restricted Data” to request the data. The  
499 reference genome GRCh38 used in this study is available at  
500 <https://console.cloud.google.com/storage/browser/genomicspublic-data/resources/bro>

501 [ad/hg38/v0/](#).

502

## 503 **CRedit authorship contribution statement**

504 **Shuai Liu:** Conceptualization, Investigation, Methodology, Formal analysis,  
505 Visualization, Writing – original draft, Writing – review & editing. **Yanyan Li:**  
506 Conceptualization, Investigation, Methodology, Formal analysis, Visualization,  
507 Writing – original draft, Writing – review & editing. **Tingrui Song:**  
508 Conceptualization, Methodology, Formal analysis, Visualization, Writing – original  
509 draft, Writing – review & editing. **Jingjing Zhang:** Investigation, Methodology,  
510 Formal analysis. **Peng Zhang:** Data curation, Methodology, Formal analysis. **Huaxia**  
511 **Luo:** Methodology, Formal analysis. **Sijia Zhang:** Data curation. **Yiwei Niu:** Data  
512 curation. **Tao Xu:** Conceptualization, Funding acquisition, Project administration.  
513 **Shunmin He:** Conceptualization, Funding acquisition, Project administration, Writing  
514 – review & editing. All authors have read and approved the final manuscript.

## 515 **Competing interests**

516 The authors have declared that no competing interests exist.

## 517 **Acknowledgements**

518 We thank Zhen Xiong for thoughtful discussions and valuable comments  
519 regarding the immunogenetics. We thank the people for generously contributing  
520 samples to the NyuWa dataset. Data analysis and computing resources were supported  
521 by the Centre for Big Data Research in Health (<http://bigdata.ibp.ac.cn>), Institute of  
522 Biophysics, Chinese Academy of Sciences. This work was supported by Strategic  
523 Priority Research Program of the Chinese Academy of Sciences [XDB38040300];  
524 National Key R&D Program of China [2021YFF0703701 ,2021YFF0704500 ,  
525 2022YFC3400405]; 14th Five-year Informatization Plan of Chinese Academy of  
526 Sciences [CAS-WX2021SF-0203]; National Natural Science Foundation of China  
527 [91940306, 31970647, 32200478]; China Postdoctoral Science Foundation  
528 [2022M713311, GZC20232899]; and National Genomics Data Centre, China.

## 529 **Declaration of AI and AI-assisted technologies in the writing process**

530 In the writing process section, we did not use the AI and AI-assisted technologies

531 for generating content or images, writing code and processing data, and only apply  
532 them to the use of basic tools for checking grammar, spelling.

533 **ORCID IDs**

534 <sup>a</sup>ORCID: 0000-0001-9015-4772 (Shuai Liu).

535 <sup>b</sup>ORCID: 0000-0001-5256-6696 (Yanyan Li).

536 <sup>c</sup>ORCID: 0000-0003-2967-7704 (Tingrui Song).

537 <sup>d</sup>ORCID: 0009-0008-8916-551X (Jingjing Zhang).

538 <sup>e</sup>ORCID: 0000-0001-9303-1639 (Peng Zhang).

539 <sup>f</sup>ORCID: 0000-0001-9944-0345 (Huaxia Luo).

540 <sup>g</sup>ORCID: 0000-0001-9943-3073 (Sijia Zhang).

541 <sup>h</sup>ORCID: 0000-0002-9694-8159 (Yiwei Niu).

542 <sup>i</sup>ORCID: 0000-0002-8260-9754 (Tao Xu).

543 <sup>j</sup>ORCID: 0000-0002-7294-0865 (Shunmin He).

544

## 545 **References**

- 546 [1] Chaplin DD. Overview of the Immune Response. *J Allergy Clin Immunol*  
547 2010;125:S3-23. <https://doi.org/10.1016/j.jaci.2009.12.980>.
- 548 [2] Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, et al.  
549 Gene map of the extended human MHC. *Nat Rev Genet* 2004;5:889–99.  
550 <https://doi.org/10.1038/nrg1489>.
- 551 [3] Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V, Balloux F.  
552 Pathogen-Driven Selection and Worldwide HLA Class I Diversity. *Curr Biol*  
553 2005;15:1022–7. <https://doi.org/10.1016/j.cub.2005.04.050>.
- 554 [4] Manczinger M, Boross G, Kemény L, Müller V, Lenz TL, Papp B, et al. Pathogen  
555 diversity drives the evolution of generalist MHC-II alleles in human populations.  
556 *PLOS Biol* 2019;17:e3000131. <https://doi.org/10.1371/journal.pbio.3000131>.
- 557 [5] Radwan J, Babik W, Kaufman J, Lenz TL, Winternitz J. Advances in the  
558 Evolutionary Understanding of MHC Polymorphism. *Trends Genet* 2020;36:298–311.  
559 <https://doi.org/10.1016/j.tig.2020.01.008>.
- 560 [6] Karlsson EK, Kwiatkowski DP, Sabeti PC. Natural selection and infectious  
561 disease in human populations. *Nat Rev Genet* 2014;15:379–93.  
562 <https://doi.org/10.1038/nrg3734>.
- 563 [7] Gourraud P-A, Khankhanian P, Cereb N, Yang SY, Feolo M, Maiers M, et al.  
564 HLA Diversity in the 1000 Genomes Dataset. *PLOS ONE* 2014;9:e97282.  
565 <https://doi.org/10.1371/journal.pone.0097282>.
- 566 [8] Pillai NE, Okada Y, Saw W-Y, Ong RT-H, Wang X, Tantoso E, et al. Predicting  
567 HLA alleles from high-resolution SNP data in three Southeast Asian populations.  
568 *Hum Mol Genet* 2014;23:4443–51. <https://doi.org/10.1093/hmg/ddu149>.
- 569 [9] Lande A, Andersen I, Egeland T, Lie BA, Viken MK. HLA -A, -C,  
570 -B,-DRB1,-DQB1 and-DPB1 allele and haplotype frequencies in 4514 healthy  
571 Norwegians. *Hum Immunol* 2018;79:527–9.  
572 <https://doi.org/10.1016/j.humimm.2018.04.012>.
- 573 [10] Mimori T, Yasuda J, Kuroki Y, Shibata TF, Katsuoka F, Saito S, et al.

- 574 Construction of full-length Japanese reference panel of class I HLA genes with  
575 single-molecule, real-time sequencing. *Pharmacogenomics J* 2019;19:136–46.  
576 <https://doi.org/10.1038/s41397-017-0010-4>.
- 577 [11]Nordin J, Ameer A, Lindblad-Toh K, Gyllensten U, Meadows JRS. SweHLA: the  
578 high confidence HLA typing bio-resource drawn from 1000 Swedish genomes. *Eur J*  
579 *Hum Genet* 2020;28:627–35. <https://doi.org/10.1038/s41431-019-0559-2>.
- 580 [12]Tokić S, Žižkova V, Štefanić M, Glavaš-Obrovac L, Marci S, Samardžija M, et  
581 al. HLA-A, -B, -C, -DRB1, -DQA1, and -DQB1 allele and haplotype frequencies  
582 defined by next generation sequencing in a population of East Croatia blood donors.  
583 *Sci Rep* 2020;10:5513. <https://doi.org/10.1038/s41598-020-62175-9>.
- 584 [13]Luo Y, Kanai M, Choi W, Li X, Sakaue S, Yamamoto K, et al. A high-resolution  
585 HLA reference panel capturing global population diversity enables multi-ancestry  
586 fine-mapping in HIV host response. *Nat Genet* 2021;53:1504–16.  
587 <https://doi.org/10.1038/s41588-021-00935-7>.
- 588 [14]Zhou F, Cao H, Zuo X, Zhang T, Zhang X, Liu X, et al. Deep sequencing of the  
589 MHC region in the Chinese population contributes to studies of complex disease. *Nat*  
590 *Genet* 2016;48:740–6. <https://doi.org/10.1038/ng.3576>.
- 591 [15]Pierini F, Lenz TL. Divergent Allele Advantage at Human MHC Genes:  
592 Signatures of Past and Ongoing Selection. *Mol Biol Evol* 2018;35:2145–58.  
593 <https://doi.org/10.1093/molbev/msy116>.
- 594 [16]Özer O, Lenz TL. Unique Pathogen Peptidomes Facilitate Pathogen-Specific  
595 Selection and Specialization of MHC Alleles. *Mol Biol Evol* 2021;38:4376–87.  
596 <https://doi.org/10.1093/molbev/msab176>.
- 597 [17]Ramos PS, Shedlock AM, Langefeld CD. Genetics of autoimmune diseases:  
598 insights from population genetics. *J Hum Genet* 2015;60:657–64.  
599 <https://doi.org/10.1038/jhg.2015.94>.
- 600 [18]Matzaraki V, Kumar V, Wijmenga C, Zernakova A. The MHC locus and genetic  
601 susceptibility to autoimmune and infectious diseases. *Genome Biol* 2017;18:76.  
602 <https://doi.org/10.1186/s13059-017-1207-1>.
- 603 [19]Kerner G, Neehus A-L, Philippot Q, Bohlen J, Rinchai D, Kerrouche N, et al.

- 604 Genetic adaptation to pathogens and increased risk of inflammatory disorders in  
605 post-Neolithic Europe. *Cell Genomics* 2023;3:100248.  
606 <https://doi.org/10.1016/j.xgen.2022.100248>.
- 607 [20]Zhang P, Luo H, Li Y, Wang Y, Wang J, Zheng Y, et al. NyuWa Genome resource:  
608 A deep whole-genome sequencing-based variation profile and reference panel for the  
609 Chinese population. *Cell Rep* 2021;37. <https://doi.org/10.1016/j.celrep.2021.110017>.
- 610 [21]Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al.  
611 High-coverage whole-genome sequencing of the expanded 1000 Genomes Project  
612 cohort including 602 trios. *Cell* 2022;185:3426-3440.e19.  
613 <https://doi.org/10.1016/j.cell.2022.08.004>.
- 614 [22]Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al.  
615 Insights into human genetic variation and population history from 929 diverse  
616 genomes. *Science* 2020;367:eaay5012. <https://doi.org/10.1126/science.aay5012>.
- 617 [23]Gonzalez-Galarza FF, McCabe A, Santos EJM dos, Jones J, Takeshita L,  
618 Ortega-Rivera ND, et al. Allele frequency net database (AFND) 2020 update:  
619 gold-standard data classification, open access genotype data and new query tools.  
620 *Nucleic Acids Res* 2019;gkz1029. <https://doi.org/10.1093/nar/gkz1029>.
- 621 [24]Goulder PJR, Watkins DI. Impact of MHC class I diversity on immune control of  
622 immunodeficiency virus replication. *Nat Rev Immunol* 2008;8:619–30.  
623 <https://doi.org/10.1038/nri2357>.
- 624 [25]Kuniholm MH, Kovacs A, Gao X, Xue X, Marti D, Thio CL, et al. Specific  
625 human leukocyte antigen class I and II alleles associated with hepatitis C virus  
626 viremia. *Hepatology* 2010;51:1514–22. <https://doi.org/10.1002/hep.23515>.
- 627 [26]Moore CB, John M, James IR, Christiansen FT, Witt CS, Mallal SA. Evidence of  
628 HIV-1 Adaptation to HLA-Restricted Immune Responses at a Population Level.  
629 *Science* 2002;296:1439–43. <https://doi.org/10.1126/science.1069660>.
- 630 [27]Gaudieri S, Rauch A, Park LP, Freitas E, Herrmann S, Jeffrey G, et al. Evidence  
631 of Viral Adaptation to HLA Class I-Restricted Immune Pressure in Chronic Hepatitis  
632 C Virus Infection. *J Virol* 2006;80:11094–104. <https://doi.org/10.1128/jvi.00912-06>.
- 633 [28]Lenz TL, Spirin V, Jordan DM, Sunyaev SR. Excess of Deleterious Mutations

- 634 around HLA Genes Reveals Evolutionary Cost of Balancing Selection. *Mol Biol Evol*  
635 2016;33:2555–64. <https://doi.org/10.1093/molbev/msw127>.
- 636 [29]Abadie V, Kim SM, Lejeune T, Palanski BA, Ernest JD, Tastet O, et al. IL-15,  
637 gluten and HLA-DQ8 drive tissue destruction in coeliac disease. *Nature*  
638 2020;578:600–4. <https://doi.org/10.1038/s41586-020-2003-8>.
- 639 [30]Goyette P, Boucher G, Mallon D, Ellinghaus E, Jostins L, Huang H, et al.  
640 High-density mapping of the MHC identifies a shared role for HLA-DRB1\*01:03 in  
641 inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nat*  
642 *Genet* 2015;47:172–9. <https://doi.org/10.1038/ng.3176>.
- 643 [31]Fogdell-Hahn A, Ligers A, Grønning M, Hillert J, Olerup O. Multiple sclerosis: a  
644 modifying influence of HLA class I genes in an HLA class II associated autoimmune  
645 disease. *Tissue Antigens* 2000;55:140–8.  
646 <https://doi.org/10.1034/j.1399-0039.2000.550205.x>.
- 647 [32]Moutsianas L, Jostins L, Beecham AH, Dilthey AT, Xifara DK, Ban M, et al.  
648 Class II HLA interactions modulate genetic risk for multiple sclerosis. *Nat Genet*  
649 2015;47:1107–13. <https://doi.org/10.1038/ng.3395>.
- 650 [33]Verstraeten T, Davis R, DeStefano F. Immunity to tetanus is protective against the  
651 development of multiple sclerosis. *Med Hypotheses* 2005;65:966–9.  
652 <https://doi.org/10.1016/j.mehy.2005.05.009>.
- 653 [34]Rubin K, Glazer S. The potential role of subclinical *Bordetella Pertussis*  
654 colonization in the etiology of multiple sclerosis. *Immunobiology* 2016;221:512–5.  
655 <https://doi.org/10.1016/j.imbio.2015.12.008>.
- 656 [35]Axelrad JE, Cadwell KH, Colombel J-F, Shah SC. The role of gastrointestinal  
657 pathogens in inflammatory bowel disease: a systematic review. *Ther Adv*  
658 *Gastroenterol* 2021;14:17562848211004493.  
659 <https://doi.org/10.1177/17562848211004493>.
- 660 [36]Ohta T. Slightly Deleterious Mutant Substitutions in Evolution. *Nature*  
661 1973;246:96–8. <https://doi.org/10.1038/246096a0>.
- 662 [37]Li SS, Hickey A, Shangguan S, Ehrenberg PK, Geretz A, Butler L, et al.  
663 HLA-B\*46 associates with rapid HIV disease progression in Asian cohorts and



664 prominent differences in NK cell phenotype. *Cell Host Microbe*  
665 2022;30:1173-1185.e8. <https://doi.org/10.1016/j.chom.2022.06.005>.

666 [38]Julg B, Moodley ES, Qi Y, Ramduth D, Reddy S, Mncube Z, et al. Possession of  
667 HLA Class II DRB1\*1303 Associates with Reduced Viral Loads in Chronic HIV-1  
668 Clade C and B Infection. *J Infect Dis* 2011;203:803–9.  
669 <https://doi.org/10.1093/infdis/jiq122>.

670 [39]Wilson EA, Hirneise G, Singharoy A, Anderson KS. Total predicted MHC-I  
671 epitope load is inversely associated with population mortality from SARS-CoV-2.  
672 *Cell Rep Med* 2021;2:100221. <https://doi.org/10.1016/j.xcrm.2021.100221>.

673 [40]Sanchez-Mazas A. A review of HLA allele and SNP associations with highly  
674 prevalent infectious diseases in human populations. *Swiss Med Wkly* 2020.  
675 <https://doi.org/10.4414/smw.2020.20214>.

676 [41]Casanova J-L, Abel L. Lethal Infectious Diseases as Inborn Errors of Immunity:  
677 Toward a Synthesis of the Germ and Genetic Theories. *Annu Rev Pathol Mech Dis*  
678 2021;16:23–50. <https://doi.org/10.1146/annurev-pathol-031920-101429>.

679 [42]Kalaora S, Nagler A, Nejman D, Alon M, Barbolin C, Barnea E, et al.  
680 Identification of bacteria-derived HLA-bound peptides in melanoma. *Nature*  
681 2021;592:138–43. <https://doi.org/10.1038/s41586-021-03368-8>.

682 [43]Zanelli E, Breedveld FC, de Vries RRP. HLA association with autoimmune  
683 disease: a failure to protect? *Rheumatology* 2000;39:1060–6.  
684 <https://doi.org/10.1093/rheumatology/39.10.1060>.

685 [44]Gough SCL, Simmonds MJ. The HLA Region and Autoimmune Disease:  
686 Associations and Mechanisms of Action. *Curr Genomics* 2007;8:453–65.  
687 <https://doi.org/10.2174/138920207783591690>.

688 [45]Buendía-Roldán I, Santiago-Ruiz L, Pérez-Rubio G, Mejía M, Rojas-Serrano J,  
689 Ambrocio-Ortiz E, et al. A major genetic determinant of autoimmune diseases is  
690 associated with the presence of autoantibodies in hypersensitivity pneumonitis. *Eur*  
691 *Respir J* 2020;56. <https://doi.org/10.1183/13993003.01380-2019>.

692 [46]Yang X, Garner LI, Zvyagin IV, Paley MA, Komech EA, Jude KM, et al.  
693 Autoimmunity-associated T cell receptors recognize HLA-B\*27-bound peptides.

- 694 Nature 2022;1–7. <https://doi.org/10.1038/s41586-022-05501-7>.
- 695 [47]Brinkworth JF, Barreiro LB. The contribution of natural selection to present-day  
696 susceptibility to chronic inflammatory and autoimmune disease. *Curr Opin Immunol*  
697 2014;31:66–78. <https://doi.org/10.1016/j.coi.2014.09.008>.
- 698 [48]Hertz T, Nolan D, James I, John M, Gaudieri S, Phillips E, et al. Mapping the  
699 Landscape of Host-Pathogen Coevolution: HLA Class I Binding and Its Relationship  
700 with Evolutionary Conservation in Human and Viral Proteins. *J Virol*  
701 2011;85:1310–21. <https://doi.org/10.1128/JVI.01966-10>.
- 702 [49]Roche PA, Furuta K. The ins and outs of MHC class II-mediated antigen  
703 processing and presentation. *Nat Rev Immunol* 2015;15:203–16.  
704 <https://doi.org/10.1038/nri3818>.
- 705 [50]Barker DJ, Maccari G, Georgiou X, Cooper MA, Flicek P, Robinson J, et al. The  
706 IPD-IMGT/HLA Database. *Nucleic Acids Res* 2023;51:D1053–60.  
707 <https://doi.org/10.1093/nar/gkac1011>.
- 708 [51]Yasukochi Y, Satta Y. A human-specific allelic group of the MHC DRB1 gene in  
709 primates. *J Physiol Anthropol* 2014;33:14. <https://doi.org/10.1186/1880-6805-33-14>.
- 710 [52]Ritari J, Koskela S, Hyvärinen K, FinnGen, Partanen J. HLA-disease association  
711 and pleiotropy landscape in over 235,000 Finns. *Hum Immunol* 2022;83:391–8.  
712 <https://doi.org/10.1016/j.humimm.2022.02.003>.
- 713 [53]Tomlinson IPM, Bodmer WF. The HLA system and the analysis of multifactorial  
714 genetic disease. *Trends Genet* 1995;11:493–8.  
715 [https://doi.org/10.1016/S0168-9525\(00\)89159-3](https://doi.org/10.1016/S0168-9525(00)89159-3).
- 716 [54]Pisetsky DS. Pathogenesis of autoimmune disease. *Nat Rev Nephrol*  
717 2023;19:509–24. <https://doi.org/10.1038/s41581-023-00720-1>.
- 718 [55]Wiens KE, Woyczynski LP, Ledesma JR, Ross JM, Zenteno-Cuevas R,  
719 Goodridge A, et al. Global variation in bacterial strains that cause tuberculosis disease:  
720 a systematic review and meta-analysis. *BMC Med* 2018;16:196.  
721 <https://doi.org/10.1186/s12916-018-1180-x>.
- 722 [56]Magalhães LMD, Gollob KJ, Zingales B, Dutra WO. Pathogen diversity,  
723 immunity, and the fate of infections: lessons learned from *Trypanosoma cruzi*

- 724 human–host interactions. *Lancet Microbe* 2022;3:e711–22.  
725 [https://doi.org/10.1016/S2666-5247\(21\)00265-2](https://doi.org/10.1016/S2666-5247(21)00265-2).
- 726 [57]Seal S, Dharmarajan G, Khan I. Evolution of pathogen tolerance and emerging  
727 infections: A missing experimental paradigm. *eLife* 2021;10:e68874.  
728 <https://doi.org/10.7554/eLife.68874>.
- 729 [58]Pishesha N, Harmand TJ, Ploegh HL. A guide to antigen processing and  
730 presentation. *Nat Rev Immunol* 2022;22:751–64.  
731 <https://doi.org/10.1038/s41577-022-00707-2>.
- 732 [59]Mahe E, Pugh T, Kamel-Reid S. T cell clonality assessment: past, present and  
733 future. *J Clin Pathol* 2018;71:195–200.  
734 <https://doi.org/10.1136/jclinpath-2017-204761>.
- 735 [60]Kawaguchi S, Higasa K, Shimizu M, Yamada R, Matsuda F. HLA-HD: An  
736 accurate HLA typing algorithm for next-generation sequencing data. *Hum Mutat*  
737 2017;38:788–97. <https://doi.org/10.1002/humu.23230>.
- 738 [61]Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and  
739 NetMHCIpan-4.0: improved predictions of MHC antigen presentation by concurrent  
740 motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids*  
741 *Res* 2020;48:W449–54. <https://doi.org/10.1093/nar/gkaa379>.
- 742 [62]La Porta CAM, Zapperi S. Estimating the Binding of Sars-CoV-2 Peptides to  
743 HLA Class I in Human Subpopulations Using Artificial Neural Networks. *Cell Syst*  
744 2020;11:412-417.e2. <https://doi.org/10.1016/j.cels.2020.08.011>.
- 745 [63]Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for  
746 thousands of genomes. *Nat Methods* 2012;9:179–81.  
747 <https://doi.org/10.1038/nmeth.1785>.
- 748 [64]Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, et al. Detection of  
749 human adaptation during the past 2000 years. *Science* 2016;354:6.
- 750 [65]Voight BF, Kudravalli S, Wen X, Pritchard JK. A Map of Recent Positive  
751 Selection in the Human Genome. *PLOS Biol* 2006;4:e72.  
752 <https://doi.org/10.1371/journal.pbio.0040072>.
- 753 [66]Siewert KM, Voight BF. Detecting Long-Term Balancing Selection Using Allele

754 Frequency Correlation. *Mol Biol Evol* 2017;34:2996–3005.  
755 <https://doi.org/10.1093/molbev/msx209>.

756 [67]Szpiech ZA, Hernandez RD. selscan: an efficient multithreaded program to  
757 perform EHH-based scans for positive selection. *Mol Biol Evol* 2014;31:2824–7.  
758 <https://doi.org/10.1093/molbev/msu211>.

759 [68]Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on  
760 gene expression across human tissues. *Nature* 2017;550:204–13.  
761 <https://doi.org/10.1038/nature24277>.

762 [69]Stern AJ, Wilton PR, Nielsen R. An approximate full-likelihood method for  
763 inferring selection and allele frequency trajectories from DNA sequence data. *PLOS*  
764 *Genet* 2019;15:e1008384. <https://doi.org/10.1371/journal.pgen.1008384>.

765 [70]Speidel L, Forest M, Shi S, Myers SR. A method for genome-wide genealogy  
766 estimation for thousands of samples. *Nat Genet* 2019;51:1321–9.  
767 <https://doi.org/10.1038/s41588-019-0484-x>.

768  
769

770 **Figure legends**

771 **Figure 1 The evaluation and description for genotyped HLA genes**

772 **A.** Proportion of individuals with completed HLA genotyping and the resolution of  
773 genotyped HLA alleles. The designation “not typed” signifies that the NGS reads of  
774 an individual could not be identified as any known HLA allele by HLA-HD software.

775 **B.** The accuracy of HLA genotyping based on HLA-HD. This evaluation was  
776 performed based on the SBT-sequenced benchmark HLA genotypes from 1KGP. **C.**  
777 the heterozygosity of HLA genes in different populations.

778 **Figure 2 Binding affinity between common and low-frequency HLAs in the**  
779 **Han Chinese and the epidemic pathogens**

780 **A.** Common HLAs. **B.** Low-frequency HLAs. The pathogens represented by the  
781 yellow block are viruses, those represented by the pink block are bacteria, and those  
782 represented by the blue block are parasites. A higher score indicates a greater affinity.

783 **Figure 3 Associations between the common HLA alleles in the Han Chinese and**  
784 **the autoimmune susceptibility HLA alleles**

785 The horizontal axis represents common potential adaptive HLA alleles in the Han  
786 Chinese population, and the vertical axis represents autoimmune susceptibility HLA  
787 alleles. Pearson correlation tests were performed on the genotypes of the two sets of  
788 allele sets, and “\*” indicates a significantly correlated gene pair ( $r^2 > 0.2$  and  $P <$   
789  $0.05$ ).

790 **Figure 4 The transcriptional regulation of *HLA-C* under recent positive**  
791 **selection**

792 **A.** The EHH plot of the focal marker 6:31336302A>G. **B.** The normalized effect size  
793 of 6:31336302A>G on the expression of HLA-C in tissues. **C.** EHH plot of the focal  
794 marker 6:31337864C>T. **D.** The normalized effect size of 6:31337864C>T on the  
795 expression of *HLA-C* in tissues. **E.** The binding affinity of *HLA-C* alleles with 20  
796 pathogens that secretes intracellular antigens. A higher score indicates a greater  
797 binding affinity. It is notable that the *HLA-C* allele marked with a red star represents a

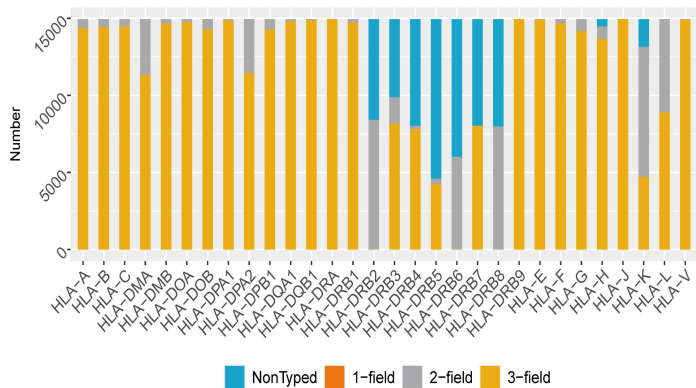
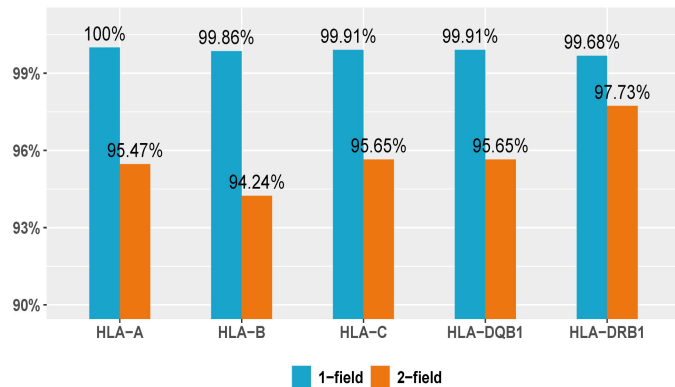
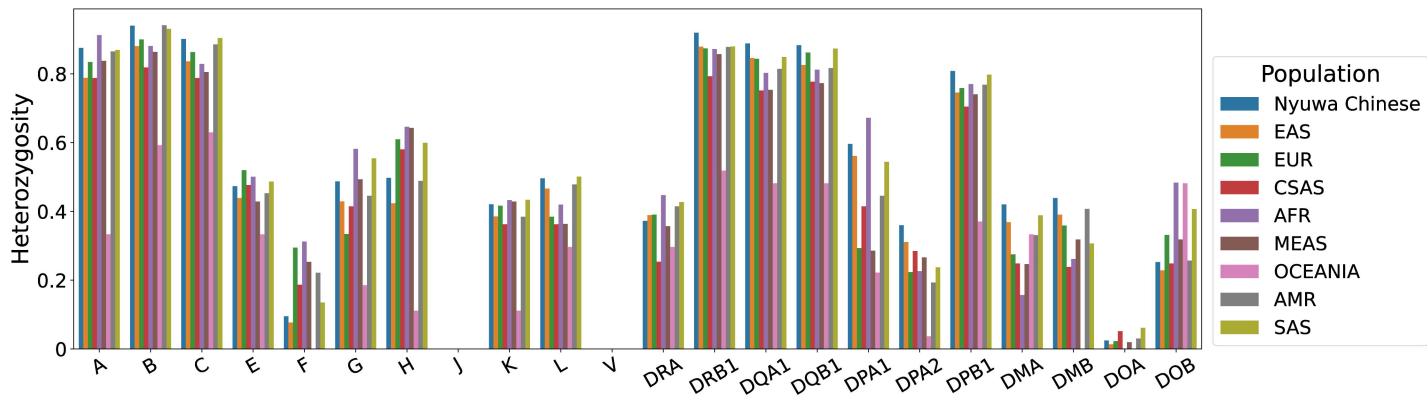
798 significant positive correlation with the favoured regulator, while the *HLA-C* allele  
799 marked with a blue star represents a significant negative correlation with the favoured  
800 regulator.

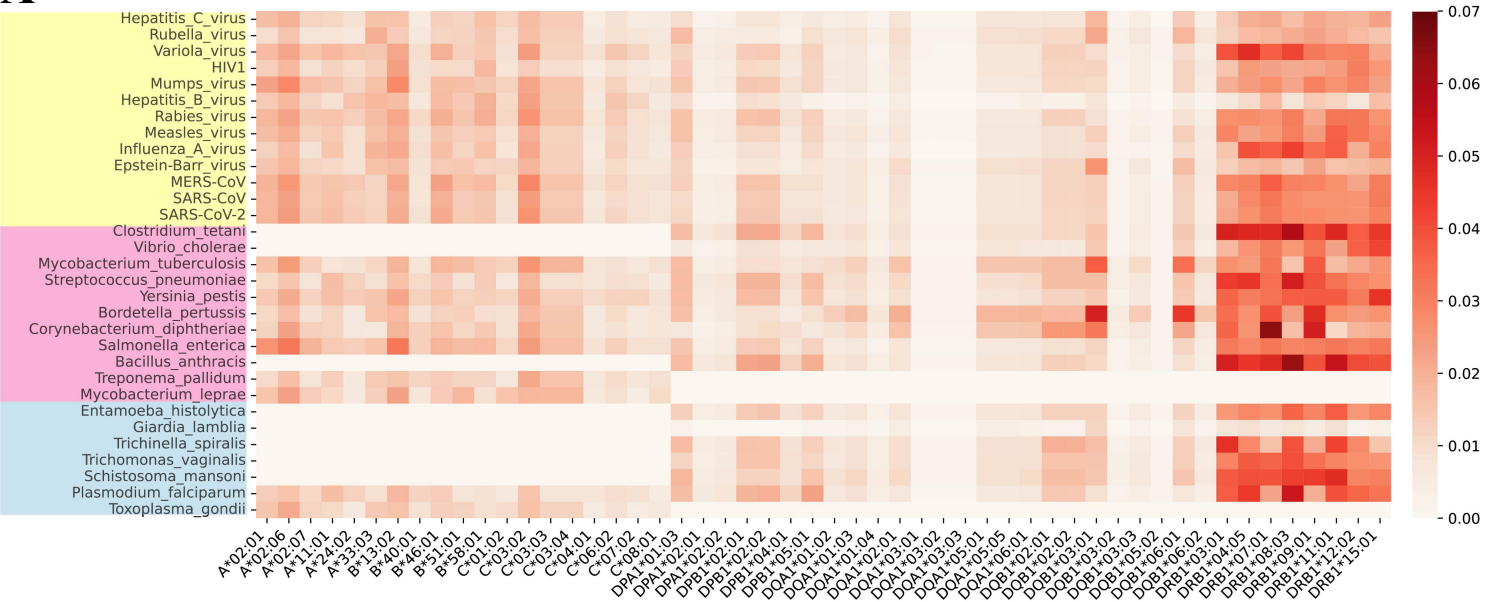
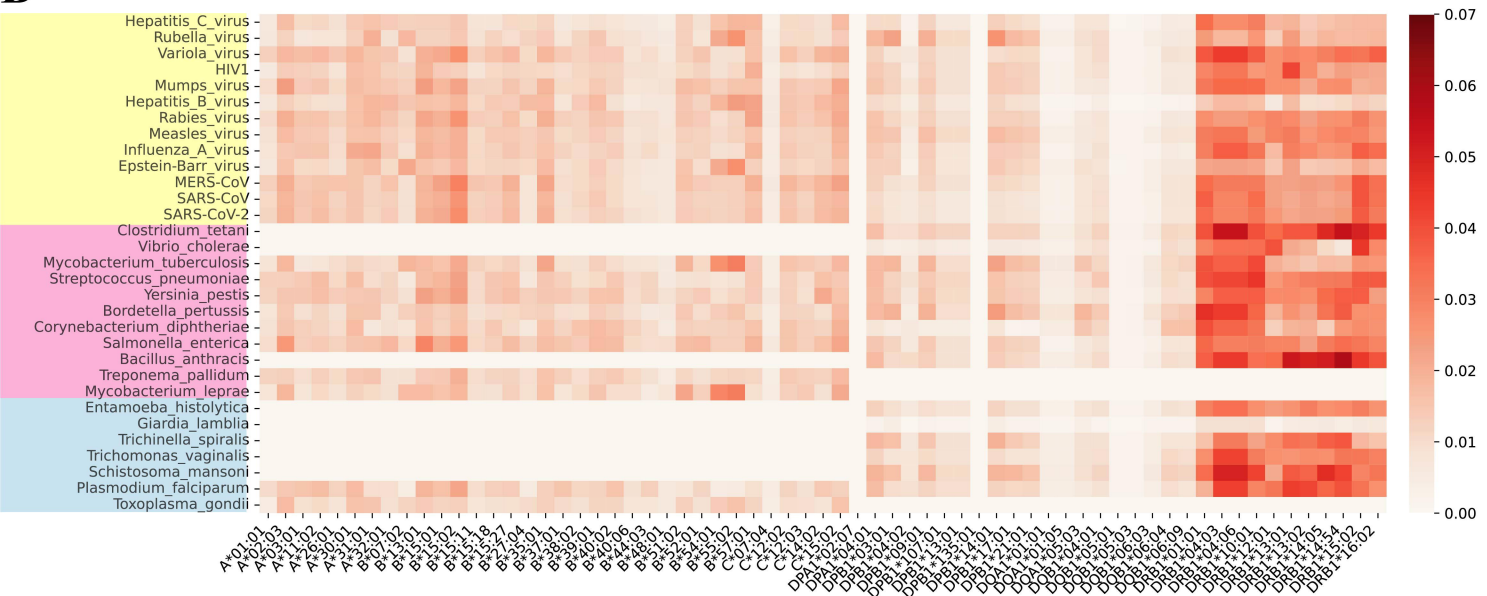
801 **Figure 5 The graphic user interface of the HLA Database**

802 **A.** A search bar is provided for the purpose of facilitating the input of search terms. **B.**  
803 The option for the field of HLA genotypes. **C.** The option for populations. **D.** Display  
804 of search results.

805

806

**A****B****C**

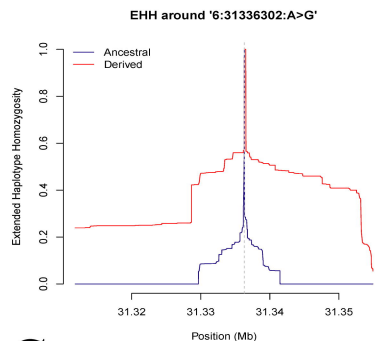
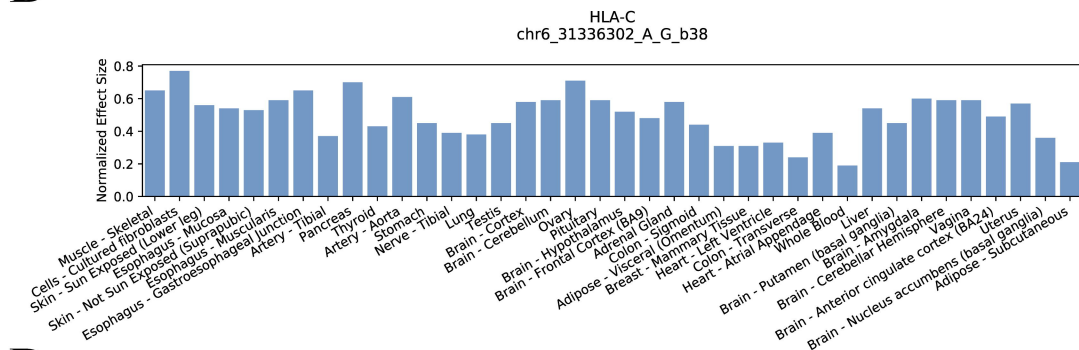
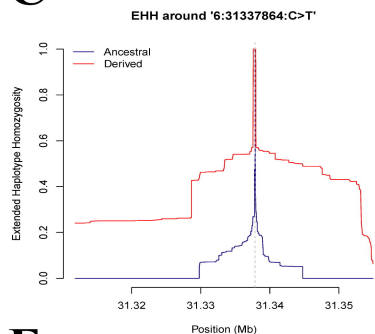
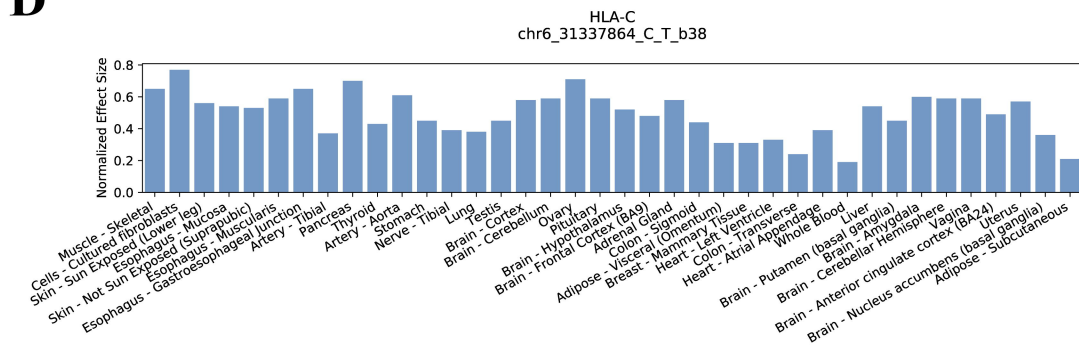
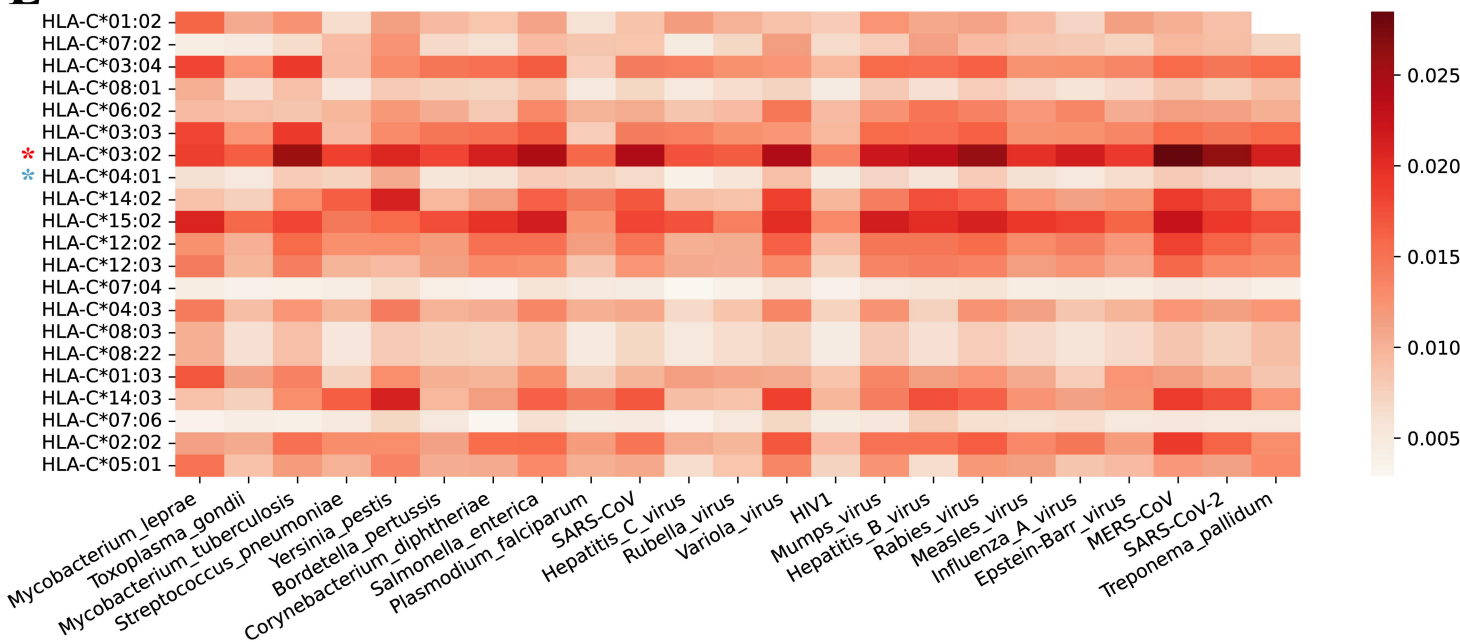
**A****B**





**Trait**

- Inflammatory bowel diseases & Multiple sclerosis & Type 1 diabetes
- Multiple sclerosis & Type 1 diabetes
- Coeliac disease
- Inflammatory bowel diseases & Type 1 diabetes & Coeliac disease
- Multiple sclerosis & Rheumatoid arthritis
- Rheumatoid arthritis
- Type 1 diabetes
- Inflammatory bowel diseases & Rheumatoid arthritis
- Inflammatory bowel diseases & Psoriasis
- Rheumatoid arthritis & Type 1 diabetes
- Psoriasis
- Multiple sclerosis

**A****B****C****D****E**

**A**

# HLA Database an integrative Human Leukocyte Antigen database

[Home](#) [Browse DB](#) [About](#)

## HLA (Human Leukocyte Antigen) Information Search

HLA gene:  Field:  Population:

Examples: A\*11,1-field,All A\*24:02,2-field,NyuWa\_Chinese A\*24:02,2-field,All DQA1\*01:02,2-field,All DQA1\*01:02,2-field,JPT A\*11,All,All

**B**

## HLA (Human Leukocyte Antigen) Information Search

HLA gene:  Field:  Population:

Examples: A\*11,1-field,All A\*24:02,2-field,NyuWa\_Chinese A\*24:02,2-field,All DQA1\*01:02,2-field,All DQA1\*01:02,2-field,JPT A\*11,All,All

**C**

## HLA (Human Leukocyte Antigen) Information Search

HLA gene:  Field:  Population:

Examples: A\*11,1-field,All A\*24:02,2-field,NyuWa\_Chinese A\*24:02,2-field,All DQA1\*01:02,2-field,All DQA1\*01:02,2-field,JPT A\*11,All,All

**D**

[Home](#) [Browse DB](#) [About](#)

## Browse HLA database

HLA gene:  Field:  Population:

HLA typing							
ID	HLA gene	Field	Allele Frequency	Homozygosity	Heterozygosity	Sample Size	Population
11837	A*24:02	2-field	0.163967	0.02965	0.26863	4013	NyuWa_Chinese(Han Chinese in NyuWa project)

Page 1 of 1

Displaying 1 to 1 of 1 items