Ancestral origins are associated with SARS-CoV-2 susceptibility and protection in a Florida patient population

Yiran Shen[1], Bhuwan Khatri[2], Santosh Rananaware[3], Danmeng Li[4], David A. Ostrov[4], Piyush K Jain[3], Christopher J. Lessard[2], Cuong Q. Nguyen[1,5,6]

[1]Department of Infectious Diseases and Immunology, College of Veterinary Medicine, University of Florida, [2]Genes and Human Disease Research Program, Oklahoma Medical Research Foundation, [3]Department of Chemical Engineering, University of Florida, [4]Department of Pathology, Immunology & Laboratory Medicine, University of Florida, [5]Department of Oral Biology, College of Dentistry, University of Florida, [6]Center of Orphaned Autoimmune Diseases, University of Florida, Gainesville, Florida, 32611-0880 USA.

Address correspondence:

Cuong Q. Nguyen, PhD

Department of Infectious Diseases and Immunology

PO Box 110880, College of Veterinary Medicine

University of Florida, Gainesville, Florida 32611-0880 USA

Telephone: 352-294-4180, Fax: 352-392-9704

Email: nguyenc@ufl.edu

Running title: The HLA association with COVID-19

Abstract

COVID-19 is caused by severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2). The severity of COVID-19 is highly variable and related to known (e.g., age, obesity, immune deficiency) and unknown risk factors. The widespread clinical symptoms encompass a large group of asymptomatic COVID-19 patients, raising a crucial question regarding genetic susceptibility, e.g., whether individual differences in immunity play a role in patient symptomatology and how much human leukocyte antigen (HLA) contributes to this. To reveal genetic determinants of susceptibility to COVID-19 severity in the population and further explore potential immune-related factors, we performed a genome-wide association study on 284 confirmed COVID-19 patients (cases) and 95 healthy individuals (controls). We compared cases and controls of European (EUR) ancestry and African American (AFR) ancestry separately. We identified two loci on chromosomes 5q32 and 11p12, which reach the significance threshold of suggestive association ($p<1x10^{-5}$ threshold adjusted for multiple trait testing) and are associated with the COVID-19 susceptibility in the European ancestry (index rs17448496: odds ratio [OR] = 0.173; 95% confidence interval [CI], 0.08–0.36 for G allele; $p=5.15\times 10^{-5}$ and index rs768632395: OR = 0.166; 95% CI, 0.07–0.35 for A allele; $p= 4.25\times10^{-6}$, respectively), which were associated with two genes, PPP2R2B at 5q32, and LRRC4C at 11p12, respectively. To explore the linkage between HLA and COVID-19 severity, we applied fine-mapping analysis to dissect the HLA association with mild and severe cases. Using *In-silico* binding predictions to map the binding of risk/protective HLA to the viral structural proteins, we found the differential presentation of viral peptides in both ancestries. Lastly, extrapolation of the identified HLA from the cohort to the

worldwide population revealed notable correlations. The study uncovers possible differences in susceptibility to COVID-19 in different ancestral origins in the genetic background, which may provide new insights into the pathogenesis and clinical treatment of the disease.

Keywords

COVID-19, SARS-CoV-2, Human leukocyte antigens, Genome-wide association study

Introduction

In December 2019, a novel coronavirus, severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2), emerged in Wuhan, Hubei Province, China, initiating a breakout of atypical acute respiratory disease, termed coronavirus disease 2019 (COVID-19). SARS-CoV-2 is a *betacoronavirus* in the family of *Coronaviridae*; the virus contains four structural proteins: S (spike), E (envelope), M (membrane), and N (nucleocapsid), sixteen non-structural proteins (nsp1−16) and eleven accessory proteins, which support viral essential physiological function and evasion from the host immune system[1]. The complex structure of the virus provides multiple possible targets for antiviral prevention and treatment; however, the lack of comprehensive knowledge of viral infection and host immune response has hampered efforts to predict the disease course and identify effective therapeutic candidates. One of the most striking features of SARS-CoV-2 is consequence variability, ranging from asymptomatic to symptomatic viral pneumonia and finally to life-threatening acute respiratory distress syndrome[2]. A majority of patients recovered during early infection, but a smaller percentage of patients were more likely to progress and eventually die from severe systemic inflammatory response syndrome. Several factors are associated with disease severity, e.g., age, gender, pre-existing conditions[3,4], and race [5–7].

To reveal the underlying pathogenesis of SARS-CoV-2 susceptibility and disease progression, genome-wide association studies (GWAS) provide additional clues regarding the pathogenesis of complex diseases by identifying potential susceptible allelic variants. Several loci on the different chromosomes have been previously reported to be associated with COVID-19 severity[8–10]. Some studies have focused on the genetic linkage between HLA alleles and SARS-CoV-2

infection. The class I (HLA-A, -B, and -C) and class II HLA (HLA-DR, -DQ, and -DP) exhibit a high degree of polymorphism, and $CD4^+$ T cells and $CD8^+$ T cells respond to pathogens by recognizing different classes of HLA molecules (I or II, respectively) on the cell surface. Specific HLA genotypes have been associated with T-cell mediated immunity and viral clearance. Several class I and II alleles have been identified to be related to SARS-CoV-2 infection, protection, and severity through *in silico* prediction[11–14], patient genotyping[15–18], and whole-genome sequencing[19]. These studies suggested that genetic variants, especially HLA alleles, were associated with disease morbidity, mortality, and prognosis.

To study the COVID-19 consequences variability, we applied GWAS on 284 SARS-CoV-2 positive samples and 89 negative samples composed of different ancestry origins to determine if a specific genetic factor was associated with susceptibility to SARS-CoV-2 infection and severity of COVID-19 on different genetic background. We also applied fine-mapping to reveal potential disease-associated HLA alleles in European and African ancestral populations. In addition, we applied *in silico* prediction and structural modeling to identify and map the structural epitopes presented by the associated protective and risk HLAs. Lastly, we extrapolated the finding to the worldwide population and the result showed significant correlations with other countries.

Materials and Methods

Study population

284 confirmed COVID-19 samples were obtained from Boca Biolistics (Pompano Beach, FL) and CTSI Biorepository at the University of Florida (Gainesville, FL). The median age of the patients was 44.7 years (range: 3-94 years). Patients had positive test results for SARS-CoV-2 by RT-PCR from nasopharyngeal swabs or tracheal aspirates. 89 healthy individuals with negative PCR tests for SARS-CoV-2 viral infection were included as controls. The median age of the control group was 60 years (range: 0-101 years). After the exclusion of samples during quality control, the final case-control data sets comprised 254 patients and 80 control participants. The study was approved by the Institutional Review Board of the University of Florida.

Sample extraction and genotyping

Clinical specimens of nasopharyngeal swabs were collected in a viral transport medium. DNA was extracted from viral transport medium or directly from tracheal aspirates by Maxwell® RSC Blood DNA Kit per manufacturer's instructions (Promega Corporation). RNase A was added to samples to remove potential viral RNA. Isolated genomic DNA was quantified by NanoDrop™ One/OneC Microvolume UV-Vis Spectrophotometer (Thermo Scientific). Genotyping was done using Axiom™ Human Genotyping SARS-CoV-2 Research Array as instructed by the manufacturer (Thermo Scientific).

Quality Control

PLINK (v1.9) [20] was used for quality control and logistic analysis of the data. The single nucleotide polymorphism (SNPs) and subjects passing the following quality control criteria [21]

were used in the downstream analysis: SNPs having Major Allele Frequency >1%, SNPs and sample each with call rate >95%, controls with Hardy-Weinberg proportion test with p>0.001 and cases and controls with differential missingness P>0.001, subjects with heterozygosity (<5 S.D. from the Mean), and one individual from the pair was removed if identity-by-descent (IBD) was >0.4.

Assessment of population stratification

Principal components between cases and controls and population substructures within the dataset were determined using EIGENSTRAT [22] and independent genotyped SNPs with r2<0.2 between variants, for this 1000 Genome reference population was used. Principle component analysis (PCA) was used to remove outliers defined by standard deviations greater than 6 (s.d. >6) from the Mean [21]. Case and control samples were plotted by PC1 and PC2.

Imputation

Whole-genome imputation was performed using TOPMed Reference Panel (TOPMed r2) in the TOPMed Imputation server [23]. The data were phased using Eagle version 2.4 and imputed using Minimac4. In addition, the HLA (chr6) region was imputed in Michigan Imputation Server [23]. The data were phased using Eagle V2.4 and imputed using Minimac4 and Four-digit Multi-ethnic HLA reference panel.

Logistic Analysis

Post imputation, a quality control measure as explained before, was used in the imputed data. Logistic regression analysis was carried out using PLINK to test for single marker SNP-COVID-19 association post imputation, adjusting for the first two principal components.

In-silico binding predictions

To predict the SARS-CoV-2 peptides in which selective HLA alleles will bind, HLA peptide-binding prediction algorithms netMHCpan (v4.1) and netMHCIIpan (v4.0) were utilized for HLA class I and class II alleles, respectively[24]. Full-length amino acid sequences of structural proteins from SARS-CoV-2 whole-genome proteome (SnapGene, EPI_ISL_7196120_B.1.1.529, EPI_ISL_7196121_B.1.1.529) were used to infer all possible potentially relevant peptides (9mers for class I and 15mers for class II). Default rank thresholds (%Rank values) were used to define strong (0.5% for netMHCpan and 2% for netMHCIIpan) and weak (2% for netMHCpan and 10% for netMHCIIpan) binders. The prediction binding pattern was compared among different alleles to predict the immunogenic portion for further application. In selecting the alleles of interest, %Rank was used as a reference value to compare the ability of the same peptide to be presented by risk and protective alleles among predicted alleles that can be presented by at least one of the candidate alleles (Class I MHC: %Rank < 2%, Class II MHC: %Rank < 10), the top three alleles have the greatest difference in presentation ability (%Rank) from the other allele are selected.

Structural modeling of SARS-CoV-2 peptide HLA molecules interactions

Models for all the HLA molecules were made using SWISS-MODEL. 7RTD was used as templates for HLA-B*27:05 and HLA-C*13:02, and 3PDO for HLA-DRB1*13:02. 6PX6 and 5KSA were used as templates for HLA-DQA1*05:01 and HLA-DQB1*03:01, respectively. Then, they were superposed into 5KSU and merged into one final model for the HLA-DQ structure in COOT. The peptide in the template structure was mutated in COOT, too. The geometry of the result complexes

was regularized in PHENIX.

Worldwide geographical comparison

The data of allele frequencies among countries/regions and ethnicities were obtained from the Allele Frequency Net Database (http://allelefrequencies.net), global susceptibility among countries/regions (cases per one million population) were obtained from Worldometer (https://www.worldometers.info), and global death rate (case-mortality) were obtained from John Hopkins Coronavirus resource center (https://coronavirus.jhu.edu/data/mortality). Databases were searched on March 24th, 2022. When multiple data points of an allele frequency were available for a country, the weighted Mean was calculated according to the sample size.

Statistical analysis

The association between the allele frequency of each HLA gene and cases and mortality were assessed by linear regression in GraphPad Prism (v9.3.1). An adjusted *p*-value of <0.05 was considered statistically significant.

Results

GWAS analysis of COVID-19 patients with European and African ancestries

The state of Florida in the United States (US) is one of the most ethnically diverse states. 53% of Floridians are White (Non-Hispanic), with 21.6% being White (Hispanic), Black or African American (Non-Hispanic) (15.2%), Asian (Non-Hispanic) (2.73%), and Other (Hispanic) (2.97%). 20.1% of Floridians were born outside the US, which is higher than the national average of 13.7% in 2019 (https://datausa.io/profile/geo/florida). Due to the ethnically diverse nature of Florida residents, we performed QC of the dataset as described previously. Following QC, we stratified the population based on their genetic ancestry (COVID-19 cases: 56% European (EUR) ancestry and 37% African (AFR) ancestry. Non-COVID-19 controls: 77% EUR ancestry and 15% AFR ancestry (**Supplementary Figures 1 and 2**). Since EUR and AFR ancestries are distinct populations, we performed GWAS analysis on the two populations separately. As presented in **Figure 1**, we found top variants rs17448496 at locus 5q32 (odds ratio [OR] = 0.173; 95% confidence interval [CI], 0.08–0.36 for G allele; $p<5.15x10^{-6}$) and rs768632395 at locus 11p12 (OR = 0.166; 95% CI, 0.07–0.35 for A allele; $p<4.25 \times 10^{-6}$) that were suggestive GWAS with COVID-19 susceptibility in the EUR ancestral patients. There were no SNP association signals in the AFR ancestral patients that met the significance threshold of suggestive association **(Figure 2)**. In summary, the limited sample size was sufficient to distinguish two different genetic ancestries. In addition, we identified two interesting SNPs localized in the EUR ancestry. These two SNPs showed OR values of less than 1, suggesting that they are possibly involved to some extent in protection against SARS-CoV-2 infection (i.e., to a greater extent, would be present in

unconfirmed healthy individuals) in contrast to the AFR ancestral patients where the absence of prominent SNPs, which can be interpreted as a protective mechanism with no apparent contribution to SARS-CoV-2 infection.

Chromosome 5q32 and 11p12

As presented in **Figure 1**, the GWAS showed a suggestive association with serine/threonine-protein phosphatase 2A (PP2A) regulatory subunit B (PPP2R2B) (rs17448496, OR = 0.173; 95% CI = 0.08–0.36 for G allele; $p<5.15x10^{-6}$)) on chromosome 5q32, and Leucine-Rich Repeat Containing 4C (LRRC4C) (rs768632395, OR = 0.166; 95% CI = 0.07–0.35 for A allele; $p<4.25 \times 10^{-6}$) on chromosome 11p12. PPP2R2B belongs to the phosphatase family, and they are involved in several biological processes. Previous reports have shown that PP2A can activate T-cell responses by inhibiting cytotoxic T-lymphocyte-associated (CTLA)-4 function or impairing programmed death-ligand (PD-L)-1 expression[25]. The PPP2R2B gene can encode the regulatory subunit B55β, forming the PP2A-B55β complex by binding to the scaffolding and catalytic subunits. PPP2R2B plays an important regulatory role in the immune system, preventing organ damage by activated T cells in chronic inflammation caused by systemic autoimmune diseases, hypermethylation of PPP2R2B can induce defective acquired apoptosis[26], and dysregulation of PPP2R2B may contribute to the development and progression of breast cancer[27]. Importantly, PPP2R2B interacted with PPP1R15A in the ERK signaling pathway, which is on chr9 and was associated with SARS-CoV-2 infection in the C5 phenotype from HGI (COVID-19 Host Genetics Initiative group)[28]. LRRC4, also known as netrin-G ligand-2 (NGL-2),

belongs to the superfamily of LRR proteins and is a receptor for netrin-G2[29], regulates excitatory synapse formation and promotes axonal differentiation. LRRC4 can also act as a tumor suppressor gene to significantly inhibit glioblastoma cell proliferation by interacting with extracellular and intracellular signaling pathways [30]. LRRC4 binds to phosphotyrosine-dependent protein kinase 1 (PDPK1), promotes NF-$_K$B activation in glioblastoma cells and secretion of Interleukin 6 (IL-6), C-C Motif Chemokine Ligand 2 (CCL2) and Interferon-gamma (IFN-γ), thereby inhibiting the expansion of tumor-infiltrating regulatory T cells and the growth of glioblastoma cells[31]. Although not as significant as the findings of the previous meta-analysis, the loci we observed in the EUR ancestral patients were associated with host-adaptive, especially T-cell-related immunity, and the lack of such a significance in AFR ancestral patients may explain the susceptibility of the population.

HLA association in COVID-19 patients with European and African ancestries

Specific HLA alleles have played significant roles in many bacterial and viral infections. We applied the fine-mapping to the extended HLA region (chromosome 6, 25 to 34 Mb) to determine their association with the EUR and AFR ancestral patients. We determined no significant SNP association signal on the HLA complex that achieved the threshold of significance for suggestive association **(Supplementary Figure 2)**. Due to the substantial overlap in bound peptides among HLA alleles and their co-dominant expression, additive GWAS association tests may not capture the full functional role of HLA in COVID-19 risk. Therefore, we further analyzed the association of COVID-19 with HLA alleles through HLA imputation. The results showed multiple alleles present

for HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, -DRB1 loci through imputation-based methods using the SNPs data from GWAS. Indeed, it is well established that minority groups of different races and ethnic groups in the US are disproportionately affected by COVID-19. Minorities endured a higher risk for infection, hospitalization, and death [6,32]. The allele distributions were compared between COVID-19 patients and control individuals in EUR and AFR ancestries, respectively. As presented in **Table 1**, significant associations ($p \leq 0.05$) between HLA-B*27:05 alleles and SARS-CoV-2 positivity were identified in the EUR ancestry, which was associated with a decreased risk of SARS-CoV-2 positivity (OR=0.17; 95% CI, 0.03–0.83; p=0.029). In the AFR ancestry, HLA-A*02:01 (OR=0.20; 95% CI, 0.05–0.80; p=0.023), -A*33:01 (OR=0.05; 95% CI, 0.005–0.71; p=0.026), -DRB1*13:02 (OR=0.19; 95% CI, 0.05–0.72; p=0.014) and -DPB1*11:01 (OR=0.16; 95% CI, 0.03–0.72; p=0.023) were associated with a decreased risk of SARS-CoV-2 positivity. Overall, individuals who carry the above class I and class II HLA alleles were less likely to be infected by SARS-CoV-2.

HLA association between severe and mild COVID-19 patients

It is well documented that COVID-19 patients exhibited a whole range of symptoms and severity. To further define the HLA association with the severity of COVID-19, we subdivided the patient cases into mild and severe according to the source of sample acquisition (saliva: mild; tracheal aspirates: hospitalized/severe) and performed the same HLA imputation. Stratified by disease severity showed significant alleles changed in both EUR and AFR ancestral patient populations. In the EUR ancestral patients, HLA-C*12:03, -B*35, -B*38:01 were associated with

an increased risk of SARS-CoV-2 severity (OR >1) (**Table 2**). Interestingly, HLA-B*38 showed a significant odds ratio (OR=1639). Whereas, as presented in **Table 3**, in the AFR ancestral patients, HLA-DQB1*03 was associated with an increased risk of SARS-CoV-2 severity. In contrast, HLA-B*58, -DRB1*13:02, and -DQB1*06 were associated with decreased risk of SARS-CoV-2 severity. The results suggest that there is a higher frequency of risk and protective alleles in mild and severe cases of different ancestral patient populations, and this frequency could explain the demographic differences in the affected population.

***In silico* mapping of peptide epitopes derived from SARS-CoV-2 structural proteins presented by protective and risk alleles in EUR ancestry**

As presented in **Table 1 and Table 2**, the protective and risk alleles of the EUR ancestry encode class I MHC molecules. To map which viral epitopes are presented by the protective and risk HLAs, we applied the prediction tool NetMHCpan - 4.1 provided by DTU Health. As a model, we chose HLA-B*27:05 (protective) and HLA-C*12:03 (risk) based on the significance and mapped the SARS-CoV-2 structural proteins (S, M, N, and E). Both stronger (%Rank less than 0.5%) and weaker (%Rank less than 2%) binders of HLA-B*27:05 (46 peptides in total) and HLA-C*12:03 (46 peptides in total) were selected and mapped across the structural proteins **(Figure 3)**. Based on the presented differences in protective and risk alleles, the top three alleles that have the most significant difference in presentation ability (%Rank) from the other allele are selected and highlighted in **Figure 3** and the sequences are listed in **Table 4**. Molecular docking combined with *in silico* mapping showed for peptides derived from S protein, protective allele

HLA-B*27:05 presents ARDLICAQK and RRARSVASQ, corresponding to amino acid (aa) positions 846-854 and 682-690 in the spike protein S2 subunit, and KHTPINLVR (aa 206-214) from N-terminal domain (NTD)**(Figure 4A-C)**. Risk allele HLA-C*12:03 presents LAATKMSEC (aa 1024-1032) and IAPGQTGKI (aa 410-418) at S2 region, and FASTEKSNI (aa 92-100) from RBD **(Figure 4D-F).** Notably, the protective allele failed to present antigen from E protein, and the risk allele shows peptides such as LTALRLCAY (aa 32-42), LAFVVFLLV (aa 21-29) and LAILTALRL (aa 31-39) across the E protein **(Figure 4G-I)**. Similar analyses were performed to determine whether HLA-B*27:05 and HLA-C*12:03 can also present the same set of structural proteins for other variants. The results indicated that similar sequences of peptide antigens of delta and omicron variants were predicted to be presented by these alleles **(Supplementary Table 1, 2)**. In summary, the results demonstrated that in the EUR ancestry group, HLA-B*27:05 (protective) and HLA-C*12:03 (risk) present multiple structural proteins of SASR-CoV-2 and the protective alleles lack presentation of the E protein.

In silico mapping of antigen derived from SARS-CoV-2 structural protein presented by protective and risk alleles in AFR ancestry

The most protective and risk alleles were class II HLA in AFR ancestry, unlike the EUR ancestry **(Table 3)**. To determine which viral epitopes are presented by the protective and risk HLAs, we again applied the prediction tool NetMHCIIpan - 4.0. As a model, we chose HLA-DRB1*13:02 (protective) and HLA-DQA1*05:01-DQB1*03:01 (risk) and mapped them for SARS-CoV-2 structural proteins. Due to only the DQB1*03 being detected in the AFR group, the

haplotype with the highest frequency was selected for DQA1 prediction. Both stronger (%Rank less than 2%) and weaker (%Rank less than 10%) binders of HLA-DRB1*13:02 (248 peptides in total) and HLA-DQB1*03:01 (206) were selected and mapped across the structural proteins **(Figure 5)**. Again, the top three alleles that have the most significant difference in presentation ability (%Rank) from the other allele were selected and highlighted in **Figure 5**, and the sequences were listed in **Table 5**. The molecular docking combined with *in silico* mapping showed for peptides derived from S protein, protective allele HLA-DRB1*13:02 tends to bind PRTFLLKYNENGTIT (aa 272-286) at NTD, KKSTNLVKNKCVNFN (aa 528-542) and KSTNLVKNKCVNFNF(aa 529-543) at RBD **(Figure 6A-C)**. While risk allele HLA-DQB1*03:01 prefers to bind ITPCSFGGVSVITPG (aa 587-601), ECDIPIGAGICASYQ (aa 661-675) and TPCSFGGVSVITPGT (aa 588-602) at beta-strand region on S2 **(Figure 6D-E)**, which is less immunogenetic, thus differences in the disease course may be the result of a combination of multiple alleles. Interestingly, different from the EUR group, only the protective allele presents the peptide such as LVKPSFYVYSRVKNL (aa 1-65), VYSRVKNLNSSRVPD (aa 58-72), and SFYVYSRVKNLNSSR (aa 55-69) from E protein **(Figure 6F-H)**. In contrast, the risk allele failed, leading to incomplete viral clearance and the subsequent induction of new mutations. We analyzed the antigen presentation of the same set of alleles for the delta and omicron variants (**Supplementary Table 3, 4**) structural proteins. We found that this property was still retained. In summary, the result demonstrated that in the AFR ancestry group, the antigens presented by protective alleles (HLA-DRB1*13:02) are more diverse, with different presentation sites against the S, M, and N proteins. In contrast, the antigens presented by risk alleles (HLA-DQB1*03:01)

are more at a single site and lack presentation of the E protein.

Association of the identified HLA alleles with worldwide COVID-19 cases and mortalities.

To further evaluate whether the identified risk and protective HLA in the studied cohort in the state of Florida can be extrapolated to determine the association with cases and mortalities worldwide, we employed linear regression to determine the association of each allele frequency, number of COVID-19 cases per one million population, and case-mortality due to COVID-19. In alleles predicted by disease severity (**Table 2 and Table 3**), We selected two protective alleles (A*11:01 and DPB1*11:01) and two risk alleles (B*38:01 and DQB1*03) based on their ORs. As presented in **Figure 7**, the protective allele A*11:01 showed a significant association. As the allele frequencies increase, the lesser case frequency per one million population (p=0.0267); however, there was no significant correlation with case mortality. There was no statistical significance with the protective DPB1*11:01 allele for case mortality and case/1M pop (**Figure 7A**). Interestingly, when we evaluated the risk B*38:01 allele, which has the largest odds ratio (OR=1639.00), it showed a positive or upward trend between the increase in the allele frequencies and rise in both case/1M population (p=0.0101) and mortality (p=0.6092) (**Figure 7B**). The risk allele DQB1*03 showed a significant association with case-mortality (p = 0.0099) but no significant correlation with case/1M population (p=0.0782). The results suggest that extrapolation of specific protective and risk HLA alleles identified in the studied AFR and EUR cohort can be applicable to determine the association with worldwide COVID-19 cases and mortalities.

Discussion

In this study, the GWAS study showed two plausible genome-wide significant associations on chromosomes 5q32 and 11p12 in EUR ancestral group in our overall susceptibility model by ancestral stratification. Using HLA imputation, the results suggest that several class I MHC alleles in EUR ancestral group and a mixed class I and II MHC alleles in AFR ancestral group were likely to be associated with disease susceptibility or severity. For example, in EUR ancestral group, HLA-B*27:05 were associated with an overall decreased risk of infection, HLA-A*11:01 were associated with less severity, while HLA-C*12:03, -B*35, -B*38:01 were associated with an increased risk of SARS-CoV-2 severity. In AFR ancestral group, HLA-A*02:01, -DRB1*13:02, and -DPB1*11:01 were associated with an overall decreased risk of SARS-CoV-2 positivity. Regarding disease severity, HLA-DRB1*13:02 and -DPB1*11:01, together with HLA-B*58, -DQB1*06, -DQA1*02:01, -DRB1*07:01 exhibited an association with less severity, whereas HLA-DQB1*03 was associated with an increased risk of SARS-CoV-2 severity. Using *in silico* prediction and modeling of SARS-CoV-2 structural proteins to identify the epitopes and binding strength with risk and protective HLA alleles showed that a different presentation pattern may activate the immune response to varying degrees, leading to changes in the course of the disease. Overall, the study sheds important insight into COVID-19 by stratifying ancestral origin, leading to better disease understanding and prevention strategies.

The first reported genome-wide association signals were at loci 3p21.31 and 9q34.2, which revealed the association of protein-coding genes that regulate viral attachment and host immune response (LC6A20, LZTFL1, CCR9, FYCO1, CXCR6, and XCR1) and ABO blood group to severe

COVID-19 disease, respectively[33]. These associations were replicated in subsequent studies[10,34,35]. Results of a genetic study of 2244 critically ill patients with COVID-19 in intensive care units across the UK identified associations on chromosomes 12q24.13, 19p13.2, 19p13.3, and 21q22.1, revealed the genes with innate immunity (IFNAR2 and OAS) and host-driven lung inflammatory injury (DPP9, TYK2, and CCR2)[8]. A large GWAS study from HGI, including three genome-wide association meta-analyses comprising 49,562 COVID-19 patients from 46 studies in 19 countries, identified 13 human genomic loci associated with infection or severe COVID-19, including TYK2, DPP9, and FOXP4 which corresponded to previously documented associations with lung or autoimmune diseases and inflammatory disorders[36]. Another GWAS study in Thai suggested a protective effect of IL17B on 5q32 against disease[37], and a recent release from IHG showed that MUC5B and ELF5 on chr11 might be associated with immune system regulation in the lungs in the development of COVID-19. Our study identified two loci on chromosomes 5q,32, and 11p12 in European ancestry who reached the significance threshold of suggestive association with SARS-CoV-2 infection. We could not define a similar association in the AFR group, and this variation in gene-level may be responsible for the difference. Our findings, by sampling in an ethnically diverse region, as the first study to evaluate different populations in one GWAS study, confirm preliminary results on the genetic determinants of COVID-19 in a diverse population and further reflect the complexity of genetic factors involved in SARS-CoV-2 infection.

HLA molecules present antigens by binding to endogenous antigenic peptides (class I) or exogenous antigenic peptides (class II) and express them as peptide-MHC complexes on the

surface of antigen-presenting cells. Previous studies from different countries have identified multiple COVID-19 susceptibility-related alleles; for example, Wang et al. identified HLA-B*15:27 alleles from a Chinese population[38], Yung et al. identified serotype B22 (HLA-B*54:01, B*56:01 and B*56:04 alleles) from Hongkong Chinese population[39]. Our study identified that HLA-B*27:05 in EUR ancestry and HLA-A*02:01, -A*33:01, -DPB1*11:01 in AFR ancestry were associated with a decreased risk of SARS-CoV-2 positivity (OR <1), which provides additional clinical revealed alleles for a diverse population. During COVID-19, HLA appears to prevent or cause further disease progression through unknown mechanisms. One piece of evidence is that the low affinity of viral peptides to bind HLA can lead to severe disease and high-affinity binding, providing better protectivity. *In silico* prediction from Nguyen et al. identified HLA-A*02:02, HLA-B*15:03, and HLA-C*12:03 have the strongest binding affinity of conserved peptides[11]. Amoroso et al. have shown that HLA-DRB1*08, which was predicted to bind SARS-CoV-2 peptides with low affinity, was correlated to mortality. To confirm this, a study with the Sardinian population found HLA-DRB1*08:01 allele only existed in hospitalized patients[16,17]. A group from South Asia found HLA-B*35 was more among the mildly infected group than the fatal group and owned a high peptide loading capacity compared to other HLA-B proteins[14,18]. Another evidence may come from cytokine storms, where an unbalanced immune response leads to higher morbidity, while a well-regulated immune response allows patients to recover more quickly. Although not studied in detail in COVID-19, it was shown in a previous study that class II MHC secretes different types of cytokines when binding to different peptides, thus triggering T cell differentiation in divergent ways (TH1/TH17 by DRB1*0401, risk; TH1/TH2 by DRB *0402, protective) in

rheumatoid arthritis [40]. In our studies, we identified in EUR group HLA-C*12:03, -B*35, -B*38:01 were associated with an increased risk of SARS-CoV-2 severity (OR >1). And in the AFR group, HLA-DQB1*03 was associated with an increased risk of SARS-CoV-2 severity, while HLA-B*58, -DRB1*13:02, and -DQB1*06 were associated with a decreased risk of SARS-CoV-2 severity. HLA-C*12:03 and -B*35 identified in the severed EUR group were previously shown to be associated with less severity, and several class II MHC molecules were associated with increased or decreased severity in the AFR group, altogether suggesting a complicated disease progression mechanism that may involve in multiple alleles as well as host immune-regulated factors.

T cells have an important role in the outcome and maintenance of SARS-CoV-2 immunity normally during viral infection by recognizing viral antigens in short peptides present by HLA. A previous study identified epitope megapools (MPs) containing SARS-CoV-2 T cell epitopes derived from various viral proteins, which successfully induced viral-specific T cells responses in patients [41] and CoVac-1 vaccinated individuals[42]. CoVac-1 is a peptide-based vaccine candidate composed of predicted MPs (S, N, N, E, and ORF8); this prediction is based on the most prominent class I HLA-A and -B and class II HLA-DR alleles to protect a broad population. The success of phase 1 clinical trial of this vaccine demonstrates the potential and importance of epitope screening, as it can stimulate long-lasting T-cell immune responses in the fight against COVID-19. We performed *in silico* epitope mapping of the identified HLAs in different ancestry groups hoping that further experiments will lead to an affected-population-based understanding of immune defense mechanisms against SARS-CoV-2 and aid in the development of vaccines and immunotherapies. To understand the disease outcome in different ancestral origins, we

applied a similar strategy; instead of utilizing common alleles, we used alleles associated with increased or decreased COVID-19 severity in both EUR and AFR groups (identified protective and risk alleles). We found the overall number of peptides that protective and risk alleles present differed, 46 vs. 151 in the EUR ancestry group and 248 vs. 206 in the AFR ancestry group. Due to the presentation difference in class I and class II MHC, the number could not simply explain the various disease outcome of different ancestry origins. In addition, the percentage of predicted stronger binders of protective and risk alleles is approximately 41% vs. 30% in the EUR ancestry group and 25% vs. 21% in the AFR ancestry group. The lesser portions to be presented in protective alleles in the AFR ancestry group might become an issue of ineffective immunity activation. Meanwhile, we observed some differences in the regions to be presented of protective and risk alleles, and we also observed that the E protein is only presented by class I MHC risk alleles in the EUR group, whereas it is presented by class II MHC protective alleles in the AFR group. This suggests that besides well-known RBD, certain regions of SARS-CoV-2 are also immunogenic, and peptides with insufficient immunogenicity should be considered for exclusion when developing vaccines or drugs for specific affected groups.

In conclusion, this study provides the first insight of group analysis regarding the effects of SARS-CoV-2 infection in different ancestry origins from the same region, including the susceptibility and disease severity. Due to the limited sample size, further validation is needed by *in vitro* experiment. This study demonstrated that the contributions of genetic factors and comorbidities are helpful in identifying potential severe COVID-19 cases.

Conflict of Interests

The authors declare no competing financial interests.

Author Contributions

YS isolated genomic DNAs and conducted the imputation. PJ and SR obtained patient samples. BK and CJL performed the GWAS analyses. YS, DO, and CN conceptualized the study, performed data analysis, and prepared the manuscript. All authors have read and approved the final manuscript.

Acknowledgments

CQN is supported financially in part by PHS grants DE028544 and DE028544-02S1 from the National Institute of Dental and Craniofacial Research. We thanked Ms. Maria Cecilia Lopez from the UF Genetics Institute for performing the GWAS assays. We appreciated Dr. Patrick Concannon at the UF Genetics Institute for the insightful discussion.

Data availability

The data supporting this study's findings are presented in the manuscript and available in Supplementary Materials. The data can also be available from the corresponding author upon request

References

1. Wang M-Y, Zhao R, Gao L-J, Gao X-F, Wang D-P, Cao J-M. SARS-CoV-2: Structure, Biology, and Structure-Based Therapeutics Development. Front Cell Infect Microbiol. 2020;10: 587269. doi:10.3389/fcimb.2020.587269
2. Siddiqi HK, Mehra MR. COVID-19 illness in native and immunosuppressed states: A clinical-therapeutic staging proposal. J Heart Lung Transplant. 2020;39: 405–407. doi:10.1016/j.healun.2020.03.012
3. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. JAMA. 2020;323: 1239–1242. doi:10.1001/jama.2020.2648
4. Ueyama H, Kuno T, Takagi H, Krishnamoorthy P, Vengrenyuk Y, Sharma SK, et al. Gender Difference Is Associated With Severity of Coronavirus Disease 2019 Infection: An Insight From a Meta-Analysis. Crit Care Explor. 2020;2: e0148. doi:10.1097/CCE.0000000000000148
5. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature. 2020;584: 430–436. doi:10.1038/s41586-020-2521-4
6. Gold JAW, Wong KK, Szablewski CM, Patel PR, Rossow J, da Silva J, et al. Characteristics and Clinical Outcomes of Adult Patients Hospitalized with COVID-19 - Georgia, March 2020. MMWR Morb Mortal Wkly Rep. 2020;69: 545–550. doi:10.15585/mmwr.mm6918e1
7. Garg S, Kim L, Whitaker M, O'Halloran A, Cummings C, Holstein R, et al. Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019. MMWR Morb Mortal Wkly Rep. 2020;69: 458–464. doi:10.15585/mmwr.mm6915e3
8. Pairo-Castineira E, Clohisey S, Klaric L, Bretherick AD, Rawlik K, Pasko D, et al. Genetic mechanisms of critical illness in COVID-19. Nature. 2021;591: 92–98. doi:10.1038/s41586-020-03065-y
9. Roberts GHL, Park DS, Coignet MV, McCurdy SR, Knight SC, Partha R, et al. AncestryDNA COVID-19 Host Genetic Study Identifies Three Novel Loci. medRxiv. 2020; doi:10.1101/2020.10.06.20205864
10. Shelton JF, Shastri AJ, Ye C, Weldon CH, Filshtein-Sonmez T, Coker D, et al. Trans-ancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity. Nat Genet. 2021;53: 801–808. doi:10.1038/s41588-021-00854-7
11. Nguyen A, David JK, Maden SK, Wood MA, Weeder BR, Nellore A, et al. Human leukocyte antigen susceptibility map for severe acute respiratory syndrome coronavirus 2. J Virol. 2020;94. doi:10.1128/JVI.00510-20
12. Romero-López JP, Carnalla-Cortés M, Pacheco-Olvera DL, Ocampo-Godínez JM, Oliva-Ramírez J, Moreno-Manjón J, et al. A bioinformatic prediction of antigen presentation from SARS-CoV-2 spike protein revealed a theoretical correlation of HLA-DRB1*01 with COVID-

19 fatality in Mexican population: An ecological approach. J Med Virol. 2021;93: 2029–2038. doi:10.1002/jmv.26561

13. Barquera R, Collen E, Di D, Buhler S, Teixeira J. … of 438 HLA proteins to complete proteomes of seven pandemic viruses and distributions of strongest and weakest HLA peptide binders in populations worldwide. Hla. 2020;
14. Huang S, Tan M. HLA class I genotypes customize vaccination strategies in immune simulation to combat COVID-19. BioRxiv. 2020; doi:10.1101/2020.11.18.388983
15. Sakuraba A, Haider H, Sato T. Population Difference in Allele Frequency of HLA-C*05 and Its Correlation with COVID-19 Mortality. Viruses. 2020;12. doi:10.3390/v12111333
16. Amoroso A, Magistroni P, Vespasiano F, Bella A, Bellino S, Puoti F, et al. HLA and AB0 Polymorphisms May Influence SARS-CoV-2 Infection and COVID-19 Severity. Transplantation. 2021;105: 193–200. doi:10.1097/TP.0000000000003507
17. Littera R, Campagna M, Deidda S, Angioni G, Cipri S, Melis M, et al. Human Leukocyte Antigen Complex and Other Immunogenetic and Clinical Factors Influence Susceptibility or Protection to SARS-CoV-2 Infection and Severity of the Disease Course. The Sardinian Experience. Front Immunol. 2020;11: 605688. doi:10.3389/fimmu.2020.605688
18. Naemi FMA, Al-Adwani S, Al-Khatabi H, Al-Nazawi A. Association between the HLA genotype and the severity of COVID-19 infection among South Asians. J Med Virol. 2021;93: 4430–4437. doi:10.1002/jmv.27003
19. Wang F, Huang S, Gao H, Zhou Y, Lai C, Li Z, et al. Initial Whole Genome Sequencing and Analysis of the Host Genetic Contribution to COVID-19 Severity and Susceptibility. medRxiv. 2020; doi:10.1101/2020.06.09.20126607
20. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81: 559–575. doi:10.1086/519795
21. Lessard CJ, Li H, Adrianto I, Ice JA, Rasmussen A, Grundahl KM, et al. Variants at multiple loci implicated in both innate and adaptive immune responses are associated with Sjögren's syndrome. Nat Genet. 2013;45: 1284–1292. doi:10.1038/ng.2792
22. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38: 904–909. doi:10.1038/ng1847
23. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48: 1284–1287. doi:10.1038/ng.3656
24. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. 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 Res. 2020;48: W449–W454. doi:10.1093/nar/gkaa379
25. Zhao H, Li D, Zhang B, Qi Y, Diao Y, Zhen Y, et al. PP2A as the Main Node of Therapeutic Strategies and Resistance Reversal in Triple-Negative Breast Cancer. Molecules. 2017;22. doi:10.3390/molecules22122277
26. Madera-Salcedo IK, Sánchez-Hernández BE, Svyryd Y, Esquivel-Velázquez M, Rodríguez-

Rodríguez N, Trejo-Zambrano MI, et al. PPP2R2B hypermethylation causes acquired apoptosis deficiency in systemic autoimmune diseases. JCI Insight. 2019;5. doi:10.1172/jci.insight.126457

27. Li Z, Li Y, Wang X, Yang Q. PPP2R2B downregulation is associated with immune evasion and predicts poor clinical outcomes in triple-negative breast cancer. Cancer Cell Int. 2021;21: 13. doi:10.1186/s12935-020-01707-9
28. COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. Eur J Hum Genet. 2020;28: 715–718. doi:10.1038/s41431-020-0636-6
29. Woo J, Kwon S-K, Kim E. The NGL family of leucine-rich repeat-containing synaptic adhesion molecules. Mol Cell Neurosci. 2009;42: 1–10. doi:10.1016/j.mcn.2009.05.008
30. Wu M, Huang C, Gan K, Huang H, Chen Q, Ouyang J, et al. LRRC4, a putative tumor suppressor gene, requires a functional leucine-rich repeat cassette domain to inhibit proliferation of glioma cells in vitro by modulating the extracellular signal-regulated kinase/protein kinase B/nuclear factor-kappaB pathway. Mol Biol Cell. 2006;17: 3534–3542. doi:10.1091/mbc.E05-11-1082
31. Li P, Feng J, Liu Y, Liu Q, Fan L, Liu Q, et al. Novel Therapy for Glioblastoma Multiforme by Restoring LRRC4 in Tumor Cells: LRRC4 Inhibits Tumor-Infitrating Regulatory T Cells by Cytokine and Programmed Cell Death 1-Containing Exosomes. Front Immunol. 2017;8: 1748. doi:10.3389/fimmu.2017.01748
32. Romano SD, Blackstock AJ, Taylor EV, El Burai Felix S, Adjei S, Singleton C-M, et al. Trends in Racial and Ethnic Disparities in COVID-19 Hospitalizations, by Region - United States, March-December 2020. MMWR Morb Mortal Wkly Rep. 2021;70: 560–565. doi:10.15585/mmwr.mm7015e2
33. Severe Covid-19 GWAS Group, Ellinghaus D, Degenhardt F, Bujanda L, Buti M, Albillos A, et al. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. N Engl J Med. 2020;383: 1522–1534. doi:10.1056/NEJMoa2020283
34. Horowitz JE, Kosmicki JA, Damask A, Sharma D, Roberts GHL, Justice AE, et al. Genome-wide analysis in 756,646 individuals provides first genetic evidence that ACE2 expression influences COVID-19 risk and yields genetic risk scores predictive of severe disease. medRxiv. 2021; doi:10.1101/2020.12.14.20248176
35. Pereira AC, Bes TM, Velho M, Marques E, Jannes CE, Valino KR, et al. Genetic risk factors and Covid-19 severity in Brazil: results from BRACOVID Study. medRxiv. 2021; doi:10.1101/2021.10.06.21264631
36. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. Nature. 2021;600: 472–477. doi:10.1038/s41586-021-03767-x
37. Chamnanphon M, Pongpanich M, Suttichet TB, Jantarabenjakul W, Torvorapanit P, Putcharoen O, et al. Host genetic factors of COVID-19 susceptibility and disease severity in a Thai population. J Hum Genet. 2022; doi:10.1038/s10038-021-01009-6
38. Wang W, Zhang W, Zhang J, He J, Zhu F. Distribution of HLA allele frequencies in 82 Chinese individuals with coronavirus disease-2019 (COVID-19). HLA. 2020;96: 194–196.

doi:10.1111/tan.13941

39. Yung Y-L, Cheng C-K, Chan H-Y, Xia JT, Lau K-M, Wong RSM, et al. Association of HLA-B22 serotype with SARS-CoV-2 susceptibility in Hong Kong Chinese patients. HLA. 2020; doi:10.1111/tan.14135
40. Luckey D, Behrens M, Smart M, Luthra H, David CS, Taneja V. DRB1*0402 may influence arthritis by promoting naive CD4+ T-cell differentiation in to regulatory T cells. Eur J Immunol. 2014;44: 3429–3438. doi:10.1002/eji.201344424
41. Grifoni A, Weiskopf D, Ramirez SI, Mateus J, Dan JM, Moderbacher CR, et al. Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals. Cell. 2020;181: 1489–1501.e15. doi:10.1016/j.cell.2020.05.015
42. Heitmann JS, Bilich T, Tandler C, Nelde A, Maringer Y, Marconato M, et al. A COVID-19 peptide vaccine for the induction of SARS-CoV-2 T cell immunity. Nature. 2022;601: 617–622. doi:10.1038/s41586-021-04232-5

Figure legends

Figure 1. Manhattan plot showed two protectivity loci with suggestive significance in EUR ancestry Two loci were found to be associated with COVID-19 susceptibility reach the significance threshold of suggestive association ($P < 1 \times 10^{-5}$ threshold adjusted for multiple trait testing) in the EUR group: the rs17448496 at locus 5q32 showed a suggestive association with serine/threonine-protein phosphatase 2A regulatory subunit B (PPP2R2B) , and the rs768632395 at locus 11p12 showed a suggestive association with Leucine Rich Repeat Containing 4C (LRRC4C). (Red: GWAS association ($P < 5 \times 10^{-8}$); Blue: suggestive association ($P < 1 \times 10^{-5}$))

Figure 2. Manhattan plot showed non-significant associated loci in AFR ancestry There were no SNP association signals in the AFR group that met the significance threshold of suggestive association: $P<1\times10^{-5}$. (Red: GWAS association ($P < 5 \times 10^{-8}$); Blue: suggestive association ($P < 1 \times 10^{-5}$))

Figure 3. Distribution of allelic presentation of peptide across the SARS-CoV-2 structural for protective (HLA-B*27:05) and risk (HLA-C*12:03) alleles in EUR ancestry group. Dark and light bars indicating the identified stronger (< 0.5 %Rank) and weaker (< 2 %Rank) binding 9 mer peptides, respectively. With green and red indicating protective and risk alleles association, respectively. Dashed lines are the selected top three peptides in each structural protein that were presented with the greatest variation in binding affinity, marked by the final tendency of alleles

(green: protective allele; red: risk allele). The relative positions are arranged by successive structural proteins in the order of S, E, M, N and relative lengths as indicated in the bottom.

Figure 4. Model of class I HLA molecule associated with protective (HLA-B*27:05) or risk (HLA-C*12:03) incidence of SARS-CoV-2 severity. (A-C) HLA-B*27:05 binds to ARDLICAQK, RRARSVASQ and KHTPINLVR derived from the S protein with affinity estimated Kd 1534 nM, 438.61 nM, 6798.81 nM respectively. (D-F) HLA-C*12:03 binds to LAATKMSEC, IAPGQTGKI and FASTEKSNI derived from the S protein with affinity estimated Kd 959.47 nM, 806.95 nM, 126.76 nM respectively. (G-I) HLA-C*12:03 binds to LTALRLCAY, LAFVVFLLV and LAILTALRL derived from the E protein with affinity estimated Kd 176.42 nM, 190.66 nM, 112.95 nM respectively. Peptides are shown as sticks modeled on the crystal structure of class I HLA molecues, different peptide backbones are represented by different colors.

Figure 5. Distribution of allelic presentation of peptide across the SARS-CoV-2 structural for protective (HLA-DRB1*13:02) and risk (HLA-DQB1*03:01) alleles in AFR ancestry group. Dark and light bars indicating the identified stronger (< 2 %Rank) and weaker (< 10 %Rank) binding 12 mer peptides, respectively. With green and red indicating protective and risk alleles association, respectively. Dashed lines are the selected top three peptides in each structural protein that were presented with the greatest variation in binding affinity, marked by the final tendency of alleles (green: protective allele; red: risk allele). The relative positions are arranged

by successive structural proteins in the order of S, E, M, N and relative lengths as indicated in the bottom.

Figure 6. Model of class II HLA molecule associated with protective (HLA-DRB1*13:02) or risk (HLA-DQB1*03:01) incidence of SARS-CoV-2 severity. (A-B) HLA-DRB1*13:02 binds to PRTFLLKYNENGTIT (LKYNENGTI), KKSTNLVKNKCVNFN (VKNKCVNF) and KSTNLVKNKCVNFNF (VKNKCVNF) derived from the S protein with affinity estimated Kd 14.09 nM, 63.11 nM, 41.78 nM respectively. (C-D) HLA-DQA1*05:01-DQB1*03:01 binds to ITPCSFGGVSVITPG (SFGGVSVIT), ECDIPIGAGICASYQ (IGAGICASY) and TPCSFGGVSVITPGT (SFGGVSVIT) derived from the S protein with affinity estimated Kd 51.75 nM, 28.32 nM, 52.35 nM respectively. (G-I) HLA-DRB1*13:02 binds to LVKPSFYVYSRVKNL (FYVYSRVKN), VYSRVKNLNSSRVPD (VKNLNSSRV) and SFYVYSRVKNLNSSR (YVYSRVKNL) derived from the E protein with affinity estimated Kd 223.87 nM, 135.65 nM, 162.90 nM respectively. Only the 9-mer core binding pepited were showed in the figures. Peptides are shown as sticks modeled on the crystal structure of class II HLA molecues, different peptide backbones are represented by different colors.

Figure 7. Worldwide HLA allele frequency and COVID-19 case/1M population and case-mortaliy. Association of protective and risk alleles in EUR and AFR cohort with worldwide case per one million population and case-mortality. Each dot represents a country plotted by average allele frequency (x-axis) with case-mortality (red, right y-axis) and case/1M population (blue, left

y-axis). The equation and values below the image (left, case/1M population; right, case-mortality) show the quadratic equation for predicting trend of linear regression, the R-squared (coefficient of determination) and the two tailed p-value for the correlation analysis. (A) Association of proteictive alleles in EUR and AFR corhort. (B) Association of risk alleles in EUR and AFR corhort.

Tables

Table 1 Significant alleles associate with either protective (OR <1) or risk (OR>1) factor overall in EUR and AFR group

Allele	Odds Ratio	Standard Error	L95	U95	P value
		EUR			
HLA-B*27:05	0.17	0.81	0.03	0.84	0.03
		AFR			
HLA-A*02:01	0.20	0.69	0.05	0.80	0.02
HLA-A*33:01	0.05	1.28	0.005	0.71	0.03
HLA-DRB1*13:02	0.19	0.67	0.05	0.72	0.01
HLA-DPB1*11:01	0.16	0.77	0.03	0.72	0.02

Table 2 Significant alleles associate with either protective (OR <1) or risk (OR>1) factor in EUR group by case severity

Allele	Odds Ratio	Standard Error	L95	U95	P value
Severe vs Control					
HLA-C*12:03	8.07	0.94	1.28	50.85	0.03
HLA-B*35	4.21	0.71	1.04	17.04	0.04
Severe vs Mild					
HLA-B*38:01	1639.00	3.65	1.29	2076000	0.04
Mild vs Control					
HLA-A*11:01	0.24	0.71	0.06	0.95	0.04

Table 3 Significant alleles associate with either protective (OR <1) or risk (OR>1) factor in AFR group by case severity

Allele	Odds Ratio	Standard Error	L95	U95	P value
Severe vs Control					
HLA-DQB1*03	115.80	1.96	2.50	5360.00	0.02
HLA-B*58	0.07	1.10	0.01	0.61	0.02
HLA-DRB1*13:02	0.09	1.14	0.01	0.89	0.04
Severe vs Mild					
HLA-DQB1*06	0.41	0.39	0.19	0.88	0.02
Mild vs Control					
HLA-DQA1*02:01	0.13	0.89	0.02	0.76	0.02
HLA-DRB1*07:01	0.15	0.85	0.03	0.80	0.03
HLA-DPB1*11:01	0.06	1.29	0.01	0.81	0.03
HLA-A*02:01	0.23	0.74	0.05	0.97	0.05
HLA-DRB1*13:02	0.26	0.68	0.07	0.99	0.05

Table 4 In silico binding prediction for protective (HLA-B*27:05) or risk (HLA-C*12:03) alleles in EUR group present structural protein of SARS-CoV-2 original strain

Peptide	Protein	Bound Preference
ARDLICAQK	S	Protective Allele
RRARSVASQ	S	Protective Allele
KHTPINLVR	S	Protective Allele
LAATKMSEC	S	Risk Allele
IAPGQTGKI	S	Risk Allele
FASTEKSNI	S	Risk Allele
LTALRLCAY	E	Risk Allele
LAFVVFLLV	E	Risk Allele
LAILTALRL	E	Risk Allele
YRINWITGG	M	Protective Allele
KKLLEQWNL	M	Protective Allele
SRYRIGNYK	M	Protective Allele
IAIAMACLV	M	Risk Allele
AAVYRINWI	M	Risk Allele
LAAVYRINW	M	Risk Allele
RRIRGGDGK	N	Protective Allele
DRLNQLESK	N	Protective Allele
GRRGPEQTQ	N	Protective Allele
FAPSASAFF	N	Risk Allele
SAFFGMSRI	N	Risk Allele
LSPRWYFYY	N	Risk Allele

Table 5 In silico binding prediction for protective (HLA-DRB1*13:02) or risk (HLA-DQA1*05:01-DQB1*03:01) alleles in AFR group present structural protein of SARS-CoV-2 original strain

Peptide	Protein	Core bound protective	Core bound risk	Bound Preference
PRTFLLKYNENGTIT	S	LKYNENGTI	LKYNENGTI	Protective Allele
KKSTNLVKNKCVNFN	S	LVKNKCVNF	LVKNKCVNF	Protective Allele
KSTNLVKNKCVNFNF	S	LVKNKCVNF	LVKNKCVNF	Protective Allele
ITPCSFGGVSVITPG	S	FGGVSVITP	SFGGVSVIT	Risk Allele
ECDIPIGAGICASYQ	S	IGAGICASY	IGAGICASY	Risk Allele
TPCSFGGVSVITPGT	S	FGGVSVITP	SFGGVSVIT	Risk Allele
LVKPSFYVYSRVKNL	E	FYVYSRVKN	YVYSRVKNL	Protective Allele
VYSRVKNLNSSRVPD	E	VKNLNSSRV	VKNLNSSRV	Protective Allele
SFYVYSRVKNLNSSR	E	YVYSRVKNL	YVYSRVKNL	Protective Allele
QFAYANRNRFLYIIK	M	YANRNRFLY	FAYANRNRF	Protective Allele
FIASFRLFARTRSMW	M	FRLFARTRS	RLFARTRSM	Protective Allele
TNILLNVPLHGTILT	M	ILLNVPLHG	ILLNVPLHG	Protective Allele
MADSNGTITVEELKK	M	ADSNGTITV	NGTITVEEL	Risk Allele
INWITGGIAIAMACL	M	ITGGIAIAM	ITGGIAIAM	Risk Allele
NWITGGIAIAMACLV	M	IAIAMACLV	ITGGIAIAM	Risk Allele
FKDQVILLNKHIDAY	N	ILLNKHIDA	VILLNKHID	Protective Allele
QVILLNKHIDAYKTF	N	ILLNKHIDA	NKHIDAYKT	Protective Allele
NFKDQVILLNKHIDA	N	VILLNKHID	FKDQVILLN	Protective Allele
LGTGPEAGLPYGANK	N	PEAGLPYGA	PEAGLPYGA	Risk Allele
GTGPEAGLPYGANKD	N	PEAGLPYGA	PEAGLPYGA	Risk Allele
YYLGTGPEAGLPYGA	N	TGPEAGLPY	PEAGLPYGA	Risk Allele

Figure 1

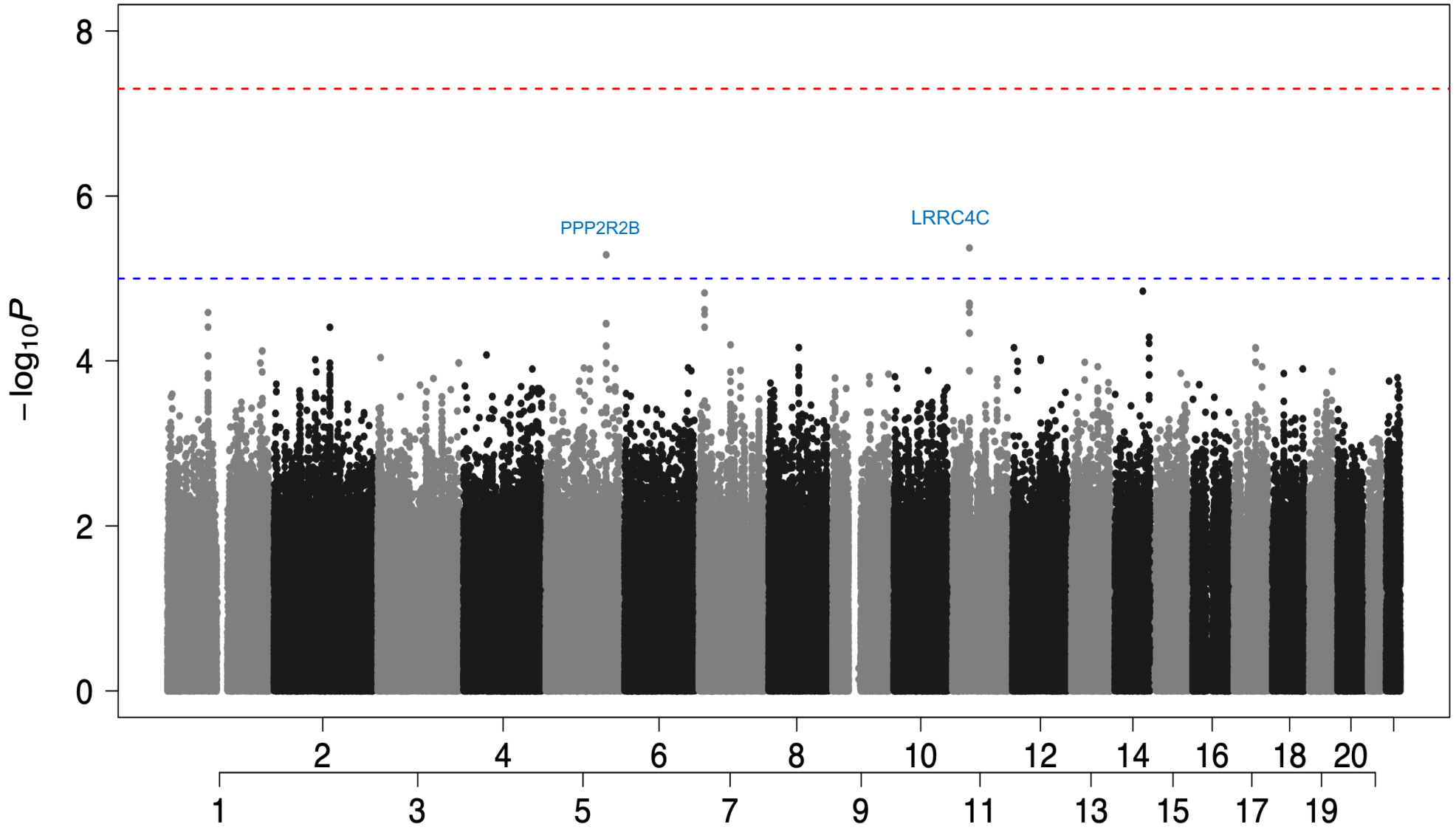

8
6
4
2
0
$-\log_{10}P$
PPP2R2B
LRRC4C
2
4
6
8
10
12
14
16
18
20
1
3
5
7
9
11
13
15
17
19

Figure 2

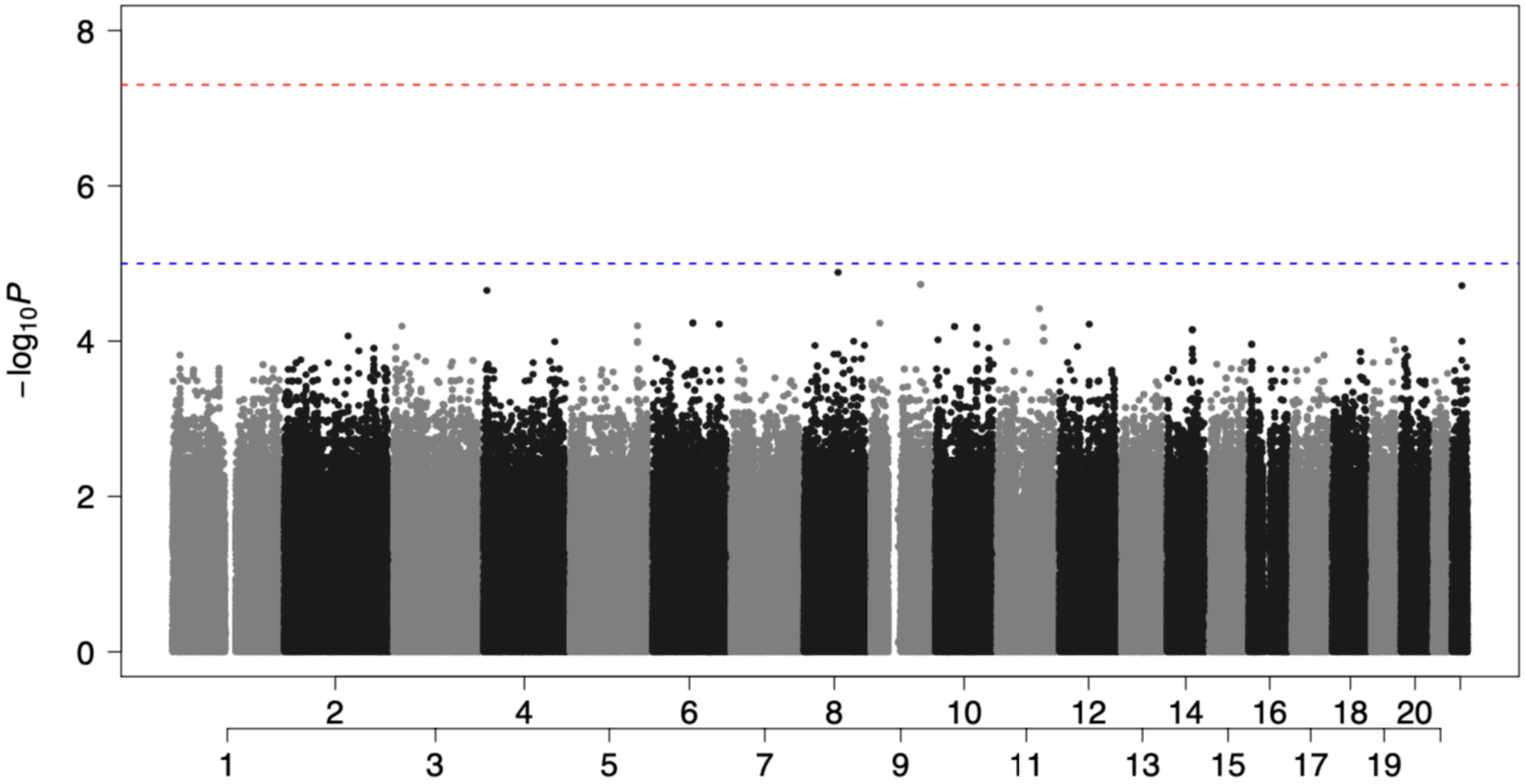


Figure 3

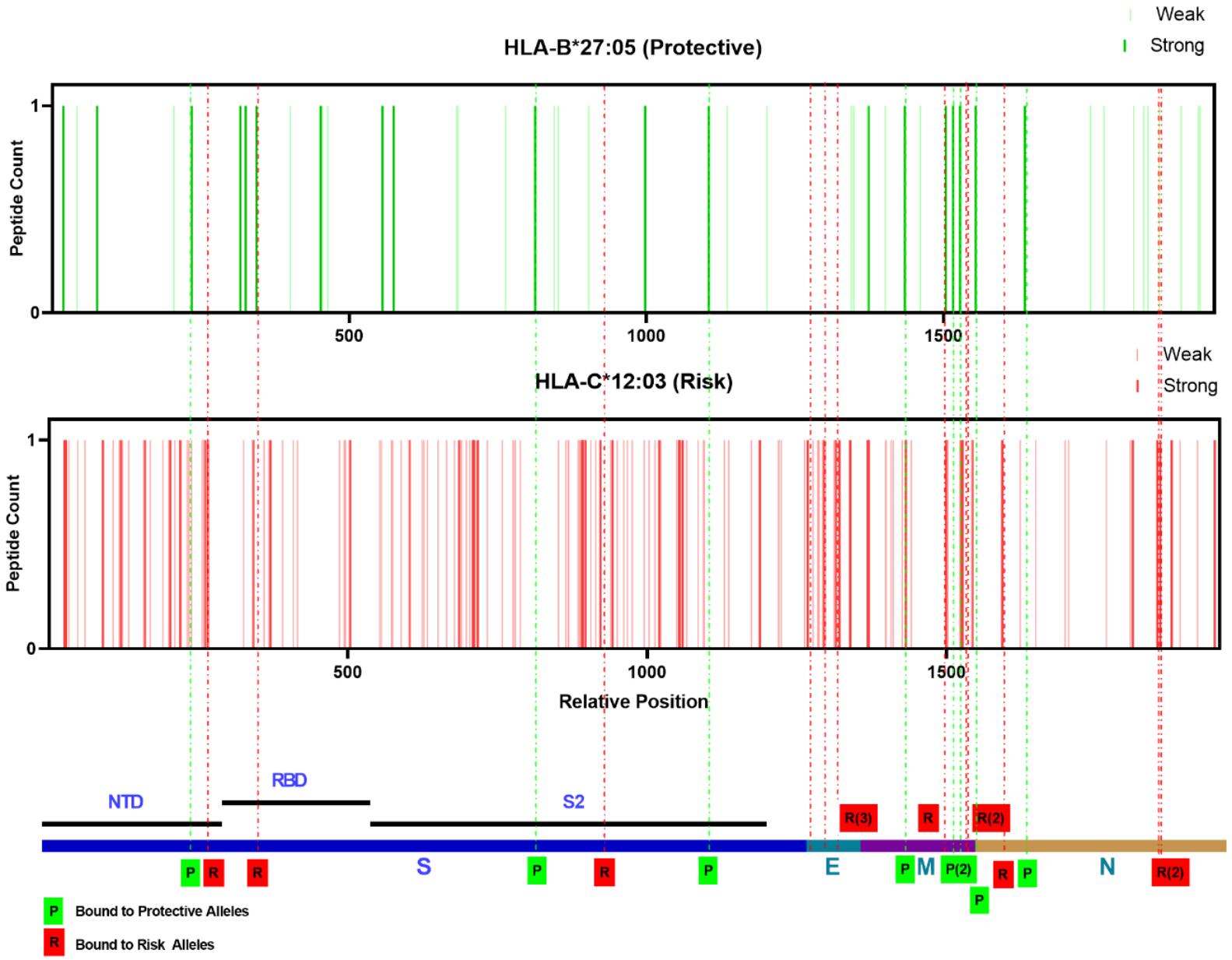

HLA-B*27:05 (Protective)
Weak
Strong
Peptide Count
1
0
500
1000
1500
HLA-C*12:03 (Risk)
Weak
Strong
Peptide Count
1
0
500
1000
1500
Relative Position
NTD
RBD
S2
S
E
M
N
R(3)
R
R(2)
P(2)
R(2)
P
R
R
P
R
P
P
P
R
P
P
Bound to Protective Alleles
R
Bound to Risk Alleles

Figure 4

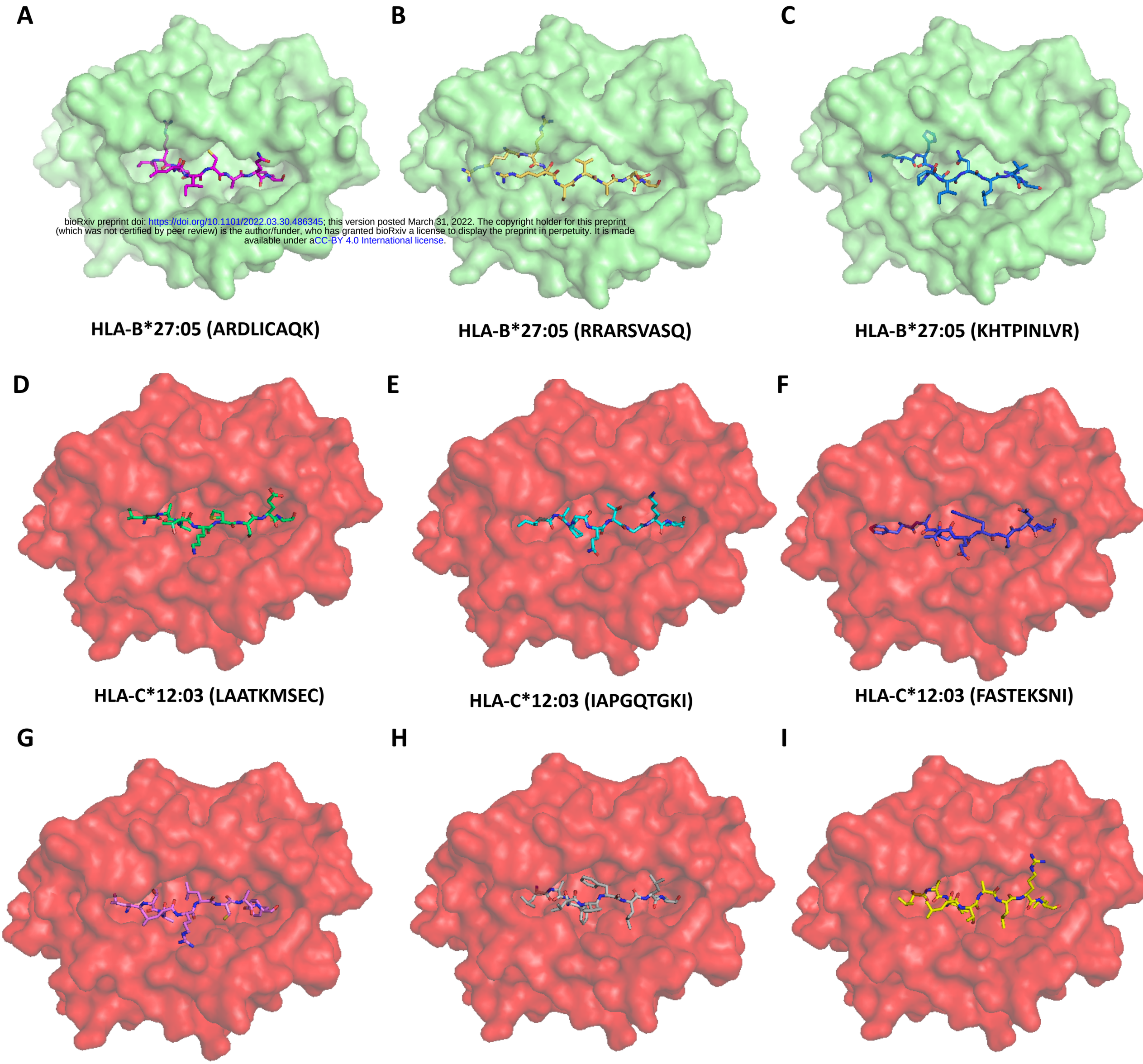


Figure 5

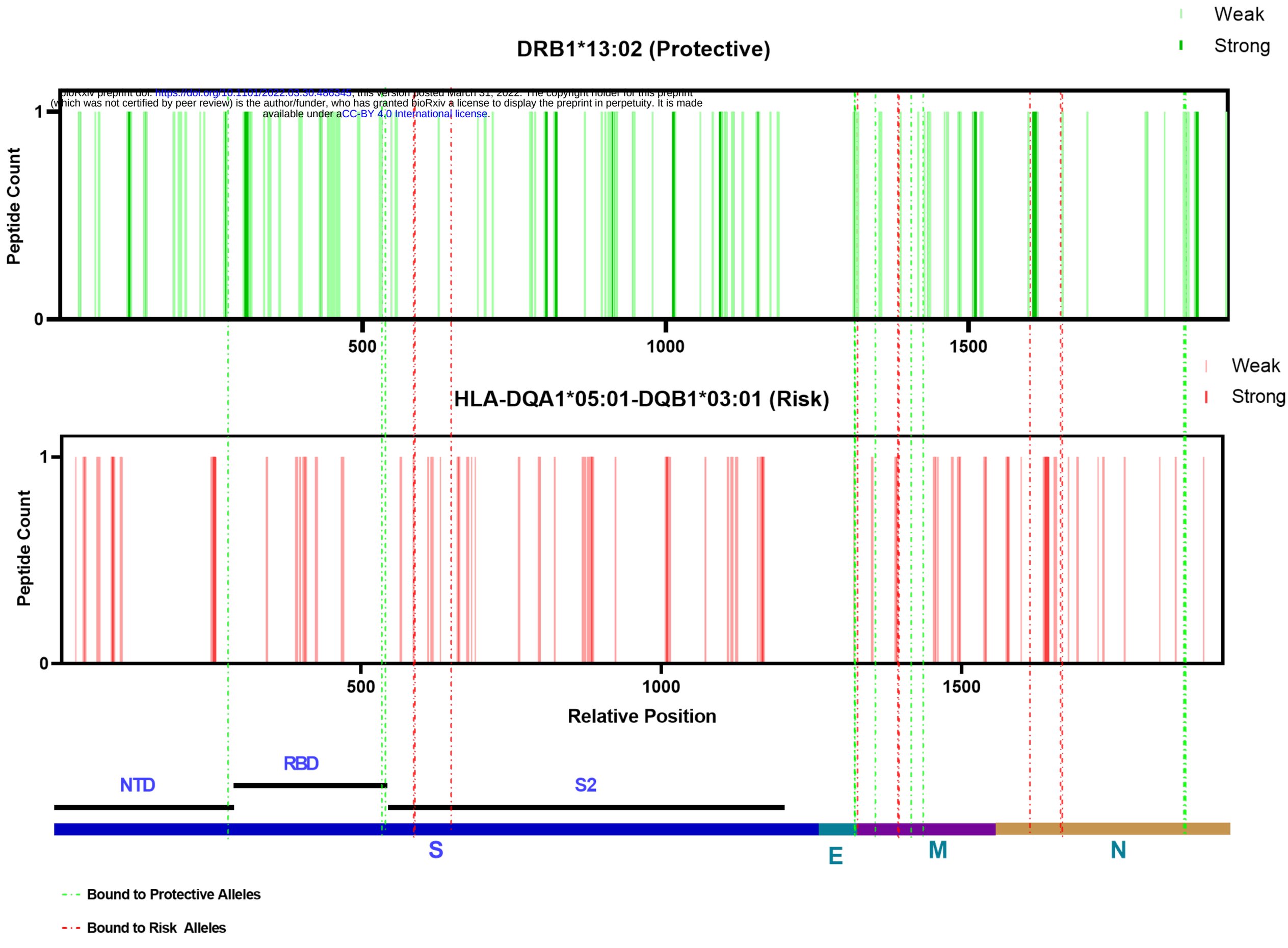
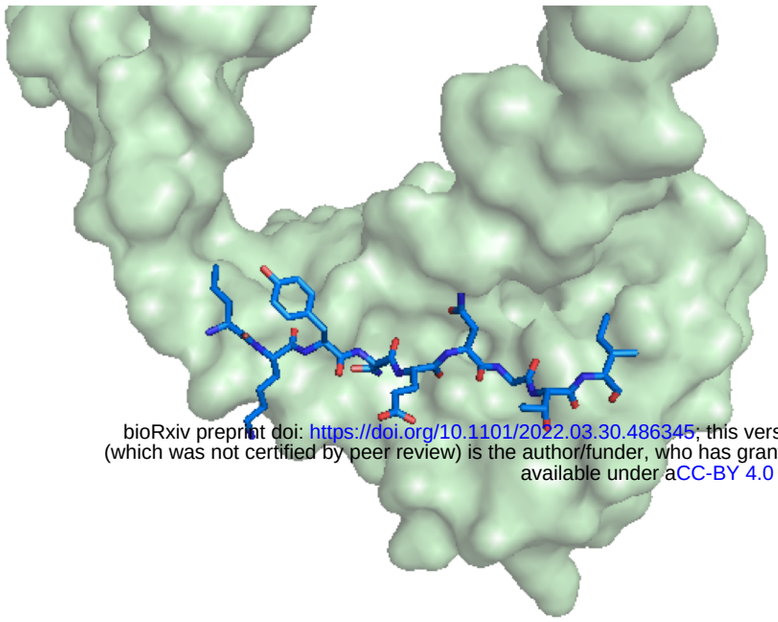


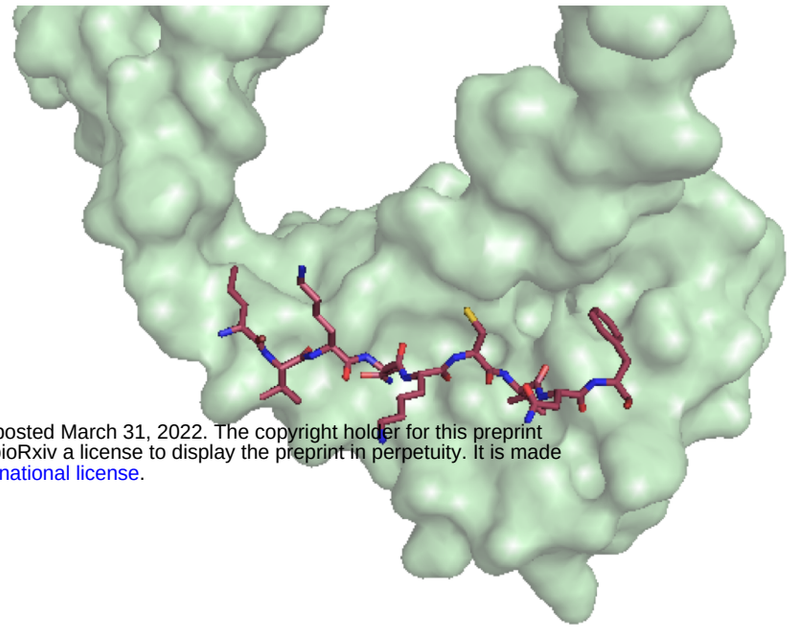
Figure 6

A



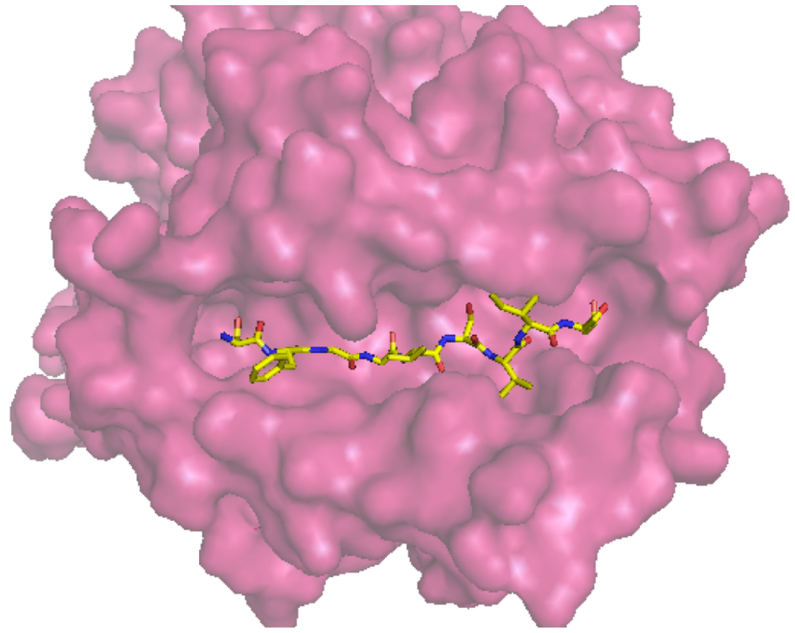
HLA-DRB1*13:02 (LKYNENGTI)

B



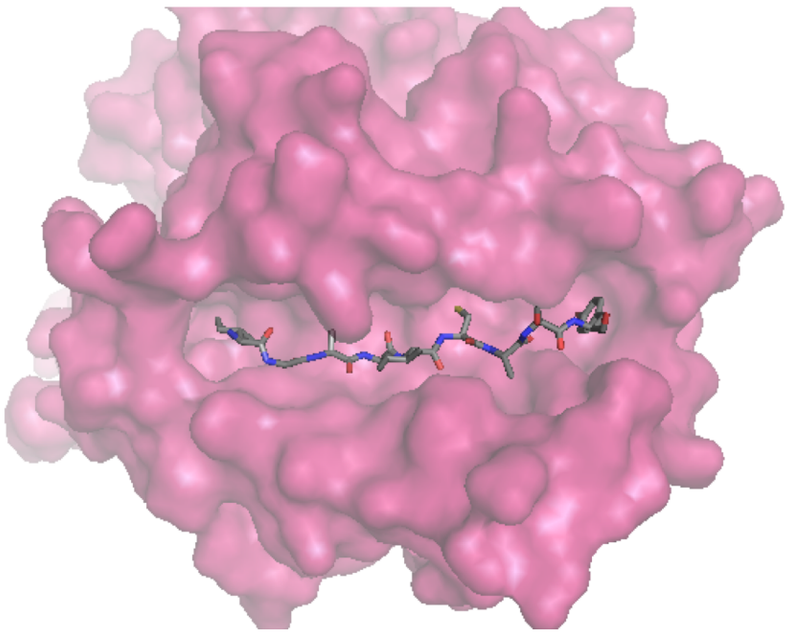
HLA-DRB1*13:02 (LVKNKCVNF)

C



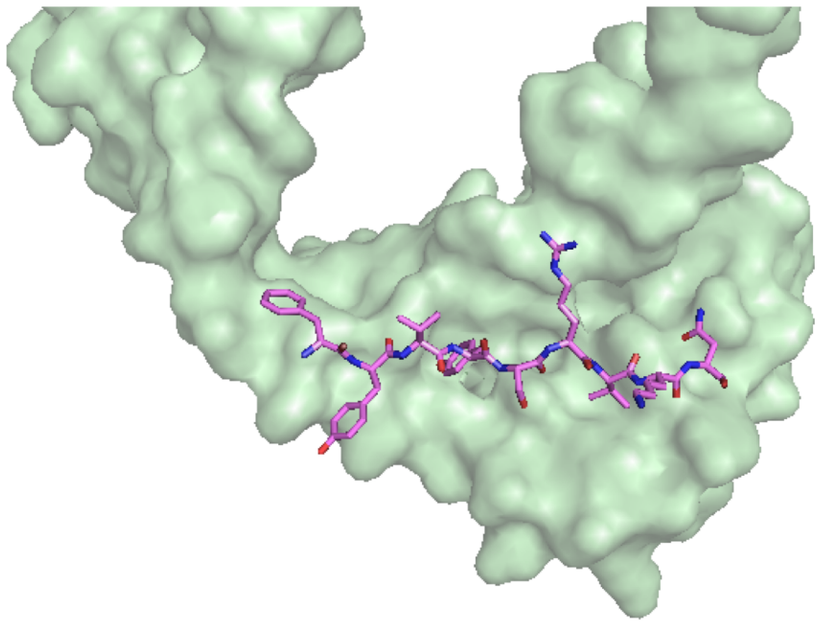
HLA-DQA1*05:01-DQB1*03:01 (SFGGVSVIT)

D



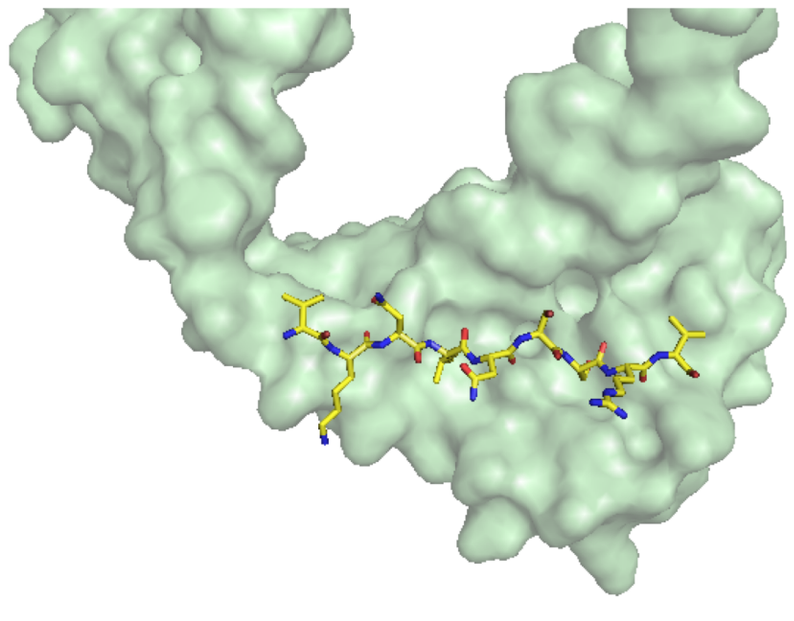
HLA-DQA1*05:01-DQB1*03:01 (IGAGICASY)

E



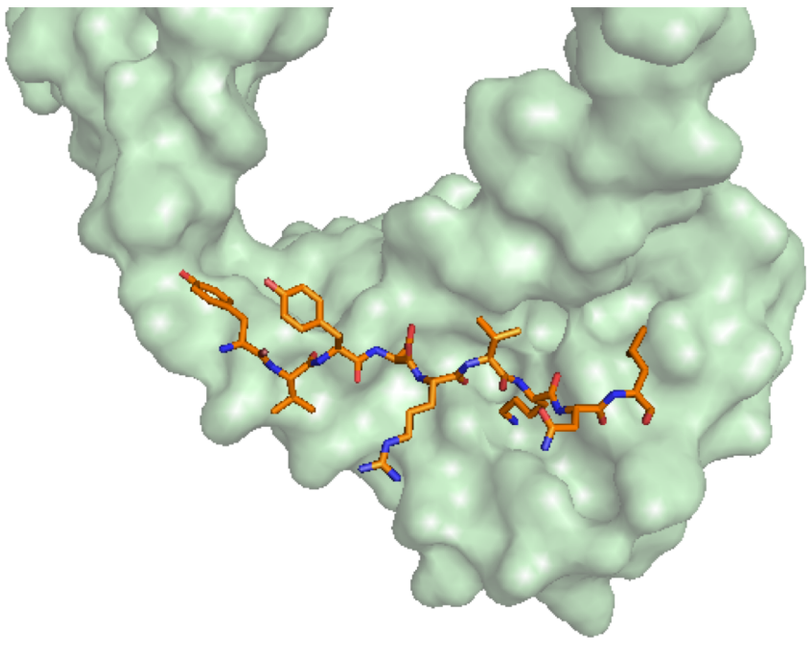
HLA-DRB1*13:02 (FYVYSRVKN)

F



HLA-DRB1*13:02 (VKNLNSSRV)

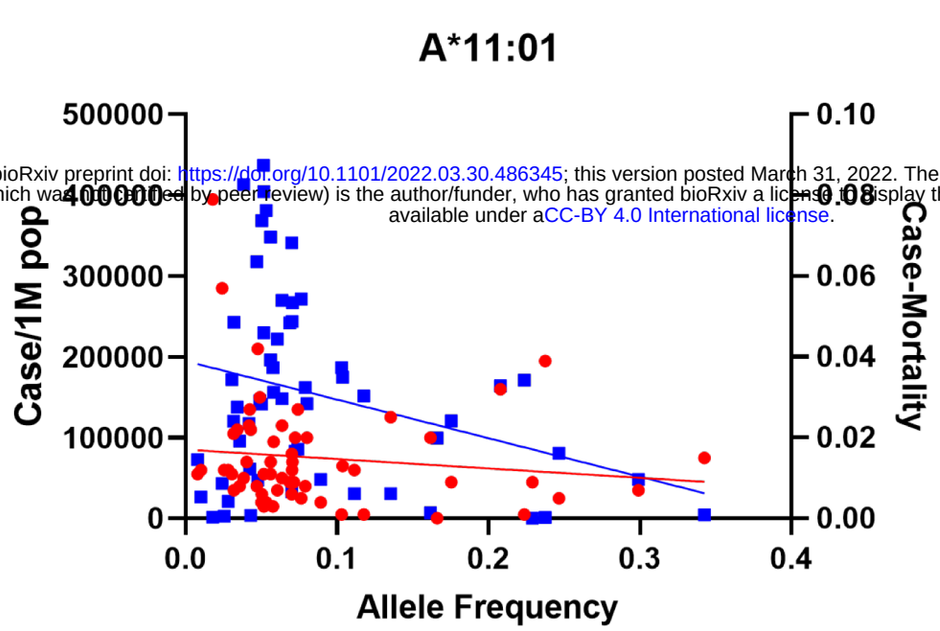
G



HLA-DRB1*13:02 (YVYSRVKNL)

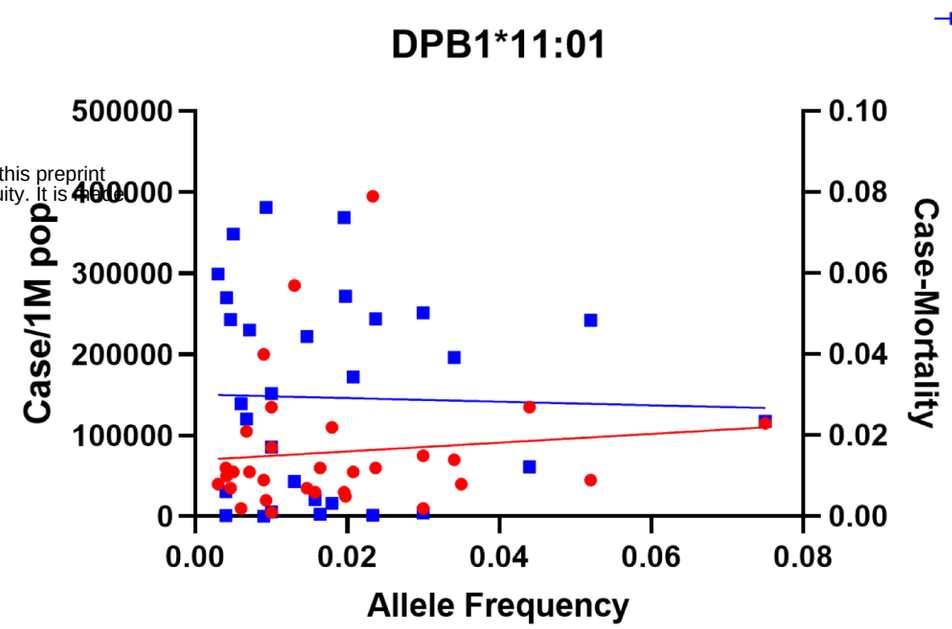
Figure 7

A. Protective alleles in EUR and AFR group



Y = -477572*X + 194762
R^2 = 0.08616
p = 0.0267

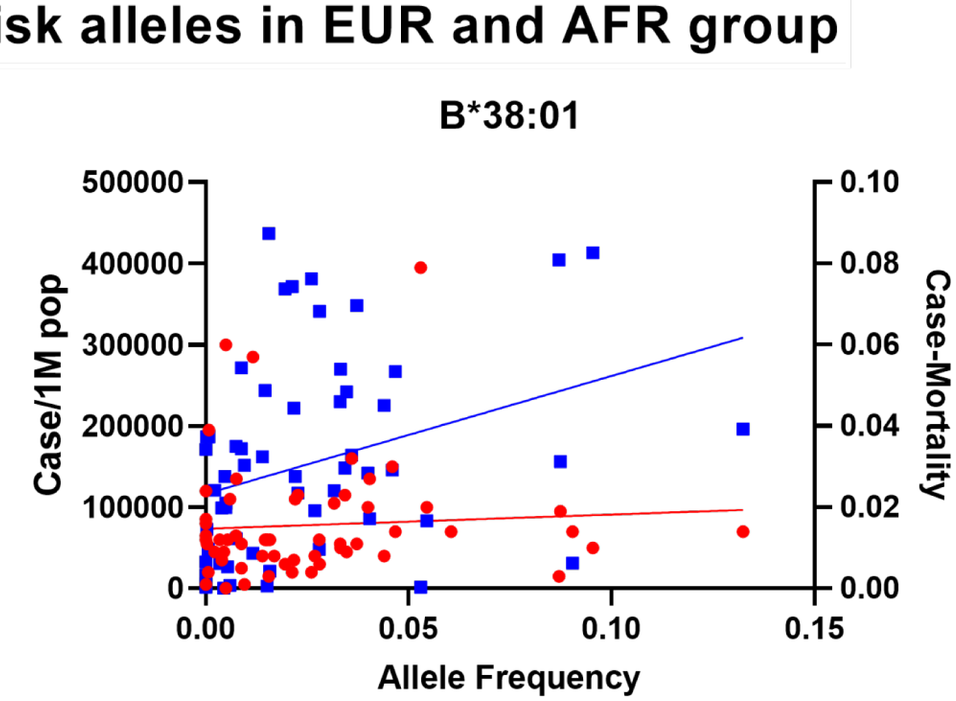
Y = -0.02326*X + 0.01701
R^2 = 0.01515
p = 0.3531



Y = -221737*X + 150340
R^2= 0.0008105
p = 0.8792

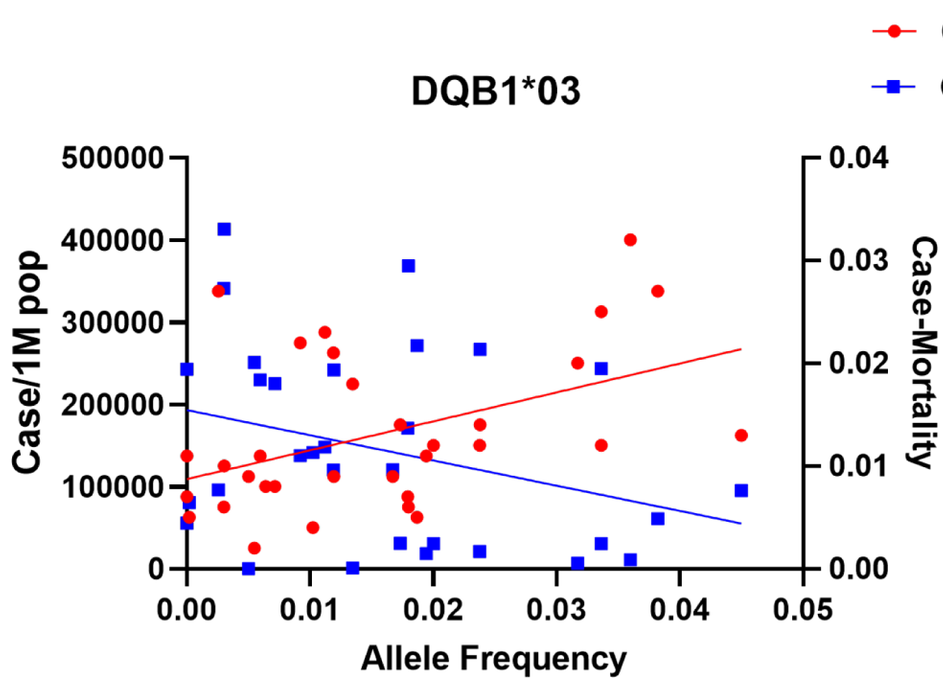
Y = 0.1080*X + 0.01386
R^2 = 0.01136
p = 0.5615

B. Risk alleles in EUR and AFR group



Y = 1452624*X + 116250
R^2 = 0.1124
p = 0.0101

Y = 0.03489*X + 0.01468
R^2 = 0.004614
p = 0.6092



Y = -3067898*X + 193178
R^2 = 0.1031
p = 0.0782

Y = 0.2813*X + 0.008725
R^2= 0.2017
p = 0.0099

Ancestral origins are associated with SARS-CoV-2 susceptibility and protection in a Florida patient population

Yiran Shen[1], Bhuwan Khatri[2], Santosh Rananaware[3], Danmeng Li[4], David A. Ostrov[4], Piyush K Jain[3], Christopher J. Lessard[2], Cuong Q. Nguyen[1,5,6]

[1]Department of Infectious Diseases and Immunology, College of Veterinary Medicine, University of Florida, [2]Genes and Human Disease Research Program, Oklahoma Medical Research Foundation, [3]Department of Chemical Engineering, University of Florida, [4]Department of Pathology, Immunology & Laboratory Medicine, University of Florida, [5]Department of Oral Biology, College of Dentistry, University of Florida, [6]Center of Orphaned Autoimmune Diseases, University of Florida, Gainesville, Florida, 32611-0880 USA.

Address correspondence:

Cuong Q. Nguyen, PhD

Department of Infectious Diseases and Immunology

PO Box 110880, College of Veterinary Medicine

University of Florida, Gainesville, Florida 32611-0880 USA

Telephone: 352-294-4180, Fax: 352-392-9704

Email: nguyenc@ufl.edu

Supplementary data

Supplementary Table 1 In silico binding prediction for protective (HLA-B*27:05) or risk (HLA-C*12:03) alleles in EUR group present structural protein of SARS-CoV-2 delta strain

Peptide	Protein	Bound Preference
ARDLICAQK	S	Protective Allele
RRARSVASQ	S	Protective Allele
KHTPINLVR	S	Protective Allele
LAATKMSEC	S	Risk Allele
FASIEKSNI	S	Risk Allele
IAPGQTGKI	S	Risk Allele
YRINWITGG	E	Risk Allele
KKLLEQWNL	E	Risk Allele
SRYRIGNYK	E	Risk Allele
AAVYRINWI	M	Protective Allele
LAAVYRINW	M	Protective Allele
FAAYSRYRI	M	Protective Allele
LTALRLCAY	M	Risk Allele
LAFVVFLLV	M	Risk Allele
LAILTALRL	M	Risk Allele
RRIRGGDGK	N	Protective Allele
DRLNQLESK	N	Protective Allele
GRRGPEQTQ	N	Protective Allele
FAPSASAFF	N	Risk Allele
SAFFGMSRI	N	Risk Allele
LSPRWYFYY	N	Risk Allele

Supplementary Table 2 In silico binding prediction for protective (HLA-B*27:05) or risk (HLA-C*12:03) alleles in EUR group present structural protein of SARS-CoV-2 delta strain

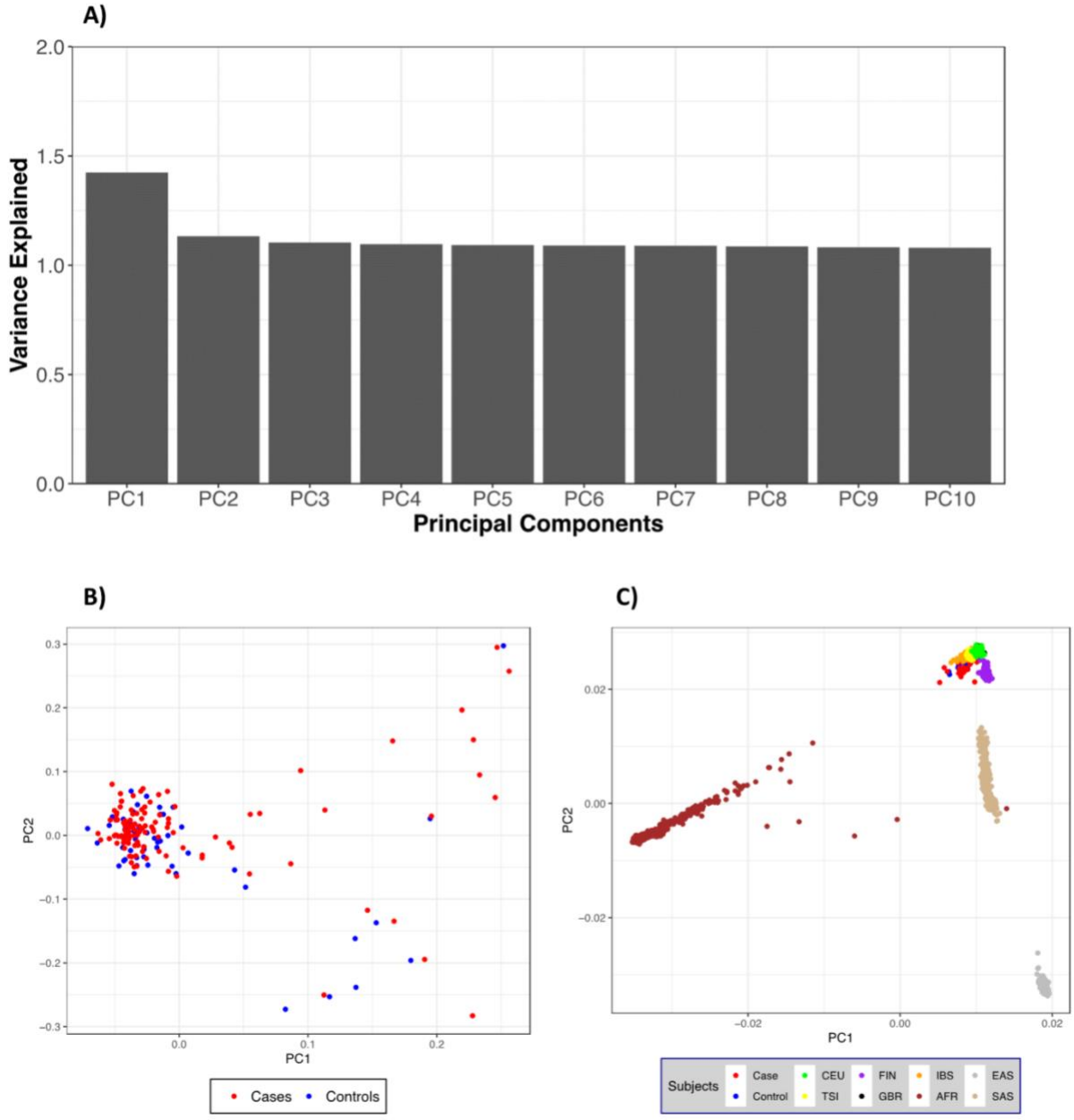
Peptide	Protein	Bound Preference
ARDLICAQK	S	Protective Allele
RRARSVASQ	S	Protective Allele
KHTPINLVR	S	Protective Allele
LAATKMSEC	S	Risk Allele
FASIEKSNI	S	Risk Allele
IAPGQTGKI	S	Risk Allele
YRINWITGG	E	Risk Allele
SRYRIGNYK	E	Risk Allele
QRVAGDSGF	E	Risk Allele
IAIAMACLV	M	Protective Allele
AAVYRINWI	M	Protective Allele
LAAVYRINW	M	Protective Allele
LAFVVFLLV	M	Risk Allele
LTALRLGAY	M	Risk Allele
LAILTALRL	M	Risk Allele
RRIRGGDGK	N	Protective Allele
DRLNQLESK	N	Protective Allele
GRRGPEQTQ	N	Protective Allele
FAPSASAFF	N	Risk Allele
SAFFGMSRI	N	Risk Allele
LSPRWYFYY	N	Risk Allele

Supplementary Table 3 In silico binding prediction for protective (HLA-DRB1*13:02) or risk (HLA-DQA1*05:01-DQB1*03:01) alleles in AFR group present structural protein of SARS-CoV-2 delta strain

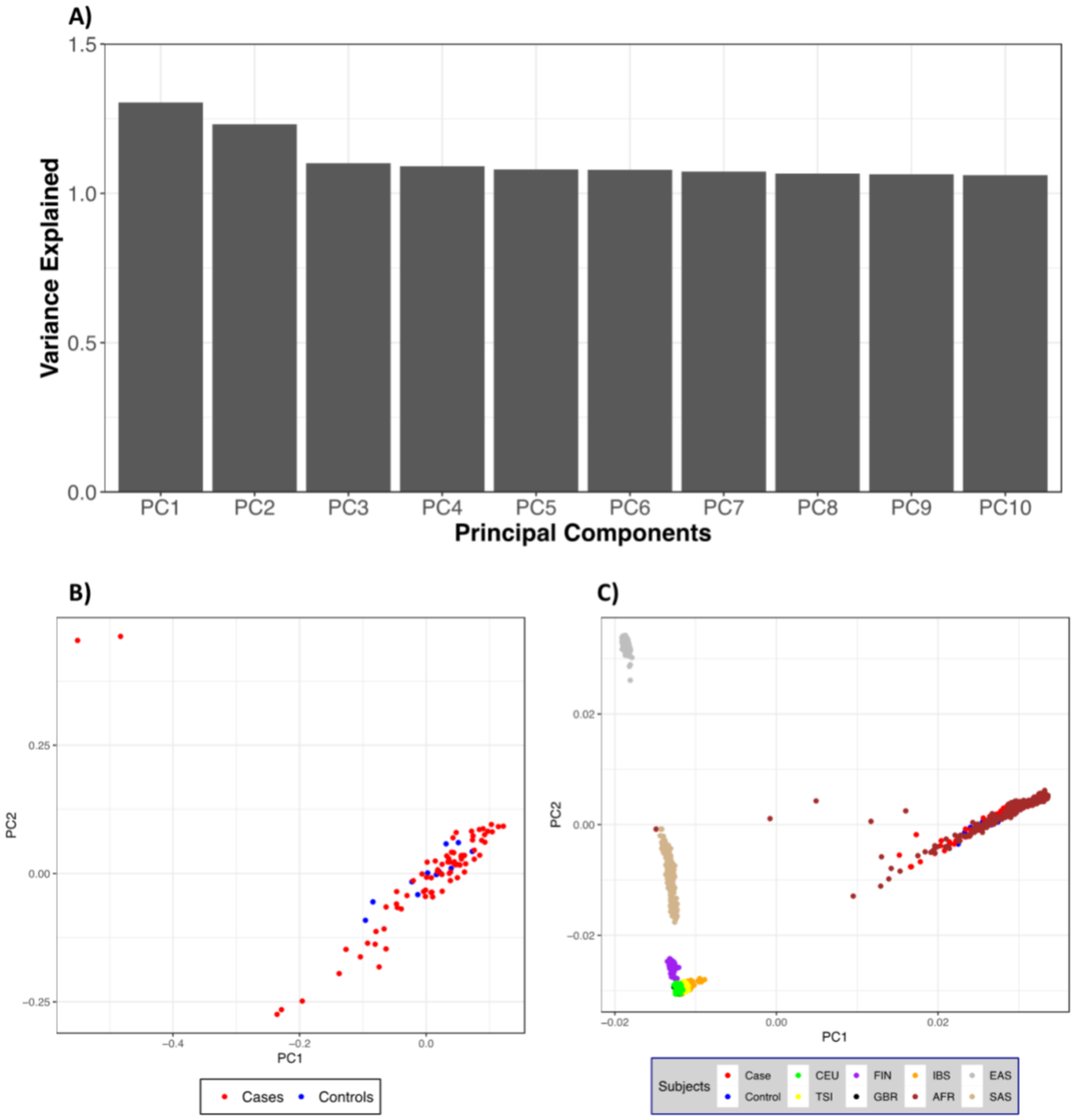
Peptide	Protein	Core bound protective	Core bound risk	Bound Preference
VGGNYNYRYRLFRKS	S	YNYRYRLFR	YNYRYRLFR	Protective Allele
PRTFLLKYNENGTIT	S	LKYNENGTI	LKYNENGTI	Protective Allele
KKSTNLVKNKCVNFN	S	LVKNKCVNF	LVKNKCVNF	Protective Allele
ITPCSFGGVSVITPG	S	FGGVSVITP	SFGGVSVIT	Risk Allele
ECDIPIGAGICASYQ	S	IGAGICASY	IGAGICASY	Risk Allele
TPCSFGGVSVITPGT	S	FGGVSVITP	SFGGVSVIT	Risk Allele
QFAYANRNRFLYIIK	E	YANRNRFLY	FAYANRNRF	Protective Allele
FIASFRLFARTRSMW	E	FRLFARTRS	RLFARTRSM	Protective Allele
TNILLNVPLHGTILT	E	ILLNVPLHG	ILLNVPLHG	Protective Allele
INWITGGIATAMACL	M	ITGGIATAM	ITGGIATAM	Protective Allele
MADSNGTITVEELKK	M	ADSNGTITV	NGTITVEEL	Protective Allele
NWITGGIATAMACLV	M	IATAMACLV	ITGGIATAM	Protective Allele
LVKPSFYVYSRVKNL	M	FYVYSRVKN	YVYSRVKNL	Risk Allele
VYSRVKNLNSSRVPD	M	VKNLNSSRV	VKNLNSSRV	Risk Allele
SFYVYSRVKNLNSSR	M	YVYSRVKNL	YVYSRVKNL	Risk Allele
FTALTQHGKEGLKFP	N	LTQHGKEGL	ALTQHGKEG	Protective Allele
FKDQVILLNKHIDAY	N	ILLNKHIDA	VILLNKHID	Protective Allele
QVILLNKHIDAYKTF	N	ILLNKHIDA	NKHIDAYKT	Protective Allele
LGTGPEAGLPYGANK	N	PEAGLPYGA	PEAGLPYGA	Risk Allele
GTGPEAGLPYGANKD	N	PEAGLPYGA	PEAGLPYGA	Risk Allele
YYLGTGPEAGLPYGA	N	TGPEAGLPY	PEAGLPYGA	Risk Allele

Supplementary Table 4 In silico binding prediction for protective (HLA-DRB1*13:02) or risk (HLA-DQA1*05:01-DQB1*03:01) alleles in AFR group present structural protein of SARS-CoV-2 omicron strain

Peptide	Protein	Core bound protective	Core bound risk	Bound Preference
PRTFLLKYNENGTIT	S	LKYNENGTI	LKYNENGTI	Protective Allele
KKSTNLVKNKCVNFN	S	LVKNKCVNF	LVKNKCVNF	Protective Allele
KSTNLVKNKCVNFNF	S	LVKNKCVNF	LVKNKCVNF	Protective Allele
ITPCSFGGVSVITPG	S	FGGVSVITP	SFGGVSVIT	Risk Allele
ECDIPIGAGICASYQ	S	IGAGICASY	IGAGICASY	Risk Allele
TPCSFGGVSVITPGT	S	FGGVSVITP	SFGGVSVIT	Risk Allele
QFAYANRNRFLYIIK	E	YANRNRFLY	FAYANRNRF	Protective Allele
FIASFRLFARTRSMW	E	FRLFARTRS	RLFARTRSM	Protective Allele
TNILLNVPLHGTILT	E	ILLNVPLHG	ILLNVPLHG	Protective Allele
INWITGGIAIAMACL	M	ITGGIAIAM	ITGGIAIAM	Protective Allele
NWITGGIAIAMACLV	M	IAIAMACLV	ITGGIAIAM	Protective Allele
RINWITGGIAIAMAC	M	ITGGIAIAM	ITGGIAIAM	Protective Allele
LVKPSFYVYSRVKNL	M	FYVYSRVKN	YVYSRVKNL	Risk Allele
VYSRVKNLNSSRVPD	M	VKNLNSSRV	VKNLNSSRV	Risk Allele
SFYVYSRVKNLNSSR	M	YVYSRVKNL	YVYSRVKNL	Risk Allele
FKDQVILLNKHIDAY	N	ILLNKHIDA	VILLNKHID	Protective Allele
QVILLNKHIDAYKTF	N	ILLNKHIDA	NKHIDAYKT	Protective Allele
NFKDQVILLNKHIDA	N	VILLNKHID	FKDQVILLN	Protective Allele
LGTGPEAGLPYGANK	N	PEAGLPYGA	PEAGLPYGA	Risk Allele
GTGPEAGLPYGANKD	N	PEAGLPYGA	PEAGLPYGA	Risk Allele
YYLGTGPEAGLPYGA	N	TGPEAGLPY	PEAGLPYGA	Risk Allele



Supplementary Figure 1. Quality Control and population stratification in EUR ancestry (A) Differential principal components analysis based on variance; (B) Case (red) and control (blue) samples were plotted by PC1 and PC2; (C) 1000 genome reference population was used and plotted by PC1 and PC2, case and controls samples were overlaid with the reference population to show ethnic distribution.



Supplementary Figure 2. Quality Control and population stratification in AFR ancestry. (A) Differential principal components analysis based on variance; (B) Case (red) and control (blue) samples were plotted by PC1 and PC2; (C) 1000 genome reference population was used and plotted by PC1 and PC2, case and controls samples were overlaid with the reference population to show ethnic distribution.