Inferring HLA Class I Binding Selection Pressure on the SARS-CoV-2 Spike Protein

Thomas Atkins, Jonas Braun, Denis Faerberg, Masha Larina
QCB First Year Students

November 2023
Option 2: Independent Research

1 Introduction

1.1 The HLA Class I System

Human Leukocyte Antigen (HLA) class I is a class of proteins responsible for presenting peptides to T-cells in order to elicit an adaptive immune response to a foreign pathogen [1]. In short, all nucleated cells continually digest a sample of the cell's proteome into short peptides, often 9 amino acids in length. These peptides then bind to HLA class I, which presents them on the surface of the cell. When a cell is infected with a virus, it will therefore present peptide fragments of the viral proteome (antigens) on the cell surface, so that CD8 T-cells with a T-cell receptor that can recognize the peptide form a complex between HLA class I and the T-cell receptor, kickstarting the adaptive immune response [2].

However, every HLA protein has a specific binding motif for peptides, meaning that a single HLA protein will only present a small subset of all possible peptides [3]. If individuals only had one copy of HLA, pathogens would easily be able to escape detection by mutating all peptides that bound to HLA. However, each individual has 6 copies of HLA class I, distributed across 3 loci: HLA-A, HLA-B, and HLA-C [4]. Furthermore, the HLA loci are the most heterogeneous genes in the human population, meaning an individual is very unlikely to inherit two copies of the same HLA allele. The most common allele, HLA-A*02:01 has a frequency of only 0.16, and there are 76 HLA alleles with frequencies greater than 0.01 in the US population (Figure 1) [5].

Because of this sequence diversity, an excess of computational tools have been developed to predict binding affinity, given an HLA allele and a peptide sequence. The most popular, NetMHCpan, utilizes multiple sources of data on peptide-HLA binding, and hosts a web server to easily make predictions [6]. This tools allows researchers to easily and quickly compute binding predictions for a given HLA allele across all peptides for an entire protein.

1.2 SARS-CoV-2 and COVID-19 Epidemiology

Understanding phylogenetic variation in different strains of COVID can help construct a more robust interpretation of our findings. A preliminary investigation of the phylogeny of SARS-CoV-2 is shown in Figure 2.

Observing the evolutionary path of the virus alone reveals the sheer mutability of the virus. Further studies have revealed the complex sequence diversity in COVID, discussing how, despite the apparently relatively low mutation rate of SARS-CoV-2, its efficacy lies in extreme selection

pressures and the introduction of multiple variants that increase the sequence complexity [8]. While further literary and experimental exploration of the true variability of the virus still needs to be done, we are able to establish a potential challenge in establishing binding selection pressure on the SARS-CoV-2 spike protein.

1.3 Previous Work on Predicting SARS-CoV-2 Binding to HLA Class I

Because of the relative ease of predicting binding affinity, researchers had created binding predictions and atlases for SARS-CoV-2 peptides only a few months after the initial genome sequence of was released [9] [10]. Using these *in silico* analyses for hypothesis generation, researchers have been able to discover many interactions between HLA and SARS-CoV-2. Most notably, researchers discovered that the allele HLA-B*15:01 recognizes a peptide that is conserved between SARS-CoV-2 and seasonal coronaviruses, leading to higher rates of asympotmatic infection in individuals with this allele [11]. However, the vast majority of work has been focused on associating HLA presentation of SARS-CoV-2 peptides with clinical outcomes.

Therefore, for our project, we propose to investigate the role of HLA binding in driving SARS-CoV-2 mutations. Our hypothesis is that peptides in the SARS-CoV-2 spike protein which bind to more common alleles of HLA will be under stronger selection pressure, and thus mutate more frequently. This could be used to predict what future variants are likely to arise, and can inform rational vaccine design.

2 Methods

Our project consists of four parts (shown in Figure 3). First, **Denis** will compute per-nucleotide mutation frequency of SARS-CoV-2 spike protein sequences from publicly available datasets (GISAID data for main strains of interest available at https://covariants.org/ and, time-permitting, more expansive GenBank data curated by Nextstrain at https://nextstrain.org/ncov/open/global/6m). Meanwhile, **Thomas** will use data from the National Marrow Donor Project (https://frequency.nmdp.org/) [5] to determine common alleles in the US population, and use NetMHCpan (https://services.healthtech.dtu.dk/services/NetMHCpan-4.1/) [6] to predict the binding affinity between the common alleles and the SARS-CoV-2 spike protein. To control for external factors, **Masha** will determine which residues in the SARS-CoV-2 spike protein are involved in binding to the ACE2 receptor (as those will likely also be under strong selection pressure). Finally, **Jonas** will build a mathematical model of the form

$$\text{mutation frequency} = \text{HLA binding frequency} + \text{ACE2 binding affinity} + \epsilon \tag{1}$$

and fit the data to the model, testing for statistical significance using bootstrap to validate the findings. All group members will be involved in writing and editing the final manuscript.

References

[1] Krensky AM. "The HLA system, antigen processing and presentation." *Kidney Int Suppl.* 1997 Mar;58:S2-7. PMID: 9067934.

[2] Rock KL, Reits E, Neefjes J. "Present Yourself! By MHC Class I and MHC Class II Molecules." *Trends Immunol.* 2016;37(11):724-737. doi:10.1016/j.it.2016.08.010

[3] Bassani-Sternberg M, Chong C, Guillaume P, et al. "Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity." *PLoS Comput Biol.* 2017;13(8):e1005725. Published 2017 Aug 23. doi:10.1371/journal.pcbi.1005725

[4] Choo SY. "The HLA system: genetics, immunology, clinical testing, and clinical implications." *Yonsei Med J.* 2007;48(1):11-23. doi:10.3349/ymj.2007.48.1.11

[5] Gragert, L., Madbouly, A., Freeman, J., & Maiers, M. (2013). "Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry." *Human Immunology*, 74(10), 1313–1320. http://dx.doi.org/10.1016/j.humimm.2013.06.025.

[6] Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. "NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data." *Nucleic Acids Res.* 2020;48(W1):W449-W454. doi:10.1093/nar/gkaa379

[7] Sironi, Manuela et al. "SARS-CoV-2 and COVID-19: A genetic, epidemiological, and evolutionary perspective." *Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases* vol. 84 (2020): 104384. doi:10.1016/j.meegid.2020.104384

[8] Chan, Ernest R et al. "COVID-19 infection and transmission includes complex sequence diversity." *PLoS genetics* vol. 18,9 e1010200. 8 Sep. 2022, doi:10.1371/journal.pgen.1010200

[9] 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; 12(3):254. https://doi.org/10.3390/v12030254

[10] Nguyen A, David JK, Maden SK, Wood MA, Weeder BR, Nellore A, Thompson RF. "Human Leukocyte Antigen Susceptibility Map for Severe Acute Respiratory Syndrome Coronavirus 2." *J Virol.* 2020 Jun 16;94(13):e00510-20. doi: 10.1128/JVI.00510-20. PMID: 32303592; PMCID: PMC7307149.

[11] Augusto, D.G., Murdolo, L.D., Chatzileontiadou, D.S.M. et al. "A common allele of HLA is associated with asymptomatic SARS-CoV-2 infection." *Nature* 620, 128–136 (2023). https://doi.org/10.1038/s41586-023-06331-x

3 Appendix

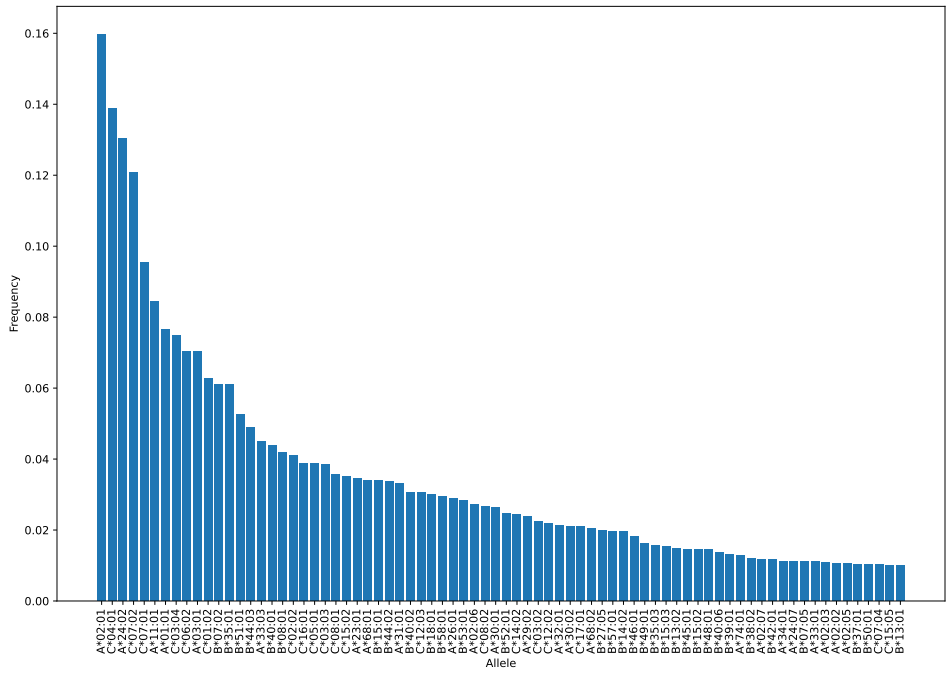

Figure 1: Distribution of all HLA alleles with frequencies ¿0.01 in the US populations. Data from the National Marrow Donor Project [5].

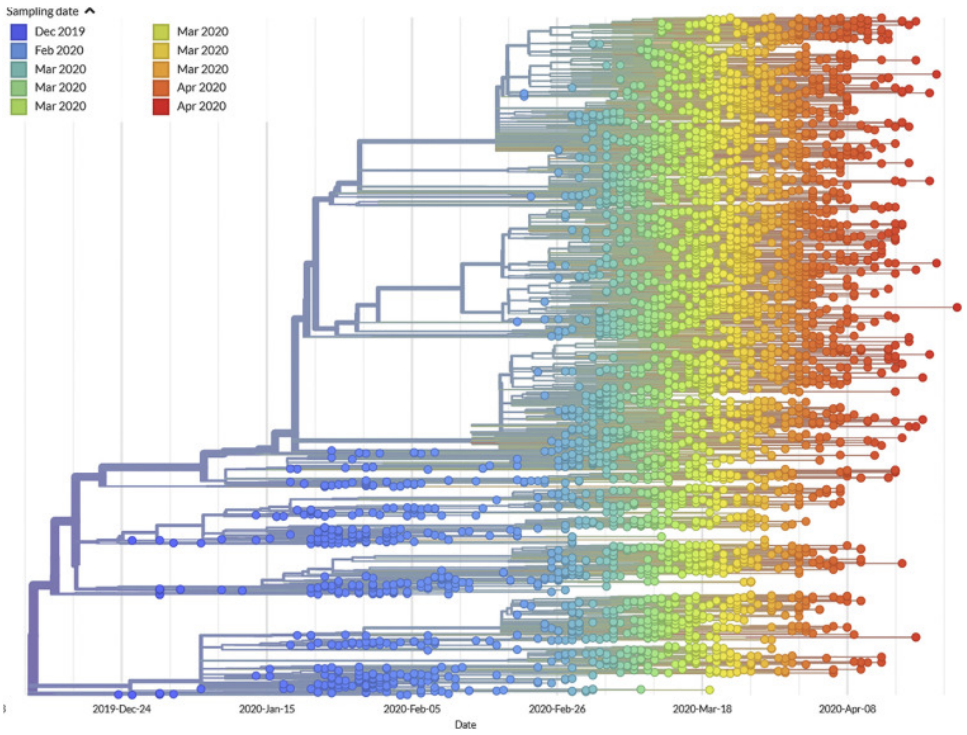

Figure 2: Reconstruction of SARS-CoV-2 phylogeny. "Isolates originated and initially diversified in China (purple), followed by multiple and independent introductions to Oceania (blue), Europe (green and yellow), and North America (red)." [7]

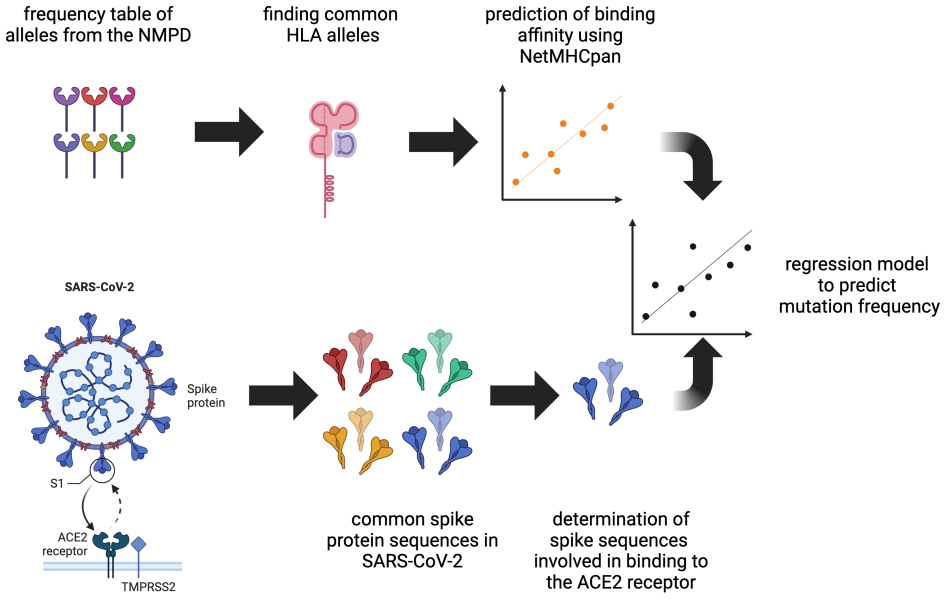

Figure 3: Overview of project workflow, created with BioRender.com.