

1 Transmembrane helices are an overlooked and  
2 evolutionarily conserved source of major  
3 histocompatibility complex class I and II epitopes

4 Richèl J.C. Bilderbeek<sup>1</sup>, Maksim V. Baranov<sup>1</sup>, Geert van den  
5 Bogaart<sup>1</sup>, and Frans Bianchi<sup>1</sup>

6 <sup>1</sup>GBB, University of Groningen, Groningen, The Netherlands

7 November 19, 2021

## Abstract

Cytolytic T cell responses are predicted to be biased towards membrane proteins. The peptide-binding grooves of the alleles of most haplotypes of histocompatibility complex class I (MHC-I) are relatively hydrophobic, therefore peptide fragments derived from human transmembrane helices (TMHs) are predicted to be presented more often as would be expected based on their abundance in the proteome. However, the physiological reason of why membrane proteins might be over-presented is unclear. In this study, we show that the predicted over-presentation of TMH-derived peptides is general, as it is predicted for bacteria and viruses and for both MHC-I and MHC-II, and confirmed by re-analysis of epitope databases. Moreover, we show that TMHs are evolutionarily more conserved, because single nucleotide polymorphisms (SNPs) are present relatively less frequently in TMH-coding chromosomal regions compared to regions coding for extracellular and cytoplasmic protein regions. Thus, our findings suggest that both cytolytic and helper T cells are more tuned to respond to membrane proteins, because these are evolutionary more conserved. We speculate that TMHs are less prone to mutations that enable pathogens to evade T cell responses.

**Keywords:** antigen presentation, membrane proteins, bioinformatics, adaptive immunity, transmembrane domain, transmembrane helix, epitopes, T lymphocyte, MHC-I, MHC-II, evolutionary conservation

## Abbreviations

Abbreviation	Full
ER	Endoplasmatic reticulum
ERAD	ER-associated degradation
HLA	Human leukocyte antigen
IEDB	Immune Epitope Database
LB	lipid body
MAP	Membrane-associated protein
MHC	Major histocompatibility complex
MVB	Multivesicular body
PLC	Peptide-loading complex
SNP	Single nucleotide polymorphism
TMH	Transmembrane helix
TMP	Transmembrane protein

# 1 Introduction

Our immune system fights diseases and infections from pathogens, such as fungi, bacteria or viruses. An important part of the acquired immune response, that develops specialized and more specific recognition of pathogens than the innate immune response, are T cells which recognize peptides, called epitopes, derived from antigenic proteins presented on Major Histocompatibility Complexes (MHC) class I and II on the cell surface.

The MHC proteins are heterodimeric complexes encoded by the HLA (Human Leukocyte Antigens) genes. In humans, the peptide binding groove of MHC-I is made by only the alpha subunit. There are three classical haplotypes of MHC-I, hallmarked by a highly polymorphic alpha chain called HLA-A, HLA-B and HLA-C, that all present epitopes to cytolytic T cells. For MHC-II, both the alpha and the beta chains contribute to the peptide binding groove. There are three classical haplotypes of MHC-II as well, called HLA-DR, HLA-DQ and HLA-DP, that all present epitopes to helper T cells. Each MHC complex can present a subset of all possible peptides. For example, HLA-A and HLA-B have no overlap in which epitopes they bind [1]. Moreover, the HLA genes of humans are highly polymorphic, with hundreds to thousands of different alleles, and each different allele presents a different subset of peptides [2].

Humans mostly express two alleles per MHC haplotype, one from the parental and one from the maternal chromosome, and therefore an individual's immune system detects only a fraction of all possible peptide fragments. However, at the population level, the coverage of pathogenic peptides that are detected is very high, because of the highly polymorphic MHC genes. It is therefore believed that MHC polymorphism improves immunity at the population level, as mutations in a protein that disrupt a particular MHC presentation at the individual level, so-called escape mutations, will not affect MHC presentation for all alleles

58 present in the population [3].

59 Many studies are aimed at identifying the repertoire of epitopes that are  
60 presented in any of the different alleles to determine which epitopes will result  
61 in an immune response, as this will for instance aid the design of vaccines.  
62 These studies have led to the development of prediction algorithms that allow  
63 for very reliable *in silico* predictions of the peptide binding affinities [4, 5, 6]. For  
64 example, S. Tang et al. [6] found that, of the 432 peptides that were predicted  
65 to bind to an MHC allele, 86% were experimentally confirmed to do so.

66 Using these prediction algorithms, we recently showed that peptides derived  
67 from transmembrane helices (TMHs) are likely to be more frequently presented  
68 by MHC-I than expected based on their abundance [7], which is in line with  
69 a previous study by Istrail et al [8], demonstrating that N-terminal signal se-  
70 quences are likely to be presented within major histocompatibility complexes,  
71 due their hydrophobic nature. Moreover, we showed that some well-known im-  
72 munodominant peptides stem from TMHs. This over-presentation is attributed  
73 to the fact that the peptide-binding groove of most MHC-I alleles is relatively  
74 hydrophobic, and therefore hydrophobic TMH-derived peptides have a higher  
75 affinity to bind than their soluble hydrophobic counterparts.

76 TMHs are hydrophobic as they need to span the hydrophobic lipid bilayer  
77 of cellular membranes. They consist of an alpha helix of, on average, 23 amino  
78 acids in length. TMHs can also be predicted with high accuracy from a pro-  
79 tein sequence by bioinformatics approaches [9, 10, 11, 12, 13, 14]. For example,  
80 a study by Jones [12] found that, from 184 transmembrane proteins (TMPs)  
81 with known topology, 80% of the TMH predictions of these proteins matched  
82 the experimental findings. TMHs are common structures in the proteins of hu-  
83 mans and microbes. Different TMH prediction tools estimate that 15-39% of all  
84 proteins in the human proteome contain at least one TMH [15]. However, the

85 physiological reason why peptides derived from TMHs would be presented more  
86 often than peptides stemming from soluble (i.e., extracellular or cytoplasmic)  
87 protein regions is unknown. In this study, we hypothesized that the presen-  
88 tation of TMH residues is evolutionarily preferred, since TMHs are less prone  
89 to undergo escape mutations. One reason to expect such a reduced variability  
90 (and hence evolutionary conservation) in TMHs, is that these are restricted in  
91 their variability by the functional requirement to span a lipid bilayer. This lim-  
92 its many of the amino acids present in TMHs to have hydrophobic side chains  
93 [16, 17]. Therefore, we speculated that the TMHs of pathogens might have a  
94 lower chance to develop escape mutations, as that will result in a dysfunctional  
95 TMH and render the protein inactive.

96 This study had two objectives. First, we aimed to generalize our findings  
97 by predicting the antigenic presentation from different kingdoms of life in both  
98 MHC-I and -II. From these *in silico* predictions, we conclude that TMH-derived  
99 epitopes from a human, viral and bacterial proteome are likely to be presented  
100 more often than expected by chance for most alleles of MHC-I and II. We con-  
101 firmed the presentation of TMH-derived peptides by re-analysis of peptides from  
102 The Immune Epitope Database (IEDB) [18]. Second, we tested our hypothe-  
103 sis that TMHs are more evolutionary conserved than soluble protein regions.  
104 Our analysis of human single nucleotide polymorphisms (SNPs) showed that  
105 random point mutations are indeed less likely to occur within TMHs. These  
106 findings strengthen the emerging notion that TMHs are important for the T  
107 cell-mediated adaptive immune system, and hence are of overlooked importance  
108 in vaccine development.

## 109 2 Methods

### 110 2.1 Predicting TMH epitopes

111 To predict how frequently epitopes overlapping with TMHs are presented, a  
112 similar analysis strategy was applied as described in [7] for several haplotypes  
113 of both MHC-I and MHC-II, and for a human, viral and bacterial proteome.  
114 To summarize, for each proteome, all possible 9-mers (for MHC-I) or 14-mers  
115 (MHC-II) were derived. For each of these peptides, we determined if it over-  
116 lapped with a predicted TMH and if it was predicted to bind to the most frequent  
117 alleles of each haplotype.

118 For MHC-I, 9-mers were used, as this is the length most frequently presented  
119 in MHC-I and was used in our earlier study [7]. For MHC-II, 14-mers were used,  
120 as this is the most frequently occurring epitope length [19]. A human (UniProt  
121 ID UP000005640.9606), viral (SARS-CoV-2, UniProt ID UP000464024) and  
122 bacterial (*Mycobacterium tuberculosis*, UniProt ID UP000001584) reference pro-  
123 teome was used. TMHMM [9] was used to predict the topology of the proteins  
124 within these proteomes. To predict the affinity of an epitope to a certain HLA  
125 allele, **EpitopePrediction** [7] for MHC-I and **MHCnuggets** [20] for MHC-II was  
126 used. Both MHC-I and MHC-II alleles were selected to have a high prevalence  
127 in the population, where the alleles of the MHC-I haplotypes are the alleles  
128 representing the 13 supertypes with over 99.6% coverage of the population's  
129 MHC-I repertoire as defined by [1] [21], and the 21 MHC-II haplotype alleles,  
130 have a phenotypic frequency of 14% or more in the human population [22].

131 We define a protein to be a binder if, for a certain haplotype, any of its 9-mer  
132 or 14-mer peptides have an IC50 value in the lowest 2% of all peptides within a  
133 *proteome* (see supplementary Tables S1 and S2 for values), this differs from our  
134 previous study where we defined a binder as having an IC50 in the lowest 2% of  
135 the peptides within a *protein*. This revised definition precludes bias of proteins

136 that give rise to no or only very few MHC epitopes. To verify that the slight  
137 change in method yields similar results, a side by side comparison is shown in  
138 the supplementary materials, Figures S4A and S4B.

## 139 **2.2 TMH epitopes obtained from experimental data**

140 To obtain experimental confirmation that peptides stemming from TMHs are  
141 presented by MHC-I and MHC-II, we mined the IEDB [18] for confirmed hu-  
142 man MHC-ligands. We queried the IEDB for all linear epitopes obtained from  
143 MHC ligand assays in healthy humans, carrying the MHC alleles as used in this  
144 study. From these epitopes, we kept those that were present exactly once in the  
145 human reference proteome with UniProt ID UP000005640.9606. We predicted  
146 the topology of the protein each epitope was found in, using TMHMM [9], from  
147 which we concluded if the epitope is overlapping with a TMH with at least 1  
148 amino acid.

149 The full analysis can be found at [https://github.com/richelbilderbeek/](https://github.com/richelbilderbeek/bbbq_article_issue_157)  
150 [bbbq\\_article\\_issue\\_157](https://github.com/richelbilderbeek/bbbq_article_issue_157).

### 151 **2.2.1 Evolutionary conservation of TMHs**

152 To determine the evolutionary conservation of TMHs, we first collected human  
153 single nucleotide polymorphisms (SNPs) resulting in a single amino acid substi-  
154 tution to determine if this occurred within a predicted TMH or not.

155 As a data source, multiple NCBI (<https://www.ncbi.nlm.nih.gov/>) databases  
156 were used: the *dbSNP* [23] database, which contains 650 million cataloged non-  
157 redundant human variations (called RefSNPs, [https://www.ncbi.nlm.nih.](https://www.ncbi.nlm.nih.gov/snp/docs/RefSNP_about/)  
158 [gov/snp/docs/RefSNP\\_about/](https://www.ncbi.nlm.nih.gov/snp/docs/RefSNP_about/)), and the databases *gene* (for gene names [24])  
159 and *protein* (for proteins sequences [25]).

160 The first query was a call to the *gene* database for the term 'membrane



161 protein' (in all fields) for the organism *Homo sapiens*. This resulted in 1,077  
162 gene IDs (on December 2020). The next query was a call to the *gene* database  
163 to obtain the gene names from the gene IDs. Per gene name, the *dbSNP* NCBI  
164 database was queried for variations associated with the gene name. As the  
165 NCBI API constrains its users to three calls per second (to assure fair use), we  
166 had to limit the extent of our analysis.

167 The number of SNPs was limited to the first 250 variations per gene, resulting  
168 in  $\approx 61k$  variations. Only variations that result in a SNP for a single amino acid  
169 substitution were analyzed, resulting in  $\approx 38k$  SNPs. The exact amounts can be  
170 found in the supplementary materials, Tables S3 and S4.

171 SNPs were picked based on ID number, which is linked to their discovery  
172 date. To verify that these ID numbers are unrelated to SNP positions, the  
173 relative positions of all analyzed SNPs in a protein were determined. This  
174 analysis showed no positional bias of the SNPs, as shown in supplementary  
175 figure S1.

176 Per SNP, the *protein* NCBI database was queried for the protein sequence.  
177 For each protein sequence, the protein topology was determined using PureseqTM.  
178 Using these predicted protein topologies, the SNPs were scored to be located  
179 within or outside TMHs.

## 180 3 Results

### 181 3.1 TMH-derived peptides are predicted to be over-presented 182 in MHC-I

183 Figure 1A shows the predicted presentation of TMH-derived peptides in MHC-  
184 I, for a human, viral and bacterial proteome. Per MHC-I allele, it shows the  
185 percentage of binders that overlap with a TMH with at least one residue. The

horizontal line shows the expected percentage of TMH-derived epitopes that would be presented, if TMH-derived epitopes would be presented just as likely as epitopes derived from soluble regions, when assuming equal incidence of soluble and TMH-derived epitope presentation. For 11 out of 13 MHC-I alleles, TMH-derived epitopes are predicted to be presented more often than the null expectation, for a human and bacterial proteome. For the viral proteome, 12 out of 13 MHC-I alleles present TMH-derived epitopes more often than expected by chance. The extent of the over-presentation between the different alleles is similar for the probed proteomes, which strengthens our previous conclusion [7] that the hydrophobicity of the MHC-binding groove is the main factor responsible for the predicted over-presentation of TMH-derived peptides.

### 3.2 TMH-derived peptides are predicted to be over-presented in MHC-II

We next wondered if the over-representation of TMH-derived peptides would also be confirmed for MHC-II. Figure 1A shows the percentages of MHC-II epitopes predicted to be overlapping with TMHs for our human, viral and bacterial proteomes. We found that TMH-derived peptides are over-presented in all of the 21 MHC-II alleles, for a human, bacterial and viral proteome, except for HLA-DRB3\*0101 in *M. tuberculosis*. See supplementary Table S5 for the exact TMH and epitope counts.

### 3.3 The over-presentation of TMH-derived peptides is caused by the hydrophobicity of the MHC peptide binding groove

For MHC-I, we previously showed that the over-presentation of TMH-derived peptides is caused by the hydrophobicity of the peptide binding grooves [7]. Fig-

ures 1B and 1C show the extent of over-presentation of TMH-derived epitopes as a function of the hydrophobicity preference score for the different human MHC alleles. An assumed linear correlation explains 88% of the variability in MHC-I. For MHC-II, 62% of the variability is explained by hydrophobicity. This indicates that TMH-derived peptides are over-presented, because the peptide binding grooves of most MHC-I and -II haplotypes are relatively hydrophobic.

### 3.4 Experimental validation of presentation of TMH-derived peptides

The Immune Epitope Database (IEDB) from the National Institutes of Health contains millions of linear epitope sequences obtained by MHC ligand assays. For the MHC alleles used in this study, we obtained 54,303 and 2,484 linear epitope sequences for the MHC-I and MHC-II alleles from human origin respectively. There are relatively few epitopes for MHC-II, as MHC-II has many more different alleles than MHC-I, whereas we selected only the human epitopes found for the 21 MHC-II alleles used in this study.

Figure (2A and S2) shows there are similar levels of over-presentation of TMH-derived epitopes between (1) the percentage of TMH-derived epitopes that is reported in the IEDB database versus (2) the percentage of TMH-derived epitopes that is predicted to be presented in MHC-I alleles. For MHC-II alleles, there were too few epitopes per MHC allele to result in an informative figure.

In figure 2B we grouped all the epitopes presented by MHC-I and MHC-II alleles by the percentage of TMH-derived epitopes, which are 22% and 10%, respectively.

These findings robustly confirm that epitopes derived from human TMHs are presented in both MHC-I and MHC-II, and support that they are over-presented. See the supplementary Table S6 for the exact values.

237 We also mined the IEDB database for epitopes for any type of T-cell response  
238 from the specified alleles, from the total reports 36% and 7% concerned TMH-  
239 derived epitopes in MHC class I and II, respectively (see Figure S3).

240 This data confirms that not only TMH derived epitopes are presented on  
241 MHC, but this also elicits T-cell mediated immune responses.

### 242 **3.5 Human TMHs are evolutionarily conserved**

243 We addressed the question whether there is an evolutionary advantage in pre-  
244 senting TMHs. We determined the conservation of TMHs by comparing the  
245 occurrences of SNPs located in TMHs or soluble protein regions for the genes  
246 coding for membrane proteins. We obtained 911 unique gene names associated  
247 with the phrase 'membrane protein', which are genes coding for both membrane-  
248 associated proteins (MAPs, which have no TMH) and transmembrane proteins  
249 (TMPs, which have at least one TMH). These genes are linked to 4,780 pro-  
250 tein isoforms, of which 2,553 are predicted to be TMPs and 2,237 proteins are  
251 predicted to be MAPs. We obtained 37,630 unique variations, of which 9,621  
252 are SNPs that resulted in a straightforward amino acids substitution, of which  
253 6,062 were located in predicted TMPs. See supplementary Tables S3 and S4 for  
254 the detailed numbers and distributions of SNPs.

255 Per protein, we calculated two percentages: (1) the percentage of a protein  
256 sequence length bearing TMHs, and (2) the percentage of SNPs located within  
257 these predicted TMHs. Each percentage pair was plotted in figure 3A. The  
258 proportion of SNPs found in TMHs varied from none (i.e., all SNPs were in  
259 soluble regions) to all (i.e., all SNPs were in TMHs). To determine if SNPs  
260 were randomly distributed over the protein, we performed a linear regression  
261 analysis, and added a 95% confidence interval on this regression. This linear fit  
262 nearly goes through the origin and has a slope below the line of equality, which

263 shows that less SNPs are found in TMHs than expected by chance.

264 We determined the probability to find the observed amount of SNPs in TMHs  
265 by chance, i.e., when assuming SNPs occur just as likely in soluble domains as  
266 in TMHs. We used a binomial Poisson distribution, where the number of trials  
267 ( $n$ ) equals the number of SNPs, which is 21,208. The probability of success  
268 for the  $i$ th TMP ( $p_i$ ), is the percentage of residues within a TMH per TMP.  
269 These percentages are shown as a histogram in figure 3B. The expected number  
270 of SNPs expected to be found in TMHs by chance equals  $\sum p \approx 4,141$ . As  
271 we observed 3,803 SNPs in TMHs, we calculated the probability of having that  
272 amount or less successes. We used the type I error cut-off value of  $\alpha = 2.5\%$ . The  
273 chance to find, within TMHs, this amount or less SNPs equals  $6.8208 \cdot 10^{-11}$ . We  
274 determined the relevance of this finding, by calculating how much less SNPs are  
275 found in TMHs, when compared to soluble regions, which is the ratio between  
276 the number of SNPs found in TMHs versus the number of SNPs as expected  
277 by chance. In effect, per 1000 SNPs found in soluble protein domains, one finds  
278 918 SNPs in TMHs, as depicted (as percentages) in figure 3C.

279 We split this analysis for TMPs containing only a single TMH (so-called  
280 single-membrane spanners) and TMPs containing multiple TMHs (multi-membrane  
281 spanners). We hypothesized that single-membrane spanners are less conserved  
282 than multi-membrane spanners, because multi-membrane spanners might have  
283 protein-protein interactions between their TMHs, for example to accommodate  
284 active sites, and thus might have additional structural constraints. From the  
285 split data, we did the same analysis as for the total TMPs. Figure 4A shows the  
286 percentages of TMHs for individual proteins as a function of the percentage of  
287 SNPs located in TMHs. For both single- and multi-spanners, a linear regression  
288 shows that less SNPs are found in TMHs, than expected by chance.

289 We also determined the probability to find the observed amount of SNPs by

chance in single- and multi-spanners. For single-spanners, we found 452 SNPs in TMH, where  $\approx 462$  were expected by chance. The chance to observe this or a lower number by chance is 0.319. As this chance was higher than our  $\alpha = 0.025$ , we consider this no significant effect. For the multi-spanners, we found 3,351 SNPs in TMH, where  $\approx 3,678$  were expected by chance. The chance to observe this or a lower number by chance is  $8.315841 \cdot 10^{-12}$ , which means this number is significantly less as explained by variation. The TMHs of multi-spanners are thus significantly more conserved than soluble protein regions, whereas this is not the case for single-spanners.

Also, for single- and multi-spanners, we determined the relevance of this finding by calculating how much less SNPs are found in TMHs when compared to soluble regions, as depicted in figure 4B. In effect, per 1,000 SNPs found in soluble protein domains, one finds 978 SNPs in TMHs of single-spanners and 911 SNPs in TMHs of multi-spanners.

## 4 Discussion

Epitope prediction is important to understand the immune system function and for the design of vaccines. In this study, we provide evidence that epitopes derived from TMHs are a major but overlooked source of MHC epitopes. Our bioinformatics predictions indicate that the TMH-derived epitope repertoire is larger than expected by chance for both MHC-I and MHC-II, regardless of the organism. Moreover, reanalysis of MHC-ligands from the IEDB database confirmed the presentation of TMH-derived epitopes. Therefore, it seems likely that TMH-derived epitopes would also result in enhanced T cell responses, although the conservation of TMHs might promote the deletion of T cells responsive to TMH-derived epitopes by central tolerance mechanisms. Finally, our SNP analysis shows that TMHs are evolutionary more conserved than solvent-exposed

316 protein regions.

## 317 **4.1 Mechanism of MHC presentation of TMH-derived epi-** 318 **topes**

319 Although our data show that TMH-derived epitopes are presented in all classi-  
320 cal MHC-I and MHC-II haplotypes, the molecular mechanisms of how integral  
321 membrane proteins are processed for MHC presentation are largely unknown [7].  
322 Most prominently, the fundamental principles of how TMHs are extracted from  
323 their hydrophobic lipid environments into the aqueous vacuolar lumen, leading  
324 to subsequent proteolytic processing are unresolved.

325 A first possibility is that the extraction of TMPs from the membrane is  
326 mediated by the ER-associated degradation (ERAD) machinery. For MHC class  
327 I (MHC-I) antigen presentation of soluble proteins, the loading of the epitope  
328 primarily occurs at the endoplasmatic reticulum (ER). The chaperones tapasin  
329 (TAPBP), ERp57 (PDIA3), and calreticulin (CALR) [26] first assemble and  
330 stabilize the heavy and light chains of MHC-I. Later, this complex binds to the  
331 transporter associated with antigen processing (TAP) leading to the formation of  
332 the so-called peptide-loading complex (PLC). The PLC drives import of peptides  
333 into the ER and mediates their subsequent loading into the peptide-binding  
334 groove of MHC-I [27]. Membrane proteins first will have to be extracted from  
335 the membrane before they become amenable to this MHC-I loading by the  
336 PLC. In the ER, this process can be orchestrated by the ERAD machinery,  
337 consisting of several chaperones that recognize TMPs, ubiquitinate them, and  
338 extract them from the ER membrane into the cytosol (retrotranslocation) for  
339 proteasomal degradation [28, 29]. Similar to the peptides generated from soluble  
340 proteins, the TMP-derived peptides might then be re-imported by TAP into the  
341 ER for MHC-I loading. This ERAD-driven antigen retrotranslocation might be

342 facilitated by lipid bodies (LBs) [30], since LBs can serve as cytosolic sites for  
343 ubiquitination of ER-derived cargo [31].

344 A second possibility is that TMPs are proteolytically processed by intramem-  
345 brane proteases that cleave TMHs while they are still membrane embedded.  
346 Supporting this hypothesis is the well-established notion that peptides gener-  
347 ated by signal peptide peptidases (SPPs), an important class of intramembrane  
348 proteases that cleave TMH-like signal sequences, are presented on a specialized  
349 class of MHC-I called HLA-E [32]. The loading of peptides generated by SPP  
350 onto MHC-I does not depend on the proteasome and TAP, possibly because  
351 the peptides are directly released into the lumen of the ER [32]. However, this  
352 mechanism cannot explain how most membrane proteins can be processed for  
353 antigen presentation, because SPPs only cleave TMH-like signal sequences at  
354 their C-termini, and N-terminal domains will hence not be removed. Neverthe-  
355 less, the presentation of peptides with a high hydrophobicity index was shown  
356 to be independent of TAP as well [33], suggesting that the TMH peptides might  
357 perhaps be released directly in the ER lumen by other intramembrane proteases.

358 A third possibility is that peptide processing and MHC-loading occur in  
359 multivesicular bodies (MVBs) [32]. TMPs can be routed from the plasma mem-  
360 brane and other organelles by vesicular trafficking to endosomes. Eventually,  
361 these TMPs can be sorted by the endosomal sorting complexes required for  
362 transport (ESCRT) pathway into luminal invaginations that pinch off from the  
363 limiting membrane and form intraluminal vesicles. This thus results in MVBs  
364 where the membrane proteins destined for degradation are located in intralumi-  
365 nal vesicles. Upon the fusion of MVBs with lysosomes, the entire intraluminal  
366 vesicles including the TMPs are degraded [34]. Via this mechanism, TMPs  
367 might well be processed for antigen presentation, particularly since the loading  
368 of MHC-II molecules is well understood to occur in MVBs [35, 36, 37]. However,



369 such processing of membrane proteins in MVBs for antigen presentation poses  
 370 a problem, because complexes of HLA-DR with its antigen-loading chaperon  
 371 HLA-DM were only observed on intraluminal vesicles, but not on the limiting  
 372 membranes of MVBs [37], indicating that epitope loading of MHC-II also oc-  
 373 curs at intraluminal vesicles. This observation hence raises the question how  
 374 the intraluminal vesicles carrying the TMPs destined for antigen presentation  
 375 can be selectively degraded, while the intraluminal vesicles carrying the MHC-II  
 376 remain intact. A second problem is that phagosomes carrying internalized mi-  
 377 crobes lack intraluminal vesicles, and it is hence unclear how TMPs from these  
 378 microbes would be routed to MVBs for MHC-II loading [37].

379 Alternatively to the enzymatic degradation of lipids in MVBs by lipases  
 380 [38, 39], they might be oxidatively degraded by reactions with radical oxygen  
 381 species produced by the NADPH oxidase NOX2 [40]. This oxidation can result  
 382 in a destabilization and disruption of membranes [40] and might thereby lead to  
 383 the extraction of TMPs. Due to the hydrophobic nature of TMHs, however, the  
 384 extracted proteins will likely aggregate and it is unclear how these aggregates  
 385 would be processed further for MHC loading.

## 386 **4.2 Evolutionary conservation of TMHs**

387 In general, one might expect that evolutionary selection shapes an immune  
 388 system where surveillance is directed towards protein regions essential for the  
 389 survival, proliferation and/or virulence or pathogenic microbes, as these will be  
 390 most conserved. In SARS-CoV-2, for example, there is preliminary evidence  
 391 that the strongest selection pressure is directed upon residues that change its  
 392 virulence [41]. These regions, however, may only account for a small part of a  
 393 pathogen’s proteome. Additionally, the structure and function of these essential  
 394 regions might differ widely between different pathogenic proteins. Because of

395 this scarcity and variance in targets, one can imagine that it will be mostly  
396 unfeasible to provide innate immune responses against such rare essential protein  
397 regions, as suggested in a study on influenza [42], where it was found that the  
398 selection pressure exerted by the immune system was either weak or absent.

399 Evolutionary selection of pathogens by a host's immune system, however, is  
400 more likely to occur for protein patterns that are general, over patterns that are  
401 rare. While essential catalytic sites in a pathogenic proteome might be relatively  
402 rare, TMHs are common and thus might be a more feasible target for evolution  
403 to respond to. Indeed, we have found the signature of evolution when both  
404 factors, that is, TMHs and catalytic sites are likely to co-occur, which is in TMPs  
405 that span the membrane at least twice. In contrast to single-spanners, where  
406 we found no significant evolutionary conservation, the TMHs of multi-spanners  
407 are more evolutionary conserved than soluble protein regions. Likely, the TMHs  
408 in many multi-spanners need to interact with each other for correct protein  
409 structure and function and they might hence be more structurally constrained  
410 compared to the TMHs of single-spanners. Thus, we speculate that the human  
411 immune system is more attentive towards TMHs in multi-spanners, as these are  
412 evolutionarily more conserved.

413 There have been more efforts to assess the conservation of TMHs, using  
414 different methodologies. One such example is a study by Stevens and Arkin [43],  
415 in which aligned protein sequence data was used. Also this study found that  
416 TMHs are evolutionarily more conserved, as the mean amino acid substitution  
417 rate in TMHs is about ten percent lower, which is a similar value as we found.  
418 Another example is a study by Oberai, et al. [44] that estimated the conservation  
419 scores for TMHs and soluble regions based on alignments of evolutionary related  
420 proteins, and also found that TMHs are more conserved, with a conservation  
421 score that was 17% higher in TMHs. Note that the last study also found that

422 mutations in human TMHs are likelier to cause a disease, in line with our  
423 conclusion that TMHs are more conserved.

424 Together, from this study, two important conclusions can be drawn. First,  
425 the MHC over-presentation of TMHs is likely a general feature and predicted  
426 to occur for most haplotypes of both MHC-I and -II and for humans as well as  
427 bacterial and viral pathogens. Second, TMHs are genuinely more evolutionary  
428 conserved than soluble protein motifs, at least in the human proteome.

## 429 **5 Acknowledgments**

430 We thank the Center for Information Technology of the University of Gronin-  
431 gen for its support and for providing access to the Peregrine high performance  
432 computing cluster. FB is funded by a Veni grant from the Netherlands Orga-  
433 nization for Scientific Research (016.Veni.192.026) and an Off-Road Grant from  
434 the Dutch Medical Science Foundation (ZonMW 04510011910005). GvdB is  
435 funded by a Young Investigator Grant from the Human Frontier Science Pro-  
436 gram (HFSP; RGY0080/2018), and a Vidi grant from the Netherlands Orga-  
437 nization for Scientific Research (NWO-ALW VIDI 864.14.001). GvdB has re-  
438 ceived funding from the European Research Council (ERC) under the European  
439 Union’s Horizon 2020 research and innovation programme (grant agreement No.  
440 862137).

## 441 **6 Data Accessibility**

442 All code, intermediate and final results are archived at [https://github.com/](https://github.com/richelbilderbeek/bbbq_article)  
443 [richelbilderbeek/bbbq\\_article](https://github.com/richelbilderbeek/bbbq_article).

## 444 7 Authors' contributions

445 RJCB and FB conceived the idea for this research. MVB helped with the  
446 proteome analysis of *M. tuberculosis*. RJCB wrote the code. RJCB, MB, GvdB  
447 and FB wrote the article.

## 448 References

- 449 [1] Ole Lund, Morten Nielsen, Can Kesmir, Anders Gorm Petersen, Claus  
450 Lundegaard, Peder Worning, Christina Sylvester-Hvid, Kasper Lamberth,  
451 Gustav Røder, Sune Justesen, et al. Definition of supertypes for HLA  
452 molecules using clustering of specificity matrices. *Immunogenetics*, 55(12):  
453 797–810, 2004.
- 454 [2] Steven GE Marsh, ED Albert, WF Bodmer, RE Bontrop, B Dupont,  
455 HA Erlich, M Fernández-Viña, DE Geraghty, R Holdsworth, CK Hurley,  
456 et al. Nomenclature for factors of the HLA system, 2010. *Tissue antigens*,  
457 75(4):291, 2010.
- 458 [3] Simone Sommer. The importance of immune gene variability (MHC) in evo-  
459 lutionary ecology and conservation. *Frontiers in zoology*, 2(1):1–18, 2005.
- 460 [4] Mette Voldby Larsen, Alina Lelic, Robin Parsons, Morten Nielsen, Ilka  
461 Hoof, Kasper Lamberth, Mark B Loeb, Søren Buus, Jonathan Bramson,  
462 and Ole Lund. Identification of CD8+ T cell epitopes in the West Nile  
463 virus polyprotein by reverse-immunology using NetCTL. *PloS one*, 5(9),  
464 2010.
- 465 [5] Ingrid MM Schellens, Can Kesmir, Frank Miedema, Debbie van Baarle,  
466 and José AM Borghans. An unanticipated lack of consensus cytotoxic T

- lymphocyte epitopes in HIV-1 databases: the contribution of prediction programs. *Aids*, 22(1):33–37, 2008.
- [6] Sheila T Tang, Krista E van Meijgaarden, Nadia Caccamo, Giuliana Gugin, Michèl R Klein, Pascale van Weeren, Fatima Kazi, Anette Stryhn, Alexander Zaigler, Ugur Sahin, et al. Genome-based in silico identification of new Mycobacterium tuberculosis antigens activating polyfunctional CD8+ T cells in human tuberculosis. *The Journal of Immunology*, 186(2):1068–1080, 2011.
- [7] Frans Bianchi, Johannes Textor, and Geert van den Bogaart. Transmembrane helices are an overlooked source of Major Histocompatibility Complex Class I epitopes. *Frontiers in immunology*, 8:1118, 2017.
- [8] Sorin Istrail, Liliana Florea, Bjarni V Halldórsson, Oliver Kohlbacher, Russell S Schwartz, Von Bing Yap, Jonathan W Yewdell, and Stephen L Hoffman. Comparative immunopeptidomics of humans and their pathogens. *Proceedings of the National Academy of Sciences*, 101(36):13268–13272, 2004.
- [9] Anders Krogh, Björn Larsson, Gunnar Von Heijne, and Erik LL Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology*, 305(3):567–580, 2001.
- [10] Lukas Käll, Anders Krogh, and Erik LL Sonnhammer. A combined transmembrane topology and signal peptide prediction method. *Journal of molecular biology*, 338(5):1027–1036, 2004.
- [11] Masafumi Arai, Hironori Mitsuke, Masami Ikeda, Jun-Xiong Xia, Takashi Kikuchi, Masanobu Satake, and Toshio Shimizu. ConPred II: a consensus

- 492 prediction method for obtaining transmembrane topology models with high  
493 reliability. *Nucleic acids research*, 32(suppl\_2):W390–W393, 2004.
- 494 [12] David T Jones. Improving the accuracy of transmembrane protein topology  
495 prediction using evolutionary information. *Bioinformatics*, 23(5):538–544,  
496 2007.
- 497 [13] Martin Klammer, David N Messina, Thomas Schmitt, and Erik LL  
498 Sonnhammer. MetaTM-a consensus method for transmembrane protein  
499 topology prediction. *BMC bioinformatics*, 10(1):314, 2009.
- 500 [14] Qing Wang, Chongming Ni, Zhen Li, Xiufeng Li, Renmin Han, Feng Zhao,  
501 Jinbo Xu, Xin Gao, and Sheng Wang. PureseqTM: efficient and accu-  
502 rate prediction of transmembrane topology from amino acid sequence only.  
503 *bioRxiv*, page 627307, 2019.
- 504 [15] Mamoun Ahram, Zoi I Litou, Ruihua Fang, and Ghaith Al-Tawallbeh.  
505 Estimation of membrane proteins in the human proteome. *In silico biology*,  
506 6(5):379–386, 2006.
- 507 [16] Tara Hessa, Nadja M Meindl-Beinker, Andreas Bernsel, Hyun Kim, Yoko  
508 Sato, Mirjam Lerch-Bader, IngMarie Nilsson, Stephen H White, and Gun-  
509 nar Von Heijne. Molecular code for transmembrane-helix recognition by  
510 the sec61 translocon. *Nature*, 450(7172):1026–1030, 2007.
- 511 [17] DT Jones, WR Taylor, and JM Thornton. A model recognition approach  
512 to the prediction of all-helical membrane protein structure and topology.  
513 *Biochemistry*, 33(10):3038–3049, 1994.
- 514 [18] Randi Vita, Swapnil Mahajan, James A Overton, Sandeep Kumar Dhanda,  
515 Sheridan Martini, Jason R Cantrell, Daniel K Wheeler, Alessandro Sette,

516 and Bjoern Peters. The immune epitope database (iedb): 2018 update.  
517 *Nucleic acids research*, 47(D1):D339–D343, 2019.

518 [19] Elin Bergseng, Siri Dørum, Magnus Ø Arntzen, Morten Nielsen, Ståle  
519 Nygård, Søren Buus, Gustavo A de Souza, and Ludvig M Sollid. Dif-  
520 ferent binding motifs of the celiac disease-associated hla molecules DQ2.5,  
521 DQ2.2, and DQ7.5 revealed by relative quantitative proteomics of endoge-  
522 nous peptide repertoires. *Immunogenetics*, 67(2):73–84, 2015.

523 [20] Xiaoshan M Shao, Rohit Bhattacharya, Justin Huang, IK Ashok Sivaku-  
524 mar, Collin Tokheim, Lily Zheng, Dylan Hirsch, Benjamin Kaminow, Ash-  
525 ton Omdahl, Maria Bonsack, et al. High-throughput prediction of MHC  
526 class I and II neoantigens with MHCnuggets. *Cancer Immunology Research*,  
527 8(3):396–408, 2020.

528 [21] A. Sette and J. Sidney. Nine major hla class i supertypes account for the  
529 vast preponderance of hla-a and -b polymorphism. *Immunogenetics*, 50:  
530 201–212, 1999.

531 [22] Jason Greenbaum, John Sidney, Jolan Chung, Christian Brander, Bjoern  
532 Peters, and Alessandro Sette. Functional classification of class II human  
533 leukocyte antigen (HLA) molecules reveals seven different supertypes and a  
534 surprising degree of repertoire sharing across supertypes. *Immunogenetics*,  
535 63(6):325–335, 2011.

536 [23] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Eliza-  
537 beth M Smigielski, and Karl Sirotkin. dbSNP: the ncbi database of genetic  
538 variation. *Nucleic acids research*, 29(1):308–311, 2001.

539 [24] Garth R Brown, Vichet Hem, Kenneth S Katz, Michael Ovetsky, Craig  
540 Wallin, Olga Ermolaeva, Igor Tolstoy, Tatiana Tatusova, Kim D Pruitt,

541 Donna R Maglott, et al. Gene: a gene-centered information resource at  
542 NCBI. *Nucleic acids research*, 43(D1):D36–D42, 2015.

543 [25] Eric W Sayers, Tanya Barrett, Dennis A Benson, Evan Bolton, Stephen H  
544 Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael  
545 DiCuccio, Scott Federhen, et al. Database resources of the national center  
546 for biotechnology information. *Nucleic acids research*, 39(suppl\_1):D38–  
547 D51, 2010.

548 [26] Kenneth L Rock, Eric Reits, and Jacques Neefjes. Present yourself! by mhc  
549 class i and mhc class ii molecules. *Trends in immunology*, 37(11):724–737,  
550 2016.

551 [27] Andreas Blees, Dovile Janulienė, Tommy Hofmann, Nicole Koller, Carla  
552 Schmidt, Simon Trowitzsch, Arne Moeller, and Robert Tampé. Structure  
553 of the human mhc-i peptide-loading complex. *Nature*, 551(7681):525–528,  
554 2017.

555 [28] G Michael Preston and Jeffrey L Brodsky. The evolving role of ubiquitin  
556 modification in endoplasmic reticulum-associated degradation. *Biochemical*  
557 *Journal*, 474(4):445–469, 2017.

558 [29] Birgit Meusser, Christian Hirsch, Ernst Jarosch, and Thomas Sommer.  
559 Erad: the long road to destruction. *Nature cell biology*, 7(8):766–772, 2005.

560 [30] Laurence Bougnères, Julie Helft, Sangeeta Tiwari, Pablo Vargas, Benny  
561 Hung-Junn Chang, Lawrence Chan, Laura Campisi, Gregoire Lauvau,  
562 Stephanie Hugues, Pradeep Kumar, et al. A role for lipid bodies in the  
563 cross-presentation of phagocytosed antigens by mhc class i in dendritic  
564 cells. *Immunity*, 31(2):232–244, 2009.



- [31] Toyoshi Fujimoto and Yuki Ohsaki. The proteasomal and autophagic pathways converge on lipid droplets. *Autophagy*, 2(4):299–301, 2006.
- [32] Cláudia C Oliveira and Thorbald van Hall. Alternative antigen processing for mhc class i: multiple roads lead to rome. *Frontiers in immunology*, 6:298, 2015.
- [33] Georg Lautscham, Sabine Mayrhofer, Graham Taylor, Tracey Haigh, Alison Leese, Alan Rickinson, and Neil Blake. Processing of a multiple membrane spanning epstein-barr virus protein for cd8+ t cell recognition reveals a proteasome-dependent, transporter associated with antigen processing-independent pathway. *The Journal of experimental medicine*, 194(8):1053–1068, 2001.
- [34] Jean Gruenberg. Life in the lumen: the multivesicular endosome. *Traffic*, 21(1):76–93, 2020.
- [35] Monique Kleijmeer, Georg Ramm, Danita Schuurhuis, Janice Griffith, Maria Rescigno, Paola Ricciardi-Castagnoli, Alexander Y Rudensky, Ferry Ossendorp, Cornelis JM Melief, Willem Stoorvogel, et al. Reorganization of multivesicular bodies regulates mhc class ii antigen presentation by dendritic cells. *The Journal of cell biology*, 155(1):53–64, 2001.
- [36] Peter J Peters, Jacques J Neefjes, Viola Oorschot, Hidde L Ploegh, and Hans J Geuze. Segregation of mhc class ii molecules from mhc class i molecules in the golgi complex for transport to lysosomal compartments. *Nature*, 349(6311):669–676, 1991.
- [37] Wilbert Zwart, Alexander Griekspoor, Coenraad Kuijl, Marije Marsman, Jacco van Rheenen, Hans Janssen, Jero Calafat, Marieke van Ham, Lennert Janssen, Marcel van Lith, et al. Spatial separation of hla-dm/hla-dr inter-

actions within miic and phagosome-induced immune escape. *Immunity*, 22  
(2):221–233, 2005.

[38] Peter Sander, Katja Becker, and Michael Dal Molin. Lipase processing of  
complex lipid antigens. *Cell chemical biology*, 23(9):1044–1046, 2016.

[39] Martine Gilleron, Marco Lepore, Emilie Layre, Diane Cala-De Paepe,  
Naila Mebarek, James A Shayman, Stéphane Canaan, Lucia Mori, Frédéric  
Carrière, Germain Puzo, et al. Lysosomal lipases plrp2 and lpla2 process  
mycobacterial multi-acylated lipids and generate t cell stimulatory anti-  
gens. *Cell chemical biology*, 23(9):1147–1156, 2016.

[40] Ilse Dingjan, Daniëlle RJ Verboogen, Laurent M Paardekooper, Natalia H  
Revelo, Simone P Sittig, Linda J Visser, Gabriele Fischer Von Mollard,  
Stefanie SV Henriët, Carl G Figdor, Martin Ter Beest, et al. Lipid perox-  
idation causes endosomal antigen release for cross-presentation. *Scientific  
reports*, 6(1):1–12, 2016.

[41] Lauro Velazquez-Salinas, Selene Zarate, Samantha Eberl, Douglas P  
Gladue, Isabel Novella, and Manuel V Borca. Positive selection of ORF3a  
and ORF8 genes drives the evolution of SARS-CoV-2 during the 2020  
COVID-19 pandemic. *bioRxiv*, 2020.

[42] Alvin X Han, Sebastian Maurer-Stroh, and Colin A Russell. Individual  
immune selection pressure has limited impact on seasonal influenza virus  
evolution. *Nature ecology & evolution*, 3(2):302–311, 2019.

[43] Timothy J Stevens and Isaiah T Arkin. Substitution rates in  $\alpha$ -helical  
transmembrane proteins. *Protein Science*, 10(12):2507–2517, 2001.

[44] Amit Oberai, Nathan H Joh, Frank K Pettit, and James U Bowie. Struc-  
tural imperatives impose diverse evolutionary constraints on helical mem-

615 brane proteins. *Proceedings of the National Academy of Sciences*, 106(42):  
616 17747–17750, 2009.

617 [45] Morten Nielsen, Claus Lundegaard, Thomas Blicher, Bjoern Peters,  
618 Alessandro Sette, Sune Justesen, Søren Buus, and Ole Lund. Quantita-  
619 tive predictions of peptide binding to any HLA-DR molecule of known  
620 sequence: NetMHCIIpan. *PLoS computational biology*, 4(7), 2008.

621 [46] Edita Karosiene, Michael Rasmussen, Thomas Blicher, Ole Lund, Søren  
622 Buus, and Morten Nielsen. NetMHCIIpan-3.0, a common pan-specific  
623 MHC class II prediction method including all three human MHC class  
624 II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics*, 65(10):  
625 711–724, 2013.

626 [47] Richèl J C Bilderbeek. tmhmm, 2019. [https://github.com/](https://github.com/richelbilderbeek/tmhmm)  
627 [richelbilderbeek/tmhmm](https://github.com/richelbilderbeek/tmhmm) [Accessed: 2019-03-08].

628 [48] Richèl J C Bilderbeek. pureseqtmr, 2020. [https://github.com/](https://github.com/richelbilderbeek/pureseqtmr)  
629 [richelbilderbeek/pureseqtmr](https://github.com/richelbilderbeek/pureseqtmr) [Accessed: 2020-05-19].

630 [49] Richèl J C Bilderbeek. netmhc2pan, 2019. [https://github.com/](https://github.com/richelbilderbeek/netmhc2pan)  
631 [richelbilderbeek/netmhc2pan](https://github.com/richelbilderbeek/netmhc2pan) [Accessed: 2019-03-08].

632 [50] Richèl J C Bilderbeek. iedbr, 2021. [https://github.com/](https://github.com/richelbilderbeek/iedbr)  
633 [richelbilderbeek/iedbr](https://github.com/richelbilderbeek/iedbr) [Accessed: 2021-11-09].

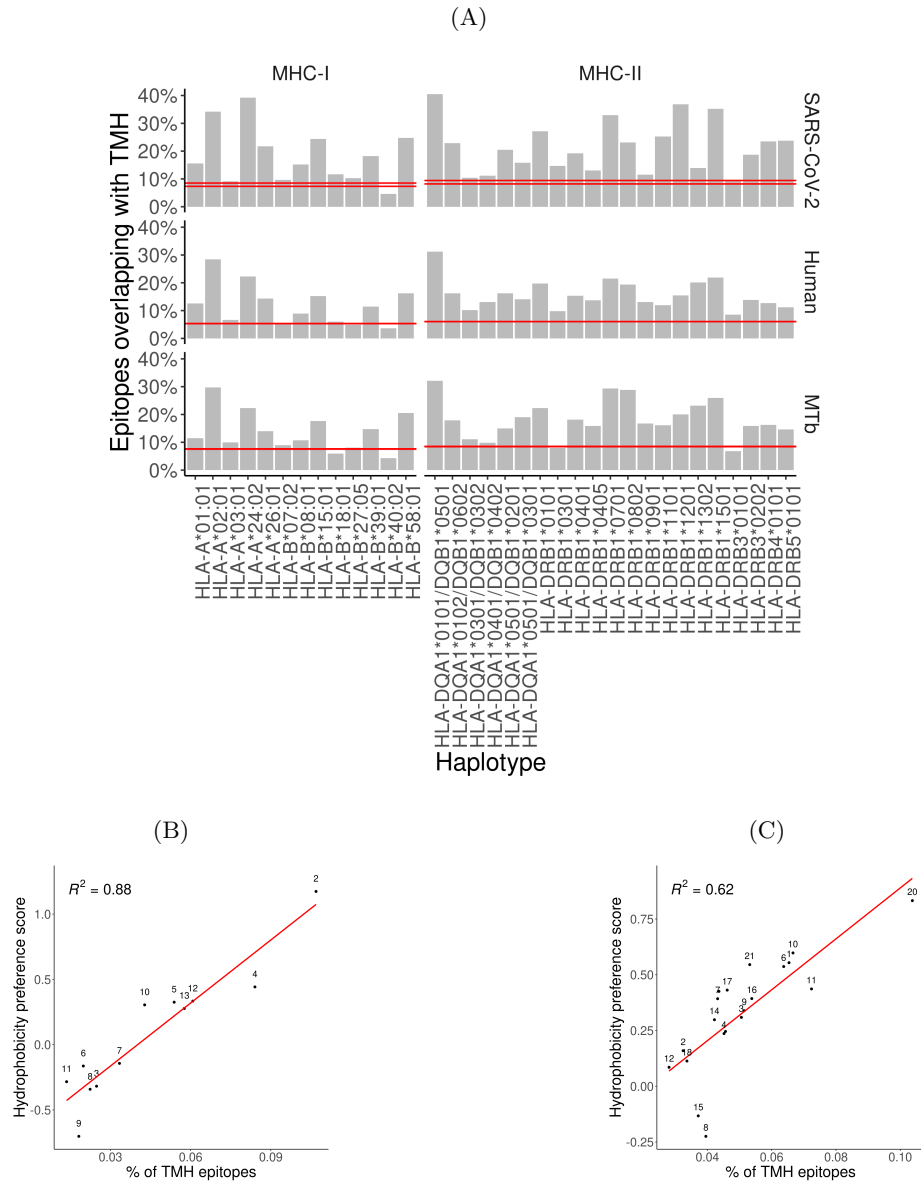
634 [51] Richèl J C Bilderbeek. sprentrez, 2021. [https://github.com/](https://github.com/richelbilderbeek/sprentrez)  
635 [richelbilderbeek/sprentrez](https://github.com/richelbilderbeek/sprentrez) [Accessed: 2021-02-09].

636 [52] Richèl J C Bilderbeek. bbbq, 2020. [https://github.com/](https://github.com/richelbilderbeek/bbbq)  
637 [richelbilderbeek/bbbq](https://github.com/richelbilderbeek/bbbq) [Accessed: 2020-09-02].

- 638 [53] Steffen Möller, Michael DR Croning, and Rolf Apweiler. Evaluation of  
639 methods for the prediction of membrane spanning regions. *Bioinformatics*,  
640 17(7):646–653, 2001.
- 641 [54] Claus Lundegaard, Ole Lund, and Morten Nielsen. Prediction of epitopes  
642 using neural network based methods. *Journal of immunological methods*,  
643 374(1-2):26–34, 2011.
- 644 [55] Morten Nielsen, Claus Lundegaard, Peder Worning, Sanne Lise Lauemøller,  
645 Kasper Lamberth, Søren Buus, Søren Brunak, and Ole Lund. Reliable  
646 prediction of T-cell epitopes using neural networks with novel sequence  
647 representations. *Protein Science*, 12(5):1007–1017, 2003.
- 648 [56] Edita Karosiene, Claus Lundegaard, Ole Lund, and Morten Nielsen.  
649 NetMHCcons: a consensus method for the major histocompatibility com-  
650 plex class I predictions. *Immunogenetics*, 64(3):177–186, 2012.
- 651 [57] Morten Nielsen, Claus Lundegaard, Peder Worning, Christina Sylvester  
652 Hvid, Kasper Lamberth, Søren Buus, Søren Brunak, and Ole Lund. Im-  
653 proved prediction of MHC class I and class II epitopes using a novel Gibbs  
654 sampling approach. *Bioinformatics*, 20(9):1388–1397, 2004.
- 655 [58] David J. Winter. rentrez: an R package for the NCBI eUtils API. *The R*  
656 *Journal*, 9:520–526, 2017.
- 657 [59] Lucia Musumeci, Jonathan W Arthur, Florence SG Cheung, Ashraful  
658 Hoque, Scott Lippman, and Juergen KV Reichardt. Single nucleotide dif-  
659 ferences (SNDs) in the dbSNP database may lead to errors in genotyping  
660 and haplotyping studies. *Human mutation*, 31(1):67–73, 2010.
- 661 [60] Ryan Hunt, Zuben E Sauna, Suresh V Ambudkar, Michael M Gottesman,

662 and Chava Kimchi-Sarfaty. Silent (synonymous) SNPs: should we care  
663 about them? *Single nucleotide polymorphisms*, pages 23–39, 2009.

## 664 8 Figures



**Figure 1: Over-presentation of TMH-derived epitopes on most MHC-I and -II haplotypes (A)** The percentage of epitopes for MHC-I and -II haplotypes that are predicted to overlap with TMHs for the proteomes of SARS-CoV-2 (top row), human (middle row) and *M. tuberculosis* (MtB; bottom row). The pair of horizontal red lines in each plot indicate the lower and upper bound of the 99% confidence interval. See supplementary Tables S5 and S7 for the exact TMH and epitope counts. **(B-C)** Correlation between the percentages of predicted TMH-derived epitopes and the hydrophobicity score of all predicted epitopes for human MHC-I **(B)** and MHC-II alleles **(C)**. Diagonal red line: linear regression analysis. Labels are shorthand for the HLA haplotypes, see the supplementary Table S8 for the names.

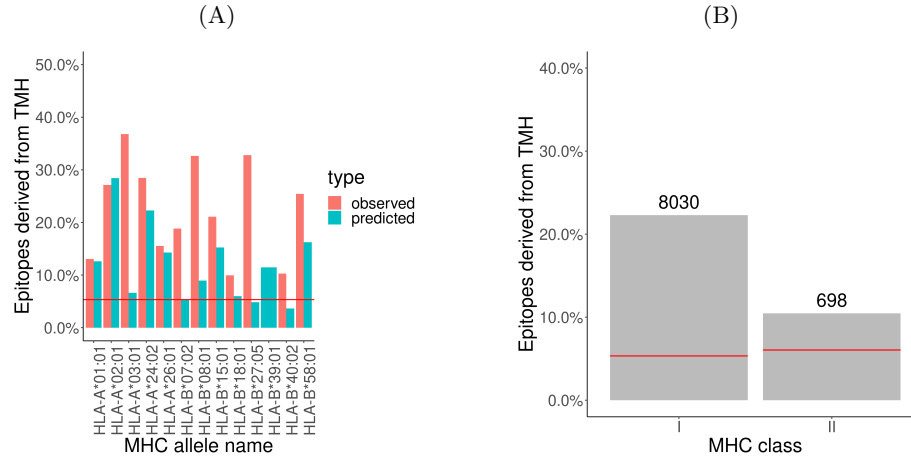


Figure 2: **Robust prediction that TMH epitopes are presented.** The percentage of epitopes for MHC-I and -II alleles that overlap with TMHs that are presented. The pair of horizontal red lines in each plot indicate the lower and upper bound of the 99% confidence interval. Note that only one line is visible as this interval is relatively narrow. Alleles are listed in Table S8). **(A)** Observed and predicted percentage of TMH-derived epitopes for MHC-I alleles. **(B)** MHC ligands from IEDB corresponding to TMH-derived epitopes. The numbers above the bars denotes the number of TMH derived epitopes obtained.



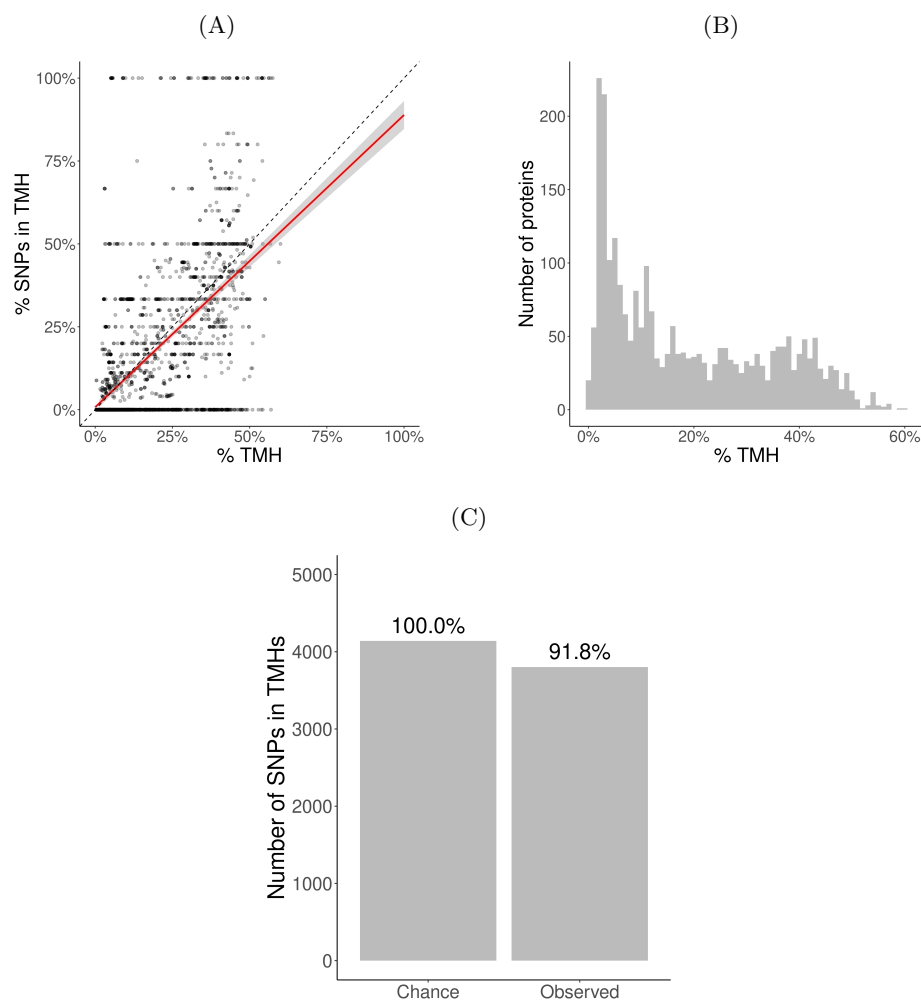
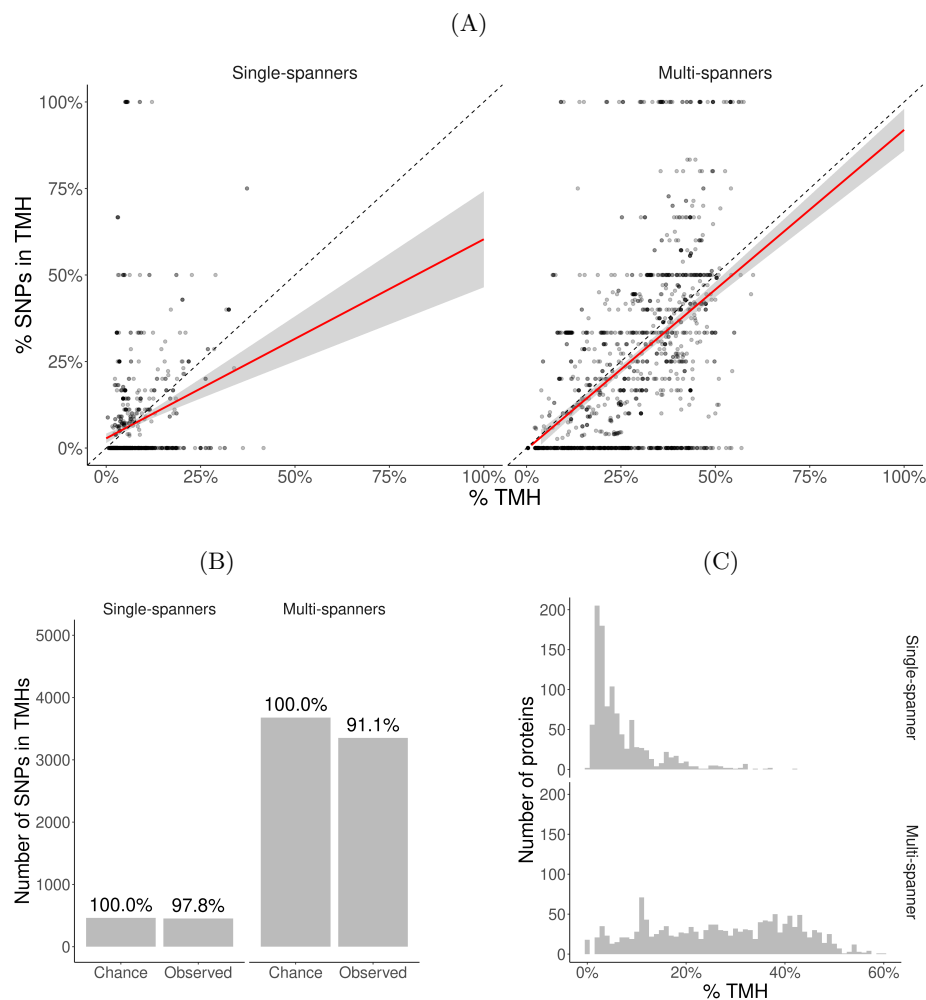


Figure 3: **Evolutionary conservation of human TMHs.** (A) Percentage of SNPs found in TMHs. Each point shows for one protein the predicted percentage of amino acids that are part of a TMH ( $x$ -axis) and the observed occurrence of SNPs being located within a TMH ( $y$ -axis). The dashed diagonal line shows the line of equality (i.e., equal conservation of TMHs and soluble protein regions). The diagonal red line indicates a linear fit, the gray area its 95% confidence interval. (B) Distribution of the percentages of TMH in the TMPs used in this study. (C) The number of SNPs in TMHs as expected by chance (left bar) and found in the dbSNP database (right bar). Percentages show the relative conservation of SNPs in TMHs found relative to stochastic chance.



**Figure 4: Membrane proteins with multiple TMHs are evolutionary more conserved than proteins with only a single TMH.** (A) Percentage of SNPs found in TMPs predicted to have only a single (left) or multiple (right) TMHs. Each point shows for one protein the predicted percentage of amino acids that are part of a TMH ( $x$ -axis) and the observed occurrence of SNPs being located within a TMH ( $y$ -axis). The dashed diagonal lines show the line of equality (i.e., equal conservation of TMHs and soluble protein regions). The diagonal red lines indicate a linear fit, the gray areas their 95% confidence intervals. (B) The number of SNPs in TMHs as expected by chance and observed in the dbSNP database, for TMPs with one TMH (single-spanners) and multiple TMHs (multi-spanners). Percentages show the relative conservation of SNPs in TMHs found relative to the stochastic chances. (C) Distribution of the proportion of amino acids residing in the plasma membrane.

Table S1: IC50 values (in nM) per haplotype below which a peptide is considered a binder. percentage used: 2

haplotype	covid	human	myco
HLA-A*01:01	1470.5912	2545.9537	2812.1714
HLA-A*02:01	118.9596	218.7274	186.7565
HLA-A*03:01	537.0144	804.7455	1544.1073
HLA-A*24:02	984.8147	1590.0623	1971.8258
HLA-A*26:01	1095.2591	1771.6924	1526.1101
HLA-B*07:02	1215.7734	705.6514	435.5361
HLA-B*08:01	886.5661	883.0951	1023.2213
HLA-B*18:01	921.4157	1063.2215	1319.0445
HLA-B*27:05	1186.0963	689.8815	475.6130
HLA-B*39:01	437.3506	484.3843	399.3873
HLA-B*40:02	585.6308	541.2392	600.1688
HLA-B*58:01	435.4693	591.0526	538.9063
HLA-B*15:01	281.9129	440.6541	482.8369

## A Supplementary materials

### A.1 IC50 values of binders per haplotype

Per target proteome (i.e. human, SARS-CoV-2, *M tuberculosis*), we collected all 9-mers (for MHC-I) and 14-mers (for MHC-II), after removing the selenoproteins and proteins that are shorter than the epitope length. From these epitopes, per MHC haplotype, we predicted the IC50 (in nM) using `epitope-prediction` (for MHC-I) and MHCnuggets (for MHC-II). Here, we show the IC50 value per haplotype that is used to determine if a peptide binds to the haplotype's MHC for MHC-I (see supplementary Table S1) and MHC-II (see supplementary Table S2).

Table S2: IC50 values (in nM) per haplotype below which a peptide is considered a binder. percentage used: 2

haplotype	covid	human	myco
HLA-DRB1*0101	7.3896	9.72	9.9600
HLA-DRB1*0301	121.8420	198.40	164.4900
HLA-DRB1*0401	59.8780	74.92	84.3112
HLA-DRB1*0405	46.2324	51.88	66.7100
HLA-DRB1*0701	17.7464	22.40	28.1700
HLA-DRB1*0802	99.7592	137.16	67.9900
HLA-DRB1*0901	42.3464	53.52	41.5400
HLA-DRB1*1101	35.9988	39.01	48.9200
HLA-DRB1*1201	194.4408	248.72	289.7300
HLA-DRB1*1302	21.1084	40.59	35.4100
HLA-DRB1*1501	32.6196	40.69	46.6700
HLA-DRB3*0101	175.2984	298.94	218.7300
HLA-DRB3*0202	176.8168	291.95	405.8724
HLA-DRB4*0101	47.6384	51.04	62.7800
HLA-DRB5*0101	32.8872	43.52	60.2312
HLA-DQA1*0501/DQB1*0201	193.1108	209.89	174.2124
HLA-DQA1*0501/DQB1*0301	51.2028	43.47	20.3200
HLA-DQA1*0301/DQB1*0302	361.8180	365.96	296.4712
HLA-DQA1*0401/DQB1*0402	214.1932	242.68	199.8912
HLA-DQA1*0101/DQB1*0501	550.4488	674.95	930.9612
HLA-DQA1*0102/DQB1*0602	157.4480	174.82	114.3512

Table S3: Amounts. raw = all variations, including DNA variations. all\_proteins = all proteins. map = membrane associated protein. tmp = transmembrane protein. in\_tmh = in transmembrane helix of TMP. in\_sol = in soluble region of TMP.

what	raw	all_proteins	map	tmp	in_tmh	in_sol
Number of variations	60931	37831	16623	21208	3803	17405
Number of unique variations	60544	37630	16606	21024	3789	17235
Number of unique SNPs	NA	9621	4219	6026	1140	4936
Number of unique gene names	953	911	457	605	325	590
Number of unique protein names	5163	4780	2227	2553	1280	2467
Percentage TMH	NA	10	0	19	26	18

Table S4: Amounts. single\_in\_tmh = in transmembrane helix of single-spanner. single\_in\_sol = in soluble region of single-spanner. multi\_in\_tmh = in transmembrane helix of multi-spanner. multi\_in\_sol = in soluble region of multi-spanner.

what	single_in_tmh	single_in_sol	multi_in_tmh	multi_in_sol
Number of variations	452	7734	3351	9671
Number of unique variations	451	7733	3338	9502
Number of unique SNPs	160	2393	994	2762
Number of unique gene names	96	282	243	344
Number of unique protein names	304	1032	976	1435
Percentage TMH	11	5	35	26

## A.2 Counts

See supplementary Tables S3 and S4 for an overview of all amounts. Note that, for the analyses using the SARS-CoV-2 virus proteome, we labeled this by its disease (covid) to prevent typos. In supplementary Table S3 there are multiple instances where the amounts are expected to add up, yet don't, as one SNP can work on multiple isoforms. For example, there are 9,621 unique SNPs found in all proteins, of which 4,219 around found in MAPs and 6,026 in TMPs. Apparently, 624 SNPs work on a set of isoforms that contains both MAPs and TMPs.

### 684 A.3 Relative positions

685 See Supplementary Figure S1 for the distribution of the relative position of the  
686 SNPs.

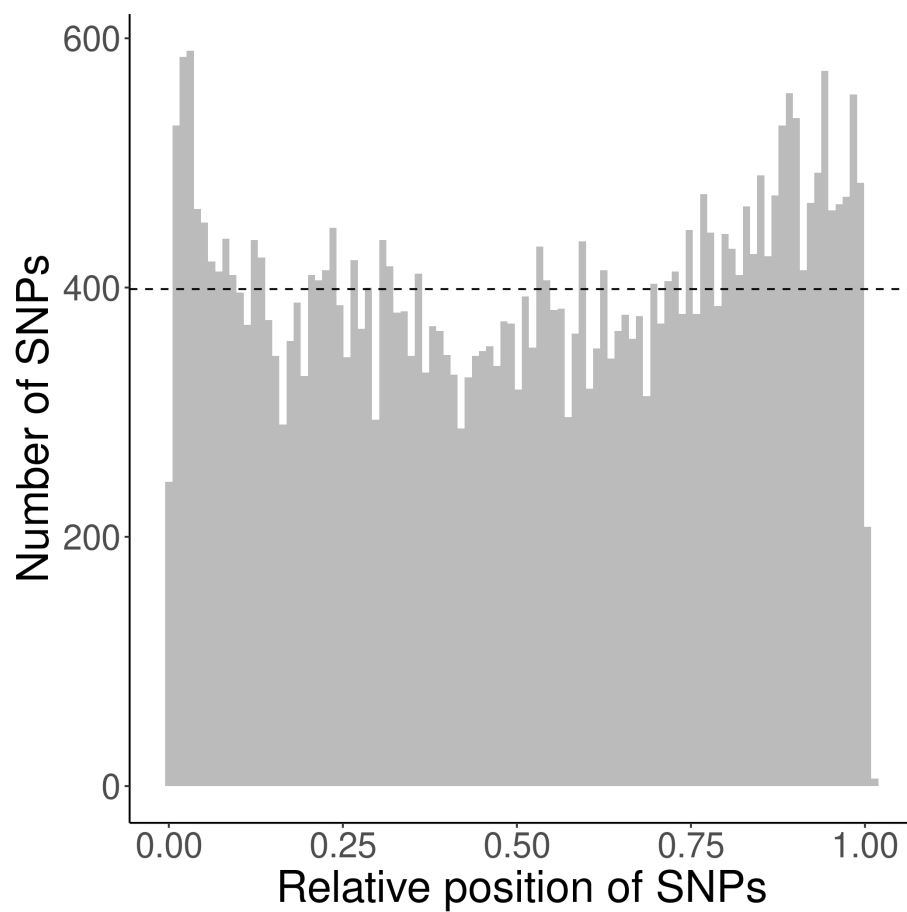


Figure S1: Distribution of the relative position of the SNPs used, where a relative position of zero denotes the first amino acid at the N-terminus, where a relative position of one indicates the last residue at the C-terminus.

Table S5: Percentage of MHC-II 14-mers overlapping with TMH. Values in brackets show the number of binders that have at least one residue overlapping with a TMH (first value) as well as the number of binders (second value). percentage used: 2

haplotype	covid	human	myco
HLA-DQA1*0101/DQB1*0501	40.433 (112/277)	31.214 (69752/223464)	32.158 (8187/25459)
HLA-DQA1*0102/DQB1*0602	22.910 (74/323)	16.167 (35753/221147)	17.950 (4608/25671)
HLA-DQA1*0301/DQB1*0302	10.381 (30/289)	10.179 (22623/222248)	11.144 (2842/25502)
HLA-DQA1*0401/DQB1*0402	11.111 (32/288)	13.135 (29319/223219)	9.890 (2524/25522)
HLA-DQA1*0501/DQB1*0201	20.430 (57/279)	16.240 (36186/222820)	14.999 (3823/25489)
HLA-DQA1*0501/DQB1*0301	15.808 (46/291)	14.106 (31046/220089)	18.969 (4878/25715)
HLA-DRB1*0101	27.119 (80/295)	19.774 (43968/222349)	22.293 (5692/25533)
HLA-DRB1*0301	14.676 (43/293)	9.801 (21831/222752)	7.956 (2025/25451)
HLA-DRB1*0401	19.231 (55/286)	15.325 (34011/221930)	18.113 (4641/25623)
HLA-DRB1*0405	12.996 (36/277)	13.684 (30380/222012)	15.837 (4036/25484)
HLA-DRB1*0701	32.877 (96/292)	21.512 (47856/222465)	29.304 (7471/25495)
HLA-DRB1*0802	23.132 (65/281)	19.339 (42859/221623)	28.805 (7358/25544)
HLA-DRB1*0901	11.565 (34/294)	13.111 (29043/221520)	16.798 (4301/25605)
HLA-DRB1*1101	25.197 (64/254)	11.924 (26582/222928)	16.103 (4101/25467)
HLA-DRB1*1201	36.897 (107/290)	15.482 (34596/223464)	20.018 (5098/25467)
HLA-DRB1*1302	13.962 (37/265)	20.121 (44798/222646)	23.141 (5935/25647)
HLA-DRB1*1501	35.206 (94/267)	21.836 (48671/222893)	25.891 (6584/25430)
HLA-DRB3*0101	9.158 (25/273)	8.496 (18884/222274)	6.819 (1740/25517)
HLA-DRB3*0202	18.657 (50/268)	13.832 (30687/221859)	15.843 (4059/25620)
HLA-DRB4*0101	23.529 (68/289)	12.749 (28376/222568)	16.221 (4131/25467)
HLA-DRB5*0101	23.776 (68/286)	11.235 (24993/222464)	14.648 (3732/25478)

#### A.4 Presentation of TMH-derived epitopes

See supplementary Table S5 for the percentage of MHC-II 14-mers overlapping with TMH.

## 690 A.5 The percentage of TMH-derived epitopes from IEDB 691 epitopes

692 We display the over-presentation of epitopes taken from the IEDB database, for  
693 two assays: an MHC ligand assay (Figure 2A) and a T cell assay (see figure S3),  
694 as a bar plot. Supplementary Table S6 below shows the exact numbers.

MHC class	Dataset	n
I	iedb_mhc_ligand	22.28% (1789/8030)
I	iedb_t_cell	35.91% (93/259)
II	iedb_mhc_ligand	10.46% (73/698)
II	iedb_t_cell	6.66% (42/631)

Table S6: Percentage of epitopes derived from a TMH for epitopes taken from the IEDB, for two different types of assays: an MHC ligand assay, as well as a T cell assay. The values between brackets show the the number of epitopes that were predicted to overlapping with a TMH per all epitopes that could be uniquely mapped to the representative human reference proteome.



## 695 A.6 Correlation of epitope presentation

696 In the main text of this research, we use two sources of epitopes to determine  
 697 if TMH-derived epitopes are presented. The first source of epitopes are all the  
 698 9-mers (for MHC-I) (and 14-mers for MHC-II) derived from a human reference  
 699 proteome, where this over-presentation is displayed in figure 1A. The second  
 700 source of epitopes are those that are present in the IEDB that are obtained  
 701 from MHC ligand assays, as displayed in figure 2A.

702 Here we correlate between the over-presentation of TMH-derived epitopes  
 703 between these two sources of data. Figure S2 shows per haplotype the percent-  
 704 age of TMH-derived epitopes, with a linear trendline.

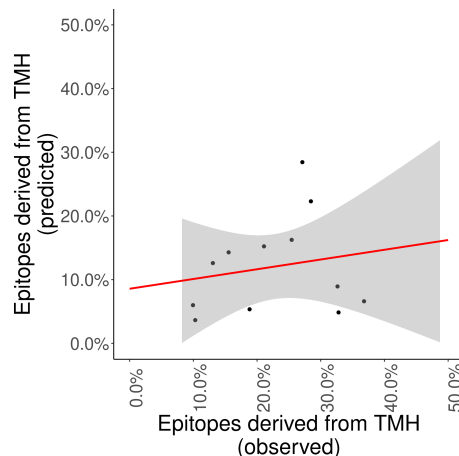


Figure S2: **TMH-derived epitopes are over-presented when using predicted as well as experimental data** For the MHC class I alleles, the over-presentation of TMH-derived epitopes is correlated between IEDB MHC ligand epitopes (horizontal axis) and the 9-mers derived from a human reference proteome (vertical axis). Alleles are listed in Table S8). The trendline shows the linear correlation between these percentages, where the gray area is the 95% confidence interval.

705 **A.7 Presentation of TMH-derived epitopes result in T cell**  
 706 **responses**

707 Figure S3 shows the percentage of TMH-derived epitopes of the reported epi-  
 708 topes from human origin for which T-cell responses were established. The data  
 709 was obtained from the IEDB and includes only the MHC alleles used in this  
 710 study. As there are many (especially class II) MHC alleles, only a small per-  
 711 centage of the full IEDB data could be used.

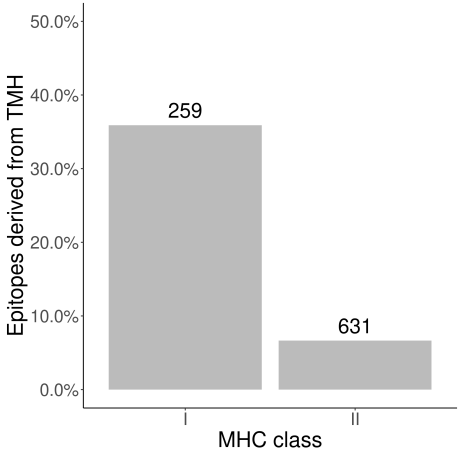


Figure S3: **TMH-derived epitopes evoke T-cell responses** The numbers above the bars denotes the number of epitopes found in the IEDB for the MHC alleles used in this study.

Table S7: Percentage of MHC-I 9-mers overlapping with TMH. Values in brackets show the number of binders that have at least one residue overlapping with a TMH (first value) as well as the number of binders (second value). percentage used: 2

haplotype	covid	human	myco
HLA-A*01:01	15.603 (44/282)	12.600 (28377/225209)	11.424 (2947/25797)
HLA-A*02:01	34.155 (97/284)	28.441 (63994/225003)	29.749 (7646/25702)
HLA-A*03:01	9.122 (27/296)	6.606 (14851/224796)	9.972 (2565/25721)
HLA-A*24:02	39.223 (111/283)	22.297 (50313/225648)	22.346 (5752/25741)
HLA-A*26:01	21.739 (65/299)	14.287 (32232/225598)	13.950 (3598/25793)
HLA-B*07:02	9.712 (27/278)	5.347 (11893/222429)	8.899 (2291/25744)
HLA-B*08:01	15.248 (43/282)	8.935 (19981/223616)	10.714 (2750/25667)
HLA-B*15:01	24.324 (72/296)	15.228 (34498/226542)	17.600 (4547/25835)
HLA-B*18:01	11.724 (34/290)	5.993 (13409/223745)	5.960 (1536/25773)
HLA-B*27:05	10.227 (27/264)	4.854 (10882/224178)	8.031 (2063/25688)
HLA-B*39:01	18.182 (50/275)	11.468 (25621/223419)	14.682 (3787/25793)
HLA-B*40:02	4.594 (13/283)	3.647 (8147/223408)	4.264 (1097/25729)
HLA-B*58:01	24.731 (69/279)	16.245 (36409/224119)	20.558 (5292/25742)

## A.8 Presentation of TMH-derived epitopes

See supplementary Table S7 for the percentage of MHC-I 9-mers overlapping with TMH.

Supplementary Table S8 shows the shorthand notation for the HLA haplotypes.

Supplementary Tables S7 and S5 show the exact number of binders, binders that overlap with TMHs and the percentage of binders that overlap with TMHs, as visualized by figure 1A.

index	haplotype_name
1	HLA-A*01:01
2	HLA-A*02:01
3	HLA-A*03:01
4	HLA-A*24:02
5	HLA-A*26:01
6	HLA-B*07:02
7	HLA-B*08:01
8	HLA-B*18:01
9	HLA-B*27:05
10	HLA-B*39:01
11	HLA-B*40:02
12	HLA-B*58:01
13	HLA-B*15:01
1	HLA-DRB1*0101
2	HLA-DRB1*0301
3	HLA-DRB1*0401
4	HLA-DRB1*0405
5	HLA-DRB1*0701
6	HLA-DRB1*0802
7	HLA-DRB1*0901
8	HLA-DRB1*1101
9	HLA-DRB1*1201
10	HLA-DRB1*1302
11	HLA-DRB1*1501
12	HLA-DRB3*0101
13	HLA-DRB3*0202
14	HLA-DRB4*0101
15	HLA-DRB5*0101
16	HLA-DQA1*0501/DQB1*0201
17	HLA-DQA1*0501/DQB1*0301
18	HLA-DQA1*0301/DQB1*0302
19	HLA-DQA1*0401/DQB1*0402
20	HLA-DQA1*0101/DQB1*0501
21	HLA-DQA1*0102/DQB1*0602

Table S8: Abbreviations of the haplotype names

## 720 A.9 Differences with Bianchi et al., 2017

721 A part of this study does the same analysis as Bianchi et al., 2017. mainly  
722 concern the use of different software and a different definition of what an MHC  
723 binder is.

724 The earlier study defined a peptide an MHC binder if *within the protein* in  
725 which it was found, is was among the peptides with the 2% lowest IC50 val-  
726 ues. This can be seen at [https://github.com/richelbilderbeek/bianchi\\_](https://github.com/richelbilderbeek/bianchi_et_al_2017/blob/master/predict-binders.R)  
727 [et\\_al\\_2017/blob/master/predict-binders.R](https://github.com/richelbilderbeek/bianchi_et_al_2017/blob/master/predict-binders.R), where the binders are written  
728 to file.

729 However, in this study, an MHC binder is defined as a peptide within a  
730 *proteome* in which it is found, that is among the peptides with the 2% lowest  
731 IC50 values. Subsection A.1 shows the IC50 values for a binder per haplotype.

732 Our previous study used the TMHMM web server to predict TMHs. The  
733 desktop version of TMHMM, however, gives an error message on the 25 seleno-  
734 proteins found in the human reference proteome. For the sake of reproducible  
735 research, we used the desktop version (as we can call it from scripts) and, due  
736 to this, we removed the selenoproteins from this analysis.

737 To verify if the previous and the current method give rise to notable differ-  
738 ence, we show a side-by-side comparison in figures S4A and S4B. The figures  
739 that haplotypes that over-present or under-present TMH-derived epitopes, do  
740 so in both studies. The extent to which TMH-derived epitopes are presented,  
741 however, is more extreme in our current setup.

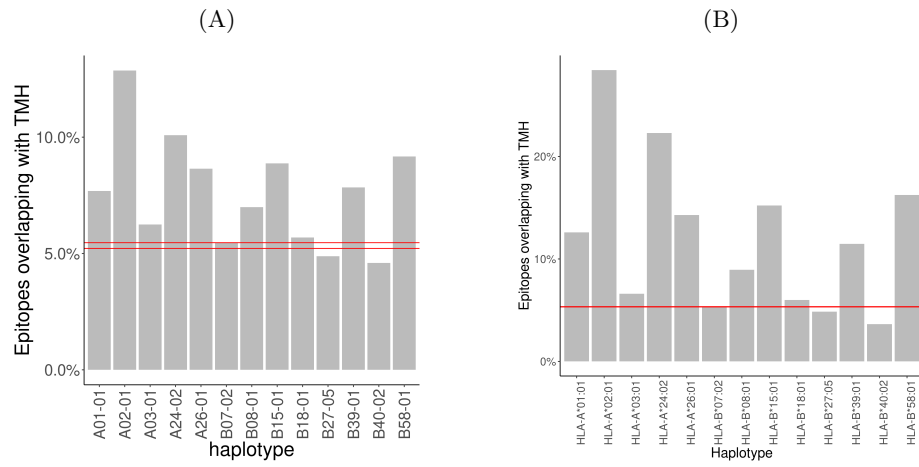


Figure S4: **(A)** Results for [7]. Dashed lines denotes the coincidence interval. **(B)** Results for this study. Dashed line denotes the percentage as expected by chance.

Goal	Tool	Reference
Predict topology	TMHMM	[9]
Predict topology	PureseqTM	[14]
Predict epitopes MHC-I	<b>epitope-prediction</b>	[7]
Predict epitopes MHC-II	NetMHCIIpan	[45, 46]
Call TMHMM from R	<b>tmhmm</b>	[47]
Call PureseqTM from R	<b>pureseqtmr</b>	[48]
Call NetMHCIIpan from R	<b>netmhc2pan</b>	[49]
Work with IEDB	<b>iedbr</b>	[50]
Work with <b>rentrez</b>	<b>sprentrez</b>	[51]
Combine all	<b>bbbq</b>	[52]

Table S9: Overview of all software used in this research.

## 742 A.10 Prediction software used

743 For this research, we needed software to predict protein topology, as well as the  
744 MHC-I and MHC-II binding affinities of epitopes. We selected our software, by  
745 searching the scientific literature to identify the most recent free and open source  
746 (FOSS) prediction software. This was done by searching for papers that (1) cite  
747 older prediction software, and (2) present a novel method to make predictions.  
748 As a starting point, per type of prediction software, a review paper was used  
749 ([53] for protein topology, [54] for MHC-I binding affinities and [55] for MHC-II  
750 binding affinities).

751 There are multiple computational tools developed to predict which parts of  
752 a protein forms a TMH. In 2001, multiple of such prediction tools have been  
753 compared [53], of which TMHMM [9] turned out to be the most accurate, as  
754 is used in the previous study [7]. However, TMHMM has a restrictive software  
755 license and is nearly two decades old. Therefore, PureseqTM [14], was also used  
756 in this study, which has been more recently developed and has a free software  
757 license.

758 For MHC-I, there are multiple computational tools developed to predict epi-  
759 topes. According to [54], at that time, NetMHCcons [56] gave the best predic-

760 tions. We used the same tool as used in our earlier study, **epitope-prediction**  
761 [7],

762 Also for MHC-II, there are multiple computational tools developed to pre-  
763 dict epitopes, such as using a trained neural network [55] or a Gibbs sam-  
764 pling approach [57]. According to [54], in 2011, from a set of multiple tools,  
765 NetMHCIipan [45, 46] made the most accurate predictions. The most recent  
766 FOSS tool available now appears to be MHCnuggets [20], which can do both  
767 MHC-I and MHC-II predictions. As we already use **epitope-prediction** [7]  
768 for MHC-I predictions, we use MHCnuggets only for MHC-II predictions.

769 To retrieve the data from the NCBI databases the **rentrez** R package [58]  
770 was used that calls the NCBI database’s API. The NCBI database provides a  
771 stable user experience for all users, by limiting its API to 3 calls per second  
772 per user. Additionally, the API splits the result of a bigger query into multiple  
773 pages, each of which needs one API call. The **sprentrez** package [51] provides  
774 for bigger queries of multiple (and delayed) API calls.

775 To retrieve the data from the IEDB databases [18], the **iedbr** R package [50]  
776 was written, to calls the IEDB database’s API. Similar to the NCBI database,  
777 the IEDB has a limit to 1 call per second per user and allows a query results to  
778 return 10k results maximally. The **iedbr** package [50] allows for bigger queries.



## 779 A.11 Prediction software written

780 The R programming language is used for the complete experiment, including the  
781 analysis. The complete experiment is bundled in the 'bbbq' R package, which  
782 is dependent on 'tmhmm', 'pureseqtmr', 'epitope-prediction' and 'mhc-nuggets-r'  
783 as described below.

784 The R package 'tmhmm' was developed to do the similar topology predic-  
785 tions as our earlier study (that used 'TMHMM'), yet in an automated way.  
786 'TMHMM' has a restrictive software license [9] and allows a user to download a  
787 pre-compiled executable after confirmation that he/she is in academia. The R  
788 package respects this restriction and allows the user to install and use TMHMM  
789 from within R, as done in this study. 'tmhmm' has been submitted to and is  
790 accepted by the Comprehensive R Archive Network (CRAN).

791 To be able to call, from R, the TMH prediction software 'PureseqTM' [14],  
792 which is written in C, the package 'pureseqtmr' has been developed. 'purese-  
793 qtmr' allows to install 'PureseqTM' and use most of its features. 'pureseqtmr'  
794 has been submitted to and is accepted by CRAN.

795 MHCnuggets is a free and open-source Python package to predict epitope  
796 affinity for many MHC-I and MHC-II variants [20]. The R package 'mhc-  
797 nuggets-r' allows one to install and use MHCnuggets from within R. Also 'mhc-  
798 nuggets-r' has been submitted to and is accepted by CRAN.

799 To reproduce the full experiment presented in this paper, the functions  
800 needed are bundled in the 'bbbq' R package. This package is too specific to  
801 be submitted to CRAN.

Table S10: Percentage of spots and spots that overlap with a TMH

target	mhc_class	n_spots	n_spots_tmh	f_tmh
covid	1	14207	1124	7.91
covid	2	14137	1245	8.81
human	1	11220940	598391	5.33
human	2	11118448	672273	6.05
myco	1	1299707	98613	7.59
myco	2	1279742	108419	8.47

## 802 A.12 Prediction of percentage of epitopes overlapping with 803 a TMH

804 Supplementary Table S10 shows an overview of the findings, where a target  
805 specifies the source of the proteome, where `covid` denotes SARS-CoV-2 and  
806 `myco` denotes *Mycobacterium tuberculosis*. `mhc_class` denotes the MHC class,  
807 `n_spots` the number of possible 9-mers (for MHC-I) or 14-mers (for MHC-II)  
808 possible. `n_spots_tmh` the number of epitopes that overlapped with a TMH  
809 that were binders. `f_tmh` the percentage of peptides that had at least 1 residue  
810 overlapping with a TMH.

### 811 **A.13 Minor methods**

812 These are details that are removed from the 'Methods' section.

813 PureseqTM does not predict the topology of proteins that have less than  
814 three amino acids. The TRDD1 ('T cell receptor delta diversity 1') protein,  
815 however, is two amino acids long. The R package `pureseqtmr`, however, predicts  
816 that mono- and di-peptides are cytosolic.

### 817 **A.14 Minor discussion**

818 These are details that are removed from the 'Discussion' section.

819 In this experiment we predicted epitopes that overlap with TMHs from a  
820 human, bacterial and viral proteome, would these proteins be expressed in a  
821 human host. Bacteria, however have different cell membranes and cell walls,  
822 hence different structural requirements for a TMH. Both topology prediction  
823 tools were trained to recognize human TMHs, thus we cannot be sure that  
824 the transmembrane regions predicted in bacterial proteins are actually part of a  
825 TMH. For the purpose of this study, we assume the error in topology predictions  
826 to be unbiased way towards topology. In other words: that a bacterial TMH is  
827 incorrectly predicted to be absent just as often as it is incorrectly predicted to  
828 be present elsewhere.

829 Regarding the evolutionary conservation of TMHs using SNPs, again, it is  
830 estimated that approximately ten percent of SNPs is a false positive that result  
831 from the methods to determine a SNP. One example is that sequence variations  
832 are incorrectly detected due to highly similar duplicated sequences [59]. We  
833 assume that these duplications occur as often in TMHs as in regions around  
834 these, hence we expect this not to affect our results.

835 In our evolutionary experiment, we removed variations that were synony-  
836 mous mutations (i.e. resulted in the same amino acid, from a different genetic

code) from our analysis. There is evidence, however, that these synonymous mutations do have an effect and may even be evolutionary selected for [60]. As the possible effect of synonymous mutations is ignored by our topology prediction software, we do so as well.

## 841 A.15 Relative presentation of TMH-derived epitopes

842 To compare the over-presentation of TMH-derived epitopes between the differ-  
 843 ent proteomes, we normalized this percentages in such a way that 1.0 is the  
 844 percentage of TMH-derived epitopes that would be expected by chance. Figure  
 845 S5 and S6 show these normalized values for the MHC-I and MHC-II haplotypes  
 846 respectively.

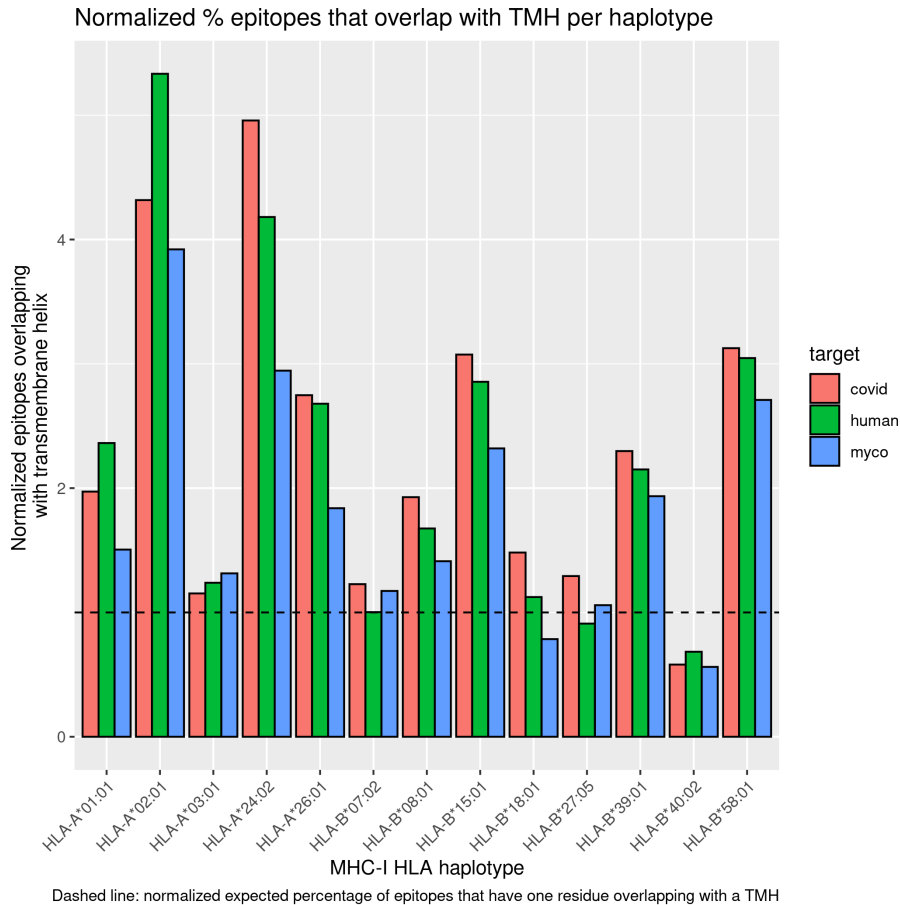


Figure S5: Normalized proportion of MHC-I epitopes overlapping with TMHs for human, viral and bacterial proteomes. Legend: covid = SARS-CoV-2, human = *Homo sapiens*, myco = *Mycobacterium tuberculosis*

847 To determine the additional over-presentation of TMH-derived epitopes in

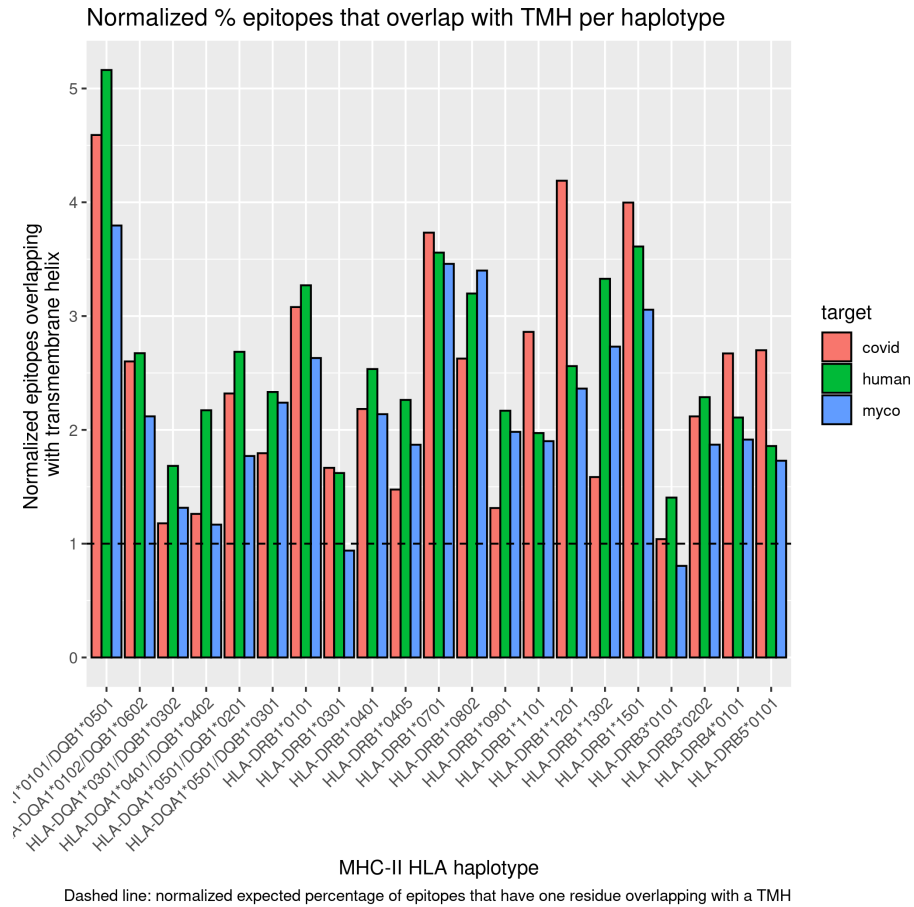


Figure S6: Normalized proportion of MHC-II epitopes overlapping with TMHs for human, viral and bacterial proteomes. Legend: covid = SARS-CoV-2, human = *Homo sapiens*, myco = *Mycobacterium tuberculosis*

848 MHC-II (as compared to MHC-I), we normalized the data to enable a side-  
 849 by-side comparison. The percentage of TMH-derived epitopes presented was  
 850 normalized to the expected percentage of TMH-derived epitopes, where 1.0  
 851 denotes that the percentage of presented TMH-derived epitopes matches the  
 852 values as expected by chance. The normalized values per haplotype are shown  
 853 in figure S7. To compare the TMH-derived over-presentation per MHC class, we  
 854 grouped the normalized values per haplotype, and plot the mean and standard  
 855 error, as shown in figure S8.

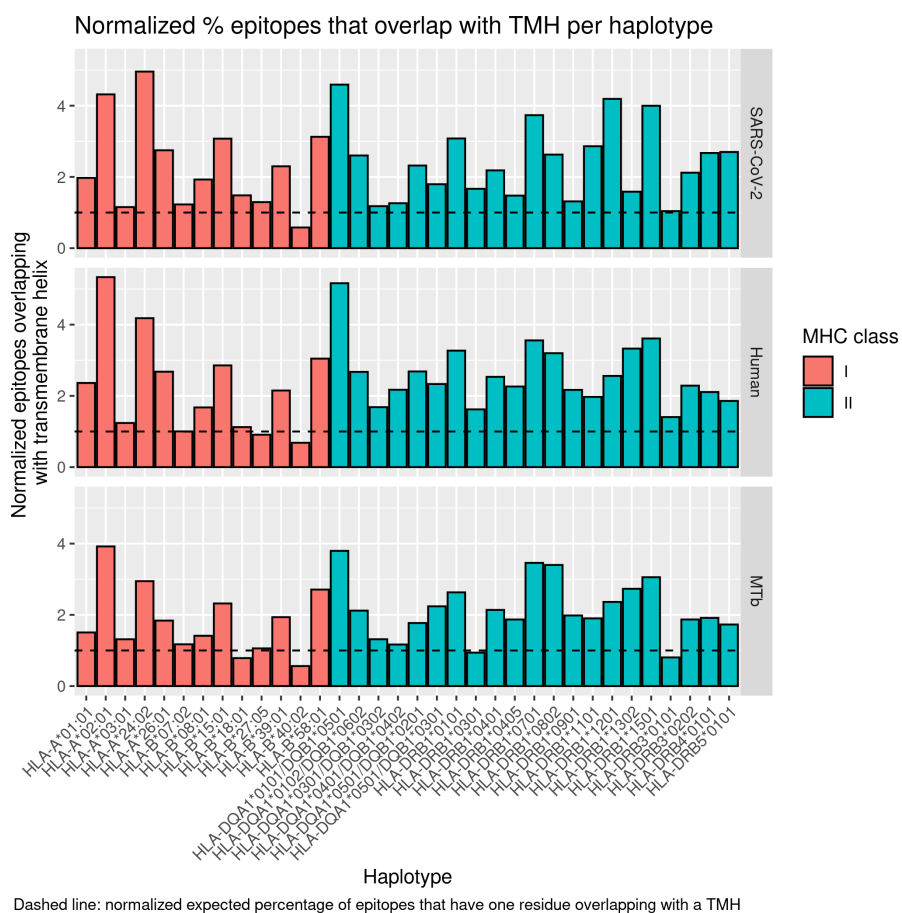


Figure S7: Normalized proportion of MHC-I and MHC-II epitopes overlapping with TMHs, for the different haplotypes and proteomes

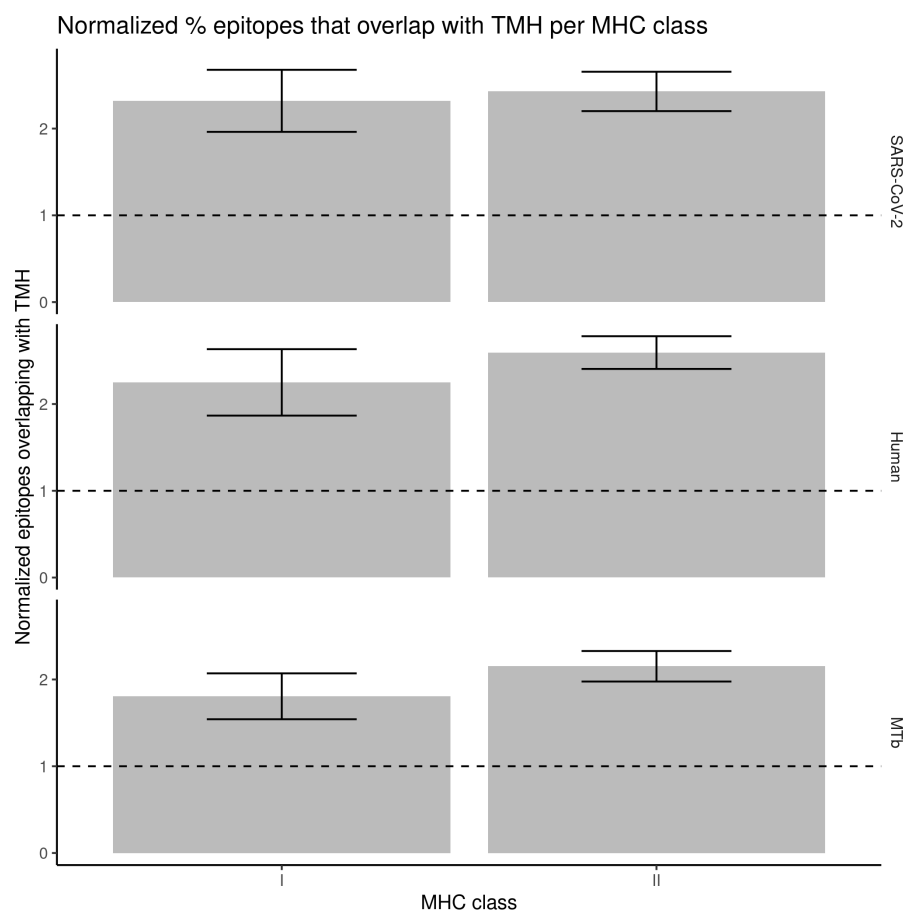


Figure S8: Normalized proportion of MHC-I and MHC-II epitopes overlapping with TMHs, for the different MHC classes and proteomes. Error bars denote the standard error.



856 **A.16 Evolutionary conservation**

857 Figure S9 shows the distribution of the number of SNPs per gene name, at the  
 858 date we started the experiment, at December 14th 2020.

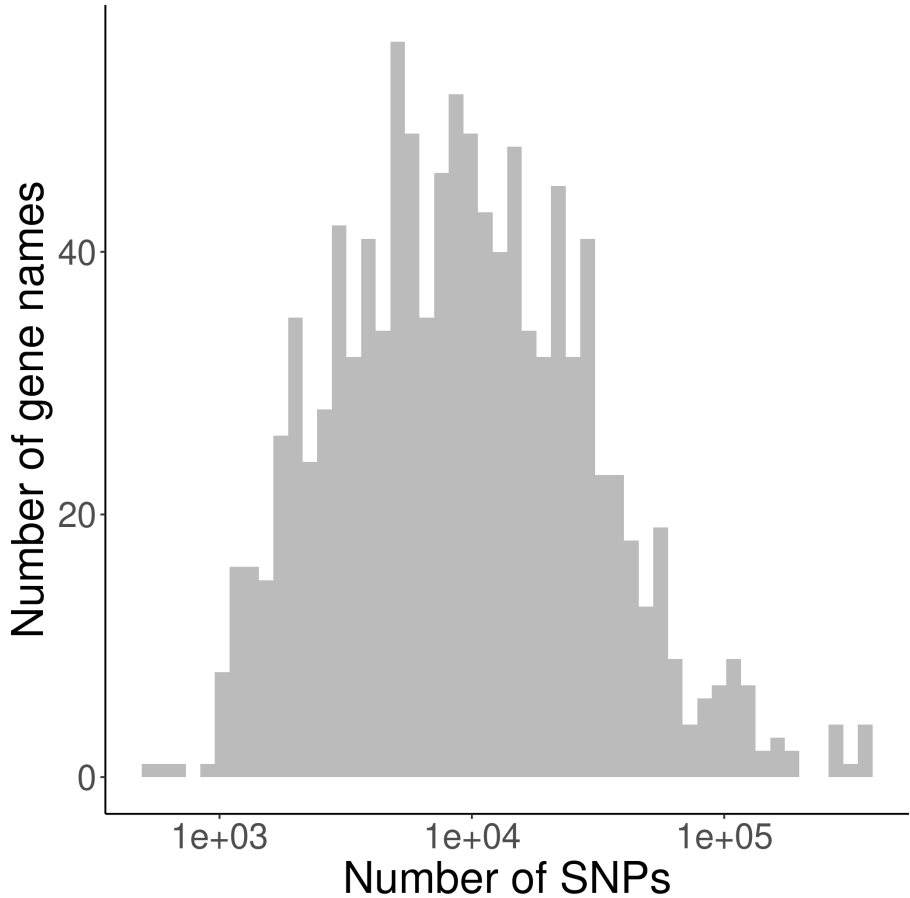


Figure S9: Distribution of the number of SNPs per gene name in the NCBI database.

859 To verify if SNPs were sampled uniformly over proteins, we show the dis-  
 860 tribution of the relative position in figure S1. We find no clear evidence of a  
 861 bias.

862 Supplementary Table S11 shows the statistics for all SNPs, where supple-

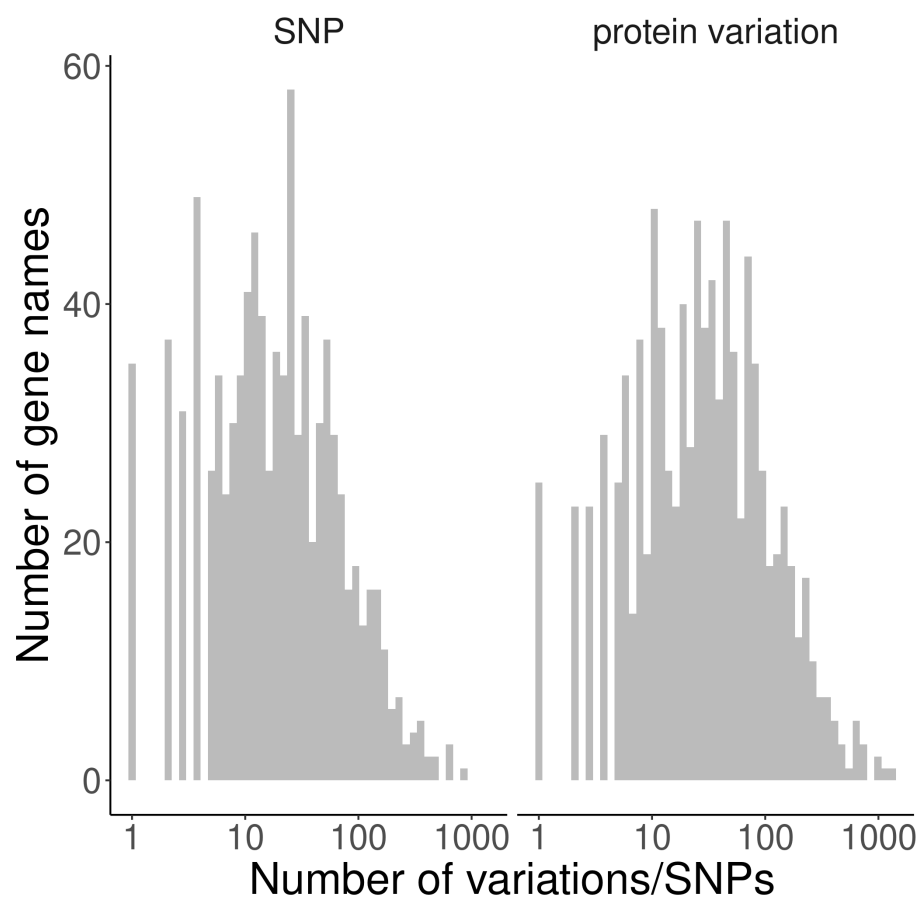


Figure S10: Distribution of the number of protein variations and SNPs per gene name processed.

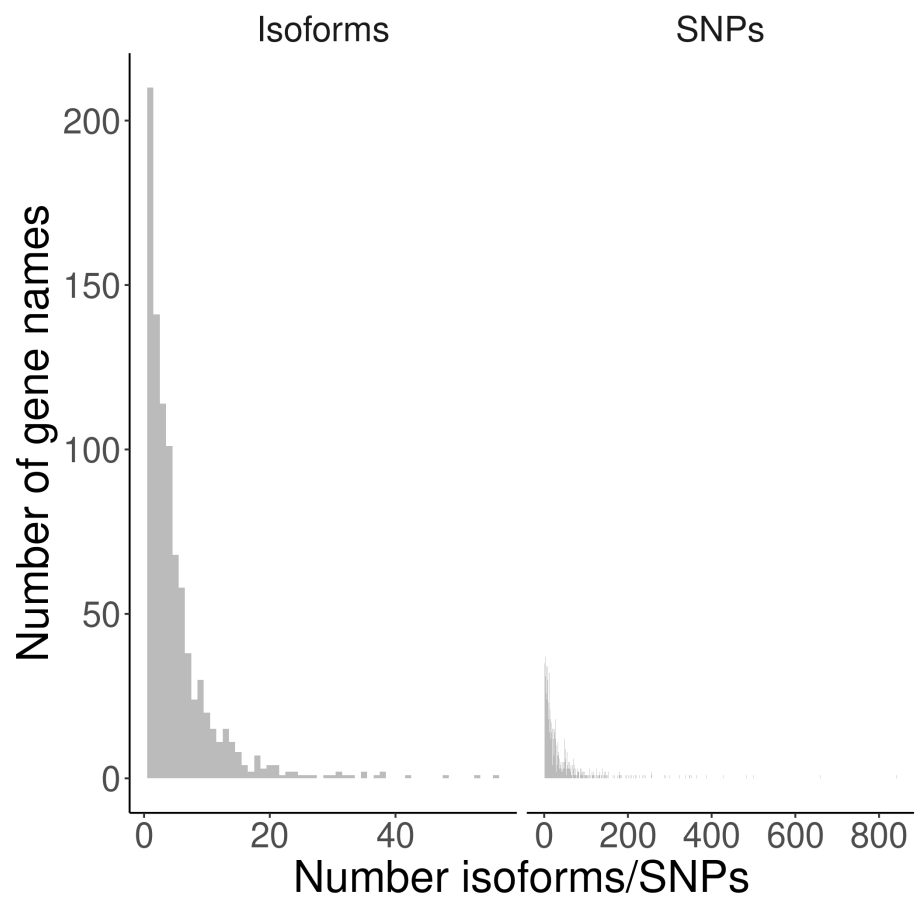


Figure S11: Histogram of the number of proteins found per gene name. Most often, a gene name is associated with one proteins.

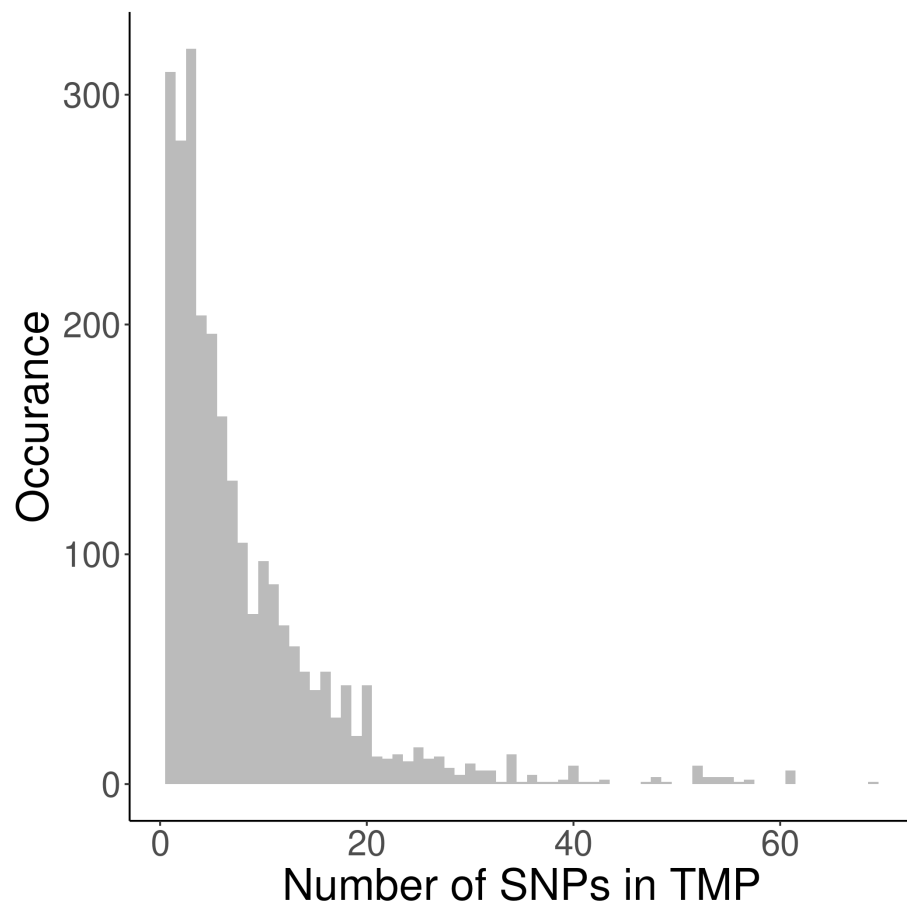


Figure S12: Histogram of the number of SNPs per trans-membrane protein.  
Dashed vertical line: average number of SNPs per TMP

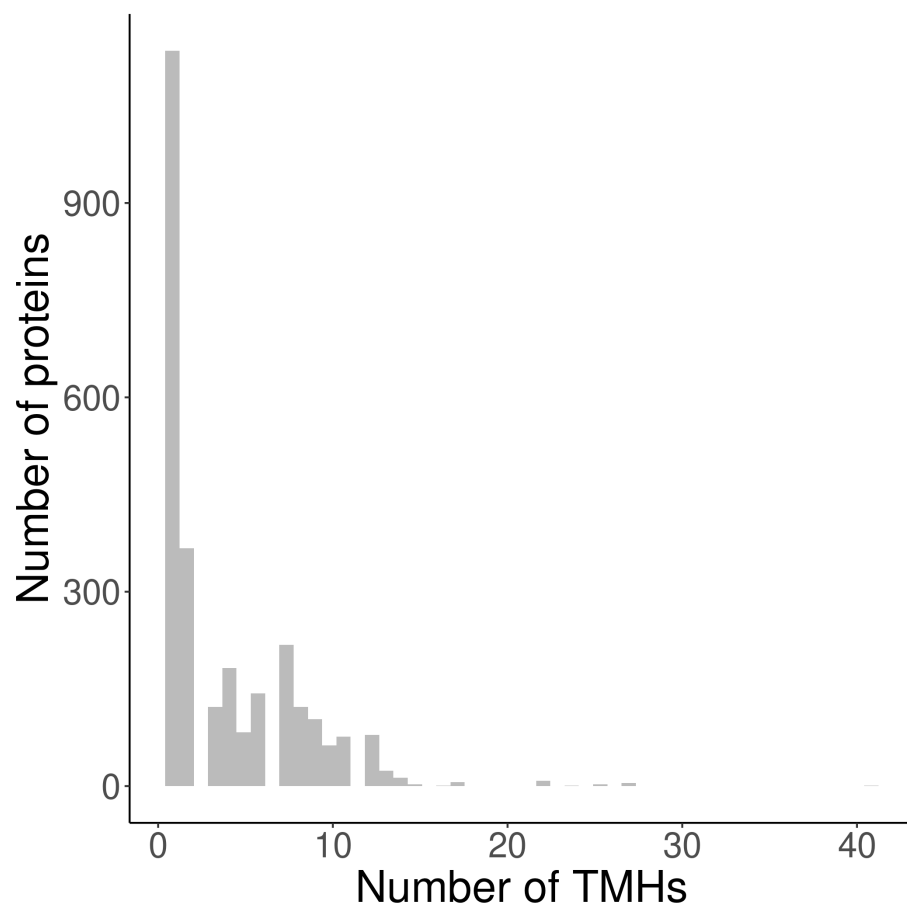


Figure S13: Histogram of the number of TMHs predicted per protein, for the trans-membrane proteins used.

Table S11: Statistics for all TMPs.  $p$  =  $p$  value.  $n$  = number of SNPs.  $n_{\text{success}}$  = number of SNPs found in TMHs (dashed blue line).  $E(n_{\text{success}})$  = expected number of SNPs to be found in TMHs (dashed line).

parameter	value
$p$	6.820823e-11
$n$	21208
$n_{\text{success}}$	3803
$E(n_{\text{success}})$	4140.56

Table S12: Statistics for the single-spanners.  $p$  =  $p$  value.  $n$  = number of SNPs in single-spanners.  $n_{\text{success}}$  = number of SNPs found in TMHs of single-spanners (dashed blue line).  $E(n_{\text{success}})$  = expected number of SNPs to be found in TMHs of single-spanners (dashed line).

parameter	value
$p$	0.3189532
$n$	8186
$n_{\text{success}}$	452
$E(n_{\text{success}})$	462.1535

mentary Tables S12 and S13 show the statistics for only single-spanners and multi-spanners respectively.

Table S13: Statistics for the multi-spanners.  $p$  =  $p$  value.  $n$  = number of SNPs in multi-spanners.  $n_{\text{success}}$  = number of SNPs found in TMHs of multi-spanners (dashed blue line).  $E(n_{\text{success}})$  = expected number of SNPs to be found in TMHs of multi-spanners (dashed line).

parameter	value
$p$	8.315841e-12
$n$	13022
$n_{\text{success}}$	3351
$E(n_{\text{success}})$	3678.406

## 865 **A.17 Presentation of TMH-derived epitopes when two amino** 866 **acids overlap**

867 In our experiment, we define a TMH-derived epitope as a peptide that overlaps  
868 with a TMH for at least one amino acid. One could argue that we should  
869 use a higher number of overlapping amino acids, so to make the epitopes more  
870 'transmembrane helix-ey'. We chose not too, for two reason: (1) epitopes that  
871 overlap with a TMH for 1 AA already, cannot be processed by the proteasome  
872 in a known and conventional way (2) whatever number of overlapping amino  
873 acids we use, we expect the pattern to be the same. However, using only 1 AA  
874 gives the most TMH-derived epitopes and hence the highest statistical power.

875 To prove this point, we did exactly the same analysis as shown in Figure  
876 1A, yet with defining a TMH-derived epitope as an epitope that overlaps with  
877 a TMH for at least 2 AAs, as shown in Figure S14. As these two figures look  
878 identical, we also added the counts as numbers, with Table S14 showing the  
879 same data as S5, except the former uses 2 AAs overlap. Likewise, Table S15  
880 showing the same data as S7, except the former uses 2 AAs overlap.

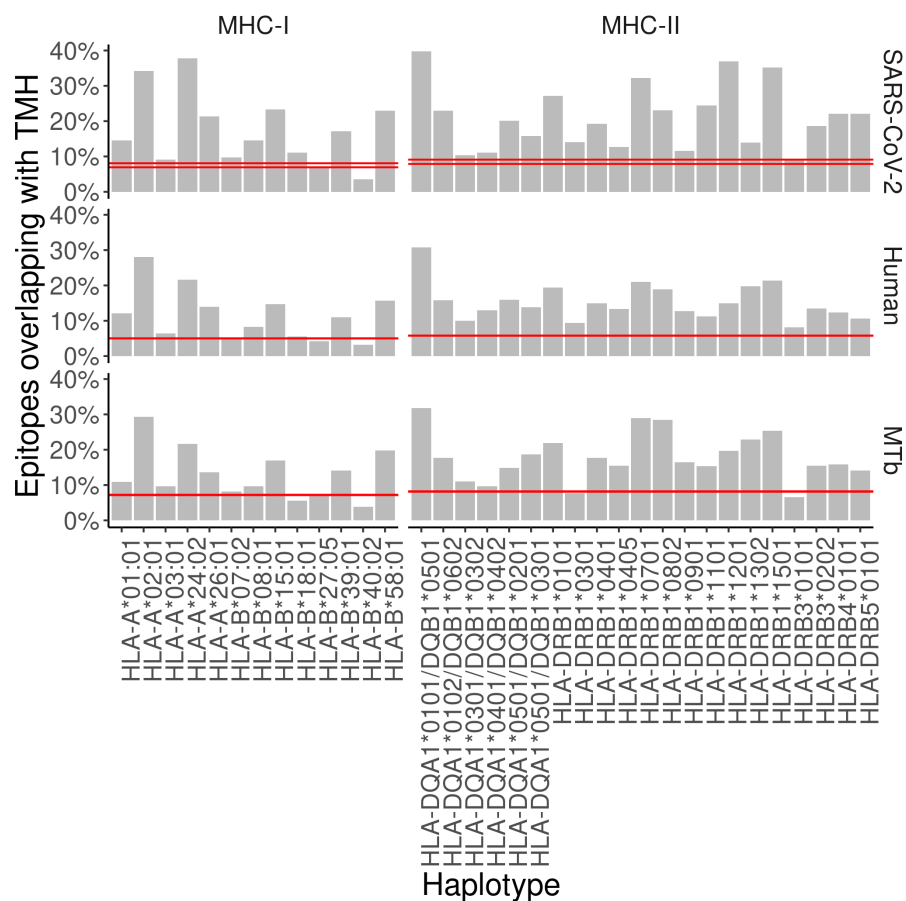


Figure S14: The percentage of epitopes for MHC-I and -II haplotypes that are predicted to overlap with TMHs (for at least two amino acids) for the proteomes of SARS-CoV-2 (top row), human (middle row) and *M. tuberculosis* (bottom row). The pair of dashed lines in each plot indicate the lower and upper bound of the 99% confidence interval. See supplementary Tables S14 and S15 for the exact TMH and epitope counts.



Table S14: Percentage of MHC-II 14-mers overlapping with TMH. Values in brackets show the number of binders that have at least two residues overlapping with a TMH (first value) as well as the number of binders (second value). percentage used: 2

haplotype	covid	human	myco
HLA-DQA1*0101/DQB1*0501	39.711 (110/277)	30.813 (68855/223464)	31.777 (8090/25459)
HLA-DQA1*0102/DQB1*0602	22.910 (74/323)	15.858 (35070/221147)	17.713 (4547/25671)
HLA-DQA1*0301/DQB1*0302	10.381 (30/289)	9.996 (22217/222248)	10.960 (2795/25502)
HLA-DQA1*0401/DQB1*0402	11.111 (32/288)	12.915 (28829/223219)	9.670 (2468/25522)
HLA-DQA1*0501/DQB1*0201	20.072 (56/279)	15.969 (35582/222820)	14.830 (3780/25489)
HLA-DQA1*0501/DQB1*0301	15.808 (46/291)	13.890 (30570/220089)	18.682 (4804/25715)
HLA-DRB1*0101	27.119 (80/295)	19.401 (43139/222349)	21.944 (5603/25533)
HLA-DRB1*0301	13.993 (41/293)	9.415 (20972/222752)	7.638 (1944/25451)
HLA-DRB1*0401	19.231 (55/286)	14.925 (33122/221930)	17.652 (4523/25623)
HLA-DRB1*0405	12.635 (35/277)	13.298 (29523/222012)	15.469 (3942/25484)
HLA-DRB1*0701	32.192 (94/292)	21.057 (46845/222465)	28.884 (7364/25495)
HLA-DRB1*0802	23.132 (65/281)	18.909 (41907/221623)	28.496 (7279/25544)
HLA-DRB1*0901	11.565 (34/294)	12.730 (28199/221520)	16.505 (4226/25605)
HLA-DRB1*1101	24.409 (62/254)	11.282 (25151/222928)	15.357 (3911/25467)
HLA-DRB1*1201	36.897 (107/290)	14.985 (33487/223464)	19.633 (5000/25467)
HLA-DRB1*1302	13.962 (37/265)	19.774 (44027/222646)	22.903 (5874/25647)
HLA-DRB1*1501	35.206 (94/267)	21.341 (47568/222893)	25.415 (6463/25430)
HLA-DRB3*0101	9.158 (25/273)	8.145 (18105/222274)	6.556 (1673/25517)
HLA-DRB3*0202	18.657 (50/268)	13.445 (29830/221859)	15.457 (3960/25620)
HLA-DRB4*0101	22.145 (64/289)	12.341 (27467/222568)	15.856 (4038/25467)
HLA-DRB5*0101	22.028 (63/286)	10.677 (23753/222464)	14.138 (3602/25478)

Table S15: Percentage of MHC-I 9-mers overlapping with TMH. Values in brackets show the number of binders that have at least two residues overlapping with a TMH (first value) as well as the number of binders (second value). percentage used: 2

haplotype	covid	human	myco
HLA-A*01:01	14.539 (41/282)	12.092 (27232/225209)	10.912 (2815/25797)
HLA-A*02:01	34.155 (97/284)	28.037 (63085/225003)	29.360 (7546/25702)
HLA-A*03:01	9.122 (27/296)	6.388 (14361/224796)	9.673 (2488/25721)
HLA-A*24:02	37.809 (107/283)	21.677 (48913/225648)	21.643 (5571/25741)
HLA-A*26:01	21.405 (64/299)	13.905 (31370/225598)	13.632 (3516/25793)
HLA-B*07:02	9.712 (27/278)	4.880 (10854/222429)	8.184 (2107/25744)
HLA-B*08:01	14.539 (41/282)	8.218 (18376/223616)	9.662 (2480/25667)
HLA-B*15:01	23.311 (69/296)	14.686 (33269/226542)	16.961 (4382/25835)
HLA-B*18:01	11.034 (32/290)	5.603 (12537/223745)	5.560 (1433/25773)
HLA-B*27:05	6.818 (18/264)	4.171 (9350/224178)	7.054 (1812/25688)
HLA-B*39:01	17.091 (47/275)	10.983 (24538/223419)	14.159 (3652/25793)
HLA-B*40:02	3.534 (10/283)	3.251 (7264/223408)	3.852 (991/25729)
HLA-B*58:01	22.939 (64/279)	15.627 (35022/224119)	19.793 (5095/25742)