Transmembrane helices are also an overlooked source of major histocompatibility complex class II epitopes and evolutionary more conserved than expected by chance

Richèl J.C. Bilderbeek[1], Maxim Baranov[1], Geert van den Bogaart[1], and Frans Bianchi[1]

[1]GBB, University of Groningen, Groningen, The Netherlands

May 14, 2020

Abstract

Transmembrane helices (TMHs) in the human proteome are an over-represented potential source of epitopes on major histocompatibility complex (MHC) class I for the majority of HLA-I haplotypes. It is unknown if the immune system is as able to detect TMHs as found in pathogens. Additionally, it is unknown if MHC-II is also likelier to detect human and pathogen TMHs than expected by chance. It is unknown if natural selection is the cause that the immune system is more attentive for TMHs This study shows that MHC-I [also has/does not have] more epitopes derived from a TMH for a pathogen proteome, when compared with a host proteome. Additionally, MHC-II binds to polypeptides derived from TMHs [less/equally/more] often than expected by chance. Lastly, using an innovative computation method, we show that natural selection on TMHs is [strong enough/still too weak] to detect. Our findings suggest that the immune system is [less/neutral/more] vigilant to TMHs than expected by chance and this has left [a clear/a weak/no] signal in the evolutionary history of the pathogen.

Keywords: antigen presentation, membrane proteins, bioinformatics, adaptive immunity, transmembrane domain, epitopes, T lymphocyte, MHC-1, MHC-I, MHC-2, MHC-II, COVID-19

1 Introduction

Immune response Our immune system fights invaders on a daily basis. These invaders can be fungi, bacteria or viruses. The innate immune response is its first general and immediate strategy, where the acquired immune response needs time to develop its specialized and more effective combat forces.

Immune response by MHC-I All nucleated cells in humans present randomly sampled polypeptides fragments to the surroundings of the cell using the Major Histocompatibility Complex (MHC) class I molecules. If a cell gets infected by a virus, also the virus' polypeptides will be presented at the cell surface. The foreign viral antigen is detected by cytotoxic T lymphocytes, which will kill the infected cell.

Immune response by MHC-II All pathogens can be detected by the foreign proteins on their (bacterial or fungal) cell walls or (viral) envelope. For bacteria and fungi, this is the main mode of their detection, as these do not infect a cell with their (foreign) DNA. T and B cells express MHC-II to detect foreign polypeptide fragments. An immune response is started when MHC-II detects a pathogen.

Classification of HLA Any human's immune system detects only a fraction of all possible polypeptide fragments. Human MHCs are also called HLAs ('Human Leukocyte Antigens') and because this research uses human hosts only, we will use these terms interchangeably. Already each MHC molecule can only bind a, possible exclusive, subset of all possible polypeptides. For example, HLA-A and HLA-B have no overlap in which peptides they bind (Lund *et al.* 2004). The HLA region of humans is highly polymorphic, making it hard to classify all of the many haplotypes. Classification of HLAs is based on algorithms that maximize the information content of each classification, such as the presence of a certain amino acid at the first (for MHC-II, Southwood *et al.* 1998) or second position (for MHC-I, Lund *et al.* 2004) of a polypeptide. However, this does not imply that all polypeptides for a classification pattern bind to the MHC. Instead, software should be used that are tailored to predict if a polypeptide binds to an HLA.

Epitope prediction It is helpful to be able to predict which polypeptides are immunogenic, in, among others, vaccine development **[RB: Reference here]**. Already for a decade, synthesized polypeptide fragments are used to experimentally determine which polypeptides are immunogenic **[RB: Reference here]**. This approach, however, is tedious and costly. It should come as no surprise that these results would be used to make in silico predictions. It may come as a surprise, however, that these predictions are actually reliable in practice Larsen *et al.* 2010; Schellens *et al.* 2008; Tang *et al.* 2011.

MHC-I epitope prediction For MHC-I, there are multiple computational tools developed to predict epitopes. According to Lundegaard *et al.* 2011, in 2011, from a set of multiple tools, NetMHCcons Karosiene *et al.* 2012 gave the best predictions. A tool developed later is `epitope-prediction` Bianchi *et al.* 2017, which uses a stabilized matrix method Kim *et al.* 2009. In this study, we will use `epitope-prediction` Bianchi *et al.* 2017.

MHC-II epitope prediction Also for MHC-II, there are multiple computational tools developed to predict epitopes, such as using a trained neural network (Nielsen *et al.* 2003) or a Gibbs sampling approach (Nielsen *et al.* 2004). According to Lundegaard *et al.* 2011, in 2011, from a set of multiple tools, NetMHCIIpan (Nielsen *et al.* 2008; Karosiene *et al.* 2013) gave rise to the most accurate predictions. Later tools are Zhang *et al.* 2013, Trolle & Nielsen 2014, Zhang *et al.* 2015 and the very recent MHCnuggetsShao *et al.* 2020. In this study, we'll be using MHCnuggets, as it is the only FOSS tool available.

TMHs Transmembrane helices are conserved structures that span a cell membrane with an alpha helix. TMHs are hydrophobic, as this is required to span the hydrophobic cellular lipid membrane. Additionally, they often have a length of 23 amino acids to be able to span the membrane. Polypeptide fragments derived from TMHs are among the most hydrophobic, together with the internals of soluble proteins, where the hydrophobicity parts guide the protein to achieve its 3D configuration. TMHs are general structures: 25 percent of the human proteome is anchored by at least one TMH. COVID-19 has 21 TMHs, making up for **[RB: ?]** percent of its entire genome.

TMH prediction There are multiple computational tools developed to detect which parts of membrane proteins are TMH In 2001, multiple tools to do so have been compared Möller *et al.* 2001, of which TMHMM Krogh *et al.* 2001 turned out to be the best. Many other tools followed, such as Phobius (Käll *et al.* 2004), ConPred II (Arai *et al.* 2004), MEMSAT3 (Jones 2007) and MetaTM (Klammer *et al.* 2009), which signals the importance of TMH predictions. Unlike TMHMM, which is still up and running, these later tools have already become obsolete. Contemporary tools that are good for use, such as the closed-source MemBrain Feng *et al.* 2020 and the FOSS tool PureseqTM Wang *et al.* 2019. In this study, we will use TMHMM for pragmatic reasons **[RB: I would love to use PureseqTM instead, but only if there is time]**.

MHC-I presents hydrophobic regions more often One might expect that the hydrophobic epitopes presented, for a same hydrophobicity, are as likely to stem from membrane proteins TMHs or from soluble proteins hydrophobic (non-TMH) regions. For MHC-I, however, it is found that the 9-mers stemming from TMHs are presented more often than expected by chance Bianchi *et al.* 2017 **[RB: but also when compensating for hydrophobicity?]**. For MHC-II, it is unknown which percentage of binders is derived from TMHs for all 13-mers present in a proteome that have the same hydrophobicity.

HLAs increase detection range An increased detection is obtained by expressing a wide variety of MHCs, which is assured by the human leukocyte antigen gene complex (HLA). However, a different HLA will result in a different variety of MHCs, that will display different polypeptide fragments. Additionally, there may have been selection the HLA to display and or detect different polypeptide fragments.

MHC-I presents TMH-derived epitopes in humans more often For MHC-I, it was found that predicted epitopes derived from human transmembrane helices (TMHs) are over-presented by all 5 HLA-A and most of 8 HLA-B super types (Bianchi *et al.* 2017). One explanation is that the presentation of TMHs may have an evolutionary advantage for the (human) host, as TMHs have a reduced variability due to the functional requirement of being able to span a lipid bilayer. Due to this, pathogens have a lower chance to develop an escape mutation, as many mutations will result in a disfunctional TMH. Note that the mechanism by which a cell presents its TMHs is yet unknown.

Does MHC-I present TMH-derived epitopes from pathogens as often? It is important that MHC-I presents both the polypeptides of the (healthy) cell, as well as possible pathogen-derived polypeptides when the cell is infected. As described above, MHC-I presents TMH-derived epitopes from the human host more often. It is unknown if MHC-I has the same dedication to present epitopes that stem from TMHs derived from proteins produced by pathogen, either would the pathogen be a virus [**RB:** $\mathcal{H}_{1,1}$] or a bacterium [**RB:** $\mathcal{H}_{1,2}$].

MHC-II is expected to present TMHs If presentation of TMHs on MHC-I would bring an evolutionary advantage in the recognition of pathogens by the immune system, it would follow that this is equally important for MHC-II, especially as the help of CD4+ T cells is needed for a long lasting CD8+ T cell response Novy *et al.* 2007. The mechanism to detect foreign TMHs by MHC-II would be unknown, similar to the discovery that MHC-I presents TMHs.

Does MHC-II present TMH-derived epitopes from pathogens as often? It is unknown if MHC-II, like MHC-I, presents TMH-derived epitopes as often, either in humans **[RB:** $\mathcal{H}_{2,1}$], bacteria **[RB:** $\mathcal{H}_{2,2}$] or viruses **[RB:** $\mathcal{H}_{2,3}$].

Selection undetectable in whole proteome The human immune system and human pathogen are in an evolutionary arms race: our immune systems is selected for the detection of pathogens, whereas pathogens are selected to avoid detection. From a pathogen's point of view, however, this struggle is of only minor importance: in seasonal influenza, for example, the selection pressure exerted by the immune system was only limited Han *et al.* 2019

In general, on would hope that evolutionary selection results in an immune system that as most attentive for loci that are essential for a virus, as these will be most conserved. In COVID-19, for example, there is preliminary evidence that the strongest selection pressure is upon residues that changes its virulence Velazquez-Salinas *et al.* 2020. These loci, however, only account for a small part of a pathogen's proteome. Additionally, these essential parts differ widely between pathogens. Because of this scarcity and variance in targets, one can imagine that the human immune system is not tailored to detect these sites, as hinted by upon by the aforementioned influenza study.

Selection may be detectable in TMHs TMHs, on the other hand, also have their function constraints, yet can occur multiple time a pathogen's proteome. One can safely assume a pathogen's proteome contains multiple TMHs. Therefore, it may be beneficial for the host if its immune system would be more attentive towards TMHs. And maybe this has already happened: MHC-I already detects hydrophobic polypeptides. This feature, however, may also be caused by selection to detect hydrophobic regions in the soluble proteins of pathogens. It is unknown, when focusing on TMHs only, if a signal of selection can be detected.

Selection needs to be additive Would a pathogen have only peptides that do not bind to any HLA, a pathogen would be undetected. In that case, one can imagine that a mutation that causes a pathogen to become detected by a host's immune system, may be selected against. Also, would a pathogen have only one epitope, one can imagine there would be selection for a mutation to lose it. If there are, however, already many epitopes present in the proteome, the selection to lose or gain an epitope is expected to have less or no effect.

Immunodominance The effective number of epitopes, however, is much lower than the actual number of possible epitopes, due to immunodominance. Immunodominance is a feature of the immune system, due to which some epitopes dominate in causing an immune response, where other (called subdominant) epitopes end up having no effect Akram & Inman 2012. The same study mentions 20 factors that influence the strength of an immune response. The most relevant factor for this study is the effect of a mutation within an epitope, such as the in vitro example described in Berkhoff *et al.* 2004, where one single mutation resulted in a 5-20% weaker responese of virus-specific CD8+ T cells **[RB: Fun experiment: do this in silico]**. Due to immunodominance, we dare assume that there can be selection for immune system avoidance, as the effective number of loci may be low and effects of a single mutation noticable.

Use of protein data in phylogenetics When using DNA sequences, one can use a skewed rate in non- versus synonymous mutation to detect the signature of selection Murrell *et al.* 2015. Using AA sequences, however, has its advantages, as its is closer to the actual phenotype selection acts upon: DNA may never be translated to RNA, or its RNA may never be transcribed Diz *et al.* 2012 **[RB: improve upon this point]**. When using a proteome in phylogenetic research, we know that the majority of proteins are selected to just maintain their function most of the time, where is the time spans there is selection, only a few AAs can actually increase the 'fitness' of the protein Anisimova & Kosiol 2009. There, when generalizing the dynamics of mostly purifying selection (to maintain a protein's function) and a short duration of positive selection, those genes that are selected cannot be detected Yang & Bielawski 2000.

Evolutionary signal of unknown strength for virus Regarding human viruses, we do not know the selection pressure exerted by the human immune system. For viral TMHs that are undetected by most human haplotypes, we expect these to be conserved, to remain avoiding detection. For viral TMHs that are detected by most human haplotypes, we expect there to be a higher mutation rate, to avoid detection. Additionally, viruses are not only selected for their evasion of an immune response, yet are selected to have a high reproductive value. Because the strength of the evolutionary signal is unknown, we use a strong data set to be able to detect it.

Evolutionary signal of unknown strength for bacteria Regarding human pathogenic bacteria, we do not know the extent to which the human immune systems selects on these, nor which loci are selected upon **[RB: so I should read the literature on that]**. Because most bacterial pathogens are generalists, that is, can infect multiple hosts, we expect the selection pressure exerted by the human immune response to be weak. For bacterial TMHs that are undetected by most human haplotypes, we expect these to be conserved, to remain avoiding detection. For bacterial TMHs that are detected by most human haplotypes, we expect there to be a higher mutation rate, to avoid detection. Also, it is unknown if bacterial TMHs are detected by MHC-II at all. Due to this, the strength of the evolutionary signal is unknown, thus we need a strong data set to be able to detect it.

2 Hypotheses

[RB: Will be moved to supplementary materials]

- $\mathcal{H}_{1,VB}$: MHC-I has the same percentage of epitopes overlapping with viral/bacterial TMHs as with human TMHs
- $\mathcal{H}_{2,HVB}$: MHC-II has the same percentage of epitopes overlapping with human/viral/bacterial TMHs as expected by chance
- $\mathcal{H}_{3,HVB}$: The mutations observed in human/viral/bacterial TMHs are caused by chance. Alternatively, these are caused by natural selection induced by binding to HLAs

3 Methods

Data sets for TMH epitopes To determine the percentages of epitopes overlapping with TMHs, we use two reference proteomes. Our viral AA sequence is represented by the proteome of the first sequenced COVID-19 strain (Wu *et al.* 2020, GenBank ID of MN908947.3, `https://www.ncbi.nlm.nih.gov/nuccore/MN908947`) For bacteria, we use the reference genome of Mycobacterium tuberculosis (`https://www.ebi.ac.uk/reference_proteomes`, UP000001584, 83332 MYCTU).

Data sets for evolutionary selection To detect an evolutionary signal for selection on binding TMHs polypeptides, we need a phylogenetic tree with many time-dated proteomes. For humans, we use the 25000 proteomes used in influenza **[RB: citation here]**. For viruses, we use the many time-dated COVID-19 proteomes. We can assume that only after the moment that COVID-19 spilled over to humans, it has mostly been the human immune system selecting against its detection, although this is only part of the selection for having a high transmission rate. For bacteria, use the proteomes of Mycobacterium **[RB: Really? Pick better set, if possible!]**.

3.1 MHC-I

To determine the percentages of MHC-I epitopes overlap with TMHs, we used the same analysis as Bianchi *et al.* 2017, except for using a viral and bacterial proteome (instead of a human proteome, as used in the original study). Bianchi and colleagues obtained a distribution of percentages of MHC-I epitopes overlap with TMHs in Homo sapiens for the different HLA haplotypes, with an average of 5.3%. We obtained a similar distribution of percentages of MHC-I epitopes that overlap with TMHs for the different HLA haplotypes, but then applied to our viral and bacterial proteome.

We compare the distributions of humans and each pathogen using a two-sample Kolmogorov-Smirnov (KS) test for a significance level $\alpha = 0.05$.

3.2 MHC-2

To determine the percentages of MHC-II epitopes overlap, we do the same analysis as MHC-I, except for MHC-II haplotypes. We picked the MHC-II alleles that occur with a frequency of at least 20% in the human population (Greenbaum *et al.* 2011, see 8 for coverage), resulting in **[RB: approx 8]** HLAs. The threshold of 20% is arbitrarily, yet chosen to result in 5 and 10 HLA variants. As Bianchi *et al.* 2017, we define the 5% of polypeptides with the lowest IC50 values as binders. We then counted the percentage of binders that have one or more amino acids that are part of a TMH.

3.2.1 Detecting evolution

Introduction To detect evolution, we determine the probability that all of the observed TMH mutations have arisen by chance, with regards to being detected/presented by the immune system, per MHC class and haplotype. We do this for human, viral and bacterial TMHs.

Gene prediction Proteogenomics is the field dedicated to, among others, predict the proteins that arise from a DNA sequence Nesvizhskii 2014. For an ab initio gene prediction of a viral genome, we searched the literature for the most recent software tool, which is Vgas (Zhang *et al.* 2019). In this article, Vgas is compared to tools, such as GeneMarkS (Besemer *et al.* 2001), Prodigal (Hyatt *et al.* 2012), GLIMMER (Delcher *et al.* 1999), RAST Aziz *et al.* 2008 and prokka (Seemann 2014). We picked the FOSS tool that scored best in the (closed source) Vgas benchmark (Zhang *et al.* 2019), which is Prodigal.

Correction against many tests Because we use 13 MHC-I and 7 MHC-II haplotypes for a significance level of 5%, by chance we expect (5% of $13 + 7$) one times to find a (MHC-I or MHC-II) haplotype that is suggested to have selection working upon it. Applying a binomial distribution ($p = 0.05$, $n = 13 + 7$, $\alpha = 0.05$), we will claim we there is some form of selection when we find 3 or more examples of having a significant result. However, because there are 2 hypotheses tested per haplotype, we will claim there is directional selection, when, for a same hypothesis, we find 2 examples of having a significant result.

System depicted as a Markov chain To calculate the probability all of our observations happened by chance, we view our experiment as a two-state

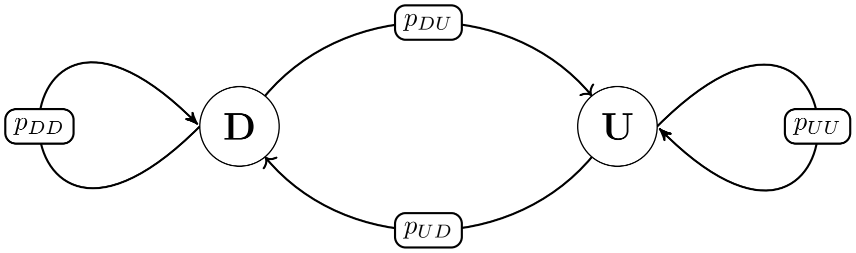


Figure 1: State transitions in TMHs caused by mutations, with respect to the detection by the immune system. Any TMH is either in a detected, D, or undetected, U, state. Mutations cause state transitions with probability p. For example, p_{DU} denotes the probability that a detected TMH mutates into an undetected one.

Markov chain, as depicted in figure 1. The two nodes depict the two possible detection state a polypeptide is in, which is either undetected U, or detected D by the immune system. The D can be either detection by an MHC-I or MHC-II haplotype. The edges depict the probability of a state transition by mutation, for example, p_{DU} denotes the probability that a detected polypeptide mutates into an undetected one. We focus on polypeptides that are TMH and we assume a mutation that renders the TMH non-functional as lethal.

Statistics When we have calculated the probability of each transition (as described below), we can simply count the number of observed transitions. Because the probability of transitions differ per polypeptide, the total system has the properties of a Poisson binomial distribution, from which we can calculate the probability of the observing our number of transitions.

Hypotheses testing Of the complete system, we test for each of the 2 state transitions (that is, we do not test for mutations that result in the same state).

For each transition rate, we know the expected and observed number of transitions. If the observed number of transitions is below the expected value, we test for finding that (low) number of transitions or less. If the observed number of transitions is above the expected value, we test for finding that (high) number of transitions or more. We use a significance level α of 5%, because we have no prior evidence.

As we are testing 2 hypotheses, we correct against type I errors (false positives), by applying the Holm-Bonferroni correction, resulting in **[RB: here is just a normal Bonferroni correction, for now]** $\alpha_c = \frac{\alpha}{2}$. If the probability of observing that few/many transition is below α_c, we will state that this unlikely situation may be caused by natural selection.

Calculating one polypeptide's probability We calculate the probability of a transition per polypeptide. For any polypeptide, we can generate all sequences that differ in only one amino acids. Of these sequences, we only keep the ones that result in a TMH. Of each of the remaining TMH sequences, we predict which mutants are detected by the immune system. Because not all AA mutations are equally likely, we correct for the AA transition rates using the FLU transition matrix (Dang *et al.* 2010), which is a transition matrix derived from observed influenza AA mutations.

Assumption that epitopes are immunogenic We assume that if a polypeptide binds to an MHC, it is always presented and always leads to an immune response. The more often this assumption is violated, the more our analysis will be weakened. Experiments in mice with MHC-I epitopes, however, indicate that (if the dose administered was high enough) there are either none or rare exceptions Sette *et al.* 1994.

Measuring strength To assess the strength of our analysis, we have applied it to four simulated data sets, which are the product of using either MHC-I or MHC-II, as well as using two different selection scenarios.

For both MHC-I and MHC-II, we use the haplotypes that have the lowest phenotypic frequency, which is **[RB: X]** for MHC-I and DPB1*0501 (21.7%) for MHC-II.

The two selection scenarios are either no selection or a two-fold selection on what we would expect. The scenario without selection, the 'null' data set, acts as a control, where we let the simulate observations occur by chance. The 'rigged' data set is a data set in which we let evolution bias the observations two-fold, that is, we let mutations that are beneficial occur twice as often as expected by chance.

All data sets consists out of 10^4 observations. Each observation starts with a randomly generated source polypeptide fragment, as well a mutant. For the 'null' data set, the mutant is randomly select. For the 'rigged' data set, beneficial mutations are twice as likely to be selected. We define a beneficial mutation as a transition from being detected to being undetected.

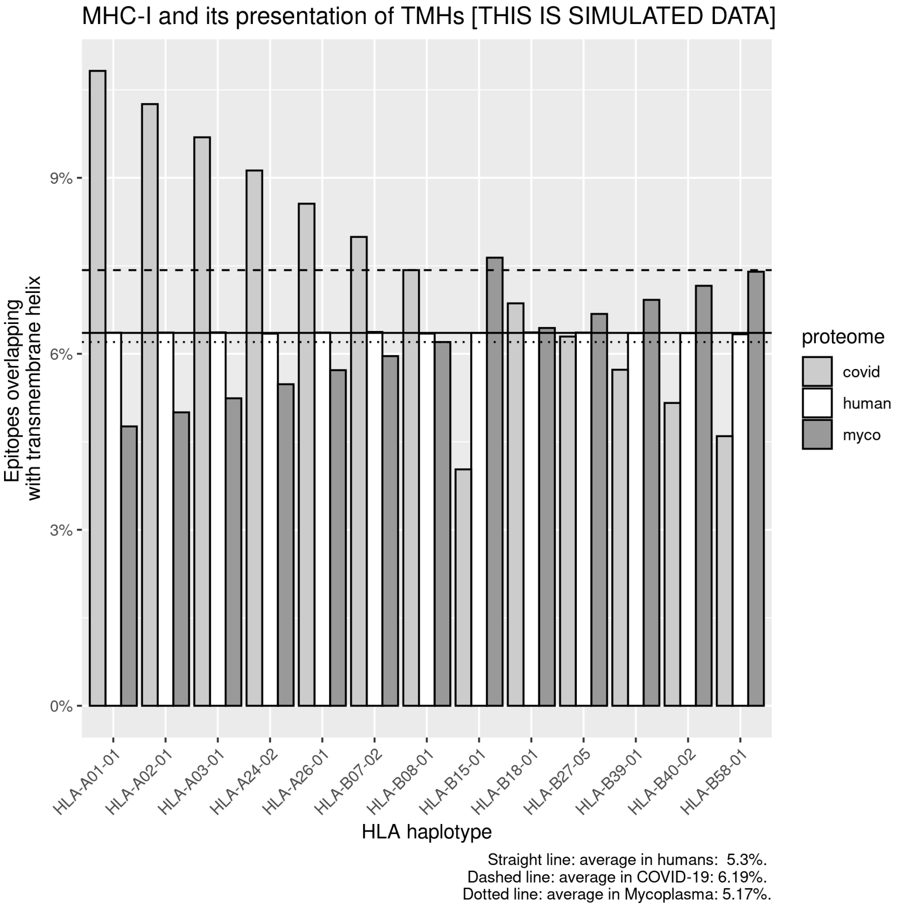
4 Results

4.1 MHC-I

Figure 2 shows the percentages of MHC-I epitopes overlapping with TMHs for our human, viral and bacterial proteome.

The KS test to determine if MHC-I has the same percentage of epitopes overlapping with viral TMHs, compared to human TMHs **[RB:** $\mathcal{H}_{1,V}$**]**, resulted in a p-value of **[RB: unknown]**, which makes us [reject/accept] the hypothesis that these percentages are sampled from the same distribution.

The KS test to determine if MHC-I has the same percentage of epitopes overlapping with bacterial TMHs, compared to human TMHs [RB: $\mathcal{H}_{1,B}$], resulted in a p-value of [RB: unknown], which makes us [reject/accept] the hypothesis that these percentages are sampled from the same distribution.

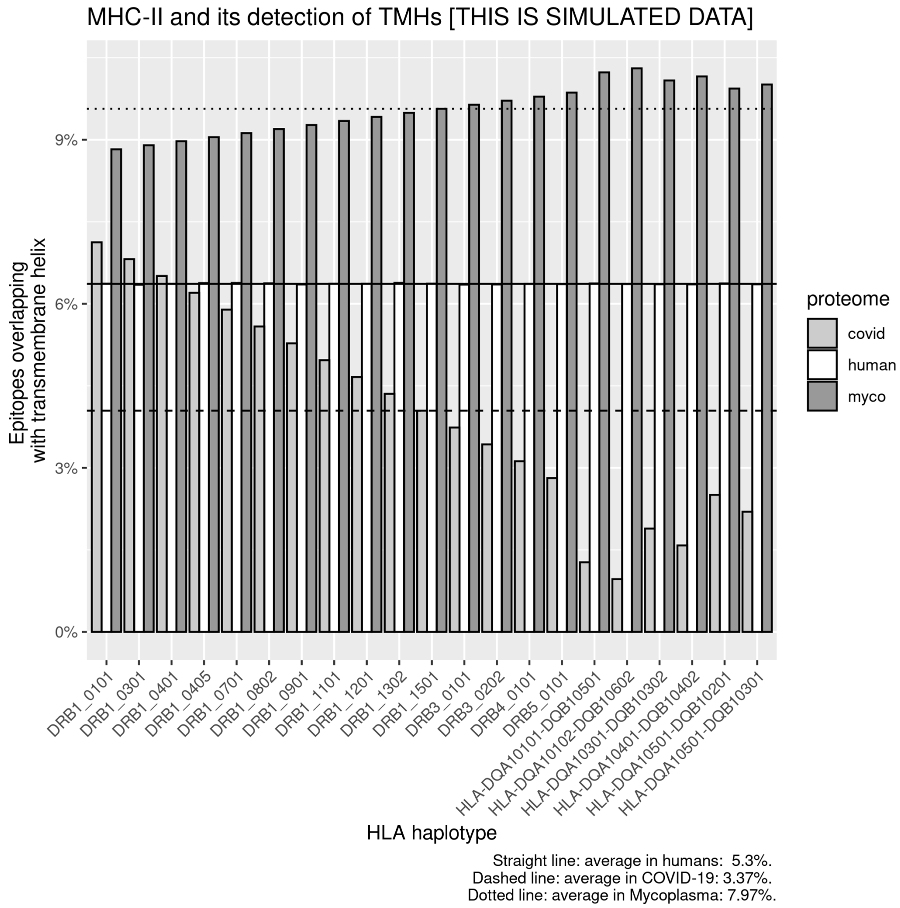


Straight line: average in humans: 5.3%.
Dashed line: average in COVID-19: 6.19%.
Dotted line: average in Mycoplasma: 5.17%.

Figure 2: Percentage of MHC-I epitopes overlapping with TMHs for a human, viral and bacterial proteome. [RB: the data underlying this figure has been simulated]

4.2 MHC-II

Figure 3 shows the percentages of MHC-II epitopes overlapping with TMHs for our human, viral and bacterial proteome.



Straight line: average in humans: 5.3%.
Dashed line: average in COVID-19: 3.37%.
Dotted line: average in Mycoplasma: 7.97%.

Figure 3: % epitopes overlapping with transmembrane helix for a human, COVID-19 and Mycobacterium proteome. **[RB: the data underlying this figure has been simulated]**

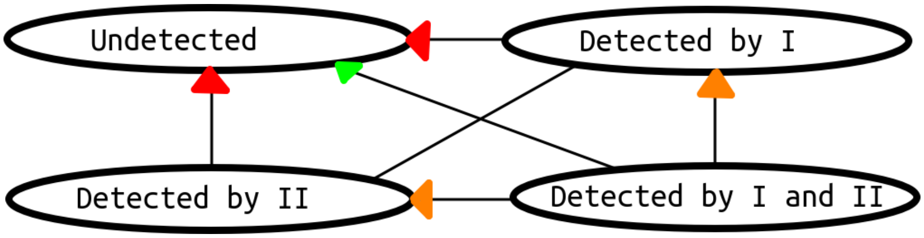


Figure 4: Transitions in TMHs that occur more often then expected by chance, for a confidence interval of 5 percent. Table 1 shows the transition counts. **[RB: stub figure]**

	Undetected	Detected by MHC-I	Detected by MHC-II	Detected by both
Undetected	700*/500	130/200	120/200	50/100
Detected by MHC-I	140*/100	50/80	5/10	5/10
Detected by MHC-II	280*/200	100/160	10/20	10/20
Detected by both	10*/5	15*/10	15*/10	10/25

Table 1: Transitions counts, where the row indicates the source state, and the column indicates the target state. First number per cell is the observed number of this state transition, where the second number is the expected number of this state transition as predicted by chance. An asterisk behind the observed count indicates that this count is unlikely to be caused by chance only. Figure 4 shows which transition counts are unlikely to be caused by chance only.

4.3 Evolutionary conservation

5 Conclusion

We found that the percentages of epitopes overlapping with TMHs for a human and viral proteome are [similar/different]. In other words, the epitopes that MHC-I presents are [as/not as] likely to be derived from TMH within either a human host and its viral pathogen.

We found that the percentages of epitopes overlapping with TMHs for a human and bacterial proteome are [similar/different]. In other words, the epitopes that MHC-I presents are [as/not as] likely to be derived from TMH within either a human host and its bacterial pathogen.

6 Discussion

We concluded that the epitopes that MHC-I presents are [as/not as] likely to be derived from TMH within either a human host and its viral pathogen. Because the full COVID-19 has 21 TMHs, the percentages of MHC-I epitopes being part of a TMH are likelier to be affected by stochasticity. We chose to use COVID-19 regardless, as the thousands of its time-dated genomic sequences are ideal for determining the evolutionary conservation of MHC-I detecting TMHs.

We concluded that the epitopes that MHC-I presents are [as/not as] likely to be derived from TMH within either a human host and its bacterial pathogen. Because a bacterium does not infect a cell, thus its polypeptides will not be presented by MHC-I, this result is [unexpected/expected]

We aimed our evolutionary experiment at TMHs, because these can be predicted well from a protein structure, are common structures and are present in all pathogens. We could have done the same experiment on beta-turn, as also these can be predicted well Petersen *et al.* 2010, are common structures and are present in all pathogens.

[RB: Note that most bacteria are opportunistic pathogens. Note that most bacteria are generalists. Note that most bacteria have different cell membranes (and walls), that may have different functional constraints than a human cell membrane]

7 Acknowledgments

We thank the Center for Information Technology of the University of Groningen for its support and for providing access to the Peregrine high performance computing cluster. Additionally, we would like to thank Sci-Hub Himmelstein *et al.* 2018 for allowing us to read paywalled articles while working from home.

8 Data Accessibility

All code is archived at `http://github.com/richelbilderbeek/someplace`, with DOI `https://doi.org/12.3456/zenodo.1234567`.

9 Authors' contributions

RJCB and FB conceived the idea for this research. RJCB wrote the code. RJCB and FB wrote the article.

References

Akram, A. & Inman, R.D. (2012) Immunodominance: a pivotal principle in host response to viral infections. *Clinical immunology*, **143**, 99–115.

Anisimova, M. & Kosiol, C. (2009) Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Molecular biology and evolution*, **26**, 255–271.

Arai, M., Mitsuke, H., Ikeda, M., Xia, J.X., Kikuchi, T., Satake, M. & Shimizu, T. (2004) Conpred ii: a consensus prediction method for obtaining transmembrane topology models with high reliability. *Nucleic acids research*, **32**, W390–W393.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M. *et al.* (2008) The rast server: rapid annotations using subsystems technology. *BMC genomics*, **9**, 75.

Bar-On, Y.M., Flamholz, A.I., Phillips, R. & Milo, R. (2020) Sars-cov-2 (covid-19) by the numbers. *arXiv preprint arXiv:200312886.*

Berkhoff, E., Boon, A., Nieuwkoop, N., Fouchier, R., Sintnicolaas, K., Osterhaus, A. & Rimmelzwaan, G. (2004) A mutation in the hla-b* 2705-restricted np383-391 epitope affects the human influenza a virus-specific cytotoxic t-lymphocyte response in vitro. *Journal of Virology*, **78**, 5216–5222.

Besemer, J., Lomsadze, A. & Borodovsky, M. (2001) Genemarks: a self-training method for prediction of gene starts in microbial genomes. implications for finding sequence motifs in regulatory regions. *Nucleic acids research*, **29**, 2607–2618.

Bianchi, F., Textor, J. & van den Bogaart, G. (2017) transmembrane helices are an overlooked source of major histocompatibility complex class i epitopes. *Frontiers in immunology*, **8**, 1118.

Dang, C.C., Le, Q.S., Gascuel, O. & Le, V.S. (2010) Flu, an amino acid substitution model for influenza proteins. *BMC evolutionary biology*, **10**, 99.

Delcher, A.L., Harmon, D., Kasif, S., White, O. & Salzberg, S.L. (1999) Improved microbial gene identification with glimmer. *Nucleic acids research*, **27**, 4636–4641.

Diz, A.P., MARTÍNEZ-FERNÁNDEZ, M. & ROLÁN-ALVAREZ, E. (2012) Proteomics in evolutionary ecology: linking the genotype with the phenotype. *Molecular ecology*, **21**, 1060–1080.

Feng, S.H., Zhang, W.X., Yang, J., Yang, Y. & Shen, H.B. (2020) Topology prediction improvement of α-helical transmembrane proteins through helix-tail modeling and multiscale deep learning fusion. *Journal of Molecular Biology*, **432**, 1279–1296.

Greenbaum, J., Sidney, J., Chung, J., Brander, C., Peters, B. & Sette, A. (2011) Functional classification of class ii human leukocyte antigen (hla) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across supertypes. *Immunogenetics*, **63**, 325–335.

Han, A.X., Maurer-Stroh, S. & Russell, C.A. (2019) Individual immune selection pressure has limited impact on seasonal influenza virus evolution. *Nature ecology & evolution*, **3**, 302–311.

Himmelstein, D.S., Romero, A.R., Levernier, J.G., Munro, T.A., McLaughlin, S.R., Tzovaras, B.G. & Greene, C.S. (2018) Sci-hub provides access to nearly all scholarly literature. *ELife*, **7**, e32822.

Hyatt, D., LoCascio, P.F., Hauser, L.J. & Uberbacher, E.C. (2012) Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics*, **28**, 2223–2230.

Jones, D.T. (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, **23**, 538–544.

Käll, L., Krogh, A. & Sonnhammer, E.L. (2004) A combined transmembrane topology and signal peptide prediction method. *Journal of molecular biology*, **338**, 1027–1036.

Karosiene, E., Lundegaard, C., Lund, O. & Nielsen, M. (2012) Netmhccons: a consensus method for the major histocompatibility complex class i predictions. *Immunogenetics*, **64**, 177–186.

Karosiene, E., Rasmussen, M., Blicher, T., Lund, O., Buus, S. & Nielsen, M. (2013) Netmhciipan-3. 0, a common pan-specific mhc class ii prediction method including all three human mhc class ii isotypes, hla-dr, hla-dp and hla-dq. *Immunogenetics*, **65**, 711–724.

Kim, Y., Sidney, J., Pinilla, C., Sette, A. & Peters, B. (2009) Derivation of an amino acid similarity matrix for peptide: Mhc binding and its application as a bayesian prior. *BMC bioinformatics*, **10**, 394.

Klammer, M., Messina, D.N., Schmitt, T. & Sonnhammer, E.L. (2009) Metatm-a consensus method for transmembrane protein topology prediction. *BMC bioinformatics*, **10**, 314.

Krogh, A., Larsson, B., Von Heijne, G. & Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of molecular biology*, **305**, 567–580.

Larsen, M.V., Lelic, A., Parsons, R., Nielsen, M., Hoof, I., Lamberth, K., Loeb, M.B., Buus, S., Bramson, J. & Lund, O. (2010) Identification of cd8+ t cell epitopes in the west nile virus polyprotein by reverse-immunology using netctl. *PloS one*, **5**.

Lund, O., Nielsen, M., Kesmir, C., Petersen, A.G., Lundegaard, C., Worning, P., Sylvester-Hvid, C., Lamberth, K., Røder, G., Justesen, S. *et al.* (2004) Definition of supertypes for hla molecules using clustering of specificity matrices. *Immunogenetics*, **55**, 797–810.

Lundegaard, C., Lund, O. & Nielsen, M. (2011) Prediction of epitopes using neural network based methods. *Journal of immunological methods*, **374**, 26–34.

Möller, S., Croning, M.D. & Apweiler, R. (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **17**, 646–653.

Murrell, B., Weaver, S., Smith, M.D., Wertheim, J.O., Murrell, S., Aylward, A., Eren, K., Pollner, T., Martin, D.P., Smith, D.M. *et al.* (2015) Gene-wide identification of episodic selection. *Molecular biology and evolution*, **32**, 1365–1371.

Nesvizhskii, A.I. (2014) Proteogenomics: concepts, applications and computational strategies. *Nature methods*, **11**, 1114.

Nielsen, M., Lundegaard, C., Blicher, T., Peters, B., Sette, A., Justesen, S., Buus, S. & Lund, O. (2008) Quantitative predictions of peptide binding to any hla-dr molecule of known sequence: Netmhciipan. *PLoS computational biology*, **4**.

Nielsen, M., Lundegaard, C., Worning, P., Hvid, C.S., Lamberth, K., Buus, S., Brunak, S. & Lund, O. (2004) Improved prediction of mhc class i and class ii epitopes using a novel gibbs sampling approach. *Bioinformatics*, **20**, 1388–1397.

Nielsen, M., Lundegaard, C., Worning, P., Lauemøller, S.L., Lamberth, K., Buus, S., Brunak, S. & Lund, O. (2003) Reliable prediction of t-cell epitopes using neural networks with novel sequence representations. *Protein Science*, **12**, 1007–1017.

Novy, P., Quigley, M., Huang, X. & Yang, Y. (2007) Cd4 t cells are required for cd8 t cell survival during both primary and memory recall responses. *The Journal of Immunology*, **179**, 8243–8251.

Petersen, B., Lundegaard, C. & Petersen, T.N. (2010) Netturnp–neural network prediction of beta-turns by use of evolutionary information and predicted protein sequence features. *PLoS One*, **5**.

Schellens, I.M., Kesmir, C., Miedema, F., van Baarle, D. & Borghans, J.A. (2008) An unanticipated lack of consensus cytotoxic t lymphocyte epitopes in hiv-1 databases: the contribution of prediction programs. *Aids*, **22**, 33–37.

Seemann, T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.

Sette, A., Vitiello, A., Reherman, B., Fowler, P., Nayersina, R., Kast, W.M., Melief, C., Oseroff, C., Yuan, L., Ruppert, J. *et al.* (1994) The relationship between class i binding affinity and immunogenicity of potential cytotoxic t cell epitopes. *The Journal of Immunology*, **153**, 5586–5592.

Shao, X.M., Bhattacharya, R., Huang, J., Sivakumar, I.A., Tokheim, C., Zheng, L., Hirsch, D., Kaminow, B., Omdahl, A., Bonsack, M. *et al.* (2020) High-throughput prediction of mhc class i and ii neoantigens with mhcnuggets. *Cancer Immunology Research*, **8**, 396–408.

Southwood, S., Sidney, J., Kondo, A., del Guercio, M.F., Appella, E., Hoffman, S., Kubo, R.T., Chesnut, R.W., Grey, H.M. & Sette, A. (1998) Several common hla-dr types share largely overlapping peptide binding repertoires. *The Journal of Immunology*, **160**, 3363–3373.

Tang, S.T., van Meijgaarden, K.E., Caccamo, N., Guggino, G., Klein, M.R., van Weeren, P., Kazi, F., Stryhn, A., Zaigler, A., Sahin, U. *et al.* (2011) Genome-based in silico identification of new mycobacterium tuberculosis antigens activating polyfunctional cd8+ t cells in human tuberculosis. *The Journal of Immunology*, **186**, 1068–1080.

Trolle, T. & Nielsen, M. (2014) Nettepi: an integrated method for the prediction of t cell epitopes. *Immunogenetics*, **66**, 449–456.

Velazquez-Salinas, L., Zarate, S., Eberl, S., Gladue, D.P., Novella, I. & Borca, M.V. (2020) Positive selection of orf3a and orf8 genes drives the evolution of sars-cov-2 during the 2020 covid-19 pandemic. *bioRxiv*.

Wang, Q., Ni, C., Li, Z., Li, X., Han, R., Zhao, F., Xu, J., Gao, X. & Wang, S. (2019) Efficient and accurate prediction of transmembrane topology from amino acid sequence only. *bioRxiv*, p. 627307.

Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y. *et al.* (2020) A new coronavirus associated with human respiratory disease in china. *Nature*, **579**, 265–269.

Yang, Z. & Bielawski, J.P. (2000) Statistical methods for detecting molecular adaptation. *Trends in ecology & evolution*, **15**, 496–503.

Zhang, K.Y., Gao, Y.Z., Du, M.Z., Liu, S., Dong, C. & Guo, F.B. (2019) Vgas: A viral genome annotation system. *Frontiers in microbiology*, **10**, 184.

Zhang, W., Liu, J., Xiong, Y., Ke, M. & Zhang, K. (2013) Predicting immunogenic t-cell epitopes by combining various sequence-derived features. *2013 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 4–9. IEEE.

Table 2: Percentage of MHC-I epitopes overlapping with transmembrane helix. **[RB: This is simulated data]**

Table 3: Kolmogorov-Smirnov test results comparing human and COVID-19 for MHC-I **[RB: Done on the simulated data]**

Zhang, W., Niu, Y., Zou, H., Luo, L., Liu, Q. & Wu, W. (2015) Accurate prediction of immunogenic t-cell epitopes from epitope sequences using the genetic algorithm-based ensemble learning. *PloS one*, **10**.

A Supplementary materials

A.1 MHC-I

A.2 MHC-II

A.3 COVID-19 genome and proteome

Tip: see Bar-On *et al.* 2020 for COVID-19 in numbers.

[RB: This is just a reminder, instead of new research. This subsection be deleted in the future.]

parameter	value_myco
statistic	0.5384615
p_value	0.04427245
alternative	two-sided
method	Two-sample Kolmogorov-Smirnov test
data_name	f_human and f_myco
alpha	0.05
n	13
verdict	Reject that the two samples are taken from a same distribution

Table 4: Kolmogorov-Smirnov test results comparing human and Mycobacterium for MHC-I **[RB: Done on the simulated data]**

Table 5: Percentage of MHC-II epitopes overlapping with transmembrane helix. [RB: This is simulated data]

parameter	value_covid
statistic	0.8571429
p_value	4.265613e-08
alternative	two-sided
method	Two-sample Kolmogorov-Smirnov test
data_name	f_human and f_covid
alpha	0.05
n	21
verdict	Reject that the two samples are taken from a same distribution

Table 6: Kolmogorov-Smirnov test results comparing human and COVID-19 for MHC-II [RB: Done on the simulated data]

parameter	value_myco
statistic	1.0000000
p_value	3.715694e-12
alternative	two-sided
method	Two-sample Kolmogorov-Smirnov test
data_name	f_human and f_myco
alpha	0.05
n	21
verdict	Reject that the two samples are taken from a same distribution

Table 7: Kolmogorov-Smirnov test results comparing human and Mycobacterium for MHC-II [RB: Done on the simulated data]

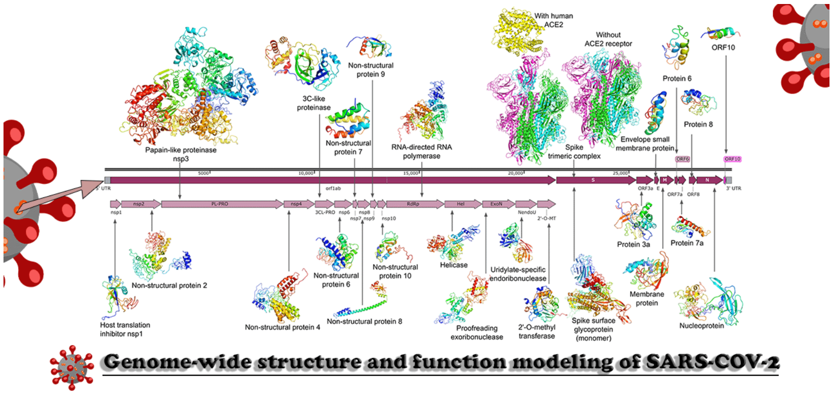


Figure 5: Overview of COVID-19 genome and proteome. From `https://zhanglab.ccmb.med.umich.edu/COVID-19`

A.4 Relation between hydrophobicity, TMH and MHC binding

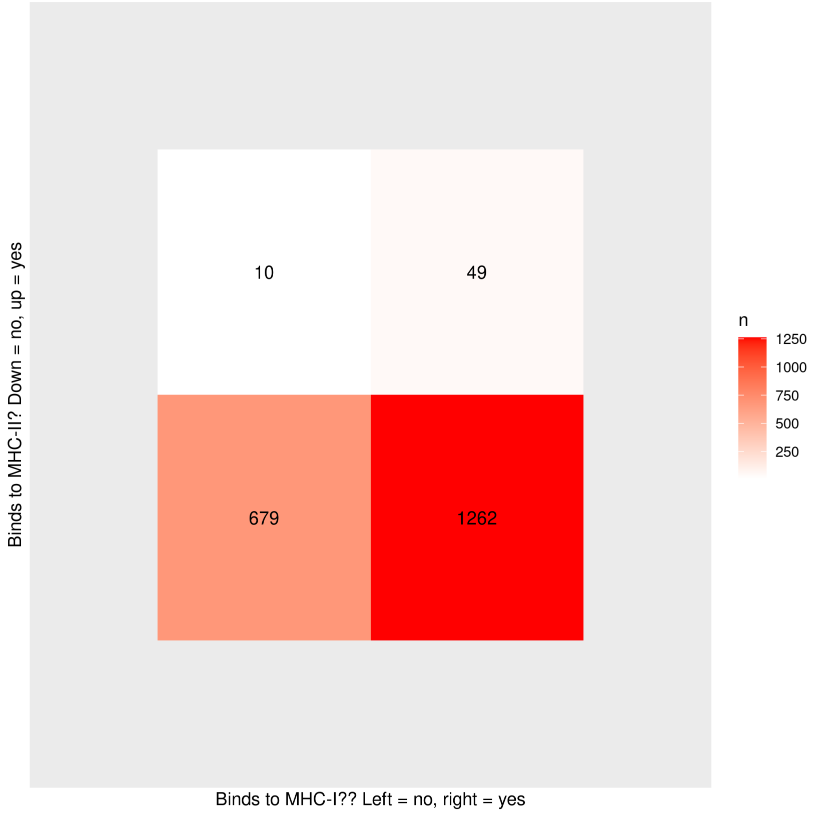


Figure 6: **[RB: Used randomly simulated polypeptides, results are real]**

A.5 MHC-II haplotype occurrences

[RB: This is just a reminder, instead of new research. This subsection be deleted in the future.]

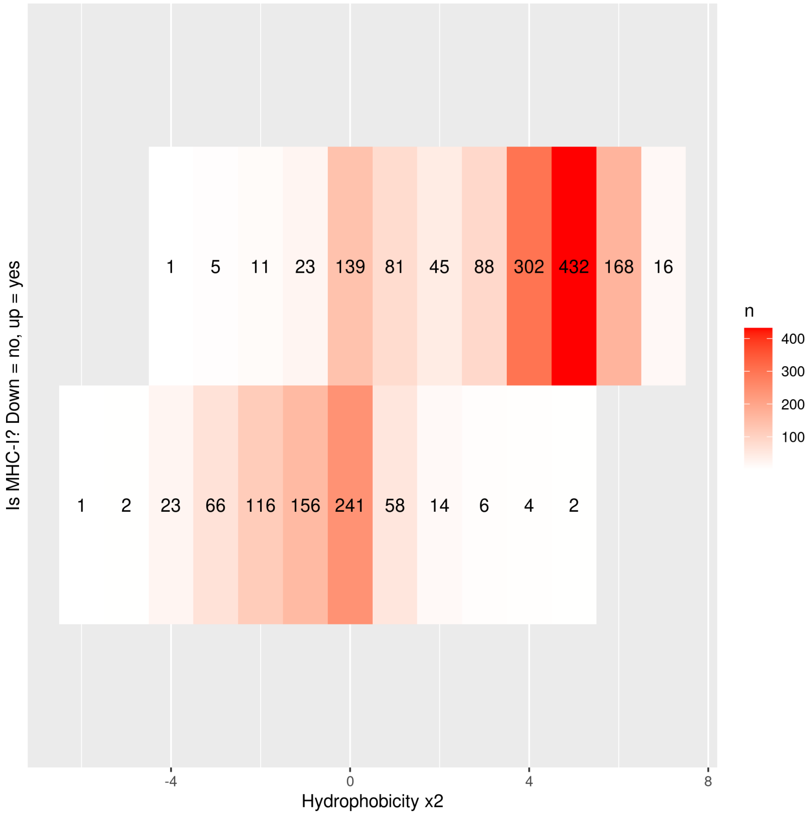


Figure 7: **[RB: Used randomly simulated polypeptides, results are real]**

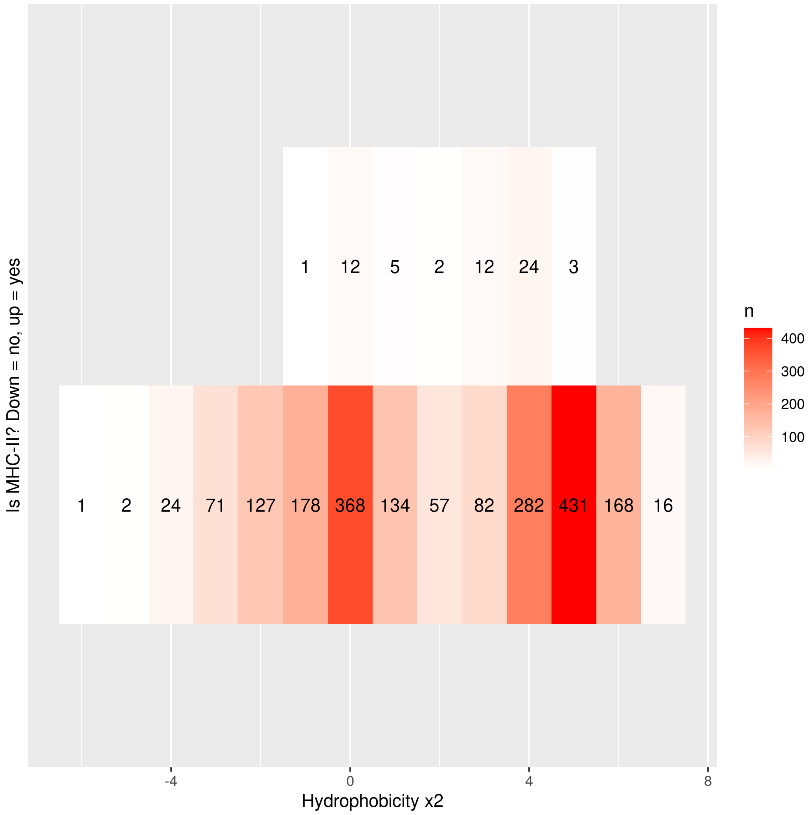


Figure 8: **[RB: Used randomly simulated polypeptides, results are real]**

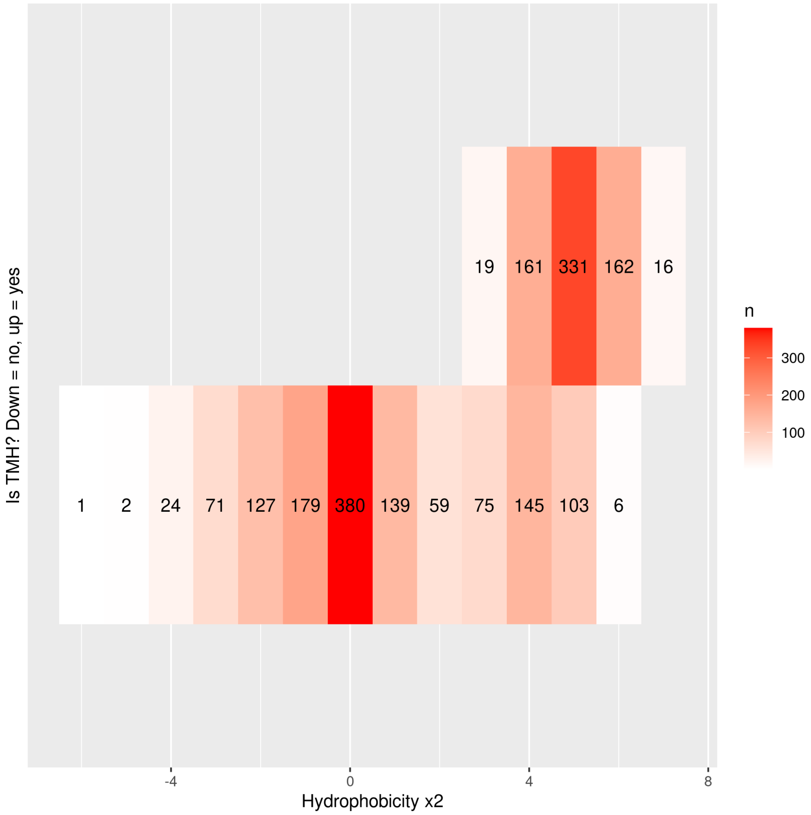


Figure 9: **[RB: Used randomly simulated polypeptides, results are real]**

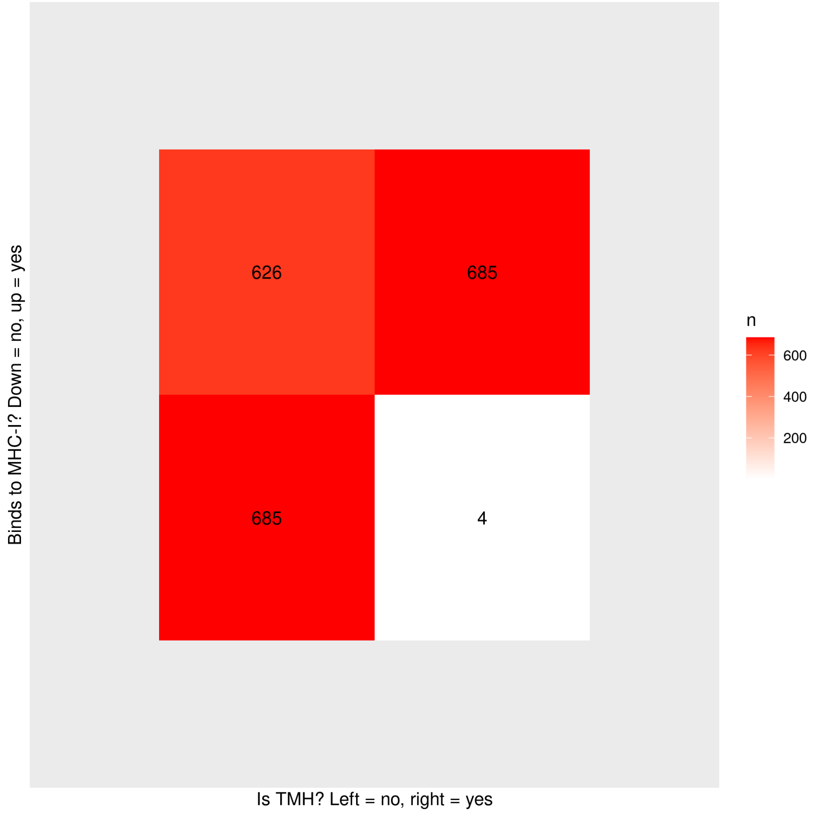


Figure 10: **[RB: Used randomly simulated polypeptides, results are real]**

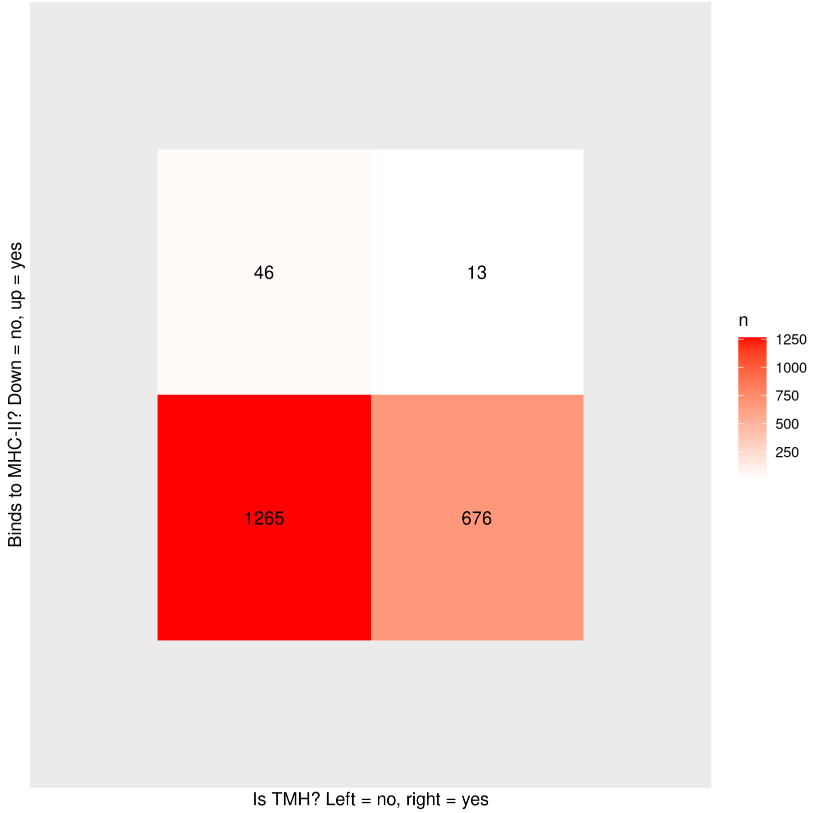


Figure 11: **[RB: Used randomly simulated polypeptides, results are real]**

Locus	Allele	Percent of haplotypes	Phenotype frequency
DRB1	DRB1*0101	2.8	5.4
	DRB1*0301	7.1	13.7
	DRB1*0401	2.3	4.6
	DRB1*0405	3.1	6.2
	DRB1*0701	7	13.5
	DRB1*0802	2.5	4.9
	DRB1*0901	3.1	6.2
	DRB1*1101	6.1	11.8
	DRB1*1201	2	3.9
	DRB1*1302	3.9	7.7
	DRB1*1501	6.3	12.2
	Total	46.2	71.1
DRB3/4/5	DRB3*0101	14	26.1
	DRB3*0202	18.9	34.3
	DRB4*0101	23.7	41.8
	DRB5*0101	8.3	16
	Total	77.3	87.7
DQA1/DQB1	DQA1*0501/DQB1*0201	5.8	11.3
	DQA1*0501/DQB1*0301	19.5	35.1
	DQA1*0301/DQB1*0302	10	19
	DQA1*0401/DQB1*0402	6.6	12.8
	DQA1*0101/DQB1*0501	7.6	14.6
	DQA1*0102/DQB1*0602	7.6	14.6
	Total	57.1	81.6
DPB1	DPB1*0101	8.4	16
	DPB1*0201	9.2	17.5
	DPB1*0401	20.1	36.2
	DPB1*0402	23.6	41.6
	DPB1*0501	11.5	21.7
	DPB1*1401	3.8	7.4
	Total	76.5	94.5

Table 8: Percentage of MHC-II haplotypes, from Greenbaum *et al.* 2011 **[RB: This is just a reminder, instead of new research. This table be deleted in the future.]**

A.6 Kolmogorov-Smirnov

[RB: This is just a reminder, instead of new research. This subsection be deleted in the future.]

The Kolmogorov-Smirnov (KS) test determines if two samples are derived from the same distribution, without making assumptions regarding the shape of that distribution.

We will reject the null hypothesis that MHC-I has the same percentage of epitopes overlapping with TMHs in Homo sapiens compared to each pathogen when the KS statistic $D_{n,m}$ follows the relationship as shown in equation 1, for a significance level $\alpha = 0.05$ and $n = m$ equals the number of HLA haplotypes.

$$D_{n,m} > \frac{1}{\sqrt{n}} \cdot \sqrt{-\ln(\frac{\alpha}{2}) \cdot \frac{1 + \frac{n}{m}}{2}} \quad (1)$$