Barra *et al. Genome Medicine* (2018) 10:84
https://doi.org/10.1186/s13073-018-0594-6

Genome Medicine

RESEARCH Open Access

Footprints of antigen processing boost MHC class II natural ligand predictions

Carolina Barra[1†], Bruno Alvarez[1†], Sinu Paul[2], Alessandro Sette[2], Bjoern Peters[2], Massimo Andreatta[1], Søren Buus[3] and Morten Nielsen[1,4*]

Abstract

Background: Major histocompatibility complex class II (MHC-II) molecules present peptide fragments to T cells for immune recognition. Current predictors for peptide to MHC-II binding are trained on binding affinity data, generated in vitro and therefore lacking information about antigen processing.

Methods: We generate prediction models of peptide to MHC-II binding trained with naturally eluted ligands derived from mass spectrometry in addition to peptide binding affinity data sets.

Results: We show that integrated prediction models incorporate identifiable rules of antigen processing. In fact, we observed detectable signals of protease cleavage at defined positions of the ligands. We also hypothesize a role of the length of the terminal ligand protrusions for trimming the peptide to the MHC presented ligand.

Conclusions: The results of integrating binding affinity and eluted ligand data in a combined model demonstrate improved performance for the prediction of MHC-II ligands and T cell epitopes and foreshadow a new generation of improved peptide to MHC-II prediction tools accounting for the plurality of factors that determine natural presentation of antigens.

Keywords: MHC-II, Binding predictions, Eluted ligands, T cell epitope, Neural networks, Antigen processing, Machine learning, Mass spectrometry

Background

Major histocompatibility complex class II (MHC-II) molecules play a central role in the immune system of vertebrates. MHC-II present exogenous, digested peptide fragments on the surface of antigen-presenting cells, forming peptide-MHC-II complexes (pMHCII). On the cell surface, these pMHCII complexes are scrutinized, and if certain stimulatory conditions are met, a T helper lymphocyte may recognize the pMHCII and initiate an immune response [1].

The precise rules of MHC class II antigen presentation are influenced by many factors including internalization and digestion of extracellular proteins, the peptide binding motif specific for each MHC class II molecule, and the transport and surface half-life of the pMHCIIs. The MHC-II binding groove, unlike MHC class I, is open at both ends. This attribute facilitates peptide protrusion out of the groove, thereby allowing longer peptides (and potentially whole proteins) to be loaded onto MHC-II molecules [2, 3]. Peptide binding to MHC-II is mainly determined by interactions within the peptide binding groove, which most commonly encompass a peptide with a consecutive stretch of nine amino acids [4]. Ligand residues protruding from either side of the MHC binding groove are commonly known as peptide flanking regions (PFRs). The PFRs are variable in length and composition and affect both the peptide MHC-II binding [5] and the subsequent interaction with T cells [6–8]. The open characteristic of the MHC-II binding groove does not constrain the peptides to a certain length, thereby increasing the diversity of sequences that a given MHC-II molecule can present. Also, MHC-II molecules are highly polymorphic, and their binding motifs have appeared to be more degenerate than MHC-I motifs [9–11].

* Correspondence: mniel@bioinformatics.dtu.dk
†Carolina Barra and Bruno Alvarez contributed equally to this work.
[1]Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, CP1650 San Martín, Argentina
[4]Department of Bio and Health Informatics, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark
Full list of author information is available at the end of the article

Considering all the aspects mentioned above, MHC-II motif characterization and rational identification of MHC-II ligands and epitopes is a highly challenging and costly endeavor. Because MHC-II is a crucial player in the exogenous antigen presentation pathway, considerable efforts have been dedicated in the past to develop efficient experimental techniques for MHC-II peptide binding quantification. The traditional approach to quantify peptide MHC-II binding relies on measuring binding affinity, either as the dissociation constant (Kd) of the complex [12, 13] or in terms of IC50 (concentration of the query peptide which displaces 50% of a bound reference peptide) [14]. To date, data repositories such as the Immune Epitope Database (IEDB) [15] have collected more than 150,000 measurements of peptide-MHC-II binding interactions. Such data have been used during the last decades to develop several prediction methods with the ability to predict binding affinities to the different alleles of MHC class II. While the accuracy of these predictors has increased substantially over the last decades due the development of novel machine learning frameworks and a growing amount of peptide binding data being available for training [16], state-of-the-art methods still fail to accurately predict accurately MHC class II ligands and T cell epitopes [17, 18].

Recent technological advances in the field of mass spectrometry (MS) have enabled the development of high-throughput assays, which in a single experiment can identify several thousands of peptides eluted of MHC molecules (reviewed in [19]). Large data sets of such naturally presented peptides have been beneficial to define more accurately the rules of peptide-MHC binding [20–26]. For several reasons, analysis and interpretation of MS eluted ligand data is not a trivial task. Firstly, because any given individual constitutively expresses multiple allelic variants of MHC molecules, thus, the ligands detected by MS are normally a mixture of specificities, each corresponding to a different MHC molecule. Secondly, MHC-II ligands can vary widely in length, and identification of the binding motifs requires a sequence alignment over a minimal binding core. Finally, data sets of MS ligands often contain contaminants and false spectrum-peptide identifications, which add a component of noise to the data. We have earlier proposed a method capable of dealing with all these issues, allowing the characterization of binding motifs and the assignment of probable MHC restrictions to individual peptides in such MS ligand data sets [27, 28].

Because naturally eluted ligands incorporate information about properties of antigen presentation beyond what is obtained from in vitro binding affinity measurements, large MS-derived sets of peptides can be used to generate more accurate prediction models of MHC antigen presentation [20, 21, 25]. As shown recently, generic machine learning tools, such as NNAlign [9, 29], can be readily applied to individual MS data sets, which in turn can be employed for further downstream analyses of the immunopeptidome [30]. The amount of MHC molecules characterized by MS eluted ligand data is, however, still limited. This has led us to suggest a machine learning framework where peptide binding data of both MS and in vitro binding assays are merged in the training of the prediction method [25]. This approach has proven highly powerful for MHC class I, but has not, to the best of our knowledge, been applied to MHC class II.

Undoubtedly, antigen processing plays a critical role in generating CD4+ T cell epitopes presented by MHC class II molecules. It is assumed that endo- and exo-peptidase activities, both before and after binding to the MHC-II molecule, play a key role in the generation and trimming of MHC class II ligands [31, 32]. However, the precise rules of MHC class II antigen processing are poorly understood. Earlier works identified patterns of protein cleavage in HLA-DR ligands; Kropshofer et al. found proline at the penultimate N and C terminal position [33], and Ciudad et al. observed aspartic acid before the cleavage site and proline next to the cut sites in HLA-DR ligands [34]. In contrast, Bird et al. suggested that endolysosomal proteases have a minor and redundant role in peptide selection leading to the conclusion that the effect of processing on the generation of antigenic peptides is “relatively non-specific” [35]. Given this context, it is perhaps not surprising that limited work has been aimed at integrating processing signals into a prediction framework for MHC-II ligands.

In this work, we have analyzed large data sets of MS MHC-II eluted ligands obtained from different research laboratories covering three HLA-DR molecules with the purpose of investigating the consistency in the data, quantifying the differences in binding motifs contained with such MS eluted data compared to traditional in vitro binding data, defining a new machine learning framework capable of integrating information from MS eluted ligand and in vitro binding data into a prediction model for MHC-II peptide interaction prediction, and finally evaluating if inclusion of potential signals from antigen processing is consistent between different data sets and can be used to boost the performance of peptide-MHCII prediction models.

Methods

Data sets

HLA class-II peptidome data were obtained from two recent MS studies. Three data sets corresponding to the HLA-DRB1*01:01: DR1Ph, DR1Pm [26], and DR1Sm [24], two to DRB1*15:01: DR15-Ph and DR15-Pm, and one to the allele DRB5*01:01: DR51 Ph (for details see Table 1). Here, the data sets with subscript h correspond

Table 1 Summary of binding affinity ("Binders") and eluted ligand ("Ligands") data sets used in this work

Binders							
Reference	Source	Allele	L11–19				
DR1 BA	Jensen et al. [36]	DRB1*01:01	9987				
DR15 BA	Jensen et al. [36]	DRB1*15:01	4466				
DR51 BA	Jensen et al. [36]	DRB5*01:01	4840				
Ligands							
Reference	Source	Allele	Cell	Unique	GC	L11–9	Random
DR1 Ph	Ooi et al. [26]	DRB1*01:01	Human	5131	4786	3992	38115
DR1 Pm	Ooi et al. [26]	DRB1*01:01	Mouse	5744	5561	5385	55710
DR1 Sm	Clement et al. [24]	DRB1*01:01	Mouse	3216	3112	2963	30510
DR15 Ph	Ooi et al. [26]	DRB1*15:01	Human	2782	1590	1390	12870
DR51 Ph		DRB5*01:01			1087	989	9315
DR15 Pm	Ooi et al. [26]	DRB1*15:01	Mouse	4810	4486	4229	42030

Binders (upper table): data set reference name ("Reference"), data source ("Source"), MHC restriction ("Allele"), and the amount of sequences in the length range of 11 to 19 amino acids ("L11–19"). Ligands (lower table): data set reference name ("Reference"), data source ("Source"), MHC restriction ("Allele"), cell line species ("Cell"), amount of unique sequences present in the data set before filtering ("Unique") and after filtering with GibbsCluster ("GC"), quantity of sequences in the 11–19mer range ("L11–19"), number of random negatives sequences added for training ("Random"). Note that the split of the Ooi et al. human data (DR15 Pm/DR51 Pm) was made using the GibbsCluster as described in the text

to the data obtained from human cell lines and data sets with the subscript m to the data obtained from human MHC-II molecules transfected into MHC-II deficient mice cell lines. Details on how the data were generated are provided in the original publications. Note that DR15 Ph and DR51 Ph data sets were obtained from a heterozygous EBV-transformed B lymphoblastoid cell line (BLCL), IHW09013 (also known as SCHU), which expresses two HLA-DR molecules, HLA-DRB1*15:01 and HLA-DRB5*01:01 (shortened here with the name DR15/51). The DR1 Ph data set was extracted from a BLCL culture as well (IHW09004). On the other hand, DR1 Pm, DR1 Sm, and DR15 Pm data sets were extracted from HLA transgenic mice, and therefore only cover the human alleles of interest. These cells are treated here as monoallelic.

MHC class II peptide binding affinity data was obtained from previous publications [36] for the alleles DR1 (DRB1*01:01, 9987 peptides), DR15 (DRB1*15:01, 4466 peptides), and DR51 (DRB5*01:01, 4840 peptides).

The MS-derived ligand data sets were filtered using the GibbsCluster-2.0 method with default settings as described earlier [30], to remove potential noise and biases imposed by some data containing multiple binding specificities. The details of the binding affinity (BA) and eluted ligand (EL) data sets are described in Table 1.

NNAlign modeling and architecture

Models predicting peptide-MHC interactions were trained as described earlier using NNAlign [29, 30]. Only ligands of length 11–19 amino acids were included in the training data. Random peptides of variable lengths derived from the non-redundant UniProt database were used as negatives. The same amount of random negatives was used for each length (11 to 19) and consisted of five times the amount of peptides for the most represented length in the positive ligand data set. Positive instances were labeled with a target value of 1, and negatives with a target value of 0. Prior to training, the data sets were clustered using the common motif approach described earlier [37] with a motif length of nine amino acids to generate five partitions for cross-validation.

Two types of model were trained: one with single data type (eluted ligand or binding affinity) input, and one with a mixed input of the two data types. Single models per each data set and allele were trained as previously described with either binding affinity or eluted ligand data as input [30]. All models were built as an ensemble of 250 individual networks generated with 10 different seeds; 2, 10, 20, 40, and 60 hidden neurons; and 5 partitions for cross-validation. Models were trained for 400 iterations, without the use of early stopping. Additional settings in the architecture of the network were used as previously described for MHC class II [30]. Combined models were trained as described earlier [25] with both binding affinity and eluted ligand data as input. Training was performed in a balanced way so that on average the same number of data points of each data type (binding affinity or eluted ligand) is used for training in each training iteration.

Novel modifications were introduced to the architecture of NNAlign to better account for specific challenges associated with MHC class II ligand data. For the network to be able to learn peptide length preferences, a "binned" encoding of the peptide length was introduced, consisting of a one-hot input vector of size nine (one neuron for each of the lengths 11 to 19). In order to guide binding core identification, a burn-in period was introduced with a limited search space for the P1 binding core position. During the burn-in period, consisting of a single learning iteration, only hydrophobic residues were allowed at the P1 binding core anchor position. Starting from the second iteration, all amino acids were allowed at the P1 position (Additional file 1: Figure S1).

NetMHCII and NetMHCIIpan

NetMHCII version 2.3 [36] and NetMHCIIpan version 3.2 [36], peptide to MHC-II binding affinity prediction algorithms were employed in this work as a benchmark comparison for the new proposed model.

Sequence logos

Sequence logos for binding motifs and context information were constructed using Seg2Logo tool using weighted Kulback-Leibler logos and excluding sequence weighting [38]. Amino acids were grouped by negatively charged (red), positively charged (blue), polar (green), or hydrophobic (black).

Performance metrics

In order to assess the performance of our new model, we employed three different and well-known metrics: AUC (area under the ROC curve), AUC 0.1 (area under the ROC curve integrated up to a false positive rate of 10%), and PPV (positive predictive value). AUC is a common performance measurement for predictive models, which takes into account the relationship between true positive rates (TPR) and false positive rates (FPR) for different prediction thresholds. AUC 0.1 is similar to AUC but focuses on the high specificity range of the ROC curve. PPV is here calculated by sorting all predictions and estimating the fraction of true positives with the top *N* predictions, where *N* is the number of positives in the benchmark data set. PPV represents a good metric to benchmark on highly unbalanced data sets like MS-derived elution data, where we have approximately ten times more negatives than positives.

Results

Data filtering and motif deconvolution

We first set out to analyze the different MS data sets of eluted ligands. Data were obtained from two recent publications: Ooi et al. [26] (termed P) and Clement et al. [24] (termed S) covering the HLA-DRB1*01:01, HLA-DRB1*15:01, and HLA-DRB5*01:01 MHC class II molecules. Data were obtained from either human (termed h) or HLA-DR transfected mouse (termed m) cell lines. Using this syntax, DR1 Ph corresponds to the HLA-DRB1*01:01 data from the human cell in the study by Ooi et al. (for more details, see the "Methods" section). Here, we applied the GibbsCluster method with default parameters for MHC class II to both filter out potential noise and to identify the binding motif(s) contained in each data set. The result of this analysis is shown in Fig. 1 and confirms the high quality of the different ligand data sets. In all data sets, less than 7% of the peptides were identified as noise (assigned to the trash cluster), and in all cases, GibbsCluster did find a solution with a number of clusters matching the number of distinct MHC specificities present in a given data set. In this context, the DR15 Ph is of particular interest, since this data set was obtained from a heterozygous cell line expressing two HLA-DR molecules, HLA-DRB1*15:01 and HLA-DRB5*01:01 (shortened here as DR15/51 Ph). Consequently, this data set contains a mixture of peptides eluted from both of these HLA-DR molecules. The GibbsCluster method was able to handle this mixed data set and correctly identified two clusters with distinct amino acid preferences at the anchor positions P1, P4, P6, and P9. Moreover, a comparison of the motifs identified from the different data sets sharing the exact same HLA-DR molecules revealed a very high degree of overlap, again supporting the high accuracy of both the MS eluted ligand data and of the GibbsCluster analysis tool.

Training prediction models on MHC class II ligand data

After filtering and deconvolution with GibbsCluster, MHC peptide binding prediction models were constructed for each of the six data sets corresponding to the majority clusters in Fig. 1. Models were trained using the NNAlign framework as described in the "Methods" section. The eluted ligand data sets (EL) were enriched with random natural peptides labeled as negatives, as described in the "Methods" section. Likewise, models were trained and evaluated on relevant and existing data sets of peptide binding affinities (BA) obtained from the IEDB [15, 36], as described in the "Methods" section. These analyses revealed a consistent and high performance for the models trained on the different eluted ligand data sets (Table 2). In accordance with what has been observed earlier for MHC class I [25], the overall cross-validated performance of models trained on binding affinity data is lower than that of models trained on eluted ligand data. Note that this observation is expected due the very different nature of the binding affinity and eluted ligand data sets: eluted ligand data are highly unbalanced, categorized, and prefiltered to remove ligands not matching the consensus binding motif.

The binding motifs captured by the different models are shown in Fig. 2. As evidenced by identical anchor positions (P1, P4, P6, and P9) and virtually identical anchor residues, highly consistent motifs were obtained from the same HLA-DR molecules irrespective of the source of the peptide (i.e., whether they were obtained from human or mouse cells, or from different laboratories). This observation to a high degree extended to the

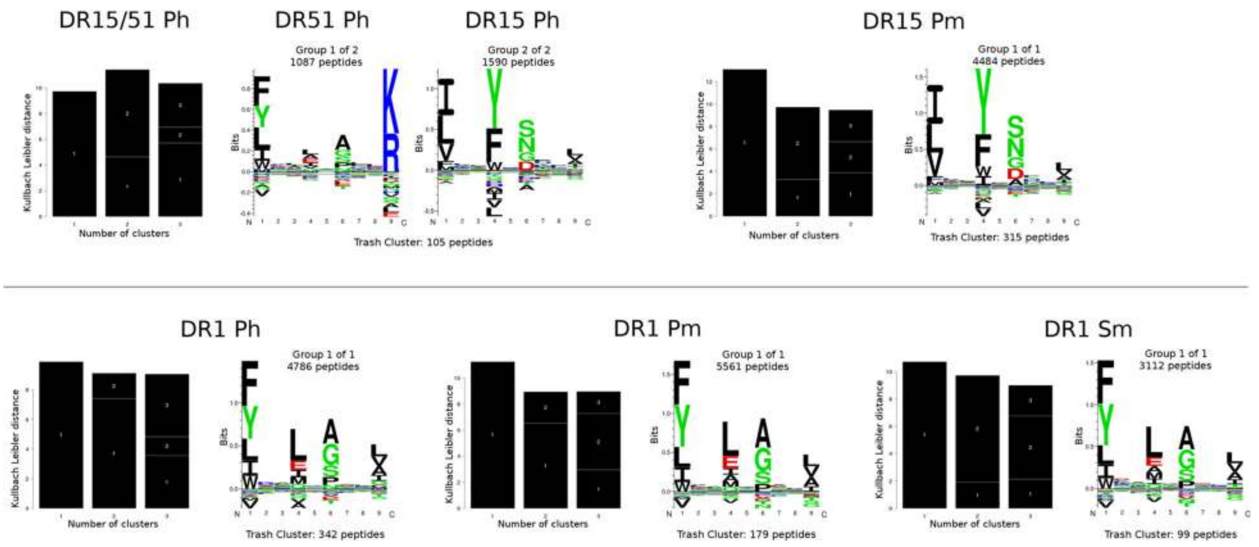

Fig. 1 GibbsCluster output for the five eluted ligand data sets employed in this work. For each set, the Kullback-Leibler distance (KLD) histogram (black bars) is displayed, which indicates the information content present in all clustering solutions (in this case, groups of one to three clusters) together with the motif logo(s) corresponding to the maximum KLD solution. The upper row gives the results for the DR15/51 data sets; the lower row for the DR1 data sets. Note that DR15 Ph was obtained from a cell line which expresses two HLA-DR molecules, HLA-DRB1*15:01 and HLA-DRB5*01:01 (DR15/51)

motifs obtained from binding affinity data, although we did observe subtle, but consistent, differences between the binding motifs derived from eluted ligand and peptide binding affinity data, exemplified for instance by the preference for E at P4 and for D at P6 in the eluted ligand motifs for DR1 and DR15, respectively. Such preferences are absent from the motifs derived from the peptide binding affinity data. To quantify differences and statistically compare the core logos shown in Fig. 2, we performed a correlation comparison of the amino acid frequency matrices of the binding motif obtained from the different models. To this end, we extracted the amino acid frequencies from the binding motifs displayed in Fig. 2, and next did a bootstrapped correlation analysis comparing the amino acid frequency values at the four anchor positions (P1, P4, P6, and P9) of the binding core between all pairs of motifs. The results of this analysis are given in Additional file 1: Figure S2 and Table S1 and show (as expected from the logo plots of Fig. 2) that the different motifs obtained from eluted ligand data for a given HLA-DR molecule are all highly similar (and statistically indistinguishable, $P > 0.05$, Student T test), whereas motif obtained from binding affinity data are significantly different ($P < 0.001$, Student T test) from those obtained from eluted ligand motifs.

Table 2 Cross-validation performance of models trained on binding affinity (BA) or eluted ligand (EL) data

BA single models			EL single models			
Training set	AUC	AUC 0.1	Training set	AUC	AUC 0.1	PPV
DR1 BA	0.84	0.374	DR1 Ph	0.96	0.819	0.773
			DR1 Pm	0.986	0.888	0.815
			DR1 Sm	0.977	0.851	0.794
DR15 BA	0.843	0.287	DR15 Ph	0.987	0.901	0.85
			DR15 Pm	0.989	0.917	0.859
DR51 BA	0.846	0.38	DR51 Ph	0.961	0.759	0.717

For BA ("BA single models"), the training performance is reported in terms of AUC and AUC 0.1. For EL ("EL single models"), values for AUC, AUC 0.1, and PPV are displayed. For references on the training sets names and compositions, refer to Table 1. For information regarding the performance metrics, see the "Performance metrics" section in the "Methods" section

Training a combined prediction model on MHC-II binding affinity and ligand elution data

Earlier work on MHC class I has demonstrated that the information contained within eluted ligand and peptide binding affinity data is, to some degree, complementary and that a prediction model can benefit from being trained integrating both data types [25]. Here, we investigate if a similar observation could be made for MHC class II. As proposed by Jurtz et al., we extended the NNAlign neural network model to handle peptides from both binding affinity and elution assays. In short, this is achieved by including an additional output neuron to the neural network prediction model allowing one prediction for each data type. In this setup, weights are shared between the input and hidden layer for the two input types (binding affinity and eluted ligand), whereas the weights connecting the hidden and output layer are specific for each input type. During neural network training, an example is randomly selected from either data set and submitted to forward and back propagation,

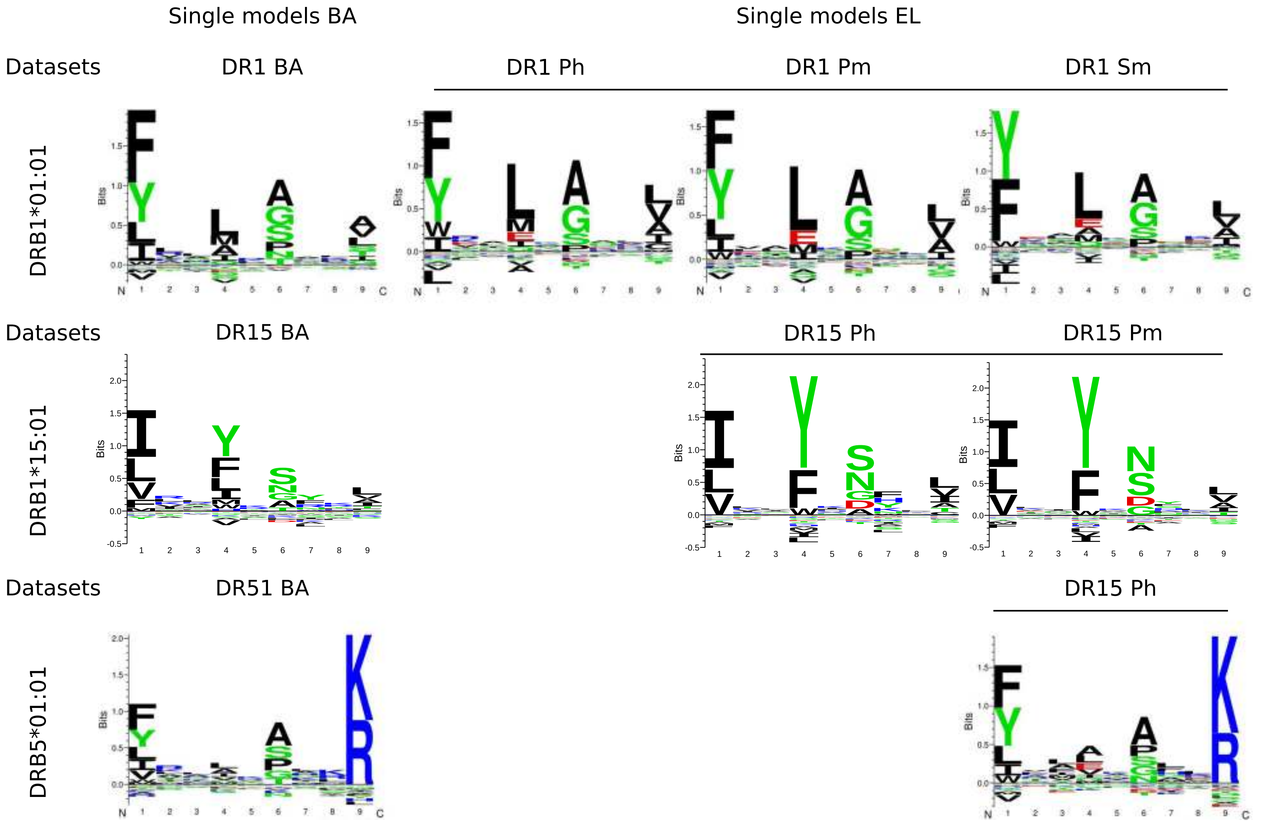

Fig. 2 Binding preferences learned by the single NNAlign [29] models trained on binding affinity (BA) or eluted ligand (EL) data. In the top row, motifs for the DRB1*01:01 allele are shown, with overlined logo plots (right) corresponding to models trained on EL data, and the non-overlined logo (left) corresponding to the BA trained model. Similarly, binding motifs for DRB1*15:01 and DRB5*01:01 are displayed in the middle and bottom row respectively, with overlined logos (right) also indicating the EL-trained model preferences, and the non-overlined logo plot (left) indicating the BA preference. Logos were constructed from the predicted binding cores in the top 1% scoring predictions of 900.000 random natural peptides for BA and from the top 0.1% scoring predictions for EL

according to the NNAlign algorithm. The weight sharing allows information to be transferred between the two data types and potentially results in a boost in predictive power (for more details on the algorithm, refer to [25]).

Models were trained and evaluated in a fivefold cross-validation manner with the same model hyper-parameters that were used for the single data type model. Comparing the performance of the single data type (Table 2), to the multiple data type models for the different data sets (Table 3), a consistent improvement in predictive performance was observed when the two data types were combined. This is the case, in particular, when looking at the PPV performance values. Here, the combined model in all cases has improved performance compared to the single data type model. This is in line to what we have previously observed for MHC class I predictions [25].

Constructing the binding motif captured by the different combined models (see Additional file 1: Figure S3) confirmed the findings from the single data type model

Table 3 Cross-validation performance for the combined NNAlign models, trained on both binding affinity (BA) and eluted ligand (EL) data

Training set		BA prediction		EL prediction		
BA	EL	AUC	AUC 0.1	AUC	AUC 0.1	PPV
DR1 BA	DR1 Ph	0.845	0.385	0.966	0.823	0.781
	DR1 Pm	0.843	0.376	0.987	0.893	0.826
	DR1 Sm	0.843	0.381	0.98	0.867	0.814
DR15 BA	DR15 Ph	0.844	0.288	0.987	0.908	0.855
	DR15 Pm	0.846	0.294	0.99	0.917	0.86
DR51 BA	DR51 Ph	0.848	0.389	0.956	0.749	0.74

Training set refers to the data set used to train the given model (BA indicated binding affinity and EL eluted ligand data). For references on the training sets names and compositions, refer to Table 1. Cross-validated performance values are reported as AUC, AUC 0.1, and PPV. For more details on these measures, refer to the "Methods" section. Note that minor variations in the BA performance values for the same molecule are due to the differences in the data partitioning in the fivefold cross-validation setup in each case (displayed in Fig. 2), with clearly defined and consistent binding motifs in all cases, and with subtle differences in the preferred amino acids at the anchor positions between motifs derived from the binding affinity and eluted ligand output value of the models.

We next turned to the issue of accurately predicting the preferred length of peptides bound to the different HLA-DR molecules. The MS eluted ligand data demonstrated a length preference for the two MHC class II molecules centered on a length around 14–16. Current prediction models such as NetMHCII and NetMHCIIpan are not able to capture this length preference and have in general a bias of assigning higher prediction values to longer peptides (data not shown). We have earlier demonstrated that including information about the peptide length in a framework integrating MS eluted ligand and peptide binding affinity data allows the model to capture the length preference of the two data types [25]. Applying a similar approach to the MHC class II data, we obtain the results shown in Fig. 3, confirming that also for class II the models are capable of approximating the preferred length preference of each molecule.

Lastly, we performed an evaluation across data sets to confirm the robustness of the results obtained and to reveal any unforeseen signal of performance overfitting. For each data set, we used the two-output model trained above to predict the other ligand data sets of the same allotype. Prior to evaluation, all data with a 9mer overlap between training and evaluation sets were removed. We observed that, in all cases, models trained on a specific data set retained high predictive performance for the prediction of ligands of the same allotype derived from a different experiment (Table 4). These results confirm the high reproducibility of the motifs across different cell lines, as well as the robustness of the prediction models derived from individual data sets.

Signals of ligand processing

Having developed improved models for prediction of MHC class II ligand binding, we next analyzed whether the models could be used to identify signals of antigen processing in the MS eluted ligand data sets. We hypothesized that information concerning antigen processing should be present in the regions around the N and C termini of the ligand. These regions comprise residues that flank the MHC binding core called peptide flanking regions (PFRs) and residues from the ligand source protein sequence located outside the ligand (see lower part of Fig. 4 for a schematic overview).

We speculate that the signals of antigen processing depend, to some degree, on the length of the PFRs on each side of the binding core. MHC-II ligands are cut and trimmed by exopeptidases, which operate according to specific motifs in prioritizing cleavage sites. However, in

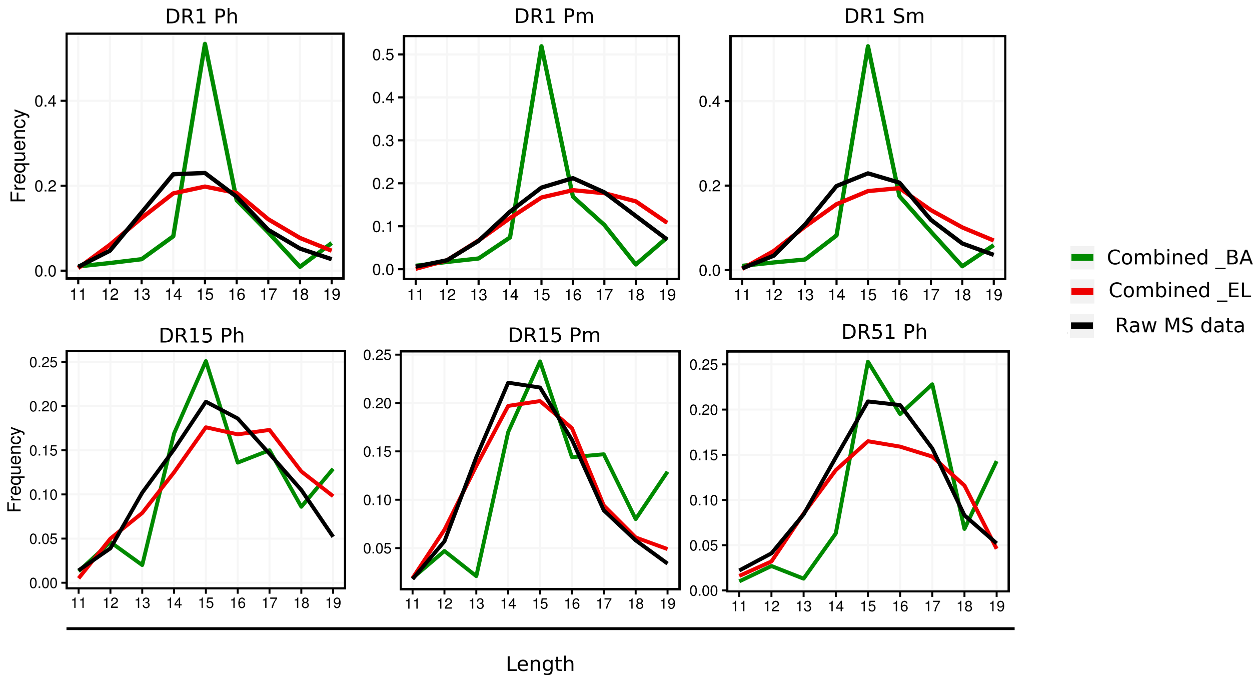

Fig. 3 Peptide length preferences learned by the six models trained on binding affinity (BA) and eluted ligand (EL) combined data. For each model, green traces represent the length histogram of the top 1% scoring predictions for the BA output neuron, on a prediction data set composed of one million random peptides; red traces refer to the length histogram of the top 0.1% scoring predictions for the EL output neuron, on the same prediction set; black traces indicate the length distribution of the raw MS data

Table 4 Independent evaluation of eluted ligand data set in terms of AUC 0.1

Training set		Eval set	BA single	BA combined	EL single	EL combined	NetMHCIIpan	NetMHCII
BA	EL							
DR1 BA	DR1 Ph	DR1 Pm	0.57	0.644	0.839	0.849	0.562	0.585
		DR1 Sm	0.573	0.648	0.81	0.813	0.576	0.578
	DR1 Pm	DR1 Ph	0.498	0.557	0.757	0.754	0.478	0.469
		DR1 Sm	0.514	0.549	0.665	0.669	0.519	0.511
	DR1 Sm	DR1 Ph	0.506	0.556	0.704	0.707	0.481	0.478
		DR1 Pm	0.568	0.636	0.824	0.835	0.54	0.57
DR15 BA	DR15 Ph	DR15 Pm	0.584	0.672	0.869	0.869	0.427	0.58
	DR15 Pm	DR15 Ph	0.615	0.71	0.888	0.889	0.459	0.583

"Training set" refers to the data sets used to train the given model (BA indicates binding affinity and EL eluted ligand data). For references on the training sets names and compositions, refer to Table 1. Cross-validated performance values are reported as AUC 0.1. "Eval set" refers to the independent eluted ligand data set from the same allotype used for evaluation. "BA single" or "EL single" refers to model trained on single data types (BA or EL respectively). "BA combined" or "EL combined" refers to the eluted ligand prediction output or binding affinity output of models trained on both data types. NetMHCIIpan or NetMHCII refers to predictions made using the NetMHCIIpan 3.2 [36], and NetMHCII 2.3 [36] publicly available prediction methods

the case of short PFRs, the MHC hinders access of the protease to the ligand, hence preventing trimming of the residues in close proximity to the MHC [39, 40]. For this reason, we expect to observe cleavage motifs only in peptides with sufficiently long PFRs, where the end-of-the-trimming signal is given by the peptide sequence rather than by MHC hindrance. To validate this hypothesis, we identified the PFRs of the ligands in the DR15 Pm EL data set, as well as three "context" residues found immediately upstream or downstream of the ligand in its source protein. To avoid over-estimation of the performance, the binding core was identified from

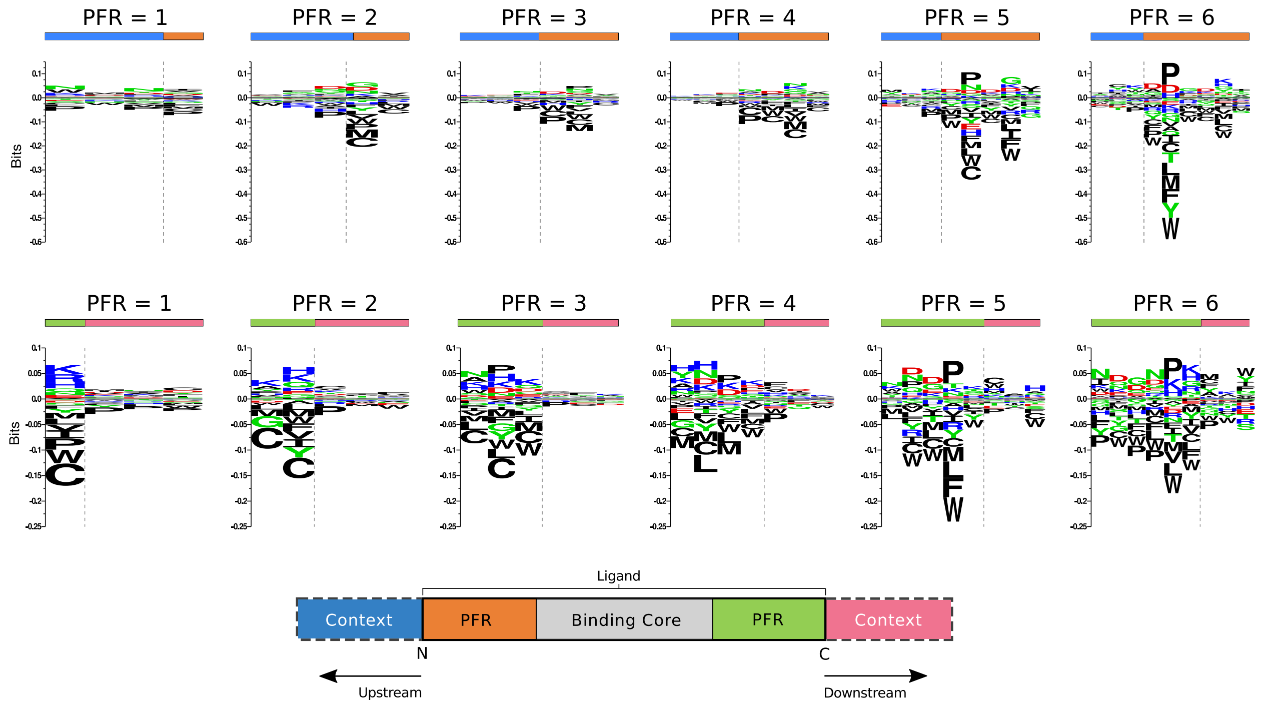

Fig. 4 Processing signals found at N and C terminus positions in the DR15 Pm data set (located at upstream and downstream regions, respectively), grouped by peptide flanking region (PFR) length. For the upstream part of the ligands (top row), the processing signal is always centered at the N terminal position, extending three positions beyond the cleavage site (upstream "context," symbolized as blue bars) and one to six positions towards the binding core, depending on the PFR length (orange bars). For the downstream region (bottom row), the disposition of elements is mirrored: the proposed processing signal is centered at C terminus and extends three positions beyond the cleavage site (downstream "context" region, pink bars) and one to six positions towards the binding core (green bars), depending on the PFR length. Amino acid background frequencies were calculated using the antigenic source protein of all the ligands present in the data set. Motifs were generated using Seq2logo, as described in the "Methods" section

the cross-validated eluted ligand predictions of the two-output model. The ligands were split into groups depending on the length of the C and N terminal PFRs, and sequence logos were generated for each ligand subset using Seq2Logo (Fig. 5).

The results displayed in Fig. 4 clearly confirm the important role of the MHC in shaping the processing signal. For both the N and C terminal data sets, we observe a clear enrichment of proline (P) at the second position from the ligand terminals only for data sets where the PFR is longer than two amino acids. This observation is confirmed from the reanalysis of a data set of peptide to HLA-DR complexes from the Protein Data Bank (PDB) previously assembled for benchmarking the accuracy for MHC-II binding core identification [41]. On this PDB data set, 29% of the entries with a N-terminal PFR longer than two amino acids contain a proline at the second position from the N terminal, and 38% of the entries with a C-terminal PFR longer than two amino acids contain a proline at the second position from the C terminal (data not shown). On the other hand, none of the bound peptides with N-terminal PFR shorter or equal than two amino acids contain a proline at the second position from N-terminal, and only 8% of peptides with C-terminal PFR shorter or equal than two amino acids exhibit a proline at the second position from the C-terminal.

To summarize these observations and construct a global motif of the processing signal, we combined the first three C and N terminal residues from all ligands with PFR length larger than two, together with the corresponding three source protein context residues at either C or N terminal side of the ligand. The processing signal at the N and C termini from DR15 Pm is shown in Fig. 5; processing motifs for all other data sets can be found in Additional file 1: Figure S4.

The processing motif confirms the strong preference for proline at the second but last position in the ligand at both N and C termini, as well as a clear signal of depletion of other hydrophobic amino acid types towards the terminals of the ligand. This cysteine depletion in the PFR is likely to be a technological artifact, as cysteines have previously been shown to be underrepresented in MS-derived peptide data sets [20, 42]. Note also that this depletion is only observed in the PFRs and not in the context residues neighboring the N and C termini. From this figure, it is also clear that processing signals present in the neighborhood (indicated as “context” in Fig. 5) of the ligand are very weak. Similar amino acid preferences were obtained in the processing motifs from the other data sets (Additional file 1: Figure S4).

Next, we investigated to what degree the processing signal was consistently identified in all data sets. To do this, the similarity between any two processing matrices was estimated in terms of the Pearson’s correlation coefficient (PCC) between the two vectors of 6*20 elements (6 positions and 20 amino acid propensity scores at each position). The result of this analysis is shown in Fig. 6 in terms of a heatmap (the processing matrices from each data set are included in Additional file 1: Figure S5).

Figure 6 exhibits a clear positive correlation between the processing motif from all the data sets involved. The mean PCC score for the matrices in Fig. 6 was 0.77 for upstream and 0.73 for downstream, with the lowest PCC = 0.59 (for the DR1 Sm and DR1 Ph pair, upstream) and the maximum PCC = 0.89 (for DR15 Pm and DR1 Ph, upstream). These results suggest that the processing signals captured are, to a large degree, MHC- and even species-independent: the correlation between the two human and mouse data sets is as high as the correlation between any two data sets within the same species. To ensure that the observed correlation is not related to

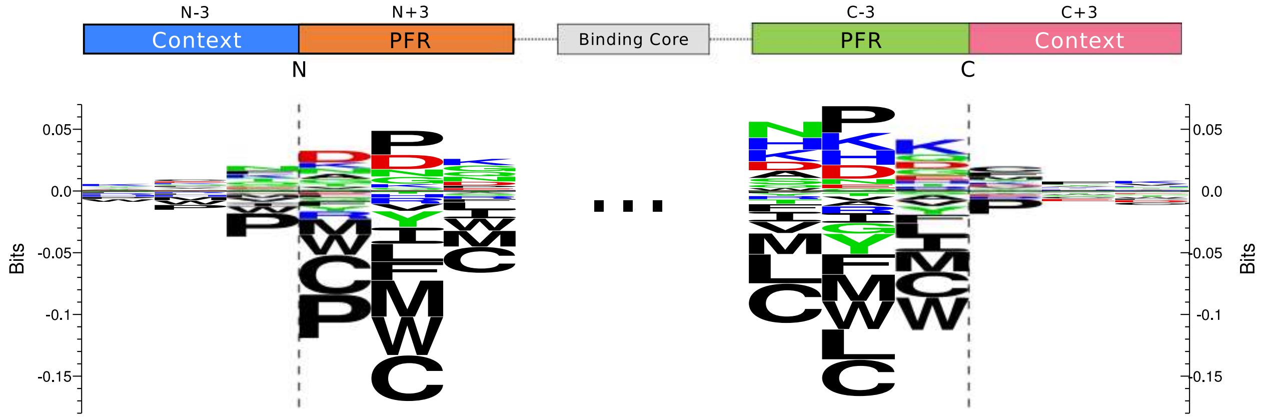

Fig. 5 Processing signals located at N and C terminal regions in the DR15 Pm data set. For each region, all ligands with PFR length lower than 3 were discarded. Then, the logos were constructed as described in the text by selecting the closest three PFR and context residues neighboring the N and C termini. For additional details on processing signal construction, refer to Fig. 4

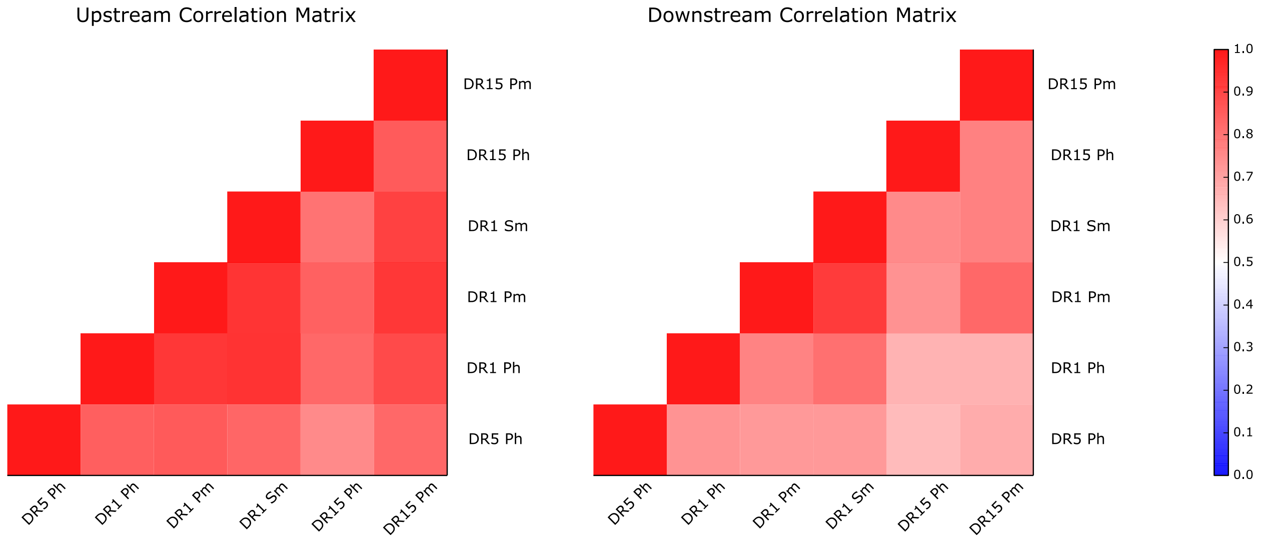

Fig. 6 Correlation between processing signals found in the six different data sets employed in this work, for upstream and downstream regions. Each matrix entry displays the Pearson correlation coefficient (PCC) value of two data sets under study. A PCC value of one corresponds to a maximum correlation, while a PCC value of zero means no correlation. Processing signals used in this figure were generated as explained in Fig. 5. All observed PCC values are statistically different from random ($P <$ 0.001, exact permutation test)

MS-derived cysteine depletion, we generated the same correlation matrices removing the cysteine contribution and observed no major differences (Additional file 1: Figure S6). These results thus strongly suggest that the observed signals are related to antigen processing.

Incorporating ligand processing into a combined predictor

Having identified consistent signals associated with antigen processing, we next investigated whether these signals could be integrated into one model to boost predictive performance. The processing signals were incorporated into the machine learning framework by complementing the encoding of each ligand with the 3 N terminal context, 3 N terminal peptide, 3 C terminal context, and 3 C terminal peptide residues (see Fig. 5). For peptide binding affinity data, the context information was presented to the neural networks with three wildcard amino acids “XXX”, corresponding to a vector of zeros. Two models were trained for each one of the allotypes considered in this work: one model including and one excluding the context information, both allowing integration of binding affinity and eluted ligand data. Prior to training, the complete set of data (binding affinity and eluted ligands for all three MHC-II molecules) was split into five partitions using the common motif approach as described in the “Methods” section. All model hyper-parameters were identical to the ones used earlier. The result of this benchmark is shown in Table 5 and confirms that the inclusion of context leads to a consistently improved predictive power of the models for all three data sets.

As an example of the processing signal captured by a model trained including context information, we constructed sequence motifs of the top 1% highest scoring peptides from a list of one million random natural peptides of length 10–25 and their context, for a combined model trained on the DR15 Pm data set (Additional file 1: Figure S7). As expected, the motif contained within the N and C terminal peptide flanks and context is close to identical to the motif described in Fig. 5.

T cell epitope prediction using the combined models

Having observed how prediction of naturally processed MHC ligands benefited from implementing ligand context features, we next wanted to evaluate if a similar gain could be observed when predicting T cell epitopes. We downloaded all available epitopes of length 14 to 19 (included) from the IEDB, for the molecules DRB1*01:01, DRB1*15:01, and DRB5*01:01. After filtering out entries

Table 5 Cross-validation performance for combined NNAlign models trained on single-allele data sets, with and without context information

	Without context		With context		
Allele	AUC 0.1	PPV	AUC 0.1	PPV	*P* value
DRB1*01:01	0.874	0.824	0.893	0.839	< 0.0001
DRB1*15:01	0.931	0.875	0.947	0.892	< 0.0001
DRB5*01:01	0.805	0.76	0.818	0.782	0.0368

“Allele” refers to the combination of all data sets for that given allele used to train the model. Cross-validated performance values are reported as AUC 0.1 and PPV. *P* values were estimated using bootstrapping. For more details on these measures, refer to the “Methods” section with post translational modifications and entries lacking information about the source protein IDs, a total of 557, 411, and 114 epitopes remained for the three DR molecules, respectively. First, we evaluated this panel of epitopes in a conventional way: digesting the epitope-source protein into overlapping peptides with the length of the epitope, predicting the peptides using the different models, and calculating the AUC (area under the receiver operator curve) per source protein-epitope pair, taking peptides identical to the epitope as positives and all other peptides in the source protein as negatives. We excluded from the evaluation data sets negative peptides that shared a common motif of nine amino acids with the epitope. Four methods were included in this benchmark: EL (the eluted ligand prediction value from the model trained on the combined data without context information), EL + context (the eluted ligand prediction value from the model trained on the combined data including context signals), NetMHCII (version 2.3), and NetMHCIIpan (version 3.2). This analysis shows, in line with what we observed earlier for the eluted ligand benchmarks, a consistent improved performance of the EL model compared to both NetMHCII and NetMHCIIpan (Fig. 7a).

The benchmark however also demonstrates a substantial drop in predictive power of the EL model when incorporating the context processing signal (EL + context). This drop is however expected since the mapped T cell epitope boundaries are not a product of natural antigen processing and presentation, but rather result from screening of overlapping peptides from a candidate antigen, or by peptides synthesized based on the results of MHC peptide binding predictions and/or in vitro binding assays. As a consequence, the N and C terminal boundaries of such epitope peptides do not necessarily contain the processing signal obtained from naturally processed ligands. However, given that the epitope was demonstrated to bind to the T cell originally induced towards a naturally processed ligand, we can assume that the sequence of the validated epitope and the original (but unknown to us) naturally processed ligand share an overlap at least corresponding to the MHC-II binding core of the validated epitope. Following this reasoning, we redefined the epitope benchmark as follows. First, we predicted a score for all 13–21mer peptides within a given source protein using the EL or EL + context models. Next, we digested the source protein into overlapping peptides of the length of the epitope and assigned a score to each of these peptides corresponding to the average prediction score of all 13–21mer peptides sharing a 9mer or more overlap with the given peptide (models where the max score was assigned were also considered, but gave consistently lower predictive performance, data not shown). Finally, we calculated as before an AUC value for the epitope-source protein pair taking peptides equal to the epitope as positives and all other peptides as negatives excluding from the

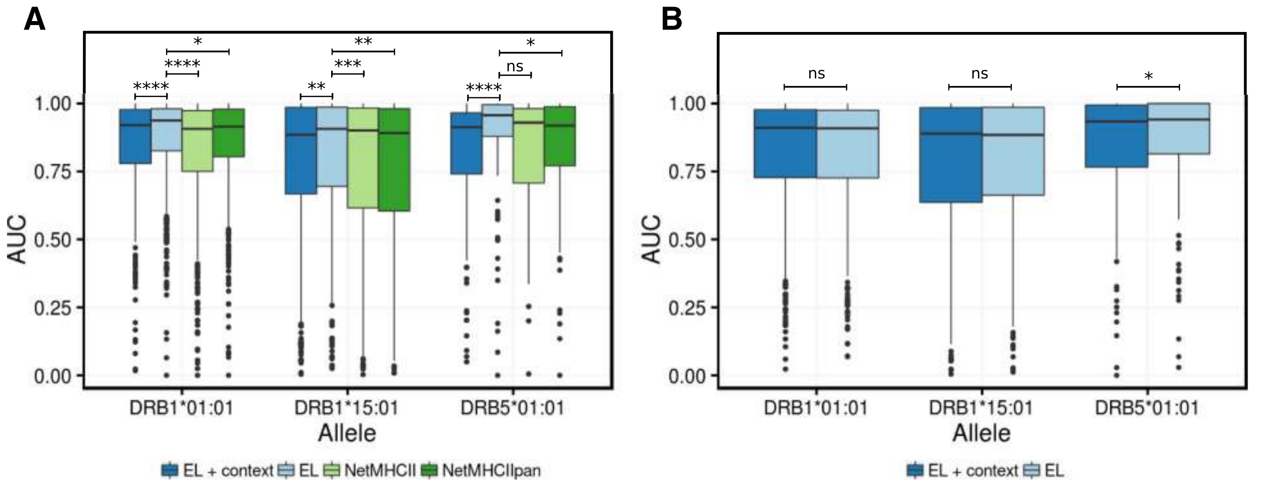

Fig. 7 Predictive performance on a panel of CD4+ T cell epitopes. The boxplots represent the distribution of AUC values over all epitope evaluation data sets restricted to a given allele comparing the different models. Middle lines in boxes correspond to median values. The height of the box represents 50% of the data. Whiskers represent 1.5 quartile range (QR) of data, and dots represent outliers of 1.5 of QR. *P* significance is calculated from Wilcoxon test. $^{ns}P > 0.05$, $^{*}P \leq 0.05$, $^{**}P \leq 0.01$, $^{***}P \leq 0.001$, $^{****}P \leq 0.0001$. In both benchmarks, an AUC value was calculated for each epitope/source protein pair by considering peptides identical to the epitope as positives and all other peptides as negatives excluding peptides with an overlap of at least nine amino acids to the epitope. **a** Comparison of the combined models developed in this study with context information (EL + context) and without context (EL) to current state-of-the-art prediction methods trained on binding affinity data only (NetMHCII-2.3 and NetMHCIIpan-3.2). **b** Comparison of EL + context and EL in a benchmark where the epitope evaluation set was constructed using the evaluation strategy accounting for ligand preference described in the text

evaluation set negative peptides sharing a common motif of nine amino acids with the epitope. The benchmark shows a comparable performance of the EL + context method vs EL method for the alleles analyzed in the study (Fig. 7b). Possible reasons for this lack of improved performance of the EL + context model are discussed below.

Discussion

Peptide binding to MHC II is arguably the most selective step in antigen presentation to CD4+ T cells. The ability to measure (and predict) specific CD4+ responses is crucial for the understanding of pathological events, such as infection by pathogens or cancerous transformations. Recent studies have also highlighted a potential role for CD4+ T cells for the development of cancer immunotherapies [43–45]. Characterizing peptide to MHC-II binding events has been a focal point of research over the last decades. Large efforts have been dedicated in conducting high-throughput, in vitro measurements of peptide MHC II interactions [46–48], and these data have been used to develop methods capable of accurately predicting the interaction of peptides to MHC II molecules from the sequence alone [29, 41, 49, 50]. While these approaches have proven highly successful as guides in the search for CD4 epitopes [51, 52], a general conclusion from these studies is that MHC II in vitro binding affinity (whether measured or predicted) is a relatively poor correlate of immunogenicity [53]. In other words, peptide binding affinity to MHC II is a necessary but not sufficient criterion for peptide immunogenicity. The same situation holds for MHC class I presented epitopes. Here, however, peptide binding to MHC I is a very strong correlate to peptide immunogenicity and can be used to discard the vast majority (99%) of the irrelevant peptide space while maintaining an extremely high (> 95%) sensitivity for epitope identification [25]. For MHC II, recent studies suggest that the corresponding numbers fall in the range 80% specificity and 50% sensitivity [36]. For these reasons, we suggest that other features than MHC II in vitro binding affinity may be critical for MHC II antigen presentation. Based on six MS MHC II eluted ligand data sets, we have here attempted to address and quantify this statement.

Firstly, we have demonstrated that the MS MHC II eluted ligand data sets employed in this work (generated by state-of-the-art technologies and laboratories) are of very high quality, with low noise levels and allowing very precise determination of MHC II binding motifs. Overall, the obtained binding motifs show overlap with the motifs identified from in vitro binding affinity data, with subtle differences at well-defined anchor positions.

Secondly, we demonstrated that high accuracy prediction models for peptide MHC II interaction can be constructed from the MS-derived MHC II eluted ligand data, that the accuracy of these models can be improved by training models integrating information from both binding affinity and eluted ligand data sets, and that these improved models can be used to identify both eluted ligands and T cell epitopes in independent data sets at an unprecedented level of accuracy. This observation strongly suggests that eluted ligand data contain information about the MHC peptide interaction that is not contained within in vitro binding affinity data. This notion is further supported by the subtle differences observed in the binding motifs derived from eluted ligand and in vitro binding affinity data. Similar observations have been made for MHC class I [20, 25]. We at this point have no evidence for the source of these differences, but a natural hypothesis would be that they are imposed by the presence of the molecular chaperones (such as HLA-DM) present in the eluted ligand but absent from in vitro binding assays. An alternative explanation could be that the eluted peptide ligands reflect peptide-MHC class II stability rather than affinity: something that would imply that stability is a better correlate of immunogenicity than affinity [54].

Thirdly, we analyzed signals potentially associated with antigen processing. Antigen-presenting cells employ multiple mechanisms to acquire and process antigens, making use of multiple proteases to digest the internalized proteins [55]. It is likely that the processing signals we observed are a combination of the cleavage specificities of several proteases operating in different stages of the presentation pathway. Looking for consistent patterns, we postulate that such processing signal should be influenced by the relative location of the peptide binding core compared to the N and C terminal of the given ligand. This is because the MHC II molecule may hinder the access of the protease, thus preventing trimming of the residues in close proximity to the MHC [39]. Investigating the data confirmed this hypothesis, and a relatively weak but consistent processing signal (with a preference for prolines at the second amino acid position from the N and C terminal of the ligand) was observed for ligands where the length of the region flanking the binding core was three amino acids or more. This observation was found consistently in all data sets independent of MHC II restriction and host species (human or mouse).

Lastly, we integrated this information associated with antigen processing into a machine learning framework and demonstrated a consistently improved predictive performance not only in terms of cross-validation but also when applied to independent evaluation data sets covering naturally processed MHC eluted ligands. However, we do not observe an improvement of the extended model for prediction of validated T cell epitopes. There are several possible reasons for this. In the first place, it is possible that epitope data have a bias towards current MHC class II binding prediction and/or in vitro binding assay methods, since researchers could use these tools to select which peptides to include in a T cell epitope screening or to define the MHC restriction element for a given positive epitope. Secondly, we have attempted a very simple strategy to assign a prediction score to each epitope. It might be that the conclusion is altered if alternative, more sophisticated mapping strategies were used. Thirdly, the reason might be biological: the antigen processing pathways predominantly utilized in cell lines used for ligand elution experiments which lead to the motifs we identified might not be the only ones generating T cell epitopes in vivo, where, e.g., cross-presentation might play a role. Finally, our prediction model still does not capture all properties that could determine T cell epitope immunogenicity. For example, HLA-DM and DO clearly have a role in regulating which peptides can be loaded onto MHC II [56, 57]; however, their contribution cannot be modeled based on existing data. Also, T cells themselves impose a level of antigen selection through the interaction between the TCR and the peptide-MHC complex. While approaches for peptide-MHC targets of TCR are beginning to appear [58], it is still unclear how they can be integrated in high-throughput approaches for the prediction of T cell epitopes. Future work is needed to disentangle these questions.

Conclusions

We have demonstrated how integrating MHC class II in vitro binding and MS eluted ligand data can boost the predictive performance for both binding affinity, eluted ligand, and T cell epitope predictions. To the best of our knowledge, we have also demonstrated for the first time how MHC II eluted ligand data can be used to extract signals of antigen processing and how these signals can be integrated into a model with improved predictive performance.

Our work is limited to three HLA-DR molecules, but the framework can be readily extended to additional molecules, once sufficient data become available. Also, it may become achievable to construct a pan-specific predictor as has been shown earlier for MHC class I [25], enabling predictions for any MHC molecule of known sequence.

Additional file

Additional file 1: Figure S1. Binding preferences learnt by each individual neural network. **Figure S2.** Pearson correlation coefficient (PCC) heatmaps comparing core motifs learnt by the single EL and BA models. **Table S1.** Standard deviations for **Figure S2. Figure S3.** Binding preferences learned by the combined NNAlign models. **Figure S4.** Processing signals located at N and C terminal regions. **Figure S5.** Heatmap representation of the processing signal's Position-Specific Scoring Matrix (PSSM). **Figure S6.** Correlation between processing signals found in all six data sets. **Figure S7.** Processing signal learned by the NNAlign model trained on DR15 Pm data set. (PDF 3717 kb)

Abbreviations

AUC 0.1: Area under the ROC curve integrated up to false positive rate of 10%; AUC: Area under the ROC curve; BA: Binding affinity data; EL: Eluted ligand data; KLD: Kullback-Leibler distance; MHC-II: Major histocompatibility complex class-II; MS: Mass spectrometry; PCC: Pearson correlation coefficient; PFRs: Peptide flanking regions; pMHCII: Peptide-MHC-II complexes; PPV: Positive predictive value; PSSM: Position-specific scoring matrix

Funding

This work was supported in part by the Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272201200010C, and by the Science and technology council of investigation (CONICET-Argentina).

Availability of data and materials

The eluted ligand data sets analyzed during the current study are available in the IEDB repository, http://www.iedb.org/reference/1031891 and http://www.iedb.org/reference/1030084.

Authors' contributions

MN designed the study. CB, BA, MA, and MN performed the experiments and statistical analysis. CB, BA, SP, AS, BP, SB, MA, and MN analyzed and interpreted the data and wrote the paper. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, CP1650 San Martín, Argentina. [2]Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, 9420 Athena Circle, La Jolla, CA 92037, USA. [3]Department of Immunology and Microbiology, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark. [4]Department of Bio and Health Informatics, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark.

Received: 25 July 2018 Accepted: 30 October 2018
Published online: 16 November 2018

References

1. Rudolph MG, Stanfield RL, Wilson IA. How TCRs bind MHCs, peptides, and coreceptors. Annu Rev Immunol. 2006;24:419–66.
2. Kim A, Hartman IZ, Poore B, Boronina T, Cole RN, Song N, et al. Divergent paths for the selection of immunodominant epitopes from distinct antigenic sources. Nat Commun. 2014;5:5369.

3. Sette A, Adorini L, Colon SM, Buus S, Grey HM. Capacity of intact proteins to bind to MHC class II molecules. J Immunol. 1989;143:1265–7.
4. Andreatta M, Jurtz VI, Kaever T, Sette A, Peters B, Nielsen M. Machine learning reveals a non-canonical mode of peptide binding to MHC class II molecules. Immunology. 2017;152:255–64.
5. Lovitch SB, Pu Z, Unanue ER. Amino-terminal flanking residues determine the conformation of a peptide-class II MHC complex. J Immunol. 2006;176:2958–68.
6. Arnold PY, La Gruta NL, Miller T, Vignali KM, Adams PS, Woodland DL, et al. The majority of immunogenic epitopes generate CD4+ T cells that are dependent on MHC class II-bound peptide-flanking residues. J Immunol. 2002;169:739–49.
7. Carson RT, Vignali KM, Woodland DL, Vignali DA. T cell receptor recognition of MHC class II-bound peptide flanking residues enhances immunogenicity and results in altered TCR V region usage. Immunity. 1997;7:387–99.
8. Godkin AJ, Smith KJ, Willis A, Tejada-Simon MV, Zhang J, Elliott T, et al. Naturally processed HLA class II peptides reveal highly conserved immunogenic flanking region sequence preferences that reflect antigen processing rather than peptide-MHC interactions. J Immunol. 2001;166:6720–7.
9. 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. PLoS One. 2011;6:e26781.
10. Hammer J, Valsasnini P, Tolba K, Bolin D, Higelin J, Takacs B, et al. Promiscuous and allele-specific anchors in HLA-DR-binding peptides. Cell. 1993;74:197–203.
11. Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, et al. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nat Biotechnol. 1999;17:555–61.
12. Roche PA, Cresswell P. High-affinity binding of an influenza hemagglutinin-derived peptide to purified HLA-DR. J Immunol. 1990;144:1849–56.
13. Hall FC, Rabinowitz JD, Busch R, Visconti KC, Belmares M, Patil NS, et al. Relationship between kinetic stability and immunogenicity of HLA-DR4/peptide complexes. Eur J Immunol. 2002;32:662–70.
14. Buus S, Sette A, Colon SM, Miles C, Grey HM. The relation between major histocompatibility complex (MHC) restriction and the capacity of Ia to bind immunogenic peptides. Science. 1987;235:1353–8.
15. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, et al. The immune epitope database (IEDB) 3.0. Nucleic Acids Res. 2015;43(Database issue):D405–12.
16. Andreatta M, Trolle T, Yan Z, Greenbaum JA, Peters B, Nielsen M. An automated benchmarking platform for MHC class II binding prediction methods. Bioinformatics. 2018;34(9):1522–8.
17. Gowthaman U, Agrewala JN. In silico tools for predicting peptides binding to HLA-class II molecules: more confusion than conclusion. J Proteome Res. 2008;7:154–63.
18. Wang P, Sidney J, Dow C, Mothé B, Sette A, Peters B. A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. PLoS Comput Biol. 2008;4:e1000048.
19. Caron E, Kowalewski DJ, Chiek Koh C, Sturm T, Schuster H, Aebersold R. Analysis of major histocompatibility complex (MHC) immunopeptidomes using mass spectrometry. Mol Cell Proteomics MCP. 2015;14:3105–17.
20. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J, et al. Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. Immunity. 2017;46:315–26.
21. Bassani-Sternberg M, Gfeller D. Unsupervised HLA peptidome deconvolution improves ligand prediction accuracy and predicts cooperative effects in peptide-HLA interactions. J Immunol. 2016;197:2492–9.
22. Bergseng E, Dørum S, Arntzen MØ, Nielsen M, Nygård S, Buus S, et al. Different binding motifs of the celiac disease-associated HLA molecules DQ2.5, DQ2.2, and DQ7.5 revealed by relative quantitative proteomics of endogenous peptide repertoires. Immunogenetics. 2015;67:73–84.
23. Chong C, Marino F, Pak H-S, Racle J, Daniel RT, Müller M, et al. High-throughput and sensitive immunopeptidomics platform reveals profound IFNγ-mediated remodeling of the HLA ligandome. Mol Cell Proteomics MCP. 2018;17(3):533–48.
24. Clement CC, Becerra A, Yin L, Zolla V, Huang L, Merlin S, et al. The dendritic cell major histocompatibility complex II (MHC II) peptidome derives from a variety of processing pathways and includes peptides with a broad Spectrum of HLA-DM sensitivity. J Biol Chem. 2016;291:5576–95.
25. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. J Immunol. 2017;199:3360–8.
26. Ooi JD, Petersen J, Tan YH, Huynh M, Willett ZJ, Ramarathinam SH, et al. Dominant protection from HLA-linked autoimmunity by antigen-specific regulatory T cells. Nature. 2017;545:243–7.
27. Andreatta M, Alvarez B, Nielsen M. GibbsCluster: unsupervised clustering and alignment of peptide sequences. Nucleic Acids Res. 2017;45(W1):W458–W463.
28. Andreatta M, Lund O, Nielsen M. Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. Bioinforma Oxf Engl. 2013;29:8–14.
29. Nielsen M, Andreatta M. NNAlign: a platform to construct and evaluate artificial neural network models of receptor-ligand interactions. Nucleic Acids Res. 2017;45(W1):W344–W349.
30. Alvarez B, Barra C, Nielsen M, Andreatta M. Computational tools for the identification and interpretation of sequence motifs in immunopeptidomes. Proteomics. 2018; In Press.
31. Blum JS, Wearsch PA, Cresswell P. Pathways of antigen processing. Annu Rev Immunol. 2013;31:443–73.
32. Lippolis JD, White FM, Marto JA, Luckey CJ, Bullock TNJ, Shabanowitz J, et al. Analysis of MHC class II antigen processing by quantitation of peptides that constitute nested sets. J Immunol. 2002;169:5089–97.
33. Kropshofer H, Max H, Halder T, Kalbus M, Muller CA, Kalbacher H. Self-peptides from four HLA-DR alleles share hydrophobic anchor residues near the NH2-terminal including proline as a stop signal for trimming. J Immunol. 1993;151:4732–42.
34. Ciudad MT, Sorvillo N, van Alphen FP, Catalán D, Meijer AB, Voorberg J, et al. Analysis of the HLA-DR peptidome from human dendritic cells reveals high affinity repertoires and nonconventional pathways of peptide generation. J Leukoc Biol. 2017;101:15–27.
35. Bird PI, Trapani JA, Villadangos JA. Endolysosomal proteases and their inhibitors in immunity. Nat Rev Immunol. 2009;9:871–82.
36. Jensen KK, Andreatta M, Marcatili P, Buus S, Greenbaum JA, Yan Z, et al. Improved methods for predicting peptide binding affinity to MHC class II molecules. Immunology. 2018;154(3):394–406.
37. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. Protein Sci Publ Protein Soc. 1992;1:409–17.
38. Thomsen MCF, Nielsen M. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. Nucleic Acids Res. 2012;40(Web Server issue):W281–7.
39. Larsen SL, Pedersen LO, Buus S, Stryhn A. T cell responses affected by aminopeptidase N (CD13)-mediated trimming of major histocompatibility complex class II-bound peptides. J Exp Med. 1996;184:183–9.
40. Mouritsen S, Meldal M, Werdelin O, Hansen AS, Buus S. MHC molecules protect T cell epitopes against proteolytic destruction. J Immunol. 1992;149:1987–93.
41. 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. Immunogenetics. 2015;67:641–50.
42. Bassani-Sternberg M, Chong C, Guillaume P, Solleder M, Pak H, Gannon PO, et al. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. PLoS Comput Biol. 2017;13:e1005725.
43. Kreiter S, Vormehr M, van de Roemer N, Diken M, Löwer M, Diekmann J, et al. Mutant MHC class II epitopes drive therapeutic immune responses to cancer. Nature. 2015;520:692–6.
44. Tran E, Turcotte S, Gros A, Robbins PF, Lu Y-C, Dudley ME, et al. Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. Science. 2014;344:641–5.
45. Zanetti M. Tapping CD4 T cells for cancer immunotherapy: the choice of personalized genomics. J Immunol. 2015;194:2049–56.
46. Justesen S, Harndahl M, Lamberth K, Nielsen L-LB, Buus S. Functional recombinant MHC class II molecules and high-throughput peptide-binding assays. Immunome Res. 2009;5:2.
47. Sidney J, Southwood S, Oseroff C, del Guercio MF, Sette A, Grey HM. Measurement of MHC/peptide interactions by gel filtration. Curr Protoc Immunol. 2001;Chapter 18(Unit 18):3.
48. 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;Chapter 18(Unit 18):3.
49. 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.

50. Singh H, Raghava GP. ProPred: prediction of HLA-DR binding sites. Bioinforma Oxf Engl. 2001;17:1236–7.
51. Gfeller D, Bassani-Sternberg M, Schmidt J, Luescher IF. Current tools for predicting cancer-specific T cell immunity. Oncoimmunology. 2016;5: e1177691.
52. Nielsen M, Lund O, Buus S, Lundegaard C. MHC class II epitope predictive algorithms. Immunology. 2010;130:319–28.
53. Backert L, Kohlbacher O. Immunoinformatics and epitope prediction in the age of genomic medicine. Genome Med. 2015;7:119.
54. Lazarski CA, Chaves FA, Jenks SA, Wu S, Richards KA, Weaver JM, et al. The kinetic stability of MHC class II:peptide complexes is a key parameter that dictates immunodominance. Immunity. 2005;23:29–40.
55. Roche PA, Furuta K. The ins and outs of MHC class II-mediated antigen processing and presentation. Nat Rev Immunol. 2015;15(4):203–16.
56. Morris P, Shaman J, Attaya M, Amaya M, Goodman S, Bergman C, Monaco JJ, Mellins E. An essential role for HLA-DM in antigen presentation by class II major histocompatibility molecules. Nature. 1994;368(6471):551–4.
57. Mellins ED, Stern LJ. HLA-DM and HLA-DO, key regulators of MHC-II processing and presentation. Curr Opin Immunol. 2014;26:115–22.
58. Lanzarotti E, Marcatili P, Nielsen M. Identification of the cognate peptide-MHC target of T cell receptors using molecular modeling and force field scoring. Mol Immunol. 2018;94:91–7.