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

Shuai Liu[1,3#], Yanyan Li[1#], Tingrui Song[1#], Jingjing Zhang[1,2], Peng Zhang[1], Huaxia Luo[1], Sijia Zhang[1,3], Yiwei Niu[1,3], Tao Xu[4,5*], Shunmin He[1,3*]

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

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

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

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

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

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

*Corresponding Authors: Tao Xu, Shunmin He

E-mail: xutao@ibp.ac.cn; heshunmin@ibp.ac.cn

ABSTRACT

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

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

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

Introduction

The immune process is of paramount importance in the intricate interplay between the *in vivo* life system and the surrounding pathogenic microbial environment [1]. The HLA genes located on chromosomal region 6p21.3, in particular the classical HLA genes (*HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1*), play a pivotal role in the identification and presentation of invasive foreign pathogens to the immune system [2]. As a result, these classical HLA genes are under intense selective pressure for resistance against pathogens, resulting in the highest degree of polymorphism observed in the human genome [3,4]. To date, two primary mechanistic theories, negative frequency-dependent selection (NFDS) and heterozygote advantage (HA), have been proposed to explain the remarkable diversity of HLA genes in the fluctuating pathogenic microbial environment [5]. Genetic predisposition to different pathogens is an area of increasing interest [6]. Although many human HLA resources have been developed [7–14], there are few studies disclosing the HLA frequency database of the Han Chinese population with high resolution at the exon level and focusing on human

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

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

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

Results

High-resolution Genotyping of HLA Genes in Multiracial Populations

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

The genotyping rates of various HLA genes within these cohorts revealed that the classical HLA genes, including *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1*, and *HLA-DPB1*, were genotyped in all samples (Figure 1A). Moreover, all these HLA genes achieved a minimum of 2-field genotyping level (amino acid resolution), with majority achieving the 3-field genotyping level (exon sequence resolution) (Figure 1A). Upon comparing our genotyping results with benchmark genotypes for *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, and *HLA-DQB1* from the 1KGP, we observed that the genotyping accuracy for these genes exceeded 99% at the 1-field genotyping level (serotyping level), and had an impressive 94% to 97% accuracy at the amino acid level. (Figure 1B). In addition, we quantified the genetic diversity by calculating the heterozygosity of HLA genes with genotyping rates surpassing 80%. The findings indicate that the seven classic HLA genes, *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, and *HLA-DPB1*, exhibit exceptionally high heterozygosity, with values exceeding 80% within the population (Figure 1C). This high level of heterozygosity underscores their critical role in the adaptation to pathogens [3,4]. The genetic architecture of HLA genes in the Han Chinese closely resembles that of East Asian population but exhibits significant differences when compared to European and African populations, in particular, the most frequent HLA haplotype (HLA-A*02:07–HLA-C*01:02–HLA-B*46:01–HLA-DRA*01:01–HLA-DRB1*09:01–HLA-DQA1*03:02–HLA-DQB1*03:03–HLA-DPA1*02:02–HLA-DPB1*05:01) in the Han Chinese only accounted for 0.19%, while the most frequent HLA haplotype (HLA-A*02:07–HLA-C*01:02–HLA-B*46:01–HLA-DRA*01:01–HLA-DRB1*14:54–HLA-DQA1*01:04–HLA-DQB1*05:02–HLA-DPA1*02:02–HLA-DPB1*02:02) in the East Asian accounted for 0.30%, and the most frequent HLA haplotype (HLA-A*03:01–HLA-C*07:02–HLA-B*07:02–HLA-DRA*01:02–HLA-DRB1*15:01–HLA-DQA1*01:02–HLA-DQB1*06:02–HLA-DPA1*01:03–HLA-DPB1*04:01) in the European accounted for 0.40%, and the most frequent HLA haplotype (HLA-A*30:01–HLA-C*17:01–HLA-B*42:01–HLA-DRA*01:02–HLA-DRB1*03:0

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

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

Pathogen-peptide-binding Landscape of HLA Elements in the Han Chinese Population

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

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

these three pathogens are also adapted to the other two.

Antagonistic Interactions between HLA Alleles involved in Adaptation and Autoimmune

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

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

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

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

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

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

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

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

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

A High-resolution HLA Allele Frequency Database

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

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

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

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

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

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

Materials and methods

Data Resource Description

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

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

HLA Genotyping

The HLA-HD [60] software (https://www.genome.med.kyoto-u.ac.jp/HLA-HD/) is a high-fidelity tool designed for high-resolution and precision HLA genotyping. In our study, HLA genes were genotyped using the aforementioned software, with all reads that were potentially derived from the HLA genes realigned to the reference sequences from the IMGT/HLA database (Release 3.45.0). To evaluate the accuracy of HLA genotyping, we employed the SBT-PCR-based "Golden Sets" HLA typing data from 1206 individuals from the 1KGP [7] (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 each correctly inferred HLA allele across all individuals (n), and divided by the total number of observations ($2n$) [13]. That is,

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

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

HLA-peptides Binding Affinity Analysis

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

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

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

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

Selection Analysis of HLA Regulators

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

Allele Frequency Trajectories Computation

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

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

Ethical statement

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

Data Availability

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

ad/hg38/v0/.

CRediT authorship contribution statement

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

Competing interests

The authors have declared that no competing interests exist.

Acknowledgements

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

Declaration of AI and AI-assisted technologies in the writing process

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

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

ORCID IDs

[a]ORCID: 0000-0001-9015-4772 (Shuai Liu).

[b]ORCID: 0000-0001-5256-6696 (Yanyan Li).

[c]ORCID: 0000-0003-2967-7704 (Tingrui Song).

[d]ORCID: 0009-0008-8916-551X (Jingjing Zhang).

[e]ORCID: 0000-0001-9303-1639 (Peng Zhang).

[f]ORCID: 0000-0001-9944-0345 (Huaxia Luo).

[g]ORCID: 0000-0001-9943-3073 (Sijia Zhang).

[h]ORCID: 0000-0002-9694-8159 (Yiwei Niu).

[i]ORCID: 0000-0002-8260-9754 (Tao Xu).

[j]ORCID: 0000-0002-7294-0865 (Shunmin He).

References

[1] Chaplin DD. Overview of the Immune Response. J Allergy Clin Immunol 2010;125:S3-23. https://doi.org/10.1016/j.jaci.2009.12.980.

[2] Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, et al. Gene map of the extended human MHC. Nat Rev Genet 2004;5:889–99. https://doi.org/10.1038/nrg1489.

[3] Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V, Balloux F. Pathogen-Driven Selection and Worldwide HLA Class I Diversity. Curr Biol 2005;15:1022–7. https://doi.org/10.1016/j.cub.2005.04.050.

[4] Manczinger M, Boross G, Kemény L, Müller V, Lenz TL, Papp B, et al. Pathogen diversity drives the evolution of generalist MHC-II alleles in human populations. PLOS Biol 2019;17:e3000131. https://doi.org/10.1371/journal.pbio.3000131.

[5] Radwan J, Babik W, Kaufman J, Lenz TL, Winternitz J. Advances in the Evolutionary Understanding of MHC Polymorphism. Trends Genet 2020;36:298–311. https://doi.org/10.1016/j.tig.2020.01.008.

[6] Karlsson EK, Kwiatkowski DP, Sabeti PC. Natural selection and infectious disease in human populations. Nat Rev Genet 2014;15:379–93. https://doi.org/10.1038/nrg3734.

[7] Gourraud P-A, Khankhanian P, Cereb N, Yang SY, Feolo M, Maiers M, et al. HLA Diversity in the 1000 Genomes Dataset. PLOS ONE 2014;9:e97282. https://doi.org/10.1371/journal.pone.0097282.

[8] Pillai NE, Okada Y, Saw W-Y, Ong RT-H, Wang X, Tantoso E, et al. Predicting HLA alleles from high-resolution SNP data in three Southeast Asian populations. Hum Mol Genet 2014;23:4443–51. https://doi.org/10.1093/hmg/ddu149.

[9] Lande A, Andersen I, Egeland T, Lie BA, Viken MK. HLA -A, -C, -B,-DRB1,-DQB1 and-DPB1 allele and haplotype frequencies in 4514 healthy Norwegians. Hum Immunol 2018;79:527–9. https://doi.org/10.1016/j.humimm.2018.04.012.

[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, Ameur 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, Marczi 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, Zhernakova 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.

Genetic adaptation to pathogens and increased risk of inflammatory disorders in post-Neolithic Europe. Cell Genomics 2023;3:100248. https://doi.org/10.1016/j.xgen.2022.100248.

[20]Zhang P, Luo H, Li Y, Wang Y, Wang J, Zheng Y, et al. NyuWa Genome resource: A deep whole-genome sequencing-based variation profile and reference panel for the Chinese population. Cell Rep 2021;37. https://doi.org/10.1016/j.celrep.2021.110017.

[21]Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell 2022;185:3426-3440.e19. https://doi.org/10.1016/j.cell.2022.08.004.

[22]Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science 2020;367:eaay5012. https://doi.org/10.1126/science.aay5012.

[23]Gonzalez-Galarza FF, McCabe A, Santos EJM dos, Jones J, Takeshita L, Ortega-Rivera ND, et al. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. Nucleic Acids Res 2019:gkz1029. https://doi.org/10.1093/nar/gkz1029.

[24]Goulder PJR, Watkins DI. Impact of MHC class I diversity on immune control of immunodeficiency virus replication. Nat Rev Immunol 2008;8:619–30. https://doi.org/10.1038/nri2357.

[25]Kuniholm MH, Kovacs A, Gao X, Xue X, Marti D, Thio CL, et al. Specific human leukocyte antigen class I and II alleles associated with hepatitis C virus viremia. Hepatology 2010;51:1514–22. https://doi.org/10.1002/hep.23515.

[26]Moore CB, John M, James IR, Christiansen FT, Witt CS, Mallal SA. Evidence of HIV-1 Adaptation to HLA-Restricted Immune Responses at a Population Level. Science 2002;296:1439–43. https://doi.org/10.1126/science.1069660.

[27]Gaudieri S, Rauch A, Park LP, Freitas E, Herrmann S, Jeffrey G, et al. Evidence of Viral Adaptation to HLA Class I-Restricted Immune Pressure in Chronic Hepatitis C Virus Infection. J Virol 2006;80:11094–104. https://doi.org/10.1128/jvi.00912-06.

[28]Lenz TL, Spirin V, Jordan DM, Sunyaev SR. Excess of Deleterious Mutations

around HLA Genes Reveals Evolutionary Cost of Balancing Selection. Mol Biol Evol 2016;33:2555–64. https://doi.org/10.1093/molbev/msw127.

[29] Abadie V, Kim SM, Lejeune T, Palanski BA, Ernest JD, Tastet O, et al. IL-15, gluten and HLA-DQ8 drive tissue destruction in coeliac disease. Nature 2020;578:600–4. https://doi.org/10.1038/s41586-020-2003-8.

[30] Goyette P, Boucher G, Mallon D, Ellinghaus E, Jostins L, Huang H, et al. High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. Nat Genet 2015;47:172–9. https://doi.org/10.1038/ng.3176.

[31] Fogdell-Hahn A, Ligers A, Grønning M, Hillert J, Olerup O. Multiple sclerosis: a modifying influence of HLA class I genes in an HLA class II associated autoimmune disease. Tissue Antigens 2000;55:140–8. https://doi.org/10.1034/j.1399-0039.2000.550205.x.

[32] Moutsianas L, Jostins L, Beecham AH, Dilthey AT, Xifara DK, Ban M, et al. Class II HLA interactions modulate genetic risk for multiple sclerosis. Nat Genet 2015;47:1107–13. https://doi.org/10.1038/ng.3395.

[33] Verstraeten T, Davis R, DeStefano F. Immunity to tetanus is protective against the development of multiple sclerosis. Med Hypotheses 2005;65:966–9. https://doi.org/10.1016/j.mehy.2005.05.009.

[34] Rubin K, Glazer S. The potential role of subclinical *Bordetella Pertussis* colonization in the etiology of multiple sclerosis. Immunobiology 2016;221:512–5. https://doi.org/10.1016/j.imbio.2015.12.008.

[35] Axelrad JE, Cadwell KH, Colombel J-F, Shah SC. The role of gastrointestinal pathogens in inflammatory bowel disease: a systematic review. Ther Adv Gastroenterol 2021;14:17562848211004493. https://doi.org/10.1177/17562848211004493.

[36] Ohta T. Slightly Deleterious Mutant Substitutions in Evolution. Nature 1973;246:96–8. https://doi.org/10.1038/246096a0.

[37] Li SS, Hickey A, Shangguan S, Ehrenberg PK, Geretz A, Butler L, et al. 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.

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.
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 NetMHCIIpan-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, Kudaravalli 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

Figure legends

Figure 1 The evaluation and description for genotyped HLA genes

A. Proportion of individuals with completed HLA genotyping and the resolution of genotyped HLA alleles. The designation "not typed" signifies that the NGS reads of an individual could not be identified as any known HLA allele by HLA-HD software. **B.** The accuracy of HLA genotyping based on HLA-HD. This evaluation was performed based on the SBT-sequenced benchmark HLA genotypes from 1KGP. **C.** the heterozygosity of HLA genes in different populations.

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

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

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

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

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

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

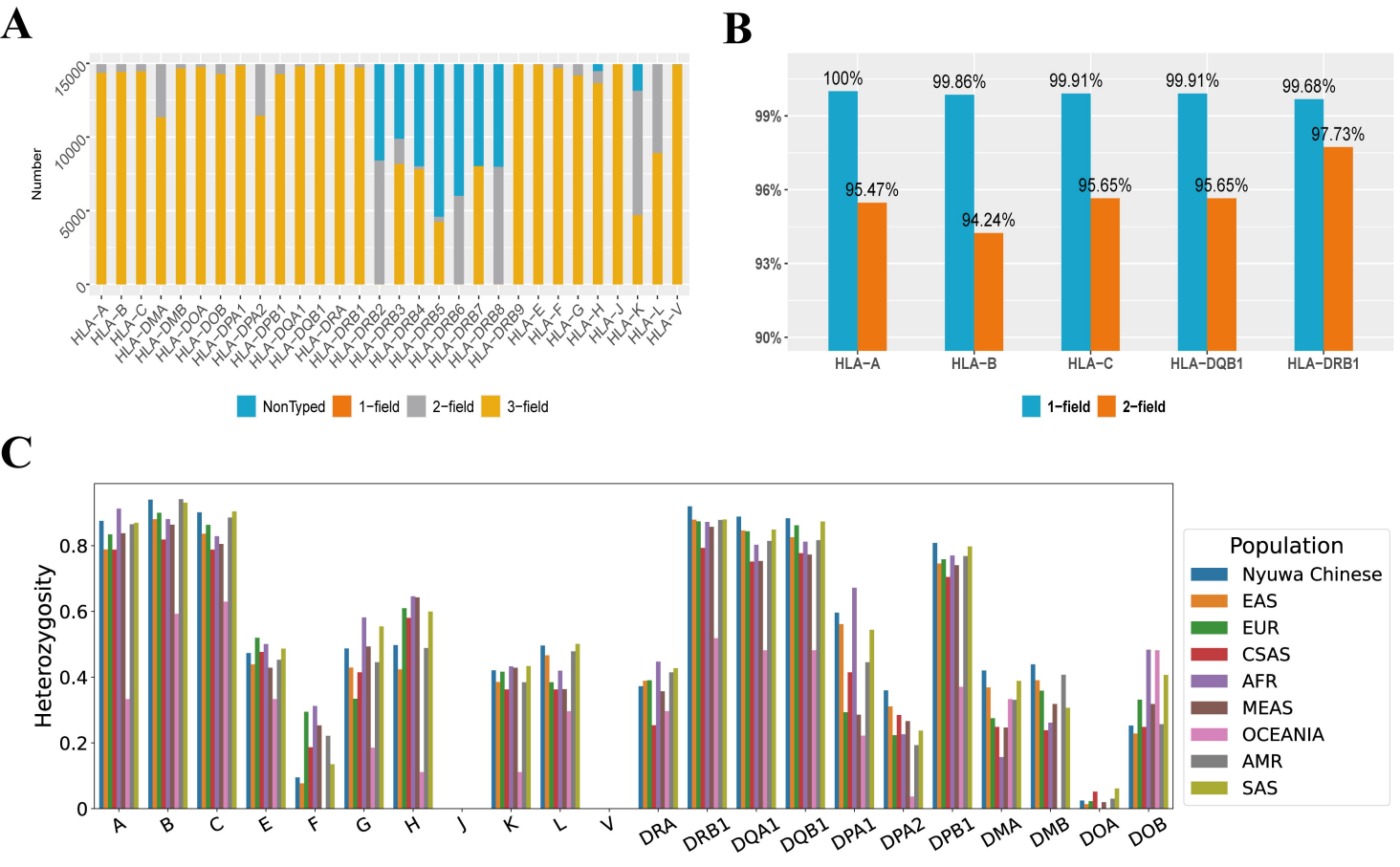
798 significant positive correlation with the favoured regulator, while the *HLA-C* allele
799 marked with a blue star represents a significant negative correlation with the favoured
800 regulator.

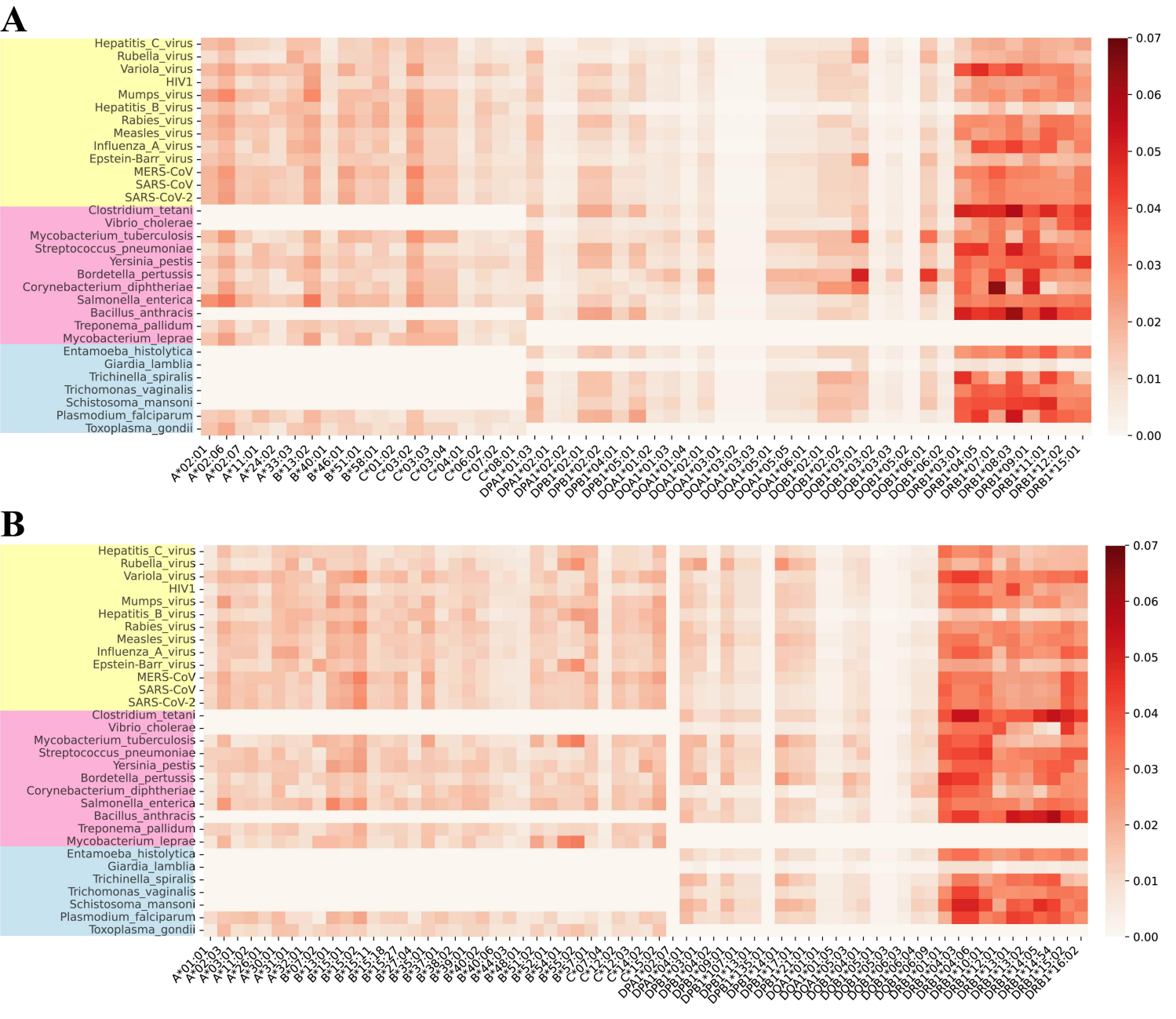
801 **Figure 5 The graphic user interface of the HLA Database**

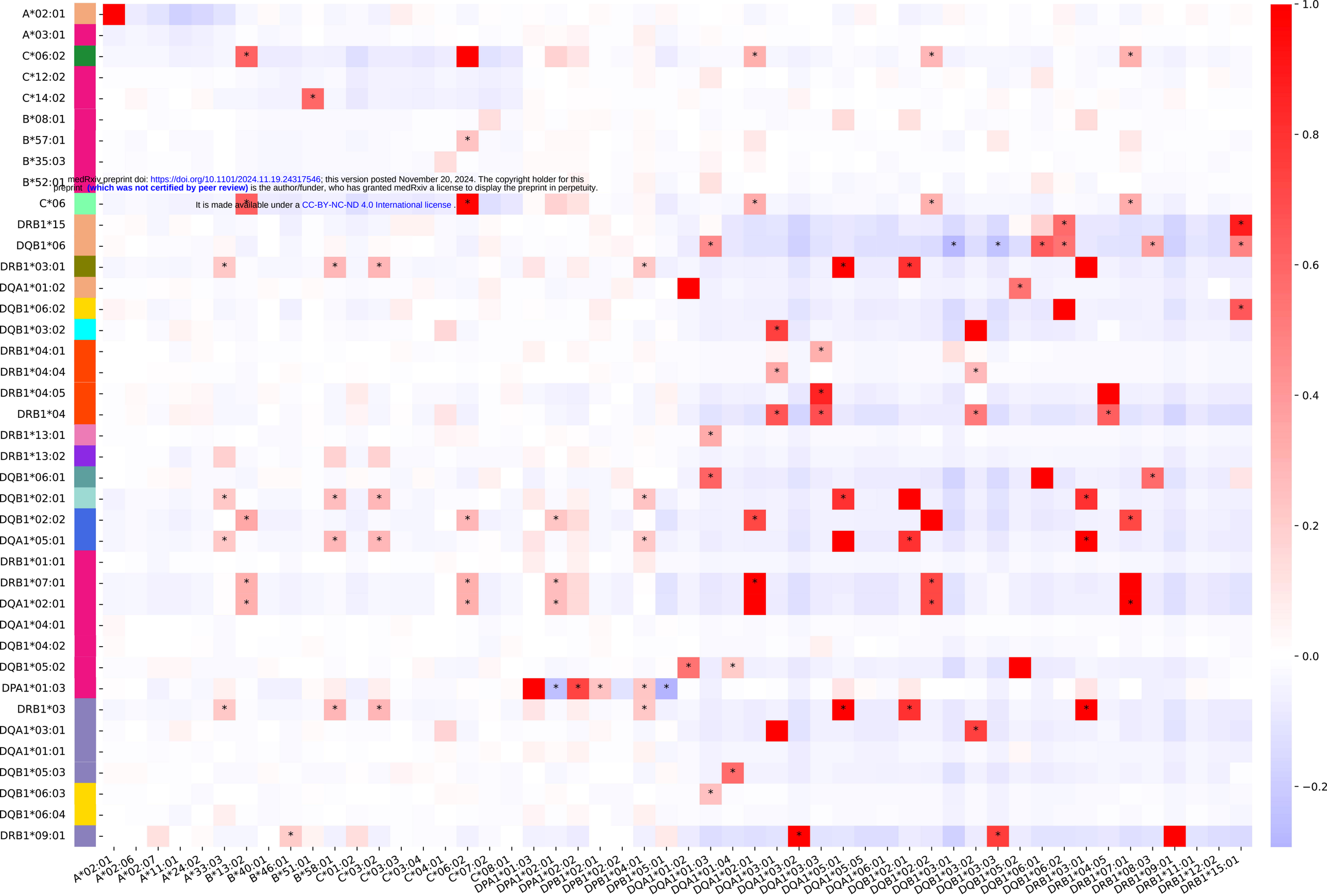
802 **A.** A search bar is provided for the purpose of facilitating the input of search terms. **B.**
803 The option for the field of HLA genotypes. **C.** The option for populations. **D.** Display
804 of search results.

805

806

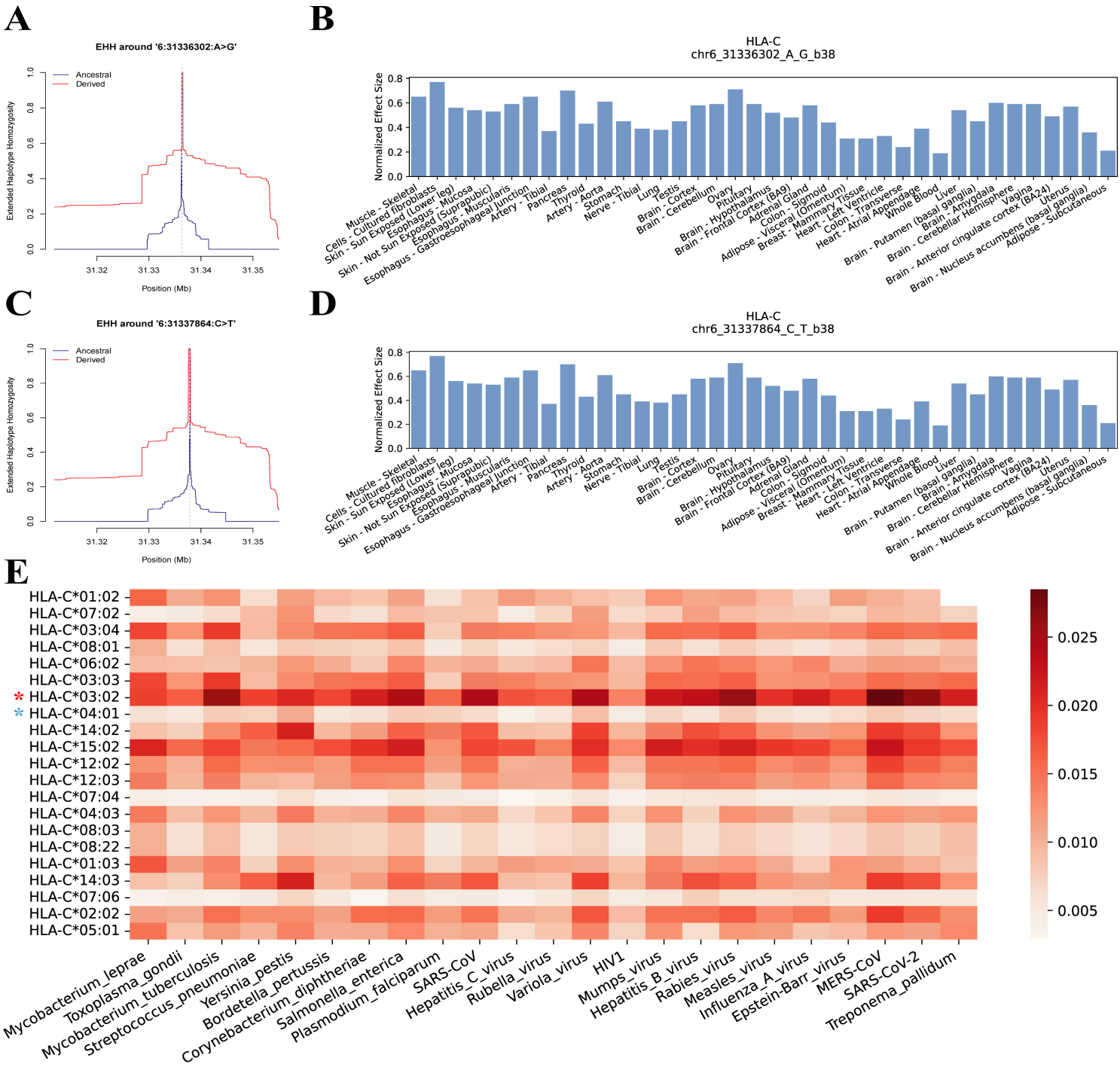

A
Number
15000
10000
5000
0
HLA-A
HLA-B
HLA-C
HLA-DMA
HLA-DMB
HLA-DOA
HLA-DOB
HLA-DPA1
HLA-DPA2
HLA-DPB1
HLA-DQA1
HLA-DQB1
HLA-DRA
HLA-DRB1
HLA-DRB2
HLA-DRB3
HLA-DRB4
HLA-DRB5
HLA-DRB6
HLA-DRB7
HLA-DRB8
HLA-DRB9
HLA-E
HLA-F
HLA-G
HLA-H
HLA-J
HLA-K
HLA-L
HLA-V
NonTyped
1-field
2-field
3-field
B
100%
99.86%
99.91%
99.91%
99.68%
95.47%
94.24%
95.65%
95.65%
97.73%
99%
96%
93%
90%
HLA-A
HLA-B
HLA-C
HLA-DQB1
HLA-DRB1
1-field
2-field
C
Heterozygosity
0
0.2
0.4
0.6
0.8
A
B
C
E
F
G
H
J
K
L
V
DRA
DRB1
DQA1
DQB1
DPA1
DPA2
DPB1
DMA
DMB
DOA
DOB
Population
Nyuwa Chinese
EAS
EUR
CSAS
AFR
MEAS
OCEANIA
AMR
SAS

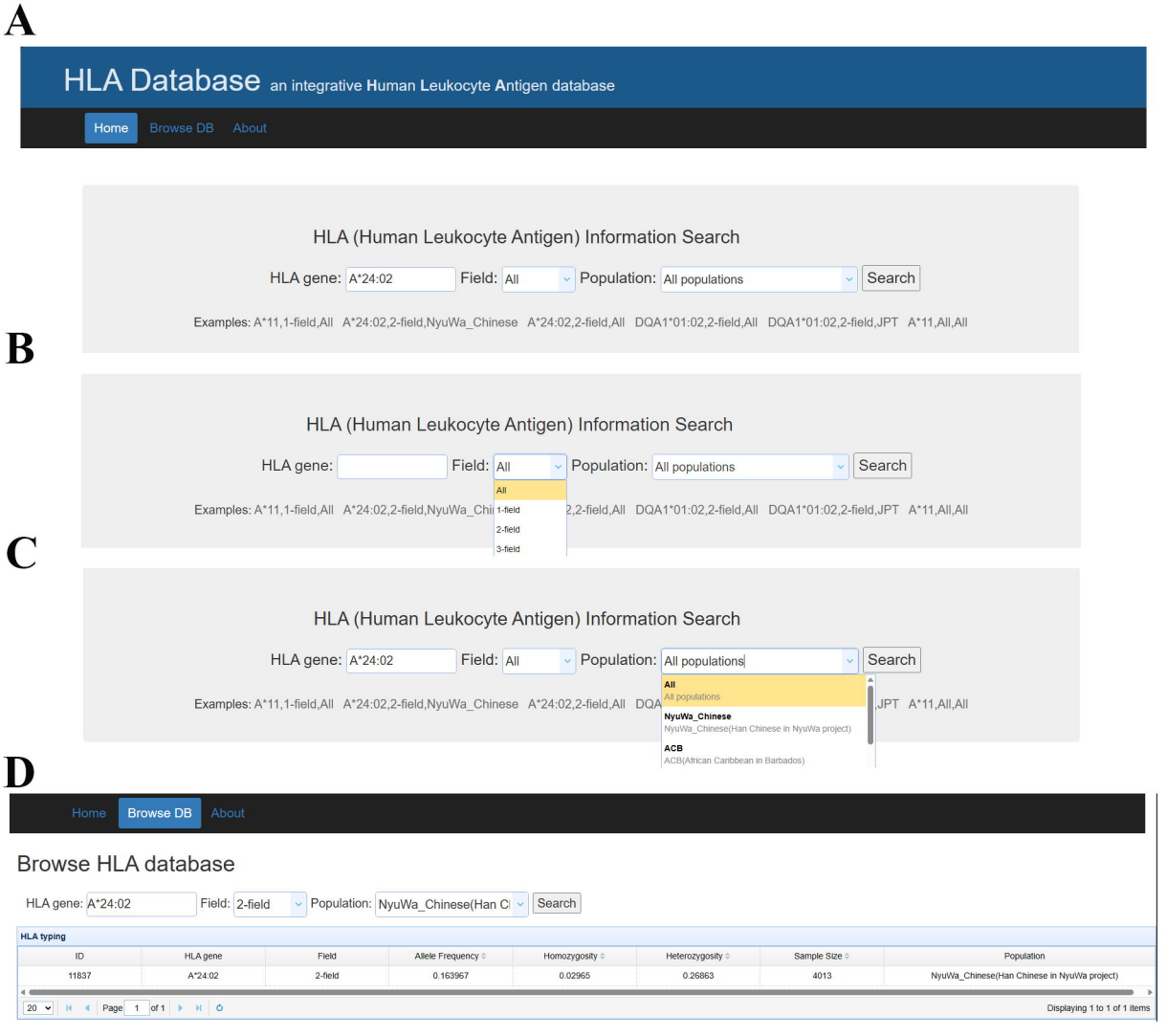

A
Hepatitis_C_virus
Rubella_virus
Variola_virus
HIV1
Mumps_virus
Hepatitis_B_virus
Rabies_virus
Measles_virus
Influenza_A_virus
Epstein-Barr_virus
MERS-CoV
SARS-CoV
SARS-CoV-2
Clostridium_tetani
Vibrio_cholerae
Mycobacterium_tuberculosis
Streptococcus_pneumoniae
Yersinia_pestis
Bordetella_pertussis
Corynebacterium_diphtheriae
Salmonella_enterica
Bacillus_anthracis
Treponema_pallidum
Mycobacterium_leprae
Entamoeba_histolytica
Giardia_lamblia
Trichinella_spiralis
Trichomonas_vaginalis
Schistosoma_mansoni
Plasmodium_falciparum
Toxoplasma_gondii
A*02:01
A*02:06
A*02:07
A*11:01
A*24:02
A*33:03
B*13:02
B*40:01
B*46:01
B*51:01
B*58:01
C*01:02
C*03:02
C*03:03
C*03:04
C*04:01
C*06:02
C*07:02
C*08:01
DPA1*01:03
DPA1*02:01
DPA1*02:02
DPB1*02:01
DPB1*02:02
DPB1*04:01
DPB1*05:01
DQA1*01:02
DQA1*01:03
DQA1*01:04
DQA1*02:01
DQA1*03:01
DQA1*03:02
DQA1*03:03
DQA1*05:01
DQA1*05:05
DQA1*06:01
DQB1*02:01
DQB1*02:02
DQB1*03:01
DQB1*03:02
DQB1*03:03
DQB1*05:02
DQB1*06:01
DQB1*06:02
DRB1*03:01
DRB1*04:05
DRB1*07:01
DRB1*08:03
DRB1*09:01
DRB1*11:01
DRB1*12:02
DRB1*15:01
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
B
Hepatitis_C_virus
Rubella_virus
Variola_virus
HIV1
Mumps_virus
Hepatitis_B_virus
Rabies_virus
Measles_virus
Influenza_A_virus
Epstein-Barr_virus
MERS-CoV
SARS-CoV
SARS-CoV-2
Clostridium_tetani
Vibrio_cholerae
Mycobacterium_tuberculosis
Streptococcus_pneumoniae
Yersinia_pestis
Bordetella_pertussis
Corynebacterium_diphtheriae
Salmonella_enterica
Bacillus_anthracis
Treponema_pallidum
Mycobacterium_leprae
Entamoeba_histolytica
Giardia_lamblia
Trichinella_spiralis
Trichomonas_vaginalis
Schistosoma_mansoni
Plasmodium_falciparum
Toxoplasma_gondii
A*01:01
A*02:03
A*03:01
A*11:02
A*26:01
A*30:01
A*31:01
A*32:01
B*07:02
B*13:01
B*15:01
B*15:02
B*15:11
B*15:18
B*27:04
B*35:01
B*37:01
B*38:02
B*39:01
B*40:02
B*40:06
B*44:03
B*48:01
B*51:02
B*52:01
B*54:01
B*55:02
C*01:02
C*07:04
C*12:03
C*14:02
C*15:02
DPA1*02:07
DPB1*03:01
DPB1*04:01
DPB1*04:02
DPB1*09:01
DPB1*10:01
DPB1*13:01
DPB1*135:01
DPB1*14:01
DPB1*17:01
DQA1*01:01
DQA1*01:05
DQA1*05:03
DQB1*02:01
DQB1*05:01
DQB1*05:03
DQB1*06:03
DQB1*06:04
DQB1*06:09
DRB1*01:01
DRB1*04:03
DRB1*04:06
DRB1*10:01
DRB1*12:01
DRB1*13:01
DRB1*13:02
DRB1*14:05
DRB1*14:54
DRB1*15:02
DRB1*16:02
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00



Trait

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


A
EHH around '6:31336302:A>G'
Ancestral
Derived
Extended Haplotype Homozygosity
Position (Mb)
B
HLA-C
chr6_31336302_A_G_b38
Normalized Effect Size
C
EHH around '6:31337864:C>T'
D
HLA-C
chr6_31337864_C_T_b38
E
HLA-C*01:02
HLA-C*07:02
HLA-C*03:04
HLA-C*08:01
HLA-C*06:02
HLA-C*03:03
HLA-C*03:02
HLA-C*04:01
HLA-C*14:02
HLA-C*15:02
HLA-C*12:02
HLA-C*12:03
HLA-C*07:04
HLA-C*04:03
HLA-C*08:03
HLA-C*08:22
HLA-C*01:03
HLA-C*14:03
HLA-C*07:06
HLA-C*02:02
HLA-C*05:01
Mycobacterium_leprae
Toxoplasma_gondii
Mycobacterium_tuberculosis
Streptococcus_pneumoniae
Yersinia_pestis
Bordetella_pertussis
Corynebacterium_diphtheriae
Salmonella_enterica
Plasmodium_falciparum
SARS-CoV
Hepatitis_C_virus
Rubella_virus
Variola_virus
HIV1
Mumps_virus
Hepatitis_B_virus
Rabies_virus
Measles_virus
Influenza_A_virus
Epstein-Barr_virus
MERS-CoV
SARS-CoV-2
Treponema_pallidum


A
HLA Database an integrative Human Leukocyte Antigen database
Home
Browse DB
About
HLA (Human Leukocyte Antigen) Information Search
HLA gene: A*24:02
Field: All
Population: All populations
Search
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: All
Population: All populations
Search
All
1-field
2-field
3-field
Examples: A*11,1-field,All A*24:02,2-field,NyuWa_Chi 2,2-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: A*24:02
Field: All
Population: All populations
Search
All
All populations
NyuWa_Chinese
NyuWa_Chinese(Han Chinese in NyuWa project)
ACB
ACB(African Caribbean in Barbados)
Examples: A*11,1-field,All A*24:02,2-field,NyuWa_Chinese A*24:02,2-field,All DQA JPT A*11,All,All
D
Home
Browse DB
About
Browse HLA database
HLA gene: A*24:02
Field: 2-field
Population: NyuWa_Chinese(Han C
Search
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)
20
Page 1 of 1
Displaying 1 to 1 of 1 items