Machine Learning Reveals a Non-Canonical Mode of Peptide Binding to MHC class II Molecules

Andreatta, Massimo; Jurtz, Vanessa Isabell; Kaever, Thomas; Sette, Alessandro; Peters, Bjoern; Nielsen, Morten

Published in:
Immunology

Link to article, DOI:
10.1111/imm.12763

Publication date:
2017

Document Version
Peer reviewed version

Link back to DTU Orbit

Citation (APA):
Andreatta, M., Jurtz, V. I., Kaever, T., Sette, A., Peters, B., & Nielsen, M. (2017). Machine Learning Reveals a Non-Canonical Mode of Peptide Binding to MHC class II Molecules. Immunology, 255-264. DOI: 10.1111/imm.12763

DR. MASSIMO ANDREATTA (Orcid ID : 0000-0002-8036-2647)

Article type : Original Article

Machine Learning Reveals a Non-Canonical Mode of Peptide Binding to MHC class II Molecules

Short title: Non-canonical peptide binding to MHC class II

Massimo Andreatta[1], Vanessa Isabell Jurtz[2], Thomas Kaever[3], Alessandro Sette[3], Bjoern Peters[3], Morten Nielsen[1,2]

[1] Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, CP1650 San Martín, Argentina

[2] Center for Biological Sequence Analysis, Department of Bio and Health Informatics, Technical University of Denmark, DK-2800 Lyngby, Denmark

[3] Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, CA92037 La Jolla, USA

Corresponding author: Massimo Andreatta, mandreatta@iibintech.com.ar - Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, Campus Miguelete, 25 de Mayo y Francia, CP1650 San Martín, Argentina

Keywords: machine learning, MHC class II, non-canonical binding, insertions, deletions

Summary

MHC class II molecules play a fundamental role in the cellular immune system: they load short peptide fragments derived from extracellular proteins and present them on the cell surface. It is currently thought that the peptide binds lying more or less flat in the MHC groove, with a fixed distance of nine amino acids between the first and last residue in contact with the MHCII. While confirming that the great majority of peptides bind to the MHC using this canonical mode, we report evidence for an alternative, less common mode of interaction. A fraction of observed ligands were shown to have an unconventional spacing of the anchor residues that directly interact with the MHC, which could only be accommodated to the canonical MHC motif either by imposing a more stretched out peptide backbone (a 8mer core) or by the peptide bulging out of the MHC groove (a 10mer core). We estimated that on average 2% of peptides bind with a core deletion, and 0.45% with a core insertion, but the frequency of such non-canonical cores was as high as 10% for certain MHCII molecules. A mutational analysis and experimental validation of a number of these anomalous ligands demonstrated that they could only fit to their MHC binding motif with a non-canonical binding core of length different from nine. This previously undescribed mode of peptide binding to MHCII molecules gives a more complete picture of peptide presentation by MHCII and allows us to model more accurately this event.

Introduction

The primary function of class II major histocompatibility complex molecules (MHCII) is to alert the immune system to the presence of a pathogen by binding and presenting short peptide fragments derived from exogenously derived proteins. Once outside the cell, T helper lymphocytes bearing receptors specific for the peptide-MHCII complex can recognize the peptide as non-self and help initiating an appropriate immune response (1–3). Structurally, MHCII consists of two non-covalently bound amino acid chains, the α and the β chain. The two domains α1 and β1 of these chains combine to form the peptide-binding groove. Because the MHCII groove is open at the extremities, the peptide ligand can freely extend outside both ends. The stretch of peptide residues directly interacting with the groove, the so-called binding core, is the main determinant of peptide binding to MHCII. Ample structural data show that the peptide binds lying flat in the groove, with an extended conformation and a fixed distance of typically nine amino acids between the first and last peptide residue in contact with the MHCII groove (4). The residues at positions P1, P4, P6 and P9 of the binding core, termed anchor residues, are normally directed towards the MHC, and they can engage in interactions with pockets in the binding groove. As the α1 and β1 chains are highly polymorphic, the preferences of these pockets in terms of which residues they can accommodate can vary greatly. As a consequence, different MHCII alleles bind distinct subsets of peptides (5).

In contrast to MHCII, the binding groove of class I MHC molecules (MHCI) is closed at both ends, with most binding peptides having a length of nine amino acids. Longer peptides are accommodated either by taking on a bulged conformation with the middle of the peptide protruding out of the MHC groove (6), or less frequently by extending outside the C or N terminals (7,8). Peptides of length eight can also bind to the MHCI, by assuming a more stretched configuration of the backbone (3). We have shown with the most recent versions of *NetMHC* (9) and *NetMHCpan* (10) that these alternative modes of binding could be modeled with deletions and insertions in the binding core of the peptide. Deletions allow removing consecutive residues in peptides longer than nine amino acids to effectively align 10mers, 11mers, etc. to the same binding core of nine positions. Ideally, the deleted positions represent the residues that bulge out of the MHCI groove, whereas an insertion mimics a more stretched-out backbone and adapts 8mers to the common window of nine residues. This strategy has proven highly successful for MHCI binding prediction as it allows training models on peptides of all lengths at the same time, resulting in improved performance for all peptide lengths (9).

NetMHCIIpan is a widely used computational method that can predict quantitative binding of peptides to any MHCII molecule of known sequence. In several benchmarks, it has been shown to be the state-of-the-art for the prediction of peptide binding to MHC class II (11–13). Similarly to the methods for peptide-MHC class I binding prediction *NetMHC* and *NetMHCpan* discussed above, the algorithm underlying *NetMHCIIpan* relies on the neural networks training pipeline *NNAlign* (14,15). One major challenge faced when training machine learning methods for the prediction of peptide binding to MHCII stems from the open binding cleft of the MHCII molecule. This open conformation makes the location of the binding core within the peptide unknown a priori. Because the peptide binding core is the main determinant of the interaction with the MHC molecule, correct alignment of peptides is essential to identify the MHCII binding motifs. The *NNAlign* method allows for such accurate alignment based on quantitative peptide-MHC binding affinity (14). It does so by encoding the amino acid sequence of the peptide and of the MHC and several other features of the training examples (e.g. peptide length, peptide flanking region (PFR) composition and PFR length) to predict quantitative measurements of peptide-MHC binding affinity.

In this work, we applied the extended *NNAlign* pipeline (version 2.0) including modes for deletions and insertions, allowing us to model the peptide-MHC class II binding event with cores of variable length. In the context of MHCII, binding cores with insertions and deletions would indicate a non-canonical mode of binding with either a bulged or stretched configuration of the peptide in the binding groove. We assessed the predictive performance of the method on several datasets, and carried out mutational studies to validate the occurrence of non-canonical binding cores of length different from nine.

Materials and Methods

Data Sets

The method was trained on the binding data used in the *NetMHCIIpan-3.0* publication (16) (available at http://www.cbs.dtu.dk/suppl/immunology/NetMHCIIpan-3.0). This set consists of quantitative peptide-MHC class II binding data from the Immune Epitope Database (IEDB) (17), comprising 52,062 affinity measurements covering 24 HLA-DR, 5 HLA-DP, 6 HLA-DQ, and 2 murine H-2 molecules. The IC50 (half inhibitory concentration) values in nanomolar were transformed using the logarithmic formula 1-log(IC50)/log(50,000) as previously described (18) to make them fall in the range between 0 and 1.

Training the Artificial Neural Networks

The method was implemented as an ensemble of feed-forward neural networks with a single hidden layer as previously described (19). Peptide and MHC sequences were presented to the input layer of each network using BLOSUM encoding, where each amino acid was encoded as the BLOSUM50 matrix score vector of 20 amino acids (20). The optimal 9-mer core of a peptide therefore required 9 × 20 = 180 input neurons. Deletions and insertions were encoded as previously described (9) : cores longer than nine amino acids were reduced to nine positions by applying consecutive deletions at all possible positions in the core; cores shorter than nine were completed by introducing the wildcard X amino acid, encoded as a vector of zeros. The length of insertions L_i and deletions L_d was encoded with four input neurons with values $L_i/(L_i+1)$, $1-L_i/(L_i+1)$, $L_d/(L_d+1)$ and $1-L_d/(L_d+1)$. Forty additional input neurons were used to encode the composition of the peptide flanking regions (PFRs), calculated as the average BLOSUM scores on a maximum window of three amino acids at either end of the binding core (19). C- and N-terminal PFR lengths (L_{PFR}) were each encoded using two input neurons with values L_{PFR} /(L_{PFR}

+ 1) and 1-L_{PFR}/(L_{PFR} + 1) respectively. The peptide length L was encoded with two input neurons taking the values L_{PEP} and 1-L_{PEP}, where L_{PEP} = 1/(1 + exp((15-L)/2)). These transformations ensure that the normalized input values to the neural networks fall in the range between 0 and 1. MHC molecules were represented in terms of a pseudo-sequence defined by polymorphic residues in potential contact with a bound peptide (21). We used the same pseudo-sequences of 34 residues for the MHC alpha and beta chains defined by Karosiene et al. (16), resulting in additional 34 × 20 = 680 inputs. As a result, the total size of the input layer amounted to 910 neurons.

The single hidden layer was composed of 10, 15, 40, or 60 hidden neurons, and network weights were initialized with 10 different random configurations for each architecture. The resulting complete ensemble was therefore composed of 200 networks (5 cross-validation folds x 4 hidden layer sizes x 10 initial weights). The output layer was composed of a single neuron having as target value the binding affinity of the training example rescaled between 0 and 1 as described in the section *Data sets*. When allowing insertions and/or deletions, networks were trained using a burn-in period, a number of initial iterations where insertions and deletions were not allowed. After the burn-in, this constraint is relaxed and the algorithm starts attempting to introduce insertions/deletions as well.

Cross-Validation Subsets

The performance of the method on the binding affinity data set was performed using a 5-fold cross-validation, where four fifths of the data were used for training and one fifth for evaluation, repeating the procedure five times for all evaluation fifths. We used the same data partitions that had been generated to train *NetMHCIIpan-3.0* (16). In order to minimize over-estimation of predictive performance, these subsets for cross-validation were created using a Hobohm1-like algorithm called "common-motif" (22). Common-motif first selects seed sequences sharing at most a continuous stretch of eight amino acids, and then splits these seed sequence randomly into five partitions. The redundant sequences, i.e. those peptides with a common subsequence of nine or more contiguous amino acids, are then added to the group containing the seed used to define their redundancy.

Reduced Subsets with Minimal Redundancy

Although the procedure outlined above reduces the overlap between partitions compared to a random split of the data, some degree of overlap is unavoidable without removing data points. We constructed a reduced data set that ensures no overlap between the partitions. The algorithm to create the subsets starts with a prioritized list of sequences, sorted by the number of MHCs that each peptide is measured in complex with. That is, we start with peptides that have affinity measurements with the largest number of MHCs and proceed to accept or reject sequences based on their similarity to the ones that were already accepted. As we proceed down the list of sequences, there are three possible outcomes for a query sequence Q:

1) Q shares no contiguous stretch >= N amino acids with any of the accepted peptides; it is accepted and assigned randomly to one of the subsets.

2) Q shares a contiguous stretch >= N amino acids only to sequences of one subset S; it is accepted and placed in the subset S. Note that it can match to several sequences, but as long as they all belong to the same subset, it is still accepted.

3) Q shares a contiguous stretch >= N amino acids to sequences belonging to more than one subset; the peptide is discarded.

Starting with a prioritized list aims at removing as few sequences as possible while at the same time ensuring no overlap >= N amino acids between the subsets.

Statistical Tests

The predictive performances of alternative methods were compared using binomial tests. Given a pair of methods, the null hypothesis is that the two methods have equal probability of returning higher PCC (or AUC) on a given MHC allele. If method 1 has higher PCC in *n1* alleles and method 2 higher PCC in *n2* alleles, we estimated the p-value of this event as the two-tailed probability of observing *n1* or more wins by chance in a binomial distribution *B(n1 + n2, 0.5)*. Ties were excluded from the counts of *n1* and *n2*.

Measurement of MHC/Peptide Interactions by Monoclonal Antibody Capture

The binding capacity of peptide ligands to bind MHC molecules was measured using direct and quantitative binding assays as previously described elsewhere (23). In brief, inhibitor peptides to be tested were prepared at 10 mg/mL in 100% DMSO and subsequently diluted to desired concentration in 0.05% NP-40 (0.05% (v/v) Nonidet P-40 (NP-40; Fluka)/PBS, pH 7.2). Then, 5μL of each peptide dose were loaded into a 96-well plate. For positive (i.e. no inhibitor peptide) and negative (no MHC) controls 5μL 0.05% NP-40 was used. For each plate, an unlabeled standard peptide (i.e. peptide with known IC50 for the tested MHC) was used. Next, radioactive labeled standard peptide (for labeling procedure refer to Support Protocol 2 (23)) was mixed with PBS (pH 7.2), MHC, protease inhibitor cocktail and either 1.6% (v/v) NP-40/PBS or 0.82% Pluronic to make reaction mix (hot mix). For plate layouts and exact amounts of ingredients used in this assay, please refer to Sidney et al (23) as they vary among different class II MHCs tested. 10μL of hot mix were immediately added to all wells but the negative control. For negative controls, 10μL of reaction mix lacking MHC (cold mix) were added. Next, 96-well plates were sealed with Costar mats (Fisher Scientific, #07-200-614) and incubated 48h at room temperature (most assays). Alternatively, some assays required 72h of incubation and/or an incubation temperature of 37°C or MHC specific modifications, e.g. adjustment of final pH to optimize reactions (23). Finally, MHC/peptide complexes were captured utilizing monoclonal antibody-coated plates (Support protocol 4 (23)) and IC50 were determined using the Topcount (PerkinElmer Instruments) benchtop microplate scintillation and luminescence counter.

Results

Modeling peptide binding to MHCII, and particularly which peptide residues are implicated in the interaction, requires an alignment of the peptide sequences. We investigated whether, allowing the peptide alignment to contain insertions and deletions, we could generate better prediction models compared to what could be obtained with an un-gapped alignment, and experimentally validated the binding mode of a set of peptides predicted to contain non-canonical binding cores.

Priming Neural Networks with a Burn-In Period

When expanding the mode of binding cores to include insertions and deletions, the solutions space becomes substantially larger. For instance, there are seven possible un-gapped 9mer cores in a 15mer peptide, but there are 45 possible 10mers with one deletion, and 56 possible

8mers with one insertion. The abundance of possible core configurations makes it likely that a naive network will get lost in local minima mainly composed of these non-9mer cores. To avoid this, we introduced a burn-in period, in which a number of initial iterations in neural network training were performed without allowing for insertions and deletions. The burn-in period has the function of guiding the neural networks onto the right path, using the prior knowledge that 9mer cores should be the norm and that cores with insertions/deletions are the exception.

We trained neural networks on the binding affinity data set (see *Materials and Methods*) with and without insertions/deletions, and with different burn-in periods. The method without insertions/deletions can be thought of as having a burn-in period equal to the total number of training iterations, and is essentially equivalent to the *NetMHCIIpan-3.0* method (16).

It is evident from Figure 1 that the burn-in iterations are crucial to drive neural network learning: training without this initial priming (burn-in=0) did not improve the performance compared to the method trained without insertions and deletions, termed NoGap in Figure 1 (p=0.26, two-tailed binomial test). With an appropriate burn-in rate (around 50-150 out of 500 total iterations) and a maximum length for deletions and insertions of one amino acid, we observed an average improvement in predictive performance in a cross-validation setup (Figure 1). The increase in Pearson Correlation Coefficient (PCC) compared to NoGap was significant, with higher performance observed on 34/37 molecules with burn-in=100 ($p=10^{-7}$, two-tailed binomial test). Similarly, when we evaluated predictive performance in terms of AUC, the method with insertions/deletions outperformed NoGap on 33 out of 37 molecules ($p=10^{-6}$), with average AUC=0.875 compared to 0.870 of the NoGap method. The method trained with at most one deletion and one insertion also outperformed neural networks trained with at most one deletion and no insertions ($p=10^{-7}$), and networks allowing no more than one insertion and no deletions ($p=10^{-6}$). Allowing longer deletions of up to 2 amino acids did not further improve cross-validated performance (p=0.32). Figure 2 summarizes these comparisons.

Insertions and Deletions are Relatively Rare and Allele-Specific

In order to estimate the frequency of the deletion and insertion events, we submitted 3.7 million (100,000 for each of the 37 molecules in the data set) random natural 15mer peptides to the neural networks. Defining peptides within the top 10th percentile as binders, we find that 2.0%

of the binders were predicted to contain a core deletion, and 0.45% presented an insertion. However, we observed that certain molecules had stronger preferences for insertions/deletions whereas they were never predicted to occur in others (Figure 3). For example, even with a relatively conservative burn-in of 100 iterations, DRB3*01:01 and DRB5*01:01 had more than 10% of their predicted binders with an optimal 8- or 10-amino acids core. Conversely, all HLA-DP and HLA-DQ showed less than 1% of predicted binders with non-canonical cores. The mouse molecules H-2-IAb and H-2-IAd also had a very low fraction of insertions/deletions. The profiles of insertion and deletion frequencies for different values of the burn-in rate are shown in Figure S1.

Data Redundancy and Its Effect on Cross-Validated Performance

The data partitions used to train *NetMHCIIpan-3.0* in cross-validation were generated using a common-motif procedure (described in *Materials and Methods* and in ref. (22)), which aimed at limiting the sequence similarity between partitions. Although this procedure reduces the overlap between partitions compared to a random split of the data, some degree of sequence similarity is unavoidable without removing data points. Indeed, on the binding affinity data set, after applying the common-motif, 45% of the sequences still shared a continuous stretch >= 9 amino acids with at least one other sequence in a different partition. Such redundancy can lead to overestimating the predictive performance in cross-validation (24), as these examples are relatively easy to predict - the neural networks have seen a very similar example in the training phase. Over-estimation of the predictive performance is expected to be more pronounced for methods with larger search spaces and model parameters. Therefore, one may argue that the gain in performance obtained by including deletions and insertions could be an artifact of data redundancy and of the increase in model complexity.

To investigate this, we generated reduced data sets that ensured no overlap >= N contiguous amino acids between any sequence in different cross-validation subsets, applying the algorithm described in *Materials and Methods*. Depending on the length of the common-motif threshold N, a different number of sequences must be removed from the data set to fulfill the no-overlap condition. Figure 4 shows the percentage of removed sequences depending on N. For common-motif lengths 3-5, as expected, the number of removed peptides decreases when considering longer cores. However, the number of peptides removed for lengths 6-10 is essentially the same (about 14-15% of the sequences). This is likely due to a significant fraction of peptides tested for binding being generated as 15-mers overlapping by 10 residues, which is a common approach to cover an antigen of interest (25,26). Therefore, for a more stringent reduction of redundancy (without losing any extra data), we continued the analysis using N=6. In other words, we removed about 7,800 sequences of the 52,062 in the training set, ensuring that no two sequences across subsets shared more than 5 consecutive amino acids. Approximately the same number of peptides would have to be removed to ensure a maximum overlap of 8 (N=9).

The MHCII-peptide binding predictor was retrained on these reduced partitions, both with and without insertions/deletions. As expected, the cross-validated performance on the data set without overlap between partitions drops significantly ($p=10^{-8}$) compared to the method trained on all data, as all the evaluation examples now have a lower degree of similarity to the training data. However, even in this extreme scenario, the performance of the method remains high with average PCC values in the order of 0.7 for all burn-in rates (Figure 5). More importantly, we observed also on these data that the method with insertions/deletions performs significantly better than the NoGap counterpart, with higher PCC for 31/37 molecules ($p=4*10^{-5}$) with a burn-in rate = 100. This demonstrates that the prediction of peptides with non-canonical binding cores was not merely a result memorized by the neural networks as outliers, and that the rules of placing insertions/deletions could be generalized and applied on a set of peptide sequences very different from the ones used to train the networks.

A Mutational Analysis Supports the Hypothesis of 10mer Binding Cores

In order to validate experimentally the occurrence of binding cores with length different from nine amino acids, we designed a number of mutations to four validated MHCII binding peptides. The four wild type (WT) peptides in Figure 6 were all predicted to contain a 10mer binding core (i.e. they comprised a deletion). In all four cases, the canonical and non-canonical predicted cores agree on the location of one anchor, and disagree on the position of a second anchor residue.

The canonical binding motif for DRB1*03:01 has anchors at P1 with preference for hydrophobic amino acids, and at P4 with a preference for D (see Figure 6). Both DRB1*03:01 binders fail to match this motif, as the D at P4 is separated by four residues from the nearest hydrophobic amino acid. Two explanations for this inconsistency are possible: the P1 anchor can tolerate non-hydrophobic amino acids, or the distance between P1 and P4 can be different from three amino acids. In the second scenario, we would predict that: 1) mutating the D at P4 to a dissimilar amino acid should abolish binding; 2) mutating the putative non-canonical P1 to a non-hydrophobic amino acid should also remove binding; 3) mutating the putative P1 of the 9mer core should not affect binding; 4) additional mutations outside of the predicted core should not have an effect on binding.

The binding motif for DRB5*01:01 has a strong hydrophobic anchor at P1, and a strong positively charged anchor at P9 (see Figure 6). Both binders are inconsistent with this motif, as the distance between the P1 and P9 in both cases is of ten residues. If these peptides interact with the MHCII with a binding core of ten amino acids, we would predict that: 1) mutating the F at P1 to a dissimilar amino acid should abolish binding; 2) mutating the P1+9 residue (i.e. the last residues in the predicted 10mer core) should prevent binding; 3) mutating the P1+8 residue (i.e. the last residue in a putative 9mer core) should not affect binding; 4) mutations outside of the predicted core(s) should not have an effect on binding.

The effects of targeted mutations on the four peptides are listed in Table 1. In most cases the measured IC50 affinity of the variants supports the presence of a 10mer binding core. Mutations on anchors of the 10mer cores abolished binding, whereas mutated anchors according to a putative 9mer core did not affect binding affinity compared to the WT. Therefore, successful binding for these peptides requires both anchor residues, and these are separated by a number of residues that is only consistent with a 10mer binding core. While the size of the effect was not identical for all mutations, the direction of the change in binding strength was consistent. For instance, peptide ISFCNANPGLMKDVA was measured to bind with affinity of 12nM to HLA-DRB5*01:01, a value that ranks this sequence among the top 1.2% of predicted binders when compared to NetMHCIIpan prediction scores for a large set of random natural 15mer peptides. However, the measured affinity of the two mutants P3F>K (522 nM) and P12K>G (599nM) translate into respectively 35% and 38% in terms of such percentile ranks; in this light, the two mutations had a very severe effect on the strength of the interaction.

The nearly three-fold reduction in binding affinity for the P4H>G mutant of MYFHKRDMRLLSLAV was somehow surprising, as the histidine at the predicted P2 of the core is not expected to play a determinant role in the binding capability of the peptide. However, glycine is a special amino acid since it only has a single hydrogen atom as its side chain. This small side chain grants high flexibility to the polypeptide chain, and glycine can therefore have rotations angles forbidden by most other amino acids. The mutation H->G could impose structure variations in the peptide conformation not captured by the prediction method. This said, the mutations at the predicted P1 and P4 of the binding core (P3F>A and P7D>V, respectively) have much more dramatic effects on the measured IC50, pointing to the dominant role of these two residues in determining the peptide-MHC interaction. Much more unexpected was the outcome for the P8D>V variant of LQIIDKI**D**AAFKVAA, which turned out as a binder (IC50=59nM) despite losing the aspartic acid (D) at the P4 anchor of the binding core (highlighted in bold). The *NNAlign* network ensemble predicts that the P8D>V mutant LQIIDKI**V**AAFKVAA has affinity of 230 nM for DRB1*03:01; this peptide probably uses an alternative binding register, with affinity comparable to the primary binding register, exploiting the D three amino acids to the left of the D>V mutation.

Discussion

The currently accepted paradigm for the event of peptide-MHC class II binding portrays the peptide lying flat in the peptide-binding groove of the MHC, extending out of both sides of the groove. The conformation of the peptide backbone is highly conserved, with a fixed distance of nine amino acids (the binding core) between the first and last peptide residue in contact with the MHC class II groove. This model is supported by a vast amount of literature and numerous crystal structures of peptide-MHC complexes (4,27), and has been tremendously useful to study the mechanisms of epitope presentation. While confirming that the great majority of peptides bind to the MHC using this canonical mode, in this work we report evidence for an alternative, less common mode of interaction. The basic idea is that a fraction of observed ligands have an unconventional spacing of the anchor residues that can only be accommodated to the canonical MHC motif with a more stretched out peptide backbone, if the anchors are closer to each other than expected, or by the peptide bulging out of the MHC groove, if the anchors are separated by too many residues.

The first, indirect evidence for a non-canonical mode of binding comes from a machine learning benchmark on a large set of binding affinity data. Allowing the peptide binding core to take on a variable length of either eight, nine or ten residues, we showed that we could build models with significantly higher predictive performance compared to having a fixed binding core of nine contiguous amino acids. Because more accurate models should correspond to better approximations of the biological system being modeled, these results suggest that such non-canonical cores do occur in practice. These observations were also confirmed in a redundancy-controlled setup, where we ensured that the improved accuracy was not an artifact of over-fitting on highly similar sequences.

Further evidence for the occurrence of the proposed non-canonical mode of binding was collected with a mutational study of four MHCII binders. These peptides were all predicted to contain a deletion (that is, a 10mer binding core). We showed that the binding measurements of the mutated peptides are only coherent with a mode of binding that involves a binding core of ten amino acids, and that a canonical 9mer binding core cannot accommodate the peculiar anchor spacing of these peptides. Taken all together, these results are strongly suggestive of a non-canonical mode of binding for MHCII ligands that, to be best of our knowledge, has never been described before.

Recent advances in mass-spectrometry have enabled large-scale analyses of the collection of peptides naturally presented by MHC molecules in a particular cell line. These technologies have a tremendous potential as they can generate thousands of datapoints in a single experiment (see for instance refs (28,29)). However, since antigen presenting cells normally express multiple MHC isoforms, the MHC restriction of each ligand is generally not known and must be assigned, either with a predictor of binding to MHC or through unsupervised clustering (30). While most studies in this field have focused on MHC class I eluted ligands, reports of class II ligandomes have started to appear (e.g ref (31)). When sufficient and reliable MHCII ligandome data become available, and the issues of restriction assignment are confidently solved, it will be essential to investigate the presence of non-canonical binding also in MHCII natural ligands. Ultimately, crystal structures of peptide-MHC class II complexes displaying a bulged mode of binding would be highly desirable, both to further confirm the occurrence of non-canonical binders and to study their possible effect on the recognition by the T-cell receptor.

The results described in this work were obtained using the fully automated *NNAlign* machine-learning pipeline (32). *NNAlign* has been previously applied to generate models of protease cleavage (15), characterization of MHC class II binding motifs (33), and it forms the foundation for the state-of-the-art prediction algorithms *NetMHC* (9), *NetMHCpan* (10), *NetMHCII* (14) and

NetMHCIIpan (12). In all these examples, it has been proven capable of detecting subtle motifs in quantitative peptide data, and of generating powerful prediction models for several biological problems. In this work, *NNAlign* highlighted anomalies in the manner a fraction of ligands bind to the MHCII, and suggested a non-canonical mode of peptide-MHC binding that was later confirmed by experimental validation. However, its applications are not limited to the MHC system, and it can be readily employed to study other kinds of receptor-ligand interactions.

Conflicts of interest: None

References

1. Cresswell P. Assembly, transport, and function of MHC class II molecules. Annu Rev Immunol. 1994;12:259–93.

2. Germain RN. MHC-dependent antigen processing and peptide presentation: providing ligands for T lymphocyte activation. Cell. 1994 Jan;76(2):287–299.

3. Rudolph MG, Stanfield RL, Wilson IA. How TCRs bind MHCs, peptides, and coreceptors. Annual Review of Immunology. 2006;24:419–466.

4. Jones EY, Fugger L, Strominger JL, Siebold C. MHC class II proteins and disease: a structural perspective. Nat Rev Immunol. 2006 Apr;6(4):271–82.

5. Greenbaum J, Sidney J, Chung J, Brander C, Peters B, Sette A. Functional classification of class II human leukocyte antigen (HLA) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across supertypes. Immunogenetics. 2011 Jun;63(6):325–35.

6. Tynan FE, Borg NA, Miles JJ, Beddoe T, El-Hassen D, Silins SL, et al. High resolution structures of highly bulged viral epitopes bound to major histocompatibility complex class I. Implications for T-cell receptor engagement and T-cell immunodominance. J Biol Chem. 2005 Jun 24;280(25):23900–9.

7. McMurtrey C, Trolle T, Sansom T, Remesh SG, Kaever T, Bardet W, et al. Toxoplasma gondii peptide ligands open the gate of the HLA class I binding groove. eLife [Internet]. 2016 Jan;5. Available from: http://elifesciences.org/lookup/doi/10.7554/eLife.12556

8. Remesh SG, Andreatta M, Ying G, Kaever T, Nielsen M, McMurtrey C, et al. Breaking confinement: unconventional peptide presentation by major histocompatibility (MHC) class I allele HLA-A*02:01. J Biol Chem. 2017 Feb 8;

9. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics. 2016 Feb;32(4):511–517.

10. Nielsen M, Andreatta M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. Genome Medicine. 2016 Dec;8(1):33.

11. Zhang L, Chen Y, Wong H-S, Zhou S, Mamitsuka H, Zhu S. TEPITOPEpan: extending TEPITOPE for peptide binding prediction covering over 700 HLA-DR molecules. PloS One. 2012;7(2):e30483.

12. Andreatta M, Karosiene E, Rasmussen M, Stryhn A, Buus S, Nielsen M. Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification; implications for the interpretation of T cell cross-reactivity. 2015;

13. Shen W-J, Wei YT, Guo X, Smale S, Wong H-S, Li SC. MHC binding prediction with KernelRLSpan and its variations. J Immunol Methods. 2014 Apr;406:10–20.

14. Nielsen M, Lund O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. BMC bioinformatics. 2009;10:296.

15. Andreatta M, Schafer-Nielsen C, Lund O, Buus S, Nielsen M. NNAlign: A web-based prediction method allowing non-expert end-user discovery of sequence motifs in quantitative peptide data. Haslam NJ, editor. PLoS ONE. 2011 Nov;6(11):e26781.

16. Karosiene E, Rasmussen M, Blicher T, Lund O, Buus S, Nielsen M. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. Immunogenetics. 2013 Oct;65(10):711–724.

17. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, et al. The immune epitope database (IEDB) 3.0. Nucleic Acids Research. 2015 Jan;43(Database issue):D405–412.

18. Nielsen M, Lundegaard C, Worning P, Lauemøller SL, Lamberth K, Buus S, et al. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Science: A Publication of the Protein Society. 2003 May;12(5):1007–1017.

19. Nielsen M, Lundegaard C, Blicher T, Peters B, Sette A, Justesen S, et al. Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. PLoS computational biology. 2008;4(7):e1000107.

20. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA. 1992;89:10915–10919.

21. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, et al. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. PloS One. 2007;2(8):e796.

22. Nielsen M, Lundegaard C, Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. BMC bioinformatics. 2007;8:238.

23. Sidney J, Southwood S, Moore C, Oseroff C, Pinilla C, Grey HM, et al. Measurement of MHC/peptide interactions by gel filtration or monoclonal antibody capture. Curr Protoc Immunol. 2013 Feb;Chapter 18:Unit 18.3.

24. Kim Y, Sidney J, Buus S, Sette A, Nielsen M, Peters B. Dataset size and composition impact the reliability of performance benchmarks for peptide-MHC binding predictions. BMC Bioinformatics. 2014 Jul 14;15:241.

25. Arlehamn CSL, Sidney J, Henderson R, Greenbaum JA, James EA, Moutaftsi M, et al. Dissecting mechanisms of immunodominance to the common tuberculosis antigens ESAT-6, CFP10, Rv2031c (hspX), Rv2654c (TB7.7), and Rv1038c (EsxJ). J Immunol. 2012 May 15;188(10):5020–31.

26. Mothé BR, Lindestam Arlehamn CS, Dow C, Dillon MBC, Wiseman RW, Bohn P, et al. The TB-specific CD4(+) T cell immune repertoire in both cynomolgus and rhesus macaques largely overlap with humans. Tuberculosis (Edinb). 2015 Dec;95(6):722–35.

27. Bjorkman PJ. Not second class: the first class II MHC crystal structure. J Immunol. 2015 Jan 1;194(1):3–4.

28. Admon A, Bassani-Sternberg M. The Human Immunopeptidome Project, a suggestion for yet another postgenome next big thing. Mol Cell Proteomics. 2011 Oct;10(10):O111.011833.

29. Sofron A, Ritz D, Neri D, Fugmann T. High-resolution analysis of the murine MHC class II immunopeptidome. Eur J Immunol. 2016 Feb;46(2):319–28.

30. Andreatta M, Lund O, Nielsen M. Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. Bioinformatics. 2013 Jan 1;29(1):8–14.

31. Mommen GPM, Marino F, Meiring HD, Poelen MCM, van Gaans-van den Brink JAM, Mohammed S, et al. Sampling From the Proteome to the Human Leukocyte Antigen-DR (HLA-DR) Ligandome Proceeds Via High Specificity. Mol Cell Proteomics. 2016 Apr;15(4):1412–23.

32. Nielsen M, Andreatta M. NNAlign: a platform to construct and evaluate artificial neural network models of receptor-ligand interactions. Nucleic Acids Res. 2017 Apr 12;

33. Andreatta M, Nielsen M. Characterizing the binding motifs of 11 common human HLA-DP and HLA-DQ molecules using NNAlign. Immunology. 2012;

Figure Legends

Figure 1: Correlation coefficient (average over 37 molecules) of the method versus the number of burn-in iterations used to prime the networks. Networks were trained in cross-validation with a maximum insertion length of one amino acid and a maximum deletion length of one amino acid. NoGap corresponds to the method trained without insertions and deletions.

Figure 2: Correlation coefficient for methods trained with different deletion (d) and insertion (i) maximum lengths. The method with at most one deletion and one insertion (d=1;i=1) had significantly higher performance than the method without insertions/deletions (d=0;i=0) with higher PCC in 34/37 molecules ($p=10^{-7}$). It also outperformed d=1;i=0 on 34/37 molecules ($p=10^{-7}$) and d=0;i=1 on 33/37 ($p=10^{-6}$) molecules. Allowing longer deletions of up to 2 amino acids (d=2;i=1) does not significantly improve cross-validated performance compared to d=1;i=1 (p=0.32). All p-values were calculated with two-tailed binomial tests.

Figure 3: Frequency of predicted binders containing deletions (black) and insertions (white) in the binding core using a burn-in of 100 iterations. While some molecules have over 10% of predicted binders with a non-canonical binding core (e.g. HLA-DRB3*01:01 and HLA-DRB5*01:01), others have no predicted binding cores with insertions or deletions (e.g. the mouse H-2 alleles). The figure was generated, for each allele, from the top 10% scoring peptides out of 100,000 natural random 15-mers.

Figure 4: Percent of peptide-MHCs that must be removed to ensure no overlap between cross-validation partitions, depending on the common-motif threshold N on the x-axis. Approximately the same number of peptides has to be removed (~14-15%) for N=6 to N=10, reflecting a possible bias in the procedures used to assay binding affinity with a sliding window over antigens of interest.

Figure 5: Correlation coefficient (average over 37 molecules) as a function of the burn-in rate for networks trained on the low redundancy data set. Networks were trained in cross-validation with a maximum insertion length of one amino acid and a maximum deletion length of one amino acid. NoGap corresponds to the method trained without insertions and deletions.

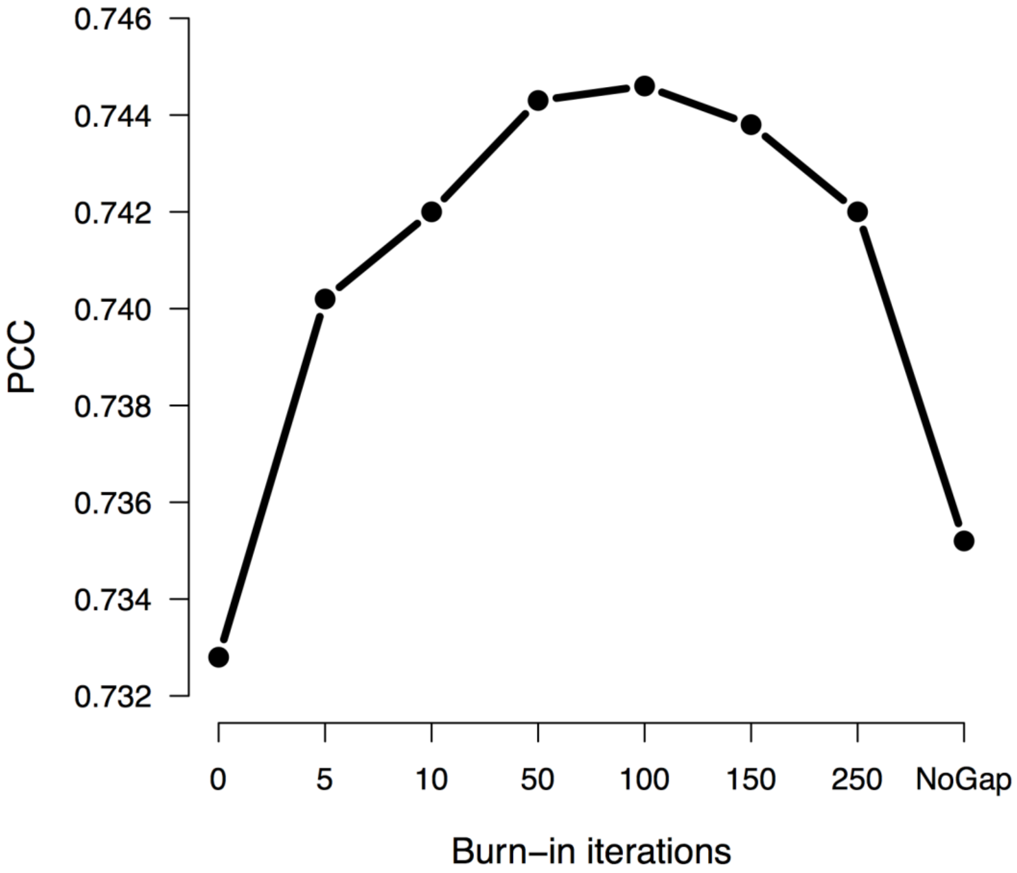
Figure 6: Predicted 9mer and 10mer binding cores for two DRB1*03:01 and two DRB5*01:01 ligands. The non-canonical spacing of the anchor residues can only be accommodated with a deletion in the binding core, which is depicted here as a protrusion of the peptide chain at the predicted position of the deletion. Reference sequence logos are from NNAlign (15).

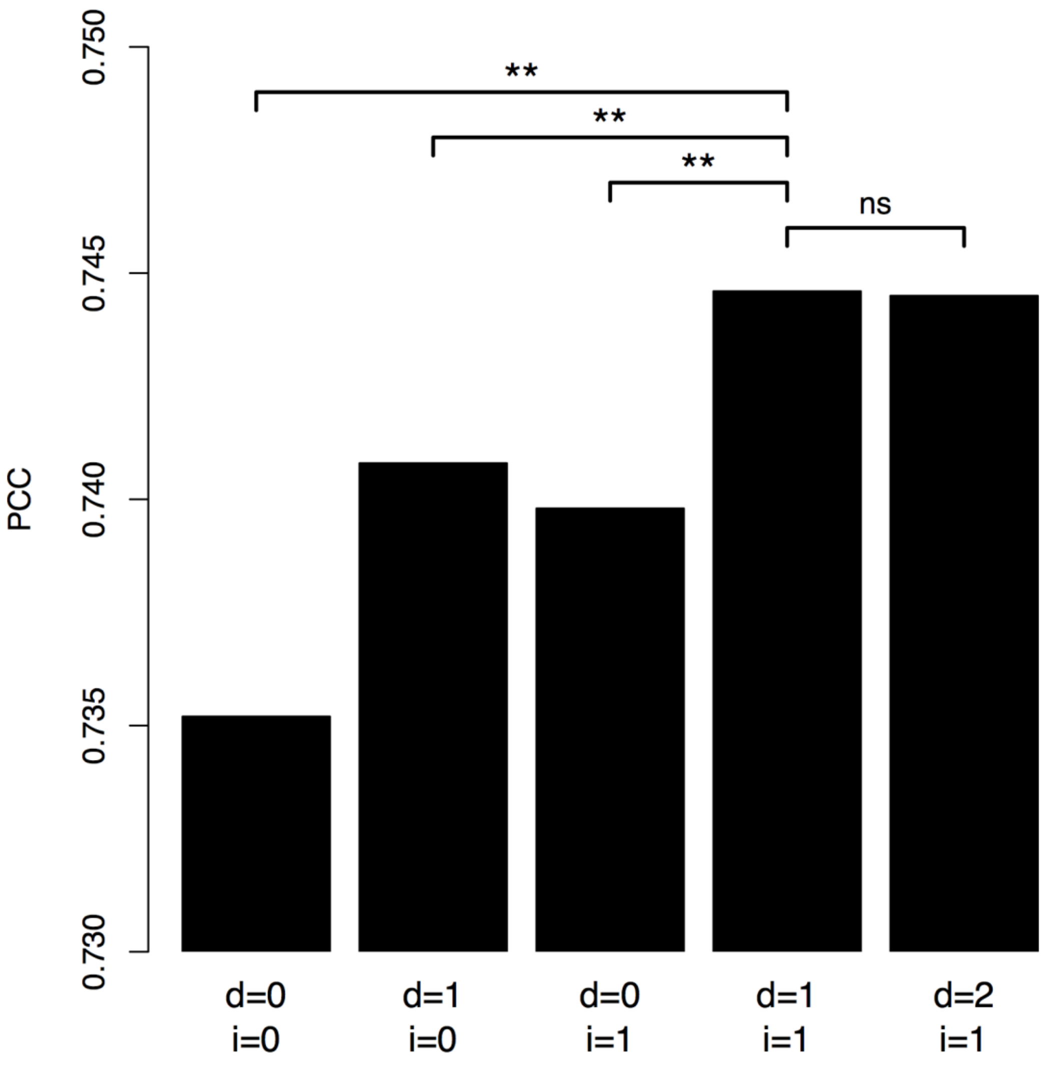
Tables

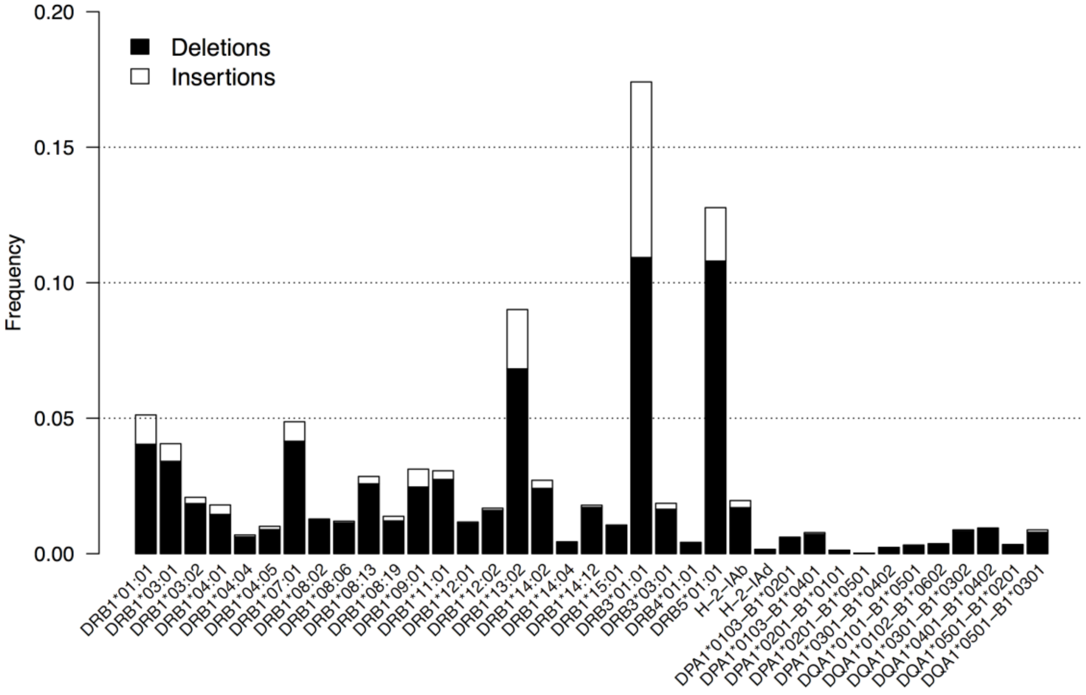
Table 1: Measured IC50 binding affinity for four MHC class II ligands in their wild type (WT) and mutated variants.

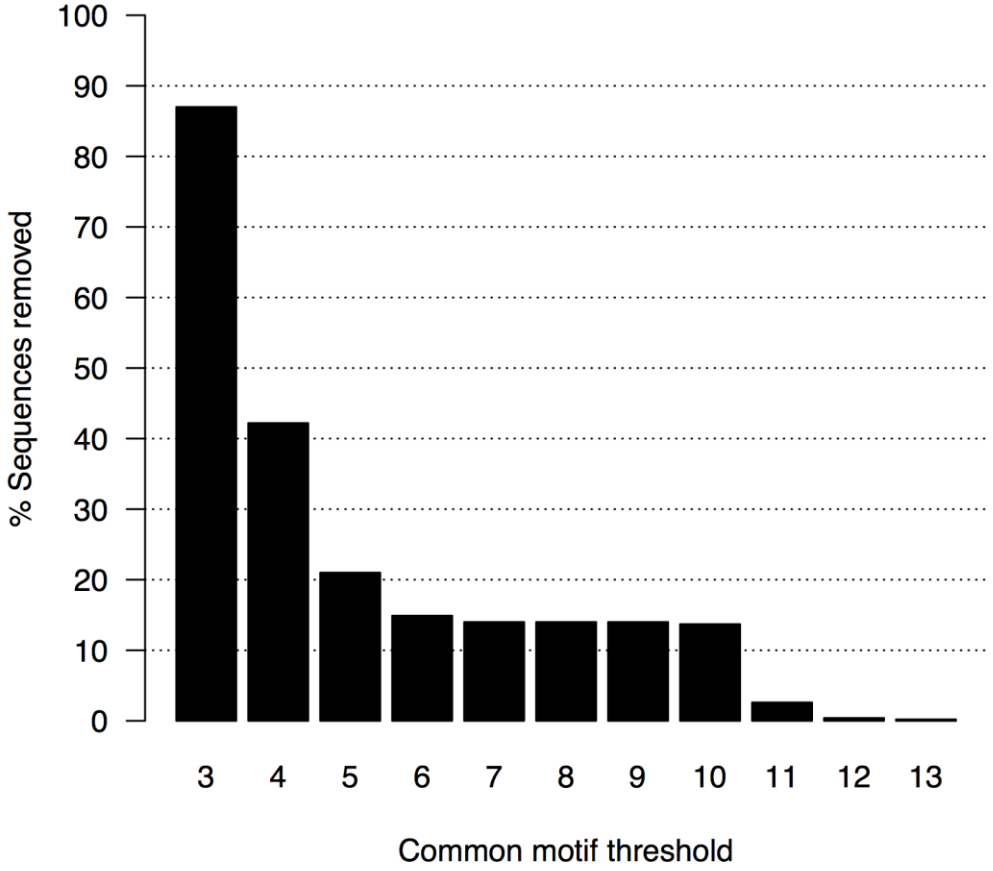
Peptide	Mut	MHC	Pred	IC50
MYFHKRDMRLLSLAV	WT	DRB1*03:01	+	272
MYFHKR**V**MRLLSLAV	P7D>V	DRB1*03:01	-	29300
MY**A**HKRDMRLLSLAV	P3F>A	DRB1*03:01	-	1780
MYF**G**KRDMRLLSLAV	P4H>G	DRB1*03:01	+	743
M**E**FHKRDMRLLSLAV	P2Y>E	DRB1*03:01	+	124
LQIIDKIDAAFKVAA	WT	DRB1*03:01	+	90
LQIIDKI**V**AAFKVAA	P8D>V	DRB1*03:01	-	59[§]
LQI**A**DKIDAAFKVAA	P4I>A	DRB1*03:01	-	539
LQII**R**KIDAAFKVAA	P5D>R	DRB1*03:01	+	89
LQ**A**IDKIDAAFKVAA	P3I>A	DRB1*03:01	+	152
RNVFDEVIPTAFKIG	WT	DRB5*01:01	+	37
RNV**R**DEVIPTAFKIG	P4F>R	DRB5*01:01	-	652
RNVFDEVIPTAF**A**IG	P13K>A	DRB5*01:01	-	666
RNVFDEVIPTA**R**KIG	P12F>R	DRB5*01:01	+	5
RN**E**FDEVIPTAFKIG	P3V>E	DRB5*01:01	+	15
ISFCNANPGLMKDVA	WT	DRB5*01:01	+	12
IS**K**CNANPGLMKDVA	P3F>K	DRB5*01:01	-	522
ISFCNANPGLM**G**DVA	P12K>G	DRB5*01:01	-	599
ISFCNANPGL**R**KDVA	P11M>R	DRB5*01:01	+	0.2
I**E**FCNANPGLMKDVA	P2S>E	DRB5*01:01	+	10

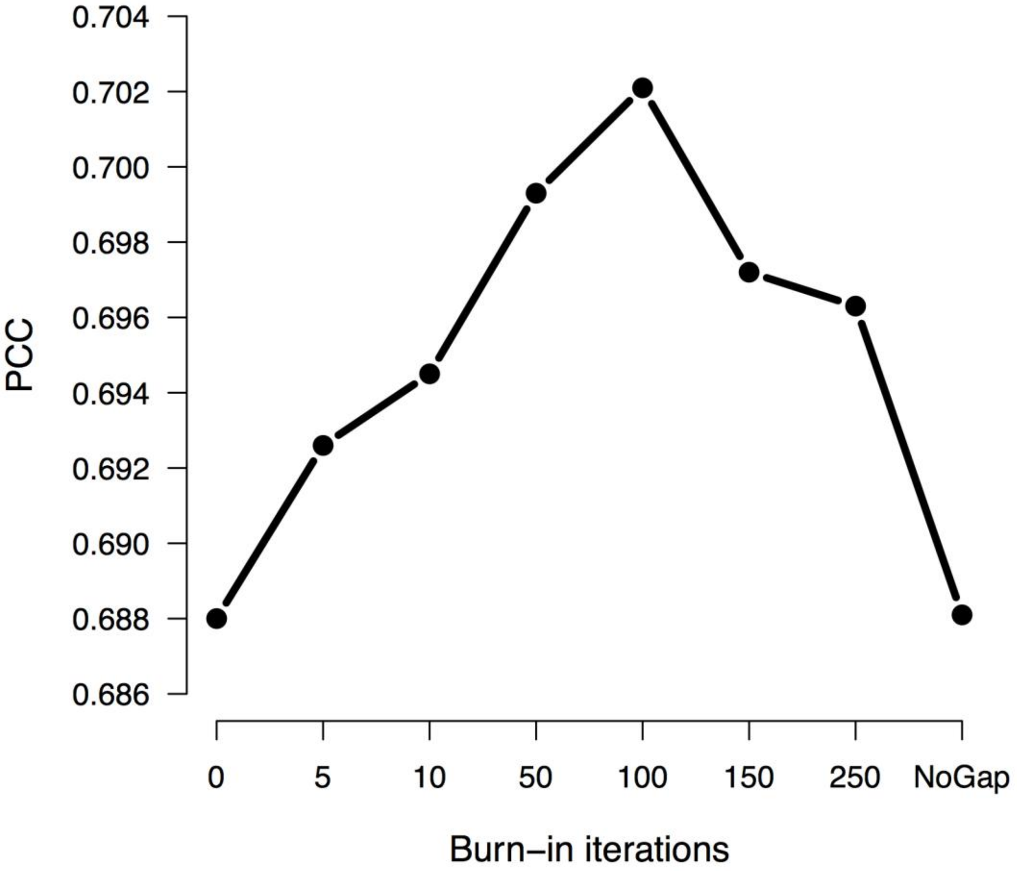
Mutations are listed in the Mut column and highlighted in bold letters in the peptide. The predicted 10mer binding core is underlined. Pred is the expected outcome if the peptide contained a 10mer binding core. [§] This mutant has a predicted secondary binding core (see text).

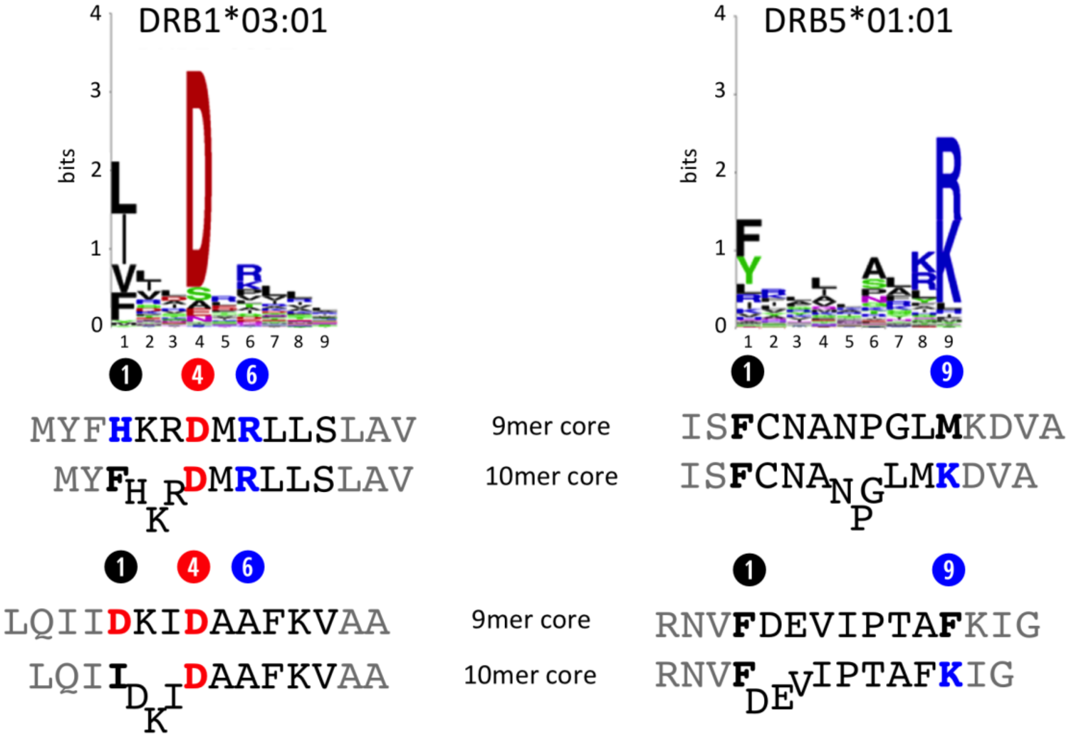

0.746
0.744
0.742
0.740
0.738
0.736
0.734
0.732
PCC
0
5
10
50
100
150
250
NoGap
Burn–in iterations


PCC
0.730
0.735
0.740
0.745
0.750
**
**
**
ns
d=0
i=0
d=1
i=0
d=0
i=1
d=1
i=1
d=2
i=1


0.20
0.15
0.10
0.05
0.00
Frequency
■ Deletions
□ Insertions
DRB1*01:01
DRB1*03:01
DRB1*03:02
DRB1*04:01
DRB1*04:04
DRB1*04:05
DRB1*07:01
DRB1*08:02
DRB1*08:06
DRB1*08:13
DRB1*08:19
DRB1*09:01
DRB1*11:01
DRB1*12:01
DRB1*12:02
DRB1*13:02
DRB1*14:02
DRB1*14:04
DRB1*14:12
DRB1*15:01
DRB3*01:01
DRB3*03:01
DRB4*01:01
DRB5*01:01
H−2−IAb
H−2−IAd
DPA1*0103−B1*0201
DPA1*0103−B1*0401
DPA1*0201−B1*0101
DPA1*0201−B1*0501
DPA1*0301−B1*0402
DQA1*0101−B1*0501
DQA1*0102−B1*0602
DQA1*0301−B1*0302
DQA1*0401−B1*0402
DQA1*0501−B1*0201
DQA1*0501−B1*0301


100
90
80
70
60
50
40
30
20
10
0
% Sequences removed
3
4
5
6
7
8
9
10
11
12
13
Common motif threshold


PCC
0.704
0.702
0.700
0.698
0.696
0.694
0.692
0.690
0.688
0.686
0
5
10
50
100
150
250
NoGap
Burn–in iterations


DRB1*03:01
DRB5*01:01
bits
4
3
2
1
0
1 2 3 4 5 6 7 8 9
1
4
6
9
MYFHKRDMRLLSLAV
MYFHKRDMRLLSLAV
9mer core
10mer core
ISFCNANPGLMKDVA
ISFCNANPGLMKDVA
LQIIDKIDAAFKVAA
LQIIDKIDAAFKVAA
RNVFDEVIPTAFKIG
RNVFDEVIPTAFKIG