Application and assessment of peptide-MHC binding affinity prediction

By: Austin Nguyen

A DISSERTATION

Presented to the Department of Biomedical Engineering and the Oregon Health & Science University School of Medicine in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

October 2022

Department of Biomedical Engineering
School of Medicine
Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the PhD dissertation of
Austin Nguyen
has been approved

_____
Mentor: Abhinav Nellore, Ph.D Assistant Professor of Biomedical Engineering

_____

Mentor: Reid Thompson, M.D., Ph.D Assistant Professor of Biomedical Engineering

_____

Chair: Jeremy Goecks, Ph.D Associate Professor of Biomedical Engineering

_____

Member: Laura Heiser, Ph.D Associate Professor of Biomedical Engineering

_____

Member: Klaus Früh, Ph.D Professor, VGTI-Vaccine and Gene Therapy Institute

_____

Member: Kyle Ellrott, Ph.D Assistant Professor of Biomedical Engineering

Table of Contents

Acknowledgments

I would like to give special thanks to the following people:

- My mentors: Dr. Reid Thompson and Dr. Abhinav Nellore
- My fellow lab members at PDXgx: Ben Weeder, Julianne David, Mary Wood, Sean Maden, Ryan Melson, Max Schreyer
- My dissertation advisory committee members: Dr. Jeremy Goecks (chair), Dr. Laura Heiser, Dr. Klaus Früh
- My defense committee member: Dr. Kyle Ellrott
- My collaborators at UCSF (Chapter 3): Tasneem Yusafali. Dr. Jill Hollenbach
- BME and OHSU support: Alex Breiding, Holly Chung, Dr. Monica Hinds, Lauren Kronebusch, Erica Hankins Regalo, Dr. Sandra Rugonyi, JoAnn Takabayashi
- And most importantly, my family and friends especially: my parents, Dr. Cate Wrege, Karen DSouza, Mary Wood, Ben Weeder, Nick Calistri, Dr. Ali Hamdan, and Maricruz Vasquez.

Abstract

Infectious diseases have historically been the leading cause of death worldwide. While cardiovascular disease has slowly overtaken infectious diseases as the leading cause of death over the last 100 years, infectious diseases continue to have a huge burden on our planet, costing millions of life years and trillions of dollars as evidenced by our most recent COVID-19 pandemic. Viral infections like SARS-CoV-2, can be detected and eliminated through the MHC class I antigen presentation pathway. Identifying which viral targets can be recognized by each person's individual immune system is critical, both for evaluating whether current treatments can work, and for developing future vaccines. In my dissertation, I developed a framework to predict and assess susceptibility to infectious disease via peptide-MHC binding. First, I assessed the binding affinity of SARS-CoV-2 peptides across a wide number of HLA -A, -B, and -C alleles and compared their ability to bind SARS-CoV-2 with the closely related SARS-CoV, validating that this binding affinity analysis predicted an allele HLA-B*46:01 as a deleterious allele for both SARS-CoV and SARS-CoV-2 as previously found in hospital case studies during the first SARS outbreak. I then explored the potential for cross-protective immunity by evaluating conserved peptides' binding potential with different HLA alleles across coronaviruses, finding that there was little to no relationship between predicted binding and level of conservation. I reported global distributions of HLA types, identifying potentially vulnerable populations to the current pandemic. In order to validate these predictions, I investigated the relationship between severity of SARS-CoV-2 disease and HLA type in 3,235 individuals with confirmed infection, finding that only the DPB1 locus to be associated with whether an individual developed symptoms after multiple comparison correction. Age, BMI, asthma status, and autoimmune disorder status were found to be comorbidities that far outweighed metrics such as patient specific predicted binding to SARS-CoV-2 peptides. To improve upon this initial framework of combining HLA genotypes with peptide-MHC binding to assess disease susceptibility as well as to

generally assess the validity of peptide-MHC binding predictors, which are widely used for investigational and therapeutic application, I investigated 4 popular peptide-MHC binding affinity predictors across a range of peptide sources and MHC class I alleles. I found significant inconsistencies in binding affinity predictions across all tools. Further, I developed and applied a method to evaluate the ability of peptide-MHC predictors to detect differences in amino acid physical properties across peptide sets, finding that while these tools were unreliable across single tool and allele combinations, for several alleles, they predicted strong preferences for specific physical properties. My work raises fundamental questions about the reliability of peptide-MHC binding prediction tools and their downstream implications. In summary, I demonstrate the potential of a framework of assessing disease susceptibility from viral peptides, show that, at least for SARS-CoV-2, the predictions from this framework were not confirmed by the clinical and survey data over a large population, discover the poor reliability of peptide-MHC predictors which may have contributed to the lack of predictive power for SARS-CoV-2, and outline a critical need to develop more accurate peptide-MHC predictors.

Chapter 1: Introduction

1.1 Introduction

March 11, 2020. After over 100,000 cases and 4,000 deaths in hundreds of countries, the WHO declared COVID-19 as a global pandemic. 2 days later, in the United States, states began to shut down to prevent the spread of the airborne virus. By March 17th, Moderna began human trials of its vaccine, but emergency use authorization (EUA) would not be granted until December 18, 2020. Social distancing measures quickly followed but these measures were not enough to stop the rampant spread of disease. While many infected individuals were either asymptomatic or had mild enough symptoms to avoid hospitalization, the threat of COVID-19 forced many individuals and families, including mine, into personal lockdowns, with the uncertainty that the measures we took would be enough to protect our family. As a family with a cancer survivor, the biggest question that we asked daily was: what can we do to protect ourselves and each other? The best we could do was quarantine to the best of our ability and take the vaccines and boosters when they were available. Even today, after 4 shots of the vaccine, with monoclonal antibody prophylactics, and regular mask usage, life has not yet returned to “normal” for us and many others.

While being immunocompromised makes one both much more likely to get and suffer more severe COVID-19 disease (and generally more severe symptoms for any disease), the first applicable question to any disease is who are the vulnerable individuals? COVID-19 is the most relevant disease to ask this question today, however, it is neither the first, nor will be the last pandemic that we face, nor does this question only apply to only infectious diseases. The second question is what are the most effective ways to reduce the effects of disease? For COVID-19, it was the vaccine, even after new variants reduced their effectiveness. For individuals with cancer, there is no one size fits all solution.

What ties an infectious disease such as SARS-CoV-2 and cancer together is the need for the immune system to recognize and rid itself of harmful cells. Here I introduce the background that motivates my dissertation research. Section 1.2 introduces antigen processing and presentation in cancer and viral infections. Section 1.3 delves into computational methods of modeling and predicting antigen presentation. Section 1.4 discusses the broad application of these computational methods and the success of their downstream products. Finally, Section 1.5 discusses the problems my research addresses and the contributions to the field made by this dissertation.

1.2 Antigen processing and presentation

Medicine has 3 components: diagnosis, treatment, and prevention of disease. Vaccines are often considered a modern medical miracle; they are simply products that teach your immune system to recognize and eliminate enemies, satisfying 2 parts of medicine: treatment and prevention. This starts with the ability of the immune system to recognize self from non-self. This process starts with antigen processing and presentation: a key process where antigens, foreign substances which can induce an immune response, are, along with our own peptides, subsequently displayed on the cell surface for T cells to identify and for our immune system to generate an immune response (1–5). In the context of infectious disease and cancer, antigen processing starts with the major histocompatibility complex (MHC) class I proteins, which bind peptides, usually between 8-12 amino acid residues, to their binding groove (6). MHC class II proteins bind longer peptides, up to 25 amino acids in length. These peptides are generated by breaking down intracellular proteins via the proteasome. Antigenic peptides are specifically generated by the immunoproteasome, a highly efficient version of the proteasome (7–9). After these peptides are generated, then processed further by cytosolic and endoplasmic reticulum aminopeptidases, they then have the opportunity to bind to an MHC class I molecule (Figure 1.1).

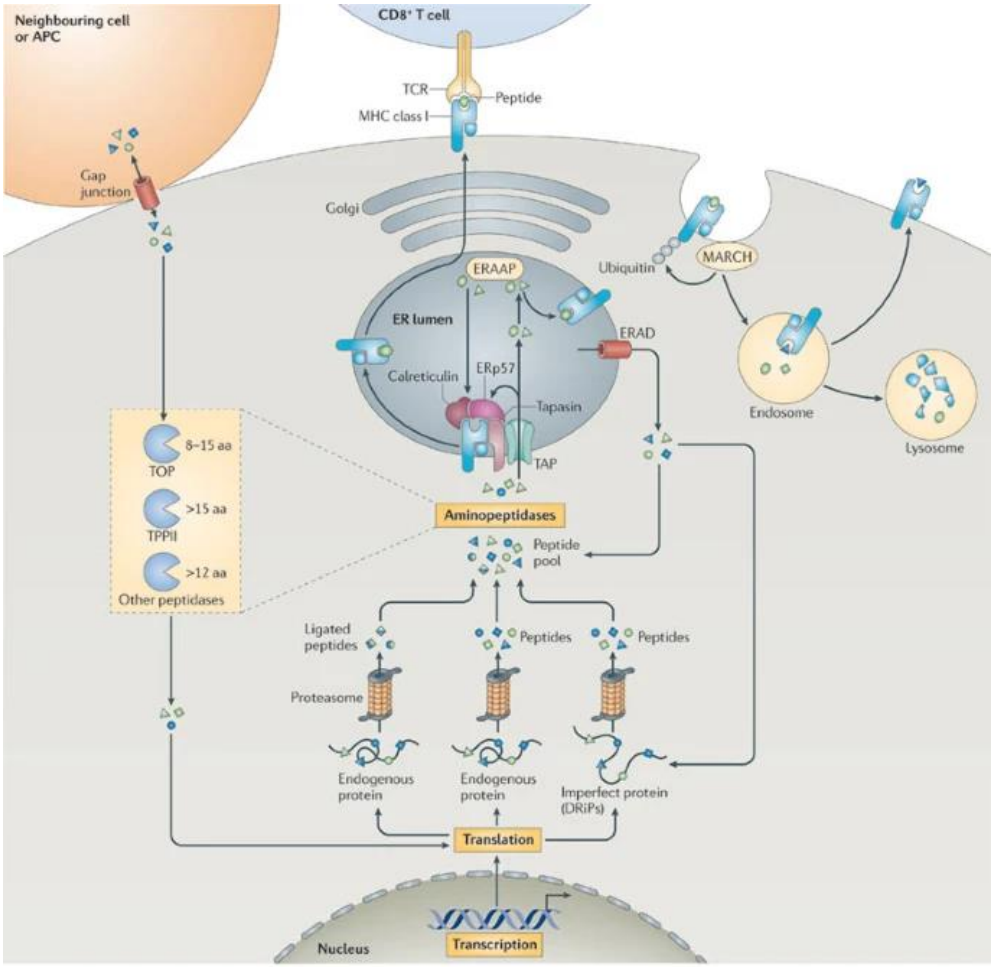


Figure 1.1. Towards a systems understanding of MHC class I and MHC class II antigen presentation (5)*. The lifecycle of an antigen, starting from transcription and translation. Proteins are broken down by proteasomal cleavage which then composes the peptide pool, which are further processed by aminopeptidases. These smaller peptides are transported by Transporters associated with Antigen Processing (TAP) protein complexes to respective MHC class I molecules for binding. Peptides that fail to bind are transported back to the cytosol and further trimmed or recycled by various peptidases.*

As the human cell fights foreign peptides from many sources, from bacteria, viruses, to cancer cells, MHC class I proteins are able to recognize a wide array of peptides. This relies on the binding of these peptides to the polymorphic region of the MHC molecule: the peptide-binding groove (6,10,11). The majority of MHC class I molecules never bind to a peptide and are ultimately degraded; this allows for almost all peptides to be “screened” by MHC class I, even when rates of

peptide synthesis are increased during events such as a viral infection (12). MHC class I proteins are encoded by some of the most highly polymorphic genes in humans, genes that contain the greatest frequency of mutations; these polymorphism gives rise to different peptide-binding grooves and thus, differing capacities to bind and recognize unique peptides (13). The majority of MHC class I proteins in humans are encoded by genes at the HLA-A, HLA-B, and HLA-C regions of chromosome 6. Nearly all individuals have two HLA-A/B/C haplotypes, giving rise to a minimum of 3 and a maximum of 6 distinct alleles (13,14). As of June 2022, there were over 24,000 different HLA class I alleles, giving the possibility for an incredibly wide and unique repertoire of peptides an individual can recognize and bind (15). Indeed, it is estimated that an average individual with 6 heterozygous alleles could present on the order of 10^{12} unique peptides (14,16).

After binding of the peptide to the MHC molecule, the peptide-MHC complex is transported to the cell surface for T cell recognition. Within the cell, this process is continuous, allowing T cells to constantly monitor the given cell's proteome. When a peptide-MHC complex that a T cell has not previously encountered during T cell maturation, for example during viral infection or oncogenesis, T cell activation can occur, and an immune response can be stimulated (3,14). The identification of these peptides is important for evaluating what can trigger an immune response; thus, antigen identification is crucial for understanding both infectious and autoimmune diseases as well as the development of immunotherapies and vaccines.

1.2.1 HLA type and disease association

Each person's immune system has its own capacity to recognize and eliminate infectious disease. With the vast variety in individuals' ability to bind peptides to their MHC repertoire, there is a likewise variety in both susceptibility and severity of disease. The association between many autoimmune disorders and HLA type has been known for decades; it accounts for half of all known genetic risk factors (17–19). The most well-studied, HLA-associated disease is Type 1 diabetes. A

subset of HLA alleles, mostly class II DQ and DR haplotypes, are between 40-50% of the familial aggregation of Type 1 Diabetes (20), with many more alleles from both class I and class II presenting smaller protective or deleterious effects.

The relationship between HLA class I and infectious disease susceptibility is well documented. For example, HLA B27,51, and 57:01 all are associated with protection from HIV infection (21,22). Hepatitis B and C are another 2 viruses that both are globally widespread and have HLA alleles associated with either increased clearance or progression (23–26). The number of both autoimmune and infectious diseases with strong HLA associations and mechanistic explanations is increasing as we gather more disease and genotype data (27). Because we have increasing evidence that some HLA alleles provide stronger protection versus specific diseases, we next want to understand what about the HLA-disease relationship makes them more protective and how we can identify vulnerable populations.

1.2.2 HLA evolution

The HLA region, like most of the human genome, is subject to selective pressures. As HLA is incredibly important for immunity and each HLA type can recognize and bind specific peptides, increased diversity may be favored at the HLA locus as there may be a wider ability to present peptides (28). While there are multiple hypotheses about the specific events that drive HLA evolution, one of those with substantial evidence is pathogen-driven selections: specific alleles are favored because of their ability to provide protection from pathogen species or strains (29–32). Further, there is evidence that suggests that the selection of alleles available in a population is dictated by both specific pathogens and diversity of pathogens in the geographical region of the population (28). For example, several studies report that specific HLA alleles provide increased resistance while other alleles provide decreased fitness against single pathogens (33–39). There is evidence that there is a positive correlation between both pathogen diversity and HLA class I promiscuity as well as pathogen

diversity and HLA variation (29,30). Heterozygotes at the MHC loci; with more available alleles and theoretically a greater range of recognition against pathogen peptides, are more likely to have higher fitness in general and against a variety of pathogens (29,31,40).

1.2.3 Antigen targets in viral infections

The antigen processing and presentation are vitally important for dealing with viral infections. Viruses are infectious agents composed of nucleic acid surrounded by a protein shell and need to infect a host cell in order to replicate. Oftentimes, viruses are cleared from the body quickly such as the common cold, or when they cause a persistent infection, they do not have any major symptoms, for example, the herpes simplex virus (41,42). However, both acute and chronic viral infections can have devastating effects on the host; SARS-CoV-2 alone has killed over 6.5 million at the time of writing, with hundreds of millions more with potential long term symptoms (43,44).

In healthy individuals, presented peptides of viral origin, both from exogenous sources through endocytosis as well as endogenous viral antigen are similarly processed, presented, and recognized by T cells (6). Typically, viral infections elicit strong cytotoxic T lymphocyte responses which can allow for the quick clearance of a virus. However, when the immune system fails to recognize the virus, as in the case where antibodies are not effective (e.g. new variants of a virus), viruses will infect host cells and reproduce (45). There are two main methods for fighting viral infection: antiviral drugs and vaccines. Antiviral drugs follow 2 strategies: targeting the host cell or the virus themselves. Drugs that target the viruses directly can target a wide variety of mechanisms such as viral attachment, entry, protease inhibitors, and these drugs are often among the top drugs by sales (46). However, for many viral infections, there are no effective antiviral drugs; in fact, there are only 12 viruses in which there are antiviral drugs and the majority of antiviral drugs are used to treat HIV, with limited effectiveness. The most effective HIV drugs are geared towards prevention of HIV integration by inhibiting integrase,

a critical viral protein that spices viral DNA into the chromosome (47). However, they have a significant number of side effects and are not yet well tolerated.

Vaccines have been present in some form since the late 1700 with the discovery of smallpox immunization. Pasteur in the late 1800s developed the first vaccines using attenuated viruses in humans, and by the mid-1980, we eradicated smallpox with worldwide vaccination programs (48). As the number of known infectious diseases has increased, so has the number and variety of vaccines; currently there are 104 FDA approved vaccines for a wide variety of viruses (49). There are a wide variety of vaccines, but they generally separate into 2 categories: antibody inducing and T cell response inducing (48–54). Not all types of vaccines activate the antigen processing and presentation pathways and subsequently T cell response. Vaccines that induce T cell response are generally effective, some of which are even capable of producing broad responses against more than one strain of disease (50,51,54–58). For example, T cell responses are vitally important in the prevention of severe COVID-19 disease, especially for variants that escape antibody response (52).

Determining viral targets that can induce a T cell response is vital; this process of determining immunogenicity again relies on the key step of peptide-MHC binding (59–62). The relationship between binding affinity and immunogenicity of approximately 100 different hepatitis B virus (HBV)-derived potential epitopes estimated that a binding affinity threshold (what concentration of peptide would be necessary to achieve 50% inhibition of the MHC molecule) of approximately 500nMdetermines the capacity of a peptide epitope to elicit a CTL response (61). Additionally, both the strength of predicted binding and the measure of dissimilarity between a foreign peptide and its human counterpart is positively correlated with the likelihood of generating an immune response. With viral epitopes generally having a larger measure of dissimilarity against human peptides, this makes them good targets for therapeutics.

1.2.4 Neoantigens in cancer

Cancer is a group of diseases where cells chronically proliferate and can spread throughout the body. Cancers are defined by a number of hallmarks including aberrant growth, resistance to cell death, angiogenesis, and invasion of healthy tissue (63–65). Genetic mutations that drive a cancer's numerous survival advantages over normal cells can be targets for both early detection and treatment of the cancer. Some of these genetic mutations produce novel peptides called neoantigens which have the ability to elicit a strong immune response via the antigen processing and presentation pathways as described above (66–69). Not all mutations produce neoantigens; only mutations that result in a novel amino acid sequence relative to the individual's regular proteome, expressed within the cancer cell, and bind with relatively high affinity to at least one of the individual's MHC molecules can produce neoantigens. Neoantigens are both targets themselves and used to evaluate which treatments may be viable for patients (67,68,70–77).

As targets, cancer immunotherapy, also known as immuno-oncology, aims to take advantage of these neoantigens by boosting and educating the immune system to recognize and eliminate cancer cells. Currently, there are 5 main forms of cancer immunotherapy: immune checkpoint inhibitors, adoptive cell therapy, monoclonal antibody, oncolytic viral therapy, and cancer vaccines (78,79). All 5 of these cancer immunotherapies, in some aspect, activate the antigen processing and presentation pathways. The targetability of these products varies widely; a peptide that is unique to the tumor and is not expressed anywhere in normal tissue is most likely to produce the strongest anti-tumor response and the lowest chance of normal tissue toxicity (67,72,80).

Not all cancer immunotherapies are effective for all patients. As biomarkers, neoantigen-related metrics such as tumor mutational burden (the number of mutations unique to the tumor as compared to normal tissue- TMB) and quantity of nonsynonymous SNVs (single nucleotide variants that change the protein sequence) have been used as predictors of immunotherapy response (81–

83). Theoretically, the higher the number of mutations, the more tumor-specific targets the immune system has the opportunity to recognize and attack, thus improving the efficacy of immunotherapy. However, numerous studies have quantified the robustness of TMB as a predictor; TMB is only effective in some populations already receiving immunotherapies but not in immunotherapy-naive populations and as a partial predictor in a small number of cancers (69,84–87).

Efforts have shown that downstream metrics of mutational burden that include adjustments for sequence novelty (measuring change from the original peptide) and MHC binding have been shown to increase predictive benefit (60,84,88). This again supports the importance of the antigen processing and presentation pathways in evaluating both targets and biomarkers in cancer.

1.3 Predicting peptide-MHC binding

For both viral infections and cancer, the ideal target is one that can generate a reliable immune response. A tight binding must occur between the peptide and MHC molecule to generate this, thus the majority of research in this area focuses on identifying these peptides which may bind to a MHC molecule. Using computational methods in order to predict the relationship between peptide and MHC molecules enables drug developers and clinicians to more rapidly identify and filter through potential therapeutic targets. This, combined with the rapidly increasing quantities and decreasing cost of next-gen sequencing, has increased the availability and cost-effectiveness of therapies such as personalized vaccines (73,89).

Initial work in predicting peptide-MHC binding started with binding motif-based models. In the late 1980s, structural characteristics of peptides binding to mouse MHC molecules were first described quantitatively (90). Subsequently, it was found that anchor positions, specific locations where the peptide would attach to pockets in the MHC molecule binding groove, were conserved across many different MHC molecules but with differing binding motifs. This led to the first models such as SYFPEITHI with position-based amino acid scoring (91).

The next generation of peptide-MHC binding affinity predictors were machine learning tools. There are 3 main types of ML peptide-MHC binding affinity predictors: predictors that are trained on eluted ligand (EL) data (ligands that are profiled using mass spectrometry from lysed antigen presenting cells), predictors trained on binding affinity (BA) data (normally obtained by observing 50% binding threshold of the peptide to MHC molecule in nM as described above), and predictors trained on mixed data (BA + EL). The first ML models suffered from a severe lack of training data; the first artificial neural network model in 1998 had fewer than 300 peptide-MHC data points (92). As binding affinity datasets became larger, models became more refined, being able to make predictions on new peptide-MHC pairs.

With the increased number of mass spectrometry studies, eluted ligand data presented another opportunity for peptide-MHC binding software developers. This data was used as a stand-alone for some tools and in tandem with binding affinity data for others (93–100). The difficulty in 2 separate types of data comes from multi-allelic ligand data, meaning an eluted ligand may have come from any number of MHC molecules. Binding affinity data is much easier to model; it is a single event with 1 peptide and 1 MHC molecule. BA + EL models tend to pseudo-assign or multi-assign eluted ligand data points to a single allele. Despite the exponential increase in the amount of training data, the total number of peptides across all sources is still only estimated to capture an extremely small proportion of the set presented by MHC, thus computational imputation will continue to be an important tool in identifying and filtering peptide targets (59,94,101,102).

There are a large number of peptide-MHC predictors in use today. Among the most commonly used are netMHCpan, MHCflurry, MHCnuggets, IEDB consensus, and SYFPEITHI(103–106).

1.3.1 MHCflurry

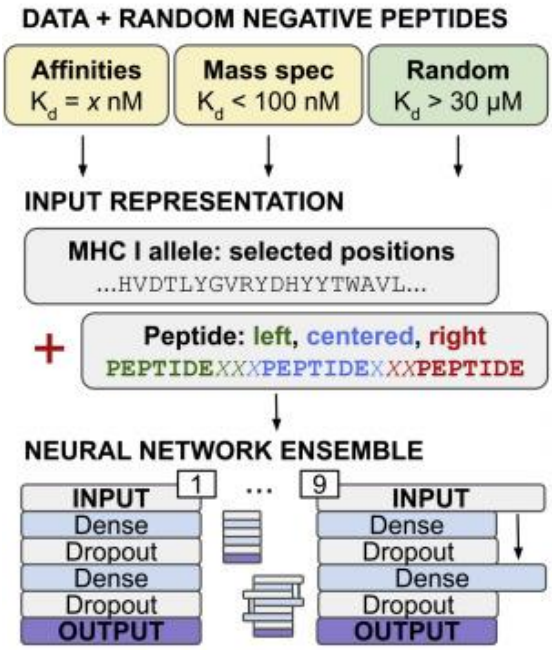


Figure 1.2. MHCflurry training and network architecture. Affinities are BA data, mass spec is eluted ligand data, and random are randomly generated peptides not found in BA or mass spec data. Peptide sequence is inputted into 3 15-mers concatenated together: left aligned, center aligned, and right aligned, creating a sequence of length 45 as input.

MHCflurry is a BA+EL model, trained specifically to discriminate published mass spectrometry data from unobserved peptides. MHCflurry uses a neural network architecture with 1 input layer, 3 hidden layers, and 1 output layer, as they found in their preliminary analysis that the deeper networks consistently outperformed shallower versions; however, the authors noted that the gains from additional layers were small compared to additional training data. MHCflurry distinguishes itself from other peptide-MHC tools by combining mass spectrometry identified peptides with unobserved decoys against the BA predictor (Figure 1.2). They found that the binding peptides in held back sets that contained established motifs were favored in the BA prediction and in a test set of held-out mass spec data, MHCflurry outperformed netMHCpan 4.0 and MixMHCpred 2 (97). A possible drawback of

using decoys, which have no prior information on whether they bind, as an automatic negative, is that for predictions outside the space of known mass spectrometry may result in false negative binders.

1.3.2 MHCnuggets

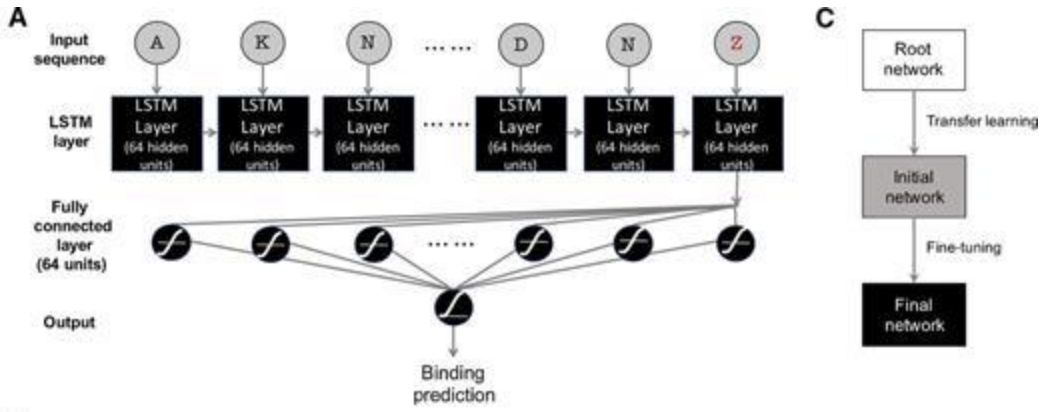


Figure 1.3. MHCnuggets network architecture. Inputs are peptide sequences of variable length. MHCnuggets uses transfer learning in order to generate predictions for alleles not directly in the training data or with fewer training examples.

MHCnuggets is a recurrent neural network with an input layer, an LSTM layer, a fully connected layer, and a single output node (Figure 1.3). They use this method in order to gain information from sequential data inputs (amino acid sequence) and can handle peptides of variable length. The authors report that MHCnuggets had higher positive predictive value (PPV) than MHCflurry and the netMHCpan suite of tools, and fewer overall binders than other methods, resulting in a smaller number of false-positive predictions for the alleles tested. They attribute this overall improvement to the LSTM network architecture, which handles a variety of lengths well without coercion to a specific k-mer length (98).

1.3.3 netMHCpan

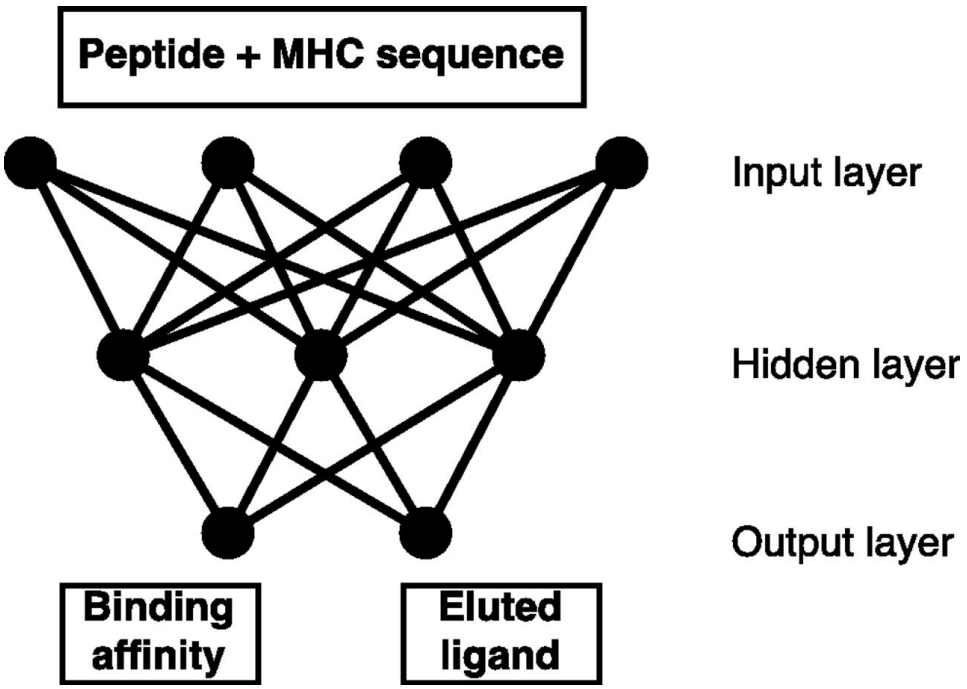


Figure 1.4. netMHCpan network architecture. netMHCpan converts all peptides into 9-mers as inputs using insertions or deletions of the original peptide sequence to the closest BLOSUM62 9-mers.

netMHCpan is an eluted ligand and binding affinity mixed model that combines MHC molecules, input data types, and multiple peptide lengths into neural network input (96,107,108). netMHCpan uses the simplest architecture of the tools mentioned: a machine learning framework consisting of 3 layers: an input layer, a single hidden layer, and an output layer (Figure 1.4). All networks were trained with back-propagation with stochastic gradient descent (96,109). Features of netMHCpan that other models such as MHCnuggets, which uses an LSTM model, do not have is the coercion of peptide input into 9-mers. This refers to the either insertion or deletion of peptides to the nearest 9-mer (95,110); for example, an 8-mer peptide input will have wildcard X amino acids attached in each possible position to make it a “9-mer”, then the highest predicted binding score of the peptides is kept as the score of the original 8-mer. The authors of netMHCpan benchmark

FRANK (per protein-based accuracy score) and PPV scores versus MHCflurry and MixMHCpred, but conclude that they are “significantly superior” over all other methods for both metrics.

1.3.4 HLAthena

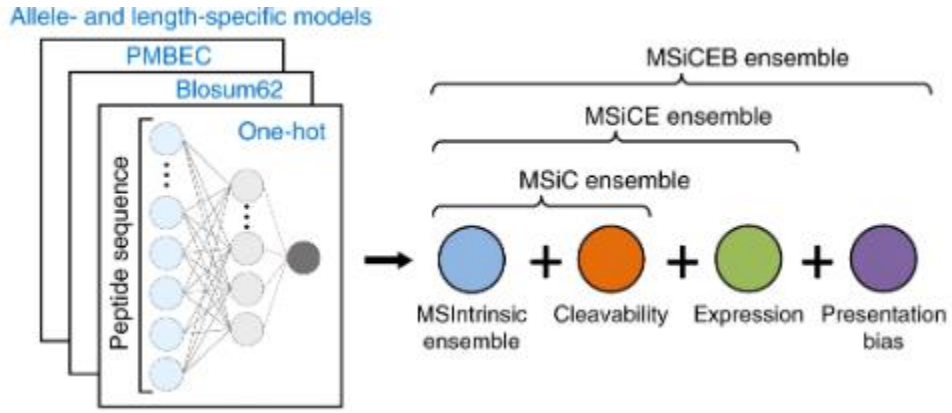


Figure 1.5. HLAthena network architecture and additional inputs. Inputs into HLAthena’s model are eluted ligands from up to 50 million single-HLA expressing cells per allele. Cleavability information was obtained from predicted cleavability (cleavnn). HLA expression and presentation bias information was obtained from mass spectrometry and ribosomal profiling respectively.

HLAthena is an eluted ligand model that is trained solely on in-house data. The authors claim that current prediction algorithms were not trained on high-quality epitope data and that with primarily BA data, “[other tools] do not account for intracellular availability of the peptide precursors or their processing by proteases” and “uneven accuracy in the prediction of epitopes binding to less common alleles in Caucasians, or those highly prevalent in other populations'' (93). These statements directly point to the heavy reliance of the aforementioned tools on the IEDB database (111) BA data, the most populous of all peptide-MHC data. HLAthena’s in-house dataset is composed of over 185,000 peptides eluted from mono-allelic cell lines, removing any possible confounding factors resulting from the use of EL multi-allelic data as with netMHCpan. At the time of publication, their mono-allelic data doubled the IEDB database’s repository of mono-allelic data. HLAthena found that their predictions for 8-mers was less accurate as those were observed to have lower cleavage scores. The authors

also suggested an additional explanation for the less accurate 8-mer prediction scores; 8-mers had the highest entropy as compared to the other k-mer lengths and integrating more features, such as cleavability, expression, and gene presentation bias, into the neural network increased performances for 8-mer predictions more than the other lengths (Figure 1.5). Benchmarking against MHCflurry, netMHCpan, and MixMHCpred, HLAthena outperformed all tools in PPV at multiple recall percentages.

1.4 Current uses of peptide-MHC binding predictors

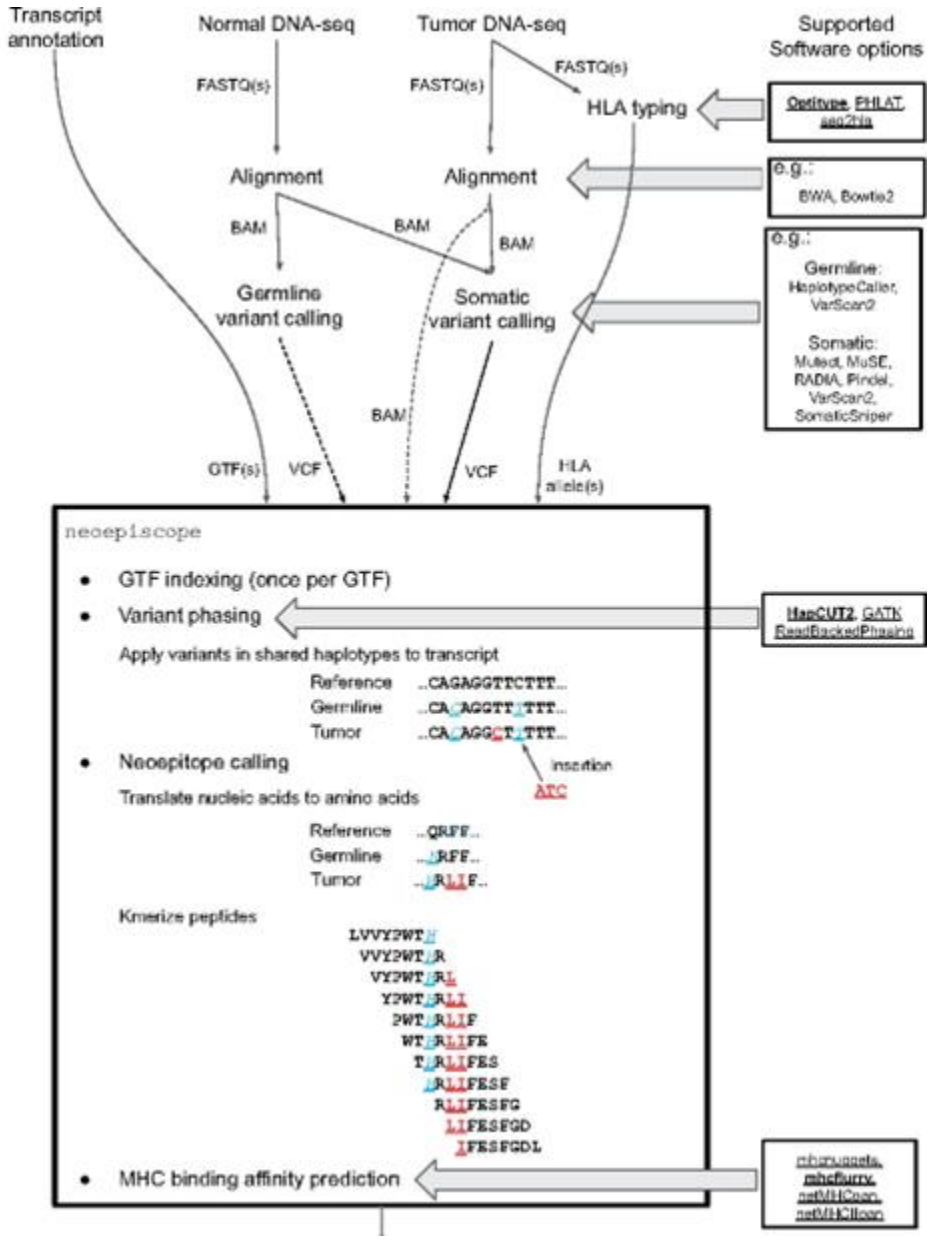


Figure 1.6 Neoepitope prediction pipeline diagram describing canonical neoepiscope workflow. Global inputs are shown at the top of the figure, connecting to neoepiscope with each option for processing listed to the right. Direct neoepiscope functionality is depicted within the outlined box, with example sequences showing both somatic (underlined) and background germline variants (underlined, italic) in a mock transcript sequence, and their translation and kmerization into short peptides (8-mers).

Peptide-MHC binding predictors have been widely used in cancer neoepitope discovery. One of the first pipelines for computationally identifying tumor neoantigens, pVAC-seq, relies on the use of epitope binding prediction software (netMHC/netMHCpan) (74). Subsequent software pipelines which used a wider variety of tools followed suit such as CloudNeo (112), MuPeXi (113), and Neoepiscope (71). The starting points for all of these software begins with tumor and normal DNA-seq, then performing alignment and variant calling to obtain lists of germline and somatic variants. The tumor variants may be validated with RNA-seq if available (Figure 1.6). neoepiscope additionally takes advantage of additional RNA-seq data to confirm any predicted phased variants in earlier steps. After a list of potential peptides is created from the lists of variants, these peptides are inputted into peptide-MHC prediction software. The majority of these tools use tools from the netMHCpan suite (71,74,112,113), with some tools such as neoepiscope offering prediction alternatives with MHCflurry and MHCnuggets.

The results from the peptide-MHC predictors are sorted and filtered, typically using a threshold of 500nM. As mentioned above, 500nM has been the accepted value for immunogenicity as it was discovered that hepatitis B epitopes elicited a cytotoxic T lymphocyte response at approximately the 500nM threshold (61). These peptides are usually filtered further by comparing the sets of tumor peptides and the normal peptides and removing any duplicates as well as by coverage if given RNA-seq data to the user specified thresholds. Finally, this end result is outputted as a list of possibly immunogenic targets.

Despite the use of the aforementioned pipelines and numerous other custom pipelines used to identify tumor-specific neoantigens across a wide variety of cancers (59,68,72,75,76,114–121), the number of successful tumor-specific vaccines is small, in no small part due the low translatability of the potential immunogenic targets. The extreme complexity of the immune system response to a vaccine, even without the added complications of cancer, involve a high degree of moving parts from B and T cell epitopes to other immune structures such as pathogen associated molecular pattern responses. While in theory, we can identify and incorporate into our models many of the factors that contribute to immunogenicity, we have not yet been able to incorporate factors such as the attractiveness of specific peptide-MHC pairs to antigen presenting cells or ability to up-regulate immune response.

Cancer therapeutics pipelines are far from the only uses of peptide-MHC binding predictors. They are used widely in transplantation and autoimmune research (122,123). The majority of the recent literature citing and using tools such as netMHCpan are studies investigating the SARS-CoV-2 virus, many of which are looking to computationally design vaccine candidates. Importantly, computational design of vaccines may allow us to match the ever-increasing number of circulating variants (124). Many of these studies lack key nuances for creating a broadly applicable vaccine. For example, some only predict against a small number or single allele and do not provide affinity predictions for other common alleles, which both broadens the number of candidates and reduces the number of individuals in which the vaccine would be most effective (125–128), nor do they evaluate peptides outside those arising from specific proteins of interest such as the spike, which severely limits the available number of peptides as it represents only a small number of possible SARS-CoV-2 peptides.

1.5 Summary

1.5.1 Challenges and opportunities

As the global burden of cancer and infectious disease increases, there will be an increased need to develop new therapies with high efficacy and equitably. There is significant potential to use peptide-MHC predictions, incorporated with more refined metrics of predicting potential immunogenicity, to more effectively narrow down targets in cancer and infectious disease, as well as identify populations that may be susceptible to both based on their HLA type. Further, as the amount of genotyping data continues to grow exponentially, the amount of personalization in evaluating individual risk factors to disease and in the development of personalized therapies can only increase. It is easy to envision a future pandemic in which a vaccine can be rapidly developed and deployed strategically first to genotypically at-risk populations when HLA typing becomes more ubiquitous.

One of the key challenges in the computational design of therapeutics for both infectious disease and cancer is establishing the reliability of peptide-MHC prediction tools. The number of successful targets out of a theoretical pool of thousands or even millions of peptides is small and the translatability of these targets into efficacious therapeutics is smaller still. While improvements have been made to these tools for nearly a decade, establishing a gold standard pipeline to predict immunogenicity has been difficult, without efforts to determine why such methods fail, especially when facing new datasets outside the space of known BA or EL peptides.

SARS-CoV-2 has given us a unique opportunity to study an infectious disease with the largest number of patients and the most amount of associated genotyping data in history. With new variants of SARS-CoV-2 and the increasing threat of a future pandemic(s), the accurate identification of targets is of escalating importance. Developing vaccines and other therapeutics in a timeline manner can and will save millions of lives. Our ability to more quickly synthesize an mRNA vaccine, a vaccine that introduces mRNA that corresponds to a viral protein which will be produced inside the

cell and recognized by the immune system, will be widely applicable to both more current and future infectious diseases, and being able to narrow down potential targets to minimize the development of this therapy will be vital to continued wellbeing for all of us.

1.5.2 Contributions

In this dissertation, I address some of the numerous questions above: how can we use people's genotyping data combined with peptide-MHC predictions to identify populations that may be at higher risk of disease, how can we evaluate peptide-MHC predictors, and how can we identify current challenges that are currently unaddressed to improve the translatability of peptide-MHC predictors. I use data from RefSeq (129) and a number of peptide-MHC binding predictors to create new metrics that may be able to identify populations at risk of severe COVID-19 disease, combine these predictions with clinical and survey data to show the relationship between HLA, peptide binding predictors, and severity of disease, and finally, extend these predictions to a larger number of viruses as well as self-antigens and randomly generated peptides in order to gain insight in the widespread applicability of peptide-MHC prediction tools. I show a number of limitations of the peptide-MHC tools and how there is a lack of reliability across tools, even for a single peptide and single allele pair, and demonstrate the relationship between predicted binding and the physical properties of peptides. Lastly, I identify implications and future directions for this work, specifically describing areas to further investigate the relationship between binding and physical properties, as well as offer possible improvements and solutions to the problem of peptide-MHC binding affinity prediction.

Chapter 2: A peptide-MHC framework for evaluating susceptibility to infectious disease.

This work has been formatted for inclusion in this dissertation from the manuscript “Human leukocyte antigen susceptibility map for SARS-CoV-2” by Austin Nguyen, Julianne K. David, Sean K. Maden, Mary A. Wood, Benjamin R. Weeder, Abhinav Nellore, Reid F. Thompson published in the Journal of Virology (130). The author of this dissertation is the primary author of the manuscript.

2.1 Abstract

Genetic variability across the three major histocompatibility complex (MHC) class I genes (human leukocyte antigen [HLA] A, B, and C) may affect susceptibility to and severity of severe acute respiratory syndrome 2 (SARS-CoV-2), the virus responsible for coronavirus disease 2019 (COVID-19). We execute a comprehensive in silico analysis of viral peptide-MHC class I binding affinity across 145 HLA -A, -B, and -C genotypes for all SARS-CoV-2 peptides. We further explore the potential for cross-protective immunity conferred by prior exposure to four common human coronaviruses. The SARS-CoV-2 proteome is successfully sampled and presented by a diversity of HLA alleles. However, we found that HLA-B*46:01 had the fewest predicted binding peptides for SARS-CoV-2, suggesting individuals with this allele may be particularly vulnerable to COVID-19, as they were previously shown to be for SARS (36). Conversely, we found that HLA-B*15:03 showed the greatest capacity to present highly conserved SARS-CoV-2 peptides that are shared among common human coronaviruses, suggesting it could enable cross-protective T-cell based immunity. Finally, we report global distributions of HLA types with potential epidemiological ramifications in the setting of the current pandemic.

2.2 Importance

Individual genetic variation may help to explain different immune responses to a virus across a population. In particular, understanding how variation in HLA may affect the course of COVID-19 could help identify individuals at higher risk from the disease. HLA typing can be fast and inexpensive. Pairing HLA typing with COVID-19 testing where feasible could improve assessment of viral severity in the population. Following the development of a vaccine against SARS-CoV-2, the virus that causes COVID-19, individuals with high-risk HLA types could be prioritized for vaccination.

2.3 Introduction

Recently, a new strain of betacoronavirus (severe acute respiratory syndrome coronavirus 2, or SARS-CoV-2) has emerged as a global pathogen, prompting the World Health Organization in January 2020 to declare an international public health emergency (131) .

In the large coronavirus family, comprising enveloped positive-strand RNA viruses, SARS-CoV-2 is the seventh encountered strain that causes respiratory disease in humans (132) ranging from mild -- the common cold -- to severe -- the zoonotic Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS-CoV). As of April 2020, there are over one million presumed or confirmed cases of coronavirus disease 19 (COVID-19) worldwide, with total deaths exceeding 50,000 (133).

While age and many comorbidities, including cardiovascular and pulmonary disease, appear to increase the severity and mortality of COVID-19 (134–139), approximately 80% of infected individuals have mild symptoms (140). As with SARS-CoV (141,142)and MERS-CoV (143,144), children seem to have low susceptibility to the disease (145–147); despite similar infection rates as adults only 5.9% of

pediatric cases are severe or critical, possibly due to lower binding ability of the ACE2 receptor in children or generally higher levels of antiviral antibodies (148). Other similarities (149–151) including genomic (152,153) and immune system response (154–162) between SARS-CoV-2 and other coronaviruses (32), especially SARS-CoV and MERS-CoV, are topics of ongoing active research, results of which may inform an understanding of the severity of infection (163) and improve the ongoing work of immune landscape profiling (164–167) and vaccine discovery (157,165,168–175) (29, 38, 42–49).

Human leukocyte antigen (HLA) alleles, which are critical components of the viral antigen presentation pathway, have been shown in previous studies to confer differential viral susceptibility and severity of disease. For instance, disease caused by the closely-related SARS-CoV shows increased severity among individuals with the HLA-B*46:01 genotype (1). Associations between HLA genotype and disease severity extend broadly to several other unrelated viruses. For example, in human immunodeficiency virus 1 (HIV-1), certain HLA types (e.g. HLA-A*02:05) may reduce risk of seroconversion (35), and in dengue virus, certain HLA alleles (e.g. HLA-A*02:07, HLA-B*51) are associated with increased secondary disease severity among ethnic Thais (34).

While a detailed clinical picture of the COVID-19 pandemic continues to emerge, there remain substantial unanswered questions regarding the role of individual genetic variability in the immune response against SARS-CoV-2. We hypothesize that individual HLA genotypes may differentially induce the T-cell mediated antiviral response, and could potentially alter the course of disease and its transmission. In this study, we performed a comprehensive in silico analysis of viral peptide-MHC class I binding affinity, across 145 different HLA types, for the entire SARS-CoV-2 proteome.

2.4 Results

2.4.1 SARS-CoV-2 presentation is similar to SARS-CoV presentation

To explore the potential for a given HLA allele to produce an antiviral response, we assessed the HLA binding affinity of all possible 8- to 12-mers from the SARS-CoV-2 proteome (n=48,395 peptides). We then removed from further consideration 16,138 peptides that were not predicted to enter the MHC class I antigen processing pathway via proteasomal cleavage. For the remaining 32,257 peptides, we repeated binding affinity predictions for a total of 145 different HLA types, and we show here the SARS-CoV-2-specific distribution of per-allele proteome presentation (predicted binding affinity threshold <500nM, Figure 2.1, Supplementary Table 2.1). Importantly, we note that the putative capacity for SARS-CoV-2 antigen presentation is unrelated to the HLA allelic frequency in the population (Figure 2.1). We identify HLA-B*46:01 as the HLA allele with the fewest predicted binding peptides for SARS-CoV-2. We performed the same analyses for the closely-related SARS-CoV proteome (Supplementary Figure 2.1) and similarly note that HLA-B*46:01 was predicted to present the fewest SARS-CoV peptides, keeping with previous clinical data associating this allele with severe disease (36).

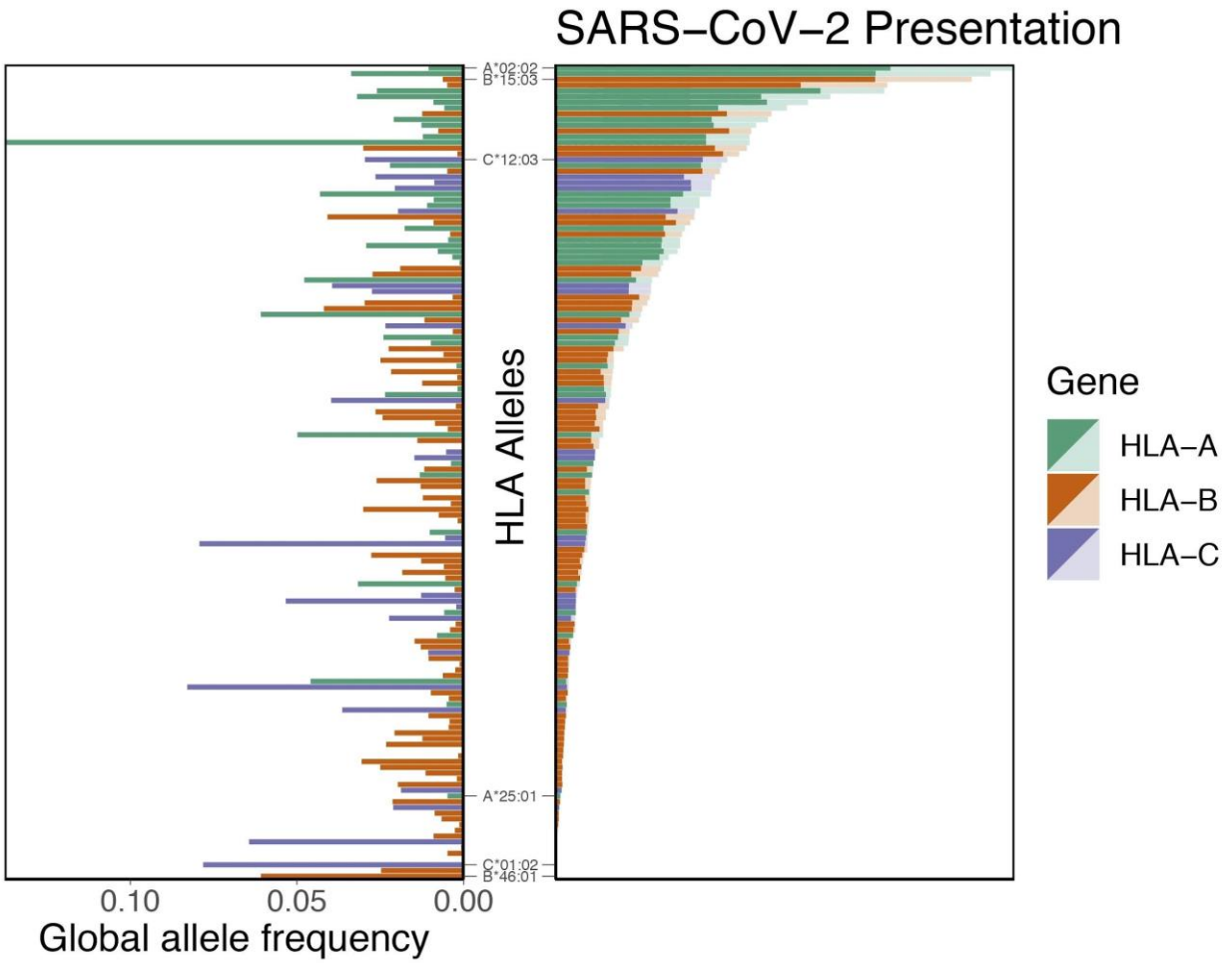


Figure 2.1: Distribution of HLA allelic presentation of 8- to 12-mers from the SARS-CoV-2 proteome. At right, the number of peptides (see Supplementary Table S1) that putatively bind to each of 145 HLA alleles is shown as a series of horizontal bars, with dark and light shading indicating the number of tightly (<50nM) and loosely (<500nM) binding peptides respectively, and with green, orange, and purple colors representing HLA-A, -B, and -C alleles, respectively. Alleles are sorted in descending order based on the number of peptides they bind (<500nM). The corresponding estimated allelic frequency in the global population is also shown (to left), with length of horizontal bar indicating absolute frequency in the population.

2.4.2 Conserved peptides do not show preferential binding

To assess the potential for cross-protective immunity conferred by prior exposure to common human coronaviruses (i.e. HKU1, OC43, NL63, and 229E), we next sought to characterize the

conservation of the SARS-CoV-2 proteome across diverse coronavirus subgenera to identify highly conserved linear epitopes. After aligning reference proteome sequence data for 5 essential viral components (ORF1ab, S, E, M, and N proteins) across 34 distinct alpha- and betacoronaviruses, including all known human coronaviruses, we identified 48 highly conserved amino acid sequence spans (Appendix 1). Acknowledging the challenges inferring cross-protective immunity among closely related peptides, we confined attention exclusively to identical peptide matches. Among conserved sequences, 44 SARS-CoV-2 sequences would each be anticipated to generate at least one 8- to 12-mer linear peptide epitope also present within at least one other common human coronavirus (Supplementary Table 2.2, Figure 2.2). In total, 564 such 8- to 12-mer peptides were found to share 100% identity with corresponding OC43, HKU1, NL63, and 229E sequences (467, 460, 179, and 157 peptides, respectively) (Supplementary Table 2.3).

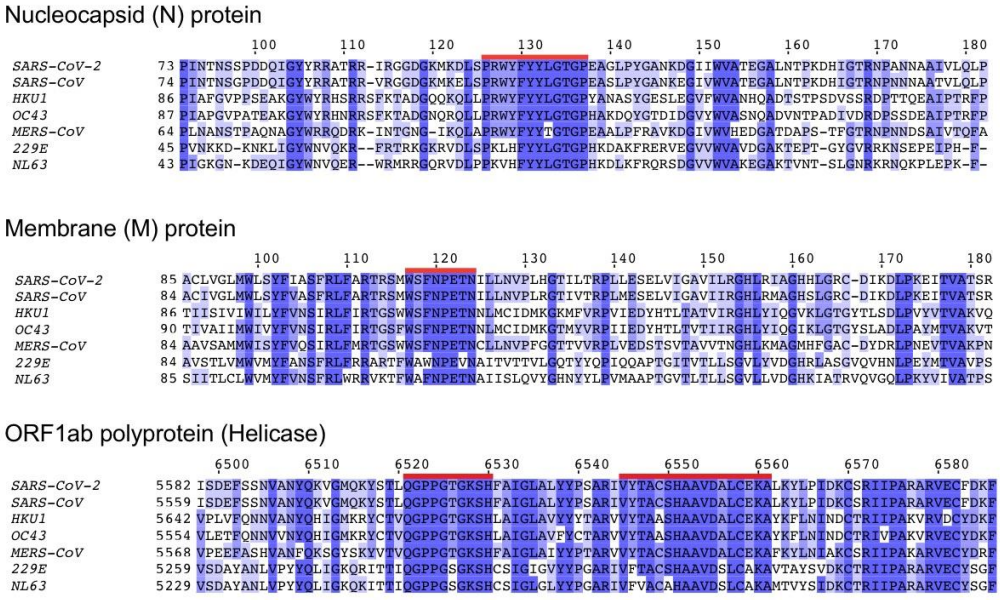


Figure 2.2: Amino acid sequence conservation of four linear peptide example sequences from three human coronavirus proteins. Protein sequence alignments are shown for nucleocapsid (N),

membrane (M), and ORF1ab polyprotein (Helicase) across all five known human betacoronaviruses (SARS-CoV-2, SARS-CoV, HKU1, OC43, and MERS-CoV) and two known human alphacoronaviruses (229E and NL63). Each row in the three depicted sequence alignments corresponds to the protein sequence from the indicated coronavirus, with starting coordinate of the viral protein sequence shown at left and position coordinates of the overall alignment displayed above. Blue shading indicates the extent of sequence identity, with the darkest blue shading indicating 100% match for that amino acid across all sequences. The four red highlighted sequences correspond to highly conserved peptides ≥8 amino acids in length (PRWYFYYLGTGP, WSFNPETN, QPPGTGKSH, VYTACSHAAVDALCEKA, see Supplementary Table 2.2).

For the subset of these potentially cross-protective peptides that are anticipated to be generated via the MHC class I antigen processing pathway, we performed binding affinity predictions across 145 different HLA-A, -B, and -C alleles. As above, we demonstrate the SARS-CoV-2-specific distribution of per-allele presentation for these conserved peptides. We found that alleles HLA-A*02:02, HLA-B*15:03, and HLA-C*12:03 were the top presenters of conserved peptides. Conversely, we note that 56 different HLA alleles demonstrated no appreciable binding affinity (<500nM) to any of the conserved SARS-CoV-2 peptides, suggesting a concomitant lack of potential for cross-protective immunity from other human coronaviruses. We note, in particular, HLA-B*46:01 is among these alleles. Note also that the putative capacity for conserved peptide presentation is unrelated to the HLA allelic frequency in the population (Figure 2.3). Moreover, we see no appreciable global correlation between conservation of the SARS-CoV-2 proteome and its predicted MHC binding affinity, suggesting a lack of selective pressure for or against the capacity to present coronavirus epitopes (p=0.27 [Fisher's exact test], Supplementary Figure 2.2).

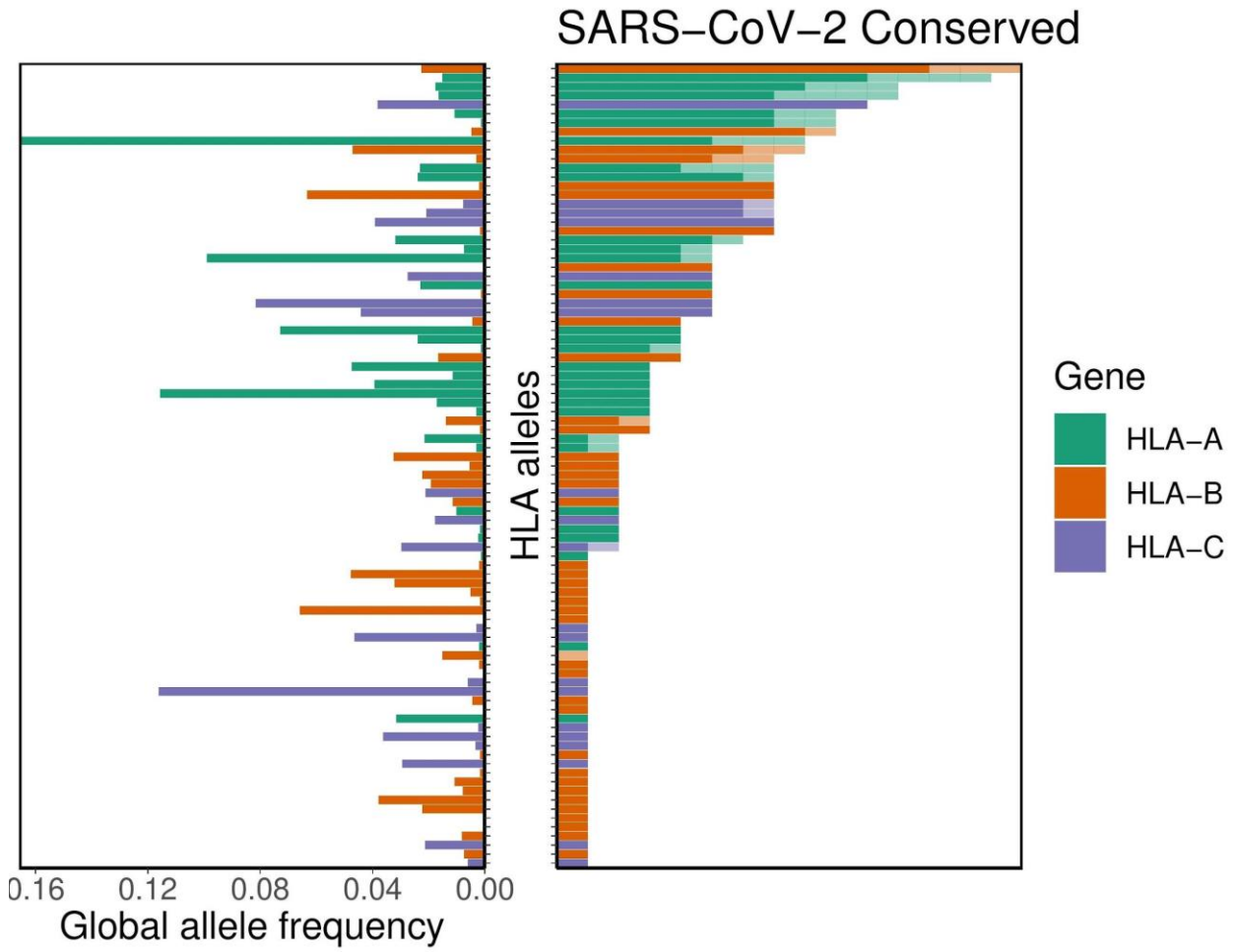


Figure 2.3: Distribution of HLA allelic presentation of highly conserved human coronavirus peptides with potential to elicit cross-protective immunity to COVID-19. At right, the number of conserved peptides (see Supplementary Table 2.3) that putatively bind to a subset of 89 HLA alleles is shown as a series of horizontal bars, with dark and light shading indicating the number of tightly (<50nM) and loosely (<500nM) binding peptides, respectively, and with green, orange, and purple colors representing HLA-A, -B, and -C alleles, respectively. Alleles are sorted in descending order based on the number of peptides they are anticipated to present (binding affinity <500nM). The corresponding allelic frequency in the global population is also shown (to left), with length of horizontal bar indicating absolute frequency in the population.

We were further interested in whether certain regions of the SARS-CoV-2 proteome showed differential presentation by the MHC class I pathway. Accordingly, we surveyed the distribution of

antigen presentation capacity across the entire proteome, highlighting its most conserved peptide sequences (Figure 2.4). Throughout the entire proteome, HLA-A and HLA-C alleles exhibited the relative largest and smallest capacity to present SARS-CoV-2 antigens, respectively. However each of the three major class I genes exhibited a very similar pattern of peptide presentation across the proteome (Supplementary Figure 2.3). We additionally note that peptide presentation appears to be independent of estimated time of peptide production during viral life cycle, with indistinguishable levels of peptide presentation of both early and late SARS-CoV-2 peptides (Supplementary Figure 2.4).

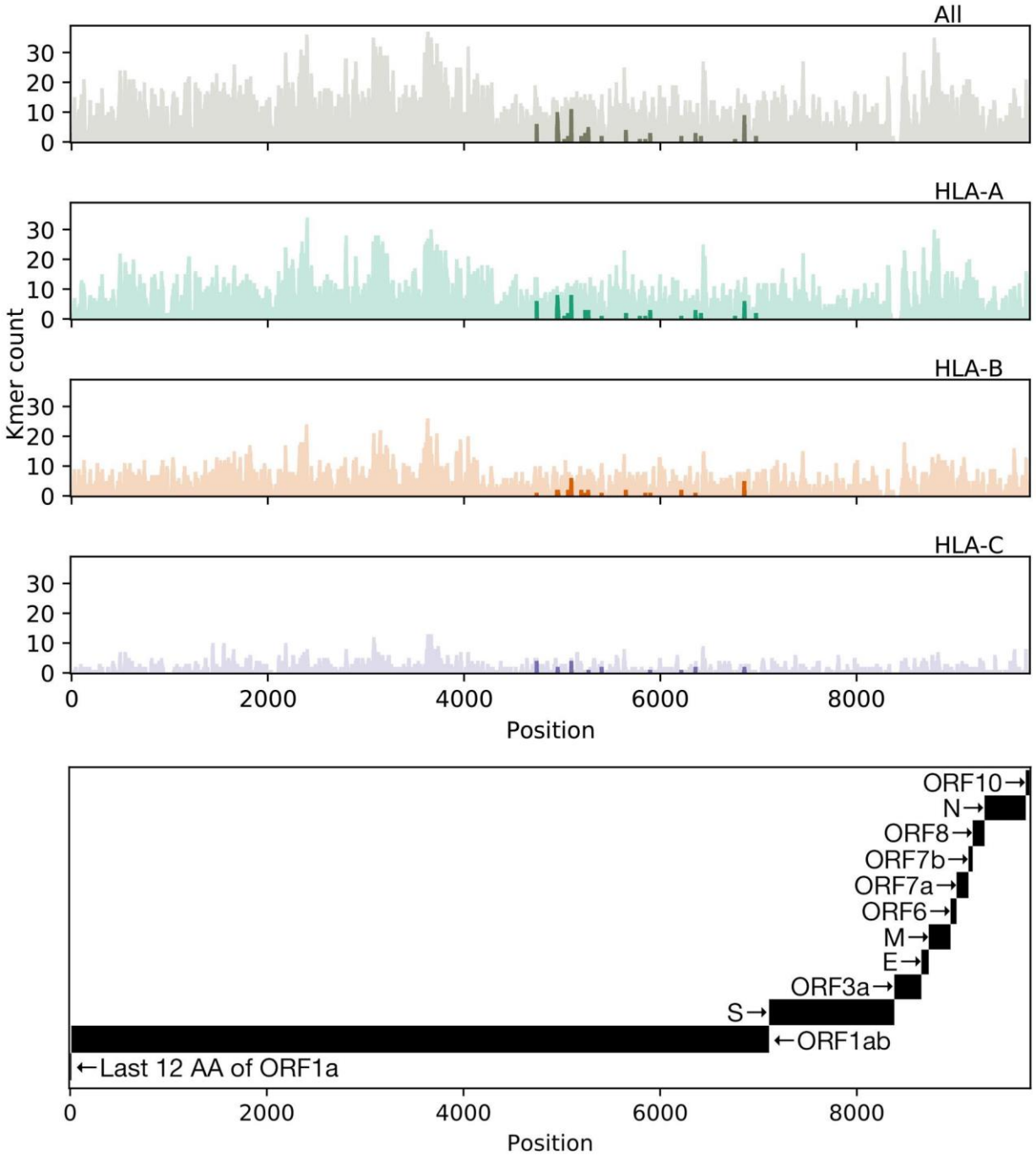
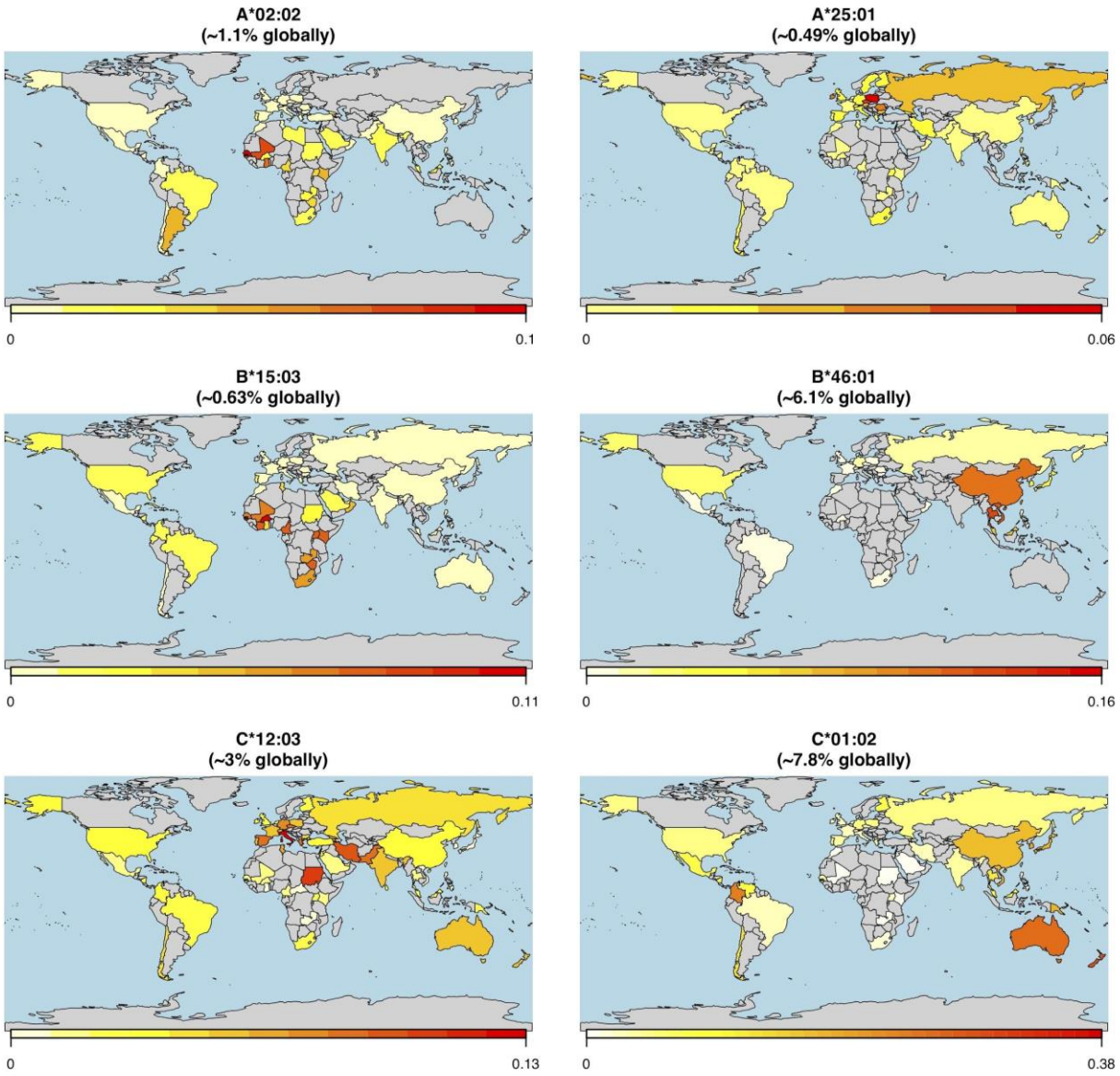


Figure 2.4: Distribution of allelic presentation of conserved 8- to 12-mers across the entire SARS-CoV-2 proteome for all HLA alleles and individually for HLA-A, HLA-B, and HLA-C (first, second, third, and fourth plots from top, respectively) with dark and light shading indicating the number of tightly (<50nM) and loosely (<500nM) binding peptides, respectively. Positions are derived from a concatenation of coding sequences (CDSes) indicated in the bottom panel. Tightly binding peptides

are confined to ORF1ab. The sequence begins with only the last 12 amino acids of ORF1a because all but the last four amino acids of ORF1a are contained in ORF1ab, and we considered binding peptides up to 12 amino acids in length.

Given the global nature of the current COVID-19 pandemic, we sought to describe population-level distributions of the HLA alleles best (and least) able to generate a repertoire of SARS-CoV-2 epitopes in support of a T-cell based immune response. While we present global maps of individual HLA allele frequencies for the full set of 145 different alleles studied herein (Appendix 2), we specifically highlight the global distributions of the three best-presenting (A*02:02, B*15:03, C*12:03) and three of the worst-presenting (A*25:01, B*46:01, C*01:02) HLA-A, -B, and -C alleles (Figure 5). Note that all allelic frequencies are aggregated by country, but implicitly reflect the distribution of HLA data available on the Allele Frequency Net Database (176).



.

*Figure 2.5: Global HLA allele frequency distribution heatmaps for six HLA-A, -B, and -C alleles. The leftmost panels show the global allele frequency distributions by country for three representative alleles (HLA-A*02:02, HLA-B*15:03, and HLA-C*12:03) with the predicted capacities to present the greatest repertoire of epitopes from the SARS-CoV-2 proteome (21.1%, 19.1%, and 7.9% of presentable epitopes, respectively). Conversely, the rightmost panels show the global allele frequency distributions by country for three representative alleles (HLA-A*25:01, HLA-B*46:01, and HLA-C*01:02) with the least predicted epitope presentation from the SARS-CoV-2 proteome (0.2%, 0%, 0% of presentable epitopes, respectively). Heatmap color corresponds to the individual HLA*

allele frequency within each country ranging from least (white/yellow) to most (red) frequent as indicated in the legend below each map.

2.4.3 Individual haplotype presentation has significant variability

Finally, we acknowledge that nearly all individuals have two HLA-A/B/C haplotypes constituting as few as three but as many as six distinct alleles, potentially buffering against the lack of presentation from a single poorly-presenting allele. We sought to describe whether allele-specific variability in SARS-CoV-2 presentation extends to full HLA haplotypes and to whole individual HLA genotypes. For six representative alleles with the highest (HLA-A*02:02, HLA-B*15:03, and HLA-C*12:03) and lowest (HLA-A*25:01, HLA-B*46:01, and HLA-C*01:02) predicted capacity for SARS-CoV-2 epitope presentation, these differences remain significant at the haplotype level, albeit with wide variability in presentation among different haplotypes (Figure 2.6). Haplotype-level data for all 145 alleles is included in Supplementary Figure 2.5 and Appendix 2. We then identified 3,382 individuals with full HLA genotype data and noted wide variability in their capacity to present peptides from the SARS-CoV-2 proteome, albeit with a small minority of individuals at either extreme (Supplementary Figure 2.6).

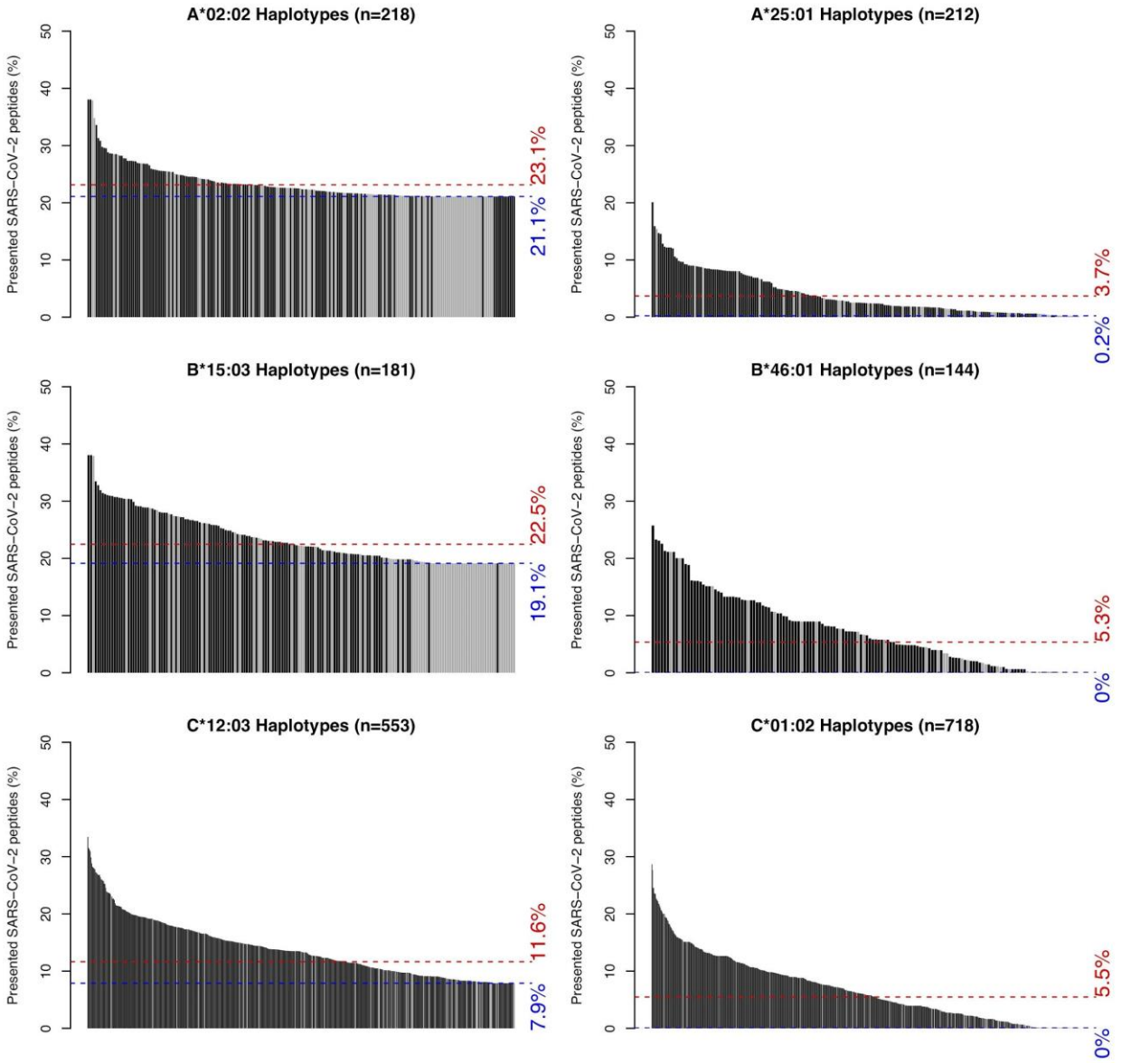


*Figure 2.6: Distributions of SARS-CoV-2 peptide presentation across HLA haplotypes. The leftmost panels show the distributions of SARS-CoV-2 peptide presentation capacity for haplotypes containing one of three representative HLA alleles (HLA-A*02:02, HLA-B*15:03, and HLA-C*12:03) with the greatest predicted repertoire of epitopes from the SARS-CoV-2 proteome. Conversely, the rightmost panels show the distributions of SARS-CoV-2 peptide presentation capacity for haplotypes containing one of three representative alleles (HLA-A*25:01, HLA-B*46:01, and HLA-C*01:02) with the least predicted epitope presentation from the SARS-CoV-2 proteome. Black and gray bars represent full or partial haplotypes, respectively. Blue and red Dashed lines represent the percent of presented SARS-*

CoV-2 peptides for the indicated allele itself (blue) and its global population frequency weighted average presentation across its observed haplotypes (red).

2.5 Discussion

To the best of our knowledge, this is the first study to evaluate per-allele viral proteome presentation across a wide range of HLA alleles using peptide-MHC binding affinity predictors. This study also introduces the relationship between coronavirus sequence conservation and MHC class I antigen presentation. We show that individual HLA, haplotype, and full genotype variability likely influence the capacity to respond to SARS-CoV-2 infection, and we note certain alleles in particular (e.g. HLA-B*46:01) that could be associated with more severe infection, as previously shown with SARS-CoV. Indeed, we further compare SARS-CoV and SARS-CoV-2 peptide presentation and note a high degree of similarity between the two across HLA types. Finally, this is the first study to report global distributions of HLA types and haplotypes with potential epidemiological ramifications in the setting of the current pandemic. We found that in general, there is no correlation between the HLA allelic frequency in the population and allelic capacity to bind SARS-CoV or SARS-CoV-2 peptides, irrespective of estimated timing of peptide production during the viral replication cycle. While we are not aware of any studies explicitly reporting the relationship between human coronavirus epitope abundance and immune response, there is data in vaccinia virus that suggests that early peptide antigens are more likely to generate CD8+ T-cell responses while antibody and CD4+ T-cell responses are more likely to target later mRNA expression with higher peptide abundance in the virion (177).

We note, however, several limitations to our work. First and foremost, while we note that a handful of our binding affinity predictions were borne out in experimentally validated SARS-CoV peptides (Supplementary Table 2.4), we acknowledge that this is an entirely in silico study. As we are

unable to obtain individual-level HLA typing and clinical outcomes data for any real-world COVID-19 populations at this time, the data presented is theoretical in nature, and subject to many of the same limitations implicit to the MHC binding affinity prediction tool(s) upon which it is based. As such, we are unable to assess the relative importance of HLA type compared to known disease-modifying risk factors such as age and clinical comorbidities. We further note that peptide-MHC binding affinity is limited as a predictor of subsequent T-cell responses (178–180), and we do not study T-cell responses herein. As such, we are ill-equipped to explore phenomena such as original antigenic sin (53,181–183), where prior exposure to closely related infection(s) may trigger T-cell anergy or immunopathogenesis in the setting of a novel infection (184–186). We explored only a limited set of 145 well-studied HLA alleles, but note that this analysis could be performed across a wider diversity of genotypes (49). Additionally, we did not assess genotypic heterogeneity or in vivo evolution of SARS-CoV-2, which could modify the repertoire of viral epitopes presented, or otherwise modulate virulence in an HLA-independent manner (187,188)(https://nextstrain.org/ncov). We also do not address the potential for individual-level genetic variation in other proteins (e.g. angiotensin converting enzyme 2 [ACE2] or transmembrane serine protease 2 [TMPRSS2], essential host proteins for SARS-CoV-2 priming and cell entry (189) to modulate the host-pathogen interface.

Unless and until the findings we present here are clinically validated, they should not be employed for any clinical purposes. However, we do at this juncture recommend integrating HLA testing into clinical trials and pairing HLA typing with COVID-19 testing where feasible to more rapidly develop and deploy predictor(s) of viral severity in the population, and potentially to tailor future vaccination strategies to genotypically at-risk populations. This approach may have additional implications for the management of a broad array of other viruses.

2.6 Materials and Methods

2.6.1 Sequence retrieval and alignments

Full polyprotein 1ab (ORF1ab), spike (S) protein, membrane (M) protein, envelope (E) protein, and nucleocapsid (N) protein sequences were obtained for each of 34 distinct but representative alpha and betacoronaviruses from a broad genus and subgenus distribution, including all known human coronaviruses (i.e. SARS-CoV, SARS-CoV-2, MERS-CoV, HKU1, OC43, NL63, and 229E). FASTA-formatted protein sequence data (full accession number list available in Supplementary Table 2.4,2.5) was retrieved from the National Center of Biotechnology Information (NCBI) (190). For each protein class (i.e. ORF1ab, S, M, E, N), all 34 coronavirus sequences were aligned using the Clustal Omega v1.2.4 multisequence aligner tool employing the following parameters: sequence type [Protein], output alignment format [clustal_num], dealign [false], mBed-like clustering guide-tree [true], mBed-like clustering iteration [true], number of combined iterations [0], maximum guide tree iterations [-1], and maximum HMM iterations [-1] (191). For the purposes of estimating time of viral peptide production, we classified ORF1a and ORF1b peptides as “early” while all other peptides produced by subgenomic mRNAs were classified as “late” (192,193).

2.6.2 Conserved peptide assessment

Aligned sequences were imported into Jalview v. 2.1.1 (194) with automated generation of the following alignment annotations: 1) sequence consensus, calculated as the percentage of the modal residue per column, 2) sequence conservation (0-11), measured as a numerical index reflecting conservation of amino acid physico-chemical properties in the alignment, 3) alignment quality (0-1), measured as a normalized sum of BLOSUM62 ratios for all residues at each position, 4) occupancy, calculated as the number of aligned residues (not including gaps) for each position. In all cases, sequence conservation was assessed for each of three groups: only human coronaviruses (n=7), all

betacoronaviruses (n=16), and combined alpha- and betacoronavirus sequences (n=34). Aligned SARS-CoV-2 sequence and all annotations were manually exported for subsequent analysis. Conserved human coronavirus peptides were defined as those with a length ≥8 consecutive amino acids, each with an agreement of SARS-CoV-2 and ≥4 other human coronavirus sequences with the consensus sequence (Supplementary Table 2.2). For each of these conserved peptides, we also assessed the component number of 8- to 12-mers sharing identical amino acid sequence between SARS-CoV-2 and each of the four other common human coronaviruses (i.e. OC43, HKU1, NL63, 229E) (Supplementary Table 2.3). For all peptides, human, beta, and combined conservation scores were obtained using a custom R v.3.6.2 script as the mean sequence conservation (minus gap penalties where relevant) (see https://github.com/pdxgx/covid19).

2.6.3 Peptide-MHC class I binding affinity predictions

FASTA-formatted input protein sequences from the entire SARS-CoV-2 and SARS-CoV proteomes were obtained from NCBI RefSeq database under accession numbers NC_045512.2 and NC_004718.3. We kmerized each of these sequences into 8- to 12-mers to assess MHC class I-peptide binding affinity across the entire proteome. MHC class I binding affinity predictions were performed using 145 different HLA alleles for which global allele frequency data was available as described previously (60) (see Supplementary Table 2.1, 2.5) with netMHCpan v4.0 (110) using the '-BA' option to include binding affinity predictions and the '-l' option to specify peptides of lengths 8-12 (Supplementary Table 2.1). Binding affinity was not predicted for peptides containing the character '|' in their sequences. Additional MHC class I binding affinity predictions were performed on all 66 MHCflurry supported alleles (--list-supported-alleles, Supplementary Table 2.6) using both MHCnuggets 2.3.2 (98) and MHCflurry 1.4.3 (195) (Supplementary Tables 2.7 and 2.8, Supplementary Figures 2.7-10). We further cross-referenced these lists of peptides with existing experimentally validated SARS-CoV epitopes present in the Immune Epitope Database (Supplementary Table 2.4) (111). We then performed consensus binding affinity predictions for the 66

supported alleles shared by all three tools by taking the union set of alleles and filtering for peptide-allele pairs matching the union set of alleles. For the SARS-CoV and SARS-CoV-2 specific distribution of per-allele proteome presentation, we exclude all peptides-allele pairs with >500nM predicted binding. In all cases, we used the netchop v3.0 (196) “C-term” model with a cleavage threshold of 0.1 to further remove any peptides that were not predicted to undergo canonical MHC class I antigen processing via proteasomal cleavage (of the peptide’s C-terminus).

2.6.4 Global HLA allele and haplotype frequencies

HLA-A, -B, and -C allele and haplotype frequency data were obtained from the Allele Frequency Net Database (176)for 805 distinct populations pertaining to 101 different countries and 2628 distinct major/minor (4-digit) alleles, corresponding to 20,478 distinct haplotypes (https://github.com/pdxgx/covid19). We also identified full HLA genotype data for 3,382 individuals whose HLA types were confined to the 145 HLA alleles studied herein. Population allele and haplotype frequency data were aggregated by country as a mean of all constituent population allele or haplotype frequencies weighted by sample size of the population, but not accounting for representative ethnic demographic size of the population. Global allele frequency maps were generated using the rworldmap v1.3-6 package (197), with total global allele and haplotype frequency estimates calculated as the mean of per-country allele and haplotype frequencies, weighted by each country’s population in 2005.

Chapter 3: The relationship between HLA genotype, peptide-MHC binding, and disease severity.

This work has been formatted for inclusion in this dissertation from the manuscript “Minimal observed impact of HLA genotype on hospitalization and severity of SARS-CoV-2 infection” by Austin Nguyen, Tasneem Yusufali, Jill A. Hollenbach, Abhinav Nellore, and Reid F. Thompson published in HLA (198). The author of this dissertation is the primary author of the manuscript.

3.1 Abstract

HLA is a critical component of the viral antigen presentation pathway. We investigated the relationship between severity of SARS-CoV-2 disease and HLA type in 3,235 individuals with confirmed SARS-CoV-2 infection. We found only the DPB1 locus to be associated with the binary outcome of whether an individual developed any COVID-19 symptoms, suggesting that HLA class II was able to initiate early immune response. The number of peptides predicted to bind to an HLA allele had no significant relationship with disease severity both when stratifying individuals by ancestry or age and in a pooled analysis. Age, BMI, asthma status, and autoimmune disorder status were predictive of severity across multiple age and individual ancestry stratifications. Overall, at the population level, we found HLA type is significantly less predictive of COVID-19 disease severity than certain demographic factors and clinical comorbidities.

3.2 Introduction

The global COVID-19 pandemic has exposed significant gaps in our ability to predict disease trajectory among individuals, with many people experiencing asymptomatic infections while others may be hospitalized with or die from COVID-19. Observational analyses have identified disease severity risk factors such as age, BMI, and sex (199–201). However, host immunogenetic factors such as human leukocyte antigen (HLA) type may help determine the severity of SARS-CoV-2 infection. HLA is a critical component of the viral antigen presentation pathway, and previous studies have shown that individual HLA alleles may confer differential susceptibility and severity across viral diseases, including SARS-CoV-2 (2,33,202–207).

While large genotype association studies have investigated the relationship between genetic variants and severity of COVID-19 disease, they have not generally implicated the HLA locus (206,208–210). Further, while a growing collection of single institution or regional hospital system-based studies have reported HLA associations with COVID-19 disease (211–214), the statistical significance of these associations does not withstand correction for multiple comparisons. The relationship between HLA genotype and severity of COVID-19 disease, especially across a large and diverse population, thus remains unclear.

In this study, we investigated the specific relationship between HLA type and COVID-19 severity in a cohort of 3,235 individuals obtained from AncestryDNA (206,215) with confirmed SARS-CoV-2 infection.

3.3 Results

We extracted basic demographic and clinical data for 3,235 individuals among the AncestryDNA cohort (206,215) with a positive SARS-CoV-2 nasal swab and classified the severity of

their COVID-19 disease according to patient survey responses (Table 1). We next assessed the extent to which these demographic and clinical features predicted COVID-19 severity, and we found comorbidities that are contributors in a linear model predicting hospitalization (Supplementary Table 3.1).

Table 1

	Severity score 1	**Severity score 2**	**Severity score 3**	**Severity score 4**	**Severity score 5**	**Severity score 6**
Count	607	965	1418	136	83	26
Male (%)	201 (18)	394 (36)	398 (36)	46 (4)	45 (4)	15 (1)
Female (%)	299 (14)	678 (32)	1020 (48)	90 (4)	38 (3)	11 (1)
EUR (%)	365 (16)	769 (33)	1008 (44)	92 (4)	61 (3)	21 (1)
AS (%)	11 (20)	19 (33)	23 (40)	3 (5)	0 (0)	1 (1)
AMR (%)	88 (13)	237 (34)	320 (46)	30 (4)	17 (2)	2 (0)
AFR (%)	36 (21)	47 (28)	67 (40)	11 (7)	5 (3)	2 (1)
Median Age	52 (19-88)	43.5 (19-85)	45 (19-85)	51 (21-86)	56 (20-84)	61.5 (29-

(range)						88)

Table 1: Demographic breakdown of the AncestryDNA cohort. Severity score 0 is left out since it corresponds to individuals who tested negative for SARS-CoV-2 infection. Severity scores 1-6 are described in Materials and Methods: Severity scoring and hospitalization.

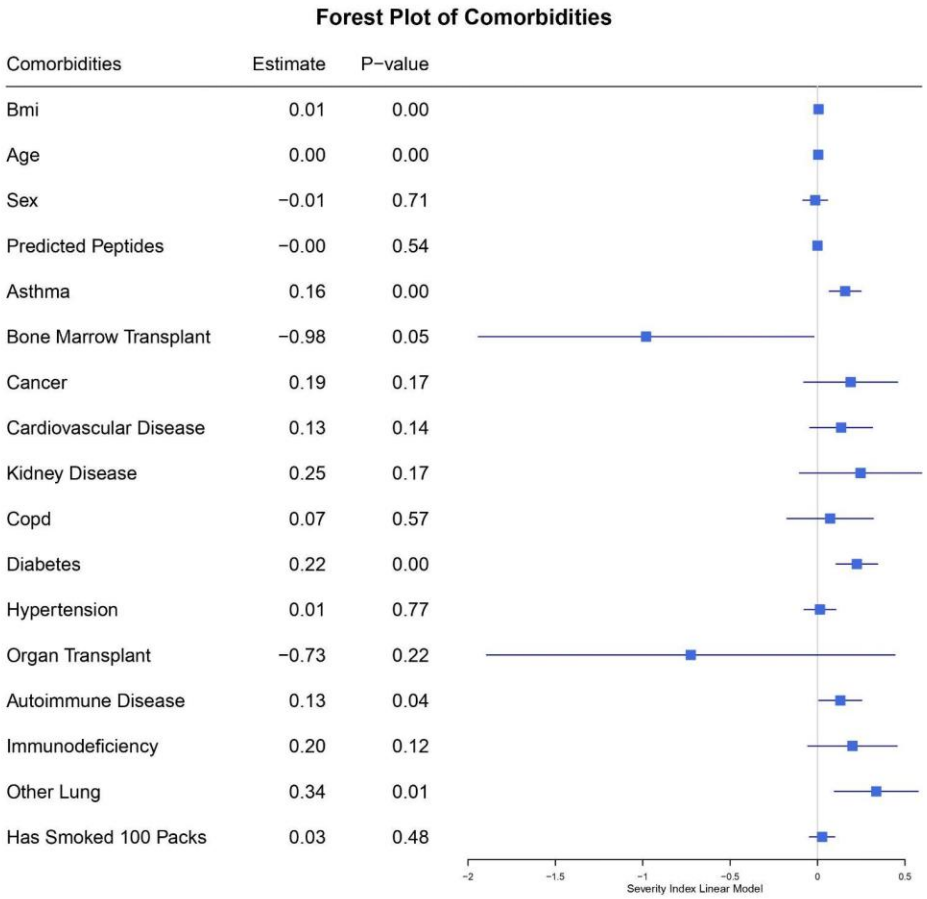


Figure 3.1: Forest plot of comorbidities including putative viral peptide presentation as a function of HLA type (Predicted Peptides) in the pooled multivariate model for predicting severity score. The estimate column shows the coefficients of variables in the linear model. Each line represents the 95% confidence interval for the estimate value. For the sex variable, female is 1 and male is 0. Positive

values for the estimate are predicted to contribute to a higher severity score and vice versa for negative values.

We next assessed the contribution of genetic variability across the HLA locus to hospitalization as a binary outcome. Using BIGDAWG for case-control association analysis (216), we found no individual HLA types or specific amino acid variants across the HLA locus that were associated with hospitalization (Supplementary File 3.1). Only the DPB1 locus (p=0.04) was found to be associated with the binary outcome of whether an individual developed any COVID-19 symptoms.

To assess an individual's capacity to present SARS-CoV-2 peptides, we computed HLA-specific MHC binding affinities of all k-mers of sizes between 8 and 12 inclusive from the SARS-CoV-2 proteome (n=48,395 unique peptides) passing a proteasomal cleavage propensity filter. We used two different predictive tools: netMHCpan and HLAthena (93,96). In agreement with our prior work (207), we find a wide variety in putative peptide presentation capacity across different HLA types (Supplementary Figure 3.1).

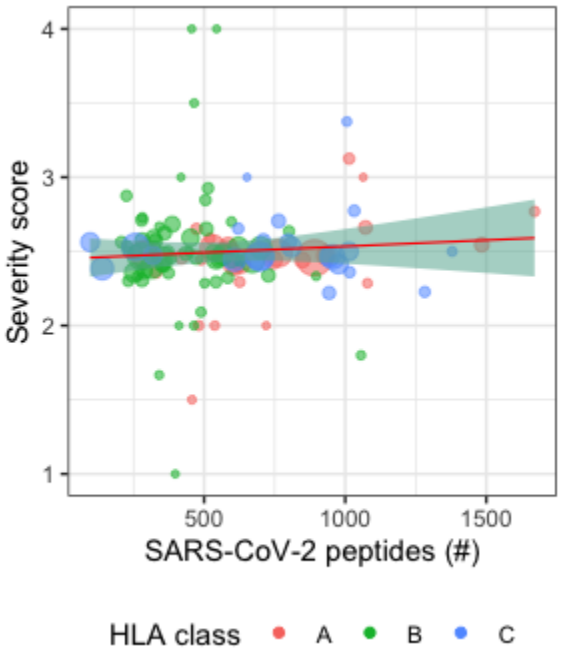


Figure 3.2: Scatter plot of HLA alleles with the number of predicted peptides vs. average severity score in the AncestryDNA dataset. Each data point represents a distinct HLA allele, with larger points representing larger numbers of individuals in the AncestryDNA dataset imputed to have the allele and the red, green, and blue colors representing HLA-A, HLA-B, and HLA-C respectively.

We next developed a pooled multivariate model of severity score, accounting for comorbidities as well as putative viral peptide presentation as a function of HLA type, and we found that age ($p < 0.01$), BMI ($p < 0.01$), asthma status ($p < 0.01$), diabetes status ($p < 0.01$), and other lung conditions ($p < 0.05$) were all predictive (Figure 3.1; Supplementary Table 1-Stratified_LM_models_corrected.csv). There was no association between the number of putatively presented class I peptides and COVID-19 severity (Figure 3.2). The significance of the association between BMI and severity of disease was diminished for age >60 years. This is consistent with CDC reports of obesity as a risk factor in hospitalization and death, specifically among individuals <65 years (201).

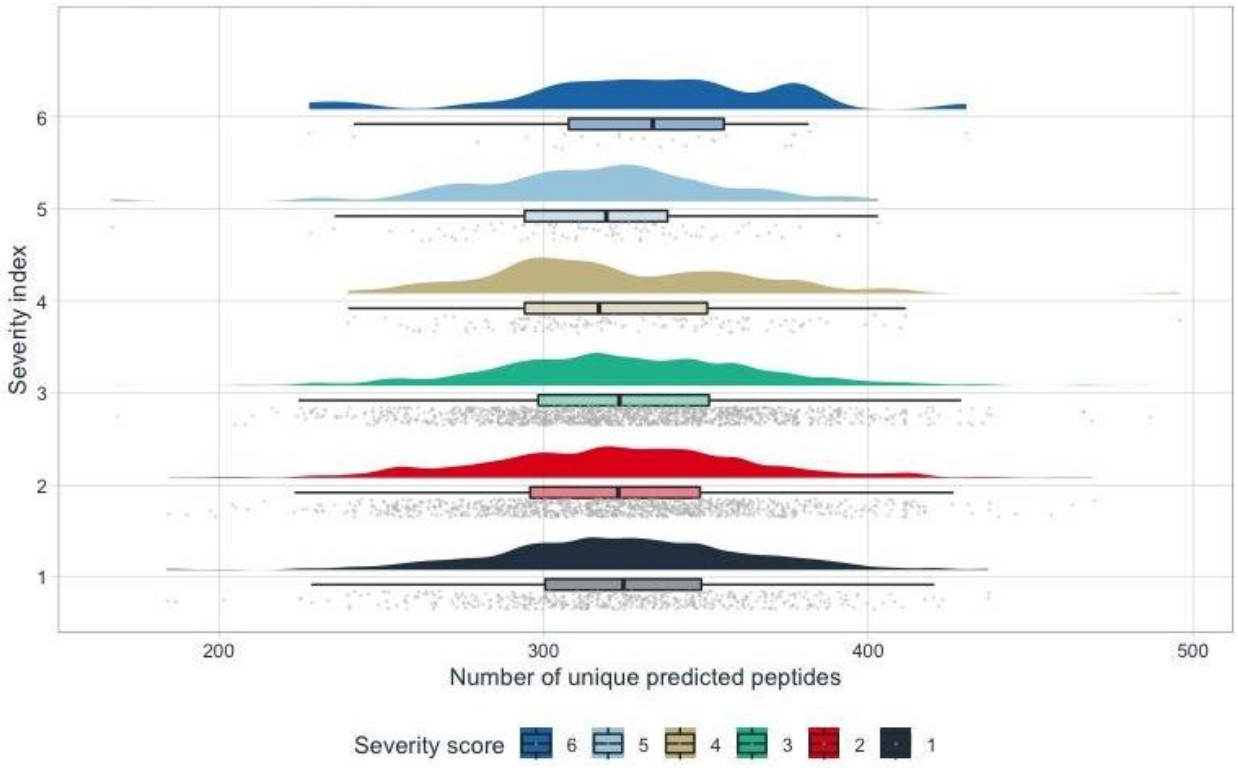


Figure 3.3: Distribution of unique predicted peptides vs. severity score. The half-violin plots represent the distribution of unique predicted peptides of the 3,235 *AncestryDNA individuals who had the corresponding severity score. The boxplot shows the IQR of the unique predicted peptides and each point in the rain cloud below the boxplot represents the number of predicted peptides of each individual.*

Predicted number of presented viral peptides demonstrated no significant relationship with disease severity when stratifying by individual ancestry., but the significance of various comorbidities was affected. Among individuals with EUR or AFR ancestry (2316 or 168 individuals, respectively) no clinical features were associated with disease severity score. , while BMI ($p<0.05$), age ($p<0.05$), and hypertension ($p<0.001$) were all predictive of disease among individuals of AS ancestry (57 individuals), and BMI ($p<0.05$) was predictive among individuals of AMR ancestry (694 individuals).

3.4 Discussion

HLA genes are generally considered important for host response to novel infectious diseases. In this study, we found that age, BMI, and other comorbidities determined clinical outcome across 3,235 individuals as described in the literature (199–201,209), and to a far greater degree than an individual's HLA-specific capacity to present SARS-CoV-2-specific peptides. Only in healthy, young (30-50) individuals did we see any type of association between HLA/peptide predictions with severity outcomes. While we previously explored the potential of HLA-peptide binding to predict COVID-19 severity (207), we do not see evidence for this phenomenon in the large real-world clinical cohort explored here. While the majority of the individuals were imputed to be of European ancestry, there were sizable numbers of individuals of Amerindian, Asian, and African descent. While Roberts et al. (206) performed a stratified GWAS analysis using this same dataset, with binary endpoints of hospitalization and whether an individual developed any COVID-19 symptoms, they did not specifically explore the role of HLA, which has a high level of variability that reduces power to detect differences in populations. Further, we investigated SARS-CoV-2 specific peptide presentation as a nonlinear function of HLA type, where some HLA types may be more similar to each other in the number of predicted peptides they can bind than they may be in canonical HLA supergroups.

We note several limitations to our work. Firstly, the proportion of SARS-CoV-2 peptides that we tested were generated through whole-peptidome *in silico* analysis of SARS-CoV-2. This may not be representative of the actual SARS-CoV-2 peptides presented in a given individual, whether due to biological sources such as viral variation, or methodological sources such as potential inaccuracies in peptide-MHC binding affinity predictions. Secondly, individuals who suffered debilitating infections may have been less likely to participate in the survey, and no individuals who died of COVID-19 were able to participate in the study, potentially resulting in an undercounting of the most severe phenotypes. Further, the cohort was primarily European, with much smaller sample sizes for African,

Asian, and Amerindian ancestry. Lastly, these data were composed entirely of the unvaccinated cohort, as this population was tested and surveyed before the release of the many SARS-CoV-2 vaccines.

A number of other studies (211–214,217–219) have examined the relationship between HLA alleles and COVID-19 severity, and few have found alleles significantly associated with severity. In the majority of these studies, the large number of possible alleles in each study reduced the statistical power to identify significant alleles after multiple testing correction. Further, a number of studies reporting statistical significant associations between severity and HLA type were regional; they tended to have more ethnically and geographically homogeneous cohorts, likely resulting in overrepresentation of some alleles. Taken together with our analysis of the AncestryDNA dataset, we suggest that the literature does not reliably support the role of HLA type in modifying real-world COVID-19 disease severity across a population. There are multiple potential explanations for this, including that the data and analyses to date do not accurately reflect the true potential disease-modifying effects of HLA genes. On an individual basis, HLA type may indeed influence the severity of COVID-19 disease; however, this hypothesis is not readily borne out at a population level, at least in this cohort. Multiple demographic features and clinical comorbidities are significantly more predictive of disease severity in a population. While we acknowledge the potential for other studies such as the COVID-19 Host Genetics Initiative to uncover a disease modifying role of HLA, future work should take a very critical and individualized approach towards evaluating any connections between HLA variation and differences in COVID-19 disease severity.

3.5 Materials and Methods

3.5.1 Genotyping

Data from 15000 individuals belonging to the Ancestry COVID-19 study was obtained through an IRB-approved project (Ancestry Human Diversity Project). The data authorized for reuse by AncestryDNA included: AncestryDNA genome-wide scale genotypes, AncestryDNA research participants' self-reported age, gender, height, weight, and smoking status, and survey answers for the AncestryDNA COVID-19 questionnaire. As reported by the AncestryDNA COVID-19 study, the study participants' genotype data were obtained using an Illumina genotyping array (Illumina OmniExpress platform) composed of 730,525 SNPs. Genotyping array data was processed by Illumina or Quest/Athena Diagnostics (206)**.**

3.5.2 Ancestry imputation

Genetic ancestry was determined using plinkQC v1.9 (220) to combine genotypes of the cohort with genotypes of a reference dataset (14) consisting of individuals of known ethnicities. Principal component analysis (PCA) on the combined genotype panel was used to detect population structure of the reference dataset to the level of continental ancestry. A→T and C→G SNPs were removed from study and reference data as they are more difficult to align and only a subset of SNPs were required for the analysis. The study data were pruned for variants in linkage disequilibrium (LD) with an $r^2 > 0.2$ in a 50kb window, and that list of pruned variants was used to reduce the size of the reference dataset. Checks were performed to ensure matching variant IDs and chromosomal positions between the study and reference dataset before merging and running PCA. Ancestries for the study population were then imputed from the principal components provided for the labeled reference dataset.

3.5.3 HLA class I + II imputation

HLA Class I/II alleles were obtained using HIBAG v1.3 (221), a prediction method for HLA imputation that utilizes large training sets with known HLA and SNP genotypes in combination with attribute bagging. Ancestry-specific pre-fit models available within HIBAG for European, Asian, African, and Amerindian populations were applied to the subgroups of distinct ancestries within the AncestryDNA cohort using 1,042 SNPs across and nearby the HLA locus.

3.5.4 Severity scoring and hospitalization

We collapsed the 10 point WHO COVID-19 Ordinal Scale of disease severity (222) into a 7-point scale to accommodate available phenotype information in the AncestryDNA COVID-19 study.

AncestryDNA Survey State	**Severity Score**	**WHO Patient State**	**WHO Score**
Uninfected	**0**	Uninfected	**0**
Asymptomatic	**1**	Asymptomatic	**1**
Symptomatic, mild symptoms	**2**	Symptomatic, no assistance needed	**2**
Symptomatic, severe symptoms	**3**	Symptomatic, assistance needed	**3**
Hospitalized, no	**4**	Hospitalized, no oxygen	**4**

oxygen			
Hospitalized, oxygen	**5**	Hospitalized, oxygen by mask/nasal prongs	**5**
		Hospitalized, oxygen by NIV/high flow	**6**
Hospitalized, ventilator	**6**	Intubation and mechanical ventilation	**7**
		Mechanical ventilation and vasopressors	**8**
		Mechanical ventilation, vasopressors, dialysis, or ECMO	**9**

The possible symptoms in the AncestryDNA cohort are fever, shortness of breath, dry cough, body aches, abdominal pain, cough producing phlegm, and nausea. There are 3 levels of severity to each symptom: normal, severe, and very severe. We defined severe symptoms (Severity Score 3) as any of the listed symptoms at the severe or very severe level. In models where we used hospitalization as an endpoint, we added hospitalization as a binary variable, with scores >=4 considered hospitalized.

Note that COVID-19 survey response data were consistent with CDC case and hospitalization rates over similar time periods, with 14% v. 12% test positivity and 11% v. 14% hospitalization rate in the AncestryDNA and CDC datasets, respectively (206,223).

3.5.5 HLA-peptide predicted binding

We obtained SARS-CoV-2 peptide sequences by k-merizing FASTA protein sequences obtained from the NCBI RefSeq database (NC_045512.2 and NC_004718.3) into 8- to 12-mers. These k-mers were filtered by NetChop v3.1 using default settings with a cutoff of 0.1. MHC class I binding affinity predictions were performed using netMHCpan v4.0 using the '-BA' option to include binding affinity predictions and the '-l' option to specify peptides 8 to 12 amino acids in length. Additional MHC class I binding affinity predictions were performed using HLAthena. For predicted peptide binding, we used the cutoff of <500nM for peptides predicted by netMHCpan v4.0 and the cutoff of >0.5 probability score for peptides predicted by HLAthena. While nearly all individuals have two HLA-A/B/C haplotypes constituting as few as three but as many as six distinct alleles, a single peptide may be predicted to bind to more than one of an individual's HLA alleles. While there is no definitive evidence that a peptide is more likely to be presented when predicted to bind to more than one allele, we wanted to capture this possibility by using 2 metrics: an overall predicted peptide value and a unique predicted peptide value. For each individual, to calculate capacity to bind SARS-CoV-2 peptides, we summed the number of predicted peptides bound to each individual's allele (min 3, max 6). For a unique-peptide specific capacity, the peptides were filtered to remove duplicates after summation.

3.5.6 Statistical analyses

We performed statistical tests for HLA vs. hospitalization using the Bridging ImmunoGenomic Data-Analysis Workflow Gaps (BIGDAWG) pipeline and a comprehensive SARS-CoV-2 peptide-genotype binding analysis for all individuals in our dataset. All statistical analyses were performed

using R version 4.0.3. For each statistical test, we performed pooled and ancestry-stratified testing. For multivariate linear modeling, we used the R function lm for multivariate regression with one of severity index, hospitalization status, or asymptomatic/symptomatic as the endpoint. Tests of Hardy-Weinberg equilibrium using Chi-squared testing for haplotypes, loci, and HLA-amino acid positions were performed using the BIGDAWG v1.3.4 R package. Note that all reported p-values have been corrected for multiple hypothesis testing, where relevant, using Benjamini-Hochberg correction.

Chapter 4: Discordant results among MHC binding affinity prediction tools.

This work has been formatted for inclusion in this dissertation from the manuscript “Discordant results among MHC binding affinity prediction tools.” by Austin Nguyen, Abhinav Nellore, and Reid F. Thompson submitted to Nucleic Acids Research. The author of this dissertation is the primary author of the manuscript.

4.1 Abstract

A large number of machine learning-based Major Histocompatibility Complex (MHC) binding affinity (BA) prediction tools have been developed and are widely used for both investigational and therapeutic applications, so it is important to explore differences in tool outputs. We examined predictions of four popular tools (netMHCpan, HLAthena, MHCflurry, and MHCnuggets) across a range of possible peptide sources (human, viral, and randomly generated) and MHC class I alleles. We uncovered inconsistencies in predictions of BA, allele promiscuity, the relationship between physical properties of peptides by source and BA predictions, as well as quality of training data. Our work raises fundamental questions about the fidelity of peptide-MHC binding prediction tools and their real-world implications.

4.2 Introduction

Human Leukocyte Antigen (HLA) alleles are critical components of the immune system’s ability to recognize and eliminate tumors and infections (224). Infectious diseases in particular are thought to be a major source of selective pressure on the Major Histocompatibility Complex (MHC) region

which encodes HLA alleles and is one of the most diverse regions of the human genome (33,202,225–229). There is large diversity in the antigenic peptide sequences which individual HLA alleles can recognize and ultimately present to the adaptive immune system (10), with a positive correlation between increased sequence diversity recognition and fitness (40).

Tools that can predict the extent to which a given HLA allele may have an affinity for a given peptide have critical implications for our ability to understand and translationally leverage antigen-specific immune response pathways. For instance, MHC binding affinity predictors have been – or otherwise have the potential to be – used to evaluate an individual or population's susceptibility to viral infection (130), to develop an understanding of specific autoimmune conditions (123), to improve transplantation technologies (122), or even to assist in the development of personalized cancer vaccines (73,230–233). Numerous peptide-MHC binding prediction tools exist, and are key components in broader antigen prediction methodologies (71,74,112,113).

The most widely adopted MHC binding prediction tools rely on neural network models trained on binding affinity (BA) and/or eluted ligand (EL) data. The most commonly cited tool, netMHCpan (96,110), uses both BA and EL data in a neural network architecture with a single hidden layer to predict allele-specific binding affinities. MHCflurry (97) attempts to improve upon netMHCpan by increasing the number of hidden layers and augmenting BA and EL training data with unobserved decoys. MHCnuggets (234) again trains on BA and EL data but uses a different architecture, with a long short-term memory layer and a fully connected layer to improve its predictions further across different peptide lengths. Lastly, HLAthena (93), while most similar in architecture to netMHCpan, relies on independently generated EL data from mono-allelic cell lines for training.

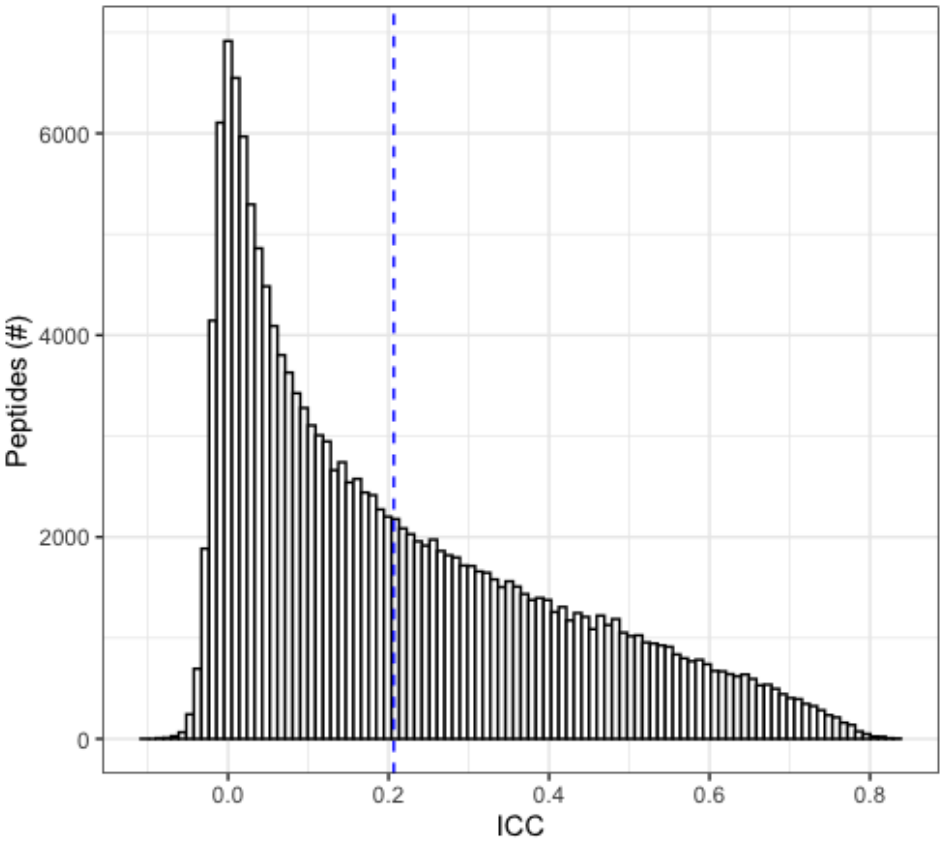
We sought to better characterize the outputs of these tools over a large and diverse set of peptides, across different tools and HLA alleles, as well as quantify the stability of these predictions. We also sought to measure allelic binding preferences and whether they may enrich for foreign v. self

peptides. In this study, we performed a comprehensive *in silico* analysis of peptides from multiple viral proteomes, the human proteome, and randomly generated peptides across HLA class I alleles.

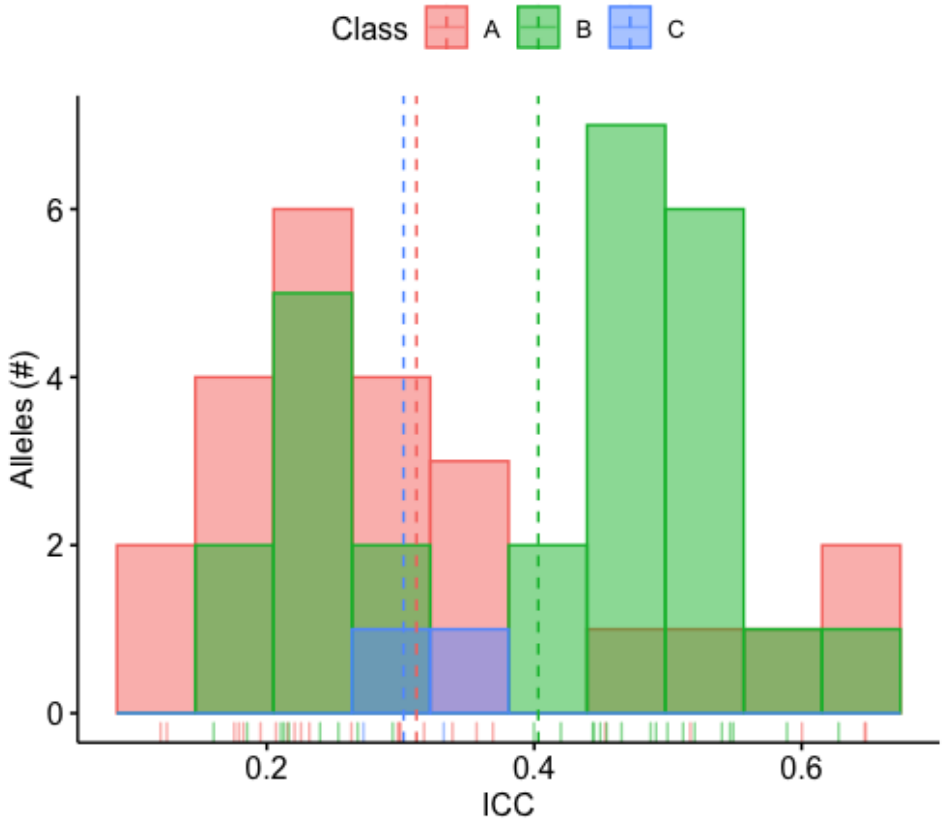
4.3 Results

4.3.1 Peptide predictions are inconsistent across tools

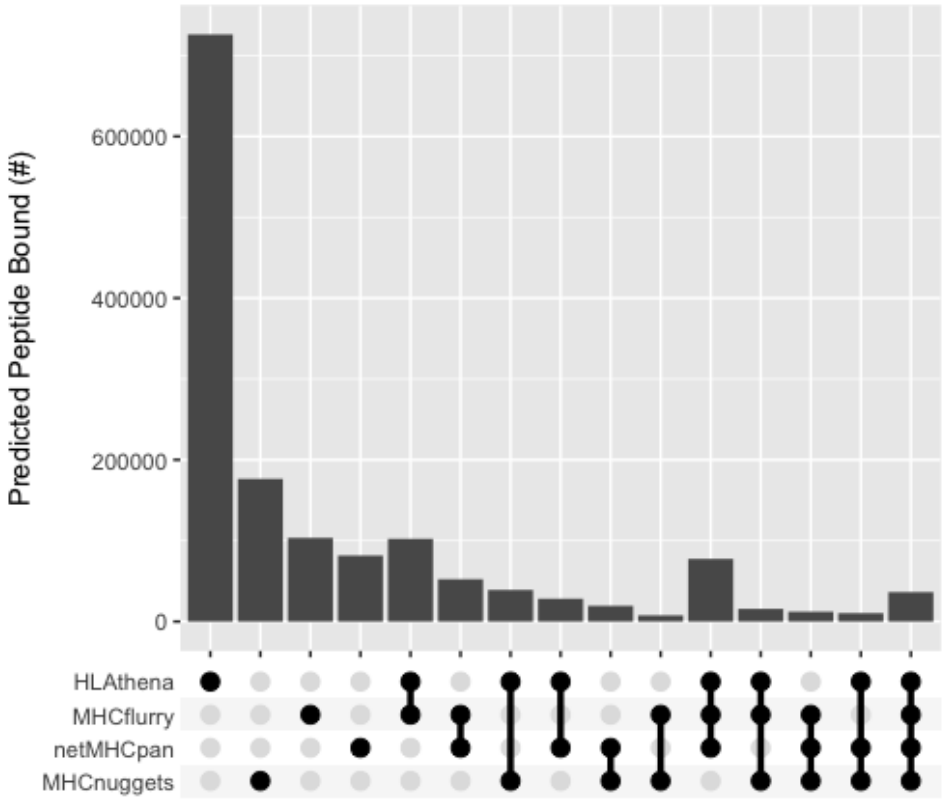
We first assessed the consistency of peptide-specific MHC I binding affinity predictions across four tools (MHCnuggets, MHCflurry, HLAthena, netMHCpan) and 52 different HLA alleles. We found substantial disagreement in peptide-specific predictions between each tool, independent of allele (Figure 4.1A), with median intraclass correlation coefficient (ICC) of 0.207 and only 0.48% of peptides having ICC > 0.75. On a per-allele basis, we found a wide range in consistency of predictions across tools, with a mean intraclass correlation as low as 0.12 for A02:07 and as high as 0.64 for A23:01 (Figure 4.1B). Among all of the peptides predicted by at least one tool to bind to at least one allele, only 7.9% were consistently predicted across all tools to bind to the same allele (Figure 4.1C).


Peptides (#)
6000
4000
2000
0
0.0
0.2
0.4
0.6
0.8
ICC

A


Class
A
B
C
Alleles (#)
6
4
2
0
0.2
0.4
0.6
ICC

B



C

Figure 4.1. Inconsistency of peptide predictions across tools. A) Histogram of intraclass correlation coefficients (ICC) calculated for a set of 1 million random peptides across four tools (MHCnuggets, MHCflurry, HLAthena, netMHCpan), with ICC calculated as the overall correlation among tools across 52 HLA alleles. The dotted vertical line indicates the median ICC value (0.207) across all peptides. B) Histogram of ICCs for 52 HLA alleles between four tools (MHCnuggets, MHCflurry, HLAthena, netMHCpan). The number of alleles is shown on the y-axis and the ICC is shown on the x-axis. The dotted lines show the mean ICC for alleles belonging to each HLA class. Red, green, and blue colors represent data from -A, -B, and -C alleles, respectively. C) Detailed comparison of the complete set of random peptides predicted to bind (binding score >=0.5) to HLA alleles according to each of four tools. Patterns of agreement or disagreement among groups of peptides predicted by different combinations of tools across 1 million random peptides are shown along each column (e.g. the first column corresponds to peptides predicted by HLAthena while the final column corresponds to peptides predicted by all tools). Each row indicates the predictions associated with the indicated tool.

The number of peptides in each column (vertical bars) corresponds to the size of the subset predicted by the indicated combination of tools.

We next investigated aggregate peptide binding predictions across different HLA alleles according to each tool. As others have noted differential HLA allelic promiscuity in peptide presentation (11,62,235,236), we too found a wide range in the proportion of peptides a given allele was predicted to bind (Supplementary Figure 4.1). We uncovered significant inconsistencies in these predictions between tools (Figure 4.2). Note that this phenomenon is independent of binding affinity threshold (Supplementary Figure 4.2).

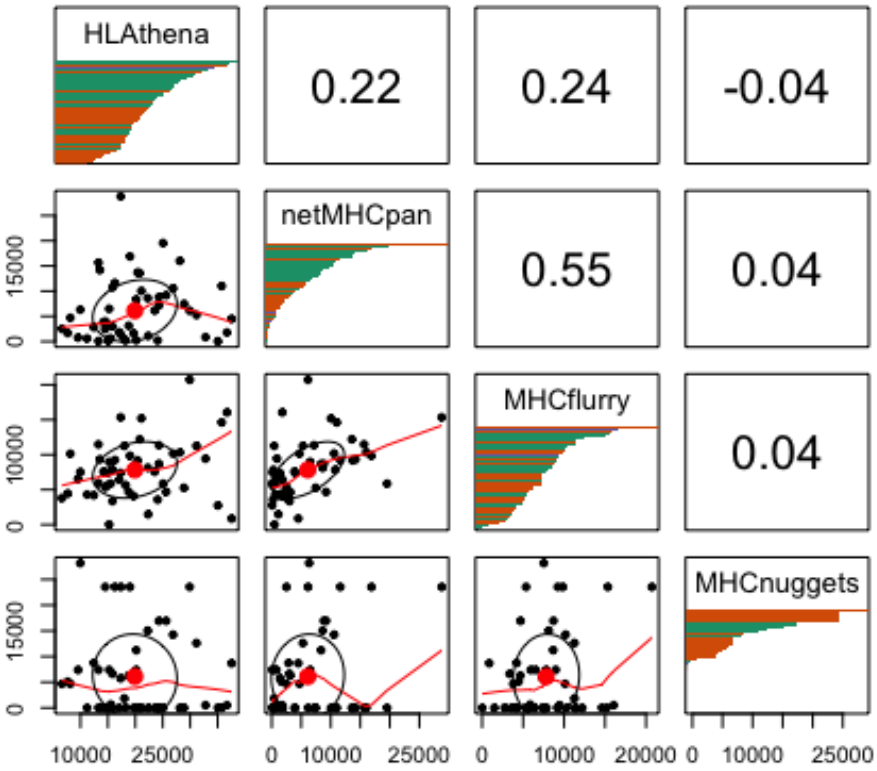
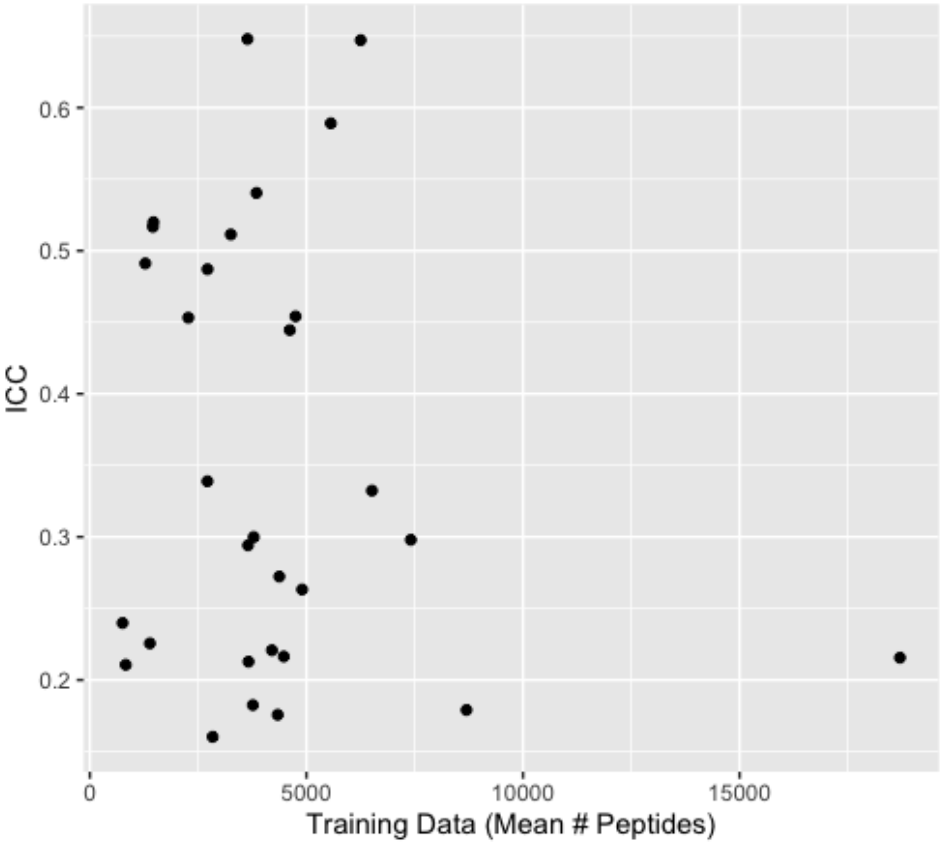


Figure 4.2: The correlation of HLA allelic presentation of 8-11mers from the random proteome between tools. The lower left grouping of plots displays scatter plots of peptides predicted to bind (>=

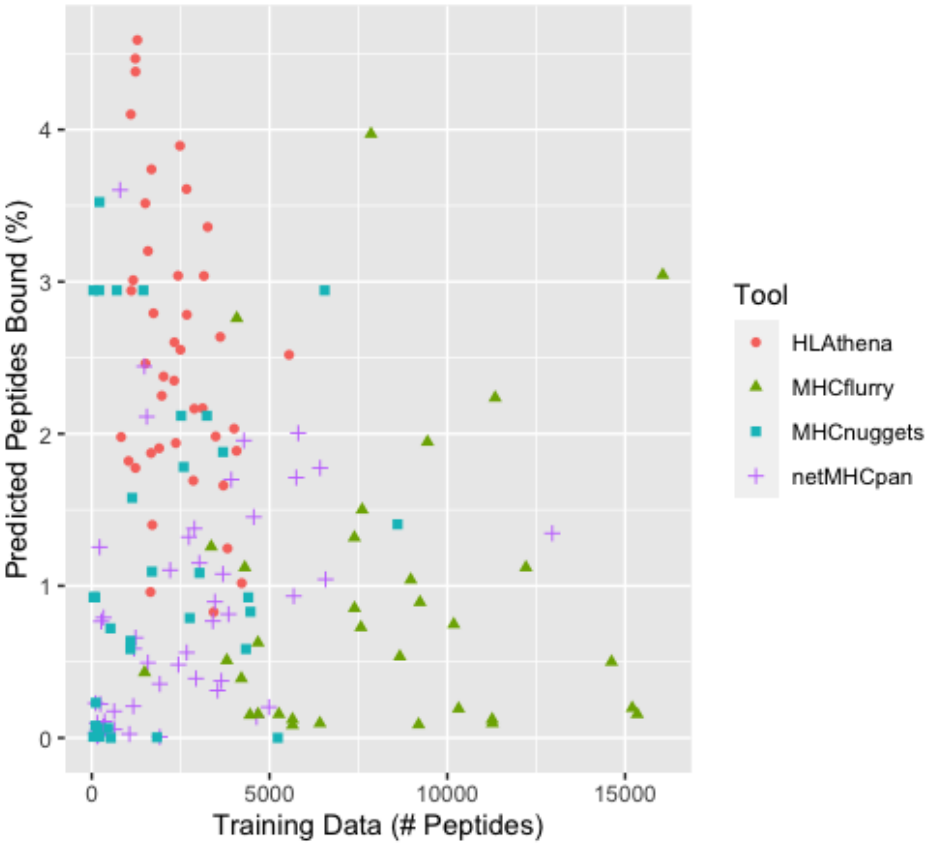
0.5 binding probability score) between 2 tools with each point representing the number of predicted binders for each HLA allele. The upper right grouping represents the Spearman correlation of the number of peptides predicted to bind to all alleles between tools. Note that MHCnuggets has a number of alleles with 0 random peptides predicted to bind. The diagonal panels show distribution of HLA allelic presentation from the random proteome for each tool. The number of peptides that putatively bind to each of the HLA alleles is shown along the x-axis as a series of horizontal bars with green, orange, and purple colors representing HLA-A, -B, and -C alleles, respectively, sorted in order of decreasing quantity of binders.

4.3.2 Amount of training data does not explain inconsistencies between tools

As each allele has a different amount of training data, we were next interested in exploring to what extent the quantity and quality of training data available to each tool might influence its allele-specific predictions. Indeed, some netMHCpan predictive models for some alleles are based on as few as 101 peptides, while others from MHCflurry are based on as many as 31,775 peptides (Supplementary Table 4.1). Note that we excluded from consideration the ~95% of alleles (4108) that were available for prediction but had no underlying allele-specific training data available (Supplementary Table 4.2). Ultimately, we found that the amount of training data available was not significantly related to the consistency of binding predictions between tools (Figure 4.3a), nor was it clearly related to the quantity of binding peptides predicted by tools (Figure 4.3b).


ICC
0.6
0.5
0.4
0.3
0.2
0
5000
10000
15000
Training Data (Mean # Peptides)

A


Predicted Peptides Bound (%)
4
3
2
1
0
0
5000
10000
15000
Training Data (# Peptides)
Tool
HLAthena
MHCflurry
MHCnuggets
netMHCpan

B

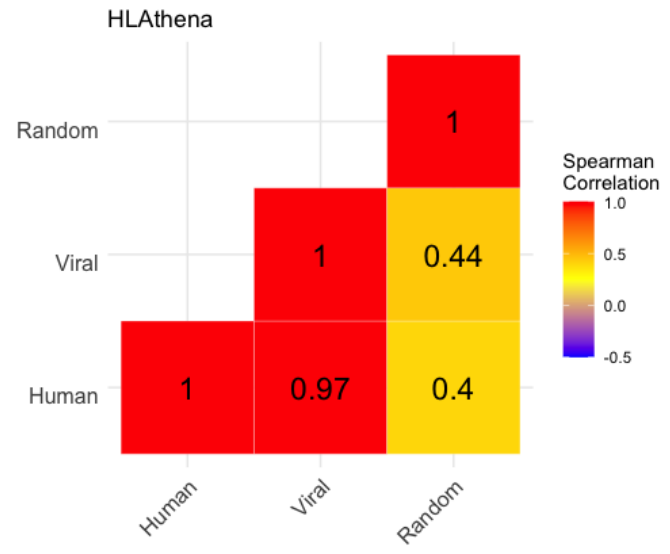
Figure 4.3. The relationship between training data and consistency of predictions. A) Scatterplot of ICC vs mean training data across 4 tools with each point representing data for a single HLA allele. The mean number of training peptides is shown on the x-axis while the ICC score is shown on the y-axis. B) Scatterplot of the relationship between training data and predicted peptide binding. The number of peptides used as training data for an allele is shown on the x-axis whereas the number of peptides predicted to bind for the same allele is shown on the y-axis. Each dot is a single allele with each color representing a different tool: red circles (HLAthena), green triangles (MHCflurry), blue squares (MHCnuggets), purple plus signs (netMHCpan). We note that netMHCpan does not make all of their training data available, thus the depicted quantity of training data represents an estimate.

4.3.3 Predicted binding quantities are similar between human and viral proteomes

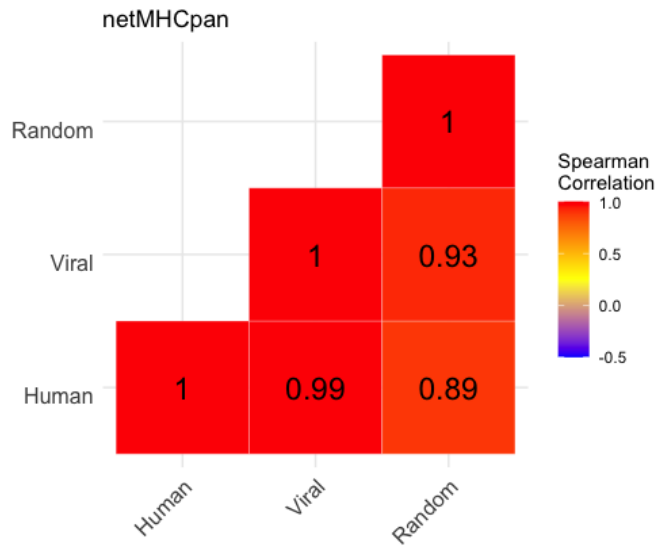
According to the pathogen driven selection theory of MHC evolution, different HLA alleles are anticipated to be particularly attuned to foreign as opposed to self-antigens (29–31,39,225,229). We therefore sought to compare the predicted capacity of different HLA alleles to present different viral vs. self-antigens. Further, we wished to establish which specific alleles had the propensity to bind a larger fraction of peptides in general (allele promiscuity) by observing the relationship between an allele's ability to bind random peptides versus peptides from a viral or human proteome.

We examined distribution of predicted allelic promiscuity across alleles for 9 sets of peptides of viral, human, and random origin (See Methods). Confining attention to human and viral proteomes, we again found a wide range in the proportion of peptides a given allele was predicted to bind and also significant inconsistencies between tools (Supplementary Figure 4.3).

We found that the alleles with highest mean binding percentage for human and viral peptides were B15:03 (2.68%) and B15:02 (2.36%) and the allele lowest mean binding percentage were B18:01 (0.24%) and A01:01 (0.33%) (Supplementary Table 4.3). No alleles were predicted by any tool to preferentially present either viral or human peptides. Further, the distribution of predicted allelic promiscuity across alleles was highly consistent between human and viral proteomes, but not when applied to a set of random peptides (Figure 4.4). We noted that this phenomenon holds for closely related viruses across all tools and to a lesser extent for more distantly related viruses (Supplementary Figure 4.4).



A



B

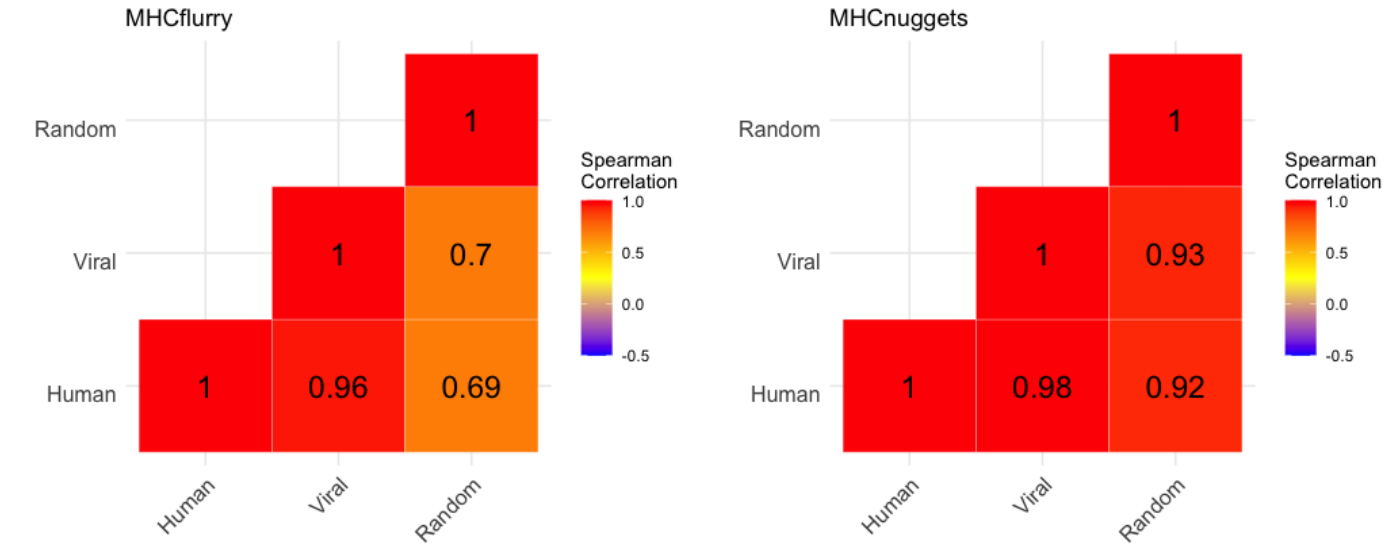


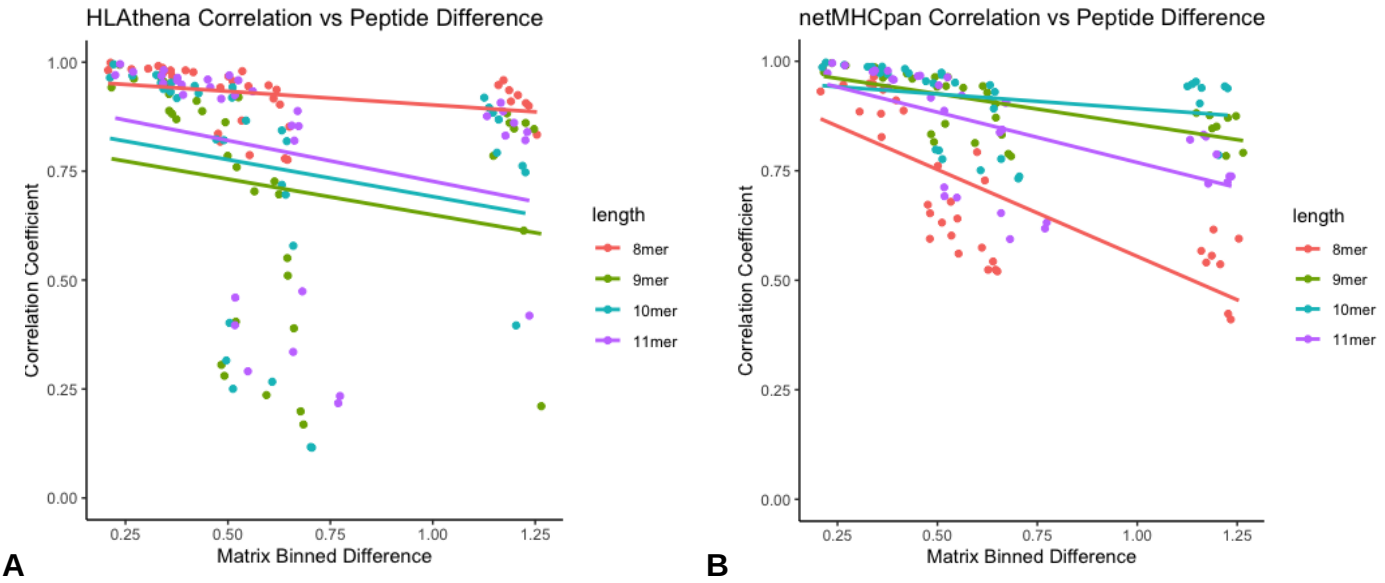
Figure 4.4. The correlation between peptide sources of predicted allelic promiscuity across alleles. A) Heatmap of spearman correlation between peptide sources for HLAthena-based predictions for human peptides, viral peptides, and randomly generated peptides. Numbers show Spearman correlation coefficients between each pair respectively, while color reflects the Spearman correlation with red approaching a Spearman correlation of 1. Analogous data is shown for netMHCpan, MHCflurry, and MHCnuggets in panels B, C, and D, respectively.

Confining attention to the 9 alleles whose predictive models were likely most robust (based on a minimum of 2000 training peptides for every tool), we again found that the distribution of predicted allelic promiscuity across alleles was consistent between closely related viruses and to a lesser extent between more distantly related viruses (Supplementary Figure 4.5).

4.3.4 Peptide physical properties are associated with allele-specific binding predictions

Reasoning that differences in peptide characteristics were the likeliest explanation for predicted differences in binding affinity between different alleles and peptide sources, we next studied the distribution of physical properties among different peptide sets. Human, viral, and random peptide sets all exhibited the same range of physical properties, but were differentially enriched among different physical properties (Supplementary Figure 4.6). Between individual peptide sets, the differential enrichment ranged from 10% (CMV v. human) to 63% (BK v. random) of peptides (Supplementary Figure 4.7).

We next sought to discover the relationship between the peptide similarity in physical property space and distribution of predicted allelic promiscuity across alleles. Across all tools, there was a positive relationship between similarity in physical property space and distribution of predicted allelic promiscuity across alleles as evidenced by the negative correlation between peptide set difference and Spearman correlation coefficient (Figure 4.6).



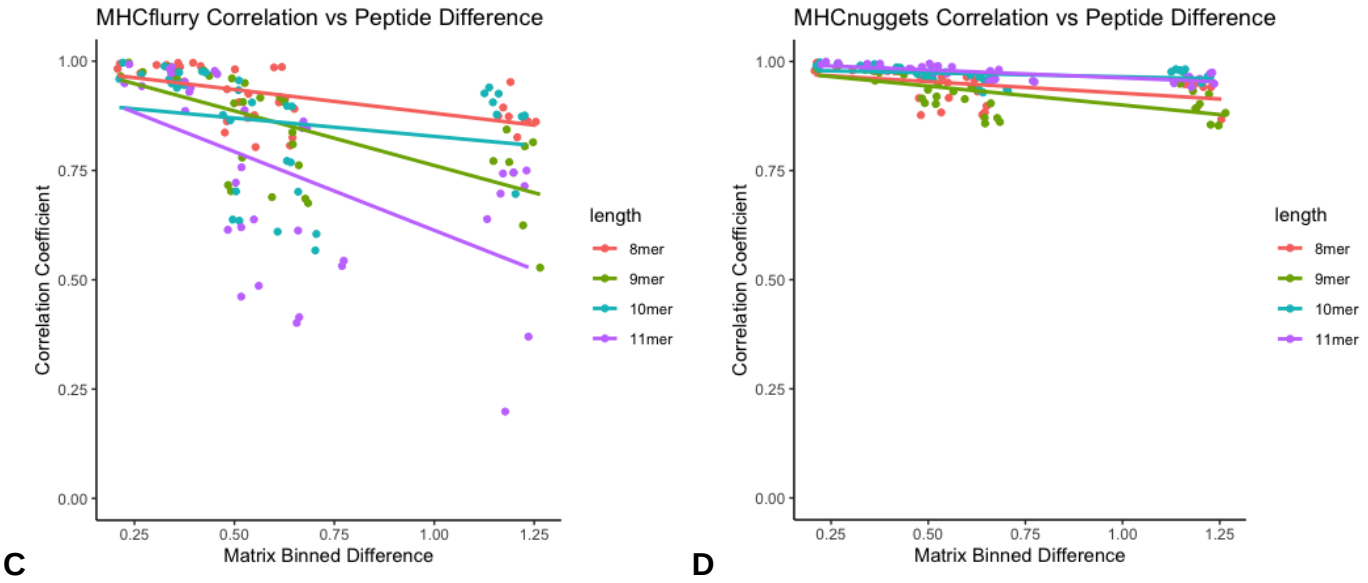


Figure 4.6. The relationship between physical property similarity vs peptide binding similarity. A) Scatterplot for HLAthena-based predictions, where each point represents predictions for a species vs species pair. Peptide dissimilarity is shown on the x-axis, whereas Spearman correlation coefficients of predicted allelic promiscuity across alleles. Color represents the length of peptide, with 8-, 9-, 10-, and 11-mers shown in red, green, blue and purple, respectively. Analogous data is shown for netMHCpan, MHCflurr, and MHCnuggets in panels B, C, and D, respectively.

Next, we found that each allele has distinct preferences for different peptide physical properties, independent of peptide length (Figure 4.7A, Supplementary Figure 4.8). Some alleles (e.g. A01:01 and B08:01) have stronger preference for certain physical properties (Figure 4.7B,C), while others (B45:01) do not have as clear of a preference (Figure 4.7D).

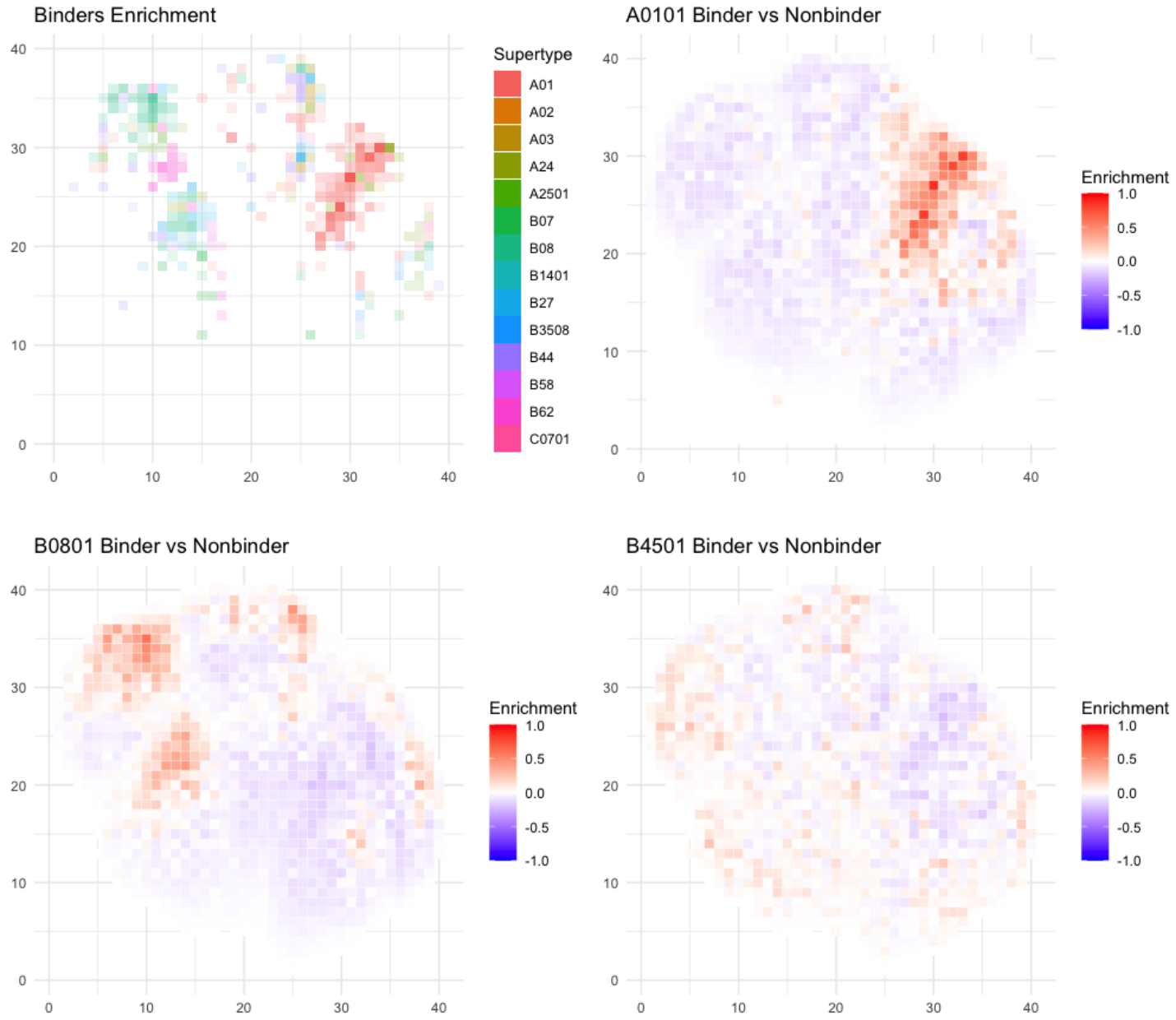


Figure 4.7. Differential distributions of physical properties for 9-mer peptides predicted to bind to HLA alleles. A) The plotting coordinates represent the first two dimensions of a UMAP transform of peptide physical properties, which is divided into 1600 (40x40) equivalently-sized square bins (see Methods). For each bin where there is at least one HLA allele with >0.2% difference in proportion of all peptides predicted to bind v. non-binders, the identity of the most enriched allele is shaded in the color corresponding to that allele's supertype as corresponding to the legend. B-D) Example plots of three different alleles (A01:01, B08:01, and B45:01) with different distributions of binders. Each box

represents enrichment as the percent peptide difference between predicted binders and non-binders for the given allele. The color scale shows the percent of peptides difference in the given box, with red meaning a larger number of predicted binders and blue meaning a larger number of predicted non-binders.

4.4 Discussion

To the best of our knowledge, this is the first study to examine the consistency of predictions of peptide-MHC binding across different tools, and to explore the quality and quantity of training data in this context. We note several limitations to this work. Firstly, we confined attention to MHC class I peptides and did not include predictions for MHC class II (237), of which there are numerous alleles. We also excluded from consideration any potential contributions of proteasomal cleavage or other antigen processing machinery to MHC binding (238–240). We did not seek to comprehensively assess all available tools for peptide-MHC binding affinity prediction, but rather confined our attention to four of the most widely used tools. The majority of our randomly generated peptides are not known to be found in nature and may not represent the optimal background distribution for measuring allele promiscuity or interrater reliability between tools primarily used for human and pathogenic peptides. While our analysis of peptides leveraged four essential and well-described amino acid physical properties, there may exist unassessed latent features that could capture additional variance and improve dimensionally-reduced comparisons. We did not assess the extent to which mass spectrometry biases in the training datasets might affect peptide-MHC predictions (241–244). Lastly, we did not evaluate individual tool performance based on known epitopes as this has been previously reported (93,96,97,103–105,110,234,245,246).

Our work raises fundamental questions about the fidelity of peptide-MHC binding prediction tools. Why, for instance, can predictions be so discordant among tools for which training datasets are

otherwise so similar? We especially worry about the real-world use of these prediction tools for alleles without any direct basis in training data. Why is the predicted range of allele promiscuity so substantial, and yet not demonstrative of any meaningful differences in enrichment between potential foreign versus self antigens? Moreover, is this differential promiscuity a universal biological phenomenon, with certain alleles being generally poor functional presenters of antigen? If this is the case, what selective advantage might have evolutionarily maintained these alleles in the population? Evaluating more viruses – as well as bacteria, fungi, and other pathogens – and linking these analyses with metrics such as evolutionary distance may give greater insight into the relationship between HLA evolution and disease.

4.5 Methods

4.5.1 Sequence retrieval, peptide filtering, and kmerization

FASTA-formatted protein sequence data was retrieved from the National Center of Biotechnology Information (NCBI) (129,190) using RefSeq as of 1-31-22 for BK, SARS-CoV-2, HHV-5, HHV-6, HSV-1, HSV-2, HSV-4, and Human. Protein sequence data was inputted into netchop v3.0 "C-term" model with a cleavage threshold of 0.1 to remove peptides that were not predicted to undergo canonical MHC class I antigen processing via proteasomal cleavage (of the peptide's C-terminus). The results from netchop v3.0 were then kmerized sequentially into 8- to 12-mers. Code used for kmerization and netchop filtering can be found at: https://github.com/pdxgx/covid19. We additionally generated a set of 1 million random peptides of length 8-12 drawn uniformly at random. Peptide sets had negligible overlap (<1% shared between human vs viral vs random peptides).

4.5.2 Peptide-MHC class I binding affinity predictions

MHC class I binding affinity predictions were performed for the peptides generated from the kmerization process above using 4 tools: netMHCpan v4.1 (96), HLAthena v1.0 (93), MHCflurry v2.0 (97), and MHCnuggets v2.3 (98). netMHCpan was run with default options with the '-l' option to specify peptides of lengths 8-12. MHCflurry was run with default options. MHCnuggets was run with default options. HLAthena was run using the dockerized version of HLAthena with default options, which predicts peptides of length 8-11. MHC class I binding affinity predictions were performed for each of 24, 26, and 2, HLA-A, -B, and -C alleles, respectively. Only alleles that were in common between all 4 tools were used (52 total alleles in common between 2489 possible alleles). Binding affinity values were converted to binding probability values for MHCflurry and MHCnuggets using 1-log(binding affinity) / log(50000) in order to match HLAthena and netMHCpan binding probability predictions. Alleles were grouped into supertypes when applicable using the HLA class I revised classification (11).

4.5.3 Dimensional reduction and binning analysis

Peptides were converted into physical property matrices using amino acid sequence mapping into a 4*kmer length matrix containing each amino acid's properties in sequence. The following physical properties of the amino acids were encoded: side chain polarity was recorded as its isoelectric point (pI) (247), the molecular volume of each side chain was recorded as its partial molar volume at 37°C (248), the hydrophobicity of each side chain was characterized by its simulated contact angle with nanodroplets of water (249) and conformational entropy was derived from peptide bond angular observations among protein sequences without observed secondary structure (250).

Each dimensional reduction was performed on the pooled set of k-mers. UMAP dimensionality was performed using uwot UMAP R implementation v0.1.11. PCA was performed using default prcomp() functions in base R v4.1.3.

For each peptide source, binned matrices were computed using the bin2() function with 40x40 (1600) bins from the Ash v1.0.15 package (251) in R v4.1.3. Bin values were then divided by the total number of peptides to create bins with the % of total peptides. In order to compare between 2 peptide sources, a matrix, called the difference matrix, is created by subtracting one matrix of a peptide source from another. Taking the absolute value of each bin in the difference matrix, then summing the values together, results in a single metric ranging from 0-2 measuring the difference in binned density between 2 peptide sources, the value 2 indicating that no peptides were shared between bins and the value 0 indicating the same percentage of peptides in every bin (Figure 4.8).

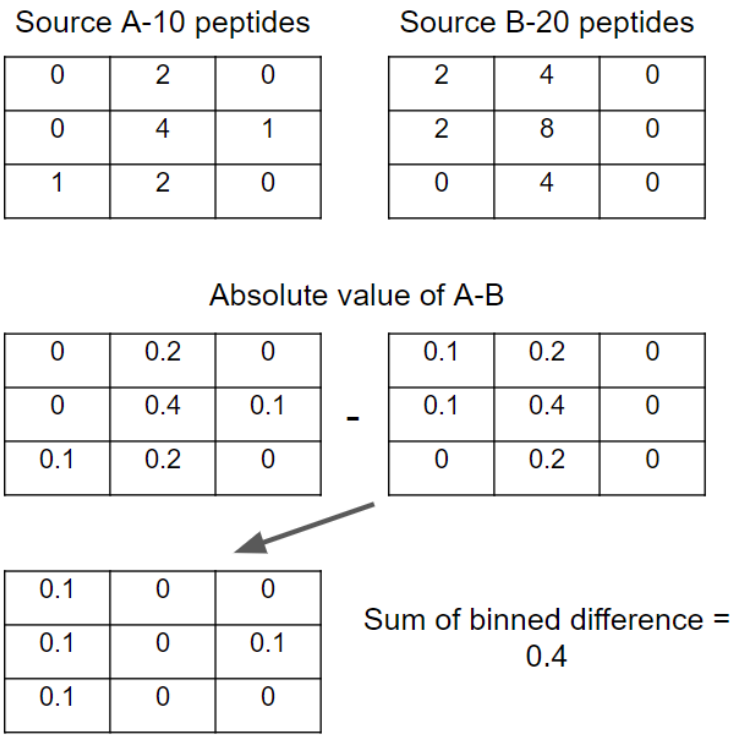


Figure 4.8.

4.5.4 Allele ordering similarity

For each allele-peptide source combination, the percentage of peptides predicted to bind with a binding probability score of 0.5 or greater was calculated for all processed peptides. 0.5 binding score is estimated to be equivalent to 250-300nM depending on the tool used. For each peptide

source, alleles were ranked from best to worst binders (most to least peptides >= 0.5 score) t. In order to compute allele ordering similarity between 2 peptide sources for a single tool, Spearman's Rank Correlation Coefficient was calculated between the 2 sets of allele ranks.

For the random group 1 vs random group 2 analysis, we conducted 100 replicates of dividing the randomly generated peptides into 2 random groups and performed a Spearman rank test of allele ordering between these groups for each of the tools.

4.5.5 Interrater reliability

Intraclass correlation coefficients (ICCs) were calculated using the ICC() function from the IRR v0.84.1 R package (252). Binding prediction scores for all 1 million randomly generated peptides were separated by tool and HLA allele, and an ICC was calculated as the interrater reliability metric between the 4 tools for each allele. ICC was also between the 4 tools on a per peptide basis, each peptide receiving a score across 4 tools using predictions separated by tool and peptide.

Conclusion

5.1 Summary

In this dissertation, I examined the utility of peptide-MHC prediction as a predictor of health outcomes and identified new challenges and opportunities in this space. I found that in SARS-CoV, a pipeline based on peptide-MHC prediction could recapitulate hospital studies that associated severity with specific HLA alleles, and applying this method to SARS-CoV-2 identified similar alleles that could be associated with more severe infection. However, when applying this pipeline to a large cohort of individuals with genotyping and clinical outcomes data, I showed that age, BMI, and other comorbidities determined the likelihood of developing symptoms and severity of disease to a far greater degree than any single allele or any metric based on potential SARS-CoV-2 peptide presentation. Only when individuals were young and had no other comorbidities did HLA have any sort of association. Further, I found the majority of literature studies on the same topic attempting to associate disease outcomes with HLA showed clear statistical biases, without applying corrections or making conclusions with limited sample size. I attempted to apply this method to a greater number of viruses, the human proteome, and a randomly generated background distribution in hopes of gleaning more insight into why there was little to no relationship between SARS-CoV-2 binding and severity of disease but found that there are fundamental questions on the fidelity of peptide-MHC binders as a whole, namely the inconsistency of peptide prediction across tools. Finally, I illustrate the relationship between similarities in physical properties and similarities in binding predictions between pairs of peptide sets.

5.2 Future directions and implications

The number of potential future directions of this work are vast. With the exponential discovery rate of viruses and the increasing number of other infectious diseases, experiments that can build

upon this work to identify important predictors of susceptibility and severity to disease based on individual genomic variation such as HLA may lead to critical developments in disease prevention and treatment.

More critical may be additional studies assessing the validity of peptide-MHC predictors. My work has demonstrated serious inconsistencies in agreement between peptide prediction. As described in Figure 4.1c, HLAthena, which evaluates its predictions as compared to other tools using PPV, shows the most peptide-MHC pairs without agreement with the other tools. Similarly, benchmarking using AUC has displayed “improvements” in accuracy despite what seems to be general unreliability. As peptide-MHC binding may be viewed more of as an outlier detection machine learning problem where the number of true negatives (non-binders) is far larger than the number of true positives (binders), it is important to evaluate accuracy by probing all of true positive, true negative, false positive, and false negative rates, possibly by using a summary metric of these such as F1 and AUPR in addition to than AUC (253) and PPV. Further, I have demonstrated the lack of improvement with increased training data for allele predictions and agreement. However, this does not entirely mean that more training data is not necessary. There is a possibility that available training data is orders of magnitude smaller than is necessary for accurate predictors, which may be why tools such as HLAthena had found that models trained going from 100 peptides to 1000 peptides was jump in improvement but not 1000 to 2000 peptides. However, it may possible that the current magnitude of data is sufficient if there were a significant number of negative binders. Conducting more experiments on potential negative binders (peptides generated at random, peptides not known to be eluted/not yet profiled by mass spectrometry) would add significant value to future work on developing new peptide-MHC predictors to balance training data, as current training data is heavily skewed towards positive examples (93,111).

Peptide-MHC binding is only one step towards assessing immunogenicity. A number of studies found that binding affinity and immunogenicity are related (61,254) and a cutoff of 500nM is often

used to identify potential T cell epitopes. These experiments, as well as other works supporting this hypothesis, have primarily been conducted on limited numbers of viral peptides. Determining better immunogenic cutoffs as well as assessing closely related but distinct metrics such as peptide-MHC stability and quantifying mathematically the probability of a peptide to be a T cell epitope, for example using some logistic regression model, would vastly increase the translatability of work in this entire field.

As certain regions in peptide space, despite the general disagreement between tools in predictions, were heavily favored for predicted binders for specific alleles, there are 2 possible implications. The first implication is that, for specific alleles, there are amino acid properties at specific positions that would cause a peptide to be more likely to bind to a specific MHC molecule. This is supported by the concept of anchor residues, where peptide residues bind to specific regions of MHC I, enhancing the stability of peptide-MHC binding (254–256). The second possible implication is that there are technical artifacts during the training of all the models, with specific amino acid inputs as predicted binders regardless of the physical properties. This phenomenon does not apply to all alleles, however, there are alleles with enriched regions of binding across alleles of different supertypes. As HLA is extremely polymorphic and assigning peptides to specific HLA alleles would result in extremely low coverage, grouping alleles into supertypes is primarily used to group alleles by "largely overlapping peptide specificity" (257). However, because there are alleles with the same enriched binding region across supertypes, we may reevaluate both assignment of these alleles into more than one supertype as well as determine binding features based on HLA mutations rather than using supertype grouping entirely. This may prove to better utilize the small amounts of training data that we do have and result in further identification of physical binding motifs that are not yet characterized by binding affinity studies.

Extending this work to more than the small number of viruses assessed here and to a wider range of alleles would provide further insight into the relationship between HLA evolution and

disease. For example, further analyses can be performed on viruses in comparison with phylogenetic distances. Combining a whole peptide set distance metric with predicted peptide binding and phylogenetic differences may yield further insight into the evolution of peptide-MHC binding. Many more disease-specific questions remain and may lead to improvements in vaccine development, and while we would all hope that this is not the case, identification and prioritization of susceptible populations to the next pandemic.

One final possible future direction starts in higher fidelity MHC-region sequencing. With increased resolution, we may be able to start developing a better sequence-to-sequence predictor of binding between peptide and MHC. While this would result in “less” training data per individual mutation in the MHC, we may glean more insight into the relationship between MHC sequence and peptide sequence, which gives potential for cross learning and each future training example adding to the accuracy of the whole model, no matter the allele.This may allow for better multi-allelic training and further enlighten specific motif binding pairs by sequence.

5.4 Concluding Remarks

On a final note, this dissertation constitutes an initial framework to predict and assess susceptibility to infectious disease. It also serves to caution when attempting to use peptide-MHC predictors as important steps in the critical healthcare decisions such as the development of therapeutics or evaluating immunotherapies. This work has demonstrated fundamental inconsistencies in commonly used peptide-MHC predictors, which are used used to evaluate an individual or population’s susceptibility to viral infection (207), to develop an understanding of specific autoimmune conditions (123), to improve transplantation technologies (122), or even to assist in the development of personalized cancer vaccines (73,230–233).

Ultimately, if we are able to improve peptide-MHC predictors and further establish our ability to predict immunogenicity on both population and patient specific levels, we will be able to improve disease outcomes for a wide variety of infectious and autoimmune disorders. One can imagine an idealized situation where we have the ability to rapidly synthesize a personal vaccine, where a patient merely has to have their HLA region sequenced and would quickly be able to receive a vaccine that would prime the patient's immune system to rapidly fight the disease or cancer.

References

1. Blum JS, Wearsch PA, Cresswell P. Pathways of Antigen Processing. Annu Rev Immunol. 2013;31(1):443–73.

2. Cruz-Tapias P, Castiblanco J, Anaya JM. Major histocompatibility complex: Antigen processing and presentation [Internet]. Autoimmunity: From Bench to Bedside [Internet]. El Rosario University Press; 2013 [cited 2021 Oct 21]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK459467/

3. Pishesha N, Harmand TJ, Ploegh HL. A guide to antigen processing and presentation. Nat Rev Immunol. 2022 Apr 13;1–14.

4. Vyas JM, Van der Veen AG, Ploegh HL. The known unknowns of antigen processing and presentation. Nat Rev Immunol. 2008 Aug;8(8):607–18.

5. Neefjes J, Jongsma MLM, Paul P, Bakke O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. Nat Rev Immunol. 2011 Dec;11(12):823–36.

6. Becar M, Kasi A. Physiology, MHC Class I. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 [cited 2022 Sep 5]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK556022/

7. Kloetzel PM. Antigen processing by the proteasome. Nat Rev Mol Cell Biol. 2001 Mar;2(3):179–88.

8. Berko D, Tabachnick-Cherny S, Shental-Bechor D, Cascio P, Mioletti S, Levy Y, et al. The Direction of Protein Entry into the Proteasome Determines the Variety of Products and Depends on the Force Needed to Unfold Its Two Termini. Mol Cell. 2012 Nov;48(4):601–11.

9. Chapiro J, Claverol S, Piette F, Ma W, Stroobant V, Guillaume B, et al. Destructive Cleavage of Antigenic Peptides Either by the Immunoproteasome or by the Standard Proteasome Results in Differential Antigen Presentation. J Immunol. 2006 Jan 15;176(2):1053–61.

10. Barbosa CRR, Barton J, Shepherd AJ, Mishto M. Mechanistic diversity in MHC class I antigen recognition. Biochem J. 2021 Dec 23;478(24):4187–202.

11. Sidney J, Peters B, Frahm N, Brander C, Sette A. HLA class I supertypes: a revised and updated classification. BMC Immunol. 2008 Jan 22;9(1):1.

12. Neefjes JJ, Ploegh HL. Allele and locus-specific differences in cell surface expression and the association of HLA class I heavy chain with β2-microglobulin: differential effects of inhibition of glycosylation on class I subunit association. Eur J Immunol. 1988;18(5):801–10.

13. Trowsdale J. HLA genomics in the third millennium. Curr Opin Immunol. 2005 Oct 1;17(5):498–504.

14. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68–74.

15. HLA Nomenclature @ hla.alleles.org [Internet]. [cited 2022 Sep 5]. Available from: http://hla.alleles.org/nomenclature/stats.html

16. Sewell AK. Why must T cells be cross-reactive? Nat Rev Immunol. 2012 Sep;12(9):669–77.

17. Seldin MF. The genetics of human autoimmune disease: A perspective on progress in the field and future directions. J Autoimmun. 2015 Nov 1;64:1–12.

18. Muñiz-Castrillo S, Vogrig A, Honnorat J. Associations between HLA and autoimmune neurological diseases with autoantibodies. Autoimmun Highlights. 2020 Jan 22;11(1):2.

19. Kelley J, de Bono B, Trowsdale J. IRIS: A database surveying known human immune system genes. Genomics. 2005 Apr 1;85(4):503–11.

20. Noble JA, Erlich HA. Genetics of Type 1 Diabetes. Cold Spring Harb Perspect Med. 2012 Jan;2(1):a007732.

21. Pereyra F, Jia X, McLaren PJ, Telenti A, de Bakker PIW, Walker BD, et al. The Major Genetic Determinants of HIV-1 Control Affect HLA Class I Peptide Presentation. Science. 2010 Dec 10;330(6010):1551–7.

22. Zhang Y, Peng Y, Yan H, Xu K, Saito M, Wu H, et al. Multilayered defence in HLA-B51 associated HIV viral control. J Immunol Baltim Md 1950. 2011 Jul 15;187(2):684–91.

23. Thio CL, Thomas DL, Karacki P, Gao X, Marti D, Kaslow RA, et al. Comprehensive Analysis of Class I and Class II HLA Antigens and Chronic Hepatitis B Virus Infection. J Virol. 2003 Nov;77(22):12083–7.

24. Yengo CK, Torimiro J, Kowo M, Lebon PA, Tiedeu BA, Luma H, et al. Variation of HLA class I (-A and -C) genes in individuals infected with hepatitis B or hepatitis C virus in Cameroon. Heliyon. 2020 Oct 1;6(10):e05232.

25. Rao X, Hoof I, van Baarle D, Keşmir C, Textor J. HLA Preferences for Conserved Epitopes: A Potential Mechanism for Hepatitis C Clearance. Front Immunol [Internet]. 2015 [cited 2019 Nov 5];6. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2015.00552/full

26. Pignatelli M, Waters J, Brown D, Lever A, Iwarson S, Schaff Z, et al. HLA class I antigens on the hepatocyte membrane during recovery from acute hepatitis B virus infection and during interferon therapy in chronic hepatitis B virus infection. Hepatology. 1986;6(3):349–53.

27. Debebe BJ, Boelen L, Lee JC, IAVI Protocol C Investigators, Thio CL, Astemborski J, et al. Identifying the immune interactions underlying HLA class I disease associations. Davenport MP, Walczak AM, Lipsitch M, editors. eLife. 2020 Apr 2;9:e54558.

28. Evolution and Diversity of the Human Leukocyte Antigen(HLA) - PMC [Internet]. [cited 2022 Sep 8]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4315060/

29. Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V, Balloux F. Pathogen-Driven Selection and Worldwide HLA Class I Diversity. Curr Biol. 2005 Jun 7;15(11):1022–7.

30. Manczinger M, Boross G, Kemény L, Müller V, Lenz TL, Papp B, et al. Pathogen diversity drives the evolution of generalist MHC-II alleles in human populations. PLoS Biol. 2019 Jan 31;17(1):e3000131.

31. Spurgin LG, Richardson DS. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc R Soc B Biol Sci. 2010 Apr 7;277(1684):979–88.

32. Tan Y, Schneider T, Leong M, Aravind L, Zhang D. Novel Immunoglobulin Domain Proteins Provide Insights into Evolution and Pathogenesis Mechanisms of SARS-Related Coronaviruses.

33. Blackwell JM, Jamieson SE, Burgner D. HLA and Infectious Diseases. Clin Microbiol Rev. 2009 Apr;22(2):370–85.

34. Stephens H a. F, Klaythong R, Sirikong M, Vaughn DW, Green S, Kalayanarooj S, et al. HLA-A and -B allele associations with secondary dengue virus infections correlate with disease severity and the infecting viral serotype in ethnic Thais. Tissue Antigens. 2002 Oct;60(4):309–18.

35. MacDonald KS, Fowke KR, Kimani J, Dunand VA, Nagelkerke NJ, Ball TB, et al. Influence of HLA supertypes on susceptibility and resistance to human immunodeficiency virus type 1 infection. J Infect Dis. 2000 May;181(5):1581–9.

36. Lin M, Tseng HK, Trejaut JA, Lee HL, Loo JH, Chu CC, et al. Association of HLA class I with severe acute respiratory syndrome coronavirus infection. BMC Med Genet. 2003 Sep 12;4:9.

37. Gough SCL, Simmonds MJ. The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action. Curr Genomics. 2007 Nov;8(7):453–65.

38. Cardozo DM, Marangon AV, Sell AM, Visentainer JEL, Souza CA de. HLA and Infectious Diseases [Internet]. HLA and Associated Important Diseases. IntechOpen; 2014 [cited 2022 Jun 22]. Available from: https://www.intechopen.com/chapters/undefined/state.item.id

39. White CF, Pellis L, Keeling MJ, Penman BS. Detecting HLA-infectious disease associations for multi-strain pathogens. Infect Genet Evol. 2020 Sep 1;83:104344.

40. Slade JWG, Watson MJ, MacDougall-Shackleton EA. “Balancing” balancing selection? Assortative mating at the major histocompatibility complex despite molecular signatures of balancing selection. Ecol Evol. 2019 Apr 13;9(9):5146–57.

41. Adams MJ, Hendrickson RC, Dempsey DM, Lefkowitz EJ. Tracking the changes in virus taxonomy. Arch Virol. 2015 May 1;160(5):1375–83.

42. Overview of Viruses and Virus Infection - PMC [Internet]. [cited 2022 Nov 11]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7149408/

43. Mortality Analyses - Johns Hopkins Coronavirus Resource Center [Internet]. [cited 2022 Sep 7]. Available from: https://coronavirus.jhu.edu/data/mortality

44. Lopez-Leon S, Wegman-Ostrosky T, Perelman C, Sepulveda R, Rebolledo PA, Cuapio A, et al. More than 50 Long-term effects of COVID-19: a systematic review and meta-analysis. medRxiv. 2021 Jan 30;2021.01.27.21250617.

45. Rosendahl Huber S, van Beek J, de Jonge J, Luytjes W, van Baarle D. T Cell Responses to Viral Infections – Opportunities for Peptide Vaccination. Front Immunol [Internet]. 2014 [cited 2022 Sep 6];5. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2014.00171

46. De Clercq E, Li G. Approved Antiviral Drugs over the Past 50 Years. Clin Microbiol Rev. 2016 Jul;29(3):695–747.

47. Métifiot M, Marchand C, Pommier Y. HIV Integrase Inhibitors: 20-Year Landmark and Challenges. Adv Pharmacol San Diego Calif. 2013;67:75–105.

48. Saleh A, Qamar S, Tekin A, Singh R, Kashyap R. Vaccine Development Throughout History. Cureus. 13(7):e16635.

49. Research C for BE and. Vaccines Licensed for Use in the United States. FDA [Internet]. 2022 Oct 17 [cited 2022 Nov 8]; Available from: https://www.fda.gov/vaccines-blood-biologics/vaccines/vaccines-licensed-use-united-states

50. Patronov A, Doytchinova I. T-cell epitope vaccine design by immunoinformatics. Open Biol. 3(1):120139.

51. Donnelly JJ, Liu MA, Ulmer JB. Antigen Presentation and DNA Vaccines. Am J Respir Crit Care Med. 2000 Oct;162(supplement_3):S190–3.

52. Wherry EJ, Barouch DH. T cell immunity to COVID-19 vaccines. Science. 2022 Aug 19;377(6608):821–2.

53. Lambert PH, Liu M, Siegrist CA. Can successful vaccines teach us how to induce efficient protective immune responses? Nature Medicine. 2005.

54. Heiny AT, Miotto O, Srinivasan KN, Khan AM, Zhang GL, Brusic V, et al. Evolutionarily Conserved Protein Sequences of Influenza A Viruses, Avian and Human, as Vaccine Targets. PLOS ONE. 2007 Nov 21;2(11):e1190.

55. Harper DM, Nieminen P, Donders G, Einstein MH, Garcia F, Huh WK, et al. The efficacy and safety of Tipapkinogen Sovacivec therapeutic HPV vaccine in cervical intraepithelial neoplasia grades 2 and 3: Randomized controlled phase II trial with 2.5 years of follow-up. Gynecol Oncol. 2019;153(3):521–9.

56. Kumar R, Qureshi H, Deshpande S, Bhattacharya J. Broadly neutralizing antibodies in HIV-1 treatment and prevention. Ther Adv Vaccines Immunother. 2018 Oct 12;6(4):61–8.

57. Nelde A, Rammensee HG, Walz JS. The Peptide Vaccine of the Future. Mol Cell Proteomics. 2021 Jan 1;20:100022.

58. Pol S. Immunotherapy of chronic hepatitis B by anti HBV vaccine. Biomed Pharmacother Biomedecine Pharmacother. 1995;49(3):105–9.

59. Croft NP, Smith SA, Pickering J, Sidney J, Peters B, Faridi P, et al. Most viral peptides displayed by class I MHC on infected cells are immunogenic. Proc Natl Acad Sci. 2019 Feb 19;116(8):3112–7.

60. Wood MA, Paralkar M, Paralkar MP, Nguyen A, Struck AJ, Ellrott K, et al. Population-level distribution and putative immunogenicity of cancer neoepitopes. BMC Cancer. 2018 Apr 13;18(1):414.

61. Sette A, Vitiello A, Reherman B, Fowler P, Nayersina R, Kast WM, et al. The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. J Immunol Baltim Md 1950. 1994 Dec 15;153(12):5586–92.

62. Paul S, Weiskopf D, Angelo MA, Sidney J, Peters B, Sette A. HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. J Immunol Baltim Md 1950. 2013 Dec 15;191(12):5831–9.

63. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000 Jan 7;100(1):57–70.

64. Hallmarks of Cancer: The Next Generation: Cell [Internet]. [cited 2019 Nov 5]. Available from: https://www.cell.com/abstract/S0092-8674(11)00127-9

65. Hanahan D. Hallmarks of Cancer: New Dimensions. Cancer Discov. 2022 Jan 12;12(1):31–46.

66. Bai P, Li Y, Zhou Q, Xia J, Wei PC, Deng H, et al. Immune-based mutation classification enables neoantigen prioritization and immune feature discovery in cancer immunotherapy. Oncoimmunology. 10(1):1868130.

67. Efremova M, Finotello F, Rieder D, Trajanoski Z. Neoantigens Generated by Individual Mutations and Their Role in Cancer Immunity and Immunotherapy. Front Immunol. 2017;8:1679.

68. Peng M, Mo Y, Wang Y, Wu P, Zhang Y, Xiong F, et al. Neoantigen vaccine: an emerging tumor immunotherapy. Mol Cancer. 2019 Aug 23;18(1):128.

69. van den Bulk J, Verdegaal EME, Ruano D, Ijsselsteijn ME, Visser M, van der Breggen R, et al. Neoantigen-specific immunity in low mutation burden colorectal cancers of the consensus molecular subtype 4. Genome Med. 2019 Dec;11(1):87.

70. Alcazer V, Bonaventura P, Tonon L, Wittmann S, Caux C, Depil S. Neoepitopes-based vaccines: challenges and perspectives. Eur J Cancer. 2019 Feb 1;108:55–60.

71. Wood MA, Nguyen A, Struck AJ, Ellrott K, Nellore A, Thompson RF. neoepiscope improves neoepitope prediction with multivariant phasing. Bioinformatics. 2020 Feb 1;36(3):713–20.

72. Neoantigens in cancer immunotherapy | Science [Internet]. [cited 2022 Sep 6]. Available from: https://www.science.org/doi/10.1126/science.aaa4971

73. Blass E, Ott PA. Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. Nat Rev Clin Oncol. 2021 Apr;18(4):215–29.

74. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, et al. pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. Genome Med. 2016 Jan 29;8(1):11.

75. Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. Nature. 2017 13;547(7662):217–21.

76. Zhou C, Zhu C, Liu Q. Toward in silico Identification of Tumor Neoantigens in Immunotherapy. Trends Mol Med. 2019 Nov 1;25(11):980–92.

77. Brennick CA, George MM, Srivastava PK, Karandikar SH. Prediction of cancer neoepitopes needs new rules. Semin Immunol. 2020 Feb 1;47:101387.

78. Immunotherapy for Cancer - NCI [Internet]. 2015 [cited 2022 Sep 6]. Available from: https://www.cancer.gov/about-cancer/treatment/types/immunotherapy

79. Borcoman E, Nandikolla A, Long G, Goel S, Le Tourneau C. Patterns of Response and Progression to Immunotherapy. Am Soc Clin Oncol Educ Book. 2018 May 23;(38):169–78.

80. Lennerz V, Fatho M, Gentilini C, Frye RA, Lifke A, Ferel D, et al. The response of autologous T cells to a human melanoma is dominated by mutated neoantigens. Proc Natl Acad Sci. 2005 Nov;102(44):16013–8.

81. Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. Genome Med. 2017 Apr 19;9(1):34.

82. Huang T, Chen X, Zhang H, Liang Y, Li L, Wei H, et al. Prognostic Role of Tumor Mutational Burden in Cancer Patients Treated With Immune Checkpoint Inhibitors: A Systematic Review and Meta-Analysis. Front Oncol [Internet]. 2021 [cited 2022 Sep 6];11. Available from: https://www.frontiersin.org/articles/10.3389/fonc.2021.706652

83. Shao C, Li G, Huang L, Pruitt S, Castellanos E, Frampton G, et al. Prevalence of High Tumor Mutational Burden and Association With Survival in Patients With Less Common Solid Tumors. JAMA Netw Open. 2020 Oct 29;3(10):e2025109.

84. Wood MA, Weeder BR, David JK, Nellore A, Thompson RF. Burden of tumor mutations, neoepitopes, and other variants are weak predictors of cancer immunotherapy response and overall survival. Genome Med. 2020 Mar 30;12(1):33.

85. Li Y, Ma Y, Wu Z, Zeng F, Song B, Zhang Y, et al. Tumor Mutational Burden Predicting the Efficacy of Immune Checkpoint Inhibitors in Colorectal Cancer: A Systematic Review and Meta-Analysis. Front Immunol [Internet]. 2021 [cited 2022 Sep 6];12. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2021.751407

86. McGrail DJ, Pilié PG, Rashid NU, Voorwerk L, Slagter M, Kok M, et al. High tumor mutation burden fails to predict immune checkpoint blockade response across all cancer types. Ann Oncol. 2021 May 1;32(5):661–72.

87. Meng G, Liu X, Ma T, Lv D, Sun G. Predictive value of tumor mutational burden for immunotherapy in non-small cell lung cancer: A systematic review and meta-analysis. PLOS ONE. 2022 Feb 3;17(2):e0263629.

88. Strickler JH, Hanks BA, Khasraw M. Tumor Mutational Burden as a Predictor of Immunotherapy Response: Is More Always Better? Clin Cancer Res. 2021 Mar 1;27(5):1236–41.

89. Comber JD, Philip R. MHC class I antigen presentation and implications for developing a new generation of therapeutic vaccines. Ther Adv Vaccines. 2014 May;2(3):77–89.

90. Sette A, Buus S, Colon S, Miles C, Grey HM. I-Ad-binding peptides derived from unrelated protein antigens share a common structural motif. J Immunol. 1988 Jul 1;141(1):45–8.

91. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanović S. SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics. 1999 Nov;50(3–4):213–9.

92. Application of an artificial neural network to predict specific class I MHC binding peptide sequences | Nature Biotechnology [Internet]. [cited 2022 Sep 8]. Available from: https://www.nature.com/articles/nbt0898-753

93. Sarkizova S, Klaeger S, Le PM, Li LW, Oliveira G, Keshishian H, et al. A large peptidome dataset improves HLA class I epitope prediction across most of the human population. Nat Biotechnol. 2020 Feb;38(2):199–209.

94. Nielsen M, Andreatta M, Peters B, Buus S. Immunoinformatics: Predicting Peptide–MHC Binding. Annu Rev Biomed Data Sci. 2020 Jul 20;3(1):191–215.

95. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics. 2016 Feb 15;32(4):511–7.

96. 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 Jul 2;48(W1):W449–54.

97. O'Donnell TJ, Rubinsteyn A, Laserson U. MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. Cell Syst. 2020 Jul;11(1):42-48.e7.

98. Shao XM, Bhattacharya R, Huang J, Sivakumar IKA, Tokheim C, Zheng L, et al. High-throughput prediction of MHC class I and class II neoantigens with MHCnuggets. Cancer Immunol Res. 2019 Dec 23;canimm.0464.2019.

99. Antunes DA, Devaurs D, Moll M, Lizée G, Kavraki LE. General Prediction of Peptide-MHC Binding Modes Using Incremental Docking: A Proof of Concept. Sci Rep. 2018 Mar 12;8(1):4327.

100. Farrell D. epitopepredict: a tool for integrated MHC binding prediction. Gigabyte. 2021 Feb 24;2021:1–14.

101. Marino F, Chong C, Michaux J, Bassani-Sternberg M. High-Throughput, Fast, and Sensitive Immunopeptidomics Sample Processing for Mass Spectrometry. Methods Mol Biol Clifton NJ. 2019;1913:67–79.

102. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing | Nature [Internet]. [cited 2022 Sep 8]. Available from: https://www.nature.com/articles/nature14001

103. Trolle T, Metushi IG, Greenbaum JA, Kim Y, Sidney J, Lund O, et al. Automated benchmarking of peptide-MHC class I binding predictions. Bioinformatics. 2015 Jul 1;31(13):2174–81.

104. Zhao W, Sher X. Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes. PLOS Comput Biol. 2018 Nov 8;14(11):e1006457.

105. Paul S, Croft NP, Purcell AW, Tscharke DC, Sette A, Nielsen M, et al. Benchmarking predictions of MHC class I restricted T cell epitopes in a comprehensively studied model system. PLOS Comput Biol. 2020 May 26;16(5):e1007757.

106. Bhattacharya R, Sivakumar A, Tokheim C, Guthrie VB, Anagnostou V, Velculescu VE, et al. Evaluation of machine learning methods to predict peptide binding to MHC Class I proteins [Internet]. bioRxiv; 2017 [cited 2022 Sep 8]. p. 154757. Available from: https://www.biorxiv.org/content/10.1101/154757v2

107. NetMHC pan 4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data | bioRxiv [Internet]. [cited 2019 Nov 5]. Available from: https://www.biorxiv.org/content/10.1101/149518v1.full

108. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W509-512.

109. NNAlign_MA; MHC Peptidome Deconvolution for Accurate MHC Binding Motif Characterization and Improved T-cell Epitope Predictions - PMC [Internet]. [cited 2022 Sep 8]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6885703/

110. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J Immunol. 2017 Nov 1;199(9):3360–8.

111. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. 2019 Jan 8;47(D1):D339–43.

112. Bais P, Namburi S, Gatti DM, Zhang X, Chuang JH. CloudNeo: a cloud pipeline for identifying patient-specific tumor neoantigens. Bioinforma Oxf Engl. 2017 Oct 1;33(19):3110–2.

113. Bjerregaard AM, Nielsen M, Hadrup SR, Szallasi Z, Eklund AC. MuPeXI: prediction of neo-epitopes from tumor sequencing data. Cancer Immunol Immunother CII. 2017 Sep;66(9):1123–30.

114. Rajasagi M, Shukla SA, Fritsch EF, Keskin DB, DeLuca D, Carmona E, et al. Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia. Blood. 2014 Jul 17;124(3):453–62.

115. Naito T, Okada Y. HLA imputation and its application to genetic and molecular fine-mapping of the MHC region in autoimmune diseases. Semin Immunopathol. 2022 Jan 1;44(1):15–28.

116. Fleri W, Paul S, Dhanda SK, Mahajan S, Xu X, Peters B, et al. The Immune Epitope Database and Analysis Resource in Epitope Discovery and Synthetic Vaccine Design. Front Immunol [Internet]. 2017 [cited 2022 Sep 9];8. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2017.00278

117. Gubin MM, Zhang X, Schuster H, Caron E, Ward JP, Noguchi T, et al. Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. Nature. 2014 Nov;515(7528):577–81.

118. Sahin U, Derhovanessian E, Miller M, Kloke BP, Simon P, Löwer M, et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. Nature. 2017 Jul;547(7662):222–6.

119. Yadav M, Jhunjhunwala S, Phung QT, Lupardus P, Tanguay J, Bumbaca S, et al. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. Nature. 2014 Nov;515(7528):572–6.

120. Flower DR, Davies MN, Doytchinova IA. Identification of Candidate Vaccine Antigens In Silico. Immunomic Discov Adjuv Candidate Subunit Vaccines. 2012 Sep 28;5:39–71.

121. Hu W, He M, Li L. HLA class I restricted epitopes prediction of common tumor antigens in white and East Asian populations: Implication on antigen selection for cancer vaccine design. PLoS ONE. 2020 Feb 27;15(2):e0229327.

122. Geneugelijk K, Thus KA, Spierings E. Predicting Alloreactivity in Transplantation. J Immunol Res. 2014 Apr 28;2014:e159479.

123. Mishto M, Mansurkhodzhaev A, Rodriguez-Calvo T, Liepe J. Potential Mimicry of Viral and Pancreatic β Cell Antigens Through Non-Spliced and cis-Spliced Zwitter Epitope Candidates in Type 1 Diabetes. Front Immunol [Internet]. 2021 [cited 2022 Sep 29];12. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2021.656451

124. Dearlove B, Lewitus E, Bai H, Li Y, Reeves DB, Joyce MG, et al. A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. Proc Natl Acad Sci. 2020 Sep 22;117(38):23652–62.

125. Kumar A, Rathi E, Kini SG. Computational design of a broad-spectrum multi-epitope vaccine candidate against seven strains of human coronaviruses. 3 Biotech. 2022 Aug 23;12(9):240.

126. Motozono C, Toyoda M, Zahradnik J, Saito A, Nasser H, Tan TS, et al. SARS-CoV-2 spike L452R variant evades cellular immunity and increases infectivity. Cell Host Microbe. 2021 Jul 14;29(7):1124-1136.e11.

127. Shomuradova AS, Vagida MS, Sheetikov SA, Zornikova KV, Kiryukhin D, Titov A, et al. SARS-CoV-2 Epitopes Are Recognized by a Public and Diverse Repertoire of Human T Cell Receptors. Immunity. 2020 Dec 15;53(6):1245-1257.e5.

128. Mahapatra SR, Sahoo S, Dehury B, Raina V, Patro S, Misra N, et al. Designing an efficient multi-epitope vaccine displaying interactions with diverse HLA molecules for an efficient humoral and cellular immune response to prevent COVID-19 infection. Expert Rev Vaccines. 2020 Sep 1;19(9):871–85.

129. Brister JR, Ako-Adjei D, Bao Y, Blinkova O. NCBI viral genomes resource. Nucleic Acids Res. 2015 Jan;43(Database issue):D571-577.

130. 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 [Internet]. 2020 Apr 17 [cited 2022 Jul 19]; Available from: https://journals.asm.org/doi/10.1128/JVI.00510-20

131. Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV) [Internet]. [cited 2022 Sep 4]. Available from: https://www.who.int/news/item/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov)

132. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. 2020 Feb 20;382(8):727–33.

133. Ritchie H, Mathieu E, Rodés-Guirao L, Appel C, Giattino C, Ortiz-Ospina E, et al. Coronavirus Pandemic (COVID-19). Our World Data [Internet]. 2020 Mar 5 [cited 2022 Sep 11]; Available from: https://ourworldindata.org/coronavirus

134. Caramelo F, Ferreira N, Oliveiros B. Estimation of risk factors for COVID-19 mortality-preliminary results [Internet]. 2020. Available from: medRxiv.

135. Jain V, Yuan JM. Systematic review and meta-analysis of predictive symptoms and comorbidities for severe COVID-19 infection.

136. Yang X, Yu Y, Xu J, Shu H, Xia J, Liu H, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir Med. 2020 May;8(5):475–81.

137. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China. JAMA. 2020 Mar 17;323(11):1061–9.

138. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet Lond Engl. 2020 Mar 28;395(10229):1054–62.

139. Guan W jie, Liang W hua, Zhao Y, Liang H rui, Chen Z sheng, Li Y min, et al. Comorbidity and its impact on 1590 patients with COVID-19 in China: a nationwide analysis. Eur Respir J. 2020 May 14;55(5):2000547.

140. Novel Coronavirus Pneumonia Emergency Response Epidemiology Team. Zhonghua Liu Xing Bing Xue Za Zhi. 2020;41:145–51.

141. Lau JTF, Lau M, Kim JH, Wong E, Tsui HY, Tsang T, et al. In: Probable Secondary Infections in Households of SARS Patients in Hong Kong Emerging Infectious Diseases. 2004.

142. Denison MR. Severe acute respiratory syndrome coronavirus pathogenesis, disease and vaccines: an update. Pediatr Infect J. 2004;23:S207–14.

143. Thabet F, Chehab M, Bafaqih H, AlMohaimeed S. Middle East respiratory syndrome coronavirus in children. Saudi Med J. 2015;

144. Al-Tawfiq JA, Kattan RF, Memish ZA. Middle East respiratory syndrome coronavirus disease is rare in children: An update from Saudi Arabia. World J Clin Pediatr. 2016;5:391–6.

145. Cao Q, Chen YC, Chen CL, Chiu CH. SARS-CoV-2 infection in children: Transmission dynamics and clinical characteristics. J Formos Med Assoc. 2020;

146. Tang A, Xu W, Shen M, Chen P, Li G, Liu Y, et al. A retrospective study of the clinical characteristics of COVID-19 infection in 26 children.

147. Lu X, Zhang L, Du H, Zhang J, Li YY, Qu J, et al. 2020.

148. Eastin C, Eastin T. Epidemiological characteristics of 2143 pediatric patients with 2019 coronavirus disease in China. J Emerg Med. 2020 Apr;58(4):712–3.

149. Ashour HM, Elkhatib WF, Rahman MM, Elshabrawy HA. Insights into the Recent 2019 Novel Coronavirus (SARS-CoV-2) in Light of Past Human Coronavirus Outbreaks. Pathogens. 2020;9.

150. Yang Y, Peng F, Wang R, Guan K, Jiang T, Xu G, et al. The deadly coronaviruses: The 2003 SARS pandemic and the 2020 novel coronavirus epidemic in China. J Autoimmun. 2020;

151. Ge Y, Tian T, Huang S, Wan F, Li J, Li S, et al. A data-driven drug repositioning framework discovered a potential therapeutic agent targeting COVID-19.

152. Ceraolo C, Giorgi FM. Genomic variance of the 2019-nCoV coronavirus. J Med Virol. 2020;92:522–8.

153. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 2020;395:565–74.

154. Prompetchara E, Ketloy C, Palaga T. Immune responses in COVID-19 and potential vaccines: Lessons learned from SARS and MERS epidemic. Asian Pac J Allergy Immunol. 2020;

155. Zheng Z, Monteil VM, Maurer-Stroh S, Yew CW, Leong C, Mohd-Ismail NK, et al. Monoclonal antibodies for the S2 subunit of spike of SARS-CoV cross-react with the newly-emerged SARS-CoV-2.

156. Yang YY, C S, J L, J Y, M Y, F W, et al. Exuberant elevation of IP-10.

157. Ahmed SF, Quadeer AA, McKay MR. Preliminary Identification of Potential Vaccine Targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies. Viruses; 2020.

158. Li G, Fan Y, Lai Y, Han T, Li Z, Zhou P, et al. Coronavirus infections and immune responses. J Med Virol. 2020;92:424–32.

159. Lokugamage KG, Schindewolf C, Menachery VD. SARS-CoV-2 sensitive to type I interferon pretreatment.

160. Wang C, Li W, Drabek D, Okba NMA, Haperen R, M AD, et al. A human monoclonal antibody blocking SARS-CoV-2 infection.

161. Lv H, Wu NC, Tsang OTY, Yuan M, P RA, Leung WS, et al. Cross-reactive antibody response between SARS-CoV-2 and SARS-CoV infections.

162. Yuan M, Wu NC, Zhu X, Lee CCD, So RTY, Lv H, et al. A highly conserved cryptic epitope in the receptor-binding domains of SARS-CoV-2 and SARS-CoV.

163. Tetro JA. Is COVID-19 receiving ADE from other coronaviruses? Microbes Infect. 2020;22:72–3.

164. Breadth of concomitant immune responses underpinning viral clearance and patient recovery in a non-severe case of COVID-19.

165. Zhu J, Kim J, Xiao X, Wang Y, Luo D, Chen R, et al. Profiling the Immune Vulnerability Landscape of the 2019 Novel Coronavirus.

166. Ni L, Ye F, Cheng ML, Feng Y, Deng YQ, Zhao H, et al. Detection of SARS-CoV-2-Specific Humoral and Cellular Immunity in COVID-19 Convalescent Individuals. Immunity. 2020 Jun 16;52(6):971-977.e3.

167. Chen G, Wu D, Guo W, Cao Y, Huang D, Wang H, et al. Clinical and immunological features of severe and moderate coronavirus disease 2019. J Clin Invest. 2020 May 1;130(5):2620–9.

168. Designing of a next generation multiepitope based vaccine (MEV) against SARS-COV-2: Immunoinformatics and in silico approaches.

169. Fast E, Chen B. Potential T-cell and B-cell Epitopes of 2019-nCoV.

170. Abdelmageed MI, Abdelmoneim AH, Mustafa MI, Elfadol NM, Murshed NS, Shantier SW, et al. Design of multi epitope-based peptide vaccine against E protein of human COVID-19: An immunoinformatics approach.

171. Baruah V, Bose S. Immunoinformatics-aided identification of T cell and B cell epitopes in the surface glycoprotein of 2019-nCoV. J Med Virol. 2020;92:495–500.

172. Robson B. Computers and viral diseases. Preliminary bioinformatics studies on the design of a synthetic vaccine and a preventative peptidomimetic antagonist against the SARS-CoV-2 (2019-nCoV, COVID-19) coronavirus. Comput Biol Med. 2020;

173. Dhama K, Sharun K, Tiwari R, Dadar M, Malik YS, Singh KP, et al. COVID-19, an emerging coronavirus infection: advances and prospects in designing and developing vaccines, immunotherapeutics, and therapeutics. Hum Vaccin Immunother. 2020;1–7.

174. Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. Cell Host Microbe; 2020.

175. Campbell KM, Steiner G, Wells DK, Ribas A, Kalbasi A. Prediction of SARS-CoV-2 epitopes across 9360 HLA class I alleles.

176. González-Galarza FF, Takeshita LYC, Santos EJM, Kempson F, Maia MHT, Silva ALS, et al. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. Nucleic Acids Research; 2015.

177. Sette A, Grey H, Oseroff C, Peters B, Moutaftsi M, Crotty S, et al. Definition of epitopes and antigens recognized by vaccinia specific immune responses: their conservation in variola virus sequences, and use as a model system to study complex pathogens. Vaccine. 2009;27 Suppl 6:G21–6.

178. Gálvez J, Gálvez JJ, García-Peñarrubia P. Is TCR/pMHC Affinity a Good Estimate of the T-cell Response? An Answer Based on Predictions From 12 Phenotypic Models. Front Immunol. 2019;

179. Zehn D, Lee SY, Bevan MJ. Complete but curtailed T-cell response to very low-affinity antigen. Nature; 2009.

180. Sibener LV, Fernandes RA, Kolawole EM, Carbone CB, Liu F, McAffee D, et al. Isolation of a Structural Mechanism for Uncoupling T Cell Receptor Signaling from Peptide-MHC Binding. Cell. 2018 Jul 26;174(3):672-687.e27.

181. Vatti A, Monsalve DM, Pacheco Y, Chang C, Anaya JM, Eric Gershwin M. Original antigenic sin: A comprehensive review. J Autoimmun. 2017;

182. Park MS, Kim JI, Park S, Lee I, Park MS. Original Antigenic Sin Response to RNA Viruses and Antiviral Immunity. Immune Network; 2016.

183. Singh RAK, Rodgers B JR, M.A. The Role of T Cell Antagonism and Original Antigenic Sin in Genetic Immunization. J Immunol. 2002;

184. Weiskopf D, Angelo MA, Azeredo EL, Sidney J, Greenbaum JA, Fernando AN, et al. Comprehensive analysis of dengue virus-specific responses supports an HLA-linked protective role for CD8 T cells. In: Proceedings of the National Academy of Sciences. 2013.

185. Klenerman P, Zinkernagel RM. Original antigenic sin impairs cytotoxic T lymphocyte responses to viruses bearing variant epitopes. Nature; 1998.

186. Rothman AL. Immunity to dengue virus: a tale of original antigenic sin and tropical cytokine storms. Nat Rev Immunol. 2011;

187. Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L, et al. Genomic Diversity of Severe Acute Respiratory Syndrome-Coronavirus 2 in Patients With Coronavirus Disease 2019. Clin Infect Dis Off Publ Infect Dis Soc Am. 2020 Jul 28;71(15):713–20.

188. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev. 2020;

189. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell. 2020;

190. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016 Jan 4;44(D1):D733-745.

191. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011 Jan;7(1):539.

192. Baranov PV, Henderson CM, Anderson CB, Gesteland RF, Atkins JF, Howard MT. Programmed ribosomal frameshifting in decoding the SARS-CoV genome. Virology. 2005;332:498–510.

193. Taiaroa G, Rawlinson D, Featherstone L, Pitt M, Caly L, Druce J, et al. Direct RNA sequencing and early evolution of SARS-CoV-2. 2020.

194. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009 May 1;25(9):1189–91.

195. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. Cell Syst. 2018 Jul 25;7(1):129-132.e4.

196. Nielsen M, Lundegaard C, Lund O, Keşmir C. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. Immunogenetics. 2005;57:33–41.

197. South A. rworldmap : a new R package for mapping global data. R J. 2011;3(1):35.

198. Nguyen A, Yusufali T, Hollenbach JA, Nellore A, Thompson RF. Minimal observed impact of HLA genotype on hospitalization and severity of SARS-CoV-2 infection. HLA. 2022 Jun;99(6):607–13.

199. Li X, Xu S, Yu M, Wang K, Tao Y, Zhou Y, et al. Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. J Allergy Clin Immunol. 2020 Jul 1;146(1):110–8.

200. Wolff D, Nee S, Hickey NS, Marschollek M. Risk factors for Covid-19 severity and fatality: a structured literature review. Infection. 2021 Feb 1;49(1):15–28.

201. Kompaniyets L. Body Mass Index and Risk for COVID-19–Related Hospitalization, Intensive Care Unit Admission, Invasive Mechanical Ventilation, and Death — United States, March–December 2020. MMWR Morb Mortal Wkly Rep [Internet]. 2021 [cited 2021 Sep 16];70. Available from: https://www.cdc.gov/mmwr/volumes/70/wr/mm7010e4.htm

202. Zernich D, Purcell AW, Macdonald WA, Kjer-Nielsen L, Ely LK, Laham N, et al. Natural HLA Class I Polymorphism Controls the Pathway of Antigen Presentation and Susceptibility to Viral Evasion. J Exp Med. 2004 Jun 28;200(1):13–24.

203. Couture A, Garnier A, Docagne F, Boyer O, Vivien D, Le-Mauff B, et al. HLA-Class II Artificial Antigen Presenting Cells in CD4+ T Cell-Based Immunotherapy. Front Immunol. 2019;10:1081.

204. Crux NB, Elahi S. Human Leukocyte Antigen (HLA) and Immune Regulation: How Do Classical and Non-Classical HLA Alleles Modulate Immune Response to Human Immunodeficiency Virus and Hepatitis C Virus Infections? Front Immunol. 2017;8:832.

205. Warren RL, Birol I. HLA alleles measured from COVID-19 patient transcriptomes reveal associations with disease prognosis in a New York cohort. PeerJ. 2021 Oct 15;9:e12368.

206. Roberts GHL, Park DS, Coignet MV, McCurdy SR, Knight SC, Partha R, et al. AncestryDNA COVID-19 Host Genetic Study Identifies Three Novel Loci [Internet]. 2020 Oct [cited 2021 Sep 16] p. 2020.10.06.20205864. Available from: https://www.medrxiv.org/content/10.1101/2020.10.06.20205864v1

207. 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. Gallagher T, editor. J Virol. 2020 Jun 16;94(13):e00510-20.

208. Li Y, Ke Y, Xia X, Wang Y, Cheng F, Liu X, et al. Genome-wide association study of COVID-19 severity among the Chinese population. Cell Discov. 2021 Aug 31;7(1):1–16.

209. Ellinghaus D, Degenhardt F, Bujanda L, Buti M, Albillos A, Invernizzi P, et al. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. N Engl J Med. 2020 Jun 17;NEJMoa2020283.

210. Velavan TP, Pallerla SR, Rüter J, Augustin Y, Kremsner PG, Krishna S, et al. Host genetic factors determining COVID-19 susceptibility and severity. EBioMedicine [Internet]. 2021 Oct 1 [cited 2021 Nov 15];72. Available from: https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(21)00422-9/fulltext

211. Langton DJ, Bourke SC, Lie BA, Reiff G, Natu S, Darlay R, et al. The influence of HLA genotype on the severity of COVID-19 infection. HLA. 2021;98(1):14–22.

212. Pisanti S, Deelen J, Gallina AM, Caputo M, Citro M, Abate M, et al. Correlation of the two most frequent HLA haplotypes in the Italian population to the differential regional incidence of Covid-19. J Transl Med. 2020 Sep 15;18(1):352.

213. Migliorini F, Torsiello E, Spiezia F, Oliva F, Tingart M, Maffulli N. Association between HLA genotypes and COVID-19 susceptibility, severity and progression: a comprehensive review of the literature. Eur J Med Res. 2021 Aug 3;26(1):84.

214. Iturrieta-Zuazo I, Rita CG, García-Soidán A, de Malet Pintos-Fonseca A, Alonso-Alarcón N, Pariente-Rodríguez R, et al. Possible role of HLA class-I genotype in SARS-CoV-2 infection and progression: A pilot study in a cohort of Covid-19 Spanish patients. Clin Immunol Orlando Fla. 2020 Oct;219:108572.

215. Ancestry COVID-19 study - EGA European Genome-Phenome Archive [Internet]. [cited 2021 Nov 12]. Available from: https://ega-archive.org/studies/EGAS00001004716

216. Pappas DJ, Marin W, Hollenbach JA, Mack SJ. Bridging ImmunoGenomic Data Analysis Workflow Gaps (BIGDAWG): An integrated case-control analysis pipeline. Hum Immunol. 2016 Mar 1;77(3):283–7.

217. Shkurnikov M, Nersisyan S, Jankevic T, Galatenko A, Gordeev I, Vechorko V, et al. Association of HLA Class I Genotypes With Severity of Coronavirus Disease-19. Front Immunol. 2021;12:423.

218. Ishii T. Human Leukocyte Antigen (HLA) Class I Susceptible Alleles Against COVID-19 Increase Both Infection and Severity Rate. Cureus [Internet]. 2020 Dec 23 [cited 2021 Oct 21];12(12). Available from: https://www.cureus.com/articles/35178-human-leukocyte-antigen-hla-class-i-susceptible-alleles-against-covid-19-increase-both-infection-and-severity-rate

219. Toyoshima Y, Nemoto K, Matsumoto S, Nakamura Y, Kiyotani K. SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. J Hum Genet. 2020;65(12):1075–82.

220. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nat Protoc. 2010 Sep;5(9):1564–73.

221. Zheng X, Shen J, Cox C, Wakefield JC, Ehm MG, Nelson MR, et al. HIBAG—HLA genotype imputation with attribute bagging. Pharmacogenomics J. 2014 Apr;14(2):192–200.

222. Marshall JC, Murthy S, Diaz J, Adhikari NK, Angus DC, Arabi YM, et al. A minimal common outcome measure set for COVID-19 clinical research. Lancet Infect Dis. 2020 Aug;20(8):e192–7.

223. Knight SC, McCurdy SR, Rhead B, Coignet MV, Park DS, Roberts GHL, et al. COVID-19 susceptibility and severity risks in a survey of over 500,000 individuals [Internet]. 2021 Jan [cited 2022 Jan 10] p. 2020.10.08.20209593. Available from: https://www.medrxiv.org/content/10.1101/2020.10.08.20209593v3

224. Dendrou CA, Petersen J, Rossjohn J, Fugger L. HLA variation and disease. Nat Rev Immunol. 2018 May;18(5):325–39.

225. Meyer D, C. Aguiar VR, Bitarello BD, C. Brandt DY, Nunes K. A genomic perspective on HLA evolution. Immunogenetics. 2018;70(1):5–27.

226. Bihl F, Frahm N, Giammarino LD, Sidney J, John M, Yusim K, et al. Impact of HLA-B Alleles, Epitope Binding Affinity, Functional Avidity, and Viral Coinfection on the Immunodominance of Virus-Specific CTL Responses. J Immunol. 2006 Apr 1;176(7):4094–101.

227. Berger CT, Carlson JM, Brumme CJ, Hartman KL, Brumme ZL, Henry LM, et al. Viral adaptation to immune selection pressure by HLA class I–restricted CTL responses targeting epitopes in HIV frameshift sequences. J Exp Med. 2010 Jan 18;207(1):61–75.

228. Schellens IM, Meiring HD, Hoof I, Spijkers SN, Poelen MCM, van Gaans-van den Brink JAM, et al. Measles Virus Epitope Presentation by HLA: Novel Insights into Epitope Selection, Dominance, and Microvariation. Front Immunol [Internet]. 2015 [cited 2019 Nov 15];6. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2015.00546/full

229. Kaufman J. Generalists and Specialists: A New View of How MHC Class I Molecules Fight Infectious Pathogens. Trends Immunol. 2018 May 1;39(5):367–79.

230. Hu Z, Ott PA, Wu CJ. Towards personalized, tumour-specific, therapeutic vaccines for cancer. Nat Rev Immunol. 2018 Mar;18(3):168–82.

231. Nelde A, Maringer Y, Bilich T, Salih HR, Roerden M, Heitmann JS, et al. Immunopeptidomics-Guided Warehouse Design for Peptide-Based Immunotherapy in Chronic Lymphocytic Leukemia. Front Immunol [Internet]. 2021 [cited 2022 Sep 30];12. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2021.705974

232. Terasaki M, Shibui S, Narita Y, Fujimaki T, Aoki T, Kajiwara K, et al. Phase I trial of a personalized peptide vaccine for patients positive for human leukocyte antigen--A24 with recurrent or progressive glioblastoma multiforme. J Clin Oncol Off J Am Soc Clin Oncol. 2011 Jan 20;29(3):337–44.

233. Kibe S, Yutani S, Motoyama S, Nomura T, Tanaka N, Kawahara A, et al. Phase II study of personalized peptide vaccination for previously treated advanced colorectal cancer. Cancer Immunol Res. 2014 Dec;2(12):1154–62.

234. Shao XM, Bhattacharya R, Huang J, Sivakumar IKA, Tokheim C, Zheng L, et al. High-Throughput Prediction of MHC Class I and II Neoantigens with MHCnuggets. Cancer Immunol Res. 2020;8:396–408.

235. Pavlos R, McKinnon EJ, Ostrov DA, Peters B, Buus S, Koelle D, et al. Shared peptide binding of HLA Class I and II alleles associate with cutaneous nevirapine hypersensitivity and identify novel risk alleles. Sci Rep. 2017 Aug 17;7(1):8653.

236. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J, et al. Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. Immunity. 2017 Feb;46(2):315–26.

237. Roche PA, Furuta K. The ins and outs of MHC class II-mediated antigen processing and presentation. Nat Rev Immunol. 2015 Apr;15(4):203–16.

238. pepsickle rapidly and accurately predicts proteasomal cleavage sites for improved neoantigen identification | Bioinformatics | Oxford Academic [Internet]. [cited 2022 Oct 2]. Available from: https://academic.oup.com/bioinformatics/article/37/21/3723/6363787

239. Ritz U, Seliger B. The Transporter Associated With Antigen Processing (TAP): Structural Integrity, Expression, Function, and Its Clinical Relevance. Mol Med. 2001 Mar;7(3):149–58.

240. López de Castro JA. How ERAP1 and ERAP2 Shape the Peptidomes of Disease-Associated MHC-I Proteins. Front Immunol [Internet]. 2018 [cited 2022 Oct 2];9. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2018.02463

241. Dincer AB, Lu Y, Schweppe DK, Oh S, Noble WS. Reducing Peptide Sequence Bias in Quantitative Mass Spectrometry Data with Machine Learning. J Proteome Res. 2022 Jul 1;21(7):1771–82.

242. Edwards NJ. Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. Mol Syst Biol. 2007 Jan;3(1):102.

243. Prakash A, Piening B, Whiteaker J, Zhang H, Shaffer SA, Martin D, et al. Assessing Bias in Experiment Design for Large Scale Mass Spectrometry-based Quantitative Proteomics*. Mol Cell Proteomics. 2007 Oct 1;6(10):1741–8.

244. Timp W, Timp G. Beyond mass spectrometry, the next step in proteomics. Sci Adv. 2020 Jan 10;6(2):eaax8978.

245. Venkatesh G, Grover A, Srinivasaraghavan G, Rao S. MHCAttnNet: predicting MHC-peptide bindings for MHC alleles classes I and II using an attention-based deep neural model. Bioinformatics. 2020 Jul;36(Suppl 1):i399–406.

246. Bhattacharya R, Sivakumar A, Tokheim C, Guthrie VB, Anagnostou V, Velculescu VE, et al. Evaluation of machine learning methods to predict peptide binding to MHC Class I proteins. bioRxiv. 2017 Jul 27;154757.

247. Lide D. CRC handbook of chemistry and physics, 1992-1993 : a ready-reference book of chemical and physical data [Internet]. 1992 [cited 2022 Sep 4]. Available from: https://www.worldcat.org/title/crc-handbook-of-chemistry-and-physics-1992-1993-a-ready-reference-book-of-chemical-and-physical-data/oclc/758080758

248. A new set of peptide-based group heat capacities for use in protein stability calculations - ScienceDirect [Internet]. [cited 2022 Sep 4]. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0022283699929522

249. Zhu C, Gao Y, Li H, Meng S, Li L, Francisco JS, et al. Characterizing hydrophobicity of amino acid side chains in a protein environment via measuring contact angle of a water nanodroplet on planar peptide network. Proc Natl Acad Sci U S A. 2016 Nov 15;113(46):12946–51.

250. Fogolari F, Corazza A, Fortuna S, Soler MA, VanSchouwen B, Brancolini G, et al. Distance-Based Configurational Entropy of Proteins from Molecular Dynamics Simulations. PloS One. 2015;10(7):e0132356.

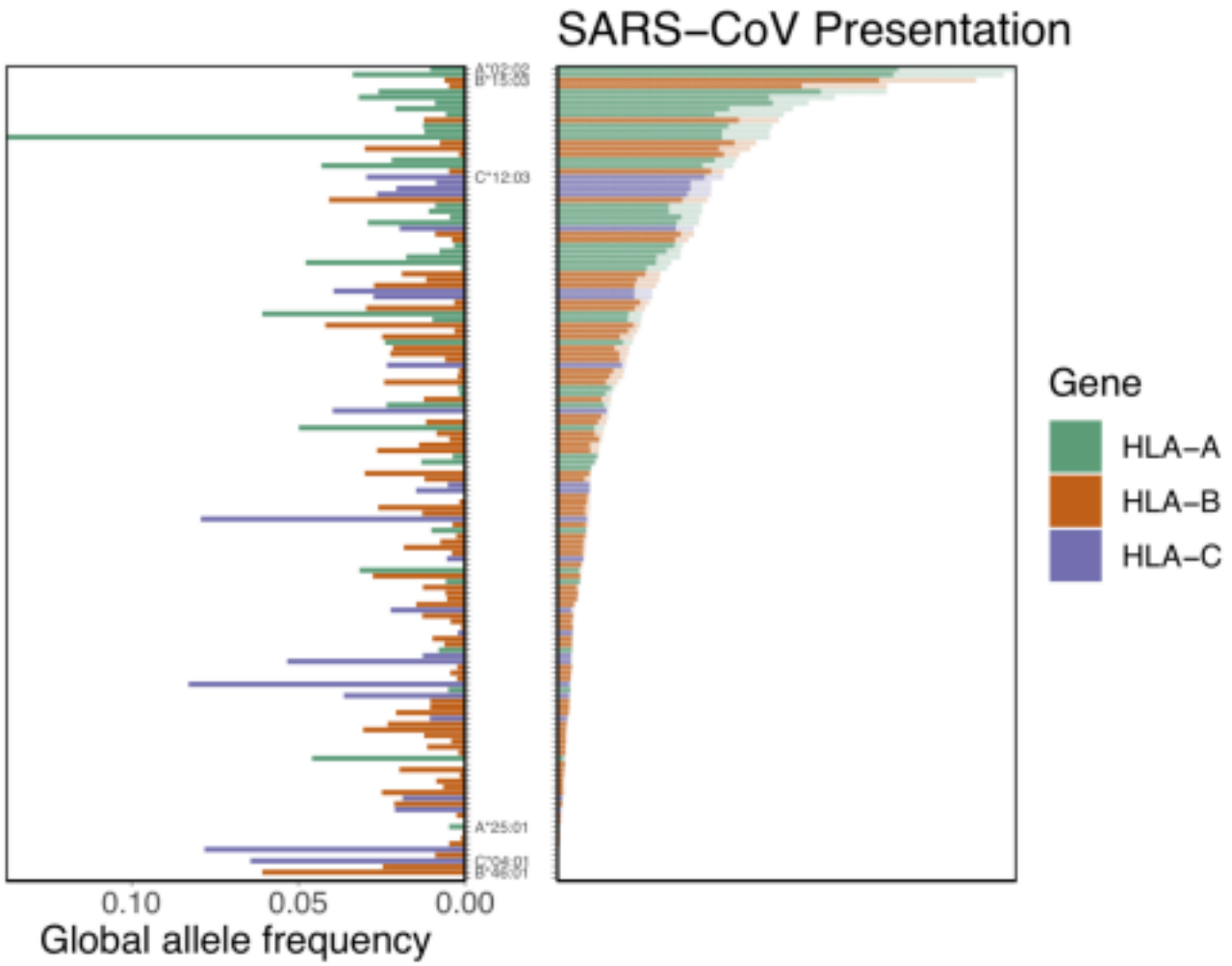
251. Kaluzny S original by DWSR port by AG adopted to recent SP by S. ash: David Scott's ASH Routines [Internet]. 2015 [cited 2022 Jul 11]. Available from: https://CRAN.R-project.org/package=ash

252. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979 Mar;86(2):420–8.

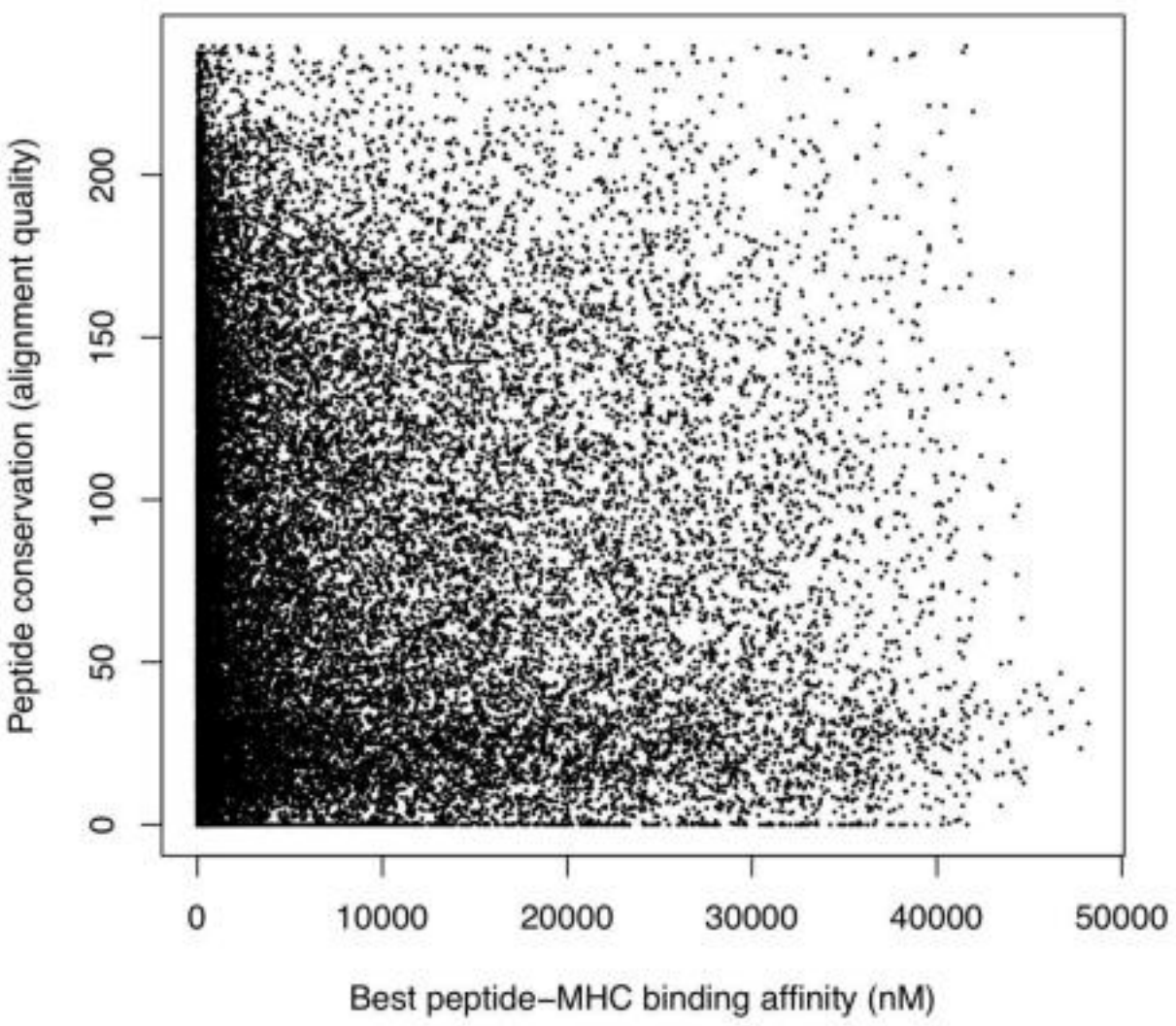
253. Ren K, Yang H, Zhao Y, Chen W, Xue M, Miao H, et al. A Robust AUC Maximization Framework With Simultaneous Outlier Detection and Feature Selection for Positive-Unlabeled Classification. IEEE Trans Neural Netw Learn Syst. 2019 Oct;30(10):3072–83.

254. Paul S, Weiskopf D, Angelo MA, Sidney J, Peters B, Sette A. HLA class I alleles are associated with peptide binding repertoires of different size, affinity and immunogenicity. J Immunol Baltim Md 1950. 2013 Dec 15;191(12):10.4049/jimmunol.1302101.

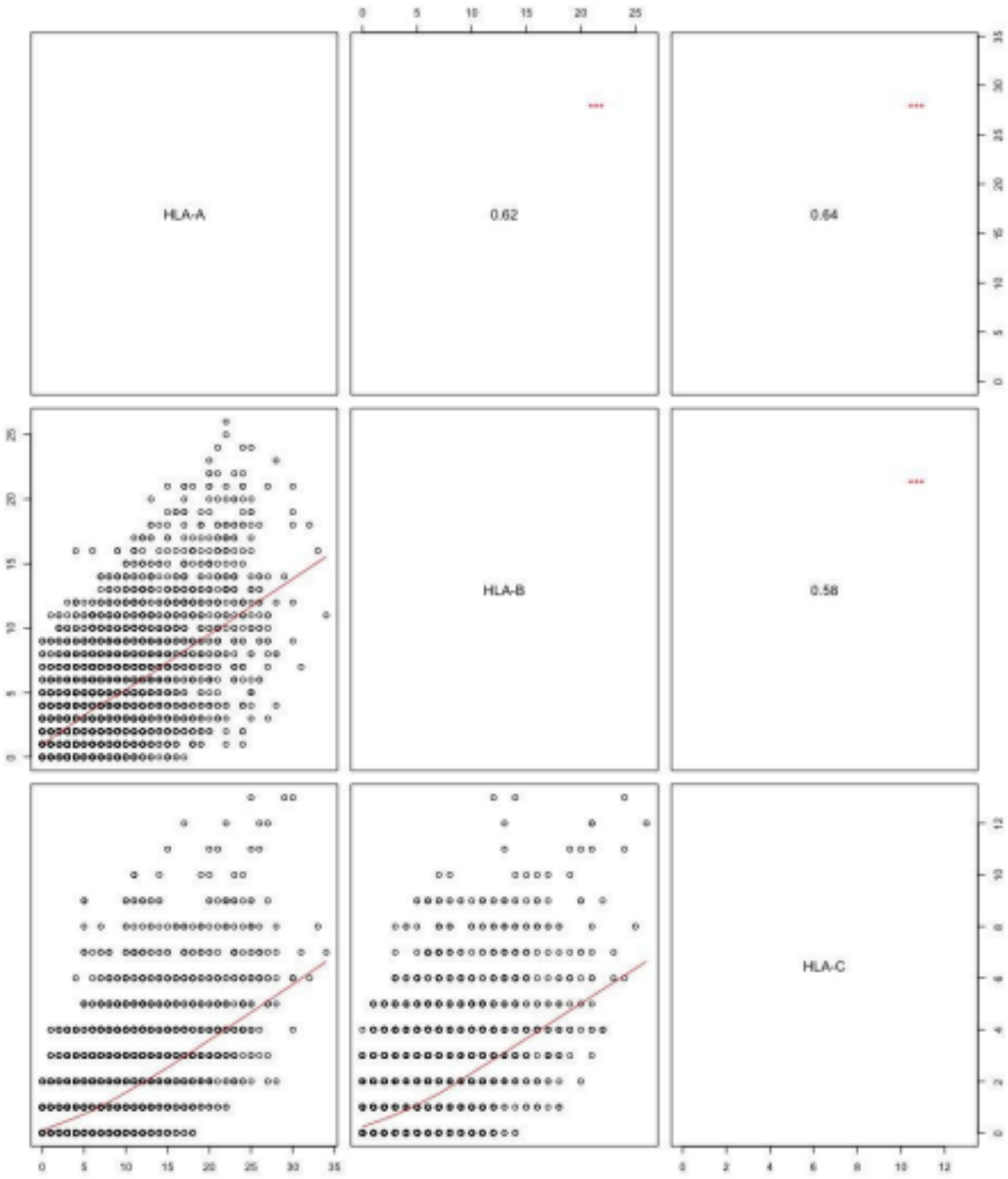
Appendix A: Supplementary Figures and Tables



Supplementary Figure 2.1: Distribution of HLA allelic presentation of 8-12mers from the SARS CoV proteome (see Supplementary Table S6). At right, the number of peptides that putatively bind to each of 145 HLA alleles is shown as a series of horizontal bars, with dark and light shading indicating the number of tightly (<50nM) and loosely (<500nM) binding peptides respectively, and with green, orange, and purple colors representing HLA-A, -B, and -C alleles, respectively. Alleles are sorted in descending order based on the number of peptides they bind (<500nM). The corresponding estimated allelic frequency in the global population is also shown (to left), with length of horizontal bar indicating absolute frequency in the population.

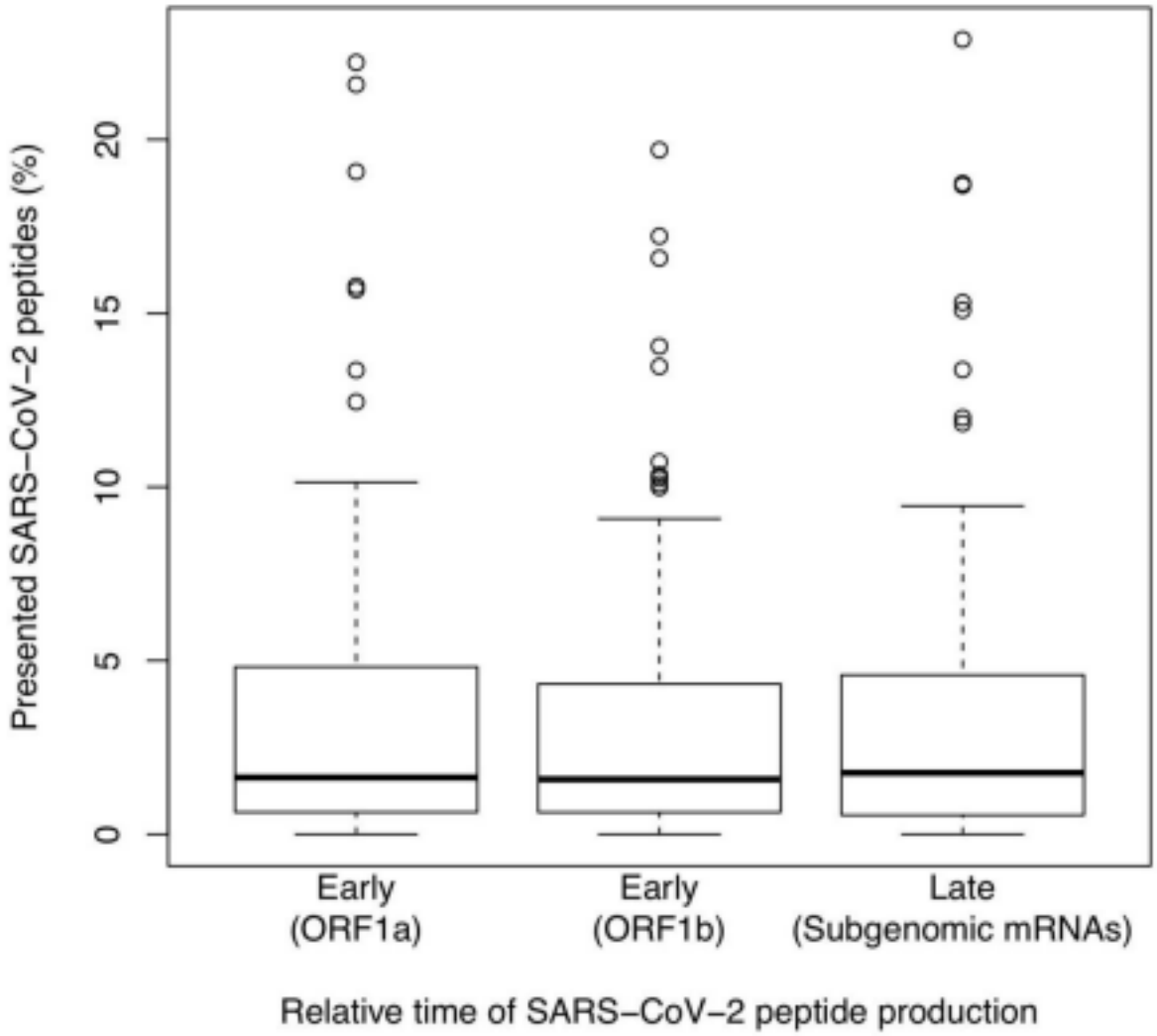


Supplementary Figure 2.2: Relationship between predicted peptide-MHC binding affinity and peptide conservation across coronaviruses. Every point represents a single unique peptide covering, together, the entirety of the SARS-CoV-2 proteome. The best predicted MHC binding affinity scores across 145 different HLA alleles are shown for each peptide along the x-axis. Sequence conservation (Clustal Omega alignment score) is shown for each peptide along the y axis.



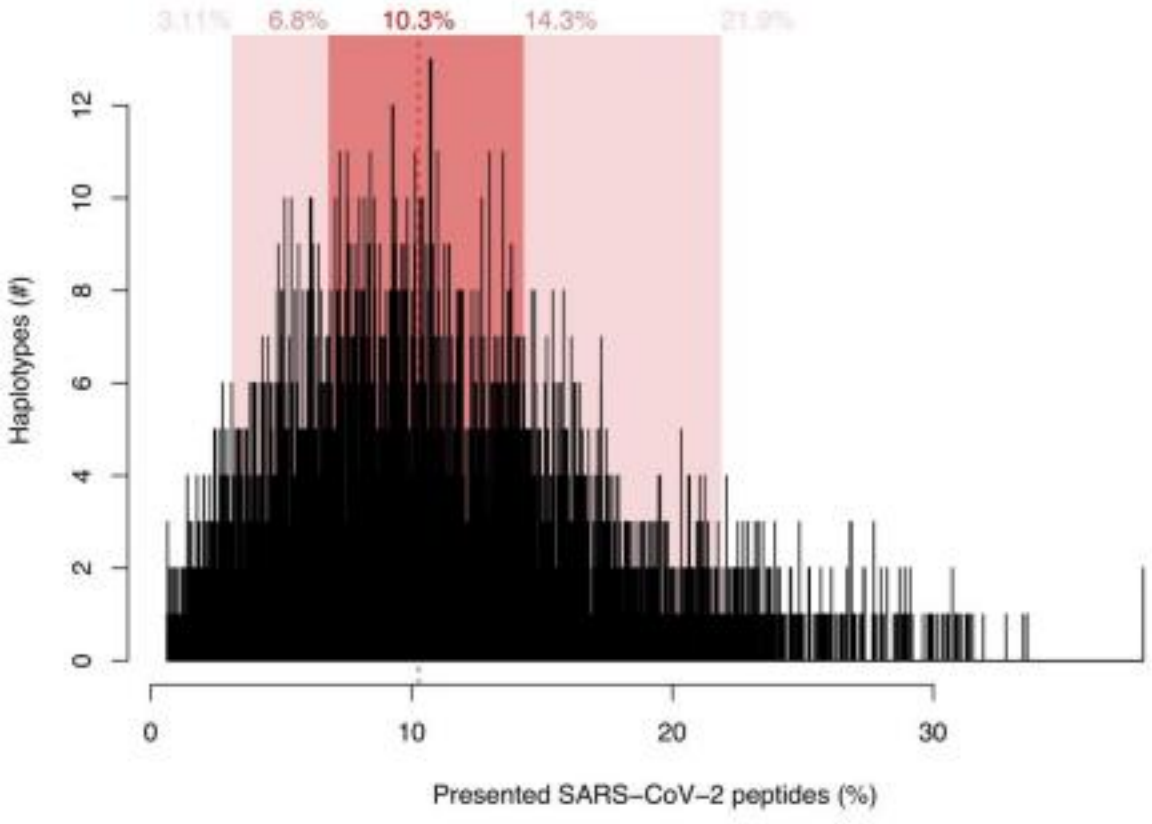
Supplementary Figure 2.3: Pairwise relationship of peptide presentation between HLA-A, -B, and -C. In the bottom left three panels, every point represents the pairwise comparison of the number of peptide-allele interactions for all position coordinates. Taken together, the position coordinates cover the entirety of the SARS-CoV-2 proteome. The top right three panels show the quantitative

*correlation scores between each pair of HLA type comparisons (*** indicates statistical significance).*



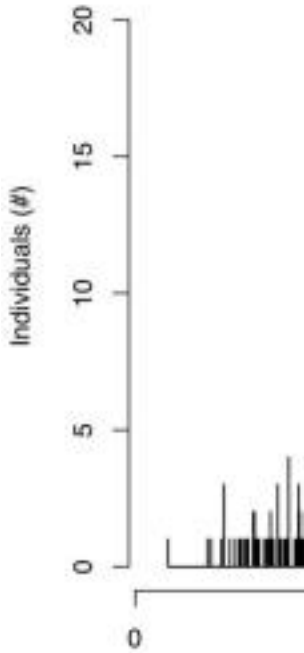
Supplementary Figure 2.4: Boxplot distributions of estimated epitope presentation across 145 HLA alleles for early and late SARS-CoV-2 peptides. Capacity for peptide presentation is shown along the

y-axis for 145 distinct HLA alleles, for three non-overlapping sets of peptides produced at different timepoints in the viral life cycle as indicated (x-axis). Y-axis percentiles are calculated as the number of peptides from the indicated compartment of the SARS-CoV-2 proteome divided by the total number of presentable 8- to 12-mer peptides from that compartment of the proteome. Dark black lines represent median values, with boxes indicating the 25% and 75% quantiles, with whiskers representing the 25% and 75% quantiles minus or plus the interquartile range, respectively, and with additional outliers shown as open circles.



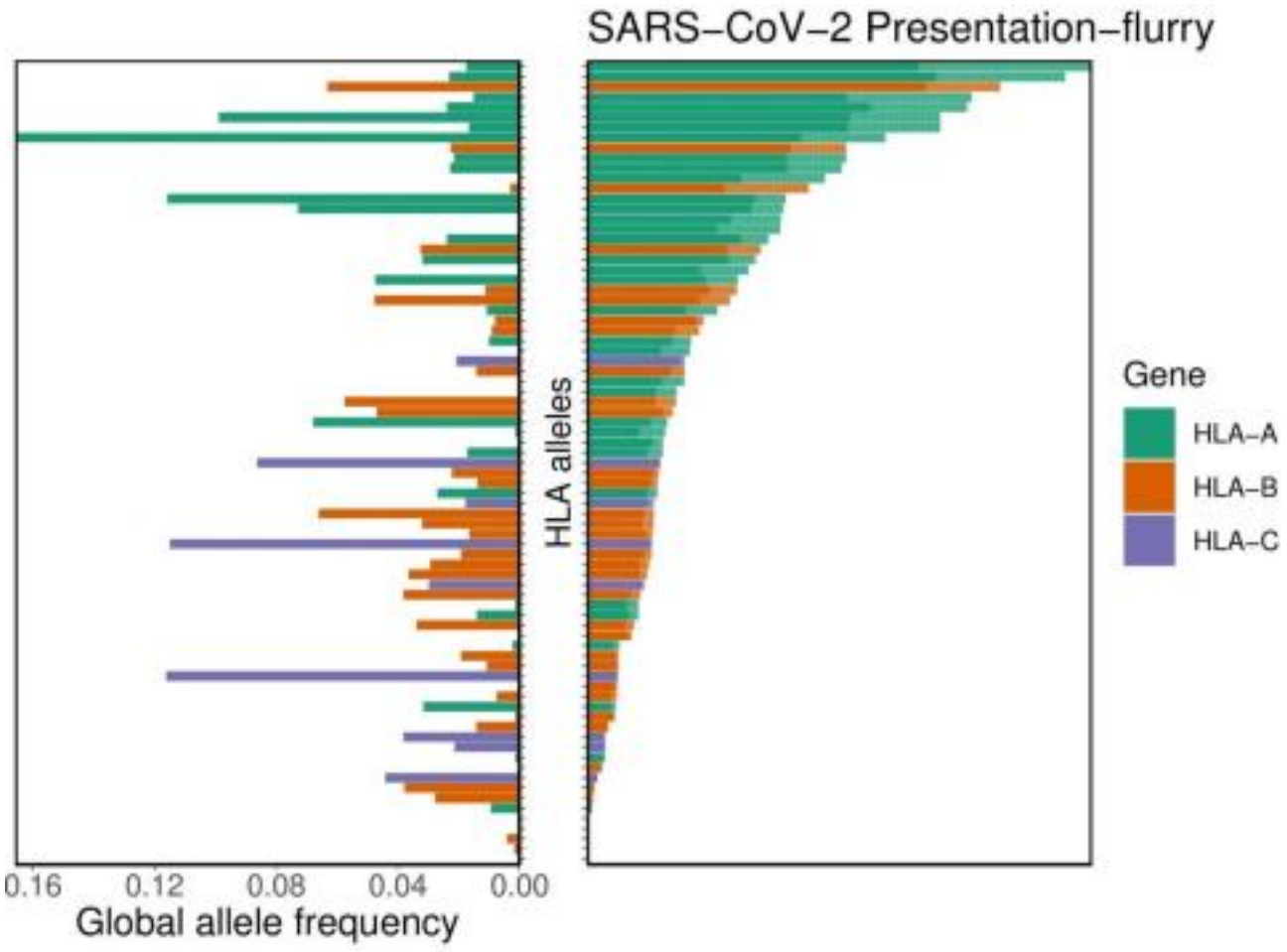
Supplementary Figure 2.5: Histogram of SARS-CoV-2 peptide presentation for 5,905 distinct

HLA-A/B/C haplotypes. Number of haplotypes are counted along the y-axis, corresponding to their individual capacity (aggregated across all their three component HLA types) to present peptides from the SARS-CoV-2 proteome, shown along the x-axis (percentile is calculated as number of unique peptides presented divided by the total number of presentable 8- to 12-mer peptides from the SARS-CoV-2 proteome). Dashed red line corresponds to the median presentation capacity, while dark and light pink highlighted regions correspond to the 25/75% and 5/95% quantiles, respectively, with numerical values shown in the upper aspect of the plotting region.



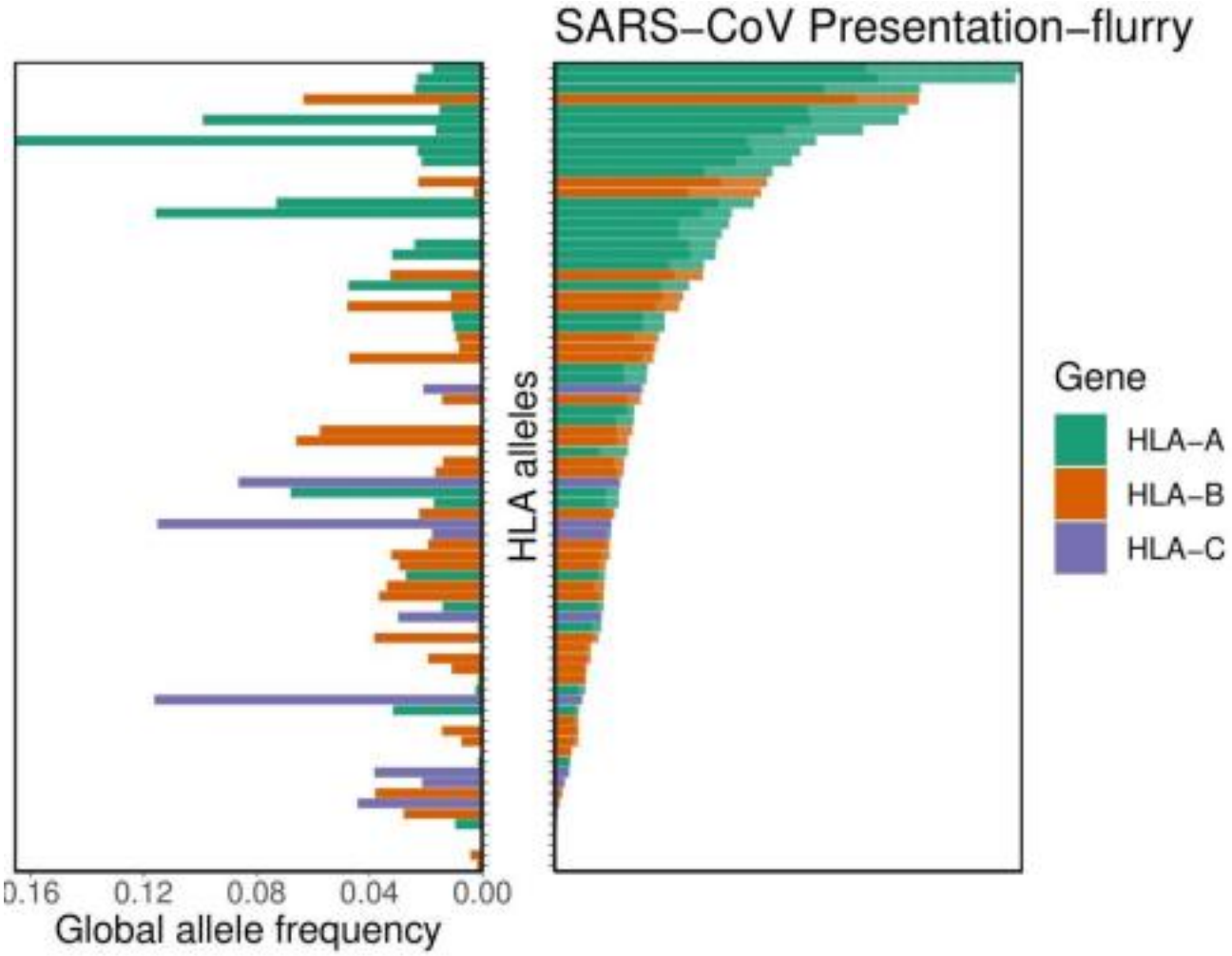
Supplementary Figure 2.6: Histogram of SARS-CoV-2 peptide presentation for 3,382 individuals'

full HLA repertoires. Individuals are counted along the y-axis, corresponding to their individual capacity (aggregated across all 6 of their HLA types) to present peptides from the SARS-CoV-2 proteome, shown along the x-axis (percentile is calculated as number of unique peptides presented divided by the total number of presentable 8- to 12-mer peptides from the SARS CoV-2 proteome). Dashed red line corresponds to the median presentation capacity, while dark and light pink highlighted regions correspond to the 25/75% and 5/95% quantiles, respectively, with numerical values shown in the upper aspect of the plotting region.



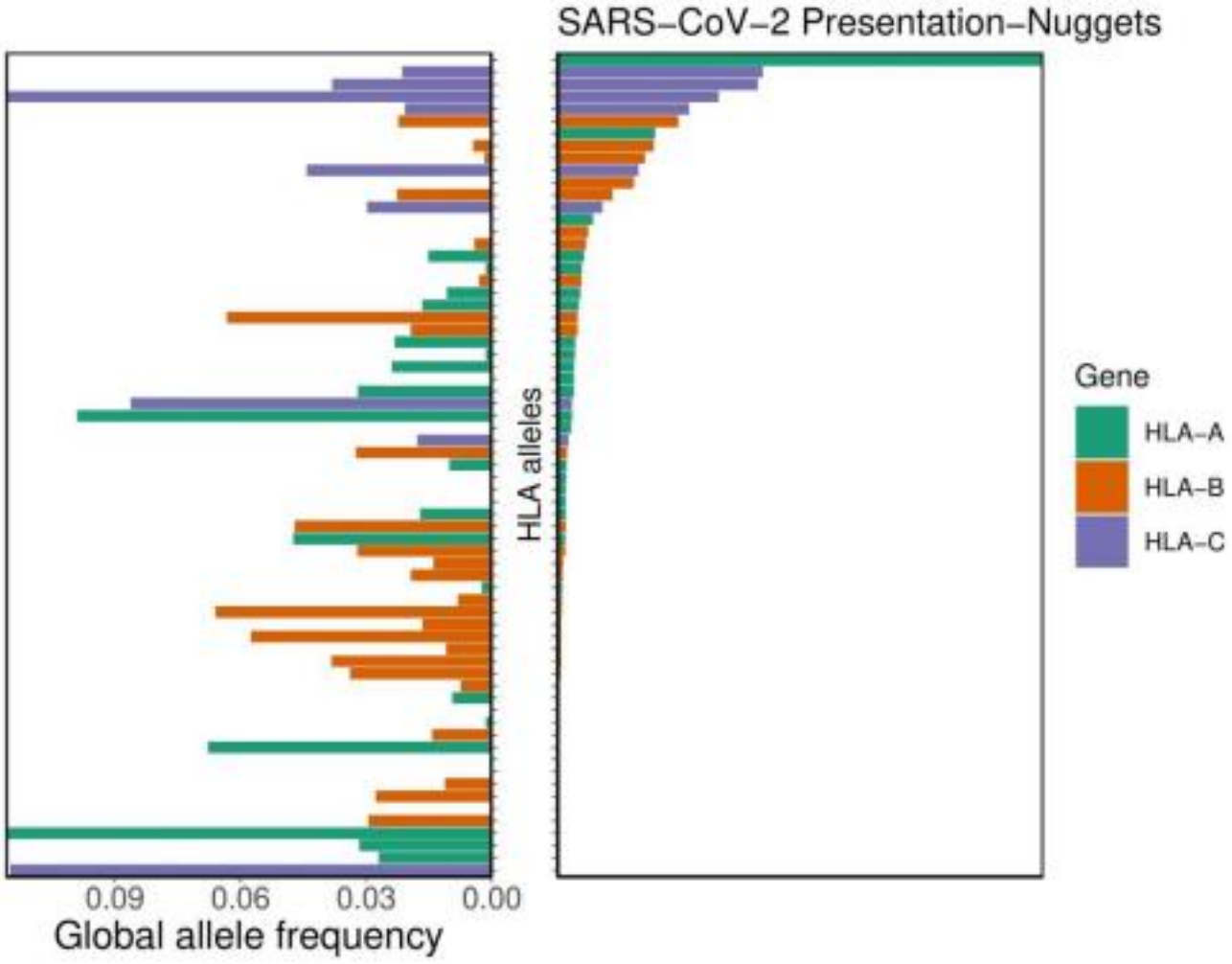
Supplementary Figure 2.7: Distribution of HLA allelic presentation of 8- to 12-mers from the SARS-CoV-2 proteome using the tool MHCflurry. At right, the number of peptides that putatively bind to

each of 66 HLA alleles is shown as a series of horizontal bars, with dark and light shading indicating the number of tightly (<50nM) and loosely (<500nM) binding peptides respectively, and with green, orange, and purple colors representing HLA-A, -B, and -C alleles, respectively. Alleles are sorted in descending order based on the number of peptides they bind (<500nM). The corresponding estimated allelic frequency in the global population is also shown (to left), with length of horizontal bar indicating absolute frequency in the population.



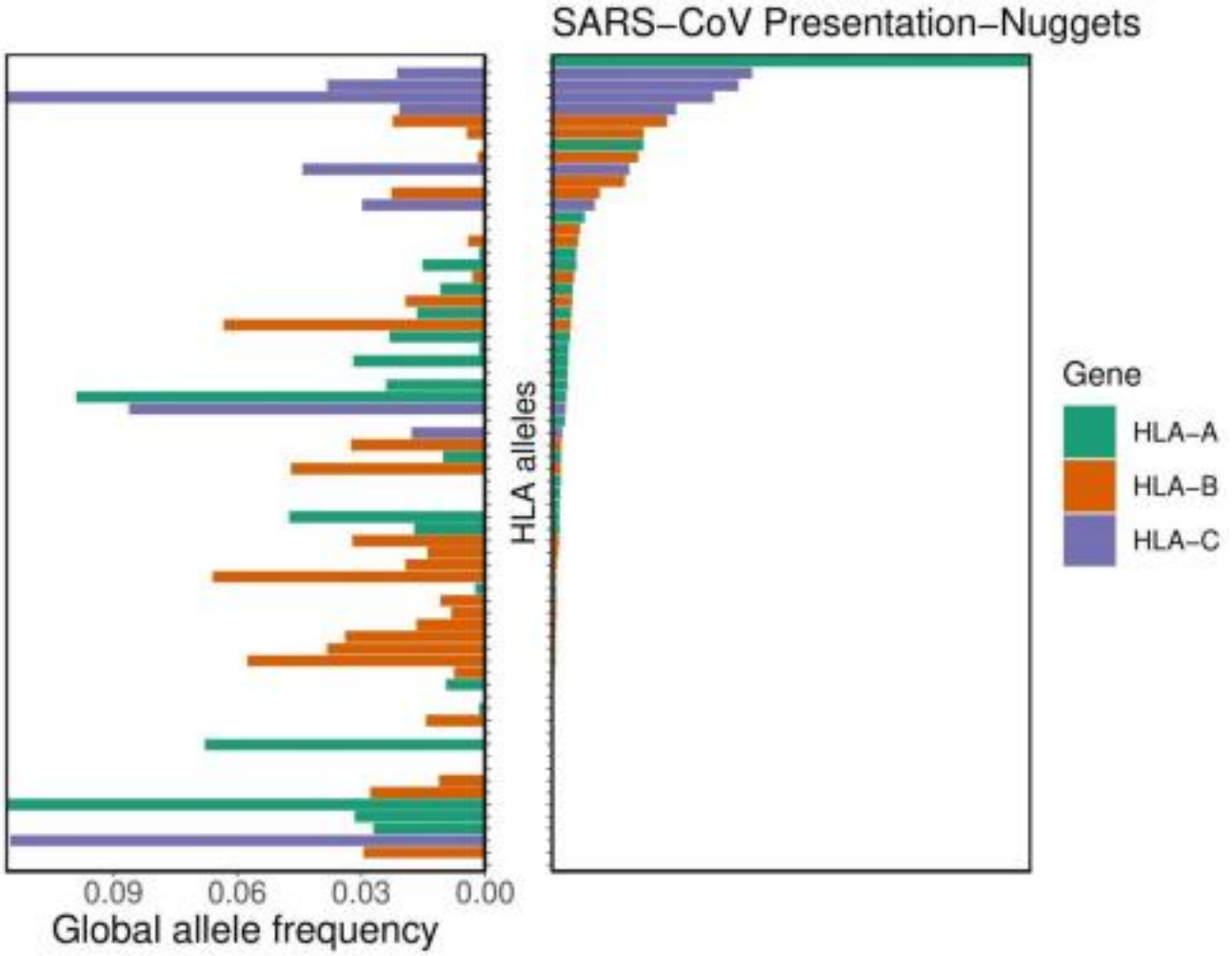
Supplementary Figure 2.8: Distribution of HLA allelic presentation of 8- to 12-mers from the SARS-CoV proteome using the tool MHCflurry. At right, the number of peptides (see Supplementary Table S8) that putatively bind to each of 66 HLA alleles is shown as a series of horizontal bars, with dark and light shading indicating the number of tightly (<50nM) and loosely (<500nM) binding peptides

respectively, and with green, orange, and purple colors representing HLA-A, -B, and -C alleles, respectively. Alleles are sorted in descending order based on the number of peptides they bind (<500nM). The corresponding estimated allelic frequency in the global population is also shown (to left), with length of horizontal bar indicating absolute frequency in the population.



Supplementary Figure 2.9: Distribution of HLA allelic presentation of 8- to 12-mers from the SARS-CoV-2 proteome using the tool MHCnuggets. At right, the number of peptides (see Supplementary Table S7) that putatively bind to each of 66 HLA alleles is shown as a series of horizontal bars, with dark and light shading indicating the number of tightly (<50nM) and loosely (<500nM) binding peptides respectively, and with green, orange, and purple colors representing HLA-A, -B, and -C alleles, respectively. Alleles are sorted in descending order based on the number of peptides they bind (<500nM). The corresponding estimated allelic frequency in the global population is also shown

(to left), with length of horizontal bar indicating absolute frequency in the population.



Supplementary Figure 2.10: Distribution of HLA allelic presentation of 8- to 12-mers from the SARS-CoV proteome using the tool MHCnuggets. At right, the number of peptides (see Supplementary Table S8) that putatively bind to each of 66 HLA alleles is shown as a series of horizontal bars, with dark and light shading indicating the number of tightly (<50nM) and loosely (<500nM) binding peptides respectively, and with green, orange, and purple colors representing HLA-A, -B, and -C alleles, respectively. Alleles are sorted in descending order based on the number of peptides they bind (<500nM). The corresponding estimated allelic frequency in the global population is also shown (to left), with length of horizontal bar indicating absolute frequency in the population.

Supplementary Table 2.1: Attached

Peptide	Source protein	Position (aa)	OC43.kmers	HKU1.kmers	NL63.kmers	229E.kmers	Quality score	Conservation.beta	Conservation.human	Conservation.combined
KHFSMMILSDD	ORF1ab (RNA polymerase)	5143-5153	10	10	10	10	239.090997	10.8181818	11	10.8181818
GPHEFCSQHTM	ORF1ab (RNA polymerase)	5200-5210	10	10	10	10	238.096775	10.8181818	10.8181818	10.8181818
YLPYPDPSRIL	ORF1ab (RNA polymerase)	5220-5230	0	10	10	6	237.241886	10.5454546	10.4545455	10.6363636
NVNRFNVAITRAK	ORF1ab (helicase)	5881-5893	20	20	20	10	237.037672	10.2307692	10.3846154	10.1538462
LKLFAAET	ORF1ab (helicase)	5454-5461	1	1	0	0	236.696679	11	10.75	10.5
LMGWDYPKCDRAMPNM	ORF1ab (RNA polymerase)	5006-5021	30	30	15	25	236.151721	10.875	10.625	10.625
CITRCNLGGAVC	ORF1ab (3'-to-5' exonuclease)	6398-6409	15	15	0	0	234.903891	10.25	10.1666667	10.0833333
VGVLTLDNQDLNG	ORF1ab (RNA polymerase)	4594-4606	20	10	20	20	234.681339	10.3846154	10.6923077	10.3076923
KAVFISPYNSQN	ORF1ab (helicase)	5832-5843	15	10	15	15	234.096112	10.4166667	10.5	10.1666667
QGSEYDYVI	ORF1ab (helicase)	5861-5869	3	3	3	3	232.915419	9.88888889	10.5555556	9.88888889
KLALGGSVAIKITE	ORF1ab (2'-O-ribose methyltransferase)	6958-6971	25	10	6	0	231.405709	10.0714286	9.92857143	9.42857143

CLFWNCNVD	ORF1ab (3'-to-5' exonuclease)	6307-6315	0	0	3	3	229.438511	9.77777778	10.2222222	10
LYYQNNVFMSE	ORF1ab (RNA polymerase)	5178-5188	10	10	6	0	229.344497	10.6363636	10.1818182	10.0909091
LYLGGMSYYC	ORF1ab (helicase)	5387-5396	6	6	0	0	227.706871	10.8	10.4	10.1
QFKHLIPLM	ORF1ab (3'-to-5' exonuclease)	6070-6078	3	1	0	0	226.330367	10.1111111	9.44444444	9.55555556
GGSLYVNKHAFHTPA	ORF1ab (3'-to-5' exonuclease)	6341-6355	20	20	30	0	225.991449	9.86666667	9.93333333	9.66666667
CFSVAALT	ORF1ab (RNA polymerase)	4787-4794	0	0	0	0	225.177139	9.75	10.5	9.625
IVCRFDTRV	ORF1ab (3'-to-5' exonuclease)	6322-6330	1	3	1	1	224.880884	10.2222222	10.4444444	10.3333333
VYTACSHAAVDALCEKA	ORF1ab (helicase)	5629-5645	15	15	0	3	222.483924	10.6470588	9.82352941	9.64705882
YVKPGGTSSGDATTAYANSVFNI	ORF1ab (RNA polymerase)	5066-5088	30	30	0	30	222.466457	10.3043478	10.0434783	9.86956522
ERFVSLAIDAYPL	ORF1ab (RNA polymerase)	5249-5261	20	0	6	6	220.82723	9.92307692	10.1538462	10.0769231
MMNVAKYTQLCQYLNT	ORF1ab (2'-O-ribose methyltransferase)	6839-6854	35	35	6	1	219.702855	9.8125	9.625	9.25
VYCPRHVI	ORF1ab (3CL)	3299-3306	1	1	0	1	219.314819	10	10.375	9.75
QGPPGTGKSH	ORF1ab (helicase)	5605-5614	6	6	0	0	215.58834	9.9	9.9	9.9
GDPAQLPAPR	ORF1ab (helicase)	5724-5733	6	6	0	0	215.58834	9.9	9.9	9.9

GAGSDKGVAPGTAVLRQWLP	ORF1ab (2'-O-ribose methyltransferase)	6869-6888	1	1	15	3	215.433072	9.05	9.25	8.75
DAIMTRCLAV	ORF1ab (3'-to-5' exonuclease)	6198-6207	6	3	0	6	214.263339	9.7	9.7	9.7
LKSIAATRGATVVIGT	ORF1ab (RNA polymerase)	4968-4983	3	3	0	0	213.559946	9.6875	9.625	9.1875
SQTSLRCG	ORF1ab (helicase)	5334-5341	1	1	0	0	208.233944	9.625	9.625	9.5
PYVCNAPGC	ORF1ab (helicase)	5371-5379	3	0	0	0	203.66079	10.1111111	9.77777778	8.88888889
TQMNLKYAISAKNRARTVAGVSI	ORF1ab (RNA polymerase)	4932-4954	70	70	0	0	202.969443	9.56521739	9.30434783	9.08695652
PPLNRNYVFTGY	ORF1ab (helicase)	5498-5509	0	0	0	0	202.868071	9.75	9.58333333	9.41666667
TLNGLWLDD	ORF1ab (3CL)	3289-3297	3	3	0	0	199.929732	9.22222222	8.88888889	8.77777778
RFYRLANECAQVLSE	ORF1ab (RNA polymerase)	5043-5057	30	30	0	0	199.400183	9.4	9.2	8.86666667
VNNLDKSAG	ORF1ab (RNA polymerase)	4887-4895	0	0	0	0	197.891889	9.11111111	9	8.88888889
PRWYFYYLGTGP	Nucleocapsid (N protein)	106-117	15	15	0	0	195.864283	10	9.83333333	8.25
FQTVKPGNFN	ORF1ab (RNA polymerase)	4799-4808	6	6	0	0	194.775547	9.7	8.8	8.5
WSFNPETN	Membrane (M protein)	110-117	1	1	0	0	194.44858	9.75	10	9.125
FGPLVRKIFVDGVPFVVS	ORF1ab (RNA polymerase)	4718-4735	10	10	0	0	194.098509	8.94444444	8.83333333	8.22222222

TGLFKDCS	ORF1ab (3'-to-5' exonuclease)	5930-5937	0	0	0	0	193.81969	8.5	8.25	7.625
LCCKCCYDHV	ORF1ab (helicase)	5349-5358	6	6	0	0	191.63408	8.8	8.8	8.8
SKEGFFTY	ORF1ab (2'-O-ribose methyltransferase)	6943-6950	0	0	0	1	191.490065	7.75	9.125	8.375
LGGLHLLIGL	ORF1ab (endoRNAse)	6697-6706	3	3	1	1	189.605719	8.5	8.5	8.4
VIDLLLDDFV	ORF1ab (endoRNAse)	6746-6755	0	6	1	1	186.536563	7.7	8.2	8.1
TVSALVYDNKL	ORF1ab (helicase)	5775-5785	0	10	0	0	184.826349	7.90909091	8.36363636	7.63636364
TNVNASSSE	ORF1ab (2'-O-ribose methyltransferase)	6993-7001	0	3	0	0	181.951097	7.88888889	8.22222222	8.11111111
WYDFVENPDI	ORF1ab (RNA polymerase)	4554-4563	6	6	0	0	179.653365	7.7	7.8	7.3
SLVLARKH	ORF1ab (RNA polymerase)	5027-5034	1	1	0	0	89.828475	4.125	4.125	4.125

Supplementary Table 2.2 SARS-CoV-2 peptides conserved across diverse coronavirus sequences. Peptide = amino acid sequence of peptide; Source protein = SARS-CoV-2 source protein containing the peptide sequence; Position (aa) = amino acid position within source protein; OC43.kmers = number of component 8-12mers also present in OC43; HKU1.kmers = number of component 8-12mers also present in HKU1; NL63.kmers = number of component 8-12mers also present in NL63; 229E.kmers = number of component 8-12mers also present in 229E; Quality score = average alignment quality score (calculated by Clustal Omega) across all constituent amino acids; Conservation.beta = BLOSUM62-based sequence conservation for all betacoronavirus sequences

(calculated by Clustal Omega), averaged across all constituent amino acids; Conservation.human = BLOSUM62-based sequence conservation for all human coronavirus sequences (calculated by Clustal Omega), averaged across all constituent amino acids; Conservation.combined = BLOSUM62-based sequence conservation for all alpha- and betacoronavirus sequences (calculated by Clustal Omega), averaged across all constituent amino acids.

Peptide	OC43	HKU1	NL63	229E	Cleaved.bound	Source
KHFSMMIL	1	1	1	1	N	KHFSMMILSDD
HFSMMILS	1	1	1	1	N	KHFSMMILSDD
FSMMILSD	1	1	1	1	N	KHFSMMILSDD
SMMILSDD	1	1	1	1	N	KHFSMMILSDD
GPHEFCSQ	1	1	1	1	N	GPHEFCSQHTM
PHEFCSQH	1	1	1	1	N	GPHEFCSQHTM
HEFCSQHT	1	1	1	1	N	GPHEFCSQHTM
EFCSQHTM	1	1	1	1	N	GPHEFCSQHTM
YLPYPDPS	0	1	1	1	N	YLPYPDPSRIL
LPYPDPSR	0	1	1	1	N	YLPYPDPSRIL
PYPDPSRI	0	1	1	1	N	YLPYPDPSRIL
YPDPSRIL	0	1	1	0	N	YLPYPDPSRIL

NVNRFNVA	1	1	1	0	N	NVNRFNVAITRAK
VNRFNVAI	1	1	1	0	N	NVNRFNVAITRAK
NRFNVAIT	1	1	1	1	N	NVNRFNVAITRAK
RFNVAITR	1	1	1	1	N	NVNRFNVAITRAK
FNVAITRA	1	1	1	1	N	NVNRFNVAITRAK
NVAITRAK	1	1	1	1	N	NVNRFNVAITRAK
LKLFAAET	1	1	0	0	N	LKLFAAET
LMGWDYPK	1	1	1	1	N	LMGWDYPKCDRAMPNM
MGWDYPKC	1	1	1	1	N	LMGWDYPKCDRAMPNM
GWDYPKCD	1	1	1	1	N	LMGWDYPKCDRAMPNM
WDYPKCDR	1	1	1	1	N	LMGWDYPKCDRAMPNM
DYPKCDRA	1	1	1	1	N	LMGWDYPKCDRAMPNM
YPKCDRAM	1	1	0	1	N	LMGWDYPKCDRAMPNM
PKCDRAMP	1	1	0	1	N	LMGWDYPKCDRAMPNM
KCDRAMPN	1	1	0	0	N	LMGWDYPKCDRAMPNM
CITRCNLG	1	1	0	0	N	CITRCNLGGAVC
ITRCNLGG	1	1	0	0	N	CITRCNLGGAVC
TRCNLGGA	1	1	0	0	N	CITRCNLGGAVC
RCNLGGAV	1	1	0	0	N	CITRCNLGGAVC

CNLGGAVC	1	1	0	0	N	CITRCNLGGAVC
VGVLTLDN	1	1	1	1	N	VGVLTLDNQDLNG
GVLTLDNQ	1	1	1	1	N	VGVLTLDNQDLNG
VLTLDNQD	1	1	1	1	N	VGVLTLDNQDLNG
LTLDNQDL	1	1	1	1	N	VGVLTLDNQDLNG
TLDNQDLN	1	0	1	1	N	VGVLTLDNQDLNG
LDNQDLNG	1	0	1	1	N	VGVLTLDNQDLNG
KAVFISPY	1	0	1	1	Y	KAVFISPYNSQN
AVFISPYN	1	1	1	1	N	KAVFISPYNSQN
VFISPYNS	1	1	1	1	N	KAVFISPYNSQN
FISPYNSQ	1	1	1	1	N	KAVFISPYNSQN
ISPYNSQN	1	1	1	1	N	KAVFISPYNSQN
QGSEYDYV	1	1	1	1	N	QGSEYDYVI
GSEYDYVI	1	1	1	1	N	QGSEYDYVI
KLALGGSV	1	0	0	0	N	KLALGGSVAIKITE
LALGGSVA	1	0	0	0	N	KLALGGSVAIKITE
ALGGSVAI	1	0	0	0	N	KLALGGSVAIKITE
LGGSVAIK	1	1	0	0	N	KLALGGSVAIKITE
GGSVAIKI	1	1	1	0	N	KLALGGSVAIKITE

GSVAIKIT	1	1	1	0	N	KLALGGSVAIKITE
SVAIKITE	1	1	1	0	N	KLALGGSVAIKITE
CLFWNCNV	0	0	1	1	N	CLFWNCNVD
LFWNCNVD	0	0	1	1	N	CLFWNCNVD
LYYQNNVF	1	1	1	0	N	LYYQNNVFMSE
YYQNNVFM	1	1	1	0	N	LYYQNNVFMSE
YQNNVFMS	1	1	1	0	Y	LYYQNNVFMSE
QNNVFMSE	1	1	0	0	N	LYYQNNVFMSE
LYLGGMSY	1	1	0	0	Y	LYLGGMSYYC
YLGGMSYY	1	1	0	0	N	LYLGGMSYYC
LGGMSYYC	1	1	0	0	N	LYLGGMSYYC
QFKHLIPL	1	0	0	0	N	QFKHLIPLM
FKHLIPLM	1	1	0	0	N	QFKHLIPLM
GGSLYVNK	1	1	1	0	N	GGSLYVNKHAFHTPA
GSLYVNKH	1	1	1	0	N	GGSLYVNKHAFHTPA
SLYVNKHA	1	1	1	0	N	GGSLYVNKHAFHTPA
LYVNKHAF	1	1	1	0	Y	GGSLYVNKHAFHTPA
YVNKHAFH	1	1	1	0	N	GGSLYVNKHAFHTPA
VNKHAFHT	1	1	1	0	N	GGSLYVNKHAFHTPA

NKHAFHTP	0	0	1	0	N	GGSLYVNKHAFHTPA
KHAFHTPA	0	0	1	0	N	GGSLYVNKHAFHTPA
IVCRFDTR	0	1	1	1	N	IVCRFDTRV
VCRFDTRV	1	1	0	0	N	IVCRFDTRV
TACSHAAV	0	0	0	1	N	VYTACSHAAVDALCEKA
ACSHAAVD	0	0	0	1	N	VYTACSHAAVDALCEKA
SHAAVDAL	1	1	0	0	Y	VYTACSHAAVDALCEKA
HAAVDALC	1	1	0	0	N	VYTACSHAAVDALCEKA
AAVDALCE	1	1	0	0	N	VYTACSHAAVDALCEKA
AVDALCEK	1	1	0	0	N	VYTACSHAAVDALCEKA
VDALCEKA	1	1	0	0	N	VYTACSHAAVDALCEKA
YVKPGGTS	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
VKPGGTSS	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
KPGGTSSG	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
PGGTSSGD	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
GGTSSGDA	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
GTSSGDAT	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
TSSGDATT	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
SSGDATTA	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI

SGDATTAY	0	0	0	1	N	YVKPGGTSSGDATTAYANSVFNI
GDATTAYA	0	0	0	1	N	YVKPGGTSSGDATTAYANSVFNI
DATTAYAN	0	0	0	1	N	YVKPGGTSSGDATTAYANSVFNI
ATTAYANS	0	0	0	1	N	YVKPGGTSSGDATTAYANSVFNI
TTAYANSV	0	0	0	1	Y	YVKPGGTSSGDATTAYANSVFNI
TAYANSVF	0	0	0	1	Y	YVKPGGTSSGDATTAYANSVFNI
AYANSVFN	0	0	0	1	N	YVKPGGTSSGDATTAYANSVFNI
YANSVFNI	0	0	0	1	N	YVKPGGTSSGDATTAYANSVFNI
ERFVSLAI	1	1	0	0	Y	ERFVSLAIDAYPL
RFVSLAID	1	1	0	0	N	ERFVSLAIDAYPL
FVSLAIDA	1	1	0	0	N	ERFVSLAIDAYPL
VSLAIDAY	1	1	1	1	N	ERFVSLAIDAYPL
SLAIDAYP	1	1	1	1	N	ERFVSLAIDAYPL
LAIDAYPL	1	1	1	1	Y	ERFVSLAIDAYPL
MMNVAKYT	1	1	0	0	N	MMNVAKYTQLCQYLNT
MNVAKYTQ	1	1	0	0	N	MMNVAKYTQLCQYLNT
NVAKYTQL	1	1	0	0	Y	MMNVAKYTQLCQYLNT
VAKYTQLC	1	1	0	0	N	MMNVAKYTQLCQYLNT
AKYTQLCQ	1	1	0	0	N	MMNVAKYTQLCQYLNT

KYTQLCQY	1	1	1	1	N	MMNVAKYTQLCQYLNT
YTQLCQYL	1	1	1	0	N	MMNVAKYTQLCQYLNT
TQLCQYLN	1	1	1	0	N	MMNVAKYTQLCQYLNT
QLCQYLNT	1	1	0	0	N	MMNVAKYTQLCQYLNT
VYCPRHVI	1	1	0	1	N	VYCPRHVI
QGPPGTGK	1	1	0	0	N	QGPPGTGKSH
GPPGTGKS	1	1	0	0	N	QGPPGTGKSH
PPGTGKSH	1	1	0	0	N	QGPPGTGKSH
GDPAQLPA	1	1	0	0	N	GDPAQLPAPR
DPAQLPAP	1	1	0	0	N	GDPAQLPAPR
PAQLPAPR	1	1	0	0	N	GDPAQLPAPR
GAGSDKGV	0	0	1	0	N	GAGSDKGVAPGTAVLRQWLP
AGSDKGVA	0	0	1	0	N	GAGSDKGVAPGTAVLRQWLP
GSDKGVAP	0	0	1	0	N	GAGSDKGVAPGTAVLRQWLP
SDKGVAPG	0	0	1	0	N	GAGSDKGVAPGTAVLRQWLP
DKGVAPGT	0	0	1	0	N	GAGSDKGVAPGTAVLRQWLP
GVAPGTAV	0	0	0	1	N	GAGSDKGVAPGTAVLRQWLP
VAPGTAVL	0	0	0	1	N	GAGSDKGVAPGTAVLRQWLP
AVLRQWLP	1	1	0	0	N	GAGSDKGVAPGTAVLRQWLP

DAIMTRCL	1	1	0	1	N	DAIMTRCLAV
AIMTRCLA	1	1	0	1	N	DAIMTRCLAV
IMTRCLAV	1	0	0	1	Y	DAIMTRCLAV
LKSIAATR	1	1	0	0	N	LKSIAATRGATVVIGT
KSIAATRG	1	1	0	0	N	LKSIAATRGATVVIGT
SQTSLRCG	1	1	0	0	N	SQTSLRCG
PYVCNAPG	1	0	0	0	N	PYVCNAPGC
YVCNAPGC	1	0	0	0	N	PYVCNAPGC
TQMNLKYA	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
QMNLKYAI	1	1	0	0	Y	TQMNLKYAISAKNRARTVAGVSI
MNLKYAIS	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
NLKYAISA	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
LKYAISAK	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
KYAISAKN	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
YAISAKNR	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
AISAKNRA	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
ISAKNRAR	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
SAKNRART	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
AKNRARTV	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI

KNRARTVA	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
NRARTVAG	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
RARTVAGV	1	1	0	0	Y	TQMNLKYAISAKNRARTVAGVSI
ARTVAGVS	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
RTVAGVSI	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
TLNGLWLD	1	1	0	0	N	TLNGLWLDD
LNGLWLDD	1	1	0	0	N	TLNGLWLDD
RFYRLANE	1	1	0	0	N	RFYRLANECAQVLSE
FYRLANEC	1	1	0	0	N	RFYRLANECAQVLSE
YRLANECA	1	1	0	0	N	RFYRLANECAQVLSE
RLANECAQ	1	1	0	0	N	RFYRLANECAQVLSE
LANECAQV	1	1	0	0	N	RFYRLANECAQVLSE
ANECAQVL	1	1	0	0	N	RFYRLANECAQVLSE
NECAQVLS	1	1	0	0	N	RFYRLANECAQVLSE
ECAQVLSE	1	1	0	0	N	RFYRLANECAQVLSE
PRWYFYYL	1	1	0	0	N	PRWYFYYLGTGP
RWYFYYLG	1	1	0	0	N	PRWYFYYLGTGP
WYFYYLGT	1	1	0	0	N	PRWYFYYLGTGP
YFYYLGTG	1	1	0	0	N	PRWYFYYLGTGP

FYYLGTGP	1	1	1	1	N	PRWYFYYLGTGP
FQTVKPGN	1	1	0	0	N	FQTVKPGNFN
QTVKPGNF	1	1	0	0	N	FQTVKPGNFN
TVKPGNFN	1	1	0	0	N	FQTVKPGNFN
WSFNPETN	1	1	0	0	N	WSFNPETN
IFVDGVPF	1	1	0	0	Y	FGPLVRKIFVDGVPFVVS
FVDGVPFV	1	1	0	0	Y	FGPLVRKIFVDGVPFVVS
VDGVPFVV	1	1	0	0	N	FGPLVRKIFVDGVPFVVS
DGVPFVVS	1	1	0	0	N	FGPLVRKIFVDGVPFVVS
LCCKCCYD	1	1	0	0	N	LCCKCCYDHV
CCKCCYDH	1	1	0	0	N	LCCKCCYDHV
CKCCYDHV	1	1	0	0	N	LCCKCCYDHV
SKEGFFTY	0	0	0	1	N	SKEGFFTY
LGGLHLLI	0	0	1	1	N	LGGLHLLIGL
GGLHLLIG	1	1	0	0	N	LGGLHLLIGL
GLHLLIGL	1	1	0	0	N	LGGLHLLIGL
VIDLLLDD	0	1	0	0	N	VIDLLLDDFV
IDLLLDDF	0	1	0	0	N	VIDLLLDDFV
DLLLDDFV	0	1	1	1	N	VIDLLLDDFV

TVSALVYD	0	1	0	0	N	TVSALVYDNKL
VSALVYDN	0	1	0	0	N	TVSALVYDNKL
SALVYDNK	0	1	0	0	N	TVSALVYDNKL
ALVYDNKL	0	1	0	0	N	TVSALVYDNKL
TNVNASSS	0	1	0	0	N	TNVNASSSE
NVNASSSE	0	1	0	0	N	TNVNASSSE
WYDFVENP	1	1	0	0	N	WYDFVENPDI
YDFVENPD	1	1	0	0	N	WYDFVENPDI
DFVENPDI	1	1	0	0	N	WYDFVENPDI
SLVLARKH	1	1	0	0	N	SLVLARKH
KHFSMMILS	1	1	1	1	N	KHFSMMILSDD
HFSMMILSD	1	1	1	1	N	KHFSMMILSDD
FSMMILSDD	1	1	1	1	N	KHFSMMILSDD
GPHEFCSQH	1	1	1	1	N	GPHEFCSQHTM
PHEFCSQHT	1	1	1	1	N	GPHEFCSQHTM
HEFCSQHTM	1	1	1	1	N	GPHEFCSQHTM
YLPYPDPSR	0	1	1	1	Y	YLPYPDPSRIL
LPYPDPSRI	0	1	1	1	N	YLPYPDPSRIL
PYPDPSRIL	0	1	1	0	N	YLPYPDPSRIL

NVNRFNVAI	1	1	1	0	Y	NVNRFNVAITRAK
VNRFNVAIT	1	1	1	0	N	NVNRFNVAITRAK
NRFNVAITR	1	1	1	1	N	NVNRFNVAITRAK
RFNVAITRA	1	1	1	1	N	NVNRFNVAITRAK
FNVAITRAK	1	1	1	1	N	NVNRFNVAITRAK
LMGWDYPKC	1	1	1	1	N	LMGWDYPKCDRAMPNM
MGWDYPKCD	1	1	1	1	N	LMGWDYPKCDRAMPNM
GWDYPKCDR	1	1	1	1	N	LMGWDYPKCDRAMPNM
WDYPKCDRA	1	1	1	1	N	LMGWDYPKCDRAMPNM
DYPKCDRAM	1	1	0	1	N	LMGWDYPKCDRAMPNM
YPKCDRAMP	1	1	0	1	N	LMGWDYPKCDRAMPNM
PKCDRAMPN	1	1	0	0	N	LMGWDYPKCDRAMPNM
CITRCNLGG	1	1	0	0	N	CITRCNLGGAVC
ITRCNLGGA	1	1	0	0	Y	CITRCNLGGAVC
TRCNLGGAV	1	1	0	0	N	CITRCNLGGAVC
RCNLGGAVC	1	1	0	0	N	CITRCNLGGAVC
VGVLTLDNQ	1	1	1	1	N	VGVLTLDNQDLNG
GVLTLDNQD	1	1	1	1	N	VGVLTLDNQDLNG
VLTLDNQDL	1	1	1	1	N	VGVLTLDNQDLNG

LTLDNQDLN	1	0	1	1	N	VGVLTLDNQDLNG
TLDNQDLNG	1	0	1	1	N	VGVLTLDNQDLNG
KAVFISPYN	1	0	1	1	N	KAVFISPYNSQN
AVFISPYNS	1	1	1	1	N	KAVFISPYNSQN
VFISPYNSQ	1	1	1	1	N	KAVFISPYNSQN
FISPYNSQN	1	1	1	1	N	KAVFISPYNSQN
QGSEYDYVI	1	1	1	1	N	QGSEYDYVI
KLALGGSVA	1	0	0	0	N	KLALGGSVAIKITE
LALGGSVAI	1	0	0	0	N	KLALGGSVAIKITE
ALGGSVAIK	1	0	0	0	Y	KLALGGSVAIKITE
LGGSVAIKI	1	1	0	0	N	KLALGGSVAIKITE
GGSVAIKIT	1	1	1	0	N	KLALGGSVAIKITE
GSVAIKITE	1	1	1	0	N	KLALGGSVAIKITE
CLFWNCNVD	0	0	1	1	N	CLFWNCNVD
LYYQNNVFM	1	1	1	0	N	LYYQNNVFMSE
YYQNNVFMS	1	1	1	0	N	LYYQNNVFMSE
YQNNVFMSE	1	1	0	0	Y	LYYQNNVFMSE
LYLGGMSYY	1	1	0	0	Y	LYLGGMSYYC
YLGGMSYYC	1	1	0	0	N	LYLGGMSYYC

QFKHLIPLM	1	0	0	0	N	QFKHLIPLM
GGSLYVNKH	1	1	1	0	N	GGSLYVNKHAFHTPA
GSLYVNKHA	1	1	1	0	N	GGSLYVNKHAFHTPA
SLYVNKHAF	1	1	1	0	Y	GGSLYVNKHAFHTPA
LYVNKHAFH	1	1	1	0	N	GGSLYVNKHAFHTPA
YVNKHAFHT	1	1	1	0	N	GGSLYVNKHAFHTPA
VNKHAFHTP	0	0	1	0	N	GGSLYVNKHAFHTPA
NKHAFHTPA	0	0	1	0	N	GGSLYVNKHAFHTPA
IVCRFDTRV	0	1	0	0	N	IVCRFDTRV
TACSHAAVD	0	0	0	1	N	VYTACSHAAVDALCEKA
SHAAVDALC	1	1	0	0	Y	VYTACSHAAVDALCEKA
HAAVDALCE	1	1	0	0	N	VYTACSHAAVDALCEKA
AAVDALCEK	1	1	0	0	Y	VYTACSHAAVDALCEKA
AVDALCEKA	1	1	0	0	N	VYTACSHAAVDALCEKA
YVKPGGTSS	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
VKPGGTSSG	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
KPGGTSSGD	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
PGGTSSGDA	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
GGTSSGDAT	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI

GTSSGDATT	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
TSSGDATTA	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
SGDATTAYA	0	0	0	1	N	YVKPGGTSSGDATTAYANSVFNI
GDATTAYAN	0	0	0	1	N	YVKPGGTSSGDATTAYANSVFNI
DATTAYANS	0	0	0	1	N	YVKPGGTSSGDATTAYANSVFNI
ATTAYANSV	0	0	0	1	Y	YVKPGGTSSGDATTAYANSVFNI
TTAYANSVF	0	0	0	1	Y	YVKPGGTSSGDATTAYANSVFNI
TAYANSVFN	0	0	0	1	N	YVKPGGTSSGDATTAYANSVFNI
AYANSVFNI	0	0	0	1	N	YVKPGGTSSGDATTAYANSVFNI
ERFVSLAID	1	1	0	0	N	ERFVSLAIDAYPL
RFVSLAIDA	1	1	0	0	N	ERFVSLAIDAYPL
FVSLAIDAY	1	1	0	0	N	ERFVSLAIDAYPL
VSLAIDAYP	1	1	1	1	N	ERFVSLAIDAYPL
SLAIDAYPL	1	1	1	1	N	ERFVSLAIDAYPL
MMNVAKYTQ	1	1	0	0	N	MMNVAKYTQLCQYLNT
MNVAKYTQL	1	1	0	0	Y	MMNVAKYTQLCQYLNT
NVAKYTQLC	1	1	0	0	N	MMNVAKYTQLCQYLNT
VAKYTQLCQ	1	1	0	0	N	MMNVAKYTQLCQYLNT
AKYTQLCQY	1	1	0	0	N	MMNVAKYTQLCQYLNT

KYTQLCQYL	1	1	1	0	Y	MMNVAKYTQLCQYLNT
YTQLCQYLN	1	1	1	0	N	MMNVAKYTQLCQYLNT
TQLCQYLNT	1	1	0	0	N	MMNVAKYTQLCQYLNT
QGPPGTGKS	1	1	0	0	N	QGPPGTGKSH
GPPGTGKSH	1	1	0	0	N	QGPPGTGKSH
GDPAQLPAP	1	1	0	0	N	GDPAQLPAPR
DPAQLPAPR	1	1	0	0	N	GDPAQLPAPR
GAGSDKGVA	0	0	1	0	N	GAGSDKGVAPGTAVLRQWLP
AGSDKGVAP	0	0	1	0	N	GAGSDKGVAPGTAVLRQWLP
GSDKGVAPG	0	0	1	0	N	GAGSDKGVAPGTAVLRQWLP
SDKGVAPGT	0	0	1	0	N	GAGSDKGVAPGTAVLRQWLP
GVAPGTAVL	0	0	0	1	N	GAGSDKGVAPGTAVLRQWLP
DAIMTRCLA	1	1	0	1	N	DAIMTRCLAV
AIMTRCLAV	1	0	0	1	Y	DAIMTRCLAV
LKSIAATRG	1	1	0	0	N	LKSIAATRGATVVIGT
PYVCNAPGC	1	0	0	0	N	PYVCNAPGC
TQMNLKYAI	1	1	0	0	Y	TQMNLKYAISAKNRARTVAGVSI
QMNLKYAIS	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
MNLKYAISA	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI

NLKYAISAK	1	1	0	0	Y	TQMNLKYAISAKNRARTVAGVSI
LKYAISAKN	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
KYAISAKNR	1	1	0	0	Y	TQMNLKYAISAKNRARTVAGVSI
YAISAKNRA	1	1	0	0	Y	TQMNLKYAISAKNRARTVAGVSI
AISAKNRAR	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
ISAKNRART	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
SAKNRARTV	1	1	0	0	Y	TQMNLKYAISAKNRARTVAGVSI
AKNRARTVA	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
KNRARTVAG	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
NRARTVAGV	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
RARTVAGVS	1	1	0	0	Y	TQMNLKYAISAKNRARTVAGVSI
ARTVAGVSI	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
TLNGLWLDD	1	1	0	0	N	TLNGLWLDD
RFYRLANEC	1	1	0	0	N	RFYRLANECAQVLSE
FYRLANECA	1	1	0	0	N	RFYRLANECAQVLSE
YRLANECAQ	1	1	0	0	N	RFYRLANECAQVLSE
RLANECAQV	1	1	0	0	N	RFYRLANECAQVLSE
LANECAQVL	1	1	0	0	N	RFYRLANECAQVLSE
ANECAQVLS	1	1	0	0	N	RFYRLANECAQVLSE

NECAQVLSE	1	1	0	0	N	RFYRLANECAQVLSE
PRWYFYYLG	1	1	0	0	N	PRWYFYYLGTGP
RWYFYYLGT	1	1	0	0	N	PRWYFYYLGTGP
WYFYYLGTG	1	1	0	0	N	PRWYFYYLGTGP
YFYYLGTGP	1	1	0	0	N	PRWYFYYLGTGP
FQTVKPGNF	1	1	0	0	N	FQTVKPGNFN
QTVKPGNFN	1	1	0	0	N	FQTVKPGNFN
IFVDGVPFV	1	1	0	0	Y	FGPLVRKIFVDGVPFVVS
FVDGVPFVV	1	1	0	0	Y	FGPLVRKIFVDGVPFVVS
VDGVPFVVS	1	1	0	0	N	FGPLVRKIFVDGVPFVVS
LCCKCCYDH	1	1	0	0	N	LCCKCCYDHV
CCKCCYDHV	1	1	0	0	N	LCCKCCYDHV
GGLHLLIGL	1	1	0	0	N	LGGLHLLIGL
VIDLLLDDF	0	1	0	0	N	VIDLLLDDFV
IDLLLDDFV	0	1	0	0	N	VIDLLLDDFV
TVSALVYDN	0	1	0	0	N	TVSALVYDNKL
VSALVYDNK	0	1	0	0	Y	TVSALVYDNKL
SALVYDNKL	0	1	0	0	N	TVSALVYDNKL
TNVNASSSE	0	1	0	0	N	TNVNASSSE

WYDFVENPD	1	1	0	0	N	WYDFVENPDI
YDFVENPDI	1	1	0	0	N	WYDFVENPDI
KHFSMMILSD	1	1	1	1	N	KHFSMMILSDD
HFSMMILSDD	1	1	1	1	N	KHFSMMILSDD
GPHEFCSQHT	1	1	1	1	N	GPHEFCSQHTM
PHEFCSQHTM	1	1	1	1	N	GPHEFCSQHTM
YLPYPDPSRI	0	1	1	1	Y	YLPYPDPSRIL
LPYPDPSRIL	0	1	1	0	N	YLPYPDPSRIL
NVNRFNVAIT	1	1	1	0	N	NVNRFNVAITRAK
VNRFNVAITR	1	1	1	0	Y	NVNRFNVAITRAK
NRFNVAITRA	1	1	1	1	N	NVNRFNVAITRAK
RFNVAITRAK	1	1	1	1	N	NVNRFNVAITRAK
LMGWDYPKCD	1	1	1	1	N	LMGWDYPKCDRAMPNM
MGWDYPKCDR	1	1	1	1	Y	LMGWDYPKCDRAMPNM
GWDYPKCDRA	1	1	1	1	N	LMGWDYPKCDRAMPNM
WDYPKCDRAM	1	1	0	1	N	LMGWDYPKCDRAMPNM
DYPKCDRAMP	1	1	0	1	N	LMGWDYPKCDRAMPNM
YPKCDRAMPN	1	1	0	0	N	LMGWDYPKCDRAMPNM
CITRCNLGGA	1	1	0	0	N	CITRCNLGGAVC

ITRCNLGGAV	1	1	0	0	Y	CITRCNLGGAVC
TRCNLGGAVC	1	1	0	0	N	CITRCNLGGAVC
VGVLTLDNQD	1	1	1	1	N	VGVLTLDNQDLNG
GVLTLDNQDL	1	1	1	1	N	VGVLTLDNQDLNG
VLTLDNQDLN	1	0	1	1	N	VGVLTLDNQDLNG
LTLDNQDLNG	1	0	1	1	N	VGVLTLDNQDLNG
KAVFISPYNS	1	0	1	1	N	KAVFISPYNSQN
AVFISPYNSQ	1	1	1	1	N	KAVFISPYNSQN
VFISPYNSQN	1	1	1	1	N	KAVFISPYNSQN
KLALGGSVAI	1	0	0	0	N	KLALGGSVAIKITE
LALGGSVAIK	1	0	0	0	N	KLALGGSVAIKITE
ALGGSVAIKI	1	0	0	0	Y	KLALGGSVAIKITE
LGGSVAIKIT	1	1	0	0	N	KLALGGSVAIKITE
GGSVAIKITE	1	1	1	0	N	KLALGGSVAIKITE
LYYQNNVFMS	1	1	1	0	N	LYYQNNVFMSE
YYQNNVFMSE	1	1	0	0	N	LYYQNNVFMSE
LYLGGMSYYC	1	1	0	0	N	LYLGGMSYYC
GGSLYVNKHA	1	1	1	0	N	GGSLYVNKHAFHTPA
GSLYVNKHAF	1	1	1	0	N	GGSLYVNKHAFHTPA

SLYVNKHAFH	1	1	1	0	N	GGSLYVNKHAFHTPA
LYVNKHAFHT	1	1	1	0	N	GGSLYVNKHAFHTPA
YVNKHAFHTP	0	0	1	0	N	GGSLYVNKHAFHTPA
VNKHAFHTPA	0	0	1	0	N	GGSLYVNKHAFHTPA
SHAAVDALCE	1	1	0	0	N	VYTACSHAAVDALCEKA
HAAVDALCEK	1	1	0	0	Y	VYTACSHAAVDALCEKA
AAVDALCEKA	1	1	0	0	N	VYTACSHAAVDALCEKA
YVKPGGTSSG	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
VKPGGTSSGD	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
KPGGTSSGDA	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
PGGTSSGDAT	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
GGTSSGDATT	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
GTSSGDATTA	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
SGDATTAYAN	0	0	0	1	N	YVKPGGTSSGDATTAYANSVFNI
GDATTAYANS	0	0	0	1	N	YVKPGGTSSGDATTAYANSVFNI
DATTAYANSV	0	0	0	1	Y	YVKPGGTSSGDATTAYANSVFNI
ATTAYANSVF	0	0	0	1	Y	YVKPGGTSSGDATTAYANSVFNI
TTAYANSVFN	0	0	0	1	N	YVKPGGTSSGDATTAYANSVFNI
TAYANSVFNI	0	0	0	1	Y	YVKPGGTSSGDATTAYANSVFNI

ERFVSLAIDA	1	1	0	0	N	ERFVSLAIDAYPL
RFVSLAIDAY	1	1	0	0	Y	ERFVSLAIDAYPL
FVSLAIDAYP	1	1	0	0	N	ERFVSLAIDAYPL
VSLAIDAYPL	1	1	1	1	N	ERFVSLAIDAYPL
MMNVAKYTQL	1	1	0	0	Y	MMNVAKYTQLCQYLNT
MNVAKYTQLC	1	1	0	0	N	MMNVAKYTQLCQYLNT
NVAKYTQLCQ	1	1	0	0	N	MMNVAKYTQLCQYLNT
VAKYTQLCQY	1	1	0	0	Y	MMNVAKYTQLCQYLNT
AKYTQLCQYL	1	1	0	0	N	MMNVAKYTQLCQYLNT
KYTQLCQYLN	1	1	1	0	Y	MMNVAKYTQLCQYLNT
YTQLCQYLNT	1	1	0	0	N	MMNVAKYTQLCQYLNT
QGPPGTGKSH	1	1	0	0	N	QGPPGTGKSH
GDPAQLPAPR	1	1	0	0	N	GDPAQLPAPR
GAGSDKGVAP	0	0	1	0	N	GAGSDKGVAPGTAVLRQWLP
AGSDKGVAPG	0	0	1	0	N	GAGSDKGVAPGTAVLRQWLP
GSDKGVAPGT	0	0	1	0	N	GAGSDKGVAPGTAVLRQWLP
DAIMTRCLAV	1	0	0	1	N	DAIMTRCLAV
TQMNLKYAIS	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
QMNLKYAISA	1	1	0	0	Y	TQMNLKYAISAKNRARTVAGVSI

MNLKYAISAK	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
NLKYAISAKN	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
LKYAISAKNR	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
KYAISAKNRA	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
YAISAKNRAR	1	1	0	0	Y	TQMNLKYAISAKNRARTVAGVSI
AISAKNRART	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
ISAKNRARTV	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
SAKNRARTVA	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
AKNRARTVAG	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
KNRARTVAGV	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
NRARTVAGVS	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
RARTVAGVSI	1	1	0	0	Y	TQMNLKYAISAKNRARTVAGVSI
RFYRLANECA	1	1	0	0	N	RFYRLANECAQVLSE
FYRLANECAQ	1	1	0	0	N	RFYRLANECAQVLSE
YRLANECAQV	1	1	0	0	Y	RFYRLANECAQVLSE
RLANECAQVL	1	1	0	0	N	RFYRLANECAQVLSE
LANECAQVLS	1	1	0	0	N	RFYRLANECAQVLSE
ANECAQVLSE	1	1	0	0	N	RFYRLANECAQVLSE
PRWYFYYLGT	1	1	0	0	N	PRWYFYYLGTGP

RWYFYYLGTG	1	1	0	0	N	PRWYFYYLGTGP
WYFYYLGTGP	1	1	0	0	N	PRWYFYYLGTGP
FQTVKPGNFN	1	1	0	0	N	FQTVKPGNFN
IFVDGVPFVV	1	1	0	0	Y	FGPLVRKIFVDGVPFVVS
FVDGVPFVVS	1	1	0	0	Y	FGPLVRKIFVDGVPFVVS
LCCKCCYDHV	1	1	0	0	N	LCCKCCYDHV
VIDLLLDDFV	0	1	0	0	Y	VIDLLLDDFV
TVSALVYDNK	0	1	0	0	N	TVSALVYDNKL
VSALVYDNKL	0	1	0	0	N	TVSALVYDNKL
WYDFVENPDI	1	1	0	0	N	WYDFVENPDI
KHFSMMILSDD	1	1	1	1	N	KHFSMMILSDD
GPHEFCSQHTM	1	1	1	1	N	GPHEFCSQHTM
YLPYPDPSRIL	0	1	1	0	Y	YLPYPDPSRIL
NVNRFNVAITR	1	1	1	0	Y	NVNRFNVAITRAK
VNRFNVAITRA	1	1	1	0	N	NVNRFNVAITRAK
NRFNVAITRAK	1	1	1	1	N	NVNRFNVAITRAK
LMGWDYPKCDR	1	1	1	1	N	LMGWDYPKCDRAMPNM
MGWDYPKCDRA	1	1	1	1	N	LMGWDYPKCDRAMPNM
GWDYPKCDRAM	1	1	0	1	N	LMGWDYPKCDRAMPNM

WDYPKCDRAMP	1	1	0	1	N	LMGWDYPKCDRAMPNM
DYPKCDRAMPN	1	1	0	0	N	LMGWDYPKCDRAMPNM
CITRCNLGGAV	1	1	0	0	N	CITRCNLGGAVC
ITRCNLGGAVC	1	1	0	0	N	CITRCNLGGAVC
VGVLTLDNQDL	1	1	1	1	N	VGVLTLDNQDLNG
GVLTLDNQDLN	1	0	1	1	N	VGVLTLDNQDLNG
VLTLDNQDLNG	1	0	1	1	N	VGVLTLDNQDLNG
KAVFISPYNSQ	1	0	1	1	N	KAVFISPYNSQN
AVFISPYNSQN	1	1	1	1	N	KAVFISPYNSQN
KLALGGSVAIK	1	0	0	0	N	KLALGGSVAIKITE
LALGGSVAIKI	1	0	0	0	N	KLALGGSVAIKITE
ALGGSVAIKIT	1	0	0	0	N	KLALGGSVAIKITE
LGGSVAIKITE	1	1	0	0	N	KLALGGSVAIKITE
LYYQNNVFMSE	1	1	0	0	N	LYYQNNVFMSE
GGSLYVNKHAF	1	1	1	0	N	GGSLYVNKHAFHTPA
GSLYVNKHAFH	1	1	1	0	N	GGSLYVNKHAFHTPA
SLYVNKHAFHT	1	1	1	0	N	GGSLYVNKHAFHTPA
LYVNKHAFHTP	0	0	1	0	N	GGSLYVNKHAFHTPA
YVNKHAFHTPA	0	0	1	0	Y	GGSLYVNKHAFHTPA

SHAAVDALCEK	1	1	0	0	N	VYTACSHAAVDALCEKA
HAAVDALCEKA	1	1	0	0	N	VYTACSHAAVDALCEKA
YVKPGGTSSGD	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
VKPGGTSSGDA	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
KPGGTSSGDAT	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
PGGTSSGDATT	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
GGTSSGDATTA	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
SGDATTAYANS	0	0	0	1	N	YVKPGGTSSGDATTAYANSVFNI
GDATTAYANSV	0	0	0	1	N	YVKPGGTSSGDATTAYANSVFNI
DATTAYANSVF	0	0	0	1	Y	YVKPGGTSSGDATTAYANSVFNI
ATTAYANSVFN	0	0	0	1	N	YVKPGGTSSGDATTAYANSVFNI
TTAYANSVFNI	0	0	0	1	Y	YVKPGGTSSGDATTAYANSVFNI
ERFVSLAIDAY	1	1	0	0	Y	ERFVSLAIDAYPL
RFVSLAIDAYP	1	1	0	0	N	ERFVSLAIDAYPL
FVSLAIDAYPL	1	1	0	0	N	ERFVSLAIDAYPL
MMNVAKYTQLC	1	1	0	0	N	MMNVAKYTQLCQYLNT
MNVAKYTQLCQ	1	1	0	0	N	MMNVAKYTQLCQYLNT
NVAKYTQLCQY	1	1	0	0	Y	MMNVAKYTQLCQYLNT
VAKYTQLCQYL	1	1	0	0	Y	MMNVAKYTQLCQYLNT

AKYTQLCQYLN	1	1	0	0	N	MMNVAKYTQLCQYLNT
KYTQLCQYLNT	1	1	0	0	N	MMNVAKYTQLCQYLNT
GAGSDKGVAPG	0	0	1	0	N	GAGSDKGVAPGTAVLRQWLP
AGSDKGVAPGT	0	0	1	0	N	GAGSDKGVAPGTAVLRQWLP
TQMNLKYAISA	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
QMNLKYAISAK	1	1	0	0	Y	TQMNLKYAISAKNRARTVAGVSI
MNLKYAISAKN	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
NLKYAISAKNR	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
LKYAISAKNRA	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
KYAISAKNRAR	1	1	0	0	Y	TQMNLKYAISAKNRARTVAGVSI
YAISAKNRART	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
AISAKNRARTV	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
ISAKNRARTVA	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
SAKNRARTVAG	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
AKNRARTVAGV	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
KNRARTVAGVS	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
NRARTVAGVSI	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
RFYRLANECAQ	1	1	0	0	N	RFYRLANECAQVLSE
FYRLANECAQV	1	1	0	0	N	RFYRLANECAQVLSE

YRLANECAQVL	1	1	0	0	Y	RFYRLANECAQVLSE
RLANECAQVLS	1	1	0	0	N	RFYRLANECAQVLSE
LANECAQVLSE	1	1	0	0	N	RFYRLANECAQVLSE
PRWYFYYLGTG	1	1	0	0	N	PRWYFYYLGTGP
RWYFYYLGTGP	1	1	0	0	N	PRWYFYYLGTGP
IFVDGVPFVVS	1	1	0	0	N	FGPLVRKIFVDGVPFVVS
TVSALVYDNKL	0	1	0	0	N	TVSALVYDNKL
NVNRFNVAITRA	1	1	1	0	N	NVNRFNVAITRAK
VNRFNVAITRAK	1	1	1	0	N	NVNRFNVAITRAK
LMGWDYPKCDRA	1	1	1	1	N	LMGWDYPKCDRAMPNM
MGWDYPKCDRAM	1	1	0	1	N	LMGWDYPKCDRAMPNM
GWDYPKCDRAMP	1	1	0	1	N	LMGWDYPKCDRAMPNM
WDYPKCDRAMPN	1	1	0	0	N	LMGWDYPKCDRAMPNM
CITRCNLGGAVC	1	1	0	0	N	CITRCNLGGAVC
VGVLTLDNQDLN	1	0	1	1	N	VGVLTLDNQDLNG
GVLTLDNQDLNG	1	0	1	1	N	VGVLTLDNQDLNG
KAVFISPYNSQN	1	0	1	1	N	KAVFISPYNSQN
KLALGGSVAIKI	1	0	0	0	N	KLALGGSVAIKITE
LALGGSVAIKIT	1	0	0	0	N	KLALGGSVAIKITE

ALGGSVAIKITE	1	0	0	0	N	KLALGGSVAIKITE
GGSLYVNKHAFH	1	1	1	0	N	GGSLYVNKHAFHTPA
GSLYVNKHAFHT	1	1	1	0	N	GGSLYVNKHAFHTPA
SLYVNKHAFHTP	0	0	1	0	N	GGSLYVNKHAFHTPA
LYVNKHAFHTPA	0	0	1	0	N	GGSLYVNKHAFHTPA
SHAAVDALCEKA	1	1	0	0	N	VYTACSHAAVDALCEKA
YVKPGGTSSGDA	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
VKPGGTSSGDAT	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
KPGGTSSGDATT	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
PGGTSSGDATTA	1	1	0	0	N	YVKPGGTSSGDATTAYANSVFNI
SGDATTAYANSV	0	0	0	1	N	YVKPGGTSSGDATTAYANSVFNI
GDATTAYANSVF	0	0	0	1	Y	YVKPGGTSSGDATTAYANSVFNI
DATTAYANSVFN	0	0	0	1	N	YVKPGGTSSGDATTAYANSVFNI
ATTAYANSVFNI	0	0	0	1	Y	YVKPGGTSSGDATTAYANSVFNI
ERFVSLAIDAYP	1	1	0	0	N	ERFVSLAIDAYPL
RFVSLAIDAYPL	1	1	0	0	Y	ERFVSLAIDAYPL
MMNVAKYTQLCQ	1	1	0	0	N	MMNVAKYTQLCQYLNT
MNVAKYTQLCQY	1	1	0	0	Y	MMNVAKYTQLCQYLNT
NVAKYTQLCQYL	1	1	0	0	N	MMNVAKYTQLCQYLNT

VAKYTQLCQYLN	1	1	0	0	N	MMNVAKYTQLCQYLNT
AKYTQLCQYLNT	1	1	0	0	N	MMNVAKYTQLCQYLNT
GAGSDKGVAPGT	0	0	1	0	N	GAGSDKGVAPGTAVLRQWLP
TQMNLKYAISAK	1	1	0	0	Y	TQMNLKYAISAKNRARTVAGVSI
QMNLKYAISAKN	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
MNLKYAISAKNR	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
NLKYAISAKNRA	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
LKYAISAKNRAR	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
KYAISAKNRART	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
YAISAKNRARTV	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
AISAKNRARTVA	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
ISAKNRARTVAG	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
SAKNRARTVAGV	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
AKNRARTVAGVS	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
KNRARTVAGVSI	1	1	0	0	N	TQMNLKYAISAKNRARTVAGVSI
RFYRLANECAQV	1	1	0	0	N	RFYRLANECAQVLSE
FYRLANECAQVL	1	1	0	0	N	RFYRLANECAQVLSE
YRLANECAQVLS	1	1	0	0	N	RFYRLANECAQVLSE
RLANECAQVLSE	1	1	0	0	N	RFYRLANECAQVLSE

PRWYFYYLGTGP	1	1	0	0	N	PRWYFYYLGTGP

Supplementary Table 2.3. Presence of 8-12mer peptides across four human coronavirus sequences. Peptide = 8-12mer peptide amino acid sequence; OC43 = value indicating presence (1) or absence (0) of kmer in OC43 proteome; HKU1 = value indicating presence (1) or absence (0) of kmer in HKU1 proteome; NL63 = value indicating presence (1) or absence (0) of kmer in NL63 proteome; 229E = value indicating presence (1) or absence (0) of kmer in 229E proteome; Cleaved.bound = value indicating whether peptide is predicted to be cleaved at C-terminus AND whether one or more HLA alleles are predicted to bind with affinity <500nM (Y=yes, N=no); Source = source peptide from which kmer was obtained.

Peptide	Validated allele(s)	Other allele(s)	Predicted allele(s)
ALNTLVKQL	A*02:01	NA	A*02:02, A*02:03
GETALALLLL	B*40:01, B*40:02, B*44:02, B*44:03	A*02:01, A*02:07, B*18:01, B*45:01, B*46:01	A*68:02, B*13:01, B*40:01, B*40:02, B*41:01, B*41:02, B*44:02, B*44:03, B*44:10, B*47:01, B*47:03
GLMWLSYFV	A*02:01, A*02:02, A*02:03, A*02:06, A*68:02	A*01:01, A*03:01, A*11:01, A*24:02, A*26:01, A*31:01, A*69:01, B*07:02, B*08:01, B*15:01, B*27:05, B*40:01, B*58:01	A*02:01, A*02:02, A*02:03, A*02:05, A*02:06, A*02:07, A*32:01, A*68:02
HLRMAGHSL	NA	NA	A*02:03, A*30:01, B*07:02, B*07:04, B*07:05, B*08:01, B*15:01, B*15:02, B*15:03, B*15:07, B*15:25, B*15:27, B*15:32, B*39:10, B*42:01, B*42:02, B*81:01, C*12:03, C*14:02, C*14:03, C*16:01
MEVTPSGTWL	B*40:01	NA	B*40:01, B*40:02, B*41:01, B*44:02, B*44:03
NLNESLIDL	A*02:01	NA	A*02:01, A*02:02, A*02:03, A*02:05, A*02:06
QFKDNVILL	A*24:02	A*01:01, A*23:01, A*26:01, A*29:02, A*30:02	A*24:03, C*07:02, C*14:02, C*14:03
RLNQLESKV	A*02:01	NA	A*02:02, A*02:03
SIVAYTMSL	A*02:01, A*02:02, A*02:03, A*02:06, A*68:02	NA	A*02:01, A*02:02, A*02:03, A*02:05, A*02:06, A*26:02, A*32:01, A*68:02, B*15:01, B*15:03,

			B*15:05, B*15:07, B*15:17, B*15:25, B*15:32, B*39:10, B*42:01, B*67:01, C*03:02, C*03:03, C*03:04, C*12:02, C*12:03, C*14:02, C*14:03, C*16:01
VLNDILSRL	A*02:01	NA	A*02:01, A*02:02, A*02:03, A*02:05, A*02:06, C*02:02, C*02:10, C*12:02, C*12:03, C*16:01, C*17:01

Supplementary Table 2.4: Prediction of validated SARS-CoV peptides from IEDB. Peptide = amino acid sequence of peptide; Validated allele(s) = experimentally validated HLA types shown to present the corresponding peptide which are also predicted to bind the same alleles in our analysis (full bold underline indicates 4-digit HLA match, while partial under underline indicates major allele match); Other allele(s) = experimentally validated HLA types shown to present the corresponding peptide which are not predicted to bind the same alleles in our analysis; Predicted allele(s) = HLA alleles predicted in silico to bind the indicated peptide with a binding affinity <500nM.

Name	Genus	Subgenus	Taxonomy ID	ORF1ab	Spike	Envelope	Membrane	Nucleocapsid
SARS-CoV-2*	Betacoronavirus	Sarbecovirus	NCBI:txid2697049	YP_009724389.1	YP_009724390.1	YP_009724392.1	YP_009724393.1	YP_009724397.2
SARS-CoV*	Betacoronavirus	Sarbecovirus	NCBI:txid694009	NP_828849.2	NP_828851.1	NP_828854.1	NP_828855.1	NP_828858.1
OC43*	Betacoronavirus	Embecovirus	NCBI:txid31631	YP_009555238.1	YP_009555241.1	YP_009555243.1	YP_009555244.1	YP_009555245.1
Bovine-CoV	Betacoronavirus	Embecovirus	NCBI:txid11128	NP_150073.2	NP_150077.1	NP_150081.1	NP_150082.1	NP_150083.1
HKU24	Betacoronavirus	Embecovirus	NCBI:txid2501960	YP_009113022.1	YP_009113025.1	YP_009113028.1	YP_009113029.1	YP_009113031.1
HKU1*	Betacoronavirus	Embecovirus	NCBI:txid290028	YP_173236.1	YP_173238.1	YP_173240.1	YP_173241.1	YP_173242.1

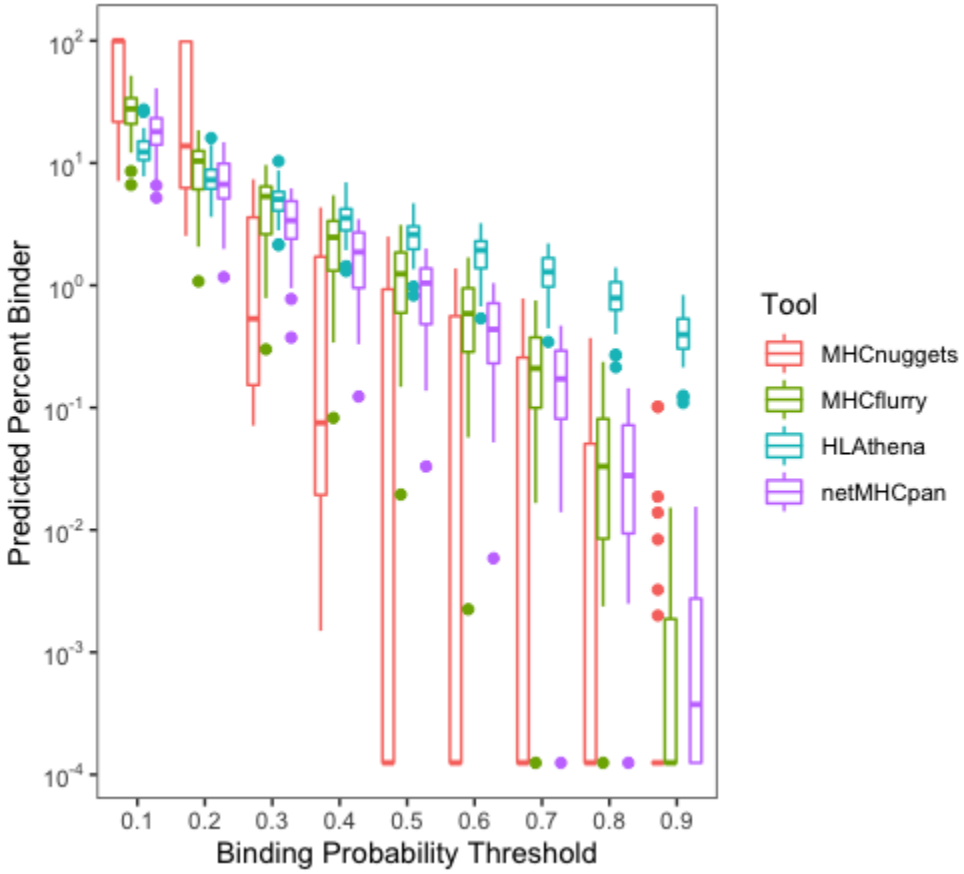
MHV	Betacoronavirus	Embecovirus	NCBI:txid11138	AAU06353.1	AAU06356.1	AAU06359.1	AAU06360.1	NP_045302.1
Rat-CoV	Betacoronavirus	Embecovirus	NCBI:txid31632	YP_003029844.1	YP_003029848.1	YP_003029850.1	YP_003029851.1	YP_003029852.1
Bat-BCoV	Betacoronavirus	Hibecovirus	NCBI:txid2501961	YP_009072438.1	YP_009072440.1	YP_009072442.1	YP_009072443.1	YP_009072446.1
Hedgehog-CoV	Betacoronavirus	Merbecovirus	NCBI:txid1965093	YP_009513008.1	YP_009513010.1	YP_009513016.1	YP_009513017.1	YP_009513018.1
MERS-CoV*	Betacoronavirus	Merbecovirus	NCBI:txid1335626	YP_009047202.1	YP_009047204.1	YP_009047209.1	YP_009047210.1	YP_009047211.1
HKU4	Betacoronavirus	Merbecovirus	NCBI:txid694007	YP_001039952.1	YP_001039953.1	YP_001039958.1	YP_001039959.1	YP_001039960.1
HKU5	Betacoronavirus	Merbecovirus	NCBI:txid694008	YP_001039961.1	YP_001039962.1	YP_001039967.1	YP_001039968.1	YP_001039969.1
GCCDC1	Betacoronavirus	Nobecovirus	NCBI:txid2501962	YP_009273004.1	YP_009273005.1	YP_009273007.1	YP_009273008.1	YP_009273009.1
HKU9	Betacoronavirus	Nobecovirus	NCBI:txid694006	YP_001039970.1	YP_001039971.1	YP_001039973.1	YP_001039974.1	YP_001039975.1
HKU14	Betacoronavirus	Unclassified	NCBI:txid1160968	YP_005454239.1	YP_005454245.1	YP_005454247.1	YP_005454248.1	YP_005454249.1
CDPHE15	Alphacoronavirus	Colacovirus	NCBI:txid1913643	YP_008439200.1	YP_008439202.1	YP_008439204.1	YP_008439205.1	YP_008439206.1
HKU10	Alphacoronavirus	Decacovirus	NCBI:txid1244203	YP_006908641.2	YP_006908642.1	YP_006908644.1	YP_006908645.1	YP_006908646.1
BtRf-AlphaCoV	Alphacoronavirus	Decacovirus	NCBI:txid2501926	YP_009199789.1	YP_009199790.1	YP_009199792.1	YP_009199793.1	YP_009199794.1
229E*	Alphacoronavirus	Duvinacovirus	NCBI:txid11137	ARU07599.1	ARU07601.1	ARU07603.1	ARU07604.1	ARU07605.1
LuchengRn-CoV	Alphacoronavirus	Luchacovirus	NCBI:txid1508224	YP_009336483.1	YP_009336484.1	YP_009336485.1	YP_009336486.1	YP_009336487.1
Ferret-CoV	Alphacoronavirus	Minacovirus	NCBI:txid1264898	YP_009256195.1	YP_009256197.1	YP_009256199.1	YP_009256200.1	YP_009256201.1
Mink-CoV	Alphacoronavirus	Minacovirus	NCBI:txid1913642	YP_009019180.1	YP_009019182.1	YP_009019184.1	YP_009019185.1	YP_009019186.1

Bat-CoV-1A	Alphacoronavirus	Minunacovirus	NCBI:txid694000	YP_001718603.1	YP_001718605.1	YP_001718607.1	YP_001718608.1	YP_001718609.1
HKU8	Alphacoronavirus	Minunacovirus	NCBI:txid694001	YP_001718610.1	YP_001718612.1	YP_001718614.1	YP_001718615.1	YP_001718616.1
BtMr-AlphaCoV	Alphacoronavirus	Myotacovirus	NCBI:txid2501927	YP_009199608.1	YP_009199609.1	YP_009199611.1	YP_009199612.1	YP_009199613.1
BtNv-AlphaCoV	Alphacoronavirus	Nyctacovirus	NCBI:txid2501928	YP_009201729.1	YP_009201730.1	YP_009201732.1	YP_009201733.1	YP_009201734.1
Porcine-EDV	Alphacoronavirus	Pedacovirus	NCBI:txid28295	NP_598309.2	NP_598310.1	NP_598312.1	NP_598313.1	NP_598314.1
BtCoV512	Alphacoronavirus	Pedacovirus	NCBI:txid693999	YP_001351683.1	YP_001351684.1	YP_001351686.1	YP_001351687.1	YP_001351688.1
HKU2	Alphacoronavirus	Rhinacovirus	NCBI:txid693998	YP_001552234.1	YP_001552236.1	YP_001552238.1	YP_001552239.1	YP_001552240.1
NL63*	Alphacoronavirus	Setracovirus	NCBI:txid277944	YP_003766.2	YP_003767.1	YP_003769.1	YP_003770.1	YP_003771.1
NL63-related	Alphacoronavirus	Setracovirus	NCBI:txid2501929	YP_009328933.1	YP_009328935.1	YP_009328937.1	YP_009328938.1	YP_009328939.1
FCoV	Alphacoronavirus	Tegacovirus	NCBI:txid12663	YP_004070193.2	YP_004070194.1	YP_004070197.1	YP_004070198.1	YP_004070199.1
TGE	Alphacoronavirus	Tegacovirus	NCBI:txid11149	NP_058422.1	NP_058424.1	NP_058426.1	NP_058427.2	NP_058428.1

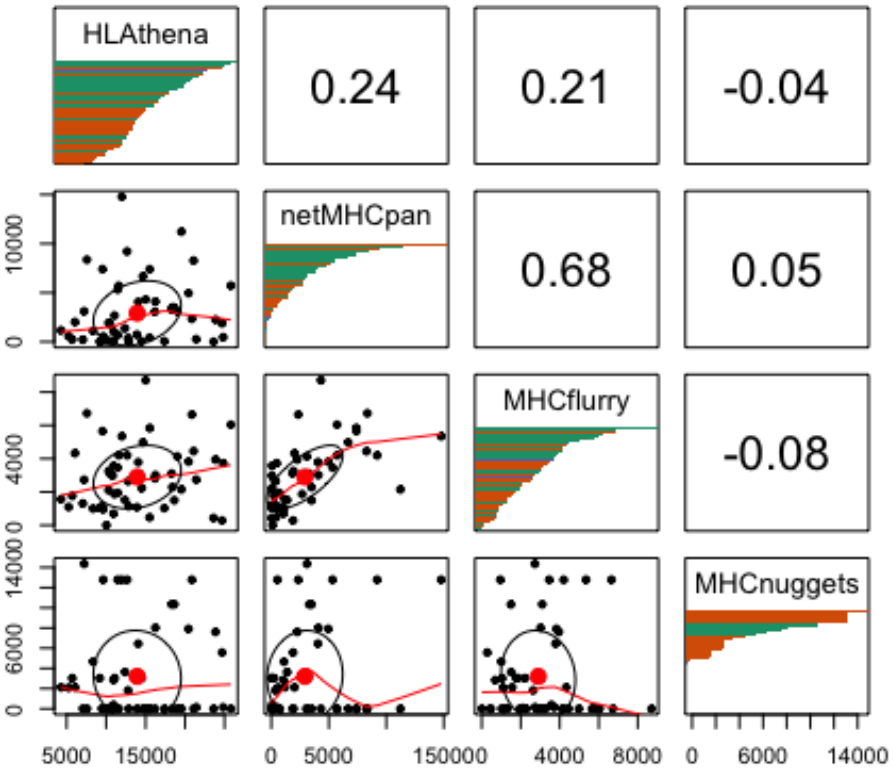
*Supplementary Table 2.5: Coronavirus taxonomy and sequence accession numbers for conserved coronavirus proteins. Name = coronavirus name used for sequence alignments (note that * indicates a known human coronavirus); Genus = Taxonomic classification as alpha- or betacoronavirus; Subgenus = further taxonomic classification at sub-genus level; Taxonomy ID = NCBI taxonomy ID accession for each virus species; ORF1ab = NCBI sequence accession number used for multi-sequence alignment for ORF1ab polyprotein; Spike = NCBI sequence accession number used for multi-sequence alignment for Spike protein; Envelope = NCBI sequence accession number used for multi-sequence alignment for Envelope protein; Membrane = NCBI sequence accession number used*

for multi-sequence alignment for Membrane protein; Nucleocapsid = NCBI sequence accession number used for multi-sequence alignment for Nucleocapsid protein.

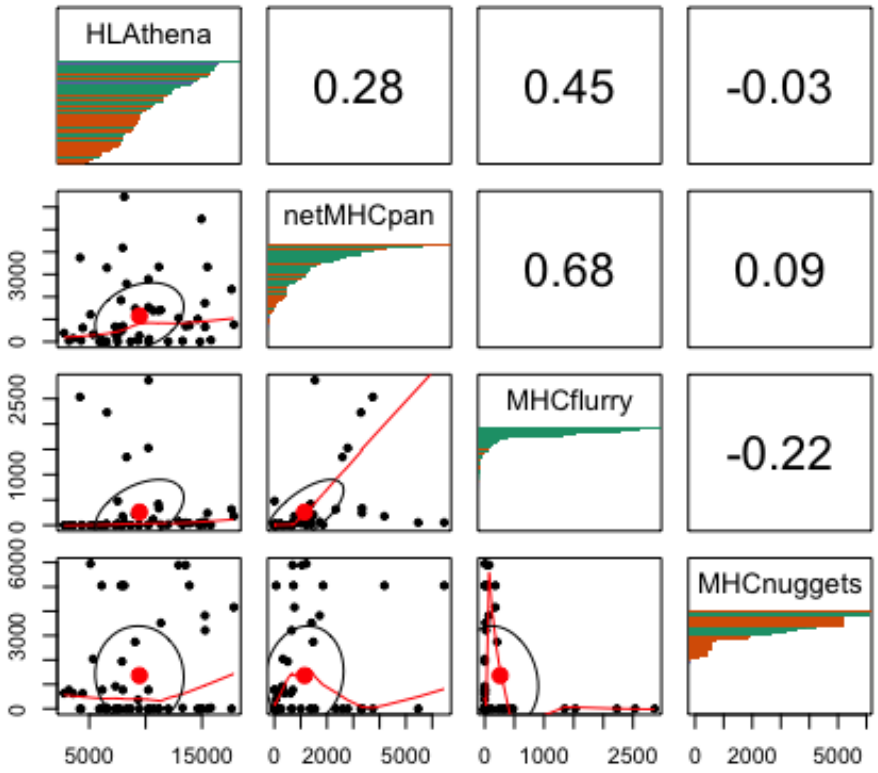
Supplementary File 3.1. Can be found at: https://onlinelibrary.wiley.com/doi/10.1111/tan.14574.



Supplementary Figure 4.1: Boxplots of the relationship between predicted binding and the threshold used to determine binding for random peptides. Each color represents a different tool with each boxplot representing the IQR of predicted percent peptides to bind for the given threshold.

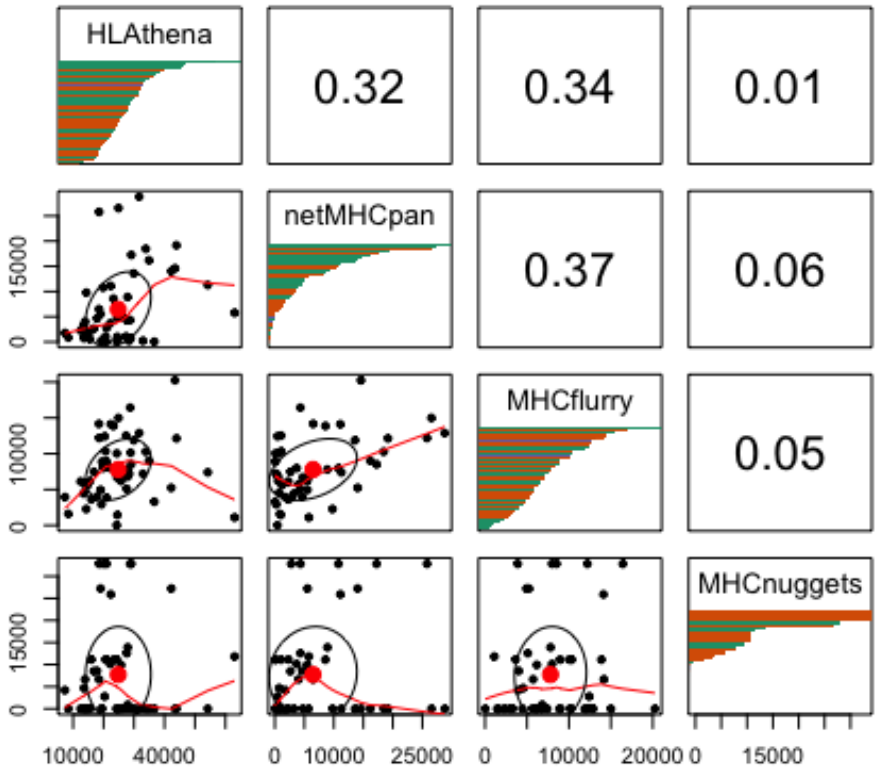

HLAthena
0.24
0.21
-0.04
netMHCpan
0.68
0.05
MHCflurry
-0.08
MHCnuggets
10000
0
4000
0
14000
0
6000
0
5000
15000
0
5000
150000
4000
8000
0
6000
14000

A

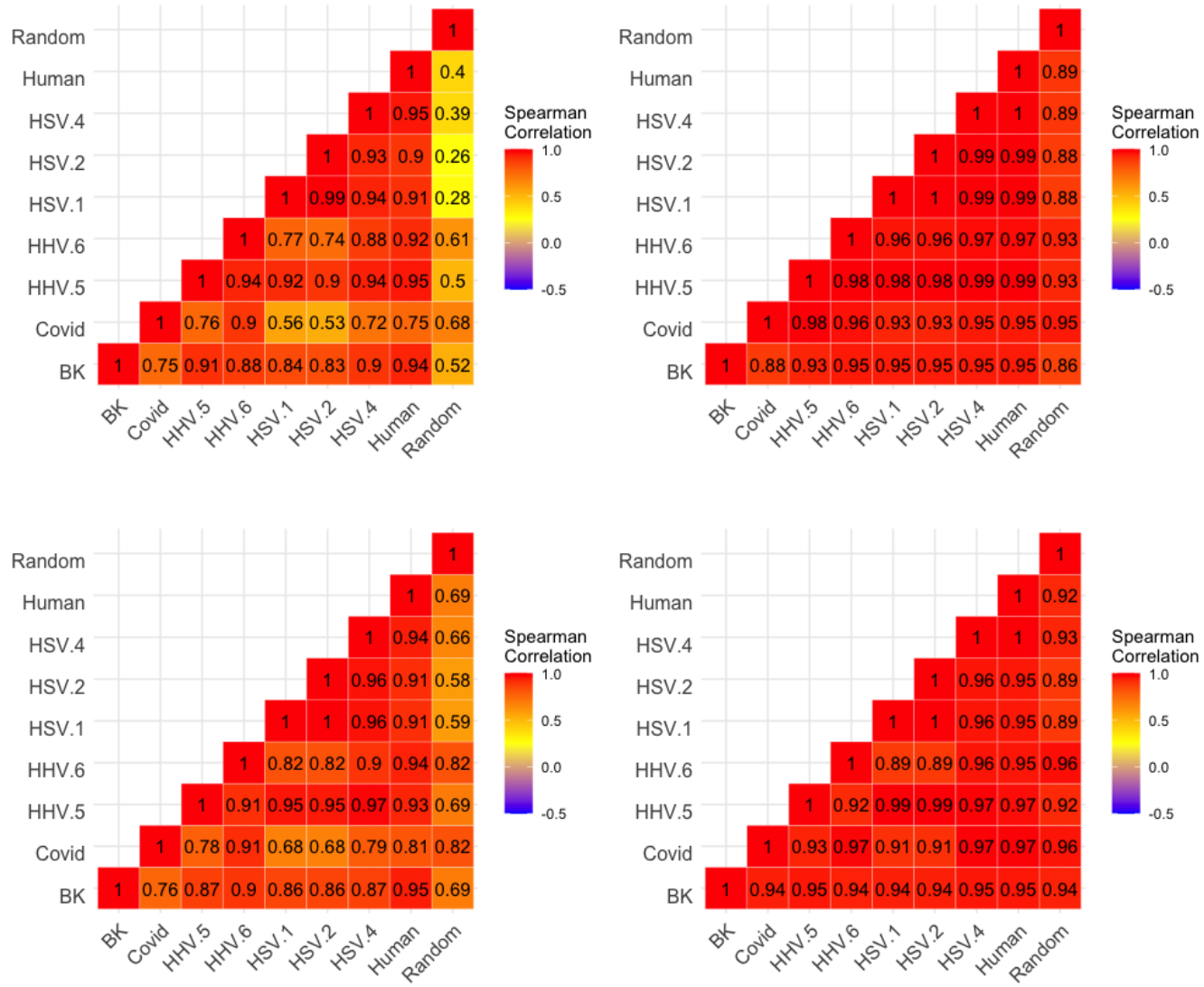


B

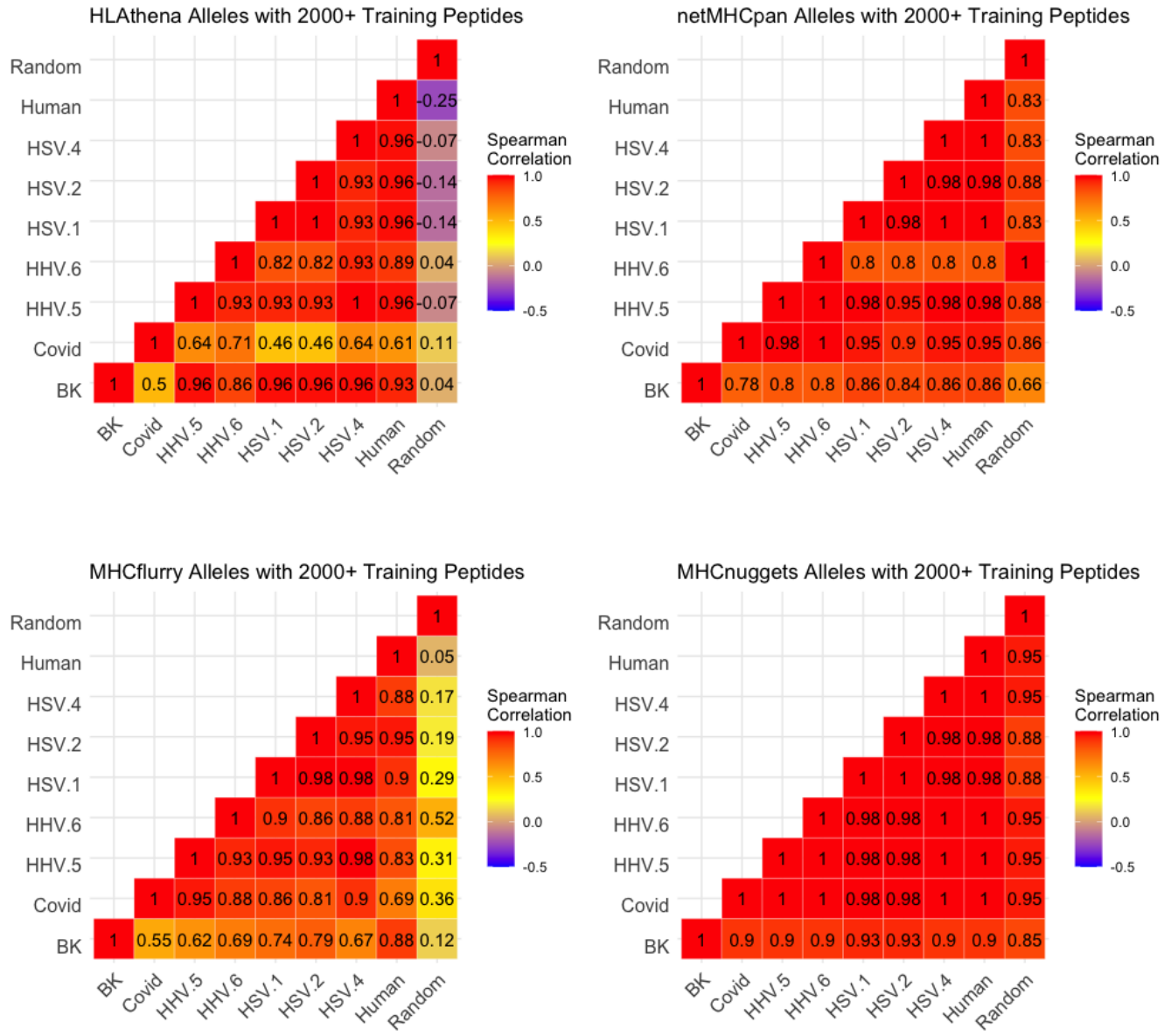
Supplementary Figure 4.2. Pairplot of HLA allelic presentation of 8-11mers from the random proteome. The lower left triangle displays scatter plots of peptides predicted to bind using 0.6 (A) and 0.7 (B) as cutoffs respectively between 2 tools with each point representing an HLA allele. The upper right triangle represents the Spearman correlation of the number of peptides predicted to bind to all alleles between tools. Note that MHCnuggets has a number of alleles with 0 random peptides predicted to bind. The diagonal panels show distribution of HLA allelic presentation from the random proteome for each tool. The number of peptides that putatively bind to each of the HLA alleles is shown along the x-axis as a series of horizontal bars with green, orange, and purple colors representing HLA-A, -B, and -C alleles, respectively, sorted in order of decreasing quantity of binders.



Supplementary Figure 4.3: Pairplot of HLA allelic presentation of 8-11mers from the human and viral proteome. The lower left triangle displays scatter plots of peptides predicted to bind (>= 0.5 binding probability score) between 2 tools with each point representing an HLA allele. The upper right triangle represents the Spearman correlation of the number of peptides predicted to bind to all alleles between tools. Note that MHCnuggets has a number of alleles with 0 random peptides predicted to bind. The diagonal panels show distribution of HLA allelic presentation from the random proteome for each tool. The number of peptides that putatively bind to each of the HLA alleles is shown along the x-axis as a series of horizontal bars with green, orange, and purple colors representing HLA-A, -B, and -C alleles, respectively, sorted in order of decreasing quantity of binders.

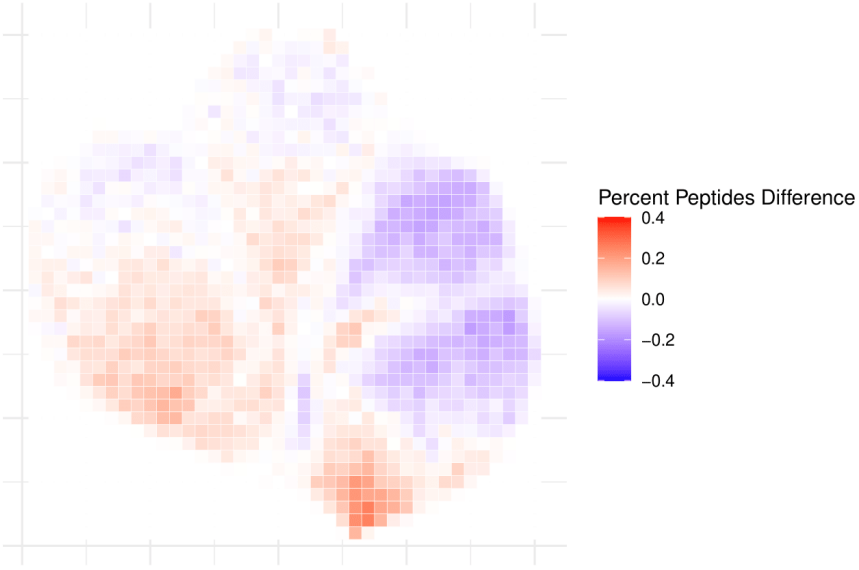


Supplementary Figure 4.4. Heatmaps of correlation between peptides for each species of predicted allelic promiscuity across alleles. A) Spearman correlation is shown between peptide sources for HLAthena-based predictions. Analogous data is shown for netMHCpan, MHCflurry, and MHCnuggets in panels B, C, and D, respectively.

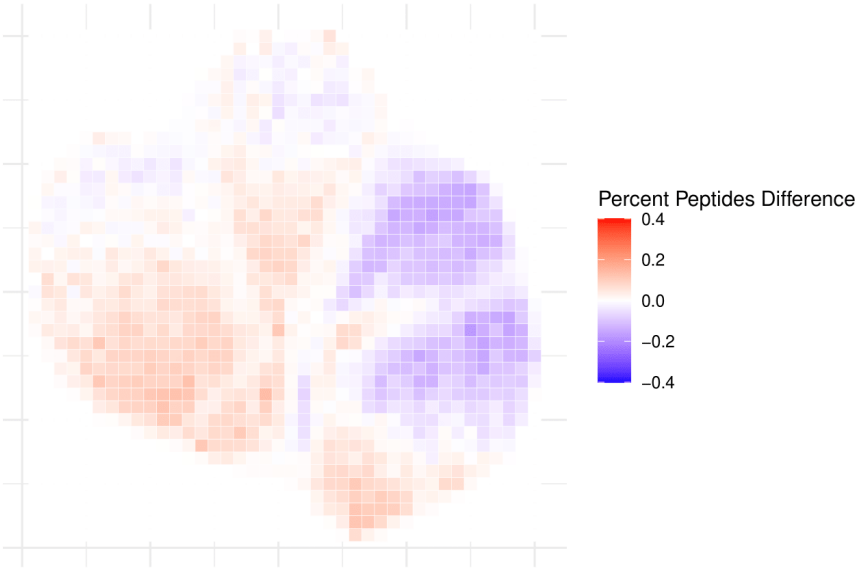


Supplementary Figure 4.5. Heatmaps of correlation between peptides for each species of predicted allelic promiscuity across alleles for which there was a minimum of 2000 peptides of training data. A) Spearman correlation is shown between peptide sources for HLAthena-based predictions. Analogous data is shown for netMHCpan, MHCflurry, and MHCnuggets in panels B, C, and D, respectively.

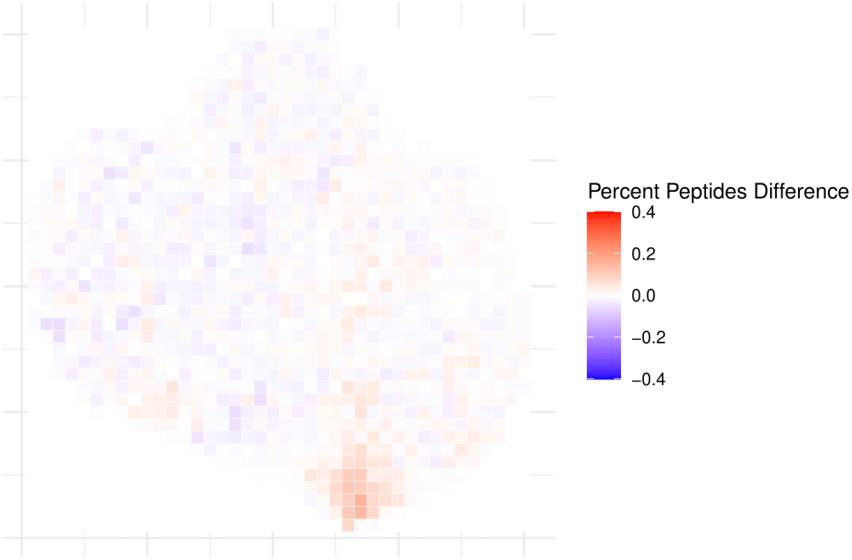
Viral vs Random 8mers



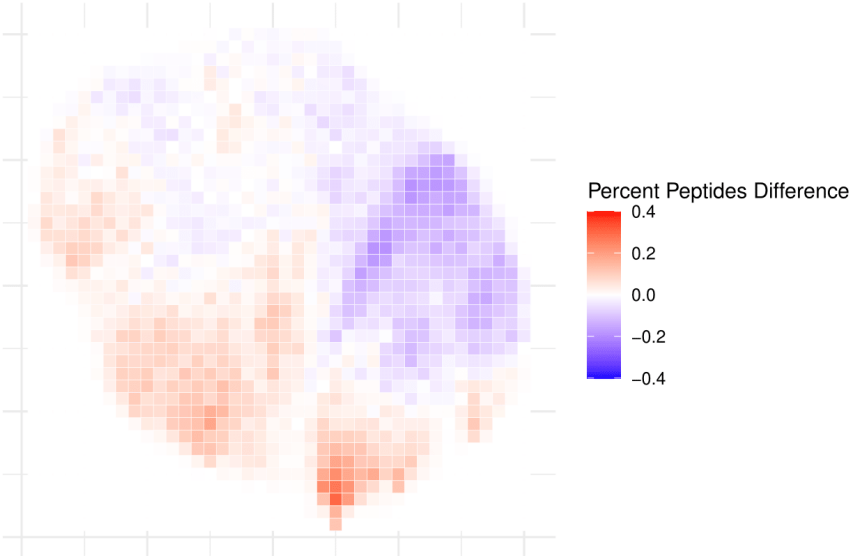
Human vs Random 8mers



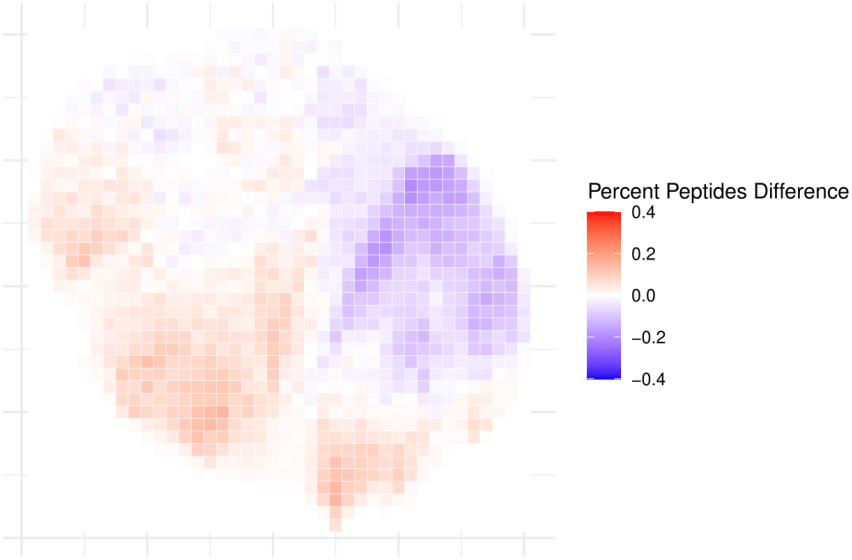
Viral vs Human 8mers



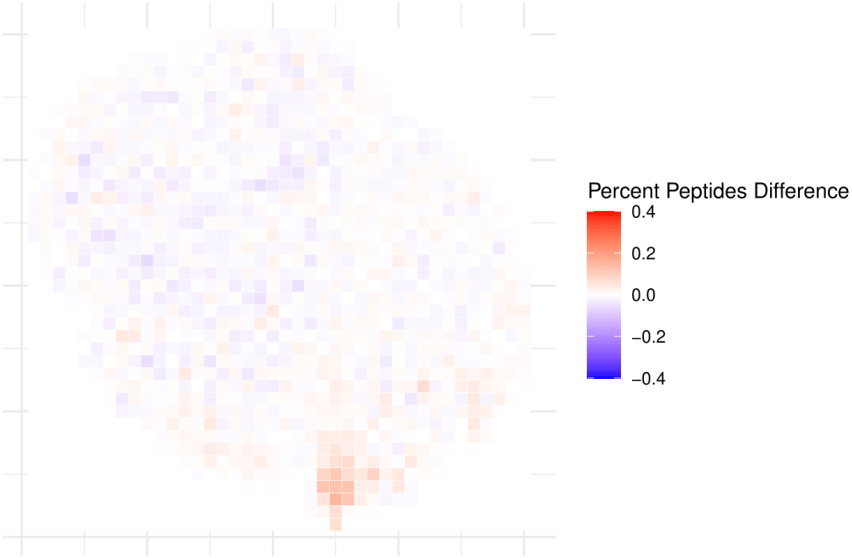
Viral vs Random 9mers



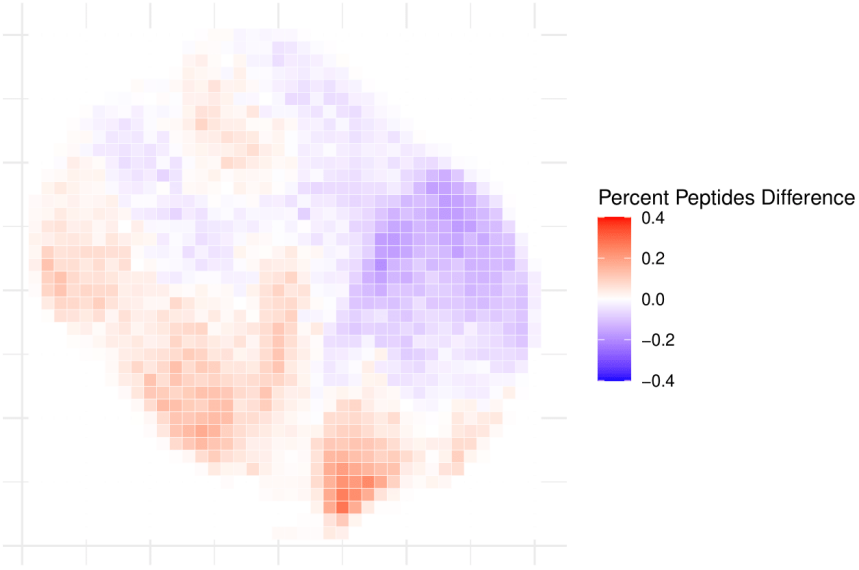
Human vs Random 9mers



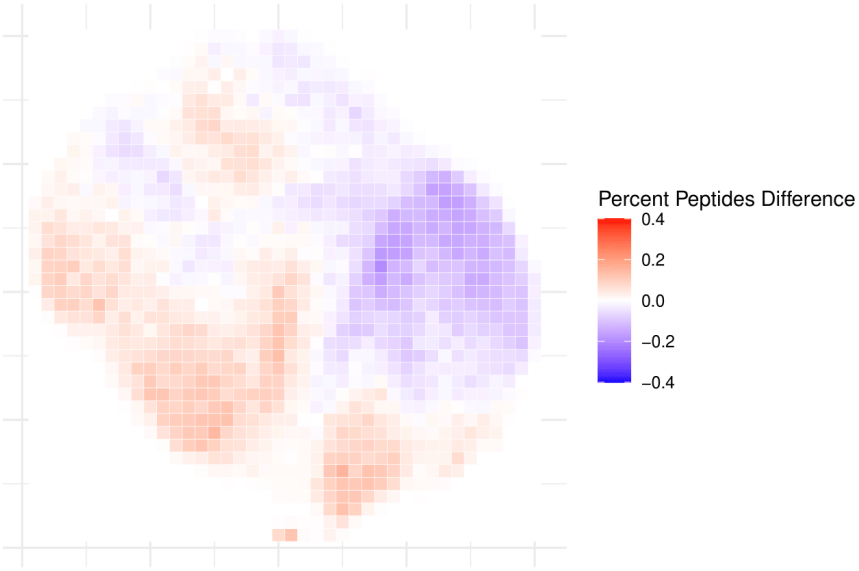
Viral vs Human 9mers



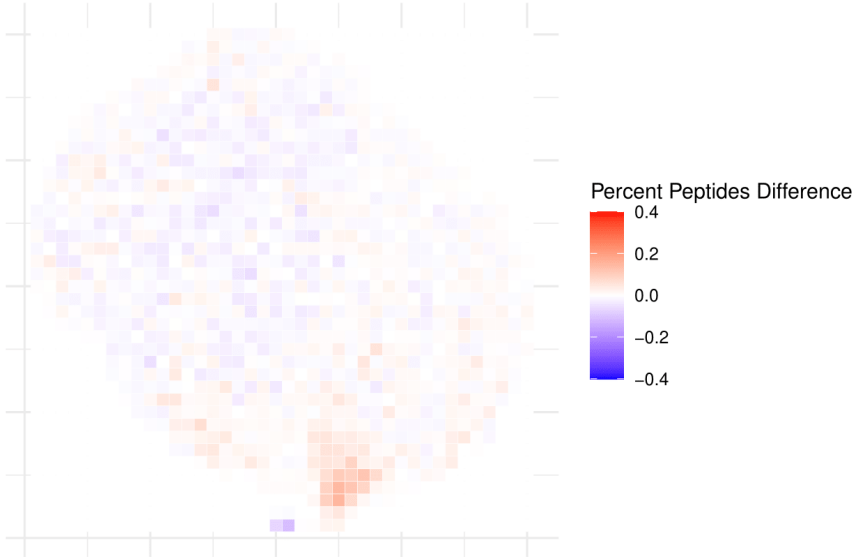
Viral vs Random 10mers



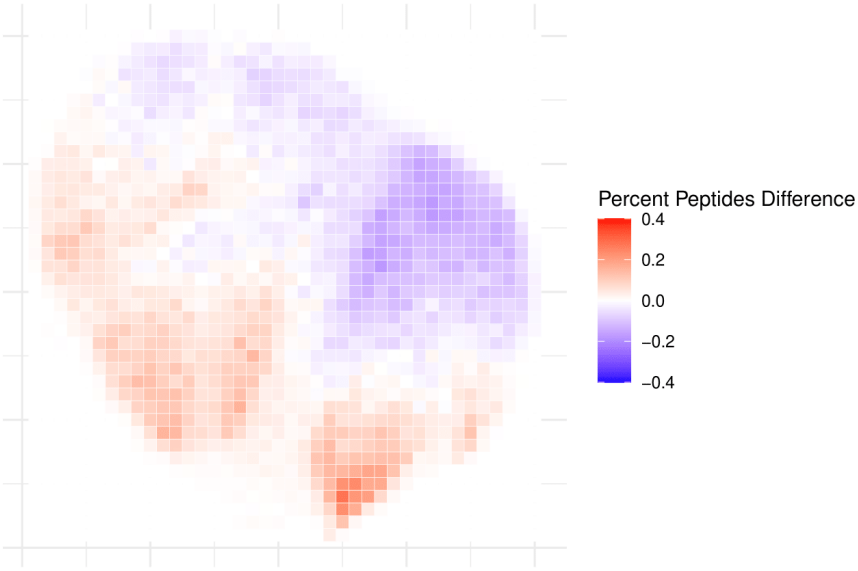
Human vs Random 10mers

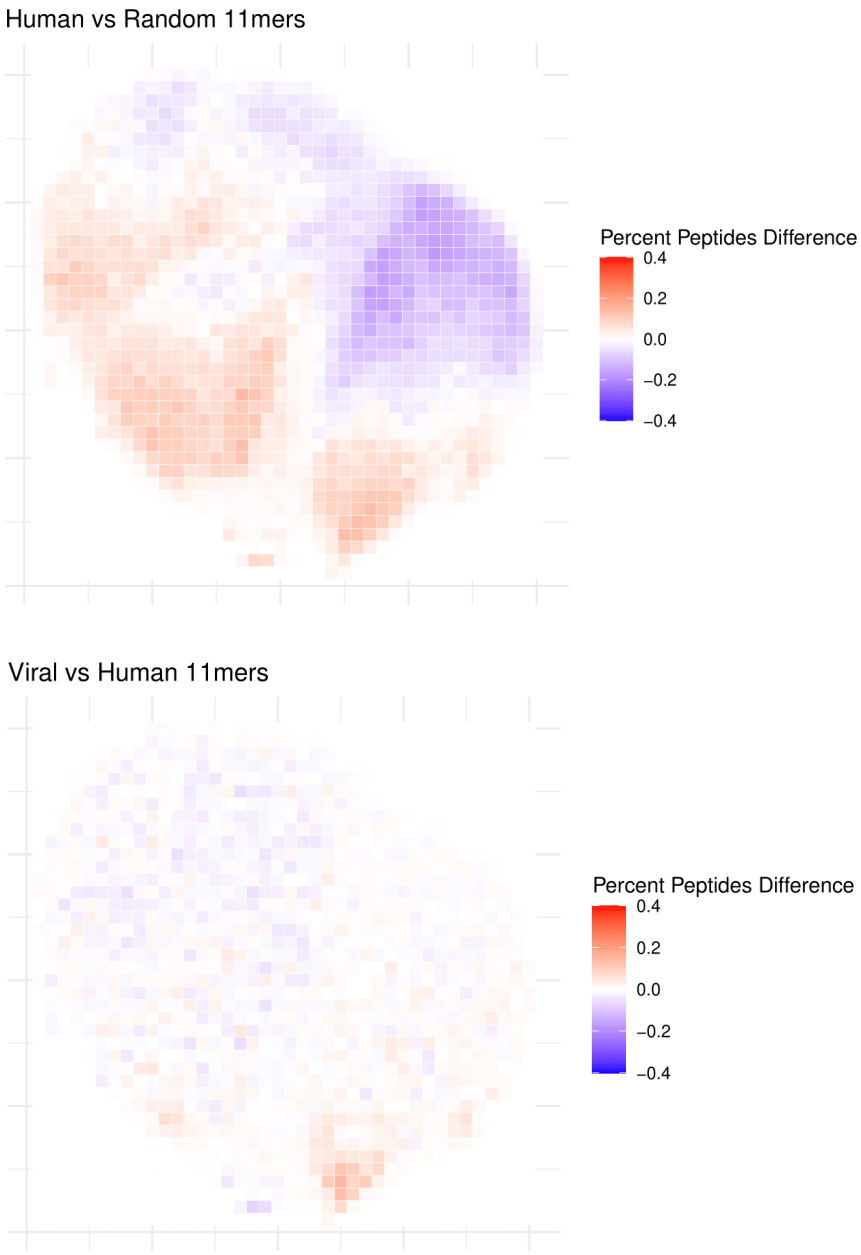


Viral vs Human 10mers

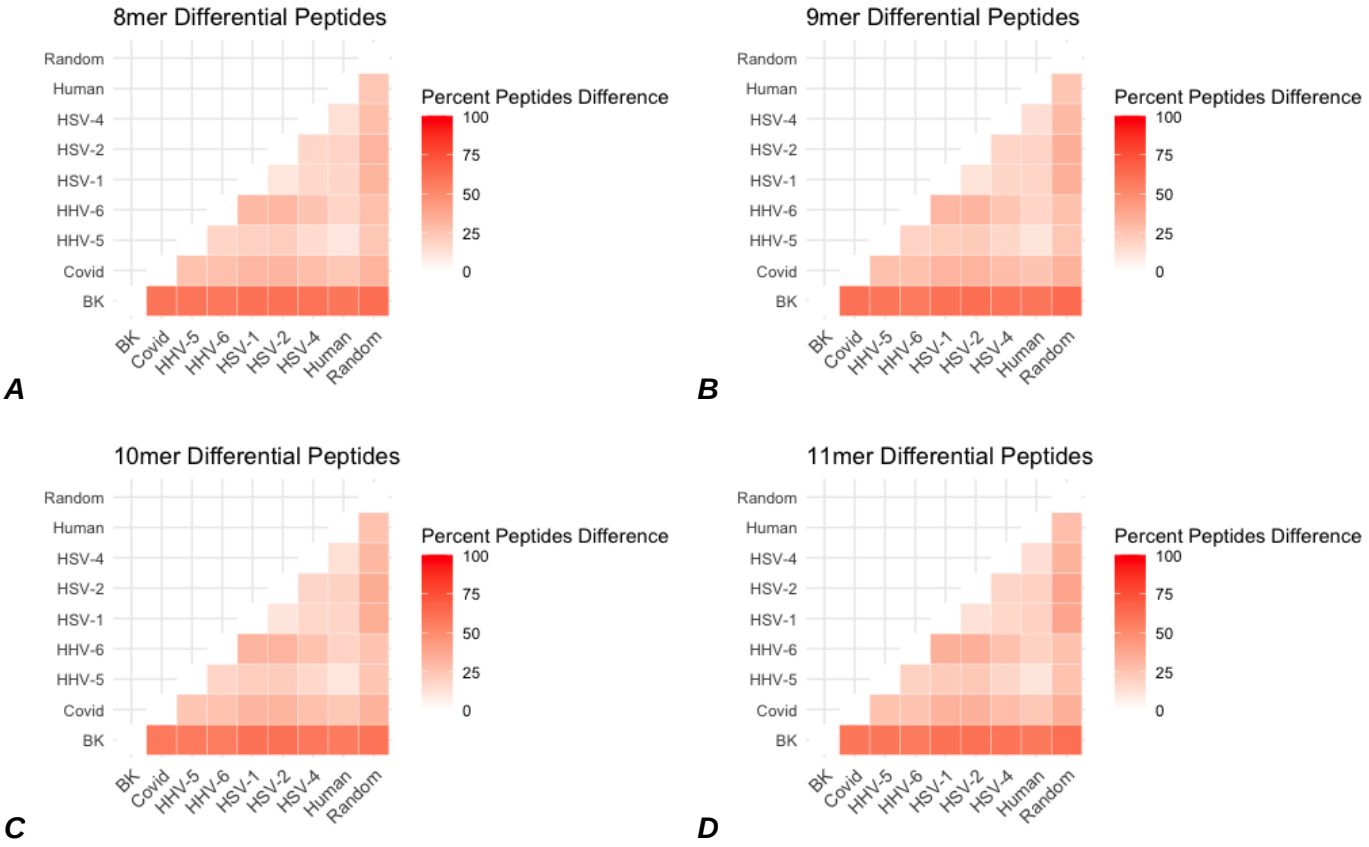


Viral vs Random 11mers

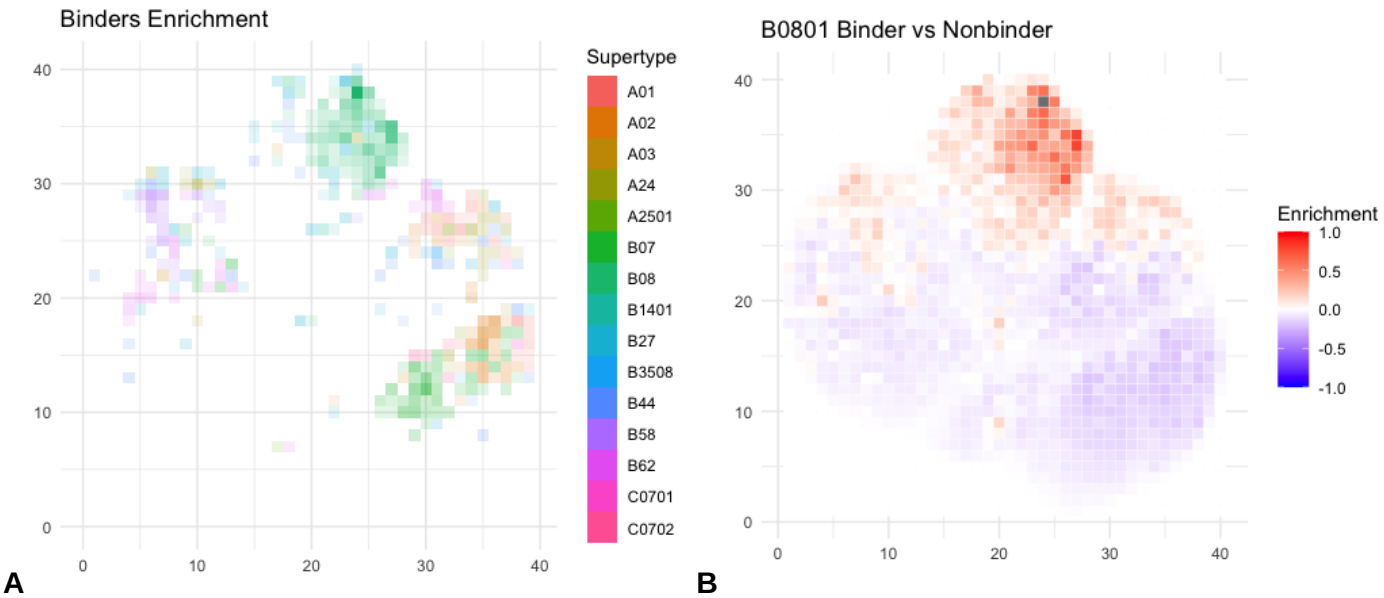


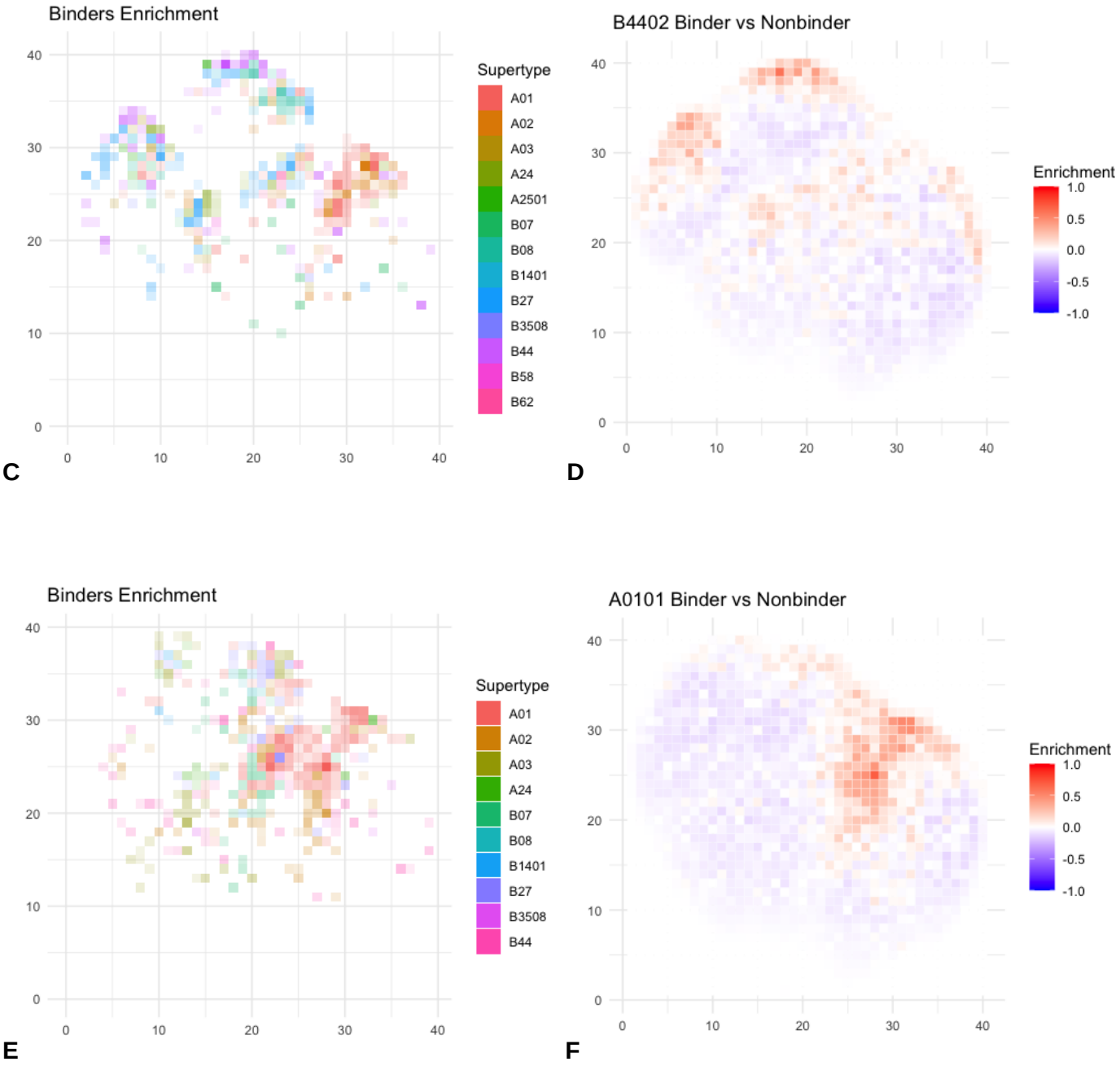


Supplementary Figure 4.6. Peptide physical property differences between different peptide sources. Each tile plot is composed of 1600 tiles, with each tile colored by the percent peptide difference between the 2 peptide sources in that particular tile. Red indicates an enrichment of the first label (e.g. viral vs human, viral enrichment will be red) while blue indicates enrichment of the second label.



Supplementary Figure 4.7. Peptide physical property difference by k-mer length. Each heatmap is the pairwise percent difference metric between each pair of peptide sets. The redder the value, the more difference in the percent difference metric.





Supplementary Figure 4.8. Differential distributions of physical properties for 8,10, and 11-mer peptides predicted to bind to HLA alleles. A,C,E) Tile plots highlighting binders enrichment 8, 10, and 11-mers respectively. The plotting coordinates represent the first two dimensions of a UMAP transform of peptide physical properties, which is divided into 1600 (40x40) equivalently-sized square bins (see Methods). For each bin where there is at least one HLA allele with >0.2% difference in proportion of all peptides predicted to bind v. non-binders, the identity of the most enriched allele is shaded in the color corresponding to that allele's supertype as corresponding to the legend. B,D,F)

Example plots of alleles with different distributions of binders for 8, 10, and 11-mers respectively. Each box represents enrichment as the percent peptide difference between predicted binders and non-binders for the given allele. The color scale shows the percent of peptides difference in the given box, with red meaning a larger number of predicted binders and blue meaning a larger number of predicted non-binders.

Supplementary Tables 4.1-4.3. Attached and will be found on bioRxiv