NNAlign_MA; semi-supervised MHC peptidome deconvolution for accurate characterization of MHC binding motifs and improved T cell epitope prediction

Bruno Alvarez [1], Birkir Reynisson [2], Carolina Barra [1], Søren Buus [3], Nicola Ternette [4], Tim Connelley [5], Massimo Andreatta [1], and Morten Nielsen* [1,2]

1. Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, San Martín, Argentina.

2. Department of Bio and Health Informatics, Technical University of Denmark, Lyngby, Denmark.

3. Department of Immunology and Microbiology, Faculty of Health Sciences, University of Copenhagen, Denmark.

4. The Jenner Institute, Nuffield Department of Medicine, Oxford, United Kingdom.

5. The Roslin Institute, Edinburgh, Midlothian, United Kingdom.

*corresponding author

Abstract

Antigen presentation by Major Histocompatibility Complex (MHC) is a cornerstone of the cellular immune system, and characterization of the rules defining which peptides are suitable for MHC presentation is of key interest for understanding T cell immunogenicity and the development of T cell driven therapeutics. The collective set of peptides presented on a cell surface by MHC molecules is known as the immunopeptidome. Due to the vast MHC polymorphism, this peptidome is unique between individuals. Current state-of-the-art liquid chromatography mass spectrometry (LC-MS) technologies allow the identification of large peptidomes specific for a given host or cell line, and numerous studies have proven this data to be a rich source of information for learning the rules of MHC-mediated antigen presentation. The data obtained with MS techniques is usually poly-specific – i.e. it contains multiple sequence motifs matching the different MHC molecules expressed in the system under investigation. Deconvolution of this poly-specificity has proven a challenge, and imposes a substantial limitation for learning MHC specific rules for antigen presentation from immunopeptidome data. This since each ligand first must be unambiguously associated to its presenting MHC molecule(s) within the haplotype of the cell line.

Here, we describe NNAlign_MA, a method that is able to address this challenge. NNAlign_MA is capable of taking a mixed training set composed of single-allele (SA) data (peptides assigned to single MHCs) and multi-allele (MA) data (peptides with multiple options for MHCs assignments) as

input, and in a fully automated manner deconvolute the individual MHC restriction of all MA sequences while learning the binding specificities of all the MHCs present in the training set.

The NNAlign_MA method was benchmarked on a large and diverse dataset, covering human MHC class I and class II, and bovine class I (BoLA) data. For all cases, the method was demonstrated to have unprecedented performance compared to state-of-the-art methods, achieving a complete deconvolution of binding motifs contained within poly-specificity MS eluted ligand data and an improved predictive performance for identification of both eluted ligands and T cell epitopes.

Given its very high flexibility and ease of use, we expect NNAlign_MA to serve as an effective tool to increase our understanding of the rules for MHC antigen presentation and guide the development of novel T cell-based therapeutics.

Introduction

Major Histocompatibility Complex (MHC) molecules play a central role in the cellular immune system as cell-surface presenters of antigenic peptides to T cell receptors (TCR). There are two main variants of MHC molecules: MHC class I, that mainly presents peptides derived from intracellular proteins; and MHC class II, that presents peptides generated predominantly from exogenous proteins. Upon presentation, the peptide-MHC complex (pMHC) is scrutinized by T cells and an immune response can be initiated if certain interactions between the pMHC and TCR are established.

The collection of peptides presented by MHC molecules is referred to as the immunopeptidome. Because of the extreme polymorphism of the MHC, the immunopeptidome presented by each individual is unique, and as a result our immune systems are individualized.

Due to the essential role of the MHC in defining immune responses, large efforts have been dedicated to understanding the rules that shape the immunopeptidome, as well as its alterations in disease – either as a result of pathogen infection or cancerous mutation[1]. A crucial step towards defining the repertoire of peptides presented by an individual is the characterization of the binding preferences of MHC molecules. The peptide-binding domain of MHC molecules consist of a groove, with specific amino acid preferences at different positions. MHC class I, by and large, loads peptides between eight and thirteen residues long[2,3]. MHC class II molecules have an open binding groove at both ends and can bind much longer peptides, and even whole proteins[4,5].

Peptide-MHC binding affinity (BA) assays represented the first attempts of studying binding preferences of different MHC molecules *in vitro*[6,7]. Despite being an informative technique, BA characterization ignores many in-vivo antigen processing and presentation features, such as crucial steps involving protein internalization, proteases digestion, peptide transport, peptide trimming, and the role of different chaperones involved in the folding of the pMHC complex[8]. Also, BA assays most often are conducted one peptide at a time, thus becoming costly, time-consuming, and low-throughput.

Recently, advances in liquid chromatography mass spectrometry (in short, LC-MS) technologies have opened a new chapter in immunopeptidomics. Several thousands of MHC-associated eluted ligands (in short, EL) can now be sequenced in a single experiment, while maintaining the biological background of the systems under study[9]. In numerous assessments, MS EL data has proven to be a rich source of information for both rational identification of T cell epitopes[10,11] and learning the rules of MHC antigen presentation[12,13].

In this context, we have demonstrated how a modeling framework that integrates both BA and EL data achieves superior predictive performance for T cell epitope discovery compared to models trained on either of the two data types alone[13,14]. In these studies, the modeling framework was an improved version of the NNAlign method[15], which incorporated two output neurons to enable prediction on both BA and EL data types. In this setup, weights are shared between the input and hidden layer for the two input types, whereas the weights connecting the hidden and output layer are specific for each input type. This weight sharing allows information to be transferred between these two data types, and results in a boost in predictive power. For MHC class I, we have demonstrated how this framework can be extended to a pan-specific model, capturing the specific antigen presentation rules for any MHC molecule with known protein sequence, including molecules characterized by limited, or even no, binding data[14,16,17].

Except genetically engineered cells, all antigen-presenting cells express multiple MHC alleles on their surface. The antibodies used to purify peptide-MHC complexes in MS EL experiment are mostly pan or locus specific, and the data generated in an MS experiment is thus inherently poly-specific – i.e. it contains ligands matching multiple binding motifs. For instance, in the context of the human immune system, each cell can express up to six different MHC class I molecules, and the immunopeptidome obtained using MS techniques will thus be a mixture of all ligands presented by such MHCs[12]. The poly-specific nature of MS EL libraries constitutes a challenge in terms of data analysis and interpretation, where, in order to learn specific MHC rules for antigen presentation, one must be able to associate each ligand to its presenting MHC molecule(s) within

the haplotype of the cell line.

Several approaches have been suggested to address this task, including experimental setups that employ cell lines expressing only one specific MHC molecule[10], and approaches aimed at inferring the MHC association using prior knowledge of MHC specificities[18] or by means of unsupervised sequence clustering[19]. For instance, GibbsCluster[20,21] has been successfully employed in multiple studies to extract binding motifs from EL datasets of several species, both for MHC class I and MHC class II[5,22–24]. A similar tool, MixMHCp[19] has been applied to the deconvolution of MHC class I EL data with performance comparable to GibbsCluster.

However, neither of the two methods are able to fully deconvolute the complete number of MHC specificities present in each dataset, especially for cell lines containing overlapping binding motifs and lowly expressed molecules (as in the case of HLA-C). Also, one must note that the association of each of the clustered solutions to a specific HLA molecule for both methods must be guided by prior knowledge of the MHC binding motifs, for instance by recurring to MHC-peptide binding predictions[14]. Thus, two methods require manual intervention for deconvolution and allele annotation.

A recent method, MixMHCpred[25], was suggested to overcome this limitation and deconvolute and assign MHC restrictions to the clustering solutions in an unsupervised manner. MixMHCpred employs a mixture model algorithm (similar to MixMHCp) to generate peptide clusters and binding motifs for a panel of poly-specificity MS datasets, and next link each cluster to an HLA molecule based on allele co-occurrence and exclusion principles. While MixMHCpred constitutes a substantial step forward for aiding our ability to analyze and interpret MS EL data, it is clear that the approach has some substantial shortcomings. First and foremost, the success of the method is tied to the ability of the underlying MixMHCp method to identify all the binding motifs in a given MS dataset. As discussed above, unsupervised clustering tends to underestimate the number of actual motifs, especially for cell lines containing MHC alleles with similar binding motifs, and for molecules characterized by low expression levels[25]. Secondly, successful HLA labeling of the obtained clusters is limited by allele co-occurrences and exclusions across multiple datasets. While one may argue that this shortcoming is destined to wane as more immunopeptidomics datasets are accumulated in public databases, there currently remain multiple cases when co-occurrence and exclusion principles fail to completely deconvolute peptidome specificities[25].

Inspired by the MixMHCpred method and earlier success of the pan-specific NNAlign framework for modeling peptide-MHC binding, in this paper we present a novel machine learning algorithm

that enables complete MHC clustering and labelling. The algorithm is an extension of the NNAlign neural network framework[15,26,27], and is capable of taking a mixed training set composed of single-allele (SA) data (peptides assigned to single MHCs) and multi-allele (MA) data (peptides with multiple options for MHCs assignments) as input and fully deconvolute the individual MHC restriction of all MA sequences while learning the binding specificities of all the MHCs present in the training set.

Results

As stated above, a key issue associated with the interpretation and analysis of LC-MS derived MHC eluted ligand datasets (EL data) stems from the challenge of deconvoluting and linking each ligand back to the presenting MHC molecule(s) of the investigated cell lines. The aim of this work is to describe a framework resolving this challenge, and also illustrate how such a framework can be applied to effectively integrate EL data in a semi-supervised manner into machine-learning models for improved prediction of MHC antigen presentation. In the following, we describe the novel framework and showcase how it compares against current state-of-the-art strategies on the analysis of three large datasets of human (class I and class II) and cattle MHC class I and derived EL datasets including evaluation on independent epitope benchmarks.

NNAlign_MA, training on multi-allele data

The algorithm is an extension of the the NNAlign neural network framework, and is capable of taking a mixed training set composed of single-allele data (SA, peptides assigned to single MHCs) and multi-allele data (MA, peptides that are assigned to multiple MHCs), and fully deconvolute the individual MHC restriction of all MA sequences while learning the binding specificities of all the MHCs present in the training set. We term this novel algorithm NNAlign_MA.

The algorithm consists of various critical steps (see Figure 1 for a schematic overview). First, a neural network is pre-trained on SA data only during a burn-in period, using the NNAlign framework underlying the NetMHCpan-4.0 method, which integrates binding affinity (BA) and EL data with sequence information of the MHC molecules. This results in a pan-specific model with power to infer binding specificities also for MHC molecules not included in the SA dataset[14,28]. After this training (from now on referred to as “pre-training”), the data in the MA datasets are annotated. That is, binding for each positive peptide in the MA dataset is predicted (using the ligand likelihood prediction value from the pre-trained model) to all the possible MHC molecules of the given cell line and the restriction is inferred from the highest prediction value (for details see Materials and Methods). For negative MA data, a random MHC molecule from the given cell line is tagged. Next, the SA and now single-MHC annotated MA data are merged, and the model is retrained on the combined data. This MHC annotation for the MA data is repeated after each training round. For details on model hyper-parameters and model training setup, see Materials and Methods.

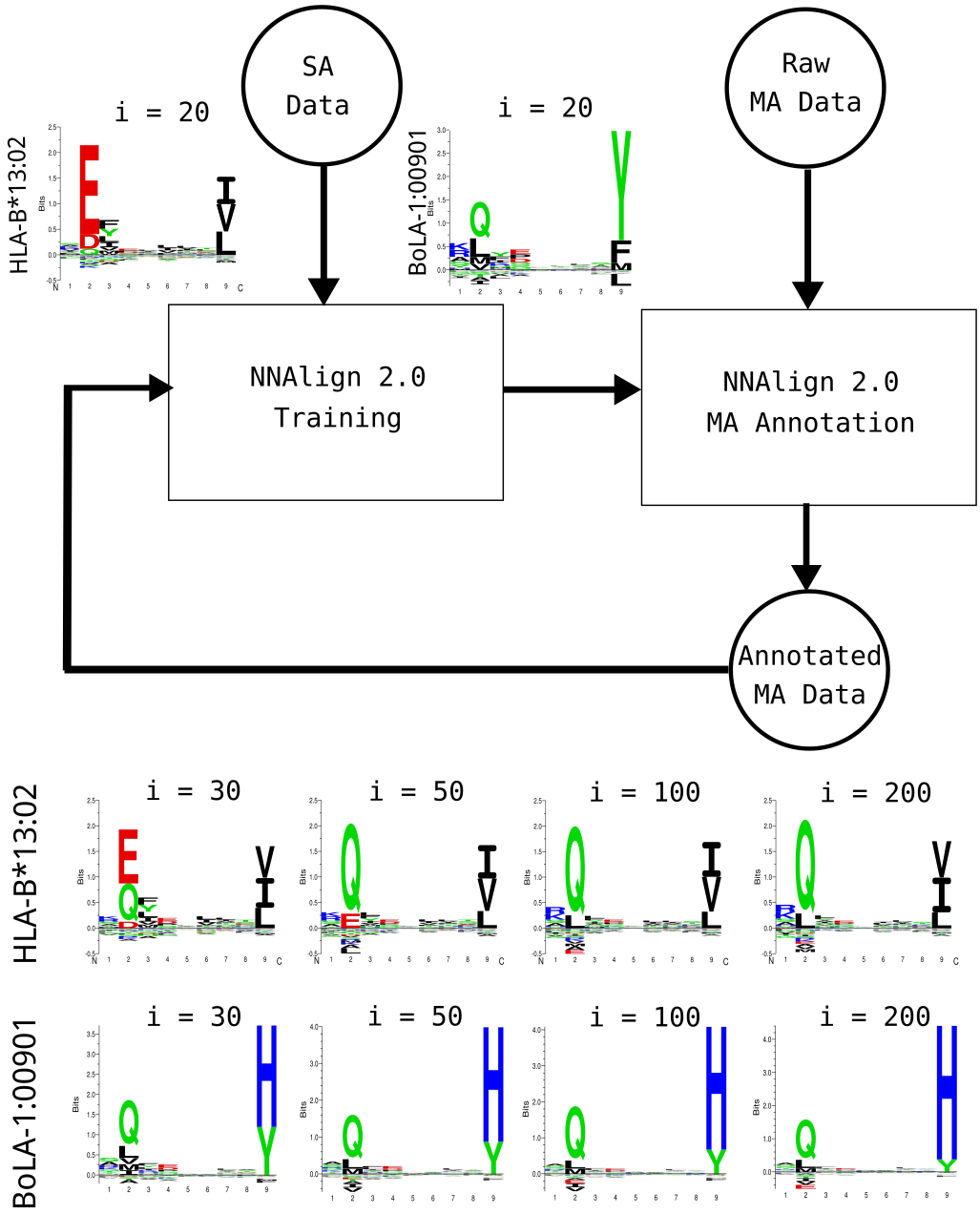


Figure 1. NNAlign_MA framework. An initial model is pre-trained using SA data, next MA data are annotated and merged with the SA data and the training repeated. In the upper left part are displayed examples of binding motifs of two MHC molecules after pre-training (i=20). In the lower part of the figure are shown the changes to predicted binding motifs of the same two MHC molecules as NNAlign_MA iterates over the data. Here i gives the number of iterations.

HLA-I benchmark

To benchmark this novel approach, we trained a model on the complete set of SA data described in the NetMHCpan-4.0 paper, combined with an extensive set of MA data covering 50 different cell lines with typed HLA allotypes described in Bassani-Sternberg et al.[25]. After training, each MA data point ends up being annotated to one single MHC molecule, and from this annotation the distribution of ligands associated with each HLA molecules was recovered. Afterwards, using the 9mer binding cores of such ligands, we constructed the logos for all the HLA alleles expressed by each of the cell lines under study (Figure 2).

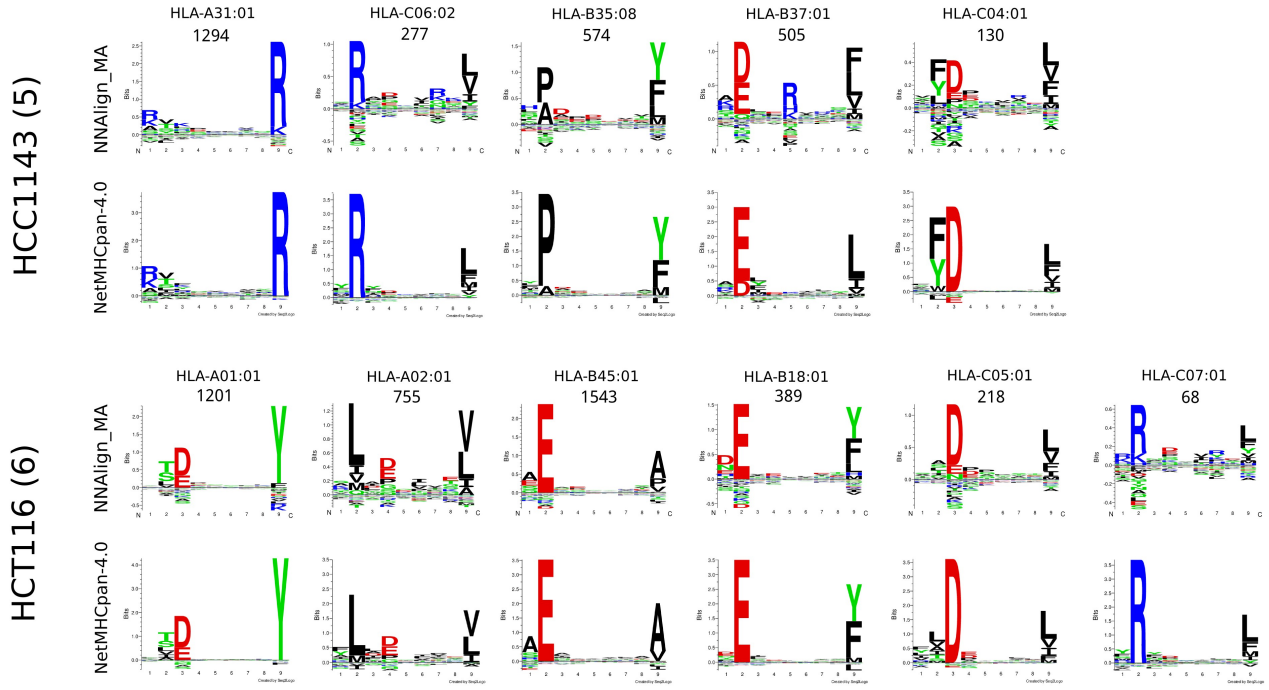


Figure 2. Sequence clustering and labeling comparison between NNAlign_MA and NetMHCpan-4.0 for the HCC1143 and HCT116 cell lines. Motifs corresponding to NNAlign_MA were constructed based on ligands from the given MA data set (cell line) predicted using cross-validation to be restricted by each HLA molecule; the quantity of peptides associated to each HLA molecule is given on top of the corresponding logos. Motifs for NetMHCpan were obtained from http://www.cbs.dtu.dk/services/NetMHCpan/logos_ps.php. The total amount of different HLAs expressed by each cell type is stated next to their names.

As seen on Figure 2, our approach was capable of not only clustering the EL data into a set of groups matching the number of expressed HLA alleles in each cell line, but also to accurately assign each group to a single corresponding allele. Referring to the original publication, MixMHCpred was capable of deconvoluting all HLA specificities of the HCC1143 cell line, but could not assign two of the five clustering groups to their corresponding allele (corresponding to HLA-A31:01 and HLA-B37:01), and for HCT116, MixMHCpred was able to deconvolute and assign only four of the six expressed alleles.

Given these highly encouraging results in terms of sequence clustering and HLA assignment, we next conducted a full scale performance evaluation of the proposed method. To this end, we first compared the performance of NNAlign_MA trained on the complete HLA-I dataset (referred to as the MA model) to the performance when trained only on the subset of SA data (referred to as the SA model). The results of this benchmark are shown in Figure 3. Here, the performance on the SA datasets of the SA model, and the performance of both the SA and MA data of the method trained on the combined SA and MA data, were extracted from the cross-validated predictive performance. These results are presented in Figure 3A, and demonstrate how the model trained on the MA data exhibited a significantly higher performance ($p < 10^{-15}$, binomial test) when evaluated on the MA

data (average AUC of 0.9660), in comparison to the SA model (average AUC of 0.9563). On the other hand, the MA and SA models showed comparable (p = 0.08, binomial test) predictive performance when evaluated on the SA data (averaged AUC of 0.9684 versus 0.9683).

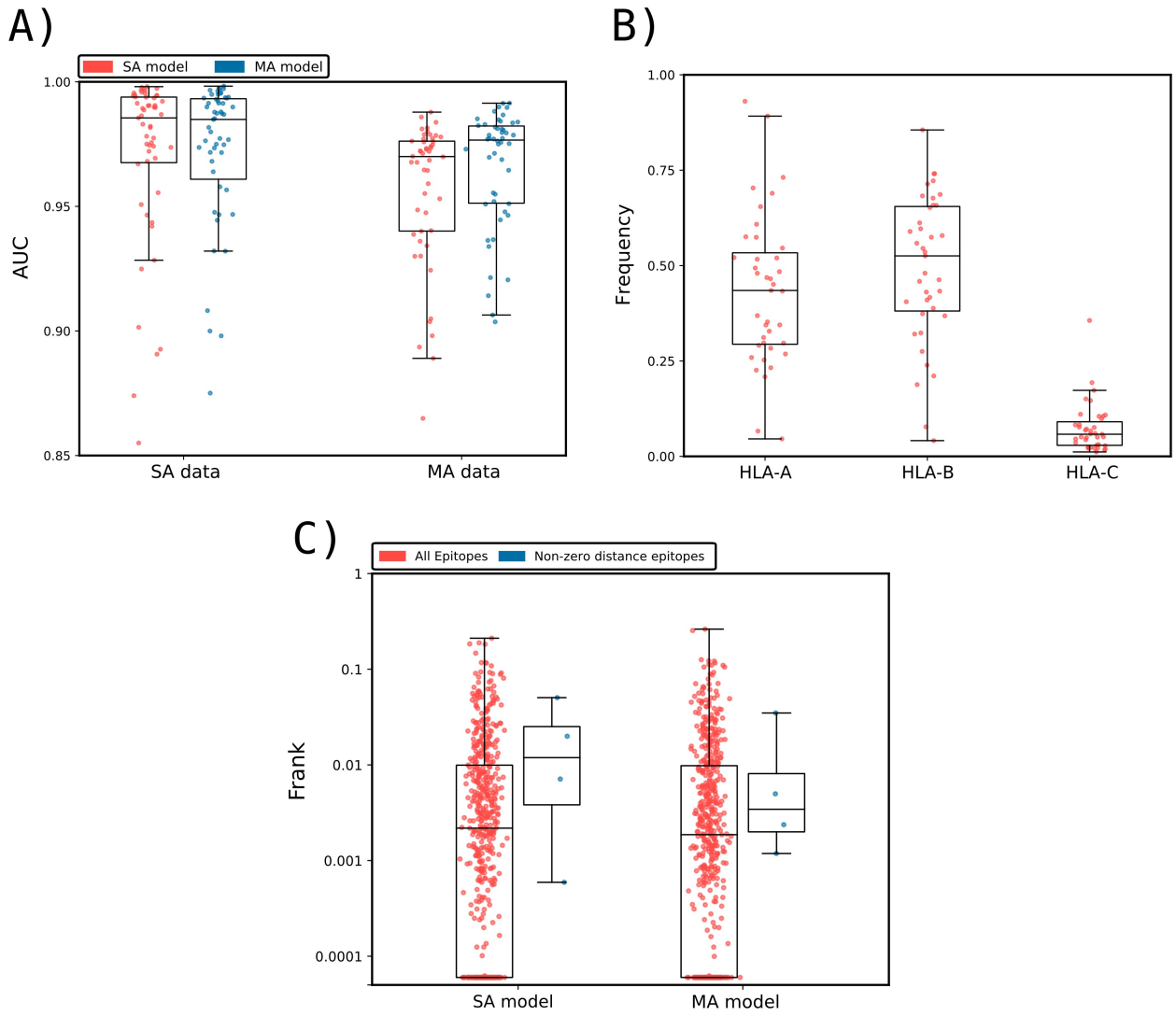


Figure 3. Benchmark of the HLA-I data. A) Performance for the Single-Allele (SA) EL and Multi-Allele (MA) datasets on the SA and MA trained models, expressed in terms of AUC. Each point corresponds to one SA or MA data set. SA data corresponds to 55 single-allele EL dataset. MA data consist of EL data from 50 different cell lines, each expressing more than one MHC molecule. To evaluate the MA data, each data point was assigned the the highest prediction value across all possible MHC restrictions in the given cell line dataset (for further details, see Materials and Methods). B) The relative peptidome size for the three loci (HLA-A, HLA-B and HLA-C) as predicted by NNAlign_MA for the different MA datasets. Each data point is the relative proportion of ligands in each MA dataset predicted to be restricted by a HLA from the given locus. The HLA restriction for each ligand data was estimated from the evaluation performance of the cross-validation as described in Material and Methods. Only datasets where HLA expression is annotated for all three loci are included. C) Frank values for the epitope evaluation dataset on models trained with only SA data (SA model) and with SA and MA data (EL model). Red dots correspond to epitopes restricted to HLAs that were part of the SA training data, while blue points refer to Frank values for epitopes with HLA restrictions absent from the SA training set. For visualization, Frank values of 0 are displayed with a value of 0.00006.

Next, we investigated how the peptidome for the EL data in each dataset was distributed among the alleles of the three loci HLA-A, HLA-B and HLA-C. To do this, we extracted the number of ligands predicted by NNAlign_MA to be restricted by HLA-A, HLA-B and HLA-C for each cell line present in the EL dataset and then calculated their corresponding relative frequencies (this is, the proportion of sequences associated to a given loci relative to the total amount of peptides present in the cell line). The result of this analysis is shown in Figure 3 B, and confirms the general notion that HLA-A and HLA-B have comparable peptidome repertoire size, whereas the peptidome size of HLA-C, in comparison, is substantially reduced.

Some clear outliers are present in the figure where either the HLA-A or HLA-B peptidome repertoires are highly reduced compared to the median value. A few such examples are the HL-60, CA46, Mel-624 and HEK293 cell lines where either the entire HLA-A or HLA-B locus appears to have been deleted or made non-functional (for details see Supplementary Table 4). These observations are in agreement with results from earlier studies, hence suggesting a strong power of NNAlign_MA also for identification of loss or down-regulation of HLA expression directly from EL datasets.

To further investigate the predictive power of this novel framework, we employed an evaluation set of epitopes of length 8-13 extracted from IEDB (see Materials and Methods for details on how this evaluation set was constructed). Here, we divided the epitope dataset into two subsets, where one set contained the epitopes restricted to HLAs that were part of the SA training dataset, and one set where the HLAs were not present in the SA training data. The results of the evaluation on these two datasets for the two methods (trained on SA and MA data) are shown in Figure 3 C. As observed from this figure, the Frank values (defined in Material and Methods) of the MA model displayed a minor but consistent and statistically significant improved performance for the epitopes restricted to HLA molecules that are part of the SA dataset (red points) (average Frank values of 0.0107 and 0.0116 for the MA and SA models, respectively, $p = 0.0215$, binomial test). Further, when evaluated on the small set of epitopes whose HLA restrictions are not present in the SA data (blue points), the performance gain of the MA model is substantially increased (average Frank values of 0.0108 and 0.0195 for the MA and SA models, respectively). These results demonstrate how the model is capable of benefiting from the integration of MA data also when it comes to prediction of T cell epitopes.

All the above experiments were conducted in a setting where the MA data shared high overlap with the SA data. By way of example, over 75% (51 of 67) of the alleles in the MA dataset were part of the SA data, and 94% (63 of 67) share a distance of less than 0.1 as measured from the similarity between pseudo sequences (for details of this distance measure see Materials and Methods), a

distance threshold earlier demonstrated to be associated with high predictive accuracy of the pan-specific prediction model[29]. Given these observations, the high performance of NNAlign method is to some extent expected.

To thus further demonstrate the predictive power of the proposed modeling framework in settings where information from SA data is limited or absent, we conducted an experiment where parts of the SA data were left out from the training data. In short, we removed all SA data for HLA molecules belonging to or similar to alleles in the HLA-A2 and HLA-A3 supertypes[30] (for details in Materials and Methods). Next, SA and MA models were trained as described above on the remaining data, and the predictive performance evaluated on the left out SA EL data. Note, that this evaluation was done respecting the data partitioning of the cross-validation to avoid introducing a bias in favor of the MA model. The performance was estimated in terms of AUC0.1 resulting in average values of 0.610 versus 0.844 for the SA and MA models respectively (for details on the performance values see Supplementary Table 5).

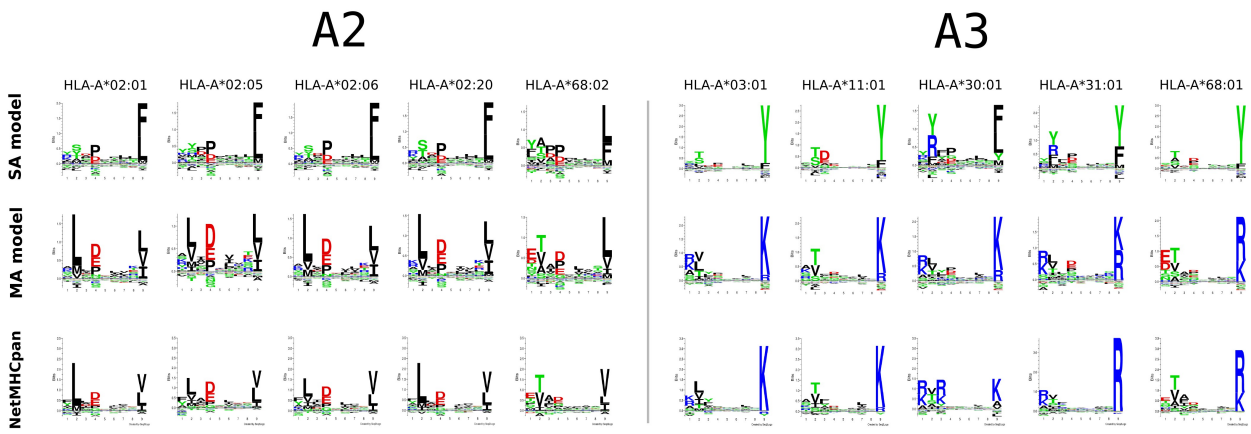


Figure 4. Binding motifs for A2 and A3 supertype alleles from the MA EL dataset predicted by the SA, MA and NetMHCpan modes. SA and MA models were trained as described in the text on SA data excluding similarity to the 10 alleles shown in the plot. Motifs for NetMHCpan were obtained from http://www.cbs.dtu.dk/services/NetMHCpan/logos_ps.php.

Also, binding motifs for the alleles in the MA data from the A2 and A3 supertypes were estimated for the two models and compared to the motifs reported by NetMHCpan (see Figure 4) demonstrating a high overlap between the MA and NetMHCpan motifs, and likewise very poor agreement to the SA motifs. These analyses demonstrate the high power of the proposed model to accurately extract binding motif information from MA datasets also in situations where no explicit information about the binding preferences of the investigated molecules is part of the SA training data.

BoLA-I benchmark

To fully appreciate the predictive power of the NNAlign_MA framework in a more challenging scenario, we next turned to the BoLA (Bovine Leukocyte Antigen) system. Due to the fact that peptide binding data is more scarce for BoLA compared to HLA, and that the relative expression of MHC molecules within a given cell line varies in a more dramatic manner for the bovine system, compared to humans, working within this system may allow us to better appreciate and outline the strength and potential limitations of the NNAlign_MA framework.

In a previous work, the NNAlign model was trained using SA BA (including binding affinity data for 7 BoLA molecules) and EL data from 3 BoLA-I homozygous cell lines describing the BoLA haplotypes A10, A14, and A18[16]. Due to the prior limitation of the NNAlign framework only admitting SA data for training, in this work the EL MA data had to be first deconvoluted using GibbsCluster and then manually annotated to the individual BoLA molecules of each cell line by visual inspection. Since this earlier publication, we have generated MA EL data for additional 5 cell lines. Using this data, we trained the NNAlign_MA framework on the SA data combined with MA data from the total of 8 BoLA cell lines (for details on these datasets, refer to Materials and Methods).

After training, we proceeded to investigate the different binding motifs captured by the model. In particular, we were interested with the motifs of the BoLA molecules shared between multiple haplotypes. One such example is BoLA-2*02501, present in the A14, and A15 haplotypes. While the motif for this molecule in our earlier work showed a strong preference for G/Q and L at P2 and P9 respectively, the new NNAlign framework captured a completely different signal, with a consistent proline (P) signal on most positions (see Supplementary Figure 1). Also, the number of ligands assigned to BoLA-2*02501 was extremely low for the three cell lines expressing this molecule (less than 1.5% in all three cases, data not shown). A prime source of this result is the very large distance (0.426) of BoLA-2*02501 to the SA training data (for details on this distance measure refer to Materials and Methods). In comparison, the maximum distance between any molecule in the MA data to the SA training data for the HLA system is less than 0.13. These large distances have two strong impacts on the predictive behavior of NNAlign_MA. First and foremost, the model pre-trained on the SA is expected to have limited power to predict the binding motif of the BoLA-2*02501 molecule (in the pre-trained model, BoLA-2*02501 has a preference for P at P2). Secondly, the prediction values of the pre-trained model will be lower for this molecule compared to molecules that share higher similarity to the SA data used for pre-training (data not shown). To deal with the latter of these two issues, we devised a rescaling scheme for the prediction score in the MA annotation step of the NNAlign_MA framework, and rescaled the raw prediction by comparing it to a score distribution obtained from a large set of random natural

peptides (for details see Materials and Methods). This score distribution was recalculated prior to each MA annotation. Including this rescaling step, the number of ligands estimated by cross-validation to be assigned to BoLA-2*02501 from the three cell lines increased to 13% on average (data not shown). However, investigating the motifs from the ligands predicted to be associated with BoLA-2*02501 from the A14 MA data to that from A12/A15 and A15 MA data, an inconsistency became apparent (see Figure 5 upper panel). Here, the motif obtained from the A14 MA data showed an additional preference for G at P2 that was completely absent for the motif obtained from the A15 and A12/A15 MA data.

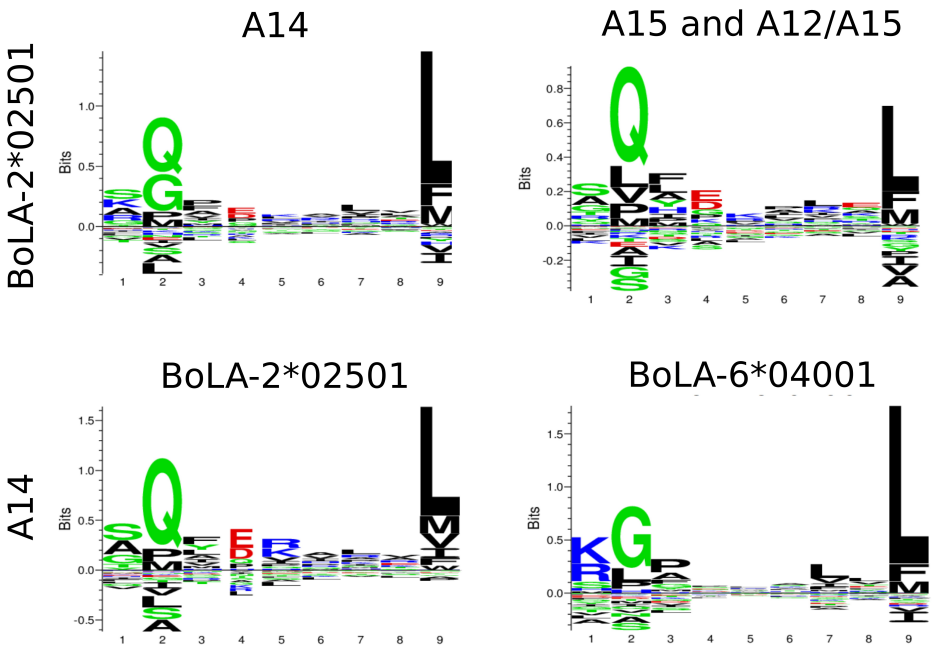


Figure 5. Upper row: binding motifs for BoLA-2*02501 obtained by NNAlign_MA trained including three BoLA alleles (BoLA-1*02301, BoLA-4*02401, and BoLA-2*02501) for the A14 MA data from ligands predicted using cross-validation to be restricted by BoLA-2*02501 from the A14 MA data (left panel), and from the A15 and A12/15 MA datasets (right panel). Lower row. Binding motifs for the BoLA-2*02501 (left panel) and BoLA-6*04001 (right panel) molecules as estimated from ligands in the A14 MA data as predicted by NNAlign_MA using cross-validation when expanding the list of alleles in A14 to include BoLA-6*04001.

Going back to the original publication describing the BoLA allele expression profile in the A14 haplotype[31], suggested an explanation to the apparent inconsistencies in the predicted BoLA-2*02501 binding motifs. In this paper, A14 was found to express 4 and not 3 BoLA alleles, as was assumed in our earlier publication and used in the first NNAlign_MA analysis. The extra allele expressed is BoLA-6*04001. After including this extra allele in the A14 haplotype and retraining the model, we arrived to the binding motifs displayed in the lower panel of Figure 5, showing a motif for BoLA-2*02501 consistent with the motif identified in the A12/A15 and A15 MA data (right upper panel of Figure 5), and a likewise well-defined motif for BoLA-6*04001 (right lower panel of Figure 5). These results clearly suggest that the motif earlier reported for BoLA-2*02501 (Supplementary

Figure 1) was a mixture of the motifs of BoLA-2*02501 and BoLA-6*04001.

Overall, these results illustrate the high predictive power of NNAlign_MA in situations where i) the MA data are characterized by molecules exclusively present in single MA datasets (as in the case for BoLA-6*04001), or ii) single molecules are shared between multiple MA datasets (as is the case for BoLA-2*02501). In both of these situations, the method was capable of correctly deconvoluting the binding motifs even though the molecules in question share limited similarity to the SA data used for pre-training and the predictive power of the pre-trained model hence is expected to be poor. Note, that the inclusion of the rescaling step in the annotation of the MA data did not alter the results from the earlier HLA analyses due to the very high overlap here between the HLAs of the MA and SA data (data not shown).

Another important observation from the results of the analysis and model training on the BoLA EL data, is the predicted motif for the BoLA-1*00901 molecule contained within the A15 haplotype. In Figure 6, we show the binding motif for this molecule, as predicted by the state-of the-art method NetBoLApan[16]. BoLA-1*00901 shares limited overlap with the SA training data (distance to the SA data D=0.137), and the motif predicted by the NNAlign_MA method after the pre-training on the SA data share, as expected, high similarity to the motif predicted by NetBoLApan. This pre-trained motif is however altered substantially after the training of the model on the MA BoLA data, resulting in a strong preference for Histidine at PΩ (Post-training motif in Figure 6). Aiming to validate the accuracy of the motif predicted for BoLA-1*00901, we performed *in vitro* binding assays of a combinatorial peptide library to the BoLA-1*00901 molecule (for details see Materials and Methods). The *in vitro* binding motif showed very high similarity to the Post-training motif predicted by NNAlign_MA, illustrating the power of the method to accurately infer binding motifs from MA data also in situations where the relevant information in the SA data is limited.

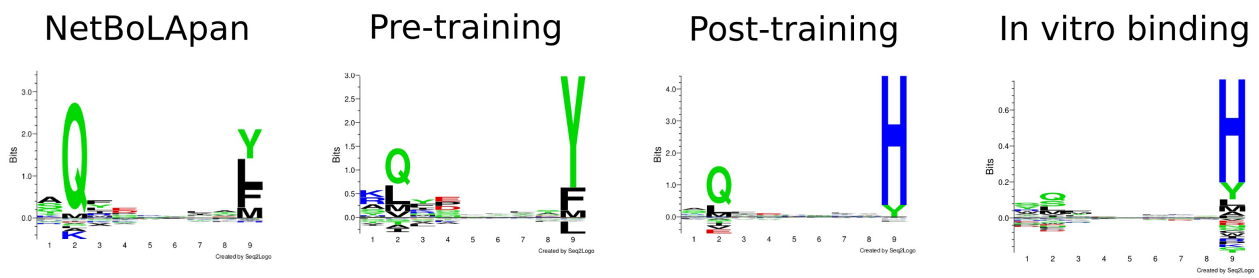


Figure 6. Binding motifs for the BoLA-1*00901 molecule estimated by different *in silico* and *in vitro* binding methods. Binding motifs for the three in silico methods were estimated from the top 0.1% of 1,000,000 random natural 9mer peptides with predicted binding by the given method, to the given BoLA molecule. The In vitro binding motifs were estimated using a position scanning combinatorial peptide library, as described in Materials and Methods. The three in silico methods are: NetBoLApan[16], trained including EL data for the cell lines A10, A14 and A18; Pre-training, the NNAlign_MA method pre-trained on SA data; Post-training, the NNAlign_MA method after completing the training including MA data.

Having demonstrated the power of the proposed model also for the BoLA system, we next turn to evaluating its predictive power on a set of experimentally validated BoLA restricted epitopes. The benchmark was performed as previously described[16,32]. In short, we predict binding for all overlapping 8−11mer peptides from the source protein of the epitopes to the known BoLA-I restriction molecule. Then, we report the performance for each epitope is reported as the Frank score. The result of this evaluation is shown in Table 1 for NNAlign_MA, NetBoLApan and earlier versions of NetMHCpan trained without the inclusion of BoLA specific EL data and confirms the high performance of the proposed model. Overall, the performance of the NNAlign_MA model is comparable to that of NetBoLApan. However, the NNAlign_MA model is trained and constructed in a fully automated and unsupervised manner, which is in strong contrast to NetBoLApan, where each EL dataset had to be first analyzed using GibbsCluster, and next each identified cluster of ligands manually assigned to a BoLA molecule expressed in the given cell line. Moreover, one important example where the two models show very different performances is the FVEGEAASH epitope, restricted by BoLA-1*00901. For this epitope, the Frank performance value of NetBoLApan was 0.121. In other words, the true positive is found 12.1% down the list of candidate peptides predicted by NetBoLApan. Including the novel BoLA EL data and training the model using the NNAlign_MA framework, the performance was improved to a Frank value of 0.000 - the epitope is the single top candidate predicted by NNAlign_MA. This result aligns very well with the experimental binding motif analysis of the BoLA-1*00901 molecule, exhibiting a strong preference for H at the C terminal (Figure 6). In summary, the results displayed in Table 1 demonstrate the very high predictive power for rational epitope discovery of the model trained including the BoLA EL data. The average Frank value of NNAlign_MA over the 16 epitopes is 0.0033, meaning that on average 99.67% of the irrelevant peptide space can be excluded by the prediction model while still identifying 100% of the epitopes. It has earlier been suggested that prediction methods trained on EL data have poor performance for identification of cysteine containing epitopes due to the bias against this amino acid in EL data[10,25]. Likewise, it has been suggested that the NNAlign framework can compensate for this bias by replacing the C residues with the wildcard character X when using the models for epitope prediction[33]. To investigate to what extend this was the case also here, we repeated the BoLA epitope benchmark replacing all cysteine residues in the source protein sequences (and epitopes) with an X. The results of this analysis are included as C2X in Table 1, and confirm in all cases an increased performance for cysteine containing epitopes leading to an overall improved average Frank performance of 0.0028.

Table 1. Predictive performance of NNAlign_MA and NetBoLApan on a set of known CD8 epitopes.

	Allele	Epitope	Antigen	Npep	BoLApan	NNAlign_MA	C2X
T. parva	BoLA-6*01301	VGYPKVKEEML	Tp1	2138	0.0098	**0.0070**	**0.0070**
	BoLA-6*04101	EELKKLGML	Tp2	662	**0.0000**	**0.0000**	**0.0000**
	BoLA-2*01201	SSHGMGKVGK	Tp2	662	0.0060	**0.0045**	0.0121
	BoLA-T2c	FAQSLVCVL	Tp2	662	0.0136	0.0151	**0.0060**
	BoLA-2*01201	QSLVCVLMK	Tp2	662	0.0015	0.0015	**0.0000**
	BoLA-AW10	TGASIQTTL	Tp4	2282	**0.0000**	**0.0000**	**0.0000**
	BoLA-1*00902	SKADVIAKY	Tp5	586	**0.0000**	**0.0000**	**0.0000**
	BoLA-T7	FISFPISL	Tp7	2850	0.0172	**0.0112**	0.0119
	BoLA-3*00101	CGAELNHFL	Tp8	1726	0.0029	0.0035	**0.0017**
	BoLA-1*02301	AKFPGMKKS	Tp9	1302	0.0054	**0.0023**	0.0031
BHV	BoLA-1:00901	FVEGEAASH	ICP4	5350	0.1215	**0.0000**	**0.0000**
	BoLA-3:00201	AGPDLQLARL	ICP4	5350	**0.0000**	**0.0000**	**0.0000**
	BoLA-3:00201	TTPEILIEL	Circ	1006	**0.0000**	**0.0000**	**0.0000**
	BoLA-3:01701	TGARAGYAA	ICP4	5350	0.0350	**0.0013**	**0.0013**
	BoLA-4:02401	GAFCPEDW	ICP22	1214	0.0066	0.0058	**0.0008**
	BoLA-2:01801	APAPSPGAL	Circ	978	0.0020	**0.0000**	**0.0000**
	Average				0.0138	0.0033	0.0028

Performance values are reported as Frank calculated as described in the text. Epitope data for *Theilera parva* (*T. parva*) and Bovine Herpes Virus (BHV) were obtained from three sources[16,32,34]. The lowest Frank value for each epitope is highlighted in bold. C2X gives the Frank values for a benchmark were all C's in the source proteins (and epitopes) were placed with X.

HLA-II benchmark

To prove that the ability of NNAlign_MA to deal with MA data also extends to MHC II, a separate study was conducted on a set of MHC II BA and EL data (for details see Materials and Methods). The results of the cross-validation evaluation of this study are depicted in Figure 7 A. This figure compares the performance (in terms of AUC) of NNAlign_MA trained on SA data only and on the full dataset including MA data (SA model and MA model respectively). Both models were evaluated on the full dataset (including SA and MA data) and the AUC performance was calculated per dataset. As for HLA I, the performance of either model was reported based on the prediction of ligands from SA data or MA data. The results of this evaluation were similar to those obtained for the HLA-I data. When evaluated on SA data, the SA model demonstrated a modest (and statistically insignificant ($p = 0.0703$, binomial test), performance gain compared to the MA model. However, when it comes to comparing performance on MA data, the MA model significantly outperformed the SA model ($p = 1.9*10^{-6}$, binomial test).

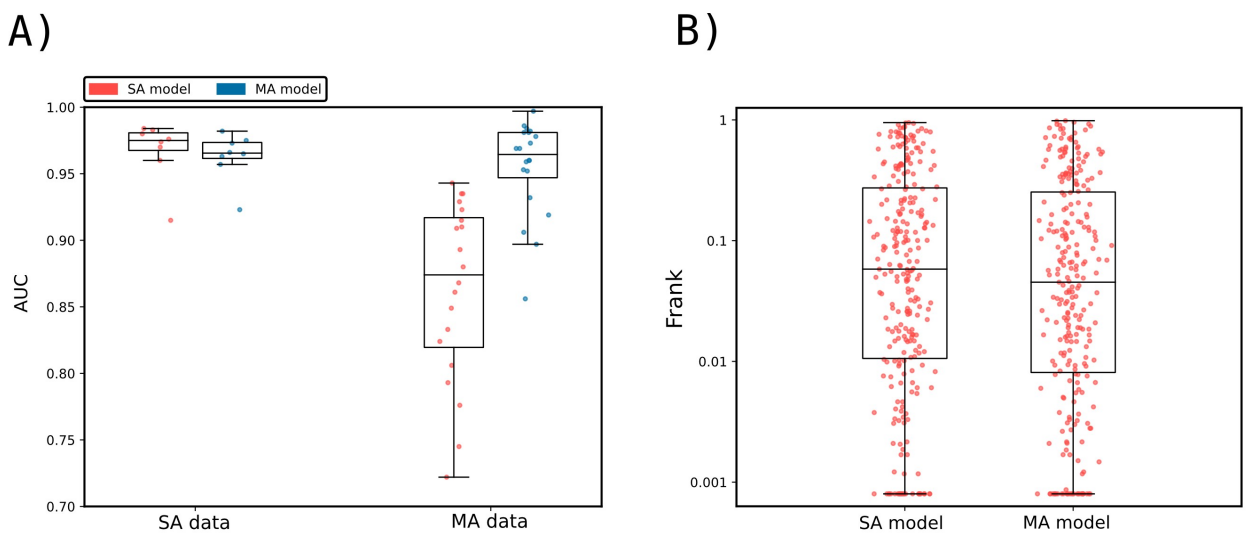


Figure 7. A) Performance of NNAlign_MA trained on SA and MA data evaluated in cross-validation. Model performance is given in terms of AUC and calculated as described in figure 4. Each point refers to a SA or MA data set. B) Epitope Frank scores of NNAlign_MA trained on SA and MA data evaluated on CD4+ epitopes from the IEDB. The plot compares the Frank distribution of the MA and SA models and each point represents the Frank of an epitope. For visualization, Frank values of 0 are displayed with a value of 0.0008.

The results of the independent CD4+ epitope benchmarking of MHC II models are depicted in Figure 7 B, where the Frank values for each epitope are shown for the SA and MA models. The median Frank is found to be significantly lower for the MA compared to the SA model (4.53 and 5.81, respectively) (p = 0.014 binomial test of a pairwise comparison of allele wise median Frank values).

Discussion

Recent advances in Mass Spectrometry have dramatically increased the throughput of immunopeptidomics experiments, with several thousands of peptides directly eluted from their cognate MHC molecule in a single experiment. This type of data has greatly changed our knowledge-base for characterizing MHC antigen processing and presentation, moving beyond the limitations imposed by the traditional *in vitro* peptide-MHC binding assays.

However, the analysis and interpretation of MS eluted ligand data is, in most situations, highly challenging. Because, in general, ligands originate from multiple MHC molecules, datasets of MS eluted ligands consist of a mixture of motifs, each corresponding to the binding specificity of one of the MHC molecule expressed by the cell line. While several tools for the deconvolution of multiple motifs have been developed, they all tend to underestimate the number of specificities in a sample, especially for haplotypes with overlapping MHC binding motifs. Even for peptidomes that can be confidently deconvoluted, the pairing between motif and MHC allele is often not trivial, and in many cases must be done manually by visual inspection – with the potential sources of error this process entails.

In this paper, we have described a fully automated approach, NNAlign_MA, that aims to resolve

these challenges. The method is an extension of the NNAlign algorithm, developed to extract and learn binding motif information from ligand datasets[26,27], either in a single receptor or pan-receptor setting. For example, the pan-receptor setting of the method was used to construct the state-of-the-art pan-specific MHC class I and class II peptide binding methods NetMHCpan[35] and NetMHCIIpan[36]. A limitation of NNAlign, when used in a pan-receptor setting, was the requirement that each data point had to be associated with one and only one receptor. Given the poly-specificity context of MS eluted ligand data, this represents a substantial limitation in the applicability of the method.

The approach taken in NNAlign_MA to resolve this limitation is very simple. The method applies a pre-training period where only single allele data (peptide data characterized by having a single MHC association) are included. After this pre-training, the multi-allele data (peptide data characterized by having two or more MHC associations) are annotated using the current prediction model to predict binding to all MHC molecules possible for the peptide, and next defining a single MHC association from the highest prediction value. In this annotation step, multi-allele data are thus caste into a single-allele format, becoming manageable by the NNAlign method and therefore enabling training on the combined dataset. This multi-allele annotation step is iteratively performed in each training cycle.

We have applied the NNAlign_MA method to analyze and interpret three large-scale multi-allele MHC eluted ligand datasets, and demonstrated its unprecedented performance compared to state-of-the-art methods. First, we applied the method to analyze a collection of multi-allele HLA MS eluted ligand data from 50 cell lines. We demonstrated how the method was capable of correctly identifying distinct binding motifs for each of the HLA molecules expressed in a given cell line. This result is in contrast to prior findings (using earlier state-of-the-art methods such as GibbsCluster and MixMHCp), where in most cases the algorithm fails to identify one or more motifs. Also, NNAlign_MA was capable of accurately associating each identified motif with a specific HLA molecule in all cases. These results highlighted the superior performance of the NNAlign_MA method compared to current state-of-the-art methods such as MixMHCpred, where the association of binding motifs to individual HLA molecules is achieved by identification of binding motifs uniquely present in individual cell line datasets.

Next, we demonstrated how the results of the NNAlign_MA annotation of the multi-allele dataset could be used to quantify the relative size of the peptidome associated with each HLA molecule in a given cell line. Overall, we found that HLA-A and HLA-B exhibit comparable peptidome repertoire sizes, whereas HLA-C molecules, in comparison, present relatively fewer peptides. These analyses further revealed examples of cell lines where either the HLA-A or HLA-B peptidome repertoires were highly reduced (or even completely empty). These were, in all cases, found to correspond to cell lines with locus-specific loss, or down-regulation, of HLA expression.

In terms of predictive performance, the models trained using the NNAlign_MA method were found

to significantly outperform conventional methods (trained on single-allele data), both when it comes to prediction of HLA eluted ligand data and known CD8 epitopes obtained from the IEDB. As expected, this performance gain was most pronounced for ligand/epitopes restricted by HLA molecules only characterized by multi-allele data.

To validate the predictive power of NNAlign_MA in a setting where novel specificities are included in the multi-allele data, a specificity leave-out experiment was performed. Here, entire HLA specificity groups were removed from the single-allele dataset, and the NNAlign_MA framework applied to analyze and characterize multi-allele data including HLA molecules from the removed specificity group. The result of this experiment confirmed the power of NNAlign_MA in accurately identifying binding motifs for individual HLA molecules in multi-allele data, also in situations where no explicit information about the binding preferences of the investigated molecules was part of the single-allele training data.

The HLA system has been studied in great detail over the past decades, and peptide-MHC binding data are available for hundreds of alleles. To explore the predictive power of the method for MHC systems characterized by limited data, we turned to the Bovine Leukocyte Antigen (BoLA) system, and applied NNAlign_MA to analyze MS MHC eluted ligands datasets from 8 cell lines expressing a total of 8 haplotypes and 16 distinct BoLA molecules. These BoLA molecules shared, for most parts, very low similarity to the molecules included in the single-allele data. We found that molecules with exceedingly large distance from the pre-training data were heavily penalized by the algorithm and were assigned few or no ligands. To address this bias, and inspired by the percentile rank score transformation applied in the NetMHC suite of prediction tools[35], we implemented a rescaling procedure of the prediction values in the annotation step. After applying such re-scaling, NNAlign_MA was demonstrated to accurately identify binding motifs in all BoLA datasets. Moreover, this analysis of the predicted binding motifs revealed an inconsistency in the BoLA allele annotation for the A14 haplotype, earlier suggested to express the alleles BoLA-1*02301, BoLA-4*02401, and BoLA-2*02501. A comparison of the predicted motif for BoLA-2*02501 from the A14 cell line, and the motifs obtained from two other cell lines also expressing BoLA-2*02501, suggested that the motif obtained from A14 was a convoluted mixture of two binding preferences. Revising the BoLA expression sequence data for A14 allowed the identification of an additional molecule, BoLA-6*04001. Including this molecule in the NNAlign_MA analysis resolved the inconsistency of the BoLA-2*02501 motifs and further identified a clear binding motif for BoLA-6*04001. These results demonstrate how NNAlign_MA is capable of correctly deconvoluting binding motifs present in multi-allele data even in situations where the molecule in question shares limited similarity to the single-allele data.

We also applied the NNAlign_MA model trained on the BoLA data to predict a set of previously described BoLA restricted CD8+ epitopes. The result of this benchmark demonstrated a very high

predictive power of the model, allowing the identification of epitopes within the top 0.3% of the peptides within the epitope source protein sequence. In practical terms, the application of the NNAlign_MA model would have allowed identification of the known CD8 epitopes by discarding the vast majority of the irrelevant peptide space and testing only 0.3% of all peptides that could be derived from the source proteins of the epitopes.

As a final validation, the NNAlign_MA framework was applied to MS EL data from MHC II. Also here, the models were evaluated in cross-validation and on an independent set of CD4+ epitopes and the results were in agreement with the result obtained for MHC I. That is, models trained on mixed SA and MA data types showed significantly improved performance (for predicting MA data) compared to models trained on SA data only, and a likewise significantly improved performance when evaluated on CD4+ epitopes.

Overall, we have evaluated the proposed NNAlign_MA framework on a large and diverse set of data, and have demonstrated how the method in all cases was capable of both achieving a complete deconvolution of binding motifs contained within poly-specific MS eluted ligand data, and how the complete deconvolution resulted in improved predictive performance for identification of both eluted ligands and T cell epitopes.

We have demonstrated how NNAlign_MA achieves binding motif deconvolution driven by similarity to MHC molecules characterized by single-specificity data, and by principles of co-occurrence and exclusion of MHC molecules between different poly-specificity MS eluted ligand dataset. These observations prompt for a direct application of the NNAlign_MA framework to effectively achieve broad and high accuracy allelic coverage for regions of the MHC repertoire with yet uncharacterized binding specificity. That is, sets of cell lines with characterized HLA expression should be selected to maximize allele co-occurrence, allele exclusion and allele similarities with the comprehensive set of available single-specificity data so that the NNAlign_MA motif deconvolution can be achieved in an optimal manner.

In conclusion, we believe NNAlign_MA offers a universal solution to the challenge of analyzing large-scale MHC peptidomics datasets and consequently affords the optimal way of exploiting the information contained in such data for improving prediction of MHC binding and antigen presentation. The modeling framework is readily extendable to include peptides with post-translational modifications[15,37], and signals from antigen processing located outside the sequence of the ligands[13]. Given its very high flexibility, we expect NNAlign_MA to serve as an effective tool to further our understanding of the rules for MHC antigen presentation, as a guide for improved T cell epitope discovery and as an aid for effective development of T cell therapeutics.

Methods

Peptide data

Several types of MHC peptide data for human (HLA) and bovine (BoLA) class I, and HLA class II were gathered to train the predictive models presented in this work. Peptide data was classified as single allele data (SA, where each peptide is associated to a single MHC restriction) and multi allele data (MA, where each peptide has multiple options for MHC restriction). MA data are generated from MS MHC ligand elution assays where most often a pan-specific antibody is applied for class I and either a pan-specific class II or a pan-DR specific antibody is applied for class II in the immuno-precipitation step leading to datasets with poly-specificities matching the MHC molecules expressed in the cell line under study. SA data were obtained from binding affinity assays, or from mass spectrometry experiments performed using genetically engineered cell lines that artificially express one single allele.

HLA class I: SA data - both binding affinity (BA), and MS MHC eluted ligands (EL) - was extracted from Jurtz et al.[14]. The MA data was collected from eight different sources[12,22,23,25,38–41]. Both datasets were filtered to include only peptides of length 8-14 amino acids. Additional information concerning the HLA class I MA data can be found in Supplementary Table 1.

HLA-II: BA data was extracted from the NetMHCIIpan-3.2 publication[36]. As for EL data, the Immune Epitope Database[42] (IEDB) was queried to identify publications with a large number of allele annotated EL data, both SA and MA[24,43–50]. Ligands were extracted from these publications, excluding any ligands with post translational modifications. Both BA and EL data was length filtered to include only peptides of length 13-21. Details on the composition of the HLA class II MA data are shown in Supplementary Table 2.

Regarding BoLA, SA data was extracted from Nielsen et al.[16] and the MA data was collected either from the same publication (data for the MHC homozygous cell lines expressing the haplotypes A10, A14, A18) or were generated for this study (data for the cell lines expressing the haplotypes A11/A11, A19/A19, A20/A20, A15/A15, and A12/A15). The novel EL data were generated as described earlier[16] by use of LC-MS using the pan anti-BoLA class I antibody IL-88 for immunoprecipitation. Sequence interpretation of MS/MS spectra was performed using a database containing all bovine protein entries in Uniprot combined (in cases of *Theileria parva* infected cell lines) with 4084 annotated *Theileria parva* protein sequences[51]. Spectral interpretation was performed using PEAKS 7.5[52] (Bioinformatics Solutions). All datasets were filtered to include only

peptides of length 8-14. A summary of the BoLA MA data is given in Supplementary Table 3.

Training data

Three training sets were constructed, one for each of the systems under study (Table 2). To ensure an unbiased performance evaluation on the MA data, duplicated entries between the SA EL and MA data were first removed from the SA EL dataset for each training set. Next, random peptides were extracted from the UniProt database and used as negative instances for the EL data in each case. Here, an equal amount of random negatives was used for each length, consisting of five times the amount of peptides for the most abundant length in the given positive EL dataset as described earlier[13,16]. This enrichment with random natural negative peptides was done for each individual SA and MA EL dataset. The amount of positive and negative peptides in each training set is shown in Table 2.

Table 2. Training data overview.

	SA (BA)			SA (EL)			MA (EL)			
	Positives	**Negatives**	**# MHCs**	**Positives**	**Negatives**	**# MHCs**	**Positives**	**Negatives**	**# MHCs**	**Lengths**
HLA-I	50,344	127,169	104	46,183	740,939	51	272,249	5,681,309	67	8-14
BoLA-I	50,361	150,833	7	84,717	1,644,976	-	79,668	1,650,380	16	8-14
HLA-II	55,178	76,185	59	32,510	337,720	8	15,494	152,445	16	13-21

For each training dataset (first column), the number of positive and negative instances is shown for each type of training data. SA: Single Allele; MA: Multi Allele; BA: Binding Affinity; EL: Eluted Ligands.

Evaluation data

For HLA class I, an independent evaluation data set of HLA restricted CD8+ epitopes was obtained from Jurtz et al.[14]. After removal of epitopes overlapping with the HLA-I training data, the final evaluation data consisted of 558 HLA-epitope entries. For BoLA-I, a set of BoLA restricted epitopes were obtained from Nielsen et al.[16]. For HLA-I and BoLA-I evaluation, the source protein sequence of each epitope was *in-silico* digested into overlapping 8-14mers, and the performance reported as the Frank score, i.e. proportion of peptides with a prediction score higher than that of the epitope[14]. Using this measure, a value of 0 corresponds to a perfect prediction (the known epitope is identified with the highest predicted binding value among all peptides found within the source protein) and a value of 0.5 to random prediction.

For HLA class II all CD4+ epitopes measured by Intracellular Cytokine Staining(ICS) assay were downloaded from the IEDB. The set was filtered to include only positive epitopes with four letter resolution HLA typing. Furthermore, epitopes overlapping with the HLA-II training data (100% identity) were removed. This resulted in a set of 271 HLA-II epitopes for evaluation. As for HLA-I, the Frank was used to validate the model performance, here *in-silico* digesting the source protein into overlapping of a length equal to that of the epitope.

NNAlign modeling and training hyperparameters

Models for peptide-MHC binding prediction were trained with hyperparameters and model architectures similar to those described earlier[13,15,53] for prediction of peptide-MHC binding based on the datasets described in Table 2. Positive instances in EL datasets (for both SA and MA) were labeled with a target value of 1, and negatives with a target value of 0.

In order to avoid performance overestimation and model overfitting, training sets were split into 5 partitions for cross-validation purposes using the common motif algorithm[54] with a motif length of 8 amino acids for class I (corresponding to the shortest binding mode for class I peptides) and 9 amino acids for class II (corresponding to the size of the class II binding core) as described earlier[13,14].

A single and simple yet highly critical step sets the updated NNAlign_MA method proposed here aside from its ancestors. To be able to accurately handle and annotate MA data, NNAlign_MA imposes a burn-in period where the method is trained only on SA data. After the burn-in period, each data point in the MA dataset is annotated by predicting binding to all possible MHC molecules defined in the MA dataset, and assigning the restriction from the highest prediction value (see prediction score rescaling for exceptions from this setup). After this annotation step, the SA and MA data are merged respecting the data partitioning to further train the algorithm.

In the case of HLA-I and BoLA-I, models were trained on the full set of SA and MA data as an ensemble of 50 individual networks, generated from 5 different seeds; 56 and 66 hidden neurons; and 5 partitions for cross-validation. Models were trained for 200 iterations (using early stopping), with a burn-in period of 20 iterations. For performance comparison, a SA-only model was trained for HLA-I using the same architecture and hyper-parameters by excluding all MA data from the cross-validation partitions, thus including only data for SA while respecting the training data structure.

Regarding HLA-II, default settings for MHC-II prediction as previously described[13,53] were used. Models were trained and evaluated on 5-fold cross-validation partitions defined by common motif clustering with a motif of length 9. The final ensemble of models consists of 250 networks (2, 10, 20, 40 and 60 hidden neurons and 10 random weight initiation seeds for each CV fold). Networks were trained for 400 iterations, without early stopping and using a burn-in period of 20.

Prediction score rescaling

To level out differences in the prediction scores between MHC alleles imposed by the differences in number of positive training examples and distance to the training data included in the SA dataset, a rescaling of the raw prediction values was implemented and applied in the MA data annotation. The rescaling was implemented as a z-score transformation of the raw prediction values using the relation $z=(p-\bar{p})/\sigma$, where p is the raw prediction value of the peptide to a given MHC molecules, and $\bar{p}$ and σ are the mean and standard deviation of the distribution of prediction values for random natural peptides for the MHC molecule. Here, the score distribution was estimated by predicting binding of 10,000 random natural 9mer peptides to MHC molecule in question. Next, the mean and standard deviation were estimated from the right-hand side of the distribution, iteratively excluding outliers (z-score < -3 or z-score > 3). This estimation of $\bar{p}$ and σ was repeated in each iteration round prior to annotating the MA data. As the rescaling is imposed to level out score differences between MHC molecules characterized in the SA training binding data and molecules from the MA data distant to the training data, the need for rescaling should be leveled out as the MA data are included in the training and the NNAlign_MA training progresses. To achieve this, the values of $\bar{p}$ and σ were modified to converge towards uniform values $\bar{p}_u$ and σ_u defined as the average of $\bar{p}$ and σ over all molecules in the MA dataset. This convergence was defined as $\bar{p}'=w*\bar{p}+(1-w)*\bar{p}_u$ and $\sigma'=w*\sigma+(1-w)*\sigma_u$, where $w=1/(1+e^{(x-75)/10})$ and x is the number of training iterations. With this relation, w is close to 1 after pretraining (x = 20), and converges to 0 as x passes 100 iterations.

Distance between pairs of MHC molecules

The distance, D, between MHC molecules was calculated as described earlier[28] from the sequence similarity between the pseudo sequences of the two molecules. Likewise, was the distance of an MHC molecule to the data used to train a given prediction model, defined as the closest distance to any MHC molecule included in the training data.

HLA models with removed branches

In order to quantify how MA data can boost the performance of a peptide-MHC predictor, we constructed additional models, where SA data associated with HLA molecules from the A2 and A3 supertypes where exclude from the training data. In short, this was achieved by first identifying the alleles in the MA data for the two supertypes[30], resulting in the following allele list: HLA-A*02:01, HLA-A*02:05, HLA-A*02:06, HLA-A*02:20, HLA-A*68:02, HLA-A*03:01, HLA-A*11:01, HLA-A*30:01, HLA-A*31:01, and HLA-A*68:01. Next, all data for alleles with a distance (see above) of less than 0.2 to any of the alleles in this list were removed from the SA data. Finally, two models were trained as described above on the remaining SA data, and one the remaining SA data

combined with the complete MA data, respecting data partitioning.

Sequence motif construction

Sequence binding motif were visualized as Kullback-Leibler logo plots using Seg2Logo[55]. Amino acids are grouped by negatively charged (red), positively charged (blue), polar (green) or hydrophobic (black). If not otherwise specified, binding motifs were generated from the top 0.1% of 200,000 random natural peptides (9mers for class I and 15mers for class II) as described earlier[14].

Model performance evaluation

For model comparison, the AUC (Area Under the ROC Curve) and AUC 0.1 (Area Under the ROC Curve integrated up to a False Positive Rate of 10%) performance measures were used. For a given model, each test set was predicted using the model trained during cross-validation. Next, all test sets were concatenated, and an AUC/AUC0.1 value was calculated for each MHC molecule/cell line identifier. In case of multi-allele data, the prediction score to each peptide was assigned as the maximal prediction value over the set of possible MHC molecules.

In vitro binding data

Recombinant BoLA-1*00901 and human beta-2 microglobulins (β2m) were produced as previously described[56]. In brief, biotinylated BoLA-1*00901 was generated in Escherichia coli, harvested as inclusion bodies, extracted into Tris-buffered 8 M urea and purified using ion exchange, hydrophobic, and gel filtration chromatographies. MHC-I heavy chain proteins were never exposed to reducing conditions, which allows for purification of highly active pre-oxidized BoLA molecules, which folds efficiently when diluted into an appropriate reaction buffer. The pre-oxidized, denatured proteins were stored at −20 °C in Tris-buffered 8 M urea. Human β2m was expressed and purified as previously described[57].

Nonameric peptide binding motifs were determined for BoLA-1*00901, using PSCPL as previously described[56,58,59]. Recombinant, biotinylated BoLA heavy chain molecules in 8 M urea were diluted at least 100-fold into PBS buffer containing 125I-labeled human β2m and peptide to initiate pMHC-I complex formation. The final concentration of BoLA was between 10 and 100 nM, depending on the specific activity of the heavy chain. The reactions were carried out in the wells of streptavidin-coated scintillation 384-well FlashPlate® PLUS microplates (Perkin Elmer, Waltham, MA). Recombinant radiolabeled human β2m and saturating concentrations (10 μM) of peptide were allowed to reach steady state by overnight incubation at 18 °C. After overnight incubation, excess unlabeled bovine β2m was added to a final concentration of 1 μM and the temperature was raised to 37 °C to initiate dissociation. pMHC-I dissociation was monitored for 24 h by consecutive measurement of the scintillation microplate on a scintillation TopCount NXT multiplate counter (Perkin Elmer, Waltham, MA). PSCPL dissociation data were analyzed as described[59]. Briefly, following background correction, the area under the dissociation curve (AUC) was calculated for

each sublibrary by summing the counts from 0 to 24 h. The relative contribution of each residue in each position (i.e., the relative binding, RB) was calculated as RB=(AUC_sublibrary/AUC_X9). The RB values were next normalized to sum to one for each peptide position and used as input to Seq2Logo to generate the in vitro BoLA-1*00901 binding-motif.

Acknowledgements

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, Bill and Melinda Gates Foundation [OPP1078791]; and by the Science and Technology Council of Investigation (CONICET-Argentina).

Author contributions

MN designed the study. BA, BR, and MN performed the experiments and statistical analysis. SB, NT and TC generated the experimental data. BA, BR, CB, MA and MN analyzed and interpreted the data and wrote the paper. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

References

1. Caron, E., Aebersold, R., Banaei-Esfahani, A., Chong, C. & Bassani-Sternberg, M. A Case for a Human Immuno-Peptidome Project Consortium. *Immunity* **47,** 203–208 (2017).
2. Trolle, T. *et al.* The Length Distribution of Class I–Restricted T Cell Epitopes Is Determined by Both Peptide Supply and MHC Allele–Specific Binding Preference. *J. Immunol.* **196,** 1480–1487 (2016).
3. Gfeller, D. *et al.* The Length Distribution and Multiple Specificity of Naturally Presented HLA-I Ligands. *J. Immunol.* **201,** 3705–3716 (2018).
4. Sette, A., Adorini, L., Colon, S. M., Buus, S. & Grey, H. M. Capacity of intact proteins to bind to MHC class II molecules. *J. Immunol.* **143,** 1265–7 (1989).
5. Mommen, G. P. M. *et al.* Sampling From the Proteome to the Human Leukocyte Antigen-DR (HLA-DR) Ligandome Proceeds *Via* High Specificity. *Mol. Cell. Proteomics* **15,** 1412–1423 (2016).
6. Buus, S., Sette, A., Colon, S. M., Miles, C. & Grey, H. M. The relation between major histocompatibility complex (MHC) restriction and the capacity of Ia to bind immunogenic peptides. *Science* **235,** 1353–8 (1987).
7. Townsend, A. *et al.* Assembly of MHC class I molecules analyzed in vitro. *Cell* **62,** 285–95 (1990).
8. Sadegh-Nasseri, S. & Kim, A. MHC Class II Auto-Antigen Presentation is Unconventional.

Front. Immunol. **6,** 372 (2015).

9. Caron, E. *et al.* Analysis of Major Histocompatibility Complex (MHC) Immunopeptidomes Using Mass Spectrometry. *Mol. Cell. Proteomics* **14,** 3105–3117 (2015).

10. Abelin, J. G. *et al.* Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity* **46,** 315–326 (2017).

11. Graham, D. B. *et al.* Antigen discovery and specification of immunodominance hierarchies for MHCII-restricted epitopes. *Nat. Med.* **24,** 1762–1772 (2018).

12. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J. & Mann, M. Mass Spectrometry of Human Leukocyte Antigen Class I Peptidomes Reveals Strong Effects of Protein Abundance and Turnover on Antigen Presentation. *Mol. Cell. Proteomics* **14,** 658–673 (2015).

13. Barra, C. *et al.* Footprints of antigen processing boost MHC class II natural ligand predictions. *Genome Med.* **10,** 84 (2018).

14. Jurtz, V. *et al.* NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J. Immunol.* **199,** 3360–3368 (2017).

15. Nielsen, M. & Andreatta, M. NNAlign: a platform to construct and evaluate artificial neural network models of receptor-ligand interactions. *Nucleic Acids Res.* **45,** W344–W349 (2017).

16. Nielsen, M., Connelley, T. & Ternette, N. Improved Prediction of Bovine Leucocyte Antigens (BoLA) Presented Ligands by Use of Mass-Spectrometry-Determined Ligand and in Vitro Binding Data. *J. Proteome Res.* **17,** 559–567 (2018).

17. DeVette, C. I. *et al.* NetH2pan: A Computational Tool to Guide MHC Peptide Prediction on Murine Tumors. *Cancer Immunol. Res.* **6,** 636–644 (2018).

18. Murphy, J. P. *et al.* MHC-I Ligand Discovery Using Targeted Database Searches of Mass Spectrometry Data: Implications for T-Cell Immunotherapies. *J. Proteome Res.* **16,** 1806–1816 (2017).

19. Bassani-Sternberg, M. & Gfeller, D. Unsupervised HLA Peptidome Deconvolution Improves Ligand Prediction Accuracy and Predicts Cooperative Effects in Peptide–HLA Interactions. *J. Immunol.* **197,** 2492–2499 (2016).

20. Andreatta, M., Lund, O. & Nielsen, M. Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. *Bioinformatics* **29,** 8–14 (2013).

21. Andreatta, M., Alvarez, B. & Nielsen, M. GibbsCluster: Unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Res.* **45,** W458–W463 (2017).

22. Ritz, D. *et al.* High-sensitivity HLA class I peptidome analysis enables a precise definition of peptide motifs and the identification of peptides from cell lines and patients' sera. *Proteomics* **16,** 1570–1580 (2016).

23. Bassani-Sternberg, M. *et al.* Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* **7,** 13404 (2016).

24. Sofron, A., Ritz, D., Neri, D. & Fugmann, T. High-resolution analysis of the murine MHC class II immunopeptidome. *Eur. J. Immunol.* **46,** 319–328 (2016).

25. Bassani-Sternberg, M. *et al.* Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLOS Comput. Biol.* **13,** e1005725 (2017).

26. Nielsen, M. & Lund, O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics* **10,** 296 (2009).

27. 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* **6,** (2011).

28. Hoof, I. *et al.* NetMHCpan, a method for MHC class i binding prediction beyond humans. *Immunogenetics* **61,** 1–13 (2009).

29. Karosiene, E., Lundegaard, C., Lund, O. & Nielsen, M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* **64,** 177–186 (2012).

30. Lund, O. *et al.* Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics* **55,** 797–810 (2004).

31. Vasoya, D. *et al.* Rapid identification of bovine MHCI haplotypes in genetically divergent cattle populations using next-generation sequencing. *Immunogenetics* **68,** 765–781 (2016).

32. Nene, V. *et al.* Designing bovine T cell vaccines via reverse immunology. *Ticks Tick. Borne. Dis.* **3,** 188–192 (2012).

33. Bjerregaard, A.-M. *et al.* An Analysis of Natural T Cell Responses to Predicted Tumor Neoepitopes. *Front. Immunol.* **8,** 1566 (2017).

34. Hart, J., MacHugh, N. D., Sheldrake, T., Nielsen, M. & Morrison, W. I. Identification of immediate early gene products of bovine herpes virus 1 (BHV-1) as dominant antigens recognized by CD8 T cells in immune cattle. *J. Gen. Virol.* **98,** 1843–1854 (2017).

35. Nielsen, M. & Andreatta, M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.* **8,** (2016).

36. Jensen, K. K. *et al.* Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* **154,** 394–406 (2018).

37. Sidney, J. *et al.* Citrullination only infrequently impacts peptide binding to HLA class II MHC. *PLoS One* **12,** e0177140 (2017).

38. Pearson, H. *et al.* MHC class I–associated peptides derive from selective regions of the human genome. *J. Clin. Invest.* **126,** 4690–4701 (2016).

39. Shraibman, B., Kadosh, D. M., Barnea, E. & Admon, A. Human Leukocyte Antigen (HLA) Peptides Derived from Tumor Antigens Induced by Inhibition of DNA Methylation for Development of Drug-facilitated Immunotherapy. *Mol. Cell. Proteomics* **15,** 3058–3070 (2016).

40. Mommen, G. P. M. *et al.* Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). *Proc. Natl. Acad. Sci.* **111,** 4507–4512 (2014).

41. Gloger, A., Ritz, D., Fugmann, T. & Neri, D. Mass spectrometric analysis of the HLA class I peptidome of melanoma cell lines as a promising tool for the identification of putative tumor-associated HLA epitopes. *Cancer Immunol. Immunother.* **65,** 1377–1393 (2016).

42. Vita, R. *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gky1006

43. Clement, C. C. *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.* **291,** 5576–5595 (2016).

44. Bergseng, E. *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* **67,** 73–84 (2015).

45. Heyder, T. *et al.* Approach for Identifying Human Leukocyte Antigen (HLA)-DR Bound Peptides from Scarce Clinical Samples. *Mol. Cell. Proteomics* **15,** 3017–3029 (2016).

46. Nelde, A. *et al.* HLA ligandome analysis of primary chronic lymphocytic leukemia (CLL) cells under lenalidomide treatment confirms the suitability of lenalidomide for combination with T-cell-based immunotherapy. *Oncoimmunology* **7,** e1316438 (2018).

47. Ooi, J. D. *et al.* Dominant protection from HLA-linked autoimmunity by antigen-specific regulatory T cells. *Nature* **545,** 243–247 (2017).

48. Karunakaran, K. P. *et al.* Identification of MHC-Bound Peptides from Dendritic Cells Infected with *Salmonella enterica* Strain SL1344: Implications for a Nontyphoidal *Salmonella* Vaccine. *J. Proteome Res.* **16,** 298–306 (2017).

49. Ting, Y. T. *et al.* The interplay between citrullination and HLA-DRB1 polymorphism in shaping peptide binding hierarchies in rheumatoid arthritis. *J. Biol. Chem.* **293,** 3236–3251 (2018).

50. Wang, Q. *et al.* Immunogenic HLA-DR-Presented Self-Peptides Identified Directly from Clinical Samples of Synovial Tissue, Synovial Fluid, or Peripheral Blood in Patients with Rheumatoid Arthritis or Lyme Arthritis. *J. Proteome Res.* **16,** 122–136 (2017).

51. Tretina, K., Pelle, R. & Silva, J. C. Cis regulatory motifs and antisense transcriptional control in the apicomplexan Theileria parva. *BMC Genomics* **17,** 128 (2016).

52. Zhang, J. *et al.* PEAKS DB: *De Novo* Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. *Mol. Cell. Proteomics* **11,** M111.010587 (2012).

53. Alvarez, B., Barra, C., Nielsen, M. & Andreatta, M. Computational Tools for the Identification and Interpretation of Sequence Motifs in Immunopeptidomes. *Proteomics* **18,** 1700252 (2018).

54. Nielsen, M., Lundegaard, C. & Lund, O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics* **8,** 238 (2007).

55. Thomsen, M. C. F. & 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.* **40,** 281–287 (2012).

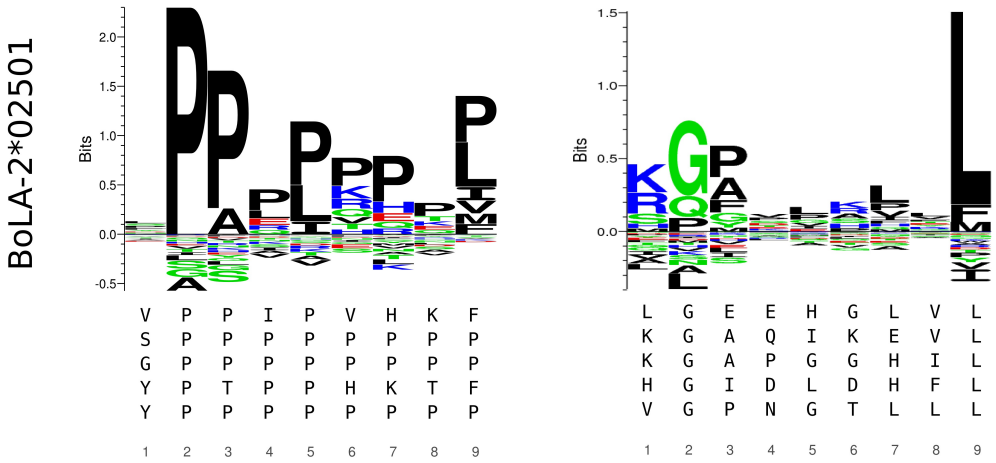
56. Hansen, A. M. *et al.* Characterization of binding specificities of bovine leucocyte class I molecules: impacts for rational epitope discovery. *Immunogenetics* **66,** 705–718 (2014).

57. Pedersen, L. E. *et al.* Porcine major histocompatibility complex (MHC) class I molecules and analysis of their peptide-binding specificities. *Immunogenetics* **63,** 821–834 (2011).

58. Harndahl, M., Rasmussen, M., Roder, G. & Buus, S. Real-time, high-throughput measurements of peptide–MHC-I dissociation using a scintillation proximity assay. *J. Immunol. Methods* **374,** 5–12 (2011).

59. Rasmussen, M. *et al.* Uncovering the Peptide-Binding Specificities of HLA-C: A General Strategy To Determine the Specificity of Any MHC Class I Molecule. *J. Immunol.* **193,** 4790–4802 (2014).

Supplementary information



Supplementary Figure 1. Different binding motifs found for the BoLA-2*02501 molecule. Left: binding preference for this molecule found by the NNAlign_MA method. Right: motif logo found in our previous work[16]. The top five repeating binding cores present in each motif alignment are shown below each logo.

Supplementary Table 1. Multi Allele (MA) data summary for the HLA-I system.

Cell line ID	Positives	Negatives	HLA-A		HLA-B		HLA-C		Source PMID
CD165	5364	132869	HLA-A*02:05	HLA-A*24:02	HLA-B*15:01	HLA-B*50:01	HLA-C*03:03	HLA-C*06:02	28832583
CM467	7401	184646	HLA-A*01:01	HLA-A*24:02	HLA-B*13:02	HLA-B*39:06	HLA-C*06:02	HLA-C*12:03	
GD149	9756	208444	HLA-A*01:01	HLA-A*24:02	HLA-B*38:01	HLA-B*44:03	HLA-C*06:02	HLA-C*12:03	
MD155	4374	108036	HLA-A*02:01	HLA-A*24:02	HLA-B*15:01	HLA-B*18:01	HLA-C*03:03	HLA-C*07:01	
PD42	2577	48693	HLA-A*02:06	HLA-A*24:02	HLA-B*07:02	HLA-B*55:01	HLA-C*01:02	HLA-C*07:02	
RA957	11037	232658	HLA-A*02:20	HLA-A*68:01	HLA-B*35:03	HLA-B*39:01	HLA-C*04:01	HLA-C*07:02	
TIL1	5445	140312	HLA-A*02:01	HLA-A*02:01	HLA-B*18:01	HLA-B*38:01	HLA-C*05:01	-	
TIL3	8799	206212	HLA-A*01:01	HLA-A*23:01	HLA-B*07:02	HLA-B*15:01	HLA-C*12:03	HLA-C*14:02	
Apher1	6145	123349	HLA-A*03:01	HLA-A*29:0	HLA-B*44:02	HLA-B*44:03	HLA-C*12:03	HLA-C*16:01	
Apher6	1962	39798	HLA-A*02:01	HLA-A*03:01	HLA-B*07:02	-	HLA-C*07:02	-	
Mel-15	21813	395324	HLA-A*03:01	HLA-A*68:01	HLA-B*27:05	HLA-B*35:03	HLA-C*02:02	HLA-C*04:01	27869121

Mel-16	11980	264233	HLA-A*01:01	HLA-A*24:02	HLA-B*07:02	HLA-B*08:01	HLA-C*07:01	HLA-C*07:02	
Mel-12	3758	88425	HLA-A*01:01	HLA-A*01:01	HLA-B*08:01	-	HLA-C*07:01	-	
Mel-8	6251	119000	HLA-A*01:01	HLA-A*03:01	HLA-B*07:02	HLA-B*08:01	HLA-C*07:01	HLA-C*07:02	
Mel-5	4749	106896	HLA-A*01:01	HLA-A*25:01	HLA-B*08:01	HLA-B*18:01	-	-	
Fibroblast	5289	122127	HLA-A*03:01	HLA-A*23:01	HLA-B*08:01	HLA-B*15:18	HLA-C*07:02	HLA-C*07:04	
HCC1143	2780	69565	HLA-A*31:01	-	HLA-B*35:08	HLA-B*37:01	HLA-C*04:01	HLA-C*06:02	
HCC1937	4976	102331	HLA-A*23:01	HLA-A*24:02	HLA-B*07:02	HLA-B*40:01	HLA-C*03:04	HLA-C*07:02	25576301
HCT116	4174	93208	HLA-A*01:01	HLA-A*02:01	HLA-B*45:01	HLA-B*18:01	HLA-C*05:01	HLA-C*07:01	
JY	2868	60863	HLA-A*02:01	-	HLA-B*07:02	-	HLA-C*07:02	-	
Bcell	12199	220971	HLA-A*01:01	HLA-A*03:01	HLA-B*07:02	HLA-B*27:05	HLA-C*02:02	HLA-C*07:01	24616531
Mel-624	2375	49050	HLA-A*02:01	HLA-A*03:01	HLA-B*07:02	HLA-B*14:01	HLA-C*07:02	HLA-C*08:01	27600516
SK-Mel-5	3293	64537	HLA-A*02:01	HLA-A*11:01	HLA-B*40:01	-	HLA-C*03:03	-	
HEK293	4972	86634	HLA-A*03:01	-	HLA-B*07:02	-	HLA-C*07:02	-	
MAVER-1	7403	171783	HLA-A*24:02	HLA-A*26:01	HLA-B*38:01	HLA-B*44:02	HLA-C*05:01	HLA-C*12:03	
HL-60	6607	115694	HLA-A*01:01	-	HLA-B*57:01	-	HLA-C*06:02	-	27600516
RPMI8226	4524	113201	HLA-A*30:01	HLA-A*68:02	HLA-B*15:03	HLA-B*15:10	HLA-C*02:10	HLA-C*03:04	
THP-1	5542	142866	HLA-A*02:01	HLA-A*24:02	HLA-B*15:11	HLA-B*35:01	HLA-C*03:03	-	
CA46	2324	62647	HLA-A*26:03	-	HLA-B*27:04	-	HLA-C*12:02	-	
LNT-229	10311	177908	HLA-A*03:01	-	HLA-B*35:01	-	HLA-C*04:01	-	
T98G	10011	216072	HLA-A*02:01	-	HLA-B*39:06	-	HLA-C*07:02	-	27412690
U-87	11585	241396	HLA-A*02:01	-	HLA-B*44:02	-	HLA-C*05:01	-	
pat-AC2	1369	32168	HLA-A*03:01	HLA-A*32:01	HLA-B*27:05	HLA-B*45:01	-	-	
pat-C	7928	149757	HLA-A*02:01	HLA-A*03:01	HLA-B*07:02	-	HLA-C*07:02	-	
pat-CELG	3814	72328	HLA-A*02:01	HLA-A*24:02	HLA-B*15:01	HLA-B*73:01	HLA-C*03:03	HLA-C*15:05	
pat-CP2	1790	36895	HLA-A*11:01	-	HLA-B*14:02	HLA-B*44:02	-	-	
pat-FL	3629	74392	HLA-A*03:01	HLA-A*11:01	HLA-B*44:03	HLA-B*50:01	-	-	
pat-J	5205	93828	HLA-A*02:01	HLA-A*03:01	HLA-B*07:02	-	HLA-C*07:02	-	
pat-JPB3	1937	35295	HLA-A*02:01	HLA-A*11:01	HLA-B*27:05	HLA-B*56:01	-	-	
pat-JT2	1467	29587	HLA-A*11:01	-	HLA-B*18:03	HLA-B*35:01	-	-	
pat-M	6959	136612	HLA-A*03:01	HLA-A*29:02	HLA-B*08:01	HLA-B*44:03	HLA-C*07:01	HLA-C*16:01	
pat-MA	3682	69891	HLA-A*02:01	HLA-A*29:02	HLA-B*44:03	HLA-B*57:01	HLA-C*07:01	HLA-C*16:01	27841757
pat-ML	3139	55262	HLA-A*02:01	HLA-A*11:01	HLA-B*40:01	HLA-B*44:03	-	-	
pat-NS2	636	15212	HLA-A*02:01	-	HLA-B*13:02	HLA-B*41:01	-	-	
pat-NT	2190	53238	HLA-A*01:01	HLA-A*32:01	-	-	-	-	
pat-PF1	4646	86859	HLA-A*01:01	HLA-A*02:01	HLA-B*07:02	HLA-B*44:03	HLA-C*07:02	HLA-C*16:01	
pat-R	4789	96011	HLA-A*03:01	HLA-A*29:02	HLA-B*08:01	HLA-B*44:03	HLA-C*07:01	HLA-C*16:01	
pat-RT	2537	49846	HLA-A*01:01	HLA-A*02:01	HLA-B*18:01	HLA-B*39:24	HLA-C*05:01	HLA-C*07:01	
pat-SR	2632	57417	HLA-A*02:01	HLA-A*23:01	HLA-B*18:01	HLA-B*44:03	-	-	
pat-ST	1256	26963	HLA-A*03:01	HLA-A*24:02	HLA-B*07:02	HLA-B*27:05	-	-	

“Positives” and “Negatives” refer to the number of positive and negative instances contained in each cell line data. Further rows show the HLA-A, HLA-B and HLA-C expressed by a given cell line, together with the Source ID for its corresponding dataset.

Supplementary Table 2. Multi Allele (MA) data summary for the HLA-II system.

Cell line ID	Positives	Negatives	HLA-DRB		Source PMID
PAT1	670	7740	DRB1*04:01	DRB1*07:01	27452731
PAT2	213	2520	DRB1*04:01	DRB1*13:01	
PAT3	145	1710	DRB1*07:01	DRB1*13:01	
PAT4	65	585	DRB1*13:01	DRB1*14:01	
PAT5	129	1305	DRB1*08:01	DRB1*13:01	
PAT7	187	2790	DRB1*03:01	DRB1*15:01	
RA1	210	2475	DRB1*04:02	DRB1*11:04	27726376
RA2	231	2430	DRB1*01:01	DRB1*04:01	
RA3	278	2925	DRB1*01:01	DRB1*04:01	
RA4	51	480	DRB1*04:01	DRB1*15:01	
RA5	121	1440	DRB1*08:01	DRB1*15:01	
LA1	240	2655	DRB1*01:01	DRB1*08:01	
LA2	118	1215	DRB1*10:01	DRB1*11:01	
LA3	408	4140	DRB1*03:01	DRB1*15:01	
LA4	257	2835	DRB1*11:01	DRB1*13:02	
LA5	327	3285	DRB1*03:01	DRB1*03:05	
LA6	200	1890	DRB1*01:03	DRB1*03:01	
LA7	682	7335	DRB1*04:03	DRB1*15:01	
LA8	1772	17460	DRB1*04:01	DRB1*10:01	
UPN1	3080	29745	DRB1*08:03	DRB1*13:01	29632711
UPN2	367	3555	DRB1*0801	DRB1*13:01	
UPN3	3216	29565	DRB1*07:01	DRB1*08:01	
DR15_human	2426	22365	DRB1*15:01	DRB5*01:01	28467828

“Positives” and “Negatives” refer to the number of positive and negative instances contained in each cell line data. Further rows show the HLA-DRB alleles expressed by a given cell line, together with the Source PMID for its corresponding dataset.

Supplementary Table 3. Multi Allele (MA) data summary for the BoLA system.

Cell line ID	Haplotype	Positives	Negatives	BoLA-1	BoLA-2	BoLA-3	BoLA-4	BoLA-6	Source PMID
2123	A12/A15	11908	310336	BoLA-1*01901	BoLA-2*00801	-	BoLA-4*02401	-	-
				BoLA-1*00901	BoLA-2*02501				
5072	A11	7469	153959	-	BoLA-2*01801	BoLA-3*01701	-	-	
2824	A19	9665	175580	-	BoLA-2*01601	-	-	BoLA-6*01402	
5350	A20	11791	274531	-	BoLA-2*02601	BoLA-3*02701	-	-	
2408	A15	12523	320919	BoLA-1*00901	BoLA-2*02501		BoLA-4*02401	-	
1011/500004	A10	10188	148801	-	BoLA-2*01201	BoLA-3*00201	-	-	29115832
641	A18	6615	80170	-	-	-	-	BoLA-6*01301	
2229/104003	A14	9509	186084	BoLA-1*02301	BoLA-2*02501	-	BoLA-4*02401	BoLA-6*04001	

“Positives” and “Negatives” refer to the number of positive and negative instances contained in each cell line data. Further rows show the BoLA-1, BoLA-2, BoLA-3, BoLA-4, and BoLA-6 alleles expressed by a given cell line, together with the Source PMID for its corresponding dataset. Allele annotation was obtained from Vasoya, D. et al.[31]

Supplementary Table 4. Cell lines with atypical HLA-A or HLA-B peptidome repertoire profiles.

Cell line ID	HLA-A	HLA-B	HLA-C	Source PMID(s)
HL-60	0.046	0.856	0.100	26992070 17083564
CA46	0.066	0.741	0.193	26992070
Mel-624	0.931	0.041	0.028	7541714
HEK293	0.892	0.077	0.031	26258424

For each cell line, the relative peptidome size for the HLA-A, HLA-B and HLA-C loci is given. Peptidome sizes were calculated as described in Figure 3 B. The last column shows the references to earlier publications describing the loss of the locus for a given cell line.

Supplementary Table 5. AUC0.1 performance values for the SA molecules left out from the training of the A2 and A3 specificity reduced SA and MA models.

	SA MODEL	MA MODEL
HLA-A*02:01	0.371	0.762
HLA-A*02:05	0.550	0.759
HLA-A*02:07	0.521	0.869
HLA-A*03:01	0.479	0.886
HLA-A*11:01	0.482	0.861
HLA-A*26:01	0.949	0.961
HLA-A*29:01	0.867	0.879
HLA-A*31:01	0.356	0.719
HLA-A*32:01	0.911	0.896

For details refer to the text. Note, that not all the molecules included in the evaluation are part of the A2 and A3 supertypes; these molecules are included because they have a distance to the A2 and A3 molecules in the MA dataset less than 0.2.