Human Leukocyte Antigen Association

with Hepatitis B Virus-Mediated Liver Cancer

by

Yan Rou Yap

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved March 2021 by the
Graduate Supervisory Committee:

Melissa Wilson, Co-Chair
Efrem Lim, Co-Chair
Kenneth Buetow

ARIZONA STATE UNIVERSITY

May 2021

ABSTRACT

Human leukocyte antigen (HLA) is a group of proteins that the human immune system uses to detect pathogens. HLA is highly polymorphic, especially in the peptide-binding groove, which allows the binding of a diverse range of peptides including peptides produced by pathogens. Hepatitis B virus (HBV), is a pathogen that can cause liver disease. Chronic HBV infection, if left untreated, can lead to hepatocellular carcinoma, the most common form of liver cancer. In this paper, the association of Class I and II HLA with HBV-mediated liver cancer in patients of East Asian and European ancestry was studied. Results showed that, in the initial combined ancestry analysis, some alleles from all HLA types are associated with HBV-mediated liver cancer. However, once stratified by population ancestry, most of the alleles are no longer significant but still associate with HBV-mediated liver cancer in the same directions. In contrast, HLA-DP is the only HLA with haplotypes that are significantly different before and after stratification by ancestry. Notably, DPA10103-DPB10401, a previously known protective haplotype in the Asian population, is associated negatively with HBV-mediated liver cancer in both East Asian and European populations. Additionally, DPA10202-DPB10501, a known risk haplotype in the Asian population, is associated positively with HBV-mediated liver cancer patients of European ancestry. To understand how HLA-DP is associated with HBV-mediated liver cancer, the binding affinity of HLA-DP to all peptides generated from HBV coding sequences of genotypes A-H was predicted. It was speculated that an individual with HLA types that can bind strongly to HBV peptides will be more likely to clear viral infection whereas an individual with HLA types that fail to bind strongly to HBV peptides will be less likely to clear viral infection,

thus developing chronic infection. Results showed that DPA10103-DPB10401 binds strongly to HBV peptides (<50nM) whereas DPA10202-DPB10501 does not bind strongly to any HBV peptides (>50nM), consistent with the speculation that the binding affinity of HBV peptides to HLA will influence the association of HLA with HBV-mediated liver cancer.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my committee for their continual guidance and patience throughout this process. I am also grateful to Wilson Lab members for their constructive feedback. I would also like to acknowledge the support I received from School of Life Sciences and ASU Research Computing. Finally, I would like to thank my partner for being my rubber duck.

TABLE OF CONTENTS

Page

LIST OF TABLES vi

LIST OF FIGURES vii

CHAPTER

1 INTRODUCTION 1

2 METHODS 5

Predict Binding Affinity of Class I and II Alleles to HBV Peptides 5

HLA Typing of TCGA Dataset 5

Comparing the Frequency of HLA Types based on Population Ancestry 6

Test Association of Class I and II HLA Alleles with HBV-mediated Liver Cancer 8

Multiple Testing Correction and Statistical Analysis 9

3 RESULTS 10

Hepatitis B Infection Status Breakdown by Population Ancestry in TCGA 10

Significant HLA-A Alleles 11

Significant HLA-B Alleles 17

Significant HLA-C Alleles 22

Significant HLA-DP Alleles 29

Significant HLA-DQ Alleles 33

Significant HLA-DR Alleles 35

Binding Affinity Distribution of Significant HLA-DP to HBV Coding Sequences 40

CHAPTER Page

4 DISCUSSION 51

REFERENCES 55

LIST OF TABLES

Table Page

1. Populations Obtained from the Allele Frequency Net Database 8
2. Comparison of Association of HLA-A Alleles with HBV-Mediated and nonHBV-Mediated Liver Cancer Patients and the General Population of East Asian and European Ancestries 14
3. Comparison of Association of HLA-B Alleles with HBV-Mediated and nonHBV-Mediated Liver Cancer Patients and the General Population of East Asian and European Ancestries 16
4. Comparison of Association of HLA-C Alleles with HBV-Mediated and nonHBV-Mediated Liver Cancer Patients and the General Population of East Asian and European Ancestries 18
5. Comparison of Association of HLA-DP Alleles with HBV-Mediated and nonHBV-Mediated Liver Cancer Patients and the General Population of East Asian and European Ancestries 18
6. Comparison of Association of HLA-DQ Alleles with HBV-Mediated and nonHBV-Mediated Liver Cancer Patients and the General Population of East Asian and European Ancestries 20
7. Comparison of Association of HLA-DR Alleles with HBV-Mediated and nonHBV-Mediated Liver Cancer Patients and the General Population of East Asian and European Ancestries 20

LIST OF FIGURES

Figure Page

1. Number of HBV-Mediated and nonHBV-Mediated Liver Cancer Patients by Ancestry 11
2. HLA-A Alleles that are Significantly Different between HBV-Mediated and nonHBV-Mediated Liver Cancer Patients in the TCGA Dataset 12
3. HLA-A Alleles that are Significantly Different between HBV-Mediated Liver Cancer Patients and the General Population of East Asian and European Ancestries..... 14
4. HLA-A Alleles that are Significantly Different between nonHBV-Mediated Liver Cancer Patients and the General Population of East Asian and European Ancestries..... 15
5. HLA-B Alleles that are Significantly Different between HBV-Mediated and nonHBV-Mediated Liver Cancer Patients in the TCGA Dataset 18
6. HLA-B Alleles that are Significantly Different between HBV-Mediated Liver Cancer Patients and the General Population of East Asian and European Ancestries..... 19
7. HLA-B Alleles that are Significantly Different between nonHBV-Mediated Liver Cancer Patients and the General Population of East Asian and European Ancestries..... 20
8. HLA-C Alleles that are Significantly Different between HBV-Mediated and nonHBV-Mediated Liver Cancer Patients in the TCGA Dataset 23

Figure Page

9. HLA-C Alleles that are Significantly Different between HBV-Mediated Liver Cancer Patients and the General Population of East Asian and European Ancestries 25
10. HLA-C Alleles that are Significantly Different between nonHBV-Mediated Liver Cancer Patients and the General Population of East Asian and European Ancestries 27
11. HLA-DP Alleles that are Significantly Different between HBV-Mediated and nonHBV-Mediated Liver Cancer Patients in the TCGA Dataset 30
12. HLA-DQ Alleles that are Significantly Different between HBV-Mediated and nonHBV-Mediated Liver Cancer Patients in the TCGA Dataset 33
13. HLA-DR Alleles that are Significantly Different between HBV-Mediated and nonHBV-Mediated Liver Cancer Patients in the TCGA Dataset 36
14. HLA-DR Alleles that are Significantly Different between HBV-Mediated Liver Cancer Patients and the General Population of East Asian and European Ancestries 37
15. HLA-DR Alleles that are Significantly Different between nonHBV-Mediated Liver Cancer Patients and the General Population of East Asian and European Ancestries 38
16. Binding Affinity Distribution of HLA-DP Haplotypes to HBV Core Protein Peptides 41
17. Binding Affinity Distribution of HLA-DP Haplotypes to HBV Precore Protein Peptides 42

Figure Page

18. Binding Affinity Distribution of HLA-DP Haplotypes to HBV Large Surface Protein Peptides 43
19. Binding Affinity Distribution of HLA-DP Haplotypes to HBV Medium Surface Protein Peptides 44
20. Binding Affinity Distribution of HLA-DP Haplotypes to HBV Small Surface Protein Peptides 45
21. Binding Affinity Distribution of HLA-DP Haplotypes to HBV Polymerase Protein Peptides 46
22. Binding Affinity Distribution of HLA-DP Haplotypes to HBV X Protein Peptides 47
23. Binding Affinity Distribution of HLA-DP Haplotypes to HBV Spliced Protein Peptides 48

CHAPTER 1

INTRODUCTION

The immune system protects the human body by recognizing and removing pathogens. Human leukocyte antigen (HLA), the human equivalent of the major histocompatibility complex (MHC), is a group of alleles that play essential roles in detecting these pathogens. Its polymorphism allows the binding of a diverse range of peptides including peptides produced by pathogens. Hepatitis B virus (HBV), is a pathogen that can cause liver disease, which could eventually lead to the development of hepatocellular carcinoma if left untreated. In this paper, we study the association of HLA alleles with HBV-mediated liver cancer and make inferences about how the binding affinity of HLA alleles may influence a person's risk of developing HBV-mediated liver cancer.

Chronic hepatitis B virus (HBV) infections contribute to roughly 50% of hepatocellular carcinoma (HCC), the most common form of liver cancer, globally (Parkin, 2006). According to a 2015 WHO estimate, 257 million people were living with chronic HBV infection, 27 million people were aware of their infection, and only 4.5 million people were on treatment (World Health Organization, 2017).

HBV is a partially double-stranded DNA virus from the *Hepadnaviridae* family (Gao & Hu, 2007). Currently, there are eight established HBV genotypes (A-H) and two relatively new genotypes (I-J) (Velkov et al., 2018). Similar to other members of the *Hepadnaviridae* family, HBV's genome consists of a relaxed circular double-stranded DNA (rcDNA) molecule of 3.2kb in length (Mohd-Ismail et al., 2019). Upon entry into the host cell, the rcDNA is transported into the nucleus to be converted into covalently closed circular DNA (cccDNA) (Mohd-Ismail et al., 2019). cccDNA is a stable episome

and it serves as a transcriptional template for the synthesis of viral RNA (Gao & Hu, 2007). Current antiviral therapy employs nucleoside and nucleotide analogues, which inhibit HBV replication; however, it does not eliminate the cccDNA, making chronic HBV infection a lifelong disease (Hu et al., 2019).

Major histocompatibility complex (MHC) is a group of genes which encode proteins that are responsible for presenting antigenic peptides to immune cells (Hickey et al., 2016; Kaufman, 2018). MHC is an important part of the jawed vertebrate immune system (Kaufman, 2018). In humans, the MHC is known as human leukocyte antigen (HLA) and it spans 3.6Mbp on chromosome 6p21 (Beck & Trowsdale, 2000).

According to Hickey et al, there are two main classes of HLA: Class I and Class II (Hickey et al., 2016). Class I HLA consists of HLA-A, -B, and -C, and are responsible for antigen presentation to CD8+ T cells (Hickey et al., 2016). A class I HLA protein is composed of two non-covalently-linked polypeptide chains α and β2m (Hickey et al., 2016). The α chain is polymorphic and consists of three globular domains α_1, α_2, and α_3; in particular, α_1 and α_2 make up the peptide-binding domain, thus allowing the presentation of a wide range of peptides (Hickey et al., 2016). β2m is not polymorphic and is not involved in the peptide-binding process directly (Hickey et al., 2016). On the other hand, class II HLA consists of HLA-DP, -DQ, and -DR and are responsible for presenting antigens to CD4+ T cells (Hickey et al., 2016). A class II HLA protein is composed of two non-covalently-linked heavy chains α and β (Hickey et al., 2016). Both chains are polymorphic and have two globular domains α_1 and α_2 and β_1 and β_2 (Hickey et al., 2016). The peptide-binding domain is made of α_1 and β_1 and can bind to diverse peptides (Hickey et al., 2016).

Particular HLA types have been shown to be associated with diseases such as diabetes, celiac disease, and multiple sclerosis (Dendrou et al., 2018). Recent studies also showed that Class II HLAs are associated with HBV infection outcome. For example, HLA-A*33:03 was significantly associated with progression to HCC in a genome-wide association study (GWAS) in Japanese and other East Asian populations (Sawai et al., 2018). In addition, a GWAS study in a Taiwanese population showed that HLA-DPA1*02:02 and HLA-DPB1*05:01 were associated positively with chronic HBV (Huang et al., 2020).

In this paper, we investigate evidence for a relationship between HLA type and incidence of HBV-mediated liver cancer. We reason that if an individual has HLA types that can bind strongly to HBV peptides, then the individual will be more likely to be able to clear acute viral infection. Alternatively, if an individual has HLA types that cannot bind HBV peptides, then the individual will be less likely to be able to clear acute viral infection, thus developing chronic infection. Under this model, among HBV-infected liver cancer patients, we expect to see a high occurrence of HLAs that have low affinity for HBV peptides, while in non-HBV-infected liver cancer patients, we expect to see a high occurrence of HLAs that are capable of binding strongly to HBV peptides.

We find that once stratified by population ancestry, only HLA-DP haplotypes are significantly different between HBV-mediated and nonHBV-mediated liver cancer patients. In particular, we demonstrated that DPA10103-DPB10401 and DPA10202-DPB10501 are associated with HBV-mediated liver cancer patients of European ancestry. Both haplotypes were shown to be associated with chronic HBV infections in Asians in previous study (Kamatani et al., 2009). Although other allele types are not significantly

different after stratification, we noted that all of these alleles still associate with HBV-mediated liver cancer as before stratification, suggesting that there these alleles may be related to HBV-mediated liver cancer but are limited by the sample size and the uneven representation of HBV-mediated patients in each population ancestry.

CHAPTER 2

METHODS

Predict Binding Affinity of Class I and II Alleles to HBV Peptides

We used netMHCpan-4.1 and netMHCIIpan-4.0 to predict the binding affinity of Class I and II HLA alleles to HBV peptides, respectively (Reynisson et al., 2020). Both software requires two inputs: sequence information and allele names. The coding sequences of hepatitis B virus genomes for genotypes A-H were downloaded from HBVdb (The Hepatitis B Virus database) and used directly as inputs (Hayer et al., 2013). Complete lists of Class I and II HLAs were downloaded from DTUHealthTech and filtered to include only human HLAs (Reynisson et al., 2020). Most alleles can be used directly as inputs except HLA-DP and HLA-DQ. These two allele types require paired input: either HLA-DPA1 and HLA-DPB1 or HLA-DQA1 and HLA-DQB1. The peptide length was set to 9mer and 15mer for Class I and II alleles, respectively. The raw binding affinity prediction score was selected as output. An XLS file was generated for each HLA allele to store the binding affinities of each allele to all possible HBV peptides. The raw binding affinity prediction score was converted to binding affinity as seen in (1) and the unit is nM (Nielsen et al., 2003) .

$$Binding\ affinity\ = e^{\,log(50\,000nM)\,*\,(1-\,raw\ binding\ affinity\ prediction\ score)} \qquad (1)$$

HLA Typing of TCGA Dataset

A total of 379 TCGA LIHC patient whole exome sequencing BAM files were obtained from NCI Genomic Data Commons (dbGaP accession #11368) (Grossman et al., 2016).

We used HLAScan version 2.1.4 to perform typing of HLA-A, B, C, DPA1, DPB1, DQA1, DQB1, and DRB1 (Ka et al., 2017) on the TCGA BAM files. Default settings were used for all analysis. The score cutoff was set to 50, the constant of the score function was set to 30, and the number of threads was set to 32. 27 samples contained unidentifiable alleles in either of the eight HLA types, therefore they were removed from downstream analysis.

Comparing the Frequency of HLA Types based on Population Ancestry

To compare the allele frequency from the TCGA dataset and the general population, we obtained samples from blood donors, bone marrow registry, and controls for disease study from the Allele Frequency Net Database (Gonzalez-Galarza et al., 2020). These samples were selected to minimize the possibility of including diseased patients who might have a biased set of HLA types.

In the original TCGA dataset, self-reported race was obtained from enrolled patients. The categories are American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or other Pacific Islander, White, Not Evaluated, and Unknown. However, the Allele Frequency Net Database lacked ethnicity information and classified the dataset based on the regions that the dataset was collected. To ensure consistency across two datasets, we replaced race and population with the 1000 Genomes Super Population categories. The categories are African, Ad Mixed American, East Asian, European, and South Asian. We obtained the ancestry information of TCGA patients from Taravella Oill et al., which was previously inferred using PopInf and used the 1000 Genomes as a reference panel (Taravella Oill et al., 2020). For the general

population dataset, we considered all populations in one region to share a common ancestry and removed any populations that could be admixed. Specifically, we selected populations from Europe, North-East Asia, and South-East Asia. The North-East Asia and South-East Asia regions were merged to create a superpopulation, East Asia, to be consistent with the ancestry inferred from PopInf.

Furthermore, any samples that do not have information pertaining to sample size or number of individuals with alleles were removed. Although the allele frequency is available, we cannot calculate the actual number of individuals with an allele because we do not know whether the allele is heterozygous or homozygous. The complete list of filtered populations and their exact names from the database are listed in Table 1.

Table 1

Populations Obtained from the Allele Frequency Net Database

East Asia	Europe
Japan pop 4 Japan pop 7 Japan South Japan Hokkaido Wajin Japan Nagano Japan pop 11 South Korea pop 4 Japan pop 6 China Hubei Han Hong Kong Chinese Hong Kong Chinese HKBMDR HLA 11 loci Hong Kong Chinese HKBMDR. DQ and DP China pop1	Austria England North West Netherlands UMCU Italy Bergamo England Lancaster England Manchester England Leeds England Newcastle England Sheffield Greece pop2 Ireland Donegal Ireland Wexford Greece pop 8 Ireland Northern Ireland South Spain (Catalunya, Navarra, Extremadura, AaragÃ³n, Cantabria, Poland BMR Wales

	Serbia pop 2 France West Spain Sevilla Italy pop 4 Spain Barcelona Spain Northwest Portugal Azores Terceira Island Czech Republic pop 2 England pop 3 France Rennes pop 2 Greece Italy Sardinia pop2 Netherlands England Czech Republic pop 3 Germany pop 3 Slovenia pop 2 Belgium pop 2 Ireland Northern pop 2 Spain Malaga Spain Malaga Romani England Bedfordshire Spain Northwest Lugo Sweden Stockholm

Note. Table 1 shows the general populations from East Asia and Europe included in the control group. We considered the populations in each region to share a common ancestry and removed any populations that could include admixed populations.

Test Association of Class I and II HLA Alleles with HBV-mediated Liver Cancer

Fisher's Exact Tests were performed using the *fisher.test* function from the baseline *stats* module in R to test for associations between HLA allele type and HBV-mediated liver cancer because some of the alleles have an expected frequency of less than 1, failing to meet Cochran's rules (Cochran, 1952, 1954). An allele type is included if it is present in at least one patient in the TCGA dataset. The null hypothesis tested is that there is no association of an HLA allele with HBV-mediated liver cancer. The alternative hypothesis

tested is that there is an association of an HLA allele with HBV-mediated liver cancer. Since there are multiple instances of zero occurrence of alleles in the raw counts, Haldane-Anscombe correction was applied by adding 0.5 to each cell before calculating the odds ratio using the *oddsratio* function from the package *fmsb* in R (Anscombe, 1956; Haldane, 1940; Nakazawa, 2019). Figures of odds ratios were plotted using the *forest* function from the package *metafor* in R (Viechtbauer, 2010)**.** Other figures were plotted using the *ggplot* function from the package *ggplot2* in R (Wickham, 2016).

Multiple Testing Correction and Statistical Analysis

All hypothesis tests were evaluated at a significance level of 0.05. Benjamini-Hochberg procedure was applied to the p-values from Fisher's Exact Tests using the *p.adjus*t function from R's *stats* module to correct for the false discovery rate given the number or tests applied. All analyses were performed in R version 3.6.3 (R Core Team, 2020).

CHAPTER 3

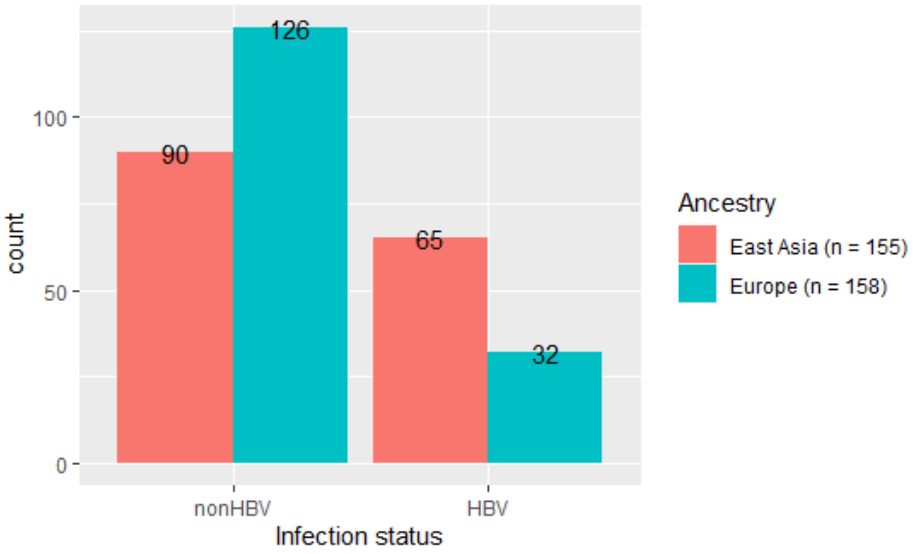
RESULTS

Hepatitis B Infection Status Breakdown by Population Ancestry in TCGA

The ancestry of TCGA patients was obtained from published data by (Taravella Oill et al., 2020). The ancestry was inferred using PopInf based on their autosomes. PopInf uses the five super populations of 1000 Genomes as a reference: African, Ad Mixed American, East Asian, European, and South Asian. Three patient samples were absent from the PopInf data, thus they were removed from downstream analysis. Due to the small sample size of the African, Ad Mixed American, and South Asian populations, only the East Asian (n = 155) and European (n = 158) populations were retained for downstream analysis. Figure 1 shows the number of HBV-mediated and non-HBV-mediated liver cancer patients by population ancestry.

Figure 1

Number of HBV-mediated and nonHBV-mediated liver cancer patients by ancestry



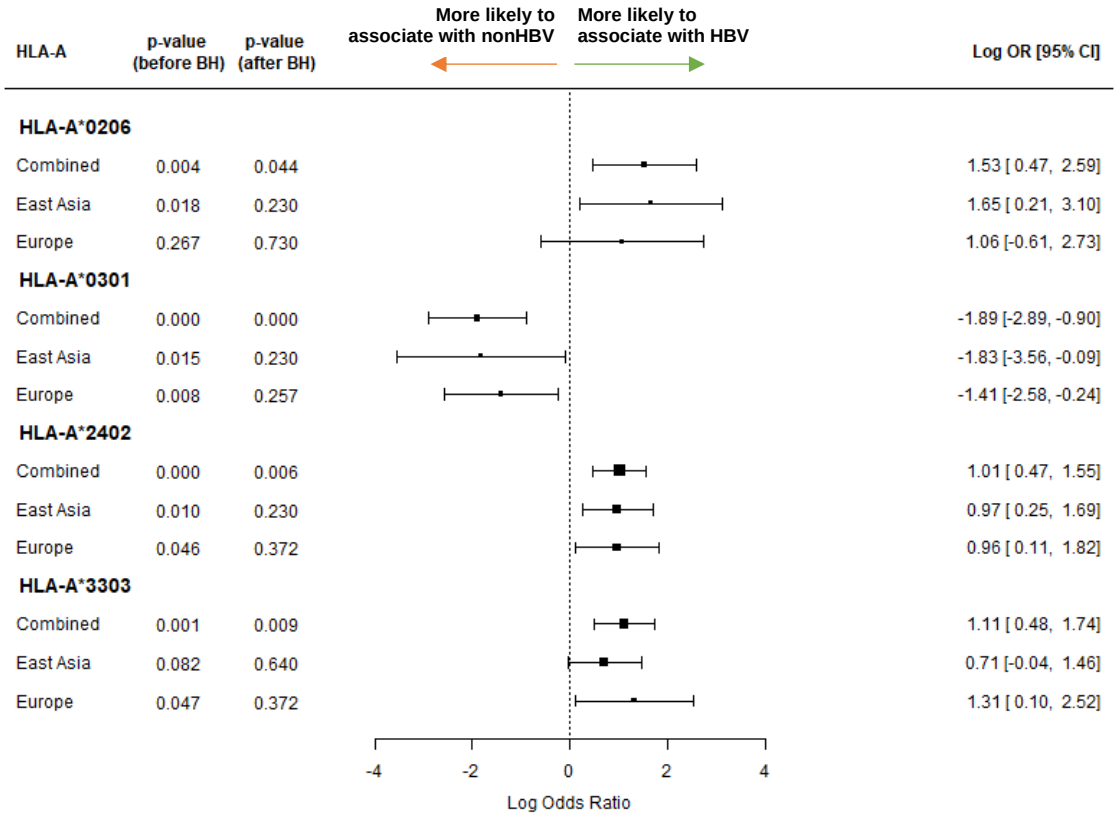
Note. The number of nonHBV-mediated liver cancer patients is higher than the number of HBV-mediated liver cancer patients in the TCGA dataset. Among the nonHBV patients, 58.3% were from Europe whereas 41.7% were from East Asia. Among the HBV patients, 67% were from East Asia whereas 33% were from Europe.

Significant HLA-A Alleles

45 HLA-A alleles were found in the TCGA dataset. Out of these 45 alleles, four alleles are significantly different between HBV-mediated and nonHBV-mediated liver cancer patients in the TCGA dataset (Figure 2). A*0206, A*2402, and A*3303 are associated positively with HBV-mediated liver cancer patients. On the other hand, HLA*0301 is associated negatively with nonHBV-mediated liver cancer patients. However, once stratified by ancestry, none of the alleles are significant but still associate with HBV-mediated liver cancer in the same direction.

Figure 2

HLA-A alleles that are significantly different between HBV-mediated and nonHBV-mediated liver cancer patients in the TCGA dataset

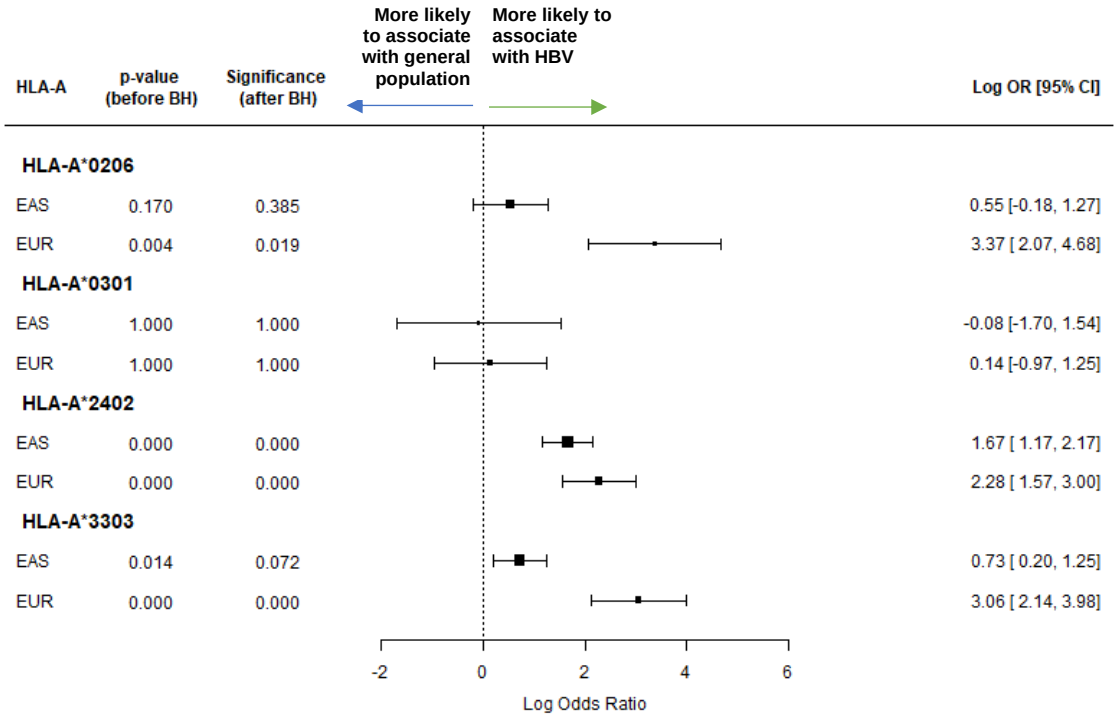


Note. Four alleles are significantly different between HBV-mediated and nonHBV-mediated liver cancer patients in the TCGA dataset. Once stratified by population ancestry, the alleles are no longer significant but their associations with HBV-mediated liver cancer remain the same direction. Positive log odds ratio indicates a positive association with HBV-mediated liver cancer whereas negative log odds ratio indicates a negative association with HBV-mediated liver cancer.

We then compared the allele frequency in HBV-mediated liver cancer patients from the TCGA dataset with the general population from Allele Frequency Net Database. We found that, in both ancestry groups, almost all alleles associate with HBV-mediated liver cancer in the same directions as in the previous comparison (Figure 3). The only exception is A*0301 in the European population, which associates negatively with HBV-mediated liver cancer in HBV vs nonHBV comparison but positively in HBV vs general population comparison.

Figure 3

HLA-A alleles that are significantly different between HBV-mediated liver cancer patients and the general population of East Asian and European ancestry

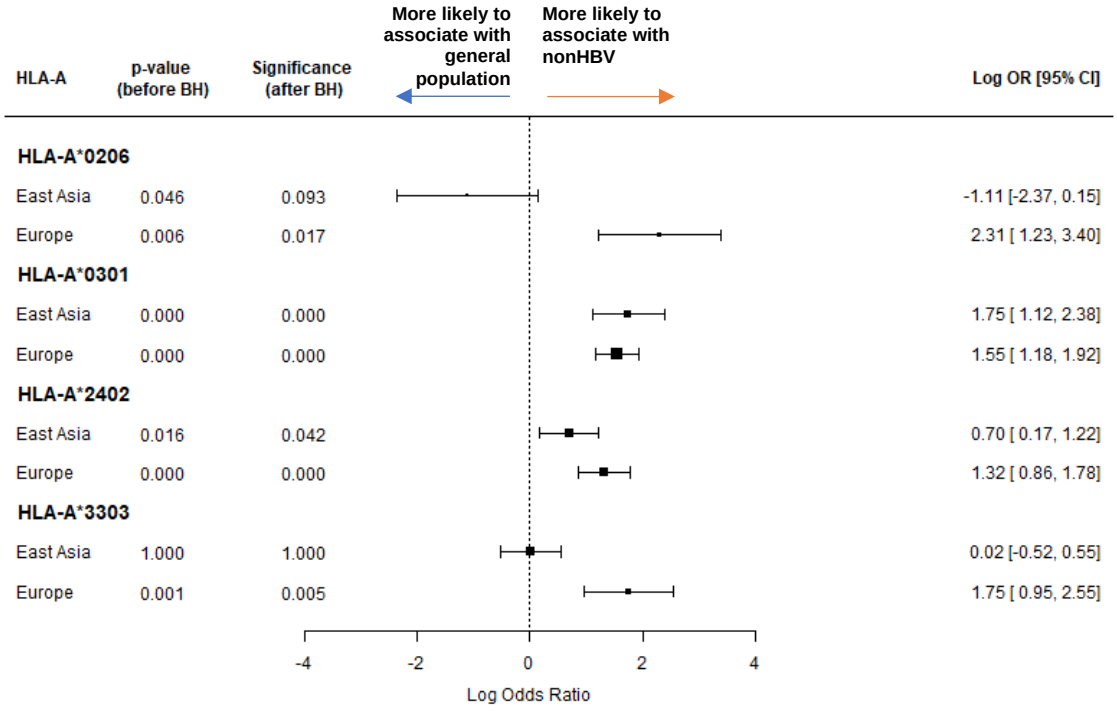


Note. Almost all alleles associate with HBV-mediated liver cancer in positive direction except A*0301 in the East Asian population. Three out of the four alleles are significantly different only in the European population.

We then compared the allele frequency in nonHBV-mediated liver cancer patients from the TCGA dataset with the general population from Allele Frequency Net Database. We found that almost all alleles associate positively with nonHBV-mediated liver cancer except A*0206 in the East Asian population, which associates positively with general population.

Figure 4

HLA-A alleles that are significantly different between nonHBV-mediated liver cancer patients and the general population of East Asian and European ancestry



Note. All alleles are significantly associated with the nonHBV-mediated liver cancer patients in the European population. In the East Asian population, three out of the four alleles associate positively with nonHBV-mediated liver cancer but only two of them are significantly different.

We summarized the results of HLA-A association with HBV-mediated liver cancer, nonHBV-mediated liver cancer, and the general population in Table 2.

Table 2

Comparison of association of HLA-A alleles with HBV-mediated liver cancer, nonHBV-mediated liver cancer, and the general population in East Asian and European populations

HLA-A	HBV vs nonHBV			HBV vs General Population		nonHBV vs General Population	
	Combined	East Asia	Europe	East Asia	Europe	East Asia	Europe
A*0206	**1.529**	1.655	1.062	0.546	**3.375**	-1.109	**2.313**
A*0301	**-1.892**	-1.828	-1.409	-0.080	0.141	**1.748**	**1.550**
A*2402	**1.012**	0.972	0.965	**1.669**	**2.285**	**0.697**	**1.320**
A*3303	**1.112**	0.710	1.310	0.729	**3.062**	0.019	**1.751**

Note. This table compares the association of HLA-A alleles with HBV-mediated liver cancer, nonHBV-mediated liver cancer, and the general population in East Asian and European populations. While there are HLA-A alleles that are significantly before stratifying by ancestry in the HBV vs nonHBV comparison, the alleles are not significantly different once ancestry is considered. When we extended the comparison to general population, most of the alleles remain in the same direction as in the HBV vs nonHBV analysis. There is a difference in allele frequency between nonHBV-mediated liver cancer and the general population.

Green: Associate positively with HBV

Orange: Associate positively with nonHBV

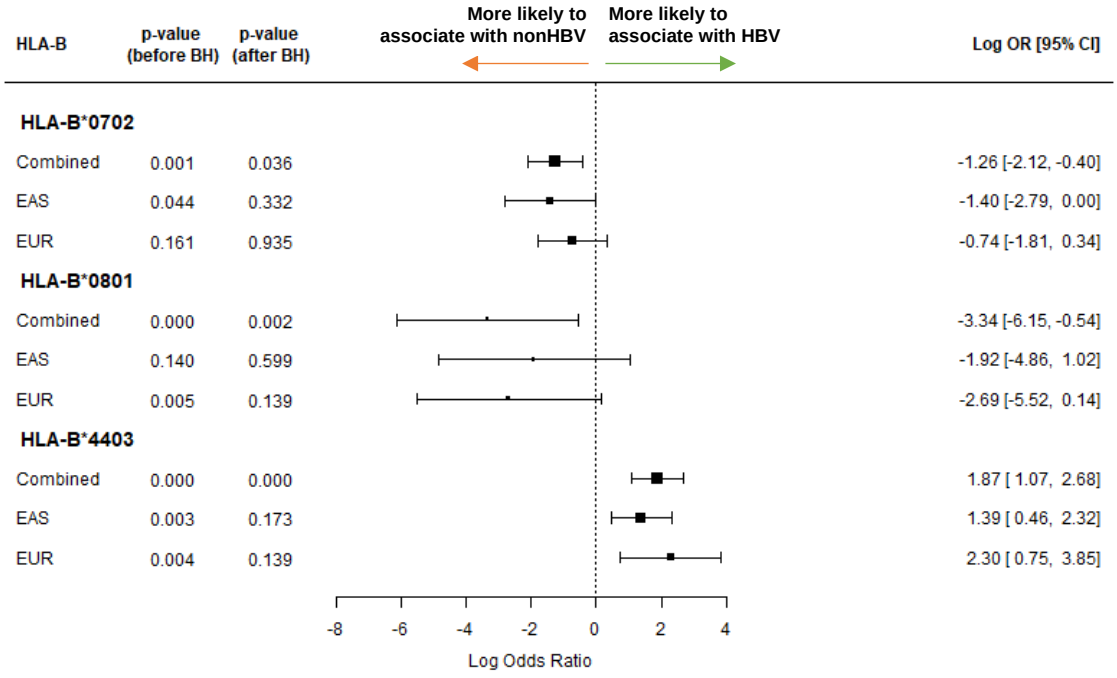
Blue: Associate positively with general population

Significant HLA-B Alleles

79 HLA-B alleles are present in the TCGA dataset. Out of the 79 alleles, three alleles are significantly different between HBV-mediated and nonHBV-mediated liver cancer patients in the TCGA dataset (Figure 5). B*0702 and B*0801 are associated positively with HBV-mediated liver cancer patients. On the other hand, B*4403 is associated negatively with nonHBV-mediated liver cancer patients. However, similar to HLA-A, once stratified by ancestry, none of the alleles remain significantly different.

Figure 5

HLA-B alleles that are significantly different between HBV-mediated liver cancer patients and nonHBV-mediated liver cancer patients

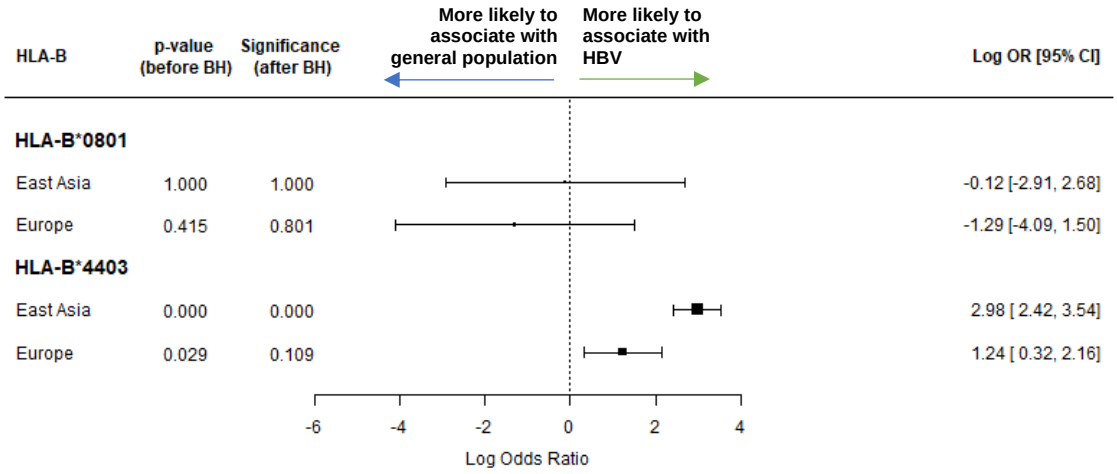


Note. Once stratified by ancestry, B*0702, B*0801, and B*4403 are no longer significantly different but remain associated with HBV-mediated liver cancer in the same direction.

We then compared the allele frequency in HBV-mediated liver cancer patients from the TCGA dataset with the general population from Allele Frequency Net Database. The allele frequency of B*0702 that met the search criteria is absent from the database, therefore it was not included in the analysis. Only B*4403 in the East Asian population is significantly different and associate positively with HBV-mediated liver cancer.

Figure 6

HLA-B alleles that are significantly different between HBV-mediated liver cancer patients and the general population of East Asian and European ancestry

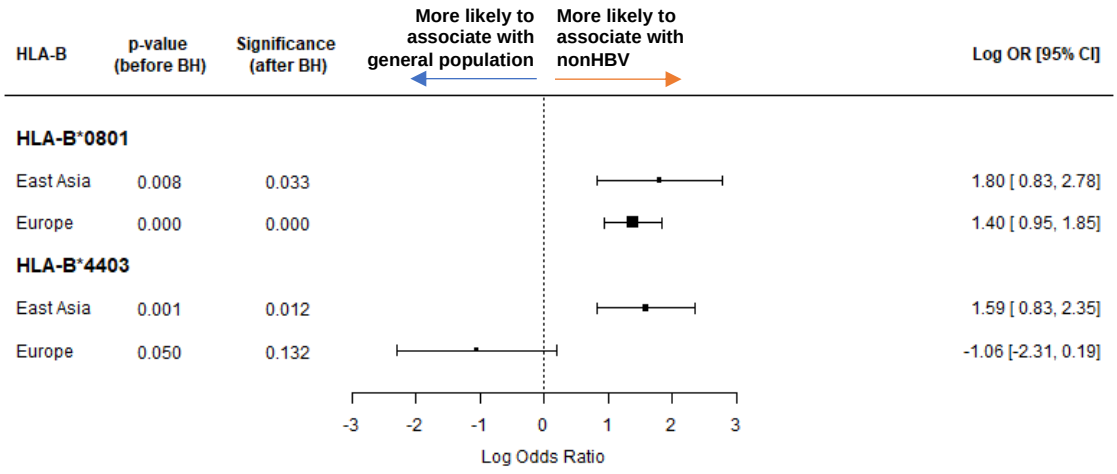


Note. The allele frequency of B*0702 that met the search criteria is absent from the database, therefore it was not included in the analysis. Only B*4403 in the East Asian population is significantly different and associate positively with HBV-mediated liver cancer.

We then tested whether the alleles that are significantly different in the previous two analyses are also significantly different between nonHBV-mediated liver cancer in TCGA and general population. We observed that B*0801 is positively associated with nonHBV-mediated liver cancer in both populations of East Asian and European ancestry (Figure 7). B*4403 is positively associated with nonHBV-mediated liver cancer in individuals of East Asian ancestry but negatively in individuals of European ancestry.

Figure 7

HLA-B alleles that are significantly different between nonHBV-mediated liver cancer patients and the general population of East Asian and European ancestry



Note. B*0702 is not included in the figure because the allele does not have corresponding population samples from the Allele Net Frequency Database that meet the search criteria. B*0801 is positively associated with nonHBV-mediated liver cancer in both East Asian and European populations. B*4403 is positively associated with nonHBV-mediated liver cancer in East Asian population but negatively in European population

We summarized the results of HLA-B association with HBV-mediated liver cancer, nonHBV-mediated liver cancer, and the general population in Table 3.

Table 3

Comparison of association of HLA-B alleles with HBV-mediated liver cancer, nonHBV-mediated liver cancer, and the general population in East Asian and European populations

HLA-B	HBV vs nonHBV			HBV vs General Population		nonHBV vs General Population	
	Combined	East Asia	Europe	East Asia	Europe	East Asia	Europe
B*0702	**-1.261**	-1.397	-0.737				
B*0801	**-3.343**	-1.919	-2.692	-0.115	-1.295	**1.804**	**1.397**
B*4403	**1.873**	1.391	2.300	**2.981**	1.240	**1.591**	-1.059

Note. This table compares the association of HLA-B alleles with HBV-mediated liver cancer, nonHBV-mediated liver cancer, and the general population in East Asian and European populations. B*0702 is missing the log odds ratio because the allele does not have corresponding population samples from the Allele Net Frequency Database that meet the search criteria.

Green: Associate positively with HBV

Orange: Associate positively with nonHBV

Blue: Associate positively with general population

Significant HLA-C Alleles

33 HLA-C alleles are present in the TCGA dataset. Out of the 33 alleles, seven alleles are significantly different between HBV-mediated and nonHBV-mediated liver cancer patients in the TCGA dataset (Figure 4, Table 4). C*0701 and C*1203 are associated negatively with HBV-mediated liver cancer patients. Once stratified by ancestry, none of the seven alleles remain significant.

Figure 8

HLA-C alleles that are significantly different between HBV-mediated and nonHBV-mediated liver cancer patients

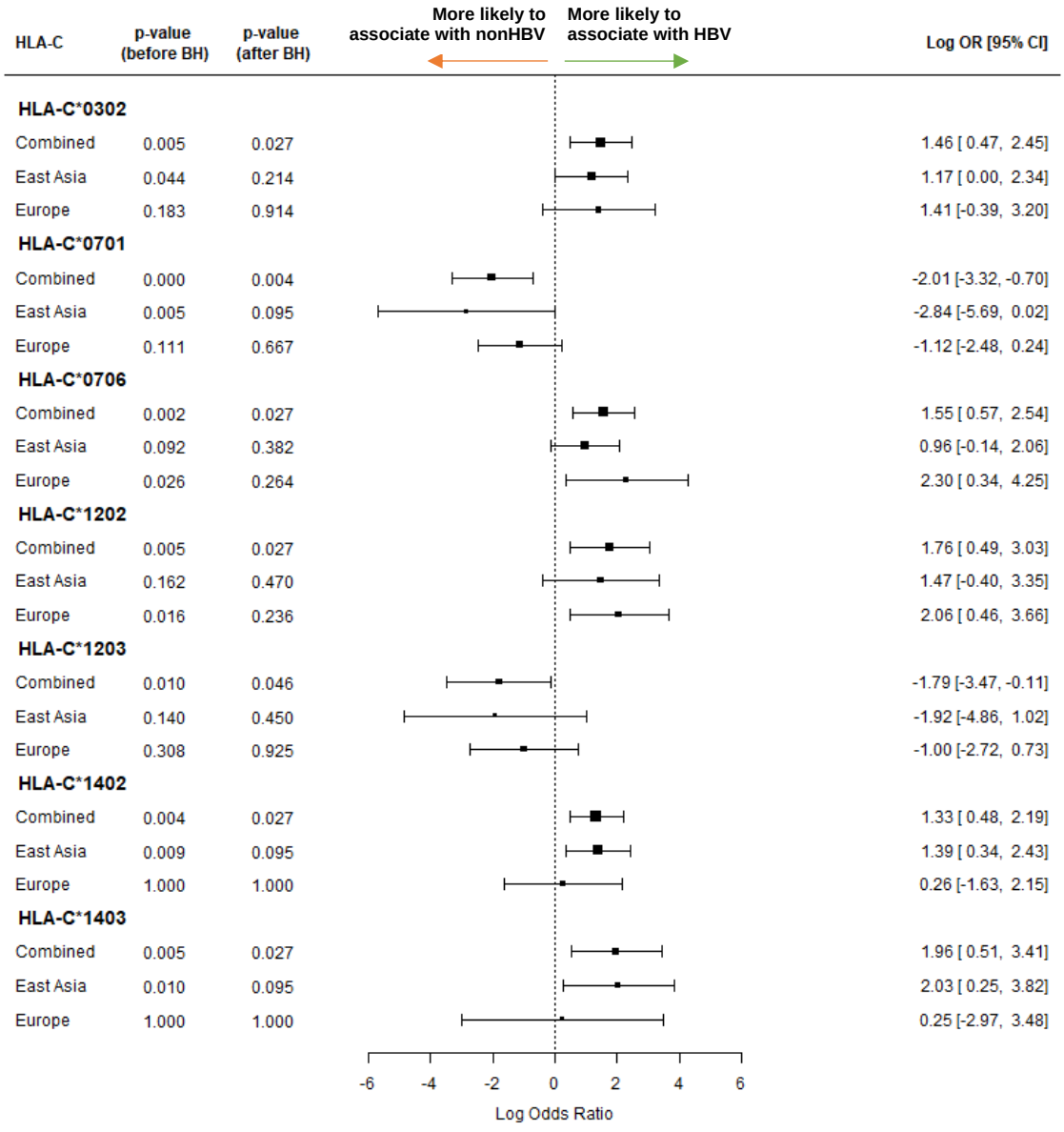


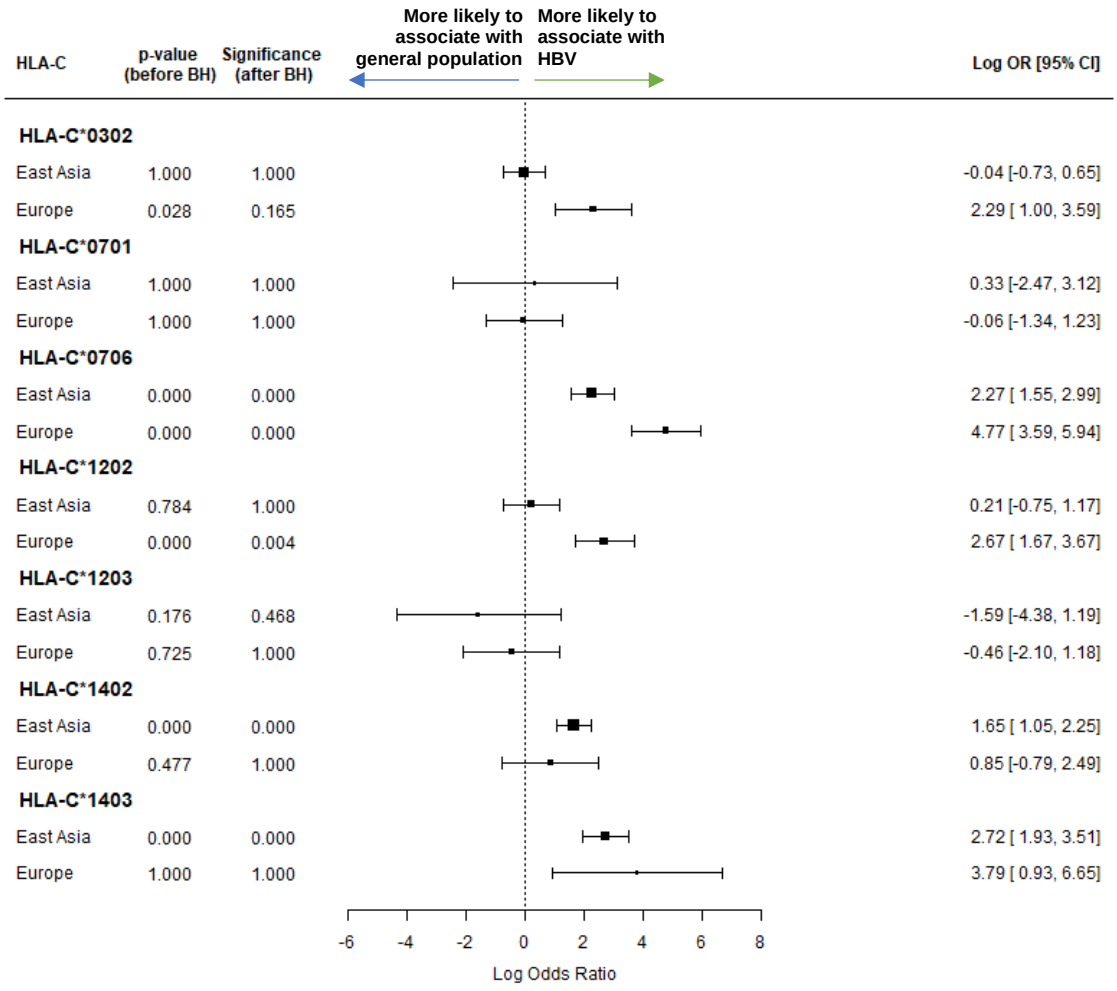
Figure 8 continued

Note. Seven HLA-C alleles are significantly different between HBV-mediated and nonHBV-mediated liver cancer patients in the TCGA dataset. However, once stratified by ancestry, the alleles are no longer significant.

We then compared the frequency of HLA-C alleles in HBV-mediated liver cancer patients from the TCGA dataset with the general population from Allele Frequency Net Database. In the East Asian population, C*0706, C*1402, and C*1403 are significantly different and positively associated with HBV-mediated liver cancer. In the European population, C*0706 and C*1202 are significantly different and positively associated with HBV-mediated liver cancer. Other alleles are not significantly different between HBV-mediated liver cancer and the general population.

Figure 9

HLA-C alleles that are significantly different between HBV-mediated liver cancer patients and the general population of East Asian and European ancestry

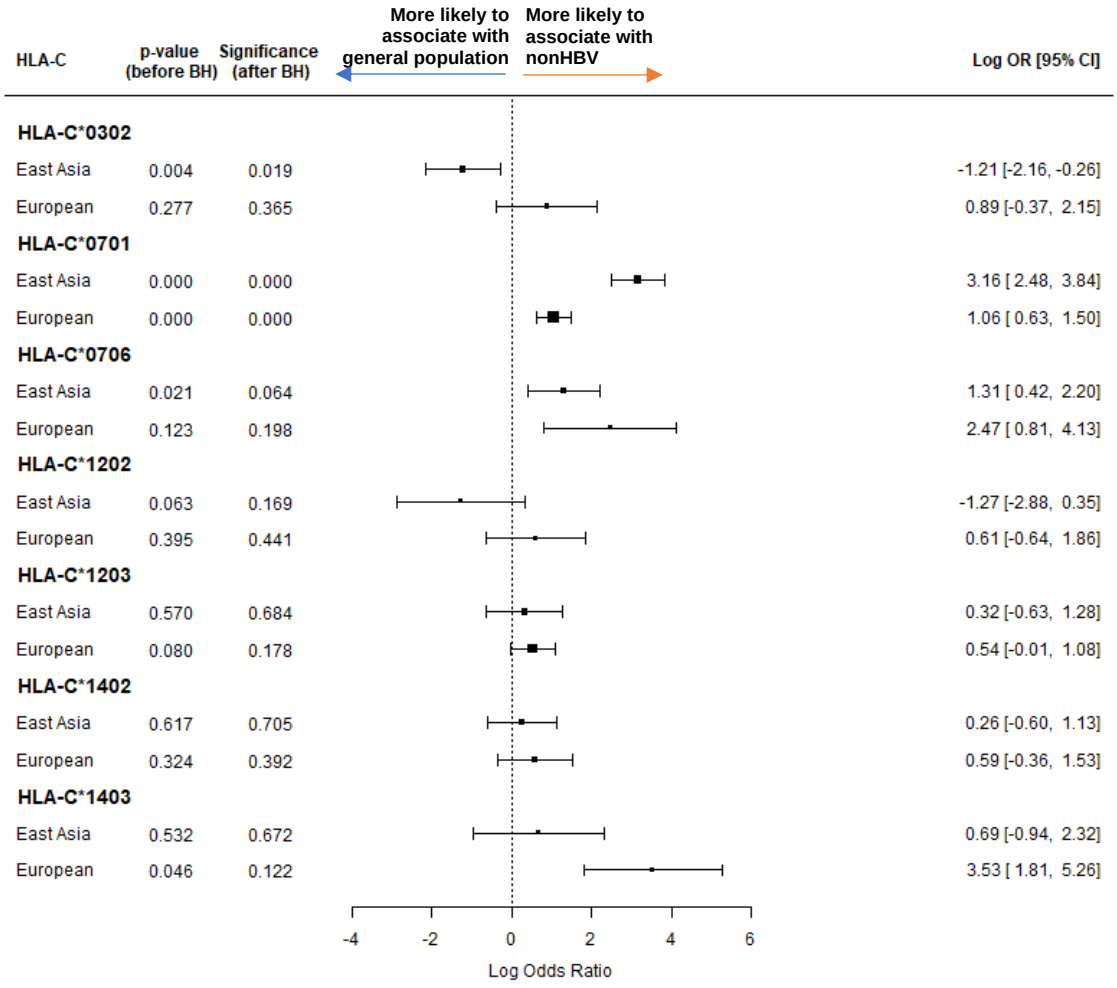


Note. In the East Asian population, C*0706, C*1402, and C*1403 are significantly different and positively associated with HBV-mediated liver cancer. In the European population, C*0706 and C*1202 are significantly different and positively associated with HBV-mediated liver cancer.

We then tested whether the alleles that are significantly different in the previous two analyses are also significantly different between nonHBV-mediated liver cancer in TCGA and general population. C*0701 is significantly different and associates positively with nonHBV-mediated liver cancer in both East Asian and European populations. In contrast, C*0302 is significantly different and associates positively with only general population of East Asian ancestry.

Figure 10

HLA-C alleles that are significantly different between nonHBV-mediated liver cancer patients and the general population of East Asian and European ancestry



Note. C*0701 is significantly different and associates positively with nonHBV-mediated liver cancer in both East Asian and European populations. In contrast, C*0302 is significantly different and associates positively with only general population of East Asian ancestry. Other alleles are not significantly different.

We summarized the results of HLA-C association with HBV-mediated liver cancer, nonHBV-mediated liver cancer, and the general population in Table 4.

Table 4

Comparison of association of HLA-C alleles with HBV-mediated liver cancer, nonHBV-mediated liver cancer, and the general population in East Asian and European populations

HLA-C	HBV vs nonHBV			HBV vs General Population		nonHBV vs General Population	
	Combined	East Asia	Europe	East Asia	Europe	East Asia	Europe
C*0302	**1.460**	1.173	1.407	-0.036	2.295	**-1.209**	0.888
C*0701	**-2.011**	-2.838	-1.120	0.325	-0.055	**3.164**	**1.065**
C*0706	**1.555**	0.961	2.295	**2.269**	**4.767**	1.308	2.471
C*1202	**1.757**	1.474	2.062	0.208	**2.671**	-1.265	0.609
C*1203	**-1.791**	-1.919	-0.996	-1.595	-0.461	0.324	0.535
C*1402	**1.333**	1.386	0.260	**1.649**	0.846	0.263	0.586
C*1403	**1.962**	2.035	0.252	**2.720**	3.785	0.686	3.533

Note. This table compares the association of HLA-C alleles with HBV-mediated liver cancer, nonHBV-mediated liver cancer, and the general population in East Asian and European populations.

Green: Associate positively with HBV

Orange: Associate positively with nonHBV

Blue: Associate positively with general population

Significant HLA-DP Alleles

Out of 86 paired HLA-DPA-DPB alleles, 6 allele pairs are significantly different between HBV-mediated and nonHBV-mediated liver cancer patients in the TCGA dataset (Figure 11). In particular, DPA10103-DPB10501, DPA10201-DPB10501, DPA10202-DPB10201, and DPA10202-DPB10501 are associated positively with HBV-mediated liver cancer patients of European ancestry but not of East Asian ancestry. On the other hand, DPA10103-DPB10401 is associated negatively with HBV-mediated liver cancer patients. The allele pair remains significant in patients of both East Asian and European ancestry after stratification.

Figure 11

HLA-DP alleles that are significantly different between HBV-mediated liver cancer patients and nonHBV-mediated liver cancer patients by ancestry

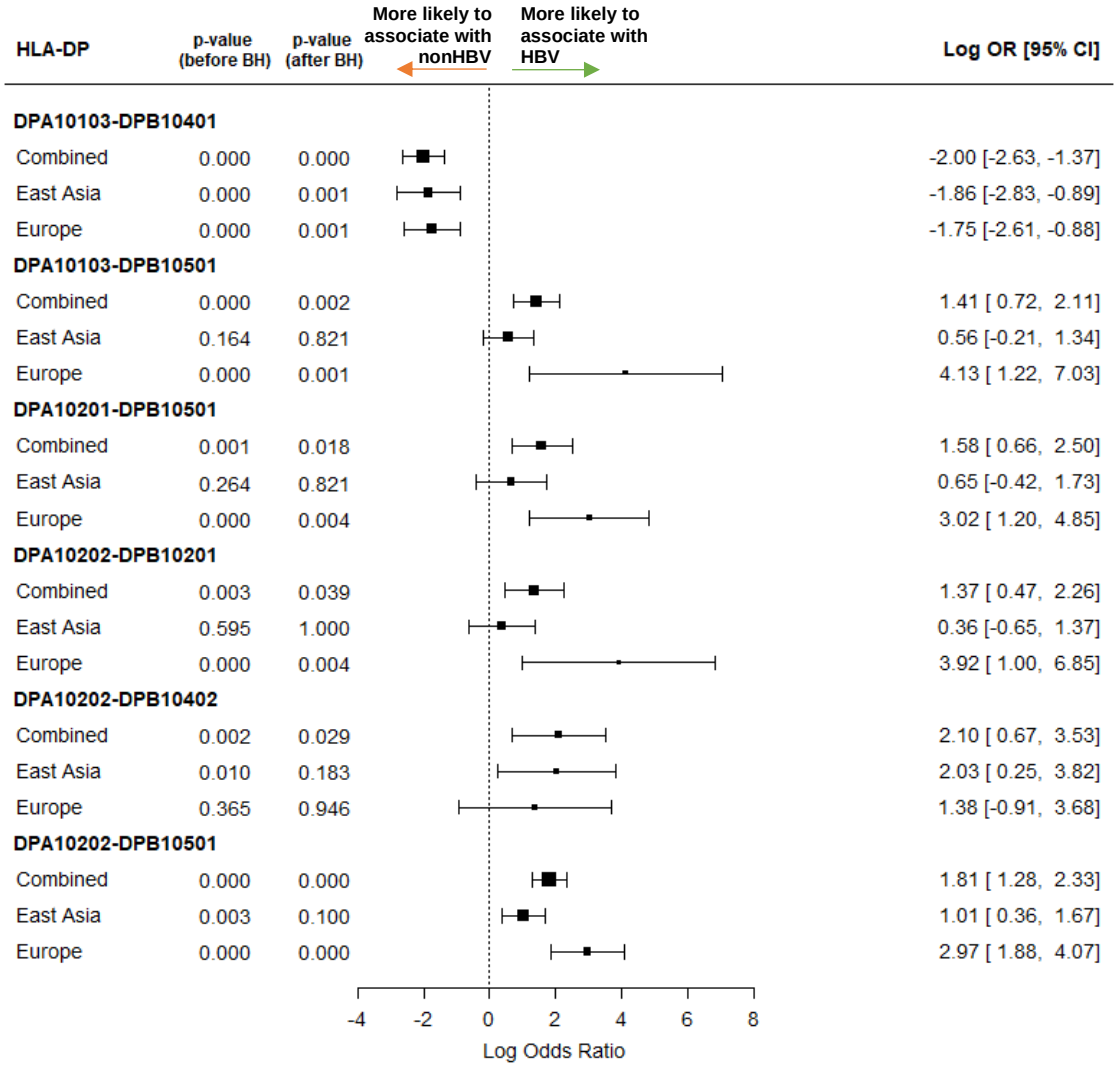


Figure 11 continued

Note. Six allele pairs are significantly different between HBV-mediated and nonHBV-mediated liver cancer patients in the TCGA dataset. Four of the six allele pairs are associated positively with HBV-mediated liver cancer patients of European ancestry but not of East Asian ancestry. In contrast, DPA10103-DPB10401 is associated negatively with HBV-mediated liver cancer patients before and after stratifying by ancestry.

We were not able to find HLA-DP haplotypes that meet the search criteria from the Allele Frequency Net Database, therefore, we did not compare the allele frequency in HBV-mediated and nonHBV-mediated liver cancer patients from the TCGA dataset to the general population. We summarized the results from HBV-mediated vs nonHBV-mediated liver cancer patients comparison in Table 5.

Table 5

Comparison of association of HLA-DP haplotypes with HBV-mediated liver cancer, nonHBV-mediated liver cancer, and the general population in East Asian and European populations

HLA-DP	HBV vs nonHBV			HBV vs General Population		nonHBV vs General Population	
	Combined	East Asia	Europe	East Asia	Europe	East Asia	Europe
DPA10103-DPB10401	**-2.00**	**-1.86**	**-1.75**				
DPA10103-DPB10501	**1.41**	0.56	**4.13**				
DPA10201-DPB10501	**1.58**	0.65	**3.02**				
DPA10202-DPB10201	**1.37**	0.36	**3.92**				
DPA10202-DPB10402	**2.10**	2.03	1.38				
DPA10202-DPB10501	**1.81**	1.01	**2.97**				

Note. HLA-DP haplotypes that meet search criteria were not found from the Allele Frequency Net Database, therefore we did not compare the association of HLA-DP alleles with HBV-mediated liver cancer, nonHBV-mediated liver cancer, and the general population in East Asian and European populations.

Green: Associate positively with HBV

Orange: Associate positively with nonHBV

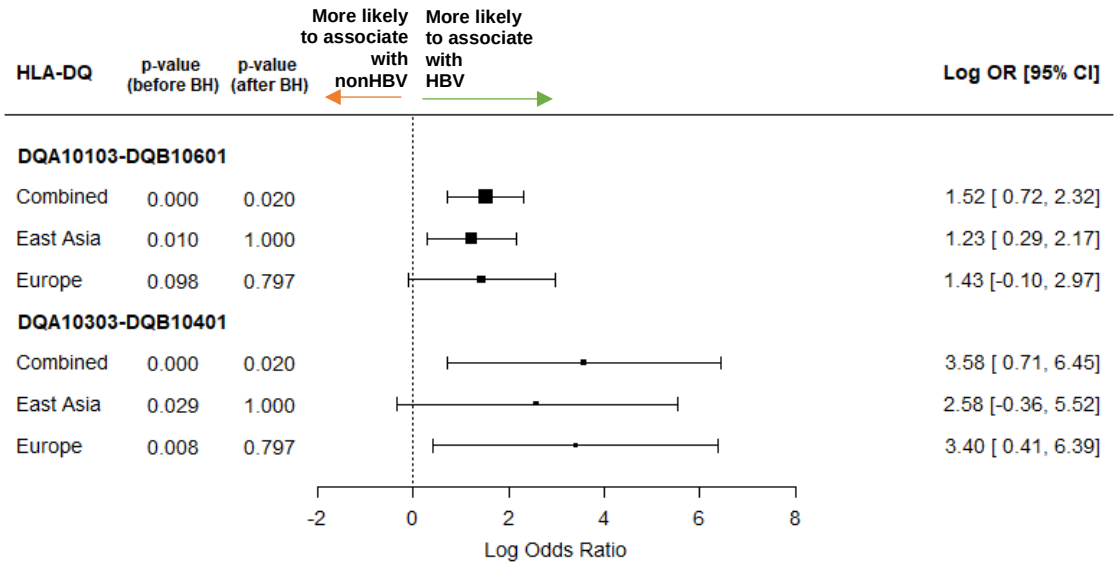
Blue: Associate positively with general population

Significant HLA-DQ Alleles

In general, out of 172 paired HLA-DQA-DQB alleles, only 2 allele pairs are significantly different between HBV-mediated and nonHBV-mediated liver cancer patients in the TCGA dataset (Figure 12). DQA*10103-DQB*10601 and DQA1*0303-DQB1*0401 are associated positively with HBV-mediated liver cancer patients. However, both allele pairs are not significant after stratifying by ancestry.

Figure 12

HLA-DQ alleles that are significantly different between HBV-mediated liver cancer patients and nonHBV-mediated liver cancer patients by ancestry



Note. Only two allele pairs are significantly different between HBV-mediated liver cancer patients and nonHBV-mediated liver cancer patients. Both allele pairs are associated positively with HBV-mediated liver cancer patients.

Similar to HLA-DP, we were not able to find HLA-DQ haplotypes that meet the search criteria from the Allele Frequency Net Database, therefore, we did not compare the allele frequency in HBV-mediated and nonHBV-mediated liver cancer patients from the TCGA dataset to the general population. We summarized the results from HBV-mediated vs nonHBV-mediated liver cancer patients comparison in Table 6.

Table 6

Comparison of association of HLA-DQ haplotypes with HBV-mediated liver cancer, nonHBV-mediated liver cancer, and the general population in East Asian and European populations

HLA-DQ	**HBV vs nonHBV**			**HBV vs General Population**		**nonHBV vs General Population**	
	Combined	**East Asia**	**Europe**	**East Asia**	**Europe**	**East Asia**	**Europe**
DQA10103-DQB10601	**1.52**	1.23	1.43				
DQA10103-DQB10401	**3.58**	2.58	3.40				

Note. HLA-DQ haplotypes that meet search criteria were not found from the Allele Frequency Net Database, therefore we did not compare the association of HLA-DP alleles with HBV-mediated liver cancer, nonHBV-mediated liver cancer, and the general population in East Asian and European populations.

Green: Associate positively with HBV

Orange: Associate positively with nonHBV

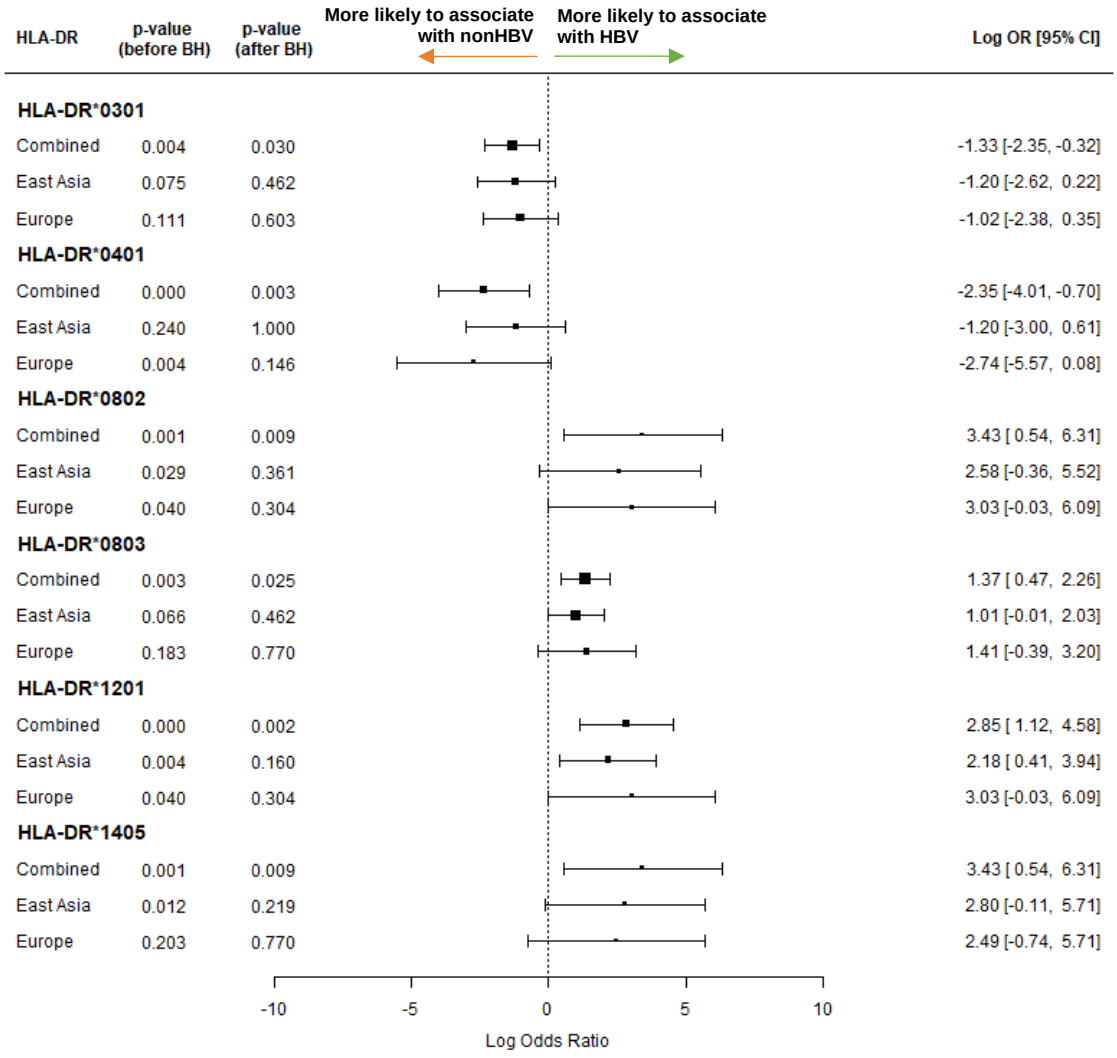
Blue: Associate positively with general population

Significant HLA-DR alleles

45 HLA-DR alleles are present in the TCGA dataset. Out of the 45 alleles, six alleles are significantly different between HBV-mediated and nonHBV-mediated liver cancer patients in the TCGA dataset (Figure 13). Only DR*0301 and DR*0401 are negatively associated with HBV liver cancer patients, the remaining alleles are positively associated with HBV liver cancer patients. Similar to class I alleles, these HLA-DR alleles are not significantly different once stratified by ancestry.

Figure 13

HLA-DR alleles that are significantly different between HBV-mediated liver cancer patients and nonHBV-mediated liver cancer patients

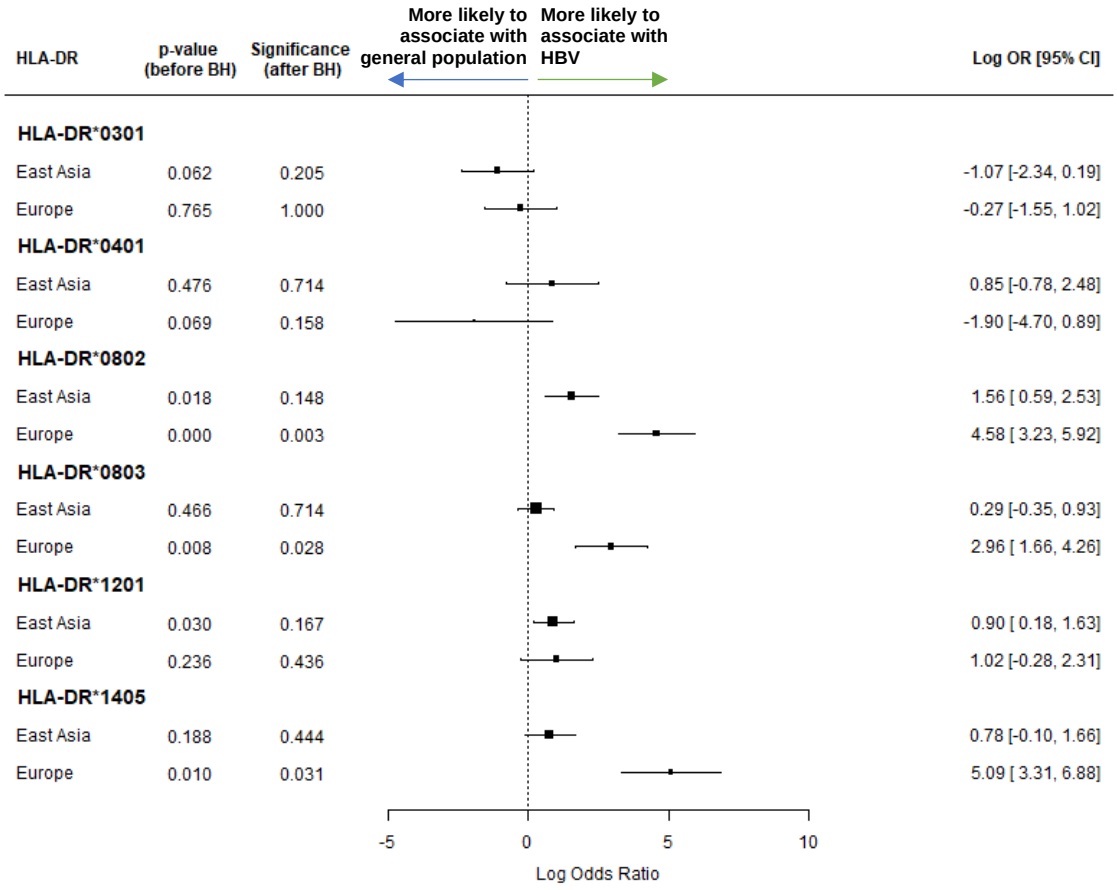


Note. Six alleles are significantly different between HBV-mediated and nonHBV-mediated liver cancer patients in the TCGA dataset. Similar to Class I HLA, once stratified by ancestry, the alleles are no longer significantly different but still associate with HBV-mediated liver cancer in the same direction.

We then compared the frequency of HLA-C alleles in HBV-mediated liver cancer patients from the TCGA dataset with the general population from Allele Frequency Net Database. DR*0802, DR*0803, and DR*1405 are significantly different and associate positively with HBV-mediated liver cancer in only the European population.

Figure 14

HLA-DR alleles that are significantly different between HBV-mediated liver cancer patients and the general population



Note. DR*0802, DR*0803, and DR*1405 are significantly different and associate positively with HBV-mediated liver cancer in only the European population.

We then tested whether the alleles that are significantly different in the previous two analyses are also significantly different between nonHBV-mediated liver cancer in TCGA and general population. Only DR*0301 and DR*0401 are significantly different and associate positively with nonHBV-mediated liver cancer patients of East Asian ancestry.

Figure 15

HLA-DR alleles that are significantly different between nonHBV-mediated liver cancer patients and the general population.

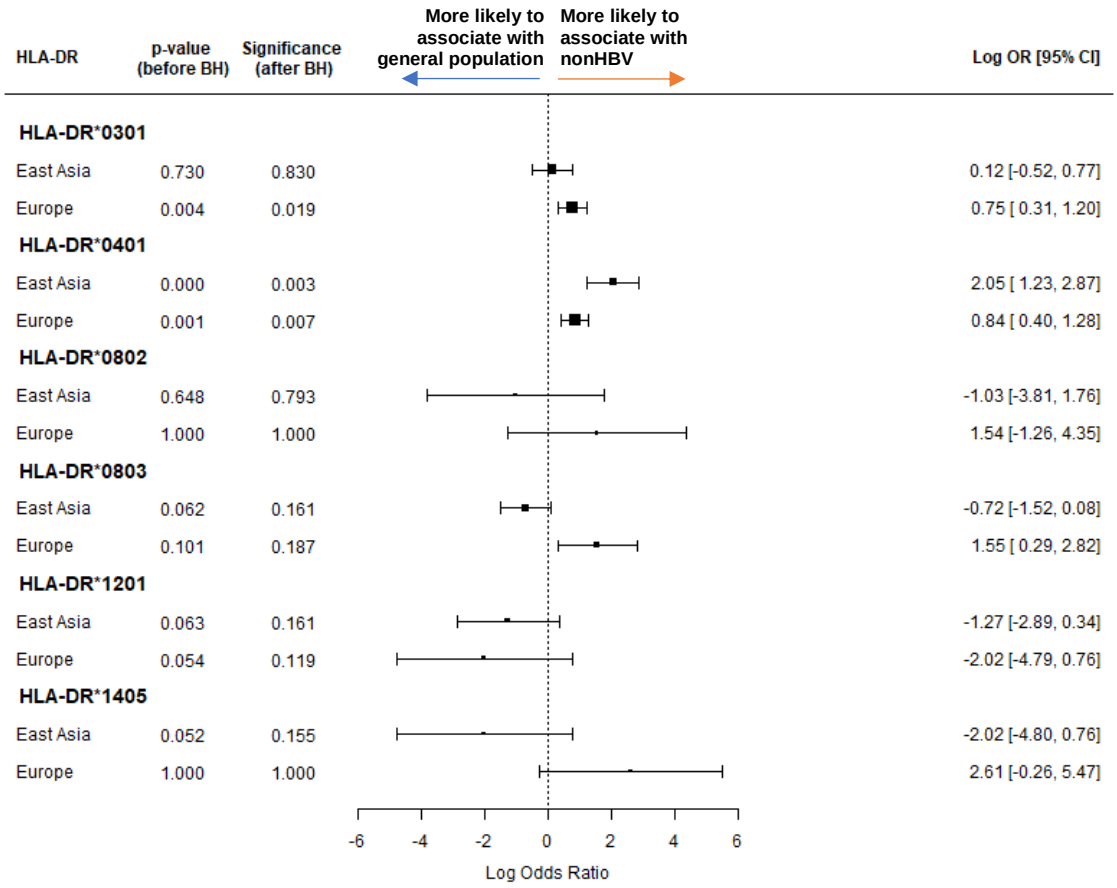


Figure 15 continued

Note. Only DR*0301 and DR*0401 are significantly different and associate positively with nonHBV-mediated liver cancer patients of East Asian ancestry.

We summarized the results of HLA-DR association with HBV-mediated liver cancer, nonHBV-mediated liver cancer, and the general population in Table 7.

Table 7

Comparison of association of HLA-DR alleles with HBV-mediated liver cancer, nonHBV-mediated liver cancer, and the general population in East Asian and European populations

HLA-DR	HBV vs nonHBV			HBV vs General Population		nonHBV vs General Population	
	Combined	East Asia	Europe	East Asia	Europe	East Asia	Europe
DR*0301	**-1.333**	-1.198	-1.019	-1.075	-0.265	0.123	**0.754**
DR*0401	**-2.353**	-1.196	-2.743	0.851	-1.905	**2.048**	**0.839**
DR*0802	**3.426**	2.584	3.032	1.556	**4.575**	-1.027	1.543
DR*0803	**1.366**	1.009	1.407	0.290	**2.961**	-0.719	1.555
DR*1201	**2.847**	2.177	3.032	0.903	1.016	-1.274	-2.016
DR*1405	**3.426**	2.801	2.489	0.781	**5.095**	-2.020	2.606

Table 7 continued

Note. This table compares the association of HLA-DR alleles with HBV-mediated liver cancer, nonHBV-mediated liver cancer, and the general population in East Asian and European populations.

Green: Associate positively with HBV

Orange: Associate positively with nonHBV

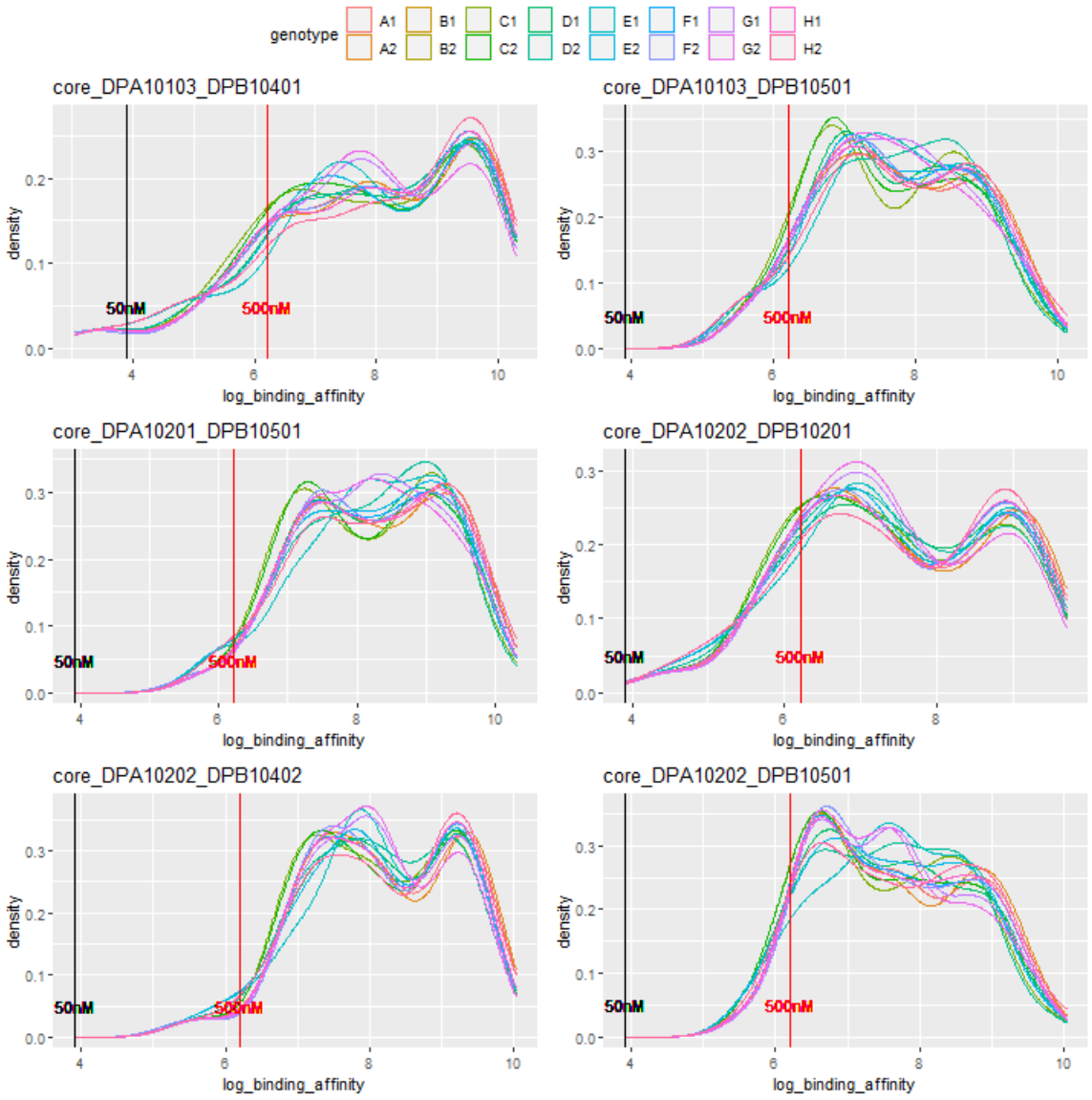
Blue: Associate positively with general population

Binding Affinity Distribution of Significant HLA-DP to HBV Coding Sequences

We plotted the binding affinity distribution of significant HLA-DP haplotypes to HBV coding sequences of 8 genotypes (Figures 16-23). Overall, the binding affinity distribution of HLA-DP haplotypes to all 16 sequences of the same protein is similar. Notably, DPA10103-DPB10401 is the only haplotype that binds strongly to at least one peptide of all HBV coding sequences, with binding affinity of less than 50nM. DPA10202-DPB10201 binds strongly to only peptides from large and medium surface proteins, polymerase, and spliced proteins. Other HLA-DP haplotypes bind to all HBV coding sequences either moderately (50nm < x < 500nM) or weakly (> 500nM).

Figure 16

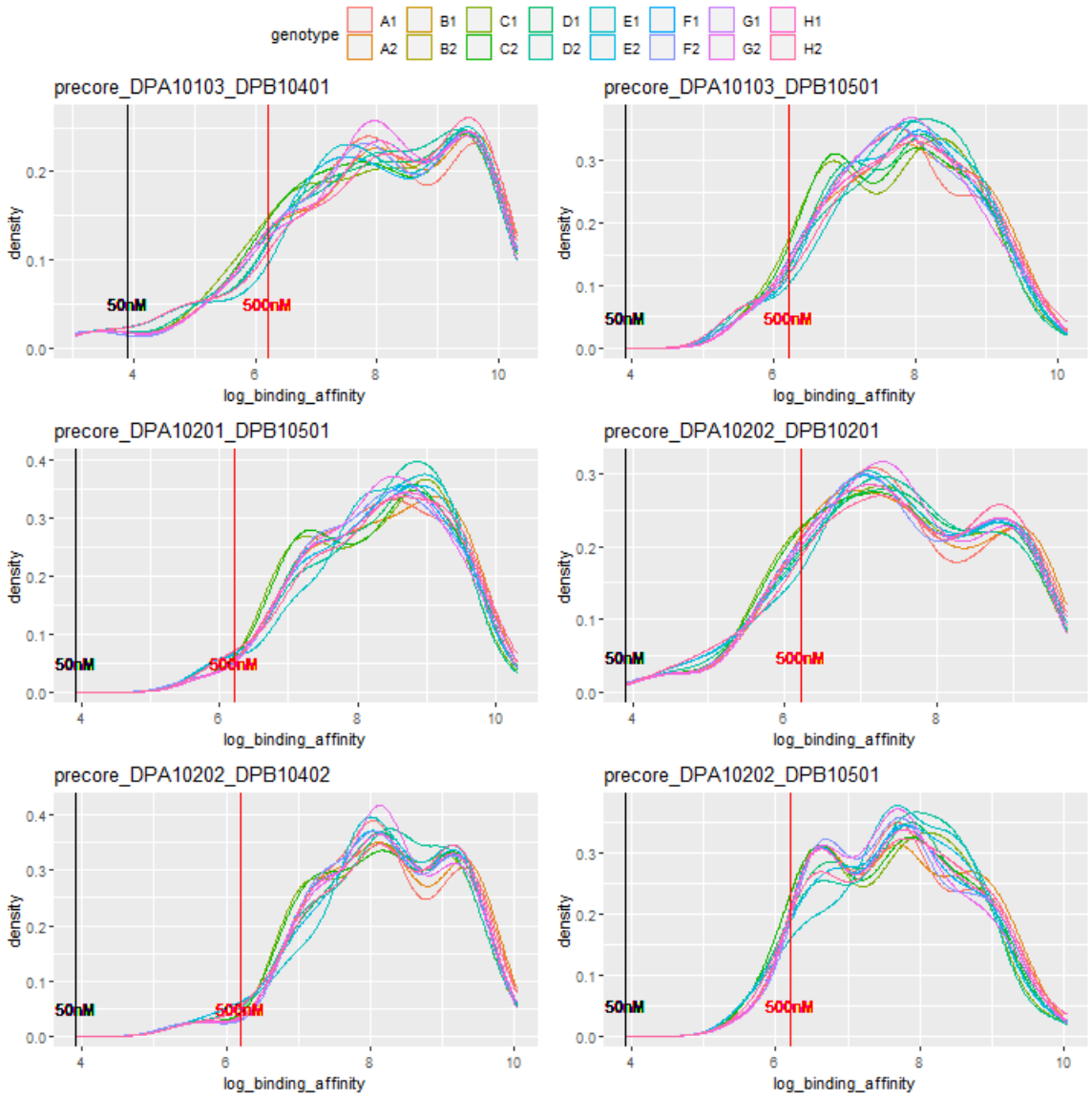
Binding affinity distribution of HLA-DP haplotypes to HBV core protein peptides



Note. DPA10103-DPB10401 binds to some of the HBV core peptides strongly whereas other HLA-DP haplotypes do not bind to any HBV core peptides strongly.

Figure 17

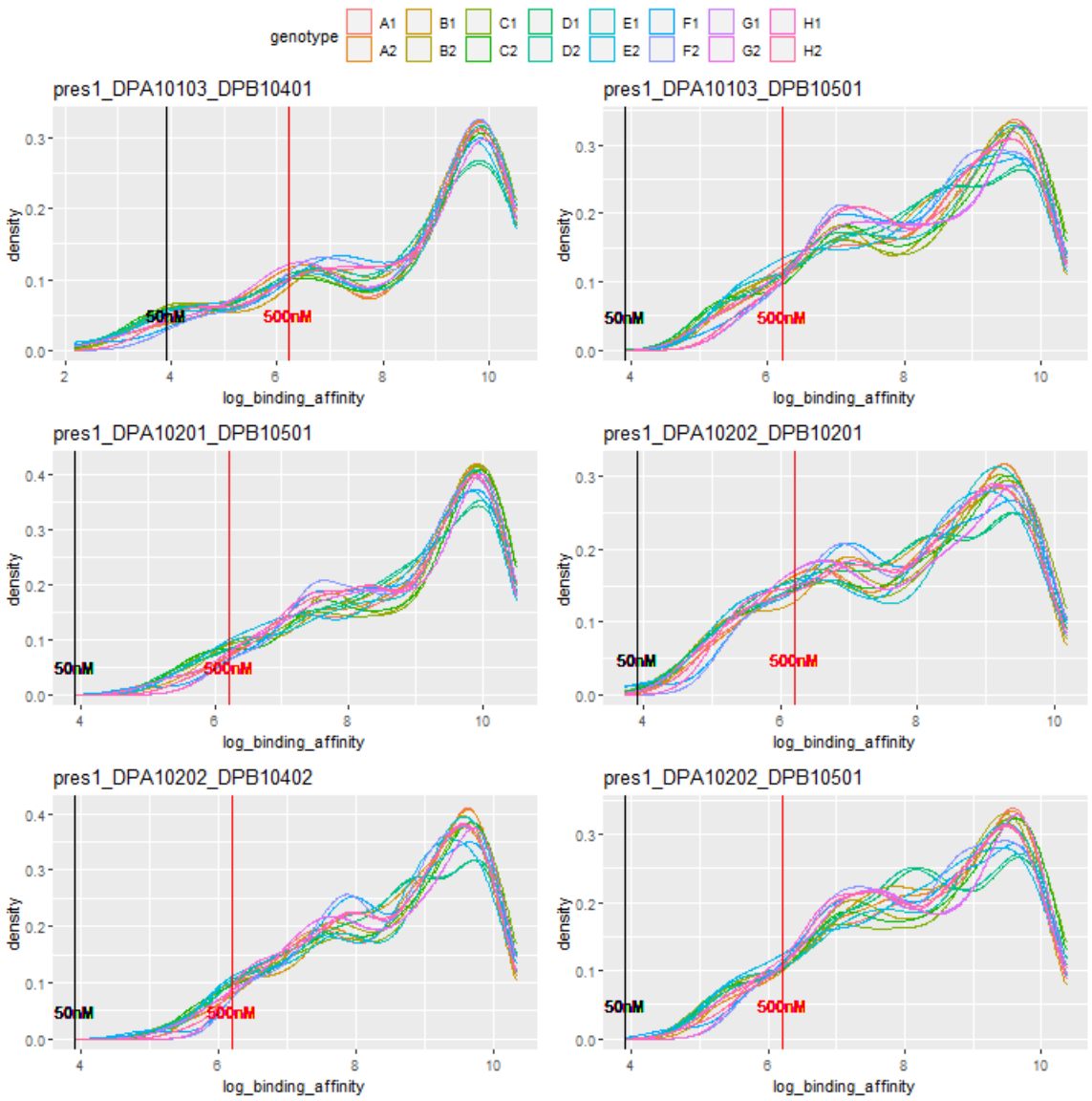
Binding affinity distribution of HLA-DP haplotypes to HBV precore protein peptides



Note. DPA10103-DPB10401 binds to some of the HBV precore peptides strongly whereas other HLA-DP haplotypes do not bind to any HBV precore peptides strongly.

Figure 18

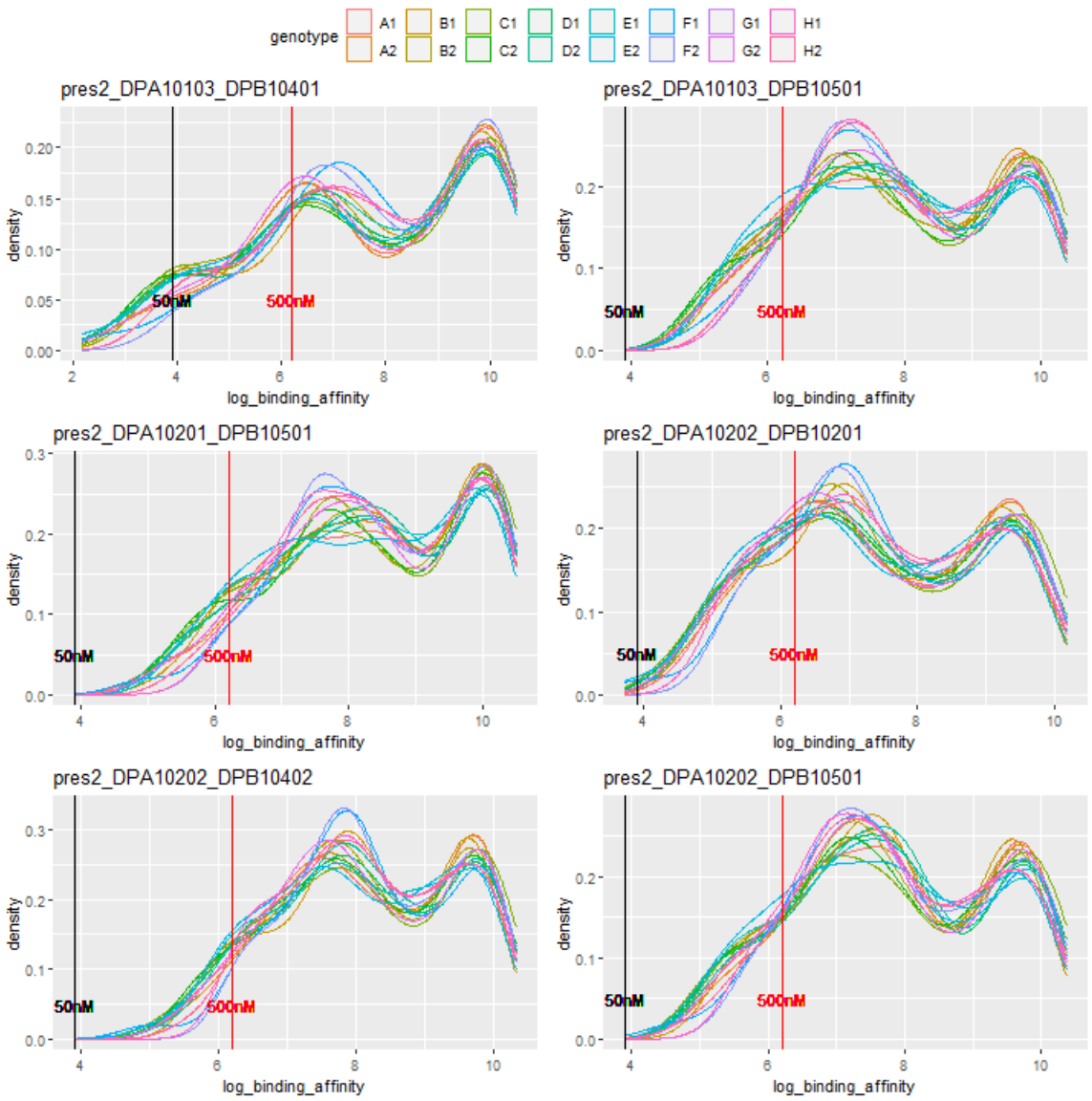
Binding affinity distribution of HLA-DP haplotypes to HBV large surface protein peptides



Note. DPA10103-DPB10401 and DPA10202-DPB10201 bind to some of the HBV large surface protein peptides strongly. Other HLA-DP haplotypes do not bind to any HBV large surface protein peptides strongly.

Figure 19

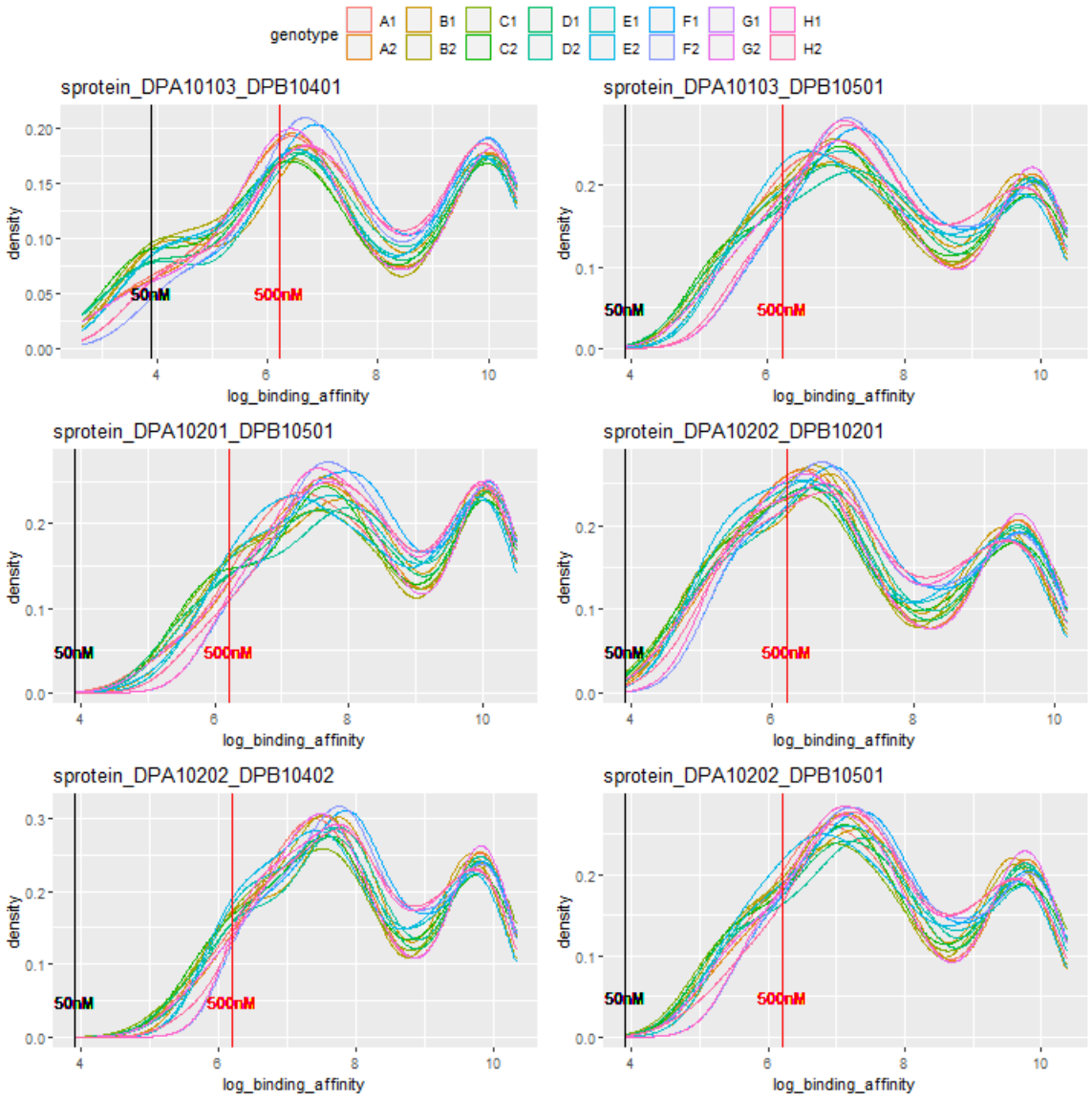
Binding affinity distribution of HLA-DP haplotypes to HBV medium surface protein peptides



Note. DPA10103-DPB10401 and DPA10202-DPB10201 binds to some of the HBV medium surface protein peptides strongly. Other HLA-DP haplotypes do not bind to any HBV medium surface protein peptides strongly.

Figure 20

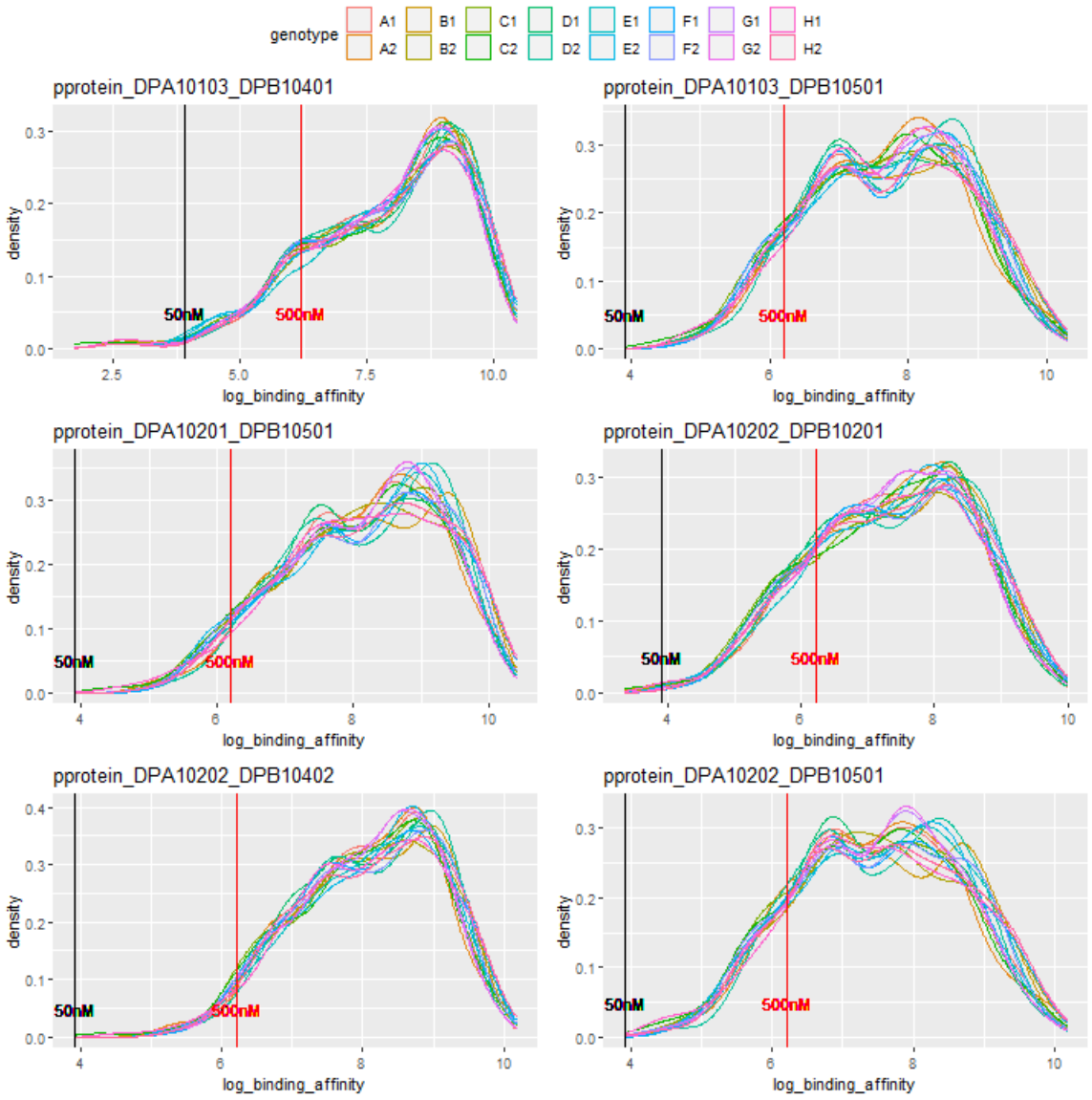
Binding affinity distribution of HLA-DP haplotypes to HBV small surface protein peptides



Note. DPA10103-DPB10401 binds to some of the HBV small surface protein peptides strongly whereas other HLA-DP haplotypes do not bind to any HBV small surface protein peptides strongly.

Figure 21

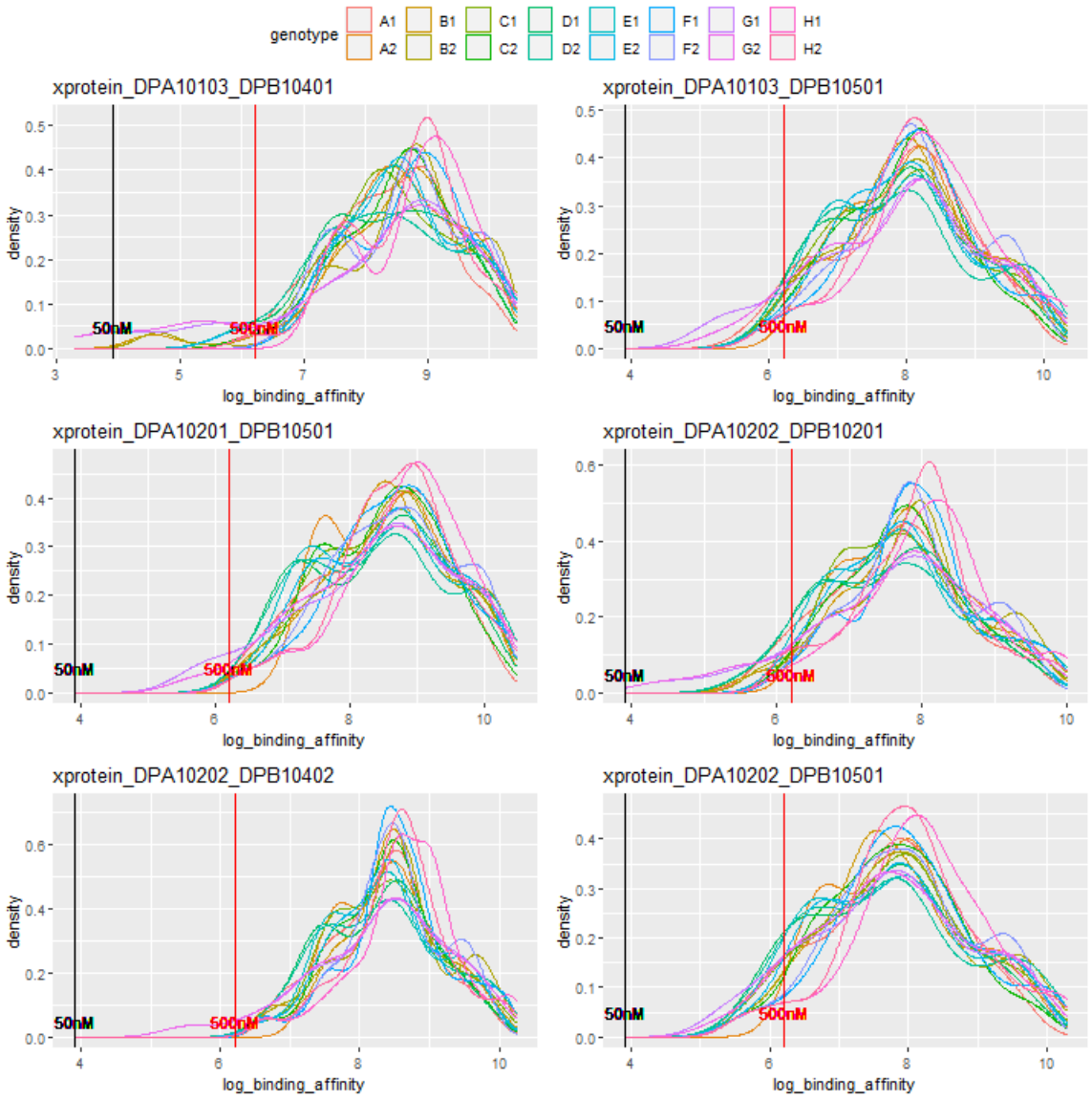
Binding affinity distribution of HLA-DP haplotypes to HBV polymerase peptides



Note. DPA10103-DPB10401 and DPA10202-DPB10201 bind to some of the HBV polymerase peptides strongly. Other HLA-DP haplotypes do not bind to any HBV polymerase peptides strongly.

Figure 22

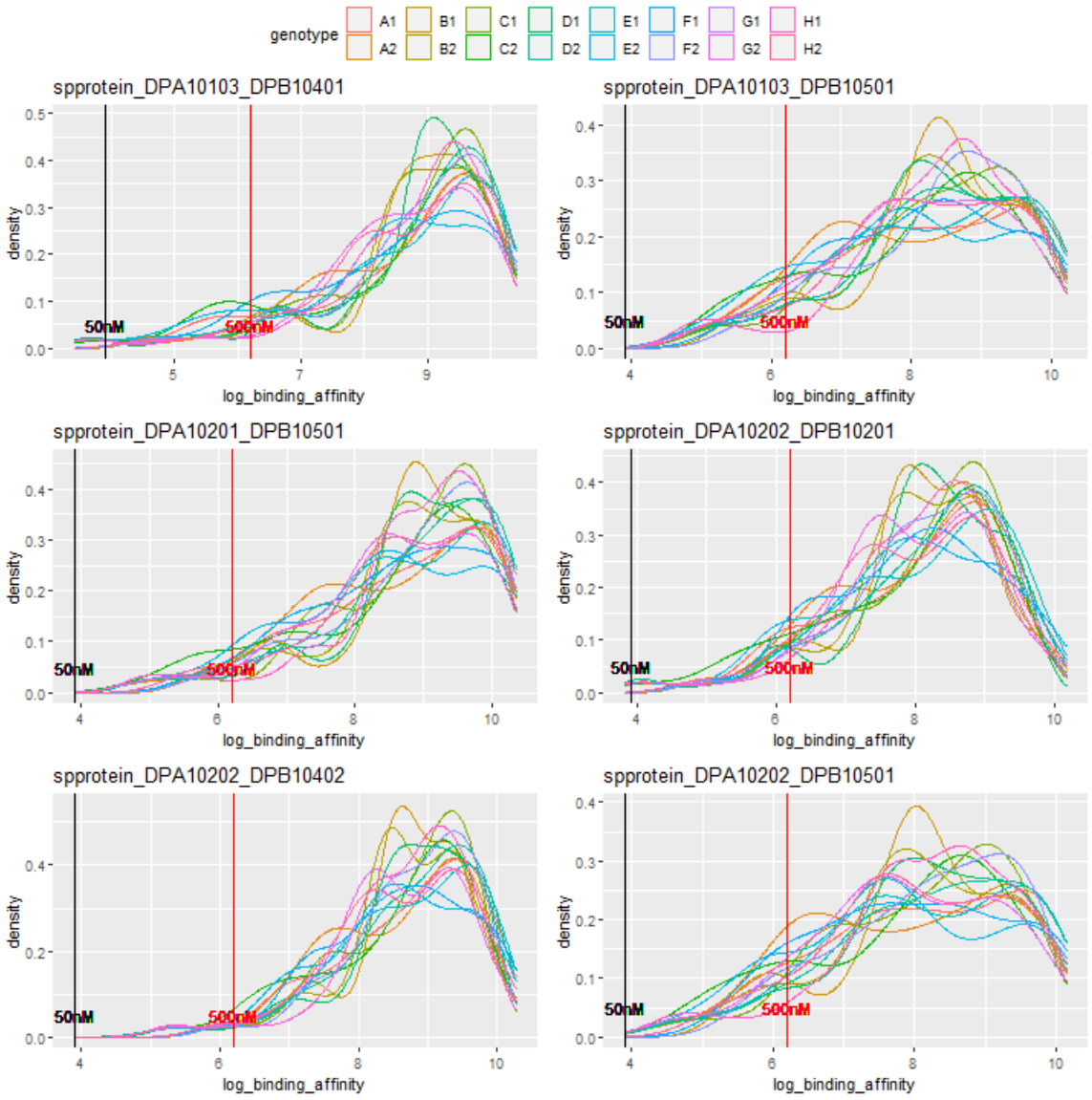
Binding affinity distribution of HLA-DP haplotypes to HBV X protein peptides



Note. DPA10103-DPB10401 binds to some of the HBV protein X peptides of all genotypes strongly whereas other HLA-DP haplotypes do not bind to any HBV protein X peptides strongly.

Figure 23

Binding affinity distribution of HLA-DP haplotypes to HBV spliced protein peptides

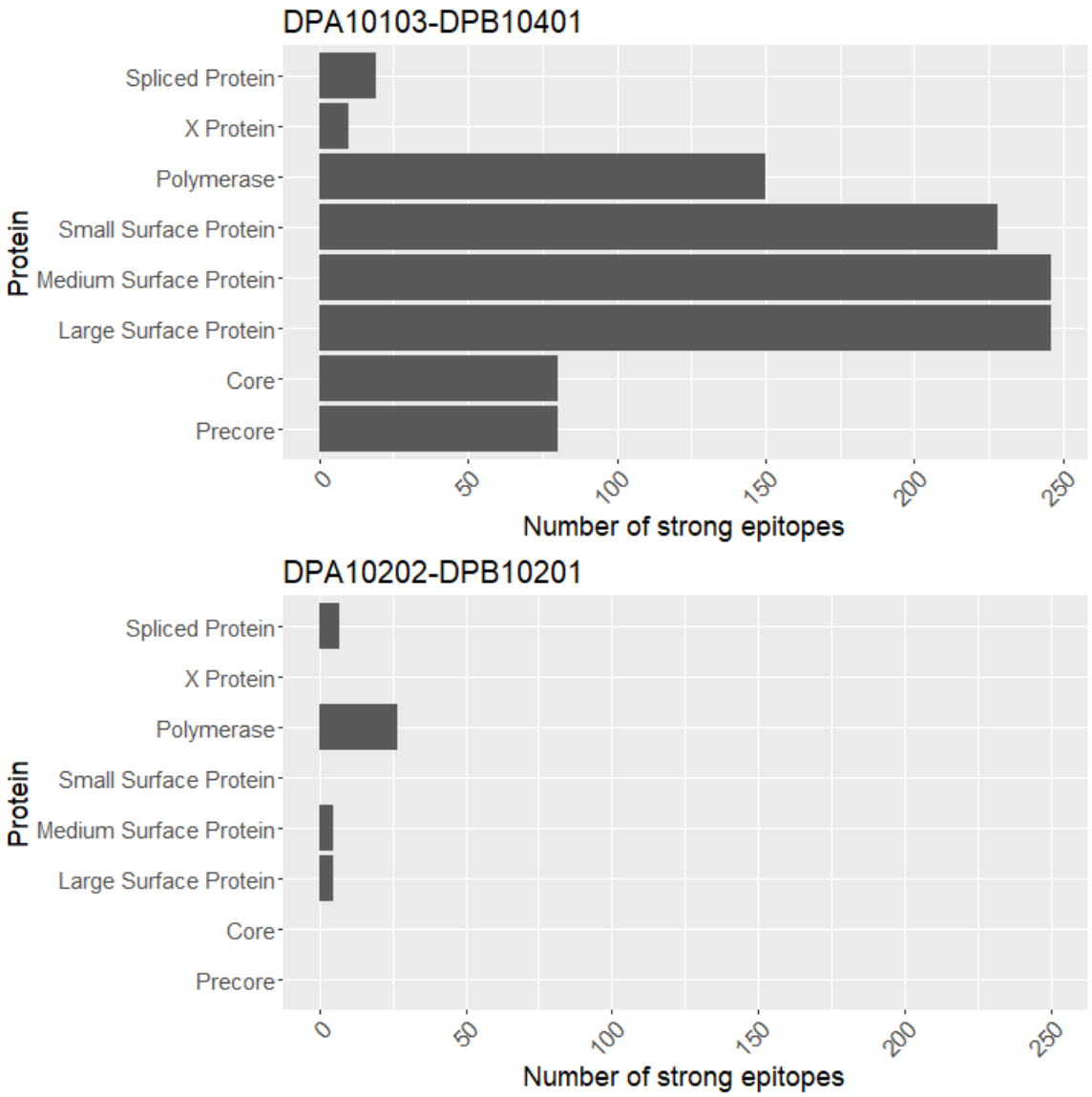


Note. DPA10103-DPB10401 and DPA10202-DPB10201 bind to some of the HBV spliced protein peptides of all genotypes strongly. Other HLA-DP haplotypes do not bind to any HBV spliced protein peptides strongly.

Since DPA10103-DPB10401 and DPA10202-DPB10201 are the only two HLA-DP haplotypes that bind to some HBV peptides strongly based on the binding affinity distribution plots, we summarized the number of HBV peptides they bind strongly in Figure 24.

Figure 24

Number of strong peptides DPA10103-DPB10401 and DPA10202-DPB10201 bind



Note. DPA10103-DPB10401 binds to peptides from all proteins strongly whereas DPA10202-DPB10201 only binds to peptides from spliced protein, polymerase, and medium and large surface proteins.

CHAPTER 4

DISCUSSION

There are three main comparison groups in this study: 1) HBV-mediated vs nonHBV-mediated liver cancer in TCGA, 2) HBV-mediated liver cancer in TCGA vs general population and 3) nonHBV-mediated-liver cancer in TCGA vs general population.

The first analysis was performed to identify if there is a difference in allele frequency between HBV-mediated and nonHBV-mediated liver cancer. We found that only HLA-DP haplotypes are significantly different before and after stratifying by population ancestry. Specifically, most of the alleles are significantly different after stratification only in the European population. DPA10103-DPB10401 is the only haplotype that is significantly different after stratification in the East Asian population. This supports previous GWAS study that the haplotype showed protective effects for HBV infections in the Asian population (Kamatani et al., 2009). Interestingly, we also observed the same effect in the European population in this study, which has not been demonstrated in the past to the best of our knowledge. We also found that DPA10202-DPB10501 is associated positively with HBV-mediated liver cancer in the European population in this study. According to Kamatani et al, DPA10202-DPB10501 is a risk haplotype for persistent HBV infection in the Asian population but this haplotype is not significantly different in the East Asian population in our study albeit in the same direction (Kamatani et al., 2009). Similarly, we found three other DP haplotypes that are significantly different in the European population but not in the East Asian population in our study. Additionally, we observed that HLA-A, -B, -C, -DQ, and -DR that are significantly different before stratifying by population ancestry are no longer significantly

different after stratifying by ancestry. However, these alleles remain associated with HBV-mediated liver cancer in the same direction as before stratifying, similar to what we observed in the HLA-DP haplotypes. This observation may result from the limitation of the sample size and the uneven representation of HBV-mediated patients in each population ancestry. Future studies would benefit from analyzing a more balanced dataset.

In the second analysis, we extended the comparison to the general population to see if the significance of the results still applies when we have healthy controls instead of liver cancer patients. We observed that most of the alleles previously significant in the combined HBV-mediated vs nonHBV-mediated analyses associate with HBV-mediated liver cancer in the same direction except A*0301-EUR, C*0302-EAS, and C*0701-EAS. In these three alleles, we also noted that the allele frequency between nonHBV-mediated liver cancer and the general population are significantly different, suggesting that general population is not a direct equivalent of the nonHBV-mediated liver cancer population and the differences between these two populations may have contributed to the change in association direction in the comparison between HBV-mediated liver cancer and the general population.

In this study, we utilized the Allele Frequency Net Database to gather the allele frequency from different populations as our healthy controls. We limited our search to apparent healthy controls including blood donors, bone marrow registry, and controls for disease study and we were not able to obtain the haplotype frequency of HLA-DP and HLA-DQ from the Allele Frequency Net Database. Therefore, we were not able to compare the allele frequency of HBV-mediated and nonHBV-mediated liver cancer

patients to the general population. For future studies, we plan to expand our data sources to include “healthy control” dataset such as the 1000 Genomes dataset and perform HLA typing to obtain the HLA-DP and HLA-DQ haplotype frequency.

To test how binding affinity is related to HBV-mediated liver cancer, we plotted the binding affinity distribution of alleles that remain significant after stratification to HBV coding sequences. In particular, we are interested in the binding affinity of HLA-DP to HBV core proteins as it has been shown previously that CD4+ response due to HBV core protein is associated with the clearance of HBV infection (Jung et al., 1995). DPA10103-DPB10401 binds strongly to HBV core peptides and is associated negatively with HBV-mediated liver cancer. In contrast, we observed that in other HLA-DP haplotypes that bind HBV core peptides either moderately or weakly, the haplotypes are associated with HBV-mediated liver cancer positively. Since Class II HLA typically presents peptide antigens to CD4+ T cells, this observation suggests that the binding affinity of HLA-DP to HBV core peptides is important in the activation of CD4+ response.

We further demonstrated in Figure 24 that DPA10103-DPB10401 binds strongly to not only HBV core peptides, but also all other HBV coding sequences. This observation can be partly explained by the structure of HBV genome, where 50% of its genome consists of overlapping reading frames (Miller et al., 1989). For example, the core protein gene has two in-frame start codons that give rise to core protein and precore protein (Miller et al., 1989). Similarly, the large surface protein overlaps the medium and small surface proteins completely and the medium surface protein sequence encompasses the small surface protein sequence (Miller et al., 1989). Therefore, when generating all

possible peptides through netMHCpan and netMHCpanII, we obtained the same peptides for these proteins. In contrast, the large surface protein is completely overlapped by polymerase but the peptides generated are not identical. This is because there is a frame shift between the coding regions for polymerase and the three surface proteins (Miller et al., 1989). Nevertheless, we observed that the peptides generated from polymerase can still be recognized by HLA-DP even though its coding sequence is constrained by another set of proteins.

Overall, our observation suggests that the binding affinity of HLA to HBV peptides may explain the association of HLA with HBV-mediated liver cancer. We speculate that if an HLA binds strongly to HBV peptides, then an individual with the HLA will be able to clear acute viral infection; if an HLA cannot bind strongly to HBV peptides, then an individual will be less likely to clear viral infection, thus developing infection. Our observations suggest that HLA-DP haplotypes are strongly implicated in HBV-mediated liver cancer etiology.

REFERENCES

Anscombe, F. J. (1956). On Estimating Binomial Response Relations. In *Biometrika* (Vol. 43, Issue 3/4, p. 461). https://doi.org/10.2307/2332926

Beck, S., & Trowsdale, J. (2000). The Human Major Histocompatibility Complex Lessons from the DNA Sequence. In *Annual Review of Genomics and Human Genetics* (Vol. 1, Issue 1, pp. 117–137). https://doi.org/10.1146/annurev.genom.1.1.117

Cochran, W. G. (1952). The chi^2 Test of Goodness of Fit. In *The Annals of Mathematical Statistics* (Vol. 23, Issue 3, pp. 315–345). https://doi.org/10.1214/aoms/1177729380

Cochran, W. G. (1954). Some Methods for Strengthening the Common χ 2 Tests. In *Biometrics* (Vol. 10, Issue 4, p. 417). https://doi.org/10.2307/3001616

Gao, W., & Hu, J. (2007). Formation of Hepatitis B Virus Covalently Closed Circular DNA: Removal of Genome-Linked Protein. In *Journal of Virology* (Vol. 81, Issue 12, pp. 6164–6174). https://doi.org/10.1128/jvi.02721-06

Gonzalez-Galarza, F. F., McCabe, A., Santos, E. J. M. D., Jones, J., Takeshita, L., Ortega-Rivera, N. D., Cid-Pavon, G. M. D., Ramsbottom, K., Ghattaoraya, G., Alfirevic, A., Middleton, D., & Jones, A. R. (2020). Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Research*, *48*(D1), D783–D788.

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a Shared Vision for Cancer Genomic Data. In *New England Journal of Medicine* (Vol. 375, Issue 12, pp. 1109–1112). https://doi.org/10.1056/nejmp1607591

Haldane, J. B. S. (1940). The Mean and Variance of | chi 2 , When Used as a Test of Homogeneity, When Expectations are Small. In *Biometrika* (Vol. 31, Issue 3/4, p. 346). https://doi.org/10.2307/2332614

Hayer, J., Jadeau, F., Deléage, G., Kay, A., Zoulim, F., & Combet, C. (2013). HBVdb: a knowledge database for Hepatitis B Virus. *Nucleic Acids Research*, *41*(Database issue), D566–D570.

Hickey, M. J., Valenzuela, N. M., & Reed, E. F. (2016). Alloantibody Generation and Effector Function Following Sensitization to Human Leukocyte Antigen. *Frontiers in Immunology*, *7*, 30.

Huang, Y.-H., Liao, S.-F., Khor, S.-S., Lin, Y.-J., Chen, H.-Y., Chang, Y.-H., Huang, Y.-H., Lu, S.-N., Lee, H.-W., Ko, W.-Y., Huang, C., Liu, P.-C., Chen, Y.-J., Wu, P.-F., Chu, H.-W., Wu, P.-E., Tokunaga, K., Shen, C.-Y., & Lee, M.-H. (2020). Large-scale genome-wide association study identifies HLA class II variants associated with chronic HBV infection: a study from Taiwan Biobank. *Alimentary Pharmacology & Therapeutics*, *52*(4), 682–691.

Hu, J., Protzer, U., & Siddiqui, A. (2019). Revisiting Hepatitis B Virus: Challenges of Curative Therapies. *Journal of Virology*, *93*(20). https://doi.org/10.1128/JVI.01032-19

Jung, M. C., Diepolder, H. M., Spengler, U., Wierenga, E. A., Zachoval, R., Hoffmann, R. M., Eichenlaub, D., Frosner, G., Will, H., & Pape, G. R. (1995). Activation of a heterogeneous hepatitis B (HB) core and e antigen-specific CD4+ T-cell population during seroconversion to anti-HBe and anti-HBs in hepatitis B virus infection. *Journal of Virology, 69*(6), 3358-3368

Kamatani, Y., Wattanapokayakit, S., Ochi, H., Kawaguchi, T., Takahashi, A., Hosono, N., Kubo, M., Tsunoda, T., Kamatani, N., Kumada, H., Puseenam, A., Sura, T., Daigo, Y., Chayama, K., Chantratita, W., Nakamura, Y., & Matsuda, K. (2009). A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians. *Nature Genetics*, *41*(5), 591–595.

Ka, S., Lee, S., Hong, J., Cho, Y., Sung, J., Kim, H.-N., Kim, H.-L., & Jung, J. (2017). HLAscan: genotyping of the HLA region using next-generation sequencing data. *BMC Bioinformatics*, *18*(1), 258.

Kaufman, J. (2018). Unfinished Business: Evolution of the MHC and the Adaptive Immune System of Jawed Vertebrates. *Annual Review of Immunology*, *36*, 383–409.

Miller, R., Kaneko, S., Chung, C. T., Girones, R., & Purcell, R. (1989). Compact Organization of the Hepatitis B Virus Genome. *Hepatology*, *9*(2), 322-327. https://doi.org/10.1002/hep.1840090226

Mohd-Ismail, N. K., Lim, Z., Gunaratne, J., & Tan, Y.-J. (2019). Mapping the Interactions of HBV cccDNA with Host Factors. *International Journal of Molecular Sciences*, *20*(17). https://doi.org/10.3390/ijms20174276

Minato Nakazawa (2019). fmsb: Functions for Medical Statistics Book with some Demographic Data. R package version 0.7.0. https://CRAN.R-project.org/package=fmsb

Nielsen, M., Lundegaard, C., Worning, P., Lauemøller, S. L., Lamberth, K., Buus, S., Brunak, S., & Lund, O. (2003). Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Science: A Publication of the Protein Society*, *12*(5), 1007–1017.

Parkin, D. M. (2006). The global health burden of infection-associated cancers in the year 2002. *International Journal of Cancer. Journal International Du Cancer*, *118*(12), 3030–3044.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Reynisson, B., Alvarez, B., Paul, S., Peters, B., & Nielsen, M. (2020). NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Research*, *48*(W1), W449–W454.

Sawai, H., Nishida, N., Khor, S.-S., Honda, M., Sugiyama, M., Baba, N., Yamada, K., Sawada, N., Tsugane, S., Koike, K., Kondo, Y., Yatsuhashi, H., Nagaoka, S., Taketomi, A., Fukai, M., Kurosaki, M., Izumi, N., Kang, J.-H., Murata, K., … Tokunaga, K. (2018). Genome-wide association study identified new susceptible genetic variants in HLA class I region for hepatitis B virus-related hepatocellular carcinoma. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-26217-7

Taravella Oill, A. M., Deshpande, A. J., Natri, H. M., & Wilson, M. A. (2020). PopInf: An Approach for Reproducibly Visualizing and Assigning Population Affiliation in Genomic Samples of Uncertain Origin. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*. https://doi.org/10.1089/cmb.2019.0434

Velkov, S., Ott, J., Protzer, U., & Michler, T. (2018). The Global Hepatitis B Virus Genotype Distribution Approximated from Available Genotyping Data. In *Genes* (Vol. 9, Issue 10, p. 495). https://doi.org/10.3390/genes9100495

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. In *Journal of Statistical Software*, *36*(3) 10.18637/jss.v036.i03

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

World Health Organization. (2017). *Global Hepatitis Report.* https://www.who.int/hepatitis/publications/global-hepatitis-report2017/en/