Immunogenetics (2019) 71:445–454
https://doi.org/10.1007/s00251-019-01122-z

ORIGINAL ARTICLE

Improved peptide-MHC class II interaction prediction through integration of eluted ligand and peptide affinity data

Christian Garde [1] · Sri H. Ramarathinam [2] · Emma C. Jappe [1,3] · Morten Nielsen [3,4] · Jens V. Kringelum [1] · Thomas Trolle [1] · Anthony W. Purcell [2]

Received: 12 April 2019 / Accepted: 31 May 2019 / Published online: 10 June 2019

Abstract

Major histocompatibility complex (MHC) class II antigen presentation is a key component in eliciting a CD4+ T cell response. Precise prediction of peptide-MHC (pMHC) interactions has thus become a cornerstone in defining epitope candidates for rational vaccine design. Current pMHC prediction tools have, so far, primarily focused on inference from in vitro binding affinity. In the current study, we collate a large set of MHC class II eluted ligands generated by mass spectrometry to guide the prediction of MHC class II antigen presentation. We demonstrate that models developed on eluted ligands outperform those developed on pMHC binding affinity data. The predictive performance can be further enhanced by combining the eluted ligand and pMHC affinity data in a single prediction model. Furthermore, by including ligand data, the peptide length preference of MHC class II can be accurately learned by the prediction model. Finally, we demonstrate that our model significantly outperforms the current state-of-the-art prediction method, NetMHCIIpan, on an external dataset of eluted ligands and appears superior in identifying CD4+ T cell epitopes.

Keywords MHC class II · Ligand prediction · CD4+ epitope · Pan method · Machine learning · Mass spectrometry · Peptidomics

Introduction

CD4+ T helper (Th) lymphocytes constitute an important subset of immune cells that help eliminate pathogens present in infected cells. Th lymphocytes carry out their important helper functions through the interaction with a specialized antigen presentation system termed the major histocompatibility complex class II (MHCII) antigen presentation pathway [1, 2]. The role of this pathway is to display digested exogenous protein antigens as peptides bound to the MHCII molecule to the CD4+ T cells together with a number of co-stimulatory factors for eliciting a profound immune response. MHCII molecules are assembled in the endoplasmic reticulum and are structurally composed of an alpha and beta chain which come together to form a binding groove [3]. Peptides capable of binding to this groove are then transferred to the cell surface of primarily professional antigen presenting cells (APCs), such as macrophages, dendritic cells (DCs), and B lymphocytes, for T cell recognition [4]. Peptides capable of T cell recognition and thus eliciting an immune response are termed T cell epitopes, and the identification of these is of great importance in the context of therapeutic research areas such as vaccine design [5, 6].

Approaches to speed up development and improve the design of vaccines and immunotherapeutics include the creation of algorithms that can accurately predict immunogenic T cell epitopes [7, 8]. With ligand binding to MHCII molecules being an important characteristic of Th cell epitopes, predicting features of this binding interaction has been of great interest, thus making peptide-MHCII (pMHCII) binding the most

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00251-019-01122-z) contains supplementary material, which is available to authorized users.

✉ Christian Garde
cg@evaxion-biotech.com

✉ Anthony W. Purcell
anthony.purcell@monash.edu

[1] Evaxion Biotech, Bredgade 34E, DK-1260 Copenhagen, Denmark

[2] Department of Biochemistry and Molecular Biology & Infection and Immunity Program, Biomedicine Discovery Institute, Monash University, Clayton, Victoria 3800, Australia

[3] Department of Bio and Health Informatics, Technical University of Denmark, DK-2800 Lyngby, Denmark

[4] Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, Buenos Aires, Argentina

studied part of antigen presentation by Th lymphocytes [9]. This has resulted in a growing amount of binding data which has made prediction algorithms increasingly reliable [10, 11].

However, despite encouraging improvements in the predictive performance of pMHCII binding algorithms, several studies have demonstrated an inferior performance to MHC class I (MHCI) binding prediction tools [9, 12]. Multiple biological factors challenge the development of accurate algorithms for predicting pMHCII binding [4]. MHCII molecules are extremely polymorphic with almost 5000 alpha and beta chain alleles known as of April 2018 [13]. The MHCII alleles are encoded by three polymorphic genes (HLA-DR, HLA-DQ, and HLA-DP) in humans, and the combinatorial of the alpha and beta chains results in a much larger complexity for MHCII molecules compared with MHCI [5]. Furthermore, the binding groove of MHCII permits the binding of a diverse range of peptides of variable lengths, partly because of its open conformation [2]. This contrasts with the MHCI molecule which accommodates peptides based on a much more stringent length criterion [1]. The pMHCII binding is primarily governed by the interactions between the MHCII binding groove and a nine amino acid “core” in the peptide. However, the open conformation of the binding groove complicates the interpretation of pMHCII measurements, as it is not clear which peptide register is bound within the groove. The high MHCII complexity underpins the need for pan prediction tools that are capable of extrapolating pMHCII binding preferences to uncharacterized MHCII molecules.

The state-of-the-art pan method for prediction of peptide binding affinity to MHCII molecules is the NetMHCIIpan tool [11, 14]. NetMHCIIpan is based on ensembles of artificial neural networks trained on quantitative binding affinity data from the Immune Epitope Database (IEDB) [10]. NetMHCIIpan accounts for the challenges that lie in the polymorphism of the MHCII molecule by including the sequence of the MHCII molecule in the model, while the identification of the 9-mer peptide cores is handled using the NNalign method [20]. However, although prediction algorithms are capable of accurately predicting in vitro pMHCII binding, its predictive performance on antigen presentation in vivo is limited. Especially, peptide processing, length preference, and loading on to the MHCII molecules are major factors not accounted for when only considering pMHCII binding. In line with earlier work on pMHCI prediction [15], we present a pan prediction framework that incorporates both pMHCII binding affinity data and mass spectrometry (MS) based eluted MHCII ligands from the surface of antigen-presenting cells. As eluted MHCII ligands undergo processing and loading on to MHCII as well as transport to the cell surface, these inherently contain essential biological information, thus addressing some of the limitations of in vitro pMHCII affinity data.

Methods

MHC class II datasets

Eluted ligands and quantitative pMHC affinity data (IC50) were retrieved from IEDB [10] and filtered to the subset associated with MHC alleles with at least a four-digit resolution on their identifiers. The IEDB data were supplemented with a large set of in-house eluted ligands.

To eliminate spurious data from the eluted ligands, we generated MHC allele-specific ligand lists and processed them using GibbsCluster [16] with one, two, three, four, and five clusters, respectively, and a discard pile. For each list of ligands, the GibbsCluster run yielding the highest Kullback-Leibler divergence was applied to flag and remove spurious ligands. The ligand length profile was computed and constrained to the subset that constitutes more than 0.5% of the total number of eluted ligands, resulting in ligands spanning from 9 to 21 amino acids. In total, the data comprised 51,269 affinity measurements across 31 MHC molecules and 142,267 eluted ligands across 25 MHC molecules. While each in vitro affinity measurement can be uniquely assigned to an MHC molecule, the in vivo experimental design may not restrict the eluted ligand to a single MHC molecule. Among the eluted ligand experiments, 15 restrict the eluted ligands to a single MHC molecule. This subset of the data will be referred to as the mono-allelic fraction.

As a consequence of the eluted ligand detection methodology, no negative data exists, which complicates training and learning from this dataset. Thus, a negative complement to the eluted ligand data was established by randomly sampling k-mers from proteins in the Swissprot database [17]. To enable the recognition of the ligand length preference, the peptides were sampled with a uniform length distribution ranging from 9 through 21 in line with earlier work [15]. For each length and for each MHC molecule, we sampled three times as many peptides as the most frequent ligand length, i.e., a total of 39 times as many random peptides as the most frequent ligand length. The dataset comprising pMHC affinity data is denoted “IC50”, the eluted ligand data is denoted “LIGAND”, and the concatenated dataset is denoted “IC50 + LIGAND.”

MHC class II peptidome analysis

Purification of HLA-bound peptides

Frozen pellets (1-5e9) from cell lines (IHW9013, IHW9022, IHW9031, IHW9087, C1R(IHW9208), JESI, or JESTHOM (IHW9004)) were ground in a Retsch Mixer Mill MM 400 under cryogenic conditions, resuspended in 1% IGEPAL (Sigma), 50 mM Tris pH 8, 150 mM NaCl, and protease inhibitors (Complete Protease Inhibitor Cocktail Tablet; Roche Molecular Biochemicals) as previously described.

Lysates were cleared by ultracentrifugation and HLA-peptide complexes purified using sequential protein A (GE Healthcare) columns bound to specific HLA mAbs. Lysates were first passed over protein A bound to LB3.1 (anti-DR), SPV-L3 (anti-DQ), and B721 (anti-DP). Bound HLA complexes were eluted from each column by acidification with 10% acetic acid. The eluted mixture of peptides and HLA chains was fractionated on a 4.6-mm internal diameter × 50-mm (or 100 mm) long reversed-phase C18 endcapped HPLC column (Chromolith Speed Rod, Merck) using an ÄKTAmicro™ HPLC system (GE Healthcare) running on a mobile phase buffer A of 0.1% TFA and buffer B of 80% ACN/0.1% TFA. The HLA-peptide mixtures were loaded onto the column at a flow rate of 1 ml/min with separation based on a gradient of 2 to 40% B for 4 min, 40 to 45% for another 4 min, and a rapid 2 min increase to 100% B. Fractions (500 μl) were collected, vacuum concentrated to 10 μl, and diluted in 0.1% formic acid to reduce the concentration of acetonitrile (ACN). HLA typing of the cell lines is provided in Table S1.

LC-MS/MS

For LC-MS/MS acquisition, peptide-containing fractions were loaded onto a microfluidic trap column packed with ChromXP C18-CL 3-μm particles (300 Å nominal pore size; equilibrated in 0.1% formic acid/2% ACN) at 5 μl/min using a NanoUltra cHiPLC system. An analytical (75 μm × 15 cm ChromXP C18-CL 3 μm, 120 A, Eksigent) microfluidic column was switched in line and peptides separated using linear gradient elution of 0–80% ACN over 90 min (300 nl/min). Separated peptides were analyzed using an AB SCIEX 5600+ TripleTOF mass spectrometer equipped with a Nanospray III ion source and accumulating up to 20 MS/MS spectra per second. The following experimental parameters were used: ion spray voltage (IS) was set at 2400 V, curtain gas at 22 L/min, ion source gas at 8 L/min, and an interface heater temperature setting of 150 °C. MS/MS switch criteria included ions of $m/z > 200$ amu, charge state + 2 to + 5, and intensity > 40 cps, and the top 20 ions meeting these criteria were selected for MS/MS per cycle.

LC-MS/MS data analysis

LC-MS/MS data was searched against the human proteome (Uniprot/Swissprot v2016_12) using the ProteinPilot™ software (AB SCIEX) [18] and resulting peptide identities subject to strict bioinformatic criteria including the use of a decoy database to calculate the false discovery rate (FDR). A 5% FDR cutoff was applied, and the filtered dataset was further analyzed manually to exclude redundant peptides and known contaminants as previously described [19]. The following protein pilot search parameters were used: no cysteine alkylation, no enzyme digestion (considers all peptide bond cleavages), instrument-specific settings fTripleTOF (MS tolerance 0.05 Da, MS/MS tolerance 0.1 Da, charge state + 2–5), species *Homo sapiens*, biological modification probabilistic features on, Swiss-Prot database (version v2016_12), thorough ID algorithm, and detected protein threshold 0.05 (see Supplementary Data).

Model framework

We adapted the NNalign approach [20] to accommodate an architecture with two output neurons. The motivation was to enable combined training on affinity and ligand data as described in [15]. Bootstrap sampling from the ligand dataset was performed at the beginning of every training epoch, such that a matched number of affinity and ligand examples were presented to the model at every epoch in a randomized order. The affinity scores were transformed to suitable target values in the range from zero to one using "1-log (affinity [nM])/log(50,000)" as we have described in earlier work [21, 22]. Since the binding strengths of the eluted ligands are unknown, we assigned the eluted ligands a target value of 1, and negative peptides were assigned a target value of 0. The input to the model was encoded similarly to the scheme employed by [23]. The MHC molecule pseudo-sequence and the 9-mer peptide cores were encoded using normalized log-odds scores derived from the BLOSUM50 substitution matrix. The flanks of the peptide cores were up to three amino acids long but constrained by the span of the peptide, i.e., a flank was shorter than three amino acids for peptide cores located at the edge of the peptide. The flanks were encoded by their average BLOSUM50 encoding. The length of the peptide core flanks was encoded by $0.85 \times L / 3 + 0.05$, with L denoting the flank length. The peptide length was encoded using $1 / (1 + \exp.((P - 15) / 2))$, with P denoting the peptide length.

Deconvolution of multiallelic data

The eluted ligands that are non-uniquely assigned to one MHC molecule were instead assigned their MHC by inference. For this purpose, model ensembles were trained for 300 epochs on the mono-allelic fraction of the data with 10 different random initialization seeds and 40 hidden neurons. Percentile rank conversion of the raw prediction scores is required in order to make the amplitude of the predictions comparable across MHC molecules [15]. The percentile rank conversion was established by predicting on 200,000 random peptides for each of the lengths 9 through 21 amino acids and each of the MHC molecules to be subject to deconvolution. MHC molecule assignment was then inferred for each ligand independently based on the best percentile rank prediction. The negative sampled peptides were randomly assigned to the MHC molecules in the same ratio as the eluted ligands. The resulting deconvoluted dataset was merged with the mono-allelic fraction and used for the nested cross-validation and to train the final model. An overview of the performance of the mono-allelic models on the data to be deconvoluted is shown in Fig. S1.

Nested cross-validation

A conservative estimate of the predictive performance is achieved by evaluating on unseen MHC molecules and unseen peptides. This has traditionally been achieved by a computationally demanding leave-one-out cross-validation strategy [14, 22].

We defined a novel nested cross-validation strategy to efficiently attain a performance estimate on unseen MHC molecules and peptides. The strategy relied on data partitioning along two axes, namely, MHC molecules and peptides, as depicted in Fig. 4. In the MHC molecule partitioning scheme, we randomly assigned each MHC molecule and its associated peptides to one partition. In this way, there are no shared MHC molecules between the partitions along this axis. The peptide partitioning was carried out by first clustering the peptides according to shared 9-mer motifs as described in [24]. Secondly, the peptide clusters were randomly agglomerated into size-balanced partitions. In this way, there are no shared 9-mer motifs between the peptide partitions. We employed five MHC molecule partitions and six peptide partitions, yielding a matrix with 30 partitions. For each rotation in the cross-validation, one of the 30 partitions was kept as the left-out evaluation set and its corresponding rows and columns in the matrix were removed since they contained shared MHC molecules or peptides. The rest of the data was used as the training set. We employed this partitioning scheme on the joint dataset “IC50 + LIGAND” and, subsequently, split it into the IC50 and LIGAND while preserving the partitioning.

For each evaluation set, we trained five models using five different random initialization seeds and used the model ensemble for prediction on the evaluation set. Models were trained with 40 hidden neurons and for 300 epochs for each of the datasets IC50, LIGAND, and IC50 + LIGAND. We constrained the evaluation to the mono-allelic subset of the data, such that the data subjected to deconvoluted does not enter into the performance estimation.

Ligand length preference

For each of the MHC molecules covered by the mono-allelic fraction of the ligand data, we predicted on a set of 200,000 random peptides for each of the lengths 9 through 21. From the top 1% of these predictions, we computed the frequency of each of the peptide lengths as a representative of the MHC length preference. This inferred ligand length preference was compared with the length distribution of the eluted ligand data.

Validation on external datasets

A final implementation was trained for 300 epochs on the IC50 + LIGAND dataset with 40 hidden neurons and 10 random seeds.

Two external datasets were established to make a comparative study between our final implementation and the state-of-the-art NetMHCIIpan [14]. The first dataset comprises an unpublished set of in-house eluted ligands and was preprocessed with GibbsCluster, as described above. The second dataset was retrieved from IEDB and comprises CD4+ T cell epitopes measured by the multimer/tetramer assay in an infectious disease setting. The CD4+ epitopes were filtered to the subset with length 15 and associated with MHC molecules at four-digit resolution. Eluted ligands and CD4+ T cell epitopes, which share a 9-mer with any peptide in our IC50 + LIGAND dataset, were removed.

Performance evaluation

The performance was evaluated using the area under the receiver operating characteristics (AUC) or the *F* rank. The *F* rank is defined as the percentile rank of the positive prediction in the complete set of k-mers from the source protein; thus, the lower the *F* rank, the better the performance. The performance was evaluated per MHC molecule to avoid coverage biases. Differences in predictive performances were tested using the binomial test.

Results

We developed a novel pan-specific model based on the NNalign framework [20] for predicting the interactions of peptides and MHC class II alleles through the integration of pMHC affinity data and MHC eluted ligands. This combined training was enabled by expanding the NNalign framework to accommodate two output neurons, one for each data type. In line with the observations on MHCI by [15], we similarly find that this architecture with a shared hidden layer enables the models to synergistically learn from both data types. A flow chart describing the model development is depicted in Fig. 1.

High-quality eluted ligand data reveal MHC binding preferences

We employed GibbsCluster to inspect the quality of the eluted ligand experiments (Fig. 2). The resulting clusters represent linearly separable specificities. Thus, we expect a single cluster to best describe the data, if data originate from an

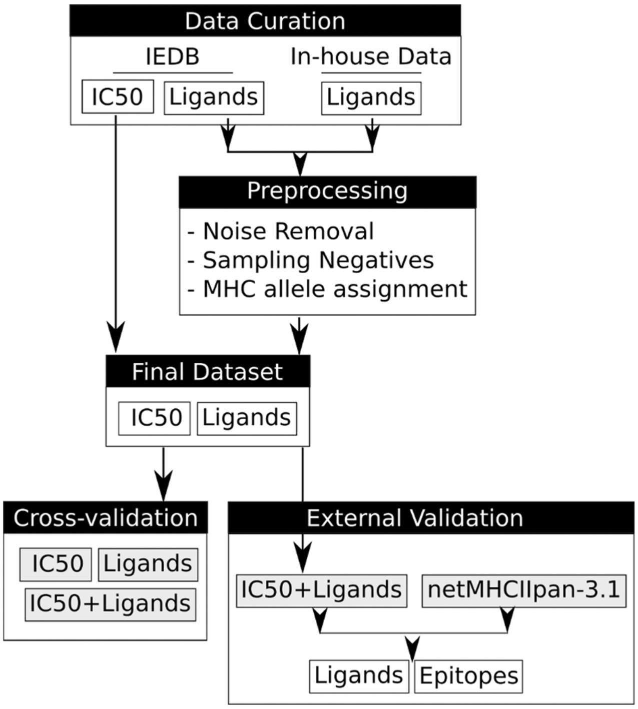

Fig. 1 Flow-chart of the study. Peptide-MHC binding affinity data (IC50) and eluted ligands (Ligands) are curated from the IEDB and in-house resources. The eluted ligand data is subject to quality assessment and filtering using GibbsCluster. A comparative study of combined training on both data types versus IC50 data or Ligands in isolation is conducted by cross-validation. Finally, a competitive study with the state-of-the-art method is carried out on external datasets of eluted ligands and CD4+ T cell epitopes, respectively. Shaded boxes represent models, while unfilled boxes represent data

experiment utilizing a mono-allelic cell line or a cell line harboring MHC molecules with similar specificity profiles. In contrast, multiple clusters indicate that multiple specificities are present in the data. We found that a single cluster yielded the highest Kullback-Leibler divergence in 23 out of the 26 curated experiments, while three experiments had two clusters as their optimum. These three cases have been highlighted with arrow heads in Fig. 2. Two of these three experiments were expected to hold two specificities as the applied cell line was HLA-typed with two MHC molecules targetable by the applied antibody. These MHC alleles are HLA-DPA1*01:03/DPB1*04:01;DPB1*03:01 and HLA-DRA*01:02/DRB1*12:01;DRB3*02:02, respectively. The last experiment with two inferred specificities should only pull down a single MHC molecule (DQA1*01:01/DQB1*05:01) based the HLA typing and DQ-specific antibody. Sequence logos for each of the Gibb's clusters for two of these cases are presented in Supplementary Figs. S3 and S4. The MHC molecules have a preference for certain ligand lengths, which can be described as a Gaussian-like length distribution centered at 15 amino acids with minor offsets and density differences between the MHC molecules (see Fig. 3b). In contrast, the pMHC affinity data is almost exclusively measured on peptides of length 15 (see Fig. 3a). The MHC molecules covered by the IC50 dataset is listed in Supplementary Table S2.

Synergistic training improves predictive performance

A conservative performance estimate relies on carrying out predictions on unseen MHC molecules and unseen peptides. This has traditionally been accomplished by a leave-one-out cross-validation strategy which is computationally demanding. We defined a new nested cross-validation scheme to more efficiently assess the pan-specific potential as illustrated in Fig. 4. Our nested cross-validation scheme requires a total of 30 partitions while the number of partitions in the leave-one-out scheme scales with the number of MHC molecules to evaluate. Models were generated based on the IC50, LIGAND, and IC50 + LIGAND datasets and evaluated for their ability to distinguish the eluted ligands from the random natural peptides. The IC50 + LIGAND model outperforms the models trained on each of the data types in isolation (IC50 and LIGAND) (see Fig. 5a). Furthermore, the output neuron designated to the ligand likelihood (IC50 + LIGAND, LIGAND) demonstrates improved performance to that of the pMHC affinity (IC50 + LIGAND, IC50). Thus, the ligand output neuron should be used for the ligand inference.

The uniform length distribution of the negative set was designed for the models to learn the specificity along two axes of information, namely, sequence binding motifs and peptide length. To test the performance of the models solely based on the sequence motif axis, we rank the prediction of the eluted ligands relative to those of all length-matched k-mers from the respective source proteins. This evaluation again demonstrates that the joint model IC50 + LIGAND has the best performance (lowest *F* rank) (see Fig. 5b).

Learning the MHC class II length preferences

Since the pMHC affinity data is almost exclusively based on measurements using peptides of length 15 as depicted in Fig. 2a, we hypothesize that it does not provide the means to learn the ligand length preference of an MHCII molecule. To investigate this, we applied our models on 200,000 artificial peptides for each of the lengths 9 through 21 and computed the length profile of the top 1% of the predictions for each MHC molecule covered by the mono-allelic fraction of the data. The inferred length preference based on the respective models is superimposed on the true length preference established from the eluted ligands (see Fig. 6a). It is clear that the model trained solely on the pMHC affinity data (IC50) infers a ramp-like length profile with a bias towards longer peptides. Without the ability to properly learn the length preference (average PCC ~0.32), the longer peptides will naturally rank better given the higher

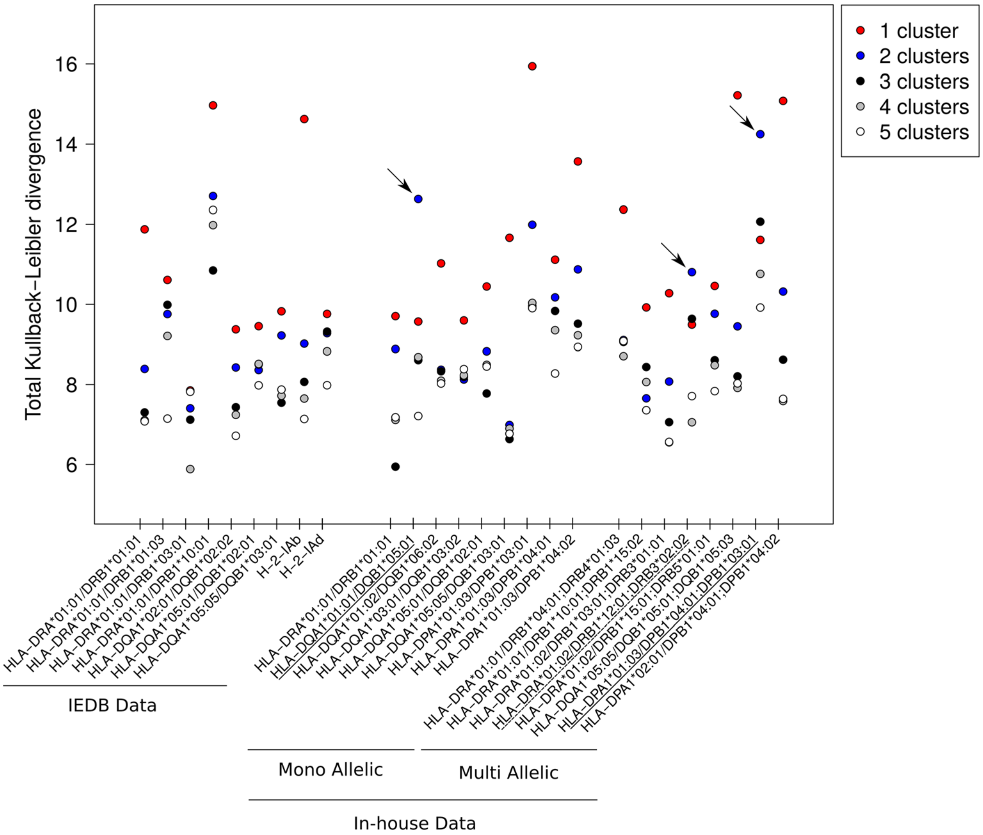

Fig. 2 Quality assessment and filtering of eluted ligand data. Assessment was done with GibbsCluster using one through five clusters. Each cluster represents a linearly separable specificity model. The highest Kullback-Leibler divergence indicates the optimal set of clusters in the dataset. Most of the cases have a single cluster as the optimum. The three cases that have two clusters as their optimum are indicated with arrows, and the MHC molecule labels are highlighted by underlining

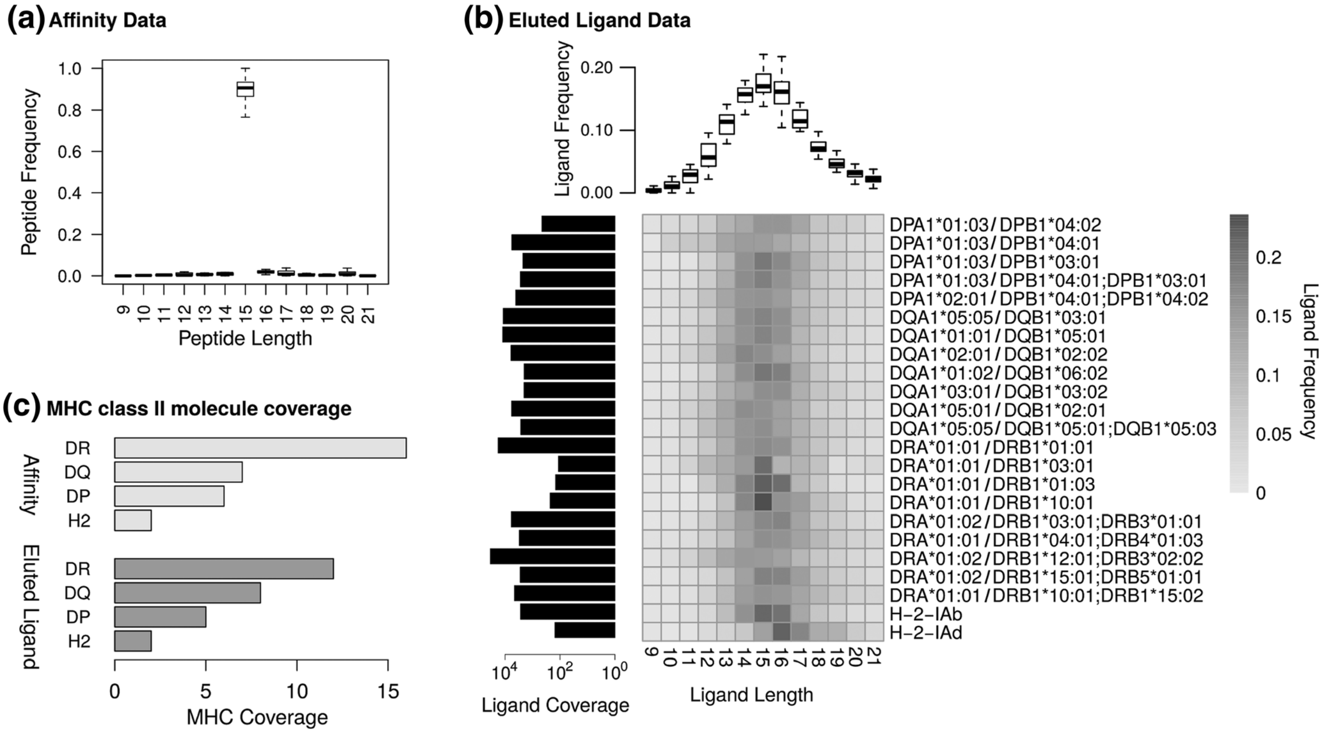

Fig. 3 Overview of curated dataset. **a** Length distribution of the peptides measured by the peptide-MHC affinity assay (IC50). **b** Length distribution and ligand coverage of the respective MHC molecules covered by the curated eluted ligand dataset. **c** MHC molecule coverage for the respective MHC loci covered by the peptide-MHC affinity data and eluted ligand data

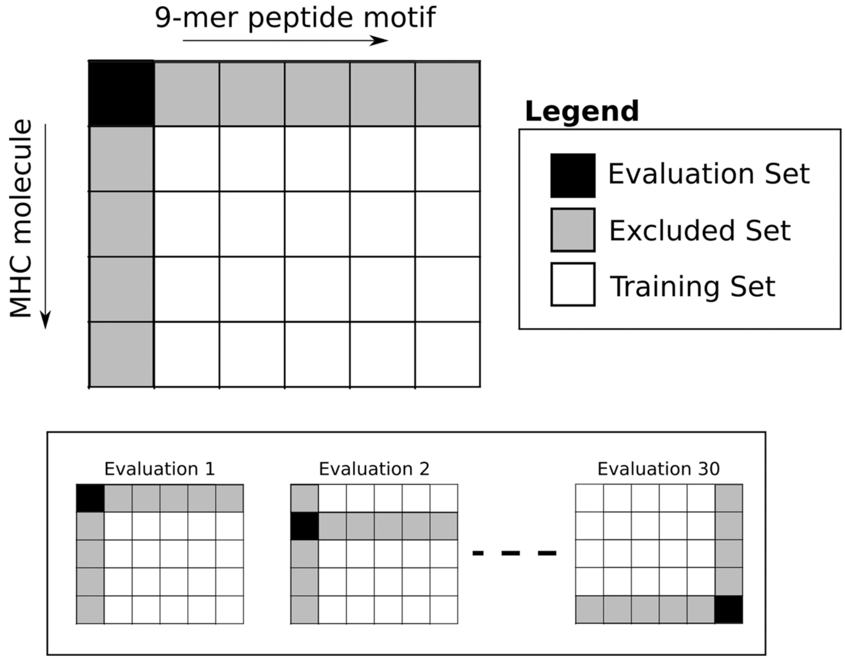

Fig. 4 Nested cross-validation strategy for efficient evaluation on unseen MHC molecules and peptides. Partitioning along two axes: MHC molecules and 9-mer peptide motifs visualized as a 2D matrix. For each rotation, one entry is selected as the evaluation set and the corresponding rows and columns are omitted from the training set, since they share MHC molecules or peptide core motifs with the evaluation set

probability of hosting a good 9-mer binding motif. In contrast, the models which learned from eluted ligand data (IC50 + LIGAND and LIGAND) have the ability to infer the length profile (average PCC ~ 0.94) (see Fig. 6b).

Validation on external datasets

Finally, a competitive study was conducted with the state-of-the-art pan method NetMHCIIpan [14], and we trained the final model implementation on the full IC50 + LIGAND dataset. A new eluted ligand dataset was generated and reduced to the subset that did not share a 9-mer peptide motif with the IC50 + LIGAND dataset. An overview of this external ligand dataset is presented in Fig. S2. Our model outperforms NetMHCIIpan-3.1 by a large margin of 9.6% *F* rank ($p < 0.05$, binomial test) (see Fig. 7a). Furthermore, similar performance is demonstrated on the external dataset as that estimated on in the cross-validation,

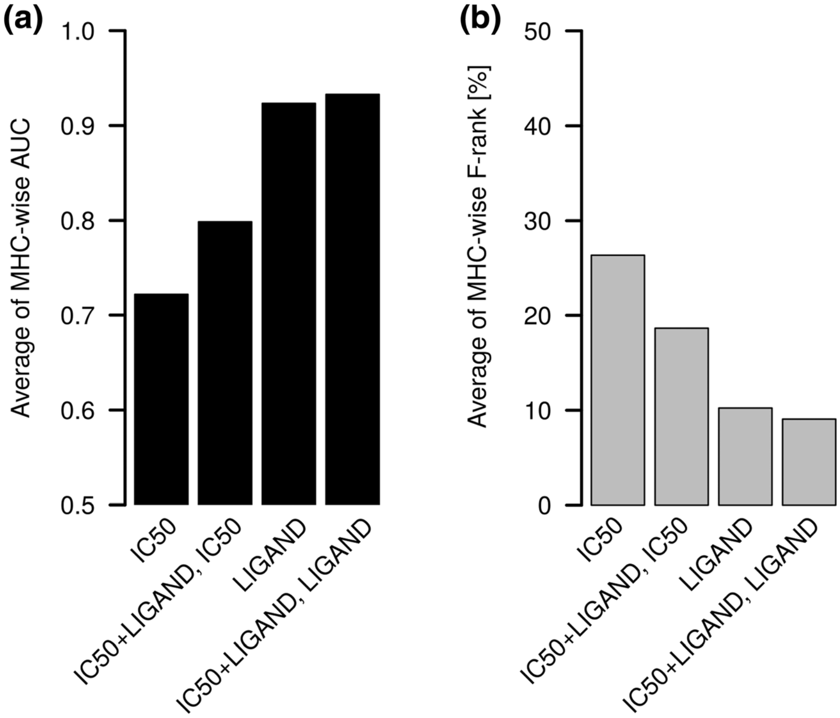

Fig. 5 Nested cross-validation experiment. **a** Receiver operating characteristics demonstrating the ability of the models to distinguish eluted ligands from random natural peptides. **b** Average rank of the predictions of the eluted ligands relative to length-matched k-mers from their source proteins with lower ranks being better. Both evaluations demonstrate an improved performance for the model based on the combined data (IC50 + LIGAND) compared with the models trained on the two data types in isolation (IC50 and LIGAND). The evaluation was constrained to the mono-allelic subset

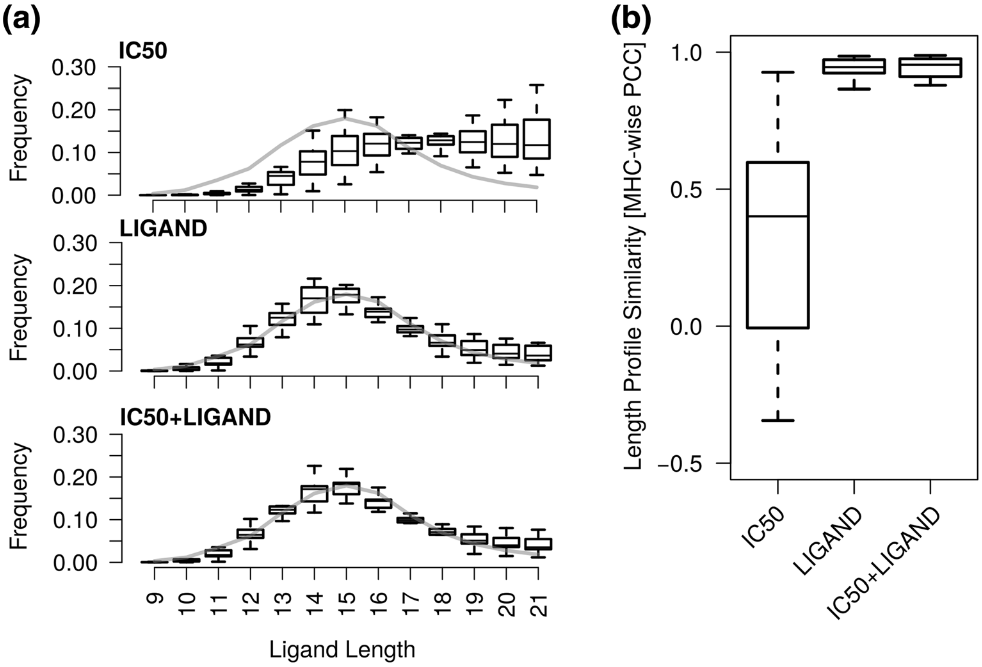

Fig. 6 Learning the ligand length preference of MHC class II. **a** Distributions of the inferred MHC preferences superimposed with the median of the length profile of the eluted ligand dataset. **b** Correlation between the ligand length profile and the inferred length profiles

which confirms the validity of the cross-validation scheme. To complement the evaluation on eluted ligands, we retrieved a dataset of CD4+ epitopes measured by the multimer/tetramer assay from the IEDB. Again, our method was found to perform better than NetMHCIIpan with the gain on the CD4+ epitopes being more moderate with a margin of 3.3% *F* rank (see Fig. 7b). It should be noted that both models have a higher variability in the CD4+ epitope performance than for the eluted ligands. Thus, on this limited benchmarking set covering seven MHC molecules, the difference is not found to be statistically significant ($p > 0.05$, binomial test).

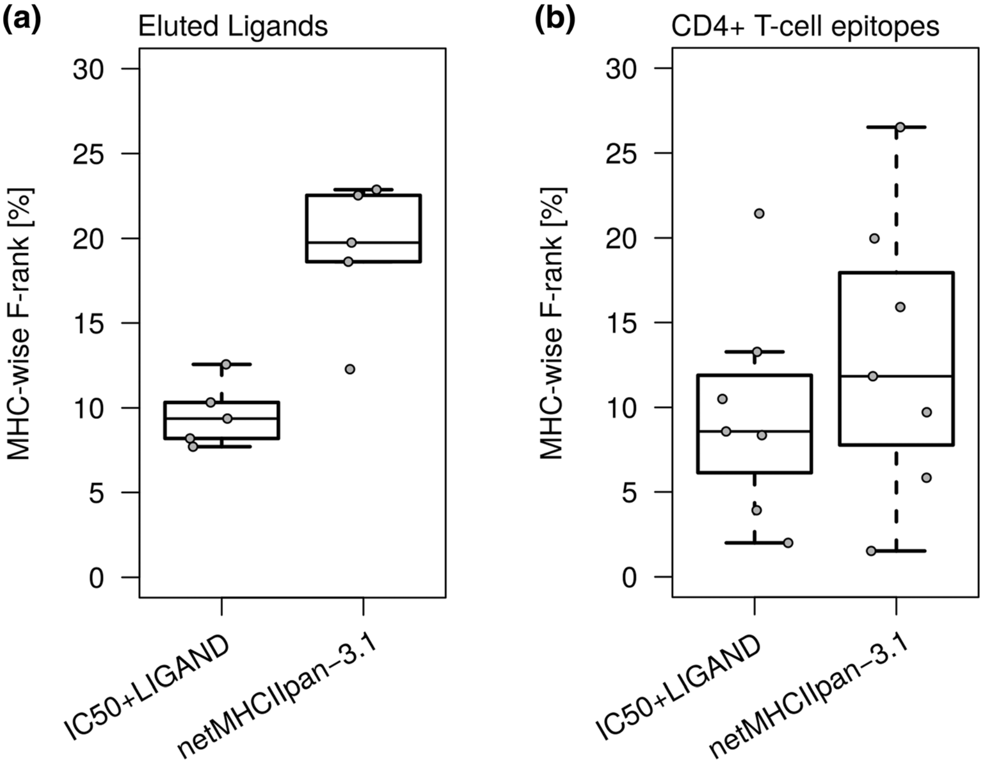

Fig. 7 Competitive evaluation with state of the art on external dataset. **a** Evaluation on an external eluted ligand dataset. **b** Evaluation on an external CD4+ epitope dataset measured with the tetramer assay. In both cases, our model outcompetes NetMHCIIpan-3.1. The performance is given as the percentile rank of the positive example in a set of length-matched k-mers from the source proteins. A lower *F* rank signifies better predictive performance

Discussion

In this study, we present a novel model for the prediction of pMHCII interactions. The model is developed by combined training on pMHC affinity data and eluted ligand data. The two datasets each have their strengths. The pMHC affinity dataset is quantitative in nature and has a larger MHC molecule coverage at the time of writing (see Fig. 3c). On the other hand, the eluted ligands represent peptides that are bound to the MHC molecule in a natural setting and thus inherently contain a wealth of information regarding peptide processing, peptide length preferences, and potentially modulated sequence specificity due to chaperone-assisted docking [4]. Additionally, the peptide coverage for each MHC molecule is larger for the eluted ligand datasets. However, the eluted ligands solely provide a snapshot of the antigen presentation, and, as of yet, they do not carry direct information on their binding strength as it is confounded by the background peptide abundance [25]. Future efforts to collect paired MS peptidomes and expression data may assist in elucidating the binding strength of eluted ligands. Collectively, we consider the eluted ligand datasets to provide the most accurate representation of antigen presentation. We demonstrate that the rules governing pMHC recognition can be synergistically learned by training on the pMHC affinity and eluted ligands simultaneously to increase predictive performance beyond that attainable by training on the two data types in isolation.

We believe that as long as eluted ligands are only covered by a limited set of MHC molecules, the pMHC affinity data can supplement the eluted ligands to assist the models to generalize better to unseen MHC molecules. Furthermore, we showcase that the increased performance can be attributed to both increased inference of the sequence specificity and the peptide length preference of the respective MHCII molecules. Notably, the MHCII molecules appear to have more similar length preferences than that observed for MHCI molecules [15]. With the open binding groove of MHCII, one may envision that the length preference is defined by the ligand processing machinery to a larger extent as compared with MHCI. Finally, we show that our model has an unprecedented performance by outcompeting the state-of-the-art NetMHCIIpan on external datasets of eluted ligands and CD4+ T cell epitopes. It should be noted that the CD4+ T cell epitopes likely have been selected based on prediction or in vitro measurements of pMHCII affinity, and they may thus yield an overestimated performance. However, the conditions are similar for the two methods, and this bias will probably disproportionally inflate the performance of NetMHCIIpan. Future efforts to improve performance should revolve around harnessing more of the information kept in the eluted ligands, e.g., ligand processing as illustrated by [26].

The holy grail of immunoinformatics is to infer epitopes to develop disease treatment regimes for, e.g., allergy, infectious diseases, autoimmune diseases, and cancer. This has especially become highly interesting with the increased focus on precision medicine [6]. To meet these challenges, a model should ultimately incorporate the recognition of pMHC by the T cell receptors (TCRs). This goal is currently limited by the lack of data on pMHC-TCR interactions and limited availability of cost-effective high-throughput assays [27]. It should be noted, however, that efforts to curate the available pMHC-TCR data are ongoing [10, 28].

Acknowledgments ECJ was supported by a PhD stipend from the Innovation Fund Denmark. AWP was supported by a Principal Research Fellowship and project grant 1022509 from the National Health and Medical Research Council of Australia (NHMRC). SHR was supported by an Australian Postgraduate Award. MN is a researcher at the Argentinean National Research Council (CONICET).

We are grateful to the members of Prof. Purcell's lab for their assistance in compiling the eluted ligand data. Furthermore, we thank the bioinformatics team at Evaxion Biotech for constructive feedback.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Wieczorek M, Abualrous ET, Sticht J, Álvaro-Benito M, Stolzenberg S et al (2017) Major histocompatibility complex (MHC) class I and MHC class II proteins: conformational plasticity in antigen presentation. Front Immunol 8:292
2. Unanue ER, Turk V, Neefjes J (2016) Variations in MHC class II antigen processing and presentation in health and disease. Annu Rev Immunol 34:265–297
3. Jones EY (1997) MHC class I and class II structures. Curr Opin Immunol 9:75–79
4. Rock KL, Reits E, Neefjes J (2016) Present yourself! By MHC class I and MHC class II molecules. Trends Immunol 37:724–737
5. Neefjes J, Jongsma MLM, Paul P, Bakke O (2011) Towards a systems understanding of MHC class I and MHC class II antigen presentation. Nat Rev Immunol 11:823–836
6. Saini SK, Rekers N, Hadrup SR (2017) Novel tools to assist neoepitope targeting in personalized cancer immunotherapy. Ann Oncol 28:xii3–xii10
7. Rosa DS, Ribeiro SP, Cunha-Neto E (2010) CD4+ T cell epitope discovery and rational vaccine design. Arch Immunol Ther Exp 58: 121–130
8. Olsen LR, Campos B, Barnkob MS, Winther O, Brusic V, Andersen MH (2014) Bioinformatics for cancer immunotherapy target discovery. Cancer Immunol Immunother 63:1235–1249
9. Andreatta M, Trolle T, Yan Z, Greenbaum JA, Peters B et al (2018) An automated benchmarking platform for MHC class II binding prediction methods. *Bioinformatics* 34:1522–1528
10. Fleri W, Paul S, Dhanda SK, Mahajan S, Xu X et al. (2017) The immune epitope database and analysis resource in epitope discovery and synthetic vaccine design. Front Immunol; 8:278
11. Jensen KK, Andreatta M, Marcatili P, Buus S, Greenbaum JA, Yan Z, Sette A, Peters B, Nielsen M (2018) Improved methods for predicting peptide binding affinity to MHC class II molecules. Immunology. 154:394–406

12. Andreatta M, Nielsen M (2015) Gapped sequence alignment using artificial neural networks: application to the MHC class i system. Bioinformatics. 32:511–517
13. EMBL-EBI (2018) IPD-IMGT/HLA database—statistics. Available at: https://www.ebi.ac.uk/ipd/imgt/hla/stats.html [Accessed July 4, 2018]
14. Andreatta M, Karosiene E, Rasmussen M, Stryhn A, Buus S, Nielsen M (2015) Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification. Immunogenetics. 67:641–650
15. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M (2017) NetMHCpan-4.0: improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. J Immunol 199:3360–3368
16. Andreatta M, Lund O, Nielsen M (2013) Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. Bioinformatics. 29:8–14
17. UniProt Consortium T (2018) UniProt: the universal protein knowledgebase. Nucleic Acids Res 46:2699–2699
18. Shilov IV, Seymour SL, Patel AA, Loboda A, Tang WH, Keating SP, Hunter CL, Nuwaysir LM, Schaeffer DA (2007) The paragon algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. Mol Cell Proteomics 6:1638–1655
19. Schittenhelm RB, Sian TC, Wilmann PG, Dudek NL, Purcell AW (2015) Revisiting the arthritogenic peptide theory: quantitative not qualitative changes in the peptide repertoire of HLA-B27 allotypes. *Arthritis Rheumatol. (Hoboken, N.J.)* 67:702–713
20. Nielsen M, Lund O (2009) NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. BMC Bioinformatics 10:296
21. Nielsen M, Lundegaard C, Worning P, Lauemøller SL, Lamberth K, Buus S, Brunak S, Lund O (2003) Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci 12:1007–1017
22. Mattsson AH, Kringelum JV, Garde C, Nielsen M (2016) Improved pan-specific prediction of MHC class I peptide binding using a novel receptor clustering data partitioning strategy. HLA 88:287–292
23. Karosiene E, Rasmussen M, Blicher T, Lund O, Buus S, Nielsen M (2013) NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. Immunogenetics 65:711–724
24. Nielsen M, Lundegaard C, Lund O (2007) Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. BMC Bioinformatics 8:238
25. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W et al (2017) Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* 46:315–326
26. Paul S, Karosiene E, Dhanda SK, Jurtz V, Edwards L, Nielsen M, Sette A, Peters B (2018) Determination of a predictive cleavage motif for eluted major histocompatibility complex class II ligands. Front Immunol 9:1795
27. Bentzen AK, Marquard AM, Lyngaa R, Saini SK, Ramskov S, Donia M, Such L, Furness AJS, McGranahan N, Rosenthal R, Straten P, Szallasi Z, Svane IM, Swanton C, Quezada SA, Jakobsen SN, Eklund AC, Hadrup SR (2016) Large-scale detection of antigen-specific T cells using peptide-MHC-I multimers labeled with DNA barcodes. Nat Biotechnol 34:1037–1045
28. Shugay M, Bagaev DV, Zvyagin IV, Vroomans RM, al CJ (2018) VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. Nucleic Acids Res 46:D419–D427