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

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

November 2023
All code and data available at `https://github.com/ThomasKAtkins/551-project`

1 Short Exploratory Data Analysis & Data Visualization

1.1 Per-Nucleotide Mutation Frequency of SARS-CoV-2 Spike Protein Sequences

We downloaded a sample SARS-CoV-2 aligned genome sequences from NextStrain (`nextstrain.org`). We chose to filter to only include sequences collected in 2020, reasoning that the adoption of SARS-CoV-2 vaccines would provide a strong enough selection pressure to novel variants as to drown out any signal related to antigen presentation by HLA. This resulted in 178 sequences.

Next steps: translate the aligned nucleotide sequences to protein sequences. From there, determine the boundary of the spike protein and isolate only those sequences. Then, calculate Shannon entropy for each site in the protein (the metric we will use to determine selection pressure).

1.2 NetMHCpan for Binding Affinity

As a proof-of-concept, we used NetMHCpan [3] to predict binding between every 9-mer of the NCBI reference spike protein for SARS-CoV-2 (YP_009724390.1) and the common allele HLA-A*02:01 [4]. Our results are shown in Figure 1. We see that the vast majority of peptides contained within the protein are not predicted as likely binders to the given allele.

Next step: repeat this prediction for every common HLA-A, HLA-B, and HLA-C allele in the US population, and average them weighted by their population frequencies to generate an overall "binding score" for each amino acid within the genome. This score will represent the positions we hypothesize to be under increased selection pressure to escape presentation to T-cells.

1.3 Residues Involved in Binding to the ACE2 Receptor

The SARS-CoV-2 spike protein is able to infect cells by binding the host receptor ACE2. We hypothesize that we can construct a more robust model to capture the mutation frequency of the SARS-CoV-2 spike protein by incorporating information about the residues in the spike protein that bind to ACE2, as residued in the receptor-inding domain are under greater selection pressure.

We use data from a study that compared the SARS-CoV-2 and SARS-CoV RBD affinity for ACE2 binding by creating single amino acid substitution mutations and observing how introducing amino acid changes affected receptor binding [5]. The following data was generated by the results from the study by Yi et al.

The RBD binding domain encompasses residues 306-527 of the SARS-CoV protein; however, receptor-binding motif (RBM) (residues 438–506) is a part of the RBD containing most of the contacting residues of SARS-CoV-2 that bind to ACE2 [1]. Because the amino acids of the RBM shared an identity of only 47.8%, the researchers selected residues for functional analysis by mutation by aligning the sequence and selecting residues that had high conservation across sequence alignments. These 19 residues that were selected were all mutated and evaluated for binding affinity with ACE2. The mutation of the following 9 amino acid

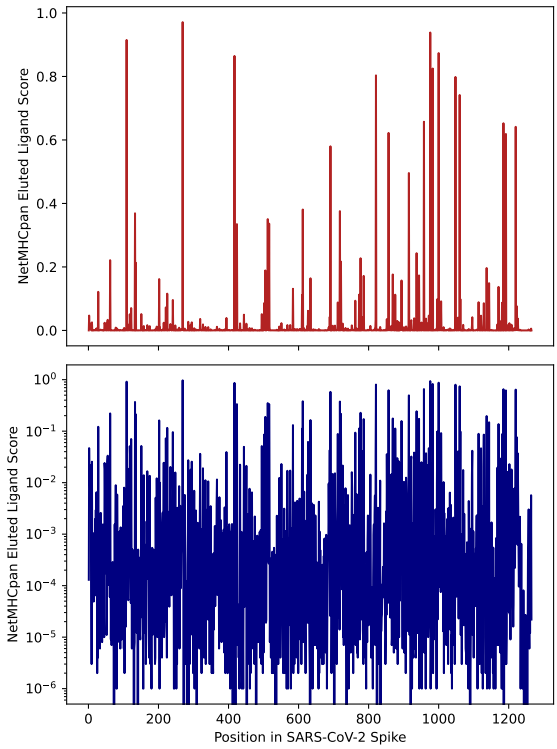

Figure 1: NetMHCpan predicted binding between every 9-mer of the reference SARS-CoV-2 spike protein and HLA-A*02:01. Raw scores are shown on top, log scores are shown on bottom.

mutations abolished binding of SARS-CoV-2 to ACE2: L455, F456, S459, Q474, A475, F486, F490, Q493 and P499. These results indicate that residues might be very important for SARS-CoV-2 binding to ACE2. The study also identified 6 substitution mutants (N439, L452, T470, E484, Q498 and N501) that enhanced finding affinity (Figure 2).

Next step: Because none of the raw data from this study was published, we will have to define and incorporate a discrete representation of these results in our final calculations.

1.4 Evaluating Statistical Significance

In our pursuit of understanding the mutation dynamics in the SARS-CoV-2 spike protein, we integrate our analyzed data on common alleles within the US population with identified residues in the spike protein crucial for ACE2 receptor binding. To assess the prevalence of common alleles, we computed the binding frequency for each available HLA allele in the US population. For the assessment of ACE2 binding affinity, we construct a binary vector highlighting residues in the SARS-CoV-2 spike protein involved in ACE2 receptor binding. The mutation frequency of the SARS-CoV-2 spike protein will be quantified using Shannon entropy.To elucidate potential correlations between mutation frequency and binding affinity in the SARS-CoV-2 spike protein, we will initiate our analysis with a linear regression model. Subsequently, model outcomes will be compared using metrics such as mean square error across various models, including Lasso, Ridge, and Random Forest Regression.

2 Anticipating Difficulties and Opportunities to Rescope

Because we are combining data from three different sources into one model, one primary difficulty we anticipate is ensuring that the data can be integrated into the model. To do this, we will ensure to align all our data to the spike protein boundaries computed from the alignment of the SARS-CoV-2-sequences from

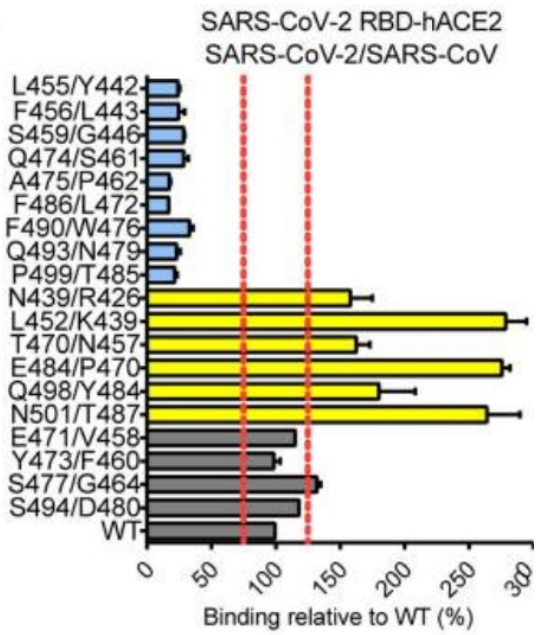

Figure 2: Figure from [5] showing the effects on binding of SARS-CoV2 with amino acid substitutions relative to WT as a percent. The red dotted lines mark 75% and 125% relative to WT.

NextStrain. Furthermore, we anticipate that the signal-to-noise ratio of the data might be too low to measure the impact of antigen presentation on selection pressure. To combat this, we can investigate obtaining more genome sequences, and looking into other confounding factors that might have been introduced in data processing (e.g., unknown bias in sequences selected to estimate mutation frequencies).

If our analysis still fails to produce substantial results, there are alternative questions that can be addressed using largely the same data - this can also be additional project scope for us, time-permitting. For instance, using the CoVariants data we can compare impact of notable mutations (there are 12 main ones that appear in many "variants of concern") on HLA affinity. Another option for us would be to use the Nextclade data on the latest curated "variant of concern" 23F (Omicron) as the frequency of mutations relative to pre- and post-vaccine strains instead of the sequence alignment. We could also attempt the analysis considering the geographic data, since many of our sources have it available. Beyond the above examples, our project has many other opportunities to pivot as well, since many high-content datasets are available and can be included or switched to fairly easily.

References

[1] Borkotoky, Subhomoi et al. "Interactions of angiotensin-converting enzyme-2 (ACE2) and SARS-CoV-2 spike receptor-binding domain (RBD): a structural perspective." Molecular biology reports vol. 50,3 (2023): 2713-2721. doi:10.1007/s11033-022-08193-4

[2] Lan, Jun et al. "Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor." Nature vol. 581,7807 (2020): 215-220. doi:10.1038/s41586-020-2180-5

[3] 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

[4] Romaniuk DS, Postovskaya AM, Khmelevskaya AA, Malko DB and Efimov GA (2019) Rapid Multiplex Genotyping of 20 HLA-A*02:01 Restricted Minor Histocompatibility Antigens. Front. Immunol. 10:1226. doi: 10.3389/fimmu.2019.01226

[5] Yi, Chunyan et al. "Key residues of the receptor binding motif in the spike protein of SARS-CoV-2 that interact with ACE2 and neutralizing antibodies." Cellular & molecular immunology vol. 17,6 (2020): 621-630. doi:10.1038/s41423-020-0458-z

3 Appendix