

# Estimating the binding of Sars-CoV-2 peptides to HLA class I in human subpopulations using artificial neural networks

Caterina A. M. La Porta, Stefano Zapperi

---

## Summary

**Initial Submission:** Received Apr. 26, 2020  
Preprint: N/A  
Scientific editor: Ernesto Andrianantoandro, Ph.D.

**First round of review:** Number of reviewers: Two  
*Two confidential, zero signed*  
Revision invited May 28, 2020  
*Major changes anticipated*  
Revision received Jun. 6, 2020

**Second round of review:** Number of reviewers: One  
*One original, zero new*  
*One confidential, zero signed*  
Accepted Aug. 13, 2020

**Data freely available:** Yes  
**Code freely available:** Yes

---

*This Transparent Peer Review Record is not systematically proofread, type-set, or edited. Special characters, formatting, and equations may fail to render properly. Standard procedural text within the editor's letters has been deleted for the sake of brevity, but all official correspondence specific to the manuscript has been preserved.*

---

## Editorial decision letter with reviewers' comments, first round of review

Dear Dr. La Porta,

I'm enclosing the comments that reviewers made on your paper, which I hope you will find useful and constructive. As you'll see, they express interest in the study, but they also have a number of criticisms and suggestions. Based on these comments, it seems premature to proceed with the paper in its current form; however, if it's possible to address the concerns raised with additional data and/or analysis, we'd be interested in considering a revised version of the manuscript.

As a matter of principle, I usually only invite a revision when I'm reasonably certain that the authors' work will align with the reviewers' concerns and produce a publishable manuscript. In the case of this manuscript, the reviewers and I have make-or-break concerns regarding data and code availability, justification for cutoffs and methodological choices, and rationale for the number of HLA alleles analyzed. In addition, I've highlighted portions of the reviews that strike me as particularly critical.

***We appreciate that the COVID-19 pandemic challenges and limits what you and your lab can do, so to make sure we're absolutely on the same page about the feasibility of revisions, let's schedule a Zoom call at our earliest mutual convenience.***

Do note that we generally consider papers through only one major round of revision, so the revised manuscript would be either accepted or rejected based on the next round of comments we receive from the reviewers. If you have any questions or concerns, please let me know. More technical information and advice about resubmission can be found below my signature. Please read it carefully, as it can save substantial time and effort later.

I look forward to seeing your revised manuscript.

All the best,

Ernesto Andrianantoandro, Ph.D.  
Scientific Editor, Cell Systems

---

### Reviewers' comments:

Reviewer #1: Heterogeneity of coronavirus peptide binding across human haplotypes

Caterina A. M. La Porta, Stefano Zapperi

## COMMENTS TO THE AUTHOR

Please be aware that a 04/01/2020 preprint identified SARS-CoV-2 epitopes across 9360 HLA class I alleles (<https://doi.org/10.1101/2020.03.30.016931>). The preprint identifies 6,748 peptide-HLA combinations that bind to HLA with < 500 nM affinity. This preprint may be relevant to your work.

## OVERVIEW

Thank you for the opportunity to review the manuscript.

La Porta et. al perform bioinformatic analyses of SARS-Cov-2 MHC class I epitopes. The authors identify HLA haplotypes with greater or lesser probability to present peptides from SARS-Cov-2 and then compare distributions across countries and also with other coronaviruses. On this basis, the authors describe a potentially important aspect of immunity to SARS-Cov-2 that varies across individuals. The work is thought-provoking, hypothesis-generating, and naturally, quite timely. The analysis appears generally technically sound.

## DATA &amp; CODE AVAILABILITY

Raw peptide affinity data are not available.

## REQUIRED MAJOR REVISIONS

None.

## MINOR REVISIONS, SUGGESTIONS &amp; COMMENTS

1. I strongly suggest that raw peptide affinity data be made available to maximize utility in this fast-paced field.
2. MHC-peptide affinity prediction binding cutoffs are certainly debatable, but a minimum of < 500nM or percentile rank < 1% would be more widely accepted. It is likely that the reported results will not differ meaningfully using one of these standard cutoffs. In the interest of time, please simply check that this is indeed the case.

Reviewer #2: Referee Report: Heterogeneity of coronavirus peptide binding across human haplotypes

Summary: the author report a calculation of predicted immunogenicity of SARS-Cov-2 Antigens on a set of 79 HLA alleles. The authors further link the potential number of predicted alleles to the immune response of an individual, and use this to infer the extent to which particular geographic regions (countries) are more or less susceptible to SARS-Cov-2 infection.

Major concerns:

1. The authors state in their introduction that HLA-A has 3285 alleles, HLA-B has 4077 alleles, and HLA-C has 2801 alleles, but in their analysis they only analyze 79 MHC I alleles, with the only rationale that "these alleles are supported by both methods". However, this represents less than 1% of known HLA alleles. Without knowing what fraction of each population this 1% covers, this raises the prospect that the author's work is not in broadly geographically representative and in fact potentially inaccurate for the geographies discussed.
2. The authors state "In the present paper, we propose a method to identify the dependence on HLA class I polymorphic alleles of the individual immune response to SARS-CoV-2. We focus on this class of MHC since they are expressed by all nucleated cells, including antigen presenting cells." This rationale to only focus on HLA class I alleles seems rather weak to me, since it is known that both CD4 and CD8 T-cells play an important role in response to COVID 19. Inclusion of MHC II predictions would be important if the authors wish to be able to make claims like those included in this paper.
3. The authors state that they use NetMHCpan and MHCFlurry, but there are many more up-to-date tools available, including MixMHCpred, NetMHCpan-4.0, and MHCFlurry-EL. This raises the likelihood that the predicted values of antigen binding in the authors manuscript are not accurate, as well.
4. The authors state "Furthermore, the distributions differ between the various alleles (Fig. S1, S2 and S3), confirming the presence of heterogeneous binding pattern. To encapsulate the binding affinity distributions into a simple parameter, we counted all the peptides displaying a strong binding affinity ( $IC_{50} < 1000Mm$ ) for each of the 79 alleles."
  - a. The authors do an admirable job showing the differences in predicted binding affinity by allele, but then disregard this analysis and simply choose a 1000 nM cutoff, seemingly out of nowhere. If the authors are to include this analysis of distribution, they should use it to support their choice of criteria for a "strong binding antigen".
  - b. More over, the choice of 1000nM is far outside the norm for this type of analysis. 500 nM is a more typical naïve choice, but it has been shown that the appropriate value of this parameter is very likely dependent pathogen-dependent. Strong support for the choice of cutoff must be given for this analysis to be valid.
5. The authors next consider 13 alleles for assessment of T-cell recognition, and show the results from 4 of these calculation. Once again, this represents an exceptionally small number of HLA alleles (~0.1%) and it is unclear what fraction of the worldwide population is included in this calculation, or what biases, if any may skew the results.
6. When it comes to understanding the geographic distribution of alleles, the authors analyze only 17 alleles. This is virtually an order of magnitude lower than the initial 79, and means that, in the context of understanding the global ability to recognize and eliminate the COVID19 pathogen, the authors consider again around 0.1% of all possible alleles. Without considering the results of many more alleles, it is the feeling of this reviewer that this calculation is certainly inaccurate and potentially deeply misleading.

7. Even if all of the above was accepted (it is not, by this reviewer), there is essentially no difference between the majority of populations considered by the authors, and **it is unclear how the error bars were generated.**

8. Data availability: the authors make no comment on the availability of data for this work, nor the code to reproduce it. **Both must be made explicitly and completely available.**

---

### Authors' response to the reviewers' first round comments

Attached.

---

### Editorial decision letter with reviewers' comments, second round of review

Dear Dr. La Porta,

I'm very pleased to let you know that the reviews of your revised manuscript are back, the peer-review process is complete, and only a few minor, editorially-guided changes are needed to move forward towards publication.

In addition to the final comments from the reviewers, I've made some suggestions about your manuscript within the "Editorial Notes" section, below. Please consider my editorial suggestions carefully, ask any questions of me that you need, make all warranted changes, and then upload your final files into Editorial Manager. ***We hope to receive your files within 5 business days, but we recognize that the COVID-19 pandemic may challenge and limit what you can do. Please email me directly if this timing is a problem or you're facing extenuating circumstances.***

I'm looking forward to going through these last steps with you. More technical information can be found below my signature, and please let me know if you have any questions.

All the best,

Ernesto Andrianantoandro, Ph.D.  
Scientific Editor, Cell Systems

---

**Editorial Notes**

*Title:* Your title is too generic. suspect it could be more effective. Please revise to better encapsulate the main advance and differentiate this paper from other similar papers. Perhaps include something about analysis in human subpopulations and geography? As you re-consider your title, note that an effective title is easily found on Pubmed and Google. A trick for thinking about titles is this: ask yourself, "How would I structure a Pubmed search to find this paper?" Put that search together and see whether it comes up is good "sister literature" for this work. If it does, feature the search terms in your title. You also may wish to consider that PubMed is sensitive to small differences in search terms. For example, "NF-kappaB" returned ~84k hits as of March, 2018, whereas "NFkappaB" only returned ~8200. Please ensure that your title contains the most effective version of the search terms you feature.

*Manuscript Text:*

House style disallows editorializing within the text (e.g. strikingly, surprisingly, importantly, etc.), especially the Results section. These terms are a distraction and they aren't needed—your excellent observations are certainly impactful enough to stand on their own. Please remove these words and others like them. "Notably" is suitably neutral to use once or twice if absolutely necessary.

*STAR Methods:*

Please revise to adhere to our most recent STAR Methods format: <https://www.cell.com/star-authors-guide>

**Thank you!**

---

**Reviewer comments:**

Reviewer #1: The authors have adequately addressed my concerns.

The Github repository provides well-annotated code necessary to reproduce the analysis and figures. It does, however, require a FASTA file of sequences that does not appear to be included in the repository, and while it would be possible to compile the sequences from the sources mentioned in the methods section, it would enhance reproducibility if this file were included alongside the repository or otherwise made available.

Thank you for the opportunity to review the manuscript.

**Reviewer #1:**

**OVERVIEW**

**La Porta et. al perform bioinformatic analyses of SARS-Cov-2 MHC class I epitopes. The authors identify HLA haplotypes with greater or lesser probability to present peptides from SARS-Cov-2 and then compare distributions across countries and also with other coronaviruses. On this basis, the authors describe a potentially important aspect of immunity to SARS-Cov-2 that varies across individuals. The work is thought-provoking, hypothesis-generating, and naturally, quite timely. The analysis appears generally technically sound.**

We thank the referee for his/her positive assessment of the paper.

**Please be aware that a 04/01/2020 preprint identified SARS-CoV-2 epitopes across 9360 HLA class I alleles (<https://doi.org/10.1101/2020.03.30.016931>). The preprint identifies 6,748 peptide-HLA combinations that bind to HLA with < 500 nM affinity. This preprint may be relevant to your work.**

We thank the referee for pointing this reference to us. We quote it in the introduction of the revised manuscript (see page 3)

**DATA & CODE AVAILABILTY**

**Raw peptide affinity data are not available.**

In the revised manuscript, we include raw peptide affinity in the supplement and deposit our analysis code into a public repository (<https://github.com/ComplexityBiosystems/hla-covid>)

**MINOR REVISIONS, SUGGESTIONS & COMMENTS**

- 1. I strongly suggest that raw peptide affinity data be made available to maximize utility in this fast-paced field.**

Raw peptide affinities will be published with the manuscript (**supplementary data 1**).

- 2. MHC-peptide affinity prediction binding cutoffs are certainly debatable, but a minimum of < 500nM or percentile rank < 1% would be more widely accepted. It is likely that the reported results will not differ meaningfully using one of these standard cutoffs. In the interest of time, please simply check that this is indeed the case.**

The referee makes a good point. We have checked that our results are robust with respect to the choice of the cutoff. To demonstrate this, we provide in the revised supplement results obtained with a cutoff of 500nM (see Fig. S5). Apart from small quantitative variations, we do not find substantial differences. Since any choice of cutoff is arbitrary, we also propose a new measure of the total binding affinity for each HLA molecule which weights the combined effect of all the peptides. A

detailed discussion is reported in the Methods at **page 9**. The ranking between the different HLA molecules is again confirmed by this cutoff-independent method (**see Fig. S6**).

#### **Reviewer #2:**

##### **Major concerns:**

**1. The authors state in their introduction that HLA-A has 3285 alleles, HLA-B has 4077 alleles, and HLA-C has 2801 alleles, but in their analysis they only analyze 79 MHC I alleles, with the only rationale that "these alleles are supported by both methods". However, this represents less than 1% of known HLA alleles. Without knowing what fraction of each population this 1% covers, this raises the prospect that the author's work is not in broadly geographically representative and in fact potentially inaccurate for the geographies discussed.**

While the 79 alleles we study represent 1% of possible alleles, they are the most frequent alleles across human populations. In particular, the HLA-A alleles that we considered are present in 80-99% of the populations, the HLA-B alleles in 65-85% and the HLA-C alleles in 45-60%. We thank the referee for pointing out that this issue that was not properly discussed in the first version of the manuscript. In the revised version of the manuscript, we report explicitly in a supplementary figure (see **Fig. S1**) the frequencies of the alleles considered in our study across different populations. The figure is discussed at **pages 3-4**.

**2. The authors state "In the present paper, we propose a method to identify the dependence on HLA class I polymorphic alleles of the individual immune response to SARS-CoV-2. We focus on this class of MHC since they are expressed by all nucleated cells, including antigen presenting cells." This rationale to only focus on HLA class I alleles seems rather weak to me, since it is known that both CD4 and CD8 T-cells play an important role in response to COVID 19. Inclusion of MHC II predictions would be important if the authors wish to be able to make claims like those included in this paper.**

The goal of the present work was to define a strategy to predict the individual susceptibility to SARS-CoV-2 infection rather than provide a complete picture of the immune response to this virus. We decided to investigate HLA class I molecules because they are expressed by all nucleated cells and because CD8+ T-cell mediated immunity is directly involved in the response to the virus. In the revised manuscript, we acknowledge that studying HLA class I molecules does not provide a complete picture of the immune response that also depends on HLA class II molecules. According to a recent review, however, the performance of peptide-MHC II binding algorithms remains considerably inferior to that of MHC class I binding predictors (Andreatta, M. et al. "An automated benchmarking platform for MHC class II binding prediction methods." *Bioinformatics* 34.9 (2018): 1522-1528.). We discuss this points and added the reference above in the discussion section (first paragraph, **page 6** and last sentence, **page 7**)

**3. The authors state that they use NetMHCPan and MHCFlurry, but there are many more up-to-date tools available, including MixMHCpred, NetMHCpan-4.0, and MHCFlurry-EL. This raises the likelihood that the predicted values of antigen binding in the authors manuscript are not accurate, as well.**



According to a very recent benchmark evaluation of 15 algorithms, NetMHCpan 4.0 (the version we use) and MHCFlurry resulted to be the most accurate methods for MHC class I binding prediction (*Paul S, Croft NP, Purcell AW, Tschärke DC, Sette A, Nielsen M, et al. (2020) Benchmarking predictions of MHC class I restricted T cell epitopes in a comprehensively studied model system. PLoS Comput Biol 16(5): e1007757.*). We quote this reference at **page 3**, first paragraph.

As discussed in the Methods section (**pages 8-9**), the two algorithms are complementary since MHCFlurry is an allele specific method and NetMHCpan4.0 is a pan allele method. To minimize spurious results, we have combined the two methods and only consider the predictions that are consistent across the two methods.

The referee suggests to use MixMHCpred and MHCFlurry-EL. MixMHCpred, however, supports only 58 alleles, so it has less coverage than the methods we use. MHCFlurry-EL does not appear to be widely used (it is the same MHCFlurry that we use, but with a different training set). In conclusions, we think that our predictions are accurate, according to the current standards.

**4. The authors state "Furthermore, the distributions differ between the various alleles (Fig. S1, S2 and S3), confirming the presence of heterogeneous binding pattern. To encapsulate the binding affinity distributions into a simple parameter, we counted all the peptides displaying a strong binding affinity ( $IC_{50} < 1000Mm$ ) for each of the 79 alleles."**

**a. The authors do an admirable job showing the differences in predicted binding affinity by allele, but then disregard this analysis and simply choose a 1000 nM cutoff, seemingly out of nowhere. If the authors are to include this analysis of distribution, they should use it to support their choice of criteria for a "strong binding antigen".**

Stimulated by this remark, we propose a more general method to quantify the binding affinities for each HLA molecule that takes into account the entire peptide binding distribution. This method does not rely on a cutoff chosen ad hoc and provides a measure of the likelihood that any of the virus peptides bind to a particular HLA molecule. When we rank HLA alleles according to this new measurement, we find a ranking that is similar to the one obtained with a cutoff. We report this result in **Fig. S6**, discuss the result at **page 4** of the result section and explain the method in details at **page 9-10** of the methods section.

**b. Moreover, the choice of 1000nM is far outside the norm for this type of analysis. 500 nM is a more typical naïve choice, but it has been shown that the appropriate value of this parameter is very likely dependent pathogen-dependent. Strong support for the choice of cutoff must be given for this analysis to be valid.**

As also requested by Referee 1, we repeat our analysis for a cutoff of 500nM and report the results in **Fig. S5**, discussed at **page 4** of the results section. We agree with the referee that any cutoff is arbitrary, but we found that our results appear to be consistent when different cutoffs are used and also when using cutoff-independent method, as discussed in the response to point 4.a.

**5. The authors next consider 13 alleles for assessment of T-cell recognition, and show the results from 4 of these calculation. Once again, this represents an exceptionally small number of HLA alleles (~0.1%) and it is unclear what fraction of the worldwide population is included in this calculation, or what biases, if any may skew the results.**

The 13 alleles considered are the only ones for which an assessment of T-cell recognition is currently feasible thanks to the netTepi algorithm. The alleles considered are among the most represented in human populations, so they represent much more than 0.1%. In particular, the HLA-A alleles are present in around 60% of the population while HLA-B alleles are present in around 30% of the population. We now report this information in the revised manuscript page (in the results at **page 5** and in the methods at **page 10**).

**6. When it comes to understanding the geographic distribution of alleles, the authors analyze only 17 alleles. This is virtually an order of magnitude lower than the initial 79, and means that, in the context of understanding the global ability to recognize and eliminate the COVID19 pathogen, the authors consider again around 0.1% of all possible alleles. Without considering the results of many more alleles, it is the feeling of this reviewer that this calculation is certainly inaccurate and potentially deeply misleading.**

There is probably a misunderstanding here. We try to explain this point more clearly: we did not analyze only 17 alleles. Out of the 79 alleles we studied, 17 of those were found to bind weakly to SARS-CoV-2 peptides (i.e. all the peptides had binding affinity larger than 1000nM). We thus quantified how many subjects in different populations displayed haplotypes containing one, two or three of these 17 alleles. As shown in **Fig. 2b**, 15-25% of the subjects of each population have at least one of these 17 alleles in their haplotype. This is a considerable and significant fraction (see the discussion in the last paragraph of **page 5**).

**7. Even if all of the above was accepted (it is not, by this reviewer), there is essentially no difference between the majority of populations considered by the authors, and it is unclear how the error bars were generated.**

Unfortunately the discussion of the method used to generate the error bars was missing from the methods. We thank the referee for pointing this out. Confidence intervals for frequencies are estimated assuming binomial statistics (see the revised methods for details, **page 10**). The fact that there are little differences among most of the populations (but not among the individuals within the population!) was not known *a priori* and is one of the new results of our study.

**8. Data availability: the authors make no comment on the availability of data for this work, nor the code to reproduce it. Both must be made explicitly and completely available.**

Raw data for the peptides binding affinities are now provided as supplementary data (**data S1**). The code used to generate the results is made available in the github repository: (<https://github.com/ComplexityBiosystems/hla-covid>)