Peptide presentation by HLA-DQ molecules is associated with the development of immune tolerance

Máté Manczinger [Corresp., 1, 2, 3] , Lajos Kemény [1, 2]

[1] Department of Dermatology and Allergology, University of Szeged, Szeged, Hungary

[2] MTA-SZTE Dermatological Research Group, University of Szeged, Szeged, Hungary

[3] Hungarian Academy of Sciences, Synthetic and Systems Biology Unit, Institute of Biochemistry, Biological Research Centre, Szeged, Hungary

Corresponding Author: Máté Manczinger
Email address: manczinger.mate@med.u-szeged.hu

HLA class II proteins are important elements of human adaptive immune recognition and are associated with numerous infectious and immune-mediated diseases. These highly variable molecules can be classified into DP, DQ and DR groups. It has been proposed that in contrast with DP and DR, epitope binding by DQ variants rather results in immune tolerance. However, the pieces of evidence are limited and controversial. We found that DQ molecules bind more human epitopes than DR. Pathogen-associated epitopes bound by DQ molecules are more similar to human proteins, than the ones bound by DR. Accordingly, DQ molecules bind epitopes of significantly different pathogen species. Moreover, the binding of autoimmunity-associated epitopes by DQ confers protection from autoimmune diseases. Our results suggest a special role of HLA-DQ in immune homeostasis and help to better understand the association of HLA molecules with infectious and autoimmune diseases.

1 **Peptide presentation by HLA-DQ molecules is associated with the development**

2 **of immune tolerance**

3 Máté Manczinger[1, 2, 3*], Lajos Kemény[1,2]

4 Affiliations:

5 [1] Department of Dermatology and Allergology, University of Szeged, Szeged, Hungary
6 [2] MTA-SZTE Dermatological Research Group, University of Szeged, Szeged, Hungary
7 [3] Synthetic and Systems Biology Unit, Institute of Biochemistry, Biological Research Centre,
8 Hungarian Academy of Sciences, Szeged, Hungary

9 Corresponding Author:
10 Máté Manczinger[1, 2, 3]

11 Email address: manczinger.mate@med.u-szeged.hu

12   **ABSTRACT**

13   HLA class II proteins are important elements of human adaptive immune recognition and are
14   associated with numerous infectious and immune-mediated diseases. These highly variable
15   molecules can be classified into DP, DQ and DR groups. It has been proposed that in contrast with
16   DP and DR, epitope binding by DQ variants rather results in immune tolerance. However, the
17   pieces of evidence are limited and controversial. We found that DQ molecules bind more human
18   epitopes than DR. Pathogen-associated epitopes bound by DQ molecules are more similar to
19   human proteins, than the ones bound by DR. Accordingly, DQ molecules bind epitopes of
20   significantly different pathogen species. Moreover, the binding of autoimmunity-associated
21   epitopes by DQ confers protection from autoimmune diseases. Our results suggest a special role
22   of HLA-DQ in immune homeostasis and help to better understand the association of HLA
23   molecules with infectious and autoimmune diseases.

INTRODUCTION

HLA molecules have an essential role in adaptive immune recognition (Trowsdale 2011). HLA class II molecules reside on antigen presenting cells and bind protein fragments of endogenous and exogenous peptides (Trowsdale 2011). The HLA-peptide complex then can be recognized by T cell receptors and induce enhanced immune response or tolerance (Robinson & Delvig 2002). The equilibrium between immune-mediated elimination and tolerance is crucial for a healthy immune homeostasis (Murphy et al. 2012).

HLA class II molecules can be classified into DP, DQ and DR groups (Trowsdale 2011). All molecules are made up by an alpha and a beta chain (Jones et al. 2006). Both chains of DP and DQ as well as the beta chain of DR are highly variable (Murphy et al. 2012). HLA class II molecules are associated with numerous diseases like autoimmunity, allergy and different kinds of infections (Karnes et al. 2017). To note, most diseases are associated with DQ and DR, while DP has a lower impact (Karnes et al. 2017). Consequently, our study focuses on the former two molecules.

Previous studies reported important specialities in the localization, expression and function of DQ molecules. HLA-DQ is more abundant in the thymus than in the periphery (Douek & Altmann 2000). Additionally, while DR is expressed in both the thymic cortex and medulla, DQ dominantly prevail in the cortex (Ishikura et al. 1987) suggesting a special - rather suppressive - role of the molecule in immune homeostasis (Altmann et al. 1991). This expression pattern is the result of many transcriptional and post-transcriptional regulatory mechanisms (Kobr et al. 1990; Miwa et al. 1987). The suppressive behavior of DQ is suggested by several findings, some of which are related to the association of HLA-DQ alleles with autoimmunity. First, risk alleles for type 1 diabetes mellitus show a recessive behavior (Todd 1990). It suggests that in case of heterozygosity, the other allele - being not associated with the disease - is able to induce tolerance in the thymus. Second, it has been shown that intrinsic stability of DQ molecules mediates susceptibility to autoimmune disorders (Miyadera et al. 2015). Unstable complexes cannot present self-peptides and, thus, central tolerance is not induced, which results in autoimmunity. Some associations between HLA-DQ and infectious diseases also suggest its dominant role in tolerance induction: it has been shown that HLA-DQ mediates non-responsiveness to *Schistosoma japonicum*, *Mycobacterium leprae* and BCG antigen through antigen-specific immune suppression (Hirayama et al. 1987; Ottenhoff et al. 1990; Salgame et al. 1991). However, the predominantly suppressive role of DQ in immune homeostasis remained controversial as DQ-mediated proliferative responses are also reported for both autoimmune and infectious diseases (Glanville et al. 2017; van Lummel et al. 2014).

Based on previous findings, we hypothesized, that HLA-DQ has an essential role in the induction of tolerance mechanisms. We carried out systematic analyses of experimental and computationally

60   predicted data to test four relevant predictions of the hypothesis: i) DQ molecules bind more human
61   epitopes, ii) DQ-bound epitopes of pathogens are more similar to human proteins, than DR-bound
62   ones, iii) DQ and DR molecules recognize epitopes of different pathogen species and iv) binding
63   of autoantigens by HLA-DQ molecules confer protection from autoimmune diseases. Our findings
64   highly supported the suppressive behavior of HLA-DQ molecules. Additionally, our results could
65   help to better understand the adaptive immune recognition of pathogens and the development of
66   autoimmune diseases.

67   **METHODS**

68   *Determining epitope sequence similarity to human proteins*

69   To determine epitope similarity to human proteins, in vitro HLA binding and T cell assay data
70   were downloaded from the Immune Epitope Database (IEDB) (Vita et al. 2015). Positive assay
71   results indicating the binding of 15 amino acids long epitopes by HLA-DQ and/or HLA-DR
72   molecules were collected. Results for human epitopes were discarded. The human reference
73   proteome was downloaded from Uniprot (The UniProt 2017) and epitope sequences found in it
74   were also discarded. Next, highly similar epitope sequences were excluded using an iterative
75   method. First, a protein distance matrix containing k-tuple distance values between all possible
76   epitope pairs was generated with Clustal Omega (Sievers et al. 2011; Yang & Zhang 2008). In
77   each iteration, epitope pair (or pairs) with the smallest distance value were identified. For each
78   member of the epitope pair (or pairs), the mean distance value to all other peptides was calculated
79   and the epitope with the smallest value was excluded. The iterations were repeated until only larger
80   than 0.5 k-tuple distance values remained in the matrix. This value corresponds to ~50% difference
81   between the two sequences. This filtering process was carried out for HLA-DQ and HLA-DR
82   epitope set separately resulting in 1476 and 4077 epitopes, respectively.

83   Each epitope sequence was decomposed to 5 amino acid long peptides (5-mers). For each 5-mer,
84   the number of times it prevails in the human reference proteome was determined. For both DQ-
85   and DR-associated epitopes, the proportion of 5-mers found for a given time in the human
86   proteome were calculated. If one 5-mer could be detected more times in epitope sequences, all
87   occurrences were taken into account. Similarly to a recent paper (Trost et al. 2012), we defined
88   rare 5-mers in three different ways: 5-mers occurring 0 times, 5-mers occurring two or fewer times
89   and 5-mers occurring five or less times in the human proteome. The epitope set containing
90   significantly less rare 5-mers was considered to be more similar to human proteins. The level of
91   significance was calculated with a randomization test. In each iteration, peptides of the original
92   epitope set were randomly assigned to DR and DQ and the 5-mer analysis was carried out on these
93   sequences. The epitope randomization and 5-mer analysis process was repeated for ten thousand
94   times. P value was defined as the probability of having larger or equal difference between the

95   proportion of rare or common 5-mers in DQ and DR-associated epitopes by chance than what we
96   found (i.e. the number of such cases divided by the total number of iterations).

97   *Determining the species specificity of DQ and DR*

98   HLA-II epitope sequences of pathogen species were downloaded from IEDB (Vita et al. 2015).
99   Obligate intracellular pathogens were excluded from the analysis and highly similar sequences
100 were discarded as described previously. Species with at least 25 epitopes available were selected
101 for further analysis. Reference proteome of all species were downloaded from the Uniprot database
102 (The UniProt 2017). Epitopes found in only one proteome (i. e. species specific epitopes) were
103 kept for further analysis. The previous filtering processes resulted in 1247 epitope sequences of 11
104 pathogens (Supplementary Table 1). Common HLA-DRB1, DQA1 and DQB1 alleles were
105 collected from the Common and Well Documented (CWD) Alleles Catalog (Mack et al. 2013).
106 All HLA-DQA1-DQB1 allele combinations were generated and forbidden allele combinations
107 were excluded (Raymond et al. 2005). The binding of each epitope by each common DRB1 allele
108 or DQA1-DQB1 allele pair was predicted with the NetMHCIIpan-3.1 computer algorithm
109 (Andreatta et al. 2015). We used the 10% rank percentile measure to define binding as suggested
110 by a recent NetMHCIIpan server update. The fraction of epitopes bound by each allele was
111 calculated for each pathogen. A matrix was created containing these values and hierarchical
112 clustering was carried out to find similar alleles based on their recognition of species. For each
113 pathogen, allele-specific recognition values were scaled and centered. Ward clustering algorithm
114 was used with the implementation of Ward's clustering criterion (Murtagh & Legendre 2014).
115 Clustering and visualization were carried out with pheatmap R library (Kolde 2012).

116 To confirm in silico results with in vitro data, we used scoring matrices. Epitopes of each species
117 were examined separately. Sequences longer than 15 amino acids were decomposed to 15 amino
118 acid long sequences and score was calculated for each resultant peptide. Epitope sets created for
119 the 5-mer analysis were used to generate scoring matrices. Before generating matrices, epitopes of
120 the examined species were excluded from the epitope sets. The prevalence of each amino acid at
121 each of the 15 positions of the epitope sequences was determined separately for the DQ and the
122 DR-associated epitopes. This resulted in two different scoring matrices, which were used to assess
123 the probability of binding the examined epitope by HLA-DQ and DR molecules. For each epitope
124 of the examined species, we calculated two scores reflecting its binding by HLA-DQ and HLA-
125 DR. The scores were determined by summing the values in the scoring matrices, which correspond
126 to the amino given amino acids at the 15 positions of the examined epitope. For example, if an
127 alanine was the first amino acid in the epitope sequence, we took the prevalence value for alanine
128 at the first position from the scoring matrix. We repeated this process for all positions and summed
129 the 15 values. Binding scores of human epitopes to HLA-DQ and HLA-DR were compared with
130 Wilcoxon rank sum test. P values were adjusted using the Benjamini-Hochberg procedure
131 (Benjamini & Hochberg 1995).

132 *Determining the relationship between auto-epitope binding and susceptibility to autoimmune*
133 *diseases*

134 Associations between HLA alleles (or certain amino acids in allele sequences) and autoimmune
135 diseases were collected from the PheWAS catalog (Karnes et al. 2017). These data were generated
136 using HLA typing of a large population with detailed disease information. To our knowledge,
137 PheWAS catalog is the only comprehensive source of HLA-disease associations. Associations
138 with P value less than 10^{-5} were considered to be significant as previously suggested (Karnes et al.
139 2017). Data about the following autoimmune diseases were collected: type 1 diabetes, Graves'
140 disease, systemic lupus erythematosus, celiac disease, multiple sclerosis, primary biliary cirrhosis,
141 systemic sclerosis, rheumatoid arthritis, juvenile rheumatoid arthritis, dermatomyositis and
142 polymyalgia rheumatica. Only associations with exact disease terms were included in the analysis
143 and terms that are only related to the diseases (for example "Type 1 diabetes with ketoacidosis")
144 were excluded. Epitope sequences associated with each disease were collected from the IEDB
145 (Vita et al. 2015). Sequences were discarded, if less than two references supported their role in
146 disease development. After excluding diseases with lack of epitope sequence data, the following
147 ones remained for further analysis: type 1 diabetes, Graves' disease, celiac disease, multiple
148 sclerosis, primary biliary cirrhosis and rheumatoid arthritis. It is important to note that the catalog
149 contains associations between diseases and individual DQA1 or DQB1 alleles, but not allele pairs.
150 However, epitope binding of DQ molecules is determined by both alpha and beta chains (Murphy
151 et al. 2012). As a solution, we selected all common allele pairs from the set we already generated
152 (described previously), which contain the given disease-associated allele. For each allele pair, we
153 determined the fraction of disease-associated epitopes bound using the NetMHCIIpan algorithm
154 as described previously (Andreatta et al. 2015). The median of these values defined the level of
155 auto-epitope binding by the original disease-associated allele. Auto-epitope binding by DRB1
156 alleles were determined by calculating the fraction of disease-associated epitopes bound by the
157 given disease-associated allele.

158 To examine associations between amino acids and autoimmune diseases, amino acid sequences of
159 DQA1 and DQB1 alleles were downloaded from the IPD-IMGT/HLA database (Robinson et al.
160 2015). For each disease-associated amino acid, we selected common allele pairs containing the
161 given amino acid in the given position. For each allele pair, we determined the bound fraction of
162 auto-epitopes and calculated the median of these values to describe the level of auto-epitope
163 binding associated with the given amino acid. The difference between auto-epitope binding by
164 susceptibility, neutral and protective alleles was detected using Kruskal-Wallis test. Pairwise
165 comparison was carried out with Wilcoxon rank sum tests. P values were adjusted using the
166 Benjamini-Hochberg procedure (Benjamini & Hochberg 1995).

167 **RESULTS**

168   *HLA-DQ molecules bind more human epitopes than HLA-DR*

169   The hypothesis that DQ molecules are associated with the induction of tolerance predicts that they
170   bind more human epitopes than DR. To test this prediction, we collected results of all in vitro
171   binding assays for human epitopes from the Immune Epitope Database (IEDB) (Vita et al. 2015).
172   We selected 1079 epitope sequences, whose binding was tested to both DQ and DR alleles. We
173   calculated the number of positive and negative assay results for DQ and DR and found a higher
174   proportion of positive binding assays for DQ alleles (OR: 1.78, Fisher exact test P: $2*10^{-22}$). This
175   result suggests a higher chance for the binding of human epitopes by DQ than by DR. To exclude
176   the possibility that this is caused by a generally higher epitope binding capacity of DQ molecules,
177   we collected assay results also for all non-human epitopes and selected 2289 sequences, whose
178   binding was tested to both DQ and DR alleles. Reassuringly, we got the opposite result: the
179   proportion of positive assays was higher for DR molecules than for DQ (OR: 1.36, Fisher exact
180   test P = $1.8*10^{-42}$).

181   *Epitopes of pathogens bound by HLA-DQ molecules are more similar to human proteins, than the*
182   *ones bound by HLA-DR*

183   We found a higher chance for binding human epitopes by DQ molecules than by DR. This suggests
184   that epitopes of pathogens, which are bound by HLA-DQ might be more similar to human proteins
185   than the ones bound by DR. To determine similarity of DQ and DR-associated epitopes to human
186   proteins, we used an established method (Trost et al. 2012). Five amino acids long peptide
187   segments (5-mers) are reported to be units of immunological recognition and protein-protein
188   interactions (Lucchese et al. 2007). We downloaded positive in vitro MHC binding and T cell
189   assay results for DQ and DR molecules from the IEDB. (Vita et al. 2015). We selected 15 amino
190   acid long sequences, excluded results for human epitopes and discarded highly similar epitope
191   sequences (see Methods). The filtering process resulted in 1476 DQ and 4077 DR-associated
192   epitopes. We decomposed each epitope to 5-mers and determined the number of times each 5-mer
193   can be found in the human proteome. Then, for both DQ and DR-associated epitopes, we calculated
194   the percentage of 5-mers that can be found for certain times in the human proteome. The
195   percentage of rare 5-mers indicates the similarity of epitope set to the human proteome: the lower
196   number of rare 5-mers can be found in epitope sequences, the more similar these epitopes are to
197   human proteins. As expected, DQ-associated epitopes contained a significantly lower number of
198   rare 5-mers, than DR-associated epitopes (Figure 1A). Moreover, common 5-mers occurring 30 or
199   more times in the human proteome could be found more frequently in DQ-associated epitopes
200   (Figure 1B).

201   *HLA-DQ and HLA-DR molecules bind different pathogen species*

202   As DQ-associated epitopes show higher similarity to human proteins, DQ and DR might be
203   responsible for the recognition of different pathogen species. To test this, we downloaded HLA-II
204   epitopes of pathogen species from IEDB. We discarded peptide sequences of obligate intracellular
205   pathogens as they are predominantly presented in an MHC-I dependent manner (Hewitt 2003). We
206   also excluded highly similar sequences from analysis (see Methods). We kept microbes having at
207   least 25 documented epitopes in IEDB and predicted the binding of each epitope by 73 common
208   HLA-DRB1 alleles and 168 common HLA-DQA1-DQB1 allele pairs using NetMHCIIpan-3.1.
209   This is reported to be the most accurate prediction algorithm for MHC class II molecules
210   (Andreatta et al. 2017). For each microbe, we determined the ratio of epitopes bound by each
211   allele. Next, we used hierarchical clustering to identify similarities between the species-preference
212   of different alleles. DQ and DRB1 alleles clearly separated from each other making up two
213   different clusters (Figure 2). Additionally, DQ and DR alleles bound markedly different microbial
214   species. We aimed to confirm these results with in vitro binding data. To this end, binding
215   probability scores were calculated for each pathogen-associated epitope (see Methods for details).
216   Briefly, we determined the prevalence of each amino acid at the 15 amino acid positions for both
217   DQ and DR-associated epitopes. We used epitope sets of the 5-mer analysis for this purpose.
218   Amino acid prevalence values were then applied to calculate scores, which reflect the binding
219   probability of a given epitope by DQ and DR molecules. Reassuringly, majority of results held
220   when using empirical binding data (Figure 2, Supplementary Table 2).

221   *The binding of auto-epitopes by HLA-DQ molecules protects from autoimmune diseases*

222   The preferred binding of human epitopes by DQ and previous evidence suggest, that these
223   molecules might play an important role in the induction of tolerance to self-epitopes.
224   Consequently, the binding of well-known epitopes of autoantigens (i.e. auto-epitopes) by DQ
225   molecules might protect from, while lack of binding might confer susceptibility to autoimmune
226   diseases. A straightforward prediction of these assumptions is that protective HLA-DQ alleles bind
227   more, while risk HLA-DQ alleles bind less auto-epitopes than neutral alleles. To test these
228   predictions, we collected 32 associations between 19 alleles of the DQA1, DQB1 and DRB1 loci
229   and six autoimmune diseases from the PheWAS catalog (Karnes et al. 2017) (Supplementary table
230   3). For each disease, we collected disease-associated epitopes from IEDB (Vita et al. 2015).
231   Epitope-binding by DQ is determined by both alpha and beta chains. Consequently, we calculated
232   the characteristic level of auto-epitope binding by a given protective or risk allele by considering
233   all DQA1-DQB1 combinations, in which the allele is included (see Methods). We found that the
234   binding of disease-associated epitopes by DQ alleles negatively correlated with the allele-
235   associated risk for autoimmune diseases (Figure 3A). This relationship was independent of the
236   autoimmune disease type and loci of DQ (DQA1 or DQB1) (Table 1). We got the same results, if
237   we used the amino acid position dataset of PheWAS catalog instead of exact alleles (Figure 3B,
238   Table 1 and Supplementary table 3).

239   As expected, protective alleles bound a higher, while susceptibility alleles bound a lower portion
240   of disease-associated epitopes than neutral ones (Kruskal-Wallis P = 9*10^{-5}, Figure 4). No
241   significant difference was found between autoantigen-binding by protective, neutral and
242   susceptibility alleles of HLA-DR (Kruskal-Wallis P = 0.18).

243   **DISCUSSION**

244   The equilibrium between immune defense mechanisms and tolerance is crucial for the homeostasis
245   of immunity. HLA molecules have a fundamental role in the regulation of these processes as HLA-
246   associated epitope presentation is one of the initial steps in the afferent arm of immune response
247   (Robinson & Delvig 2002). The different HLA class II loci are results of gene duplication events,
248   and their evolution is relatively independent from each other (Satta et al. 1994; Sommer 2005;
249   Valdes et al. 1999). Several previous studies suggested a special - potentially suppressive - role of
250   DQ molecules, but the evidence is controversial and limited (Altmann et al. 1991; Hirayama et al.
251   1987; Miyadera et al. 2015; Ottenhoff et al. 1990; Salgame et al. 1991; Todd 1990).

252   We carried out a large-scale analysis of epitope binding by DQ and DR molecules. We found a
253   higher chance for binding human epitopes by HLA-DQ than HLA-DR. We utilized an established
254   5-mer based approach to compare pathogen-associated epitopes with human proteins and found
255   that DQ-bound epitopes are more similar to self-proteins than DR-bound ones (Figure 1). Brett et
256   al. - using the same 5-mer based approach - found that bacteria causing chronic infections are more
257   similar to the human proteome (i.e. they contain a lower number of rare 5-mers) (Trost et al. 2012).
258   This is in line with another study suggesting that rare 5-mers cause a more intense immune
259   response than common ones (Patel et al. 2012). We characterized the binding of pathogen-
260   associated epitopes by DQ and DR and found that DQ and DR molecules present peptides of rather
261   different microbial species (Figure 2). While DR molecules bind pathogens associated with acute
262   infections, DQ alleles also recognize epitopes of pathogens causing chronic infections (*M.
263   tuberculosis, M. leprae*) (Russell 2011; Yamamura et al. 1991) or evading immune system and
264   having a high relapse rate after therapy (*B. pseudomallei*) (Currie et al. 2000).

265   Finally, our results suggest that DQ molecules seem to have an essential role in the development
266   of central tolerance by presenting self-epitopes. This is in line with previous findings: i) non-stable
267   binding of peptides by HLA-DQ resulted in thymic escape of autoreactive T-cells (Dendrou et al.
268   2018) and ii) autoimmune dermatitis developed in transgenic mice expressing an HLA-DQ
269   ortholog I-A^{b} complex, which can bind only one epitope (Logunova et al. 2005). In these mice,
270   autoreactive cells could be maintained due to the lack of negative selection of self-peptides by
271   MHC-II dependent presentation. We examined six different autoimmune diseases and found that
272   the hazard ratio associated with a given HLA-DQ allele inversely correlated with the proportion
273   of bound disease-associated epitopes (Figure 3, Table 1). Additionally, neutral DQ alleles bound
274   less epitopes than protective and more epitopes than risk alleles (Figure 4). This result indicates,

275   that a certain level of auto-epitope binding by DQ molecules is needed for a healthy immune-
276   homeostasis. Alleles binding less auto-epitopes might allow thymic escape of self-reactive T cells
277   and make the individual susceptible to autoimmune diseases.

278   **CONCLUSIONS**

279   Previous pieces of evidence suggested a suppressive role of HLA-DQ molecules in immune
280   homeostasis, but this hypothesis remained controversial. We tested relevant predictions of the
281   hypothesis. We found that DQ molecules bind more human epitopes than DR. Accordingly,
282   pathogen-assocated epitopes bound by DQ are more similar to human proteins than the ones bound
283   by DR. DQ molecules bind mainly epitopes of pathogens associated with chronic or relapsing
284   infectious diseases. This indicates the importance of DQ-mediated tolerance induction for the
285   immune evasion of pathogens. Our results also suggest that DQ molecules might have a more
286   important role in inducing tolerance than in activating proliferative and destructive responses
287   during the development of autoimmune diseases. All of our findings suggest an essential role of
288   HLA-DQ molecules in tolerance formation and might help to better understand the role of HLA
289   molecules in the development of infectious and autoimmune diseases.

REFERENCES

290

291 Altmann DM, Sansom D, and Marsh SG. 1991. What is the basis for HLA-DQ associations with
292   autoimmune disease? *Immunol Today* 12:267-270. 10.1016/0167-5699(91)90124-C
293 Andreatta M, Karosiene E, Rasmussen M, Stryhn A, Buus S, and Nielsen M. 2015. Accurate pan-
294   specific prediction of peptide-MHC class II binding affinity with improved binding core
295   identification. *Immunogenetics* 67:641-650. 10.1007/s00251-015-0873-y
296 Andreatta M, Trolle T, Yan Z, Greenbaum JA, Peters B, and Nielsen M. 2017. An automated
297   benchmarking platform for MHC class II binding prediction methods. *Bioinformatics*.
298   10.1093/bioinformatics/btx820
299 Benjamini Y, and Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful
300   approach to multiple testing. *Journal of the royal statistical society Series B
301   (Methodological)*:289-300.
302 Currie BJ, Fisher DA, Anstey NM, and Jacups SP. 2000. Melioidosis: acute and chronic disease,
303   relapse and re-activation. *Trans R Soc Trop Med Hyg* 94:301-304.
304 Dendrou CA, Petersen J, Rossjohn J, and Fugger L. 2018. HLA variation and disease. *Nat Rev
305   Immunol*. 10.1038/nri.2017.143
306 Douek DC, and Altmann DM. 2000. T-cell apoptosis and differential human leucocyte antigen
307   class II expression in human thymus. *Immunology* 99:249-256.
308 Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, Ji X, Han A, Krams SM, Pettus C,
309   Haas N, Arlehamn CSL, Sette A, Boyd SD, Scriba TJ, Martinez OM, and Davis MM. 2017.
310   Identifying specificity groups in the T cell receptor repertoire. *Nature* 547:94-98.
311   10.1038/nature22976
312 Hewitt EW. 2003. The MHC class I antigen presentation pathway: strategies for viral immune
313   evasion. *Immunology* 110:163-169.
314 Hirayama K, Matsushita S, Kikuchi I, Iuchi M, Ohta N, and Sasazuki T. 1987. HLA-DQ is epistatic
315   to HLA-DR in controlling the immune response to schistosomal antigen in humans. *Nature*
316   327:426-430. 10.1038/327426a0
317 Ishikura H, Ishikawa N, and Aizawa M. 1987. Differential expression of HLA-class II antigens in
318   the human thymus. Relative paucity of HLA-DQ antigens in the thymic medulla.
319   *Transplantation* 44:314-317.
320 Jones EY, Fugger L, Strominger JL, and Siebold C. 2006. MHC class II proteins and disease: a
321   structural perspective. *Nat Rev Immunol* 6:271-282. 10.1038/nri1805
322 Karnes JH, Bastarache L, Shaffer CM, Gaudieri S, Xu Y, Glazer AM, Mosley JD, Zhao S,
323   Raychaudhuri S, Mallal S, Ye Z, Mayer JG, Brilliant MH, Hebbring SJ, Roden DM,
324   Phillips EJ, and Denny JC. 2017. Phenome-wide scanning identifies multiple diseases and
325   disease severity phenotypes associated with HLA variants. *Sci Transl Med* 9.
326   10.1126/scitranslmed.aai8708
327 Kobr M, Reith W, Herrero-Sanchez C, and Mach B. 1990. Two DNA-binding proteins
328   discriminate between the promoters of different members of the major histocompatibility
329   complex class II multigene family. *Mol Cell Biol* 10:965-971.
330 Kolde R. 2012. Pheatmap: pretty heatmaps. *R package version* 61.
331 Logunova NN, Viret C, Pobezinsky LA, Miller SA, Kazansky DB, Sundberg JP, and Chervonsky
332   AV. 2005. Restricted MHC-peptide repertoire predisposes to autoimmunity. *J Exp Med*
333   202:73-84. 10.1084/jem.20050198

334 Lucchese G, Stufano A, Trost B, Kusalik A, and Kanduc D. 2007. Peptidology: short amino acid
335   modules in cell biology and immunology. *Amino Acids* 33:703-707. 10.1007/s00726-006-
336   0458-z
337 Mack SJ, Cano P, Hollenbach JA, He J, Hurley CK, Middleton D, Moraes ME, Pereira SE,
338   Kempenich JH, Reed EF, Setterholm M, Smith AG, Tilanus MG, Torres M, Varney MD,
339   Voorter CE, Fischer GF, Fleischhauer K, Goodridge D, Klitz W, Little AM, Maiers M,
340   Marsh SG, Muller CR, Noreen H, Rozemuller EH, Sanchez-Mazas A, Senitzer D,
341   Trachtenberg E, and Fernandez-Vina M. 2013. Common and well-documented HLA
342   alleles: 2012 update to the CWD catalogue. *Tissue Antigens* 81:194-203.
343   10.1111/tan.12093
344 Miwa K, Doyle C, and Strominger JL. 1987. Sequence-specific interactions of nuclear factors with
345   conserved sequences of human class II major histocompatibility complex genes. *Proc Natl*
346   *Acad Sci U S A* 84:4939-4943.
347 Miyadera H, Ohashi J, Lernmark A, Kitamura T, and Tokunaga K. 2015. Cell-surface MHC
348   density profiling reveals instability of autoimmunity-associated HLA. *J Clin Invest*
349   125:275-291. 10.1172/JCI74961
350 Murphy K, Travers P, Walport M, and Janeway C. 2012. *Janeway's immunobiology*. New York:
351   Garland Science.
352 Murtagh F, and Legendre P. 2014. Ward's hierarchical agglomerative clustering method: which
353   algorithms implement Ward's criterion? *Journal of Classification* 31:274-295.
354 Ottenhoff TH, Walford C, Nishimura Y, Reddy NB, and Sasazuki T. 1990. HLA-DQ molecules
355   and the control of Mycobacterium leprae-specific T cell nonresponsiveness in lepromatous
356   leprosy patients. *Eur J Immunol* 20:2347-2350. 10.1002/eji.1830201027
357 Patel A, Dong JC, Trost B, Richardson JS, Tohme S, Babiuk S, Kusalik A, Kung SK, and Kobinger
358   GP. 2012. Pentamers not found in the universal proteome can enhance antigen specific
359   immune responses and adjuvant vaccines. *PLoS One* 7:e43802.
360   10.1371/journal.pone.0043802
361 Raymond CK, Kas A, Paddock M, Qiu R, Zhou Y, Subramanian S, Chang J, Palmieri A, Haugen
362   E, Kaul R, and Olson MV. 2005. Ancient haplotypes of the HLA Class II region. *Genome*
363   *Res* 15:1250-1257. 10.1101/gr.3554305
364 Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, and Marsh SG. 2015. The IPD and
365   IMGT/HLA database: allele variant databases. *Nucleic Acids Res* 43:D423-431.
366   10.1093/nar/gku1161
367 Robinson JH, and Delvig AA. 2002. Diversity in MHC class II antigen presentation. *Immunology*
368   105:252-262.
369 Russell DG. 2011. Mycobacterium tuberculosis and the intimate discourse of a chronic infection.
370   *Immunol Rev* 240:252-268. 10.1111/j.1600-065X.2010.00984.x
371 Salgame P, Convit J, and Bloom BR. 1991. Immunological suppression by human CD8+ T cells
372   is receptor dependent and HLA-DQ restricted. *Proc Natl Acad Sci U S A* 88:2598-2602.
373 Satta Y, O'HUigin C, Takahata N, and Klein J. 1994. Intensity of natural selection at the major
374   histocompatibility complex loci. *Proc Natl Acad Sci U S A* 91:7184-7188.
375 Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M,
376   Soding J, Thompson JD, and Higgins DG. 2011. Fast, scalable generation of high-quality
377   protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539.
378   10.1038/msb.2011.75

379   Sommer S. 2005. The importance of immune gene variability (MHC) in evolutionary ecology and
380   conservation. *Front Zool* 2:16. 10.1186/1742-9994-2-16
381   The UniProt C. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 45:D158-
382   D169. 10.1093/nar/gkw1099
383   Todd JA. 1990. Genetic control of autoimmunity in type 1 diabetes. *Immunol Today* 11:122-129.
384   Trost B, Pajon R, Jayaprakash T, and Kusalik A. 2012. Comparing the similarity of different
385   groups   of   bacteria   to   the   human   proteome.   *PLoS   One*   7:e34007.
386   10.1371/journal.pone.0034007
387   Trowsdale   J.   2011.   The   MHC,   disease   and   selection.   *Immunol   Lett*   137:1-8.
388   10.1016/j.imlet.2011.01.002
389   Valdes AM, McWeeney SK, Meyer D, Nelson MP, and Thomson G. 1999. Locus and population
390   specific evolution in HLA class II genes. *Ann Hum Genet* 63:27-43.
391   van Lummel M, Duinkerken G, van Veelen PA, de Ru A, Cordfunke R, Zaldumbide A, Gomez-
392   Tourino I, Arif S, Peakman M, Drijfhout JW, and Roep BO. 2014. Posttranslational
393   modification of HLA-DQ binding islet autoantigens in type 1 diabetes. *Diabetes* 63:237-
394   247. 10.2337/db12-1214
395   Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, Wheeler DK, Gabbard
396   JL, Hix D, Sette A, and Peters B. 2015. The immune epitope database (IEDB) 3.0. *Nucleic
397   Acids Res* 43:D405-412. 10.1093/nar/gku938
398   Yamamura M, Uyemura K, Deans RJ, Weinberg K, Rea TH, Bloom BR, and Modlin RL. 1991.
399   Defining protective responses to pathogens: cytokine profiles in leprosy lesions. *Science*
400   254:277-279.
401   Yang K, and Zhang L. 2008. Performance comparison between k-tuple distance and four model-
402   based distances in phylogenetic tree reconstruction. *Nucleic Acids Res* 36:e33.
403   10.1093/nar/gkn075
404

Figure 1

The percentage of rare and common 5-mers in DQ- and DR-associated epitopes.

The figures show the fraction of 5-mers found for certain times in the human proteome. DQ-associated epitopes contained (A) less rare 5-mers and (B) a higher number of common 5-mers than DR-associated sequences. 5-mer composition of 815 human epitopes (green) is also shown on the figures. Dashed lines represent different cutoffs used for defining (A) rare 5-mers and (B) common 5-mers. P values represent the probability of having the same or higher difference between the number of rare and common 5-mers in DQ and DR-associated epitopes by chance (see Methods). Note, that in case of common alleles, both horizontal and vertical axes were log-transformed for better visualization.

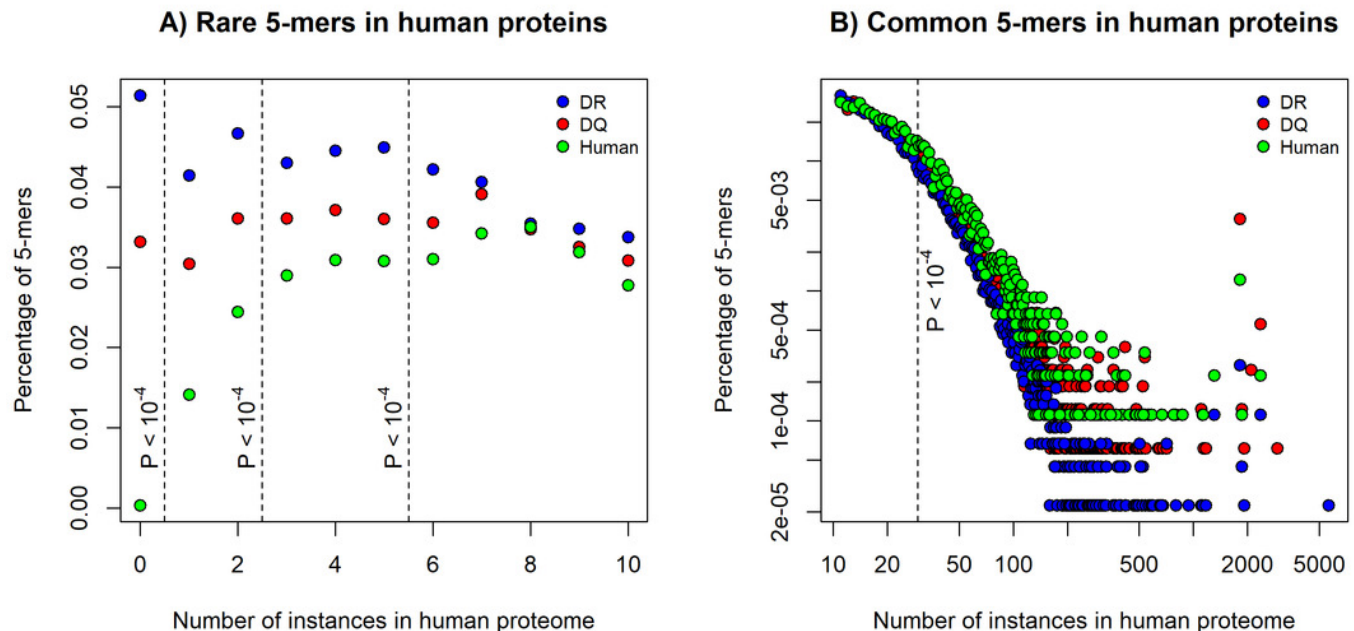



Figure 2

HLA-DQ and DR molecules bind epitopes of different microbes.

The heatmap shows epitope binding by different DQA1-DQB1 allele pairs and DRB1 alleles color coded. In case of each species, colors represent the portion of epitopes recognized by each allele or allele pair. Each row corresponds to a DRB1 allele or DQ allele pair. Rows are clustered using hierarchical clustering (see Methods). DRB1 and DQ molecules are clearly separated based on their species preference (marked with a horizontal line). Epitopes of species on the left are preferred by DR (marked with blue color in the table) and on the right are preferred by DQ molecules (marked with green color in the table). Values were centered and scaled before being clustered and visualized. The ratio between the mean proportion of epitopes bound by DQ allele pairs and DR alleles (DQ/DR) is shown in the table for each species. Note that although computational prediction indicated similar recognition of *M. tuberculosis* and *M. leprae* by DQ and DR, analysis of in vitro data showed significantly higher binding scores of these species for DQ (Supplementary Table 2). Consequently, they were classified into the DQ-associated group. P_{corr} represents FDR-corrected P values of Wilcoxon rank sum test. P_{corr} values lower than 0.1 were considered to be significant and highlighted with red color.

DQ/DR	0.16	0.22	0.34	0.78	0.8*	1.26*	1.81	2.06	2.12	2.75	3.26
P_{corr}	$1*10^{-30}$	$2*10^{-30}$	$4*10^{-26}$	0.02	0.002	$8*10^{-5}$	$3*10^{-5}$	$3*10^{-22}$	$2*10^{-6}$	$1*10^{-8}$	$1*10^{-25}$
P_{corr}(Scoring matrices)	$2*10^{-9}$	$3*10^{-7}$	$4*10^{-5}$	0.07	0.01	$2*10^{-42}$	0.91	0.56	$1*10^{-34}$	0.18	0.05

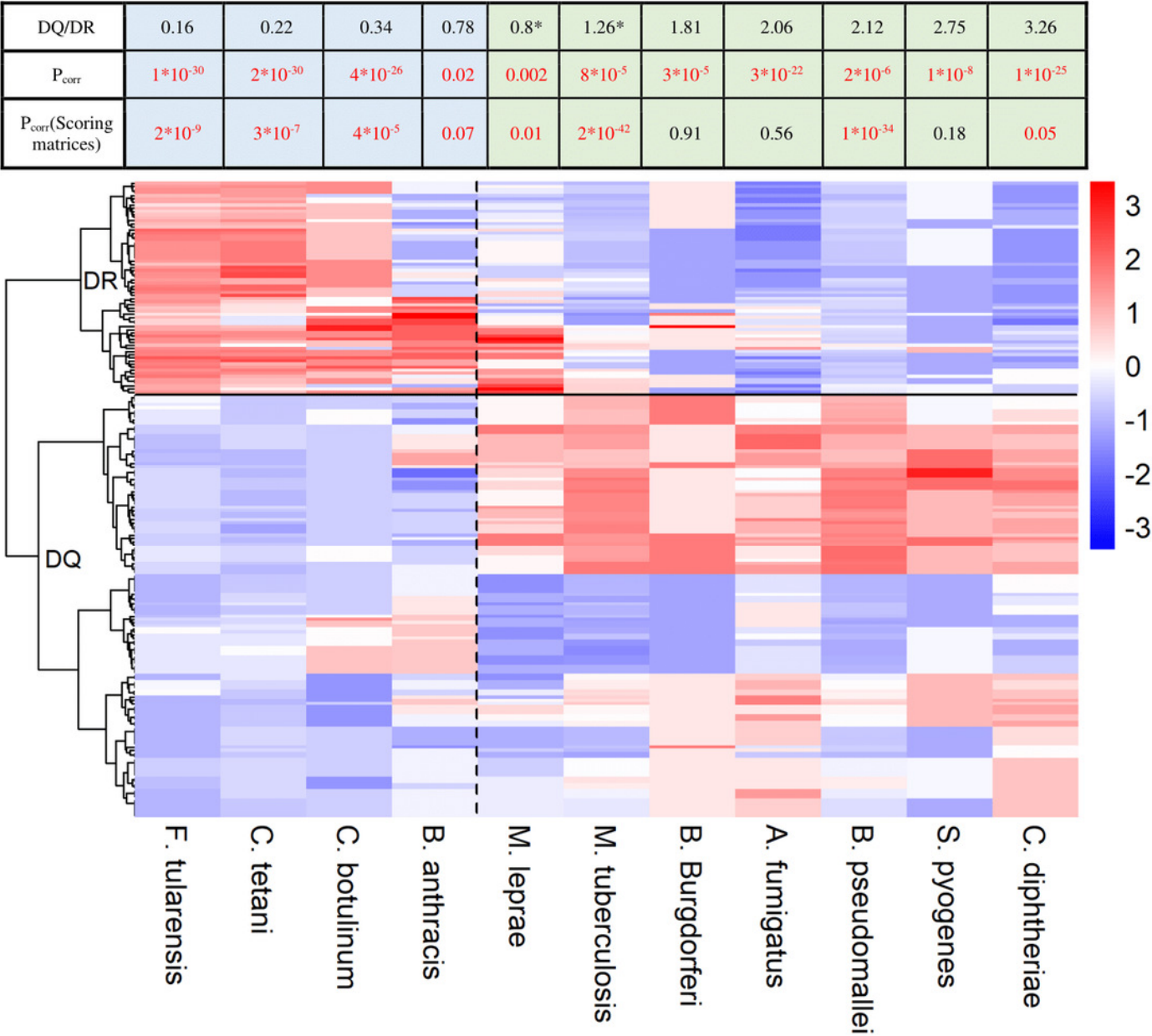



Figure 3

Binding of auto-epitopes by HLA-DQ is associated with protection from autoimmune diseases.

The binding of disease-specific auto-epitopes inversely correlated with the disease risk associated with (A) HLA-DQ alleles and (B) amino acids of HLA-DQ (Spearman's rho: -0.69 and -0.64, P = $4*10^{-4}$ and $2*10^{-8}$, respectively). Horizontal axes indicate the portion of disease-associated auto-epitopes bound by A) the given allele or B) amino acid (see Methods). Vertical axes show the risk for autoimmune disease associated with (A) the given allele or (B) amino acid. Red dashed lines represent linear regression lines. Note, that vertical axis is on a logarithmic scale.

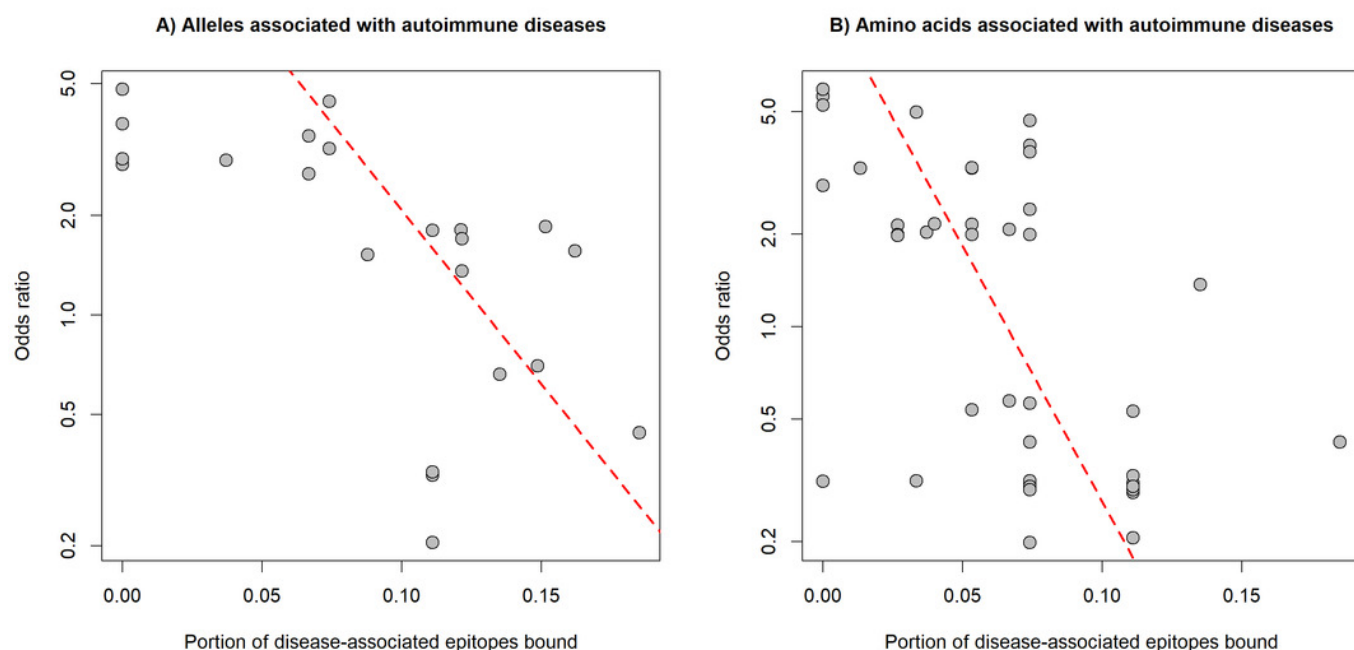
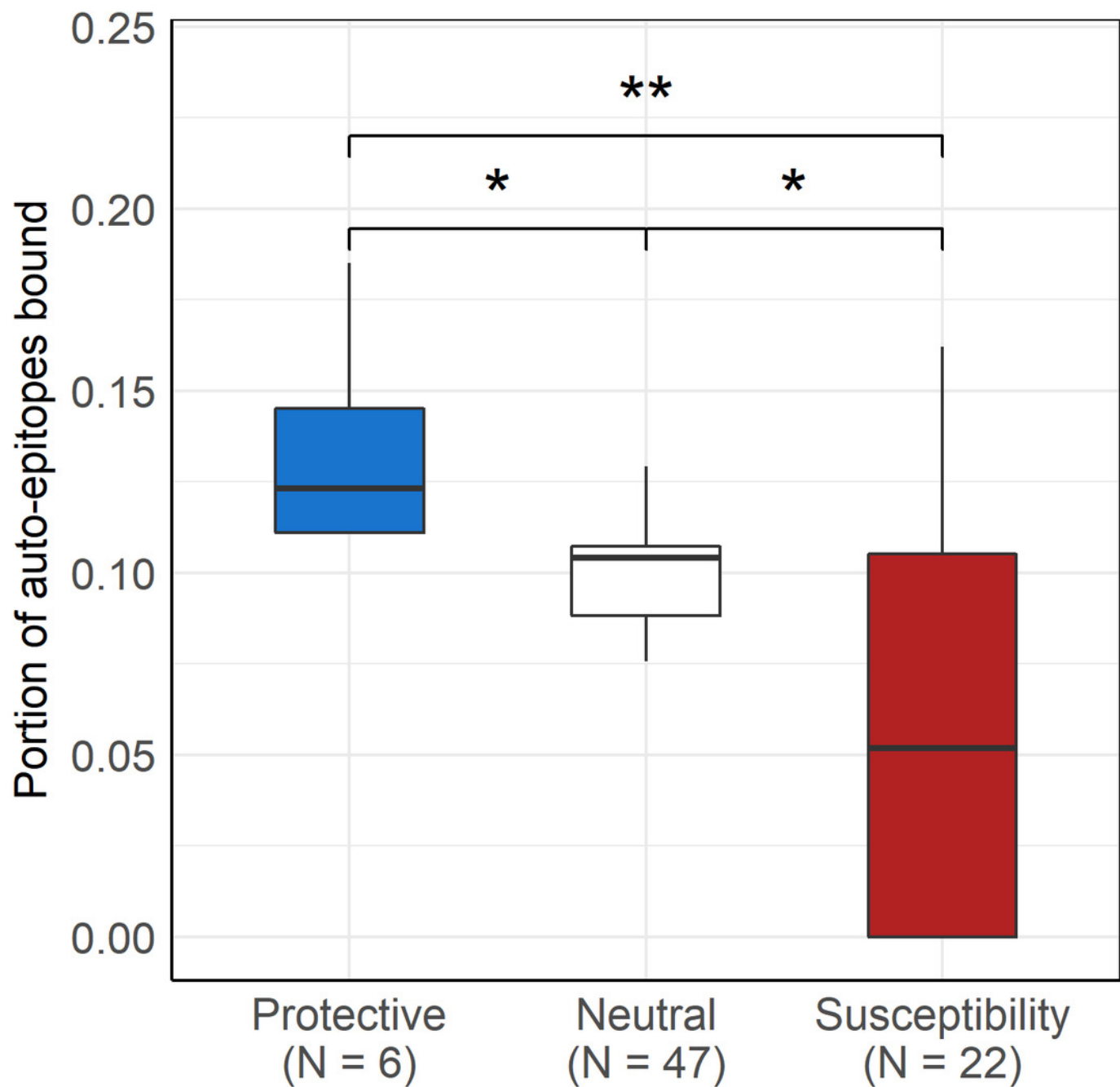



Figure 4

Protective DQ alleles bind more, while risk DQ alleles bind less autoimmune disease-associated epitopes than neutral ones.

Boxplot indicates median (horizontal line), the first and third quartile (bottom and top of boxes) and minimum and maximum values (vertical lines). The bound portion of auto-epitopes of the given disease that is associated with the given allele is shown for protective and susceptibility groups. The portion of all auto-epitopes bound by each allele is shown for the neutral allele group. * P_{corr} < 0.005 (Pairwise Wilcoxon test), ** P_{corr}<0.01 (Pairwise Wilcoxon test)

The relationship between auto-epitope binding by HLA-DQ molecules and protection from autoimmune diseases is independent of disease types and DQ chains.

To test, whether the relationship between binding auto-epitopes and risk for autoimmune diseases is caused by disease- or DQ chain-specific differences in the epitope-binding of alleles, we constructed multivariate models:

(1) $\log(OR) \sim Fr_{allele} + Disease + Chain$

(2) $\log(OR) \sim Fr_{amino\ acid} + Disease + Chain$

where *OR* is the odds ratio for developing the given disease; Fr_{allele} is the level of disease-associated auto-epitope binding by the DQ allele; $Fr_{amino\ acid}$ is the level of disease-associated auto-epitope binding by DQ alleles, which contain the given amino acid and; *Chain* (i.e. DQA1 or DQB1) and *Disease* are categorical variables. Fr_{allele} and $Fr_{amino\ acid}$ showed negative effect on *OR* after controlling for diseases and DQ chains. Significant relationship between predictor and response variables is marked with red color.

Allele associations				Amino acid associations			
Variable	Slope	Variance explained	P	Variable	Slope	Variance explained	P
Fr$_{allele}$	-	0.4	0.01	*Fr*$_{amino\ acid}$	-	0.39	0.008
Disease	NA	0.17	NA	*Disease*	NA	0.08	NA
Chain	+ (DQB1)	0.02	0.47	*Chain*	- (DQB1)	0.08	0.004

	Allele associations	Amino acid associations
R^2	0.38 (P = 0.046)	0.5 (P = 6*10^{-8})
N	22	61
BP test P	0.2	0.11

1   Variance expained: The proportion of variance in *log(OR)* explained by the given predictor variable. P: the probability of observing
2   relationship between the predictor and response variables by chance, R^2: Total variance in *log(OR)* explained by the model (P corresponds
3   to F test P value), N: number of associations between autoimmune diseases and alleles or amino acids, BP test P: the P value for Breusch-
4   Pagan test of heteroscedasticity. P values larger than 0.05 suggest lack of heteroscedasticity